POLITECNICO DI TORINO
Repository ISTITUZIONALE

Star-shaped space of solutions of the spherical negative perceptron

*Original*
Star-shaped space of solutions of the spherical negative perceptron / Livio Annesi, Brandon; Lauditi, Clarissa; Lucibello, Carlo; Malatesta, Enrico M.; Perugini, Gabriele; Pittorino, Fabrizio; Saglietti, Luca. - In: PHYSICAL REVIEW LETTERS. - ISSN 1079-7114. - 131:22(2023).

*Availability:*
This version is available at: 11583/2983674 since: 2023-11-08T22:13:22Z

*Publisher:*
APS

*Published*
DOI:

*Terms of use:*

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*
APS postprint/Author's Accepted Manuscript e postprint versione editoriale/Version of Record

(Article begins on next page)

23 November 2024

Editors' Suggestion

# Star-Shaped Space of Solutions of the Spherical Negative Perceptron

Brandon Livio Annesi,[1] Clarissa Lauditi,[2] Carlo Lucibello,[1,3] Enrico M. Malatesta,[1,3] Gabriele Perugini,[1] Fabrizio Pittorino,[4,3] and Luca Saglietti[1,3]

[1]*Department of Computing Sciences, Bocconi University, 20136 Milano, Italy*
[2]*Department of Applied Science and Technology, Politecnico di Torino, 10129 Torino, Italy*
[3]*Bocconi Institute for Data Science and Analytics, 20136 Milano, Italy*
[4]*Department of Electronics, Information, and Bioengineering, Politecnico di Milano, 20125 Milano, Italy*

(Received 17 May 2023; revised 5 September 2023; accepted 8 November 2023; published 29 November 2023)

Empirical studies on the landscape of neural networks have shown that low-energy configurations are often found in complex connected structures, where zero-energy paths between pairs of distant solutions can be constructed. Here, we consider the spherical negative perceptron, a prototypical nonconvex neural network model framed as a continuous constraint satisfaction problem. We introduce a general analytical method for computing energy barriers in the simplex with vertex configurations sampled from the equilibrium. We find that in the overparametrized regime the solution manifold displays simple connectivity properties. There exists a large geodesically convex component that is attractive for a wide range of optimization dynamics. Inside this region we identify a subset of atypical high-margin solutions that are geodesically connected with most other solutions, giving rise to a star-shaped geometry. We analytically characterize the organization of the connected space of solutions and show numerical evidence of a transition, at larger constraint densities, where the aforementioned simple geodesic connectivity breaks down.

DOI: 10.1103/PhysRevLett.131.227301

In constraint satisfaction problems, the goal is to find a configuration of the $N$ variables that satisfies a system of constraints. In the case of random instances [1] and for large size $N$, one can typically identify sharp "structural" phase transitions in the geometrical organization of the solution space [2–4]. In the past decades, statistical physics methods from spin glass theory [5] have been successfully employed to investigate the impact of these landscape features on the performance of solution-sampling algorithms [6]. A deeper understanding of this interplay in the case of continuous variables [7] is becoming a crucial prerequisite for the study of learning dynamics in neural networks.

The characterization of the manifold of low-energy lying states in neural networks has become one of the central theoretical questions of the field [8]. In typical setups, the high degree of overparametrization of the models guarantees the existence of multiple zero-energy configurations, but different local geometries induce vastly different accessibility and generalization properties [9,10]. Growing theoretical [11–16] and empirical [17–20] evidence seems to show that there exist flat degenerate areas in the landscape of neural networks. The dynamics of common stochastic gradient descent (SGD) based algorithms seem to be quickly attracted to the borders of these regions and then drift toward their core [21–23].

Linear paths between two minimizers (e.g., as given by SGD with different random initial conditions) display energy barriers. Nonetheless, zero-energy paths can be systematically constructed between them [24,25]. This surprising finding on *mode connectivity* is compatible with the hypothesis of the existence of a single connected component of zero-energy configurations, organized in an intricate network of tunnels and plateaus [26–28]. Understanding the extent of linear mode connectivity may unlock progress in some of the most debated topics in deep learning, from the "lottery ticket" hypothesis and pruning [29,30] to multitask and continual learning and ensemble methods [31,32].

Here, we consider the simplest nonconvex neural network model, the negative spherical perceptron [33–36], and characterize the connectivity properties of its solution space via *geodesic* (minimum length) paths on the high-dimensional sphere. In particular, we introduce a novel analytical method, based on the replica analysis of the model [5], that yields the typical energy barriers in the convex hull of a group of $y$ solutions sampled from the zero-energy manifold. We find that, in the low constraint density regime, the domain of solutions is *star-shaped* [40]: almost all solutions are geodesically connected through zero-energy paths to a subset of atypical high-margin solutions. This subset, which we call the "kernel," is nested in the core of the largest geodesically convex component of the solution manifold. A sketch of this geometrical organization is shown in Fig. 1.

We empirically investigate the behavior of different classes of solvers and their bias toward different regions of the solution manifold. In particular, we find that the

 © 2023 American Physical Society

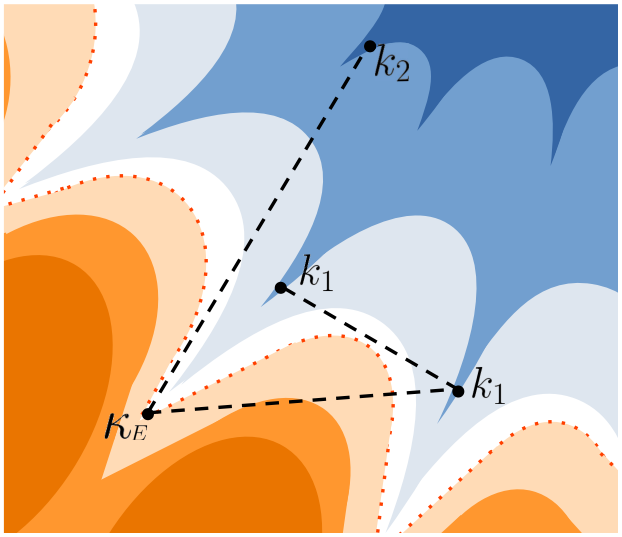


FIG. 1. Sketch of the solution space of the negative perceptron in the RS phase. The red dotted line represents the border of the connected manifold of solutions for a given margin $\kappa_E$ (white-blue region). In the orange regions, the configurations have nonzero energy. The solutions that satisfy margins larger than the one of the problem, $\kappa_E < k_1 < k_2$, are organized in a nested structure (darker shades of blue). When a typical solution with margin $\kappa_E$ is geodesically connected with a solution with margin $k_1$, an energy barrier (a crossing of the orange region) is observed. However, the $k_1$ solutions belong to a *geodesically convex* submanifold (the *geodesic path* falls within the white-blue region). Solutions with an even higher margin $k_2$, located in the *kernel*, are connected to almost any other solution. Definitions are provided in Appendix A.

dynamics of SGD with cross-entropy loss naturally flows toward the large convex component of the solution manifold. Moreover, for any solver, one can identify a phase at low constraint densities and up to a certain threshold, where the sampled solutions are geodesically connected to the most robust solutions (see below) of the problem. Above this threshold and up to the limit density for satisfiability, energy barriers are encountered, signaling the breakdown of the star-shaped organization of solutions. We compare these thresholds with the known structural transitions identified through the statistical physics analysis of this model [41].

*The model.*—The spherical perceptron is defined by $N$ weights $W_i \in \mathbb{R}$, trained to satisfy an extensive number $P = \alpha N$ of constraints

$$\Delta^\mu \equiv \boldsymbol{W} \cdot \boldsymbol{\xi}^\mu \geq \kappa_E \sqrt{N}, \qquad \mu \in [P], \tag{1}$$

where $\kappa_E$ is the margin of the problem, the $\Delta^\mu$ are called stabilities, and $\boldsymbol{\xi}^\mu \sim \mathcal{N}(0, I_N)$. The weights are also subject to the spherical constraint $\|\boldsymbol{W}\|^2 = N$. We analyze the large $N$ limit at constant $\alpha$. When a negative margin $\kappa_E < 0$ is considered, the so-called negative perceptron, linear separability of the dataset is not a necessary condition for satisfiability (SAT) and the problem is nonconvex. The negative perceptron has recently received attention in both the physics [35,41,42] and mathematics communities [43,44]. A detailed analysis of the different structural transitions affecting the solution space, as $\kappa_E$ and $\alpha$ are increased, shows that the model enters a nonergodic phase with replica symmetry breaking [details on the phase diagram in Supplemental Material [36] (SM)]. In this Letter, we further investigate the geometric properties of the ground states in the region below the critical line $\alpha_{\mathrm{dAT}}(\kappa_E)$ where replica symmetry (RS) holds.

*Organization of the solutions.*—In the RS phase, the problem is SAT with high probability for large $N$. We consider the uniform probability density over the solutions,

$$p_{\boldsymbol{\xi},\kappa_E}(\boldsymbol{W}) = \frac{1}{Z_{\boldsymbol{\xi},\kappa_E}} \delta(\|\boldsymbol{W}\|^2 - N) \prod_{\mu=1}^{P} \Theta(\boldsymbol{W} \cdot \boldsymbol{\xi}^\mu - \kappa_E \sqrt{N}), \tag{2}$$

where the partition function $Z_{\boldsymbol{\xi},\kappa_E}$ plays the role of a normalization factor. While the typical solutions obtained by sampling from (2) have minimum stabilities exactly equal to $\kappa_E$, the solution space also contains an exponential number of atypical solutions that satisfy the constraints (1) with a larger margin $k$ [45], with $\kappa_E < k \leq \kappa_{\max}(\alpha)$, where $\kappa_{\max}(\alpha)$ represents the SAT-UNSAT transition line. In order to characterize the geometry of the solution space, for a given sample $\boldsymbol{\xi}$ we consider two configurations independently sampled from $p_{\boldsymbol{\xi},k_1}$ and $p_{\boldsymbol{\xi},k_2}$, respectively, and employ the replica method [46] to compute their overlaps $q_1 = (1/N)\mathbb{E}\langle \boldsymbol{W}^1 \cdot \boldsymbol{W}^2 \rangle_{k_1,k_1}$, $p = (1/N)\mathbb{E}\langle \boldsymbol{W}^1 \cdot \boldsymbol{W}^2 \rangle_{k_1,k_2}$ and $q_2 = (1/N)\mathbb{E}\langle \boldsymbol{W}^1 \cdot \boldsymbol{W}^2 \rangle_{k_2,k_2}$. Here, $\langle \cdot \rangle_{k,k'}$ represents the average over the Cartesian product of densities (2) with the corresponding margins and $\mathbb{E}$ is the expectation over disorder $\boldsymbol{\xi}$ (see SM). For $k_1 < k_2$, we find that they satisfy the simple inequality $q_1 < p < q_2$. This ordering is compatible with a nested organization of solutions with different margins. The degree of anisotropy of this structure can be evaluated analytically (details in the SM).

*Interpolating between solutions.*—Our main analytic result is a formula for studying the typical energy landscape between groups of $y$ solutions. In particular, we consider the projection on the $N$ sphere of the $(y-1)$-simplex,

$$\boldsymbol{W}_{\boldsymbol{\gamma}} = \frac{\sqrt{N} \sum_{r=1}^{y} \gamma_r \boldsymbol{W}^r}{\| \sum_{r=1}^{y} \gamma_r \boldsymbol{W}^r \|}, \tag{3}$$

with $\gamma_r \geq 0$ and $\sum_{r=1}^{y} \gamma_r = 1$. By varying the margins $\{k_r\}_{r=1}^{y}$ of the solutions $\{\boldsymbol{W}^r\}_{r=1}^{y}$ on the vertices, different regions of the solution manifold can be explored. To obtain the asymptotic energy of the interpolating configurations with respect to the margin $\kappa_E$ of the problem, we evaluate the probability of a constraint violation:

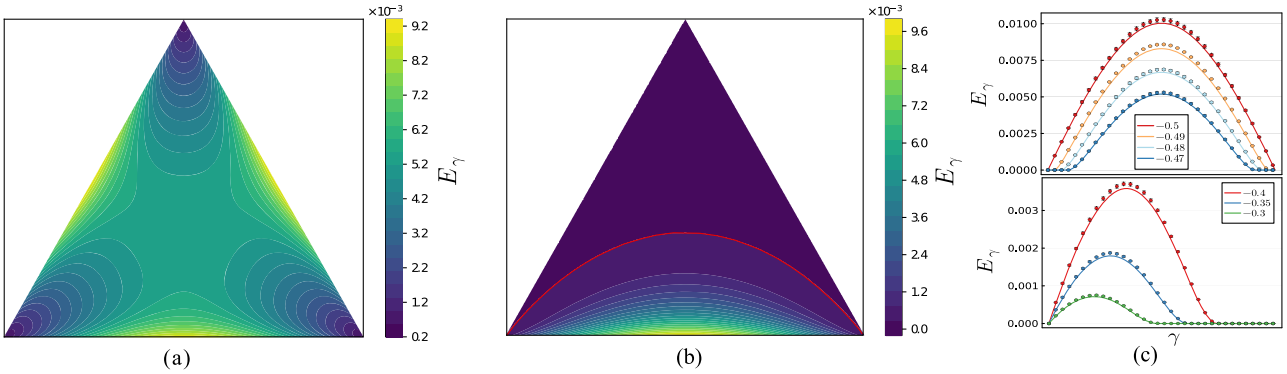


FIG. 2. Interpolating manifold between triplets of solutions for $\kappa_E = -0.5$, $\alpha = 1$. (a) Solutions with the same margin $k = \kappa_E$. (b) Solutions with different margins. The two bottom vertices are two typical solutions to the problem, i.e., $k_1 = \kappa_E$; the top vertex is sampled with $k_2 = -0.1 > \kappa_{\rm krn}(\kappa_E, \alpha) \simeq -0.171$. The red level curve separates the zero from the nonzero energy region on the simplex. (c-top) Energy along the geodesic connecting two solutions sampled with margin $k$ for $k = \kappa_E, -0.49, -0.48, -0.47$. (c-bottom) Same as the top panel but with the left endpoint margin fixed to $k_1 = \kappa_E$ and the one on the right having margin $k_2 = -0.4, -0.35, -0.3$. Lines are the theoretical predictions, dots are large-$N$ extrapolations of the numerical simulations.

$$E_{\boldsymbol{\gamma}} = \lim_{N\to+\infty} \mathbb{E}_{\boldsymbol{\xi}} \langle \Theta(-\boldsymbol{W}_{\boldsymbol{\gamma}} \cdot \boldsymbol{\xi}^\mu + \kappa_E \sqrt{N}) \rangle_{k_1,\ldots,k_y}. \quad (4)$$

In the high dimensional limit, this quantity only depends on the typical overlaps between pairs of solutions with different margins $q_{rs}$, $r, s \in [y]$. Analytic details are reported in Appendix B. With a similar approach, one can also derive the stability distribution $\Delta^\mu$ in Eq. (1) for the interpolating configurations (details in the SM).

*The largest geodesically convex component.*—We first consider the case where the vertices of the simplex are all sampled with identical margin $k_r = k$, with $\kappa_E \le k < \kappa_{dAT}(\alpha)$. One finds that the energy on the projected $(y-1)$-simplex is always strictly greater than zero when $k = \kappa_E$, while extended regions around each vertex fall to zero energy for $k > \kappa_E$. For each value of $y$ one can identify a "coalescence threshold," $\kappa_y^\star(\kappa_E, \alpha)$, corresponding to the value of the margin above which the entire $(y-1)$-simplex lies at zero energy. In particular, we find the minimum margin $\kappa_2^\star(\kappa_E, \alpha)$ that ensures *linear mode connectivity*. These thresholds are displayed in Fig. 3 as a function of $\alpha$ for $\kappa_E = -0.5$, and satisfy $\kappa_2^\star < \kappa_3^\star < \ldots < \kappa_\infty^\star$. Above the last *coalescence threshold*, $\kappa_\infty^\star$, the projected convex hull of the entire ensemble of solutions lies at zero energy: this is what we call geodesically convex component of the manifold of solutions (see also Appendix A). The size of this region can be bounded by the typical overlap $q$ between $\kappa_\infty^\star$ solutions. By inspecting the distribution of stabilities across the zero-energy manifold, one finds that the geodesic paths encounter different solutions from those of the equilibrium description at the corresponding margin (details in the SM).

In Fig. 2(a), we plot $E_{\boldsymbol{\gamma}}$ on the two-dimensional simplex, with all vertices at $k = \kappa_E$. Since the maximum energy barriers are located on the edges of the projected simplex, the minimum energy path connecting the corner solutions needs to deviate through its barycenter. Notice that, since $\kappa_2^\star < \kappa_3^\star$, as $k$ is increased from $k = \kappa_E$, the energy barriers along the edges go to zero faster than the energy at the center of the simplex. At the top of panel (c), we show how the barrier on the edges goes to zero as the margin of the vertices is increased.

*The kernel of the solution space.*—We now focus on the connectivity of solutions with different margins. Specifically, we start by considering the geodesic path between a typical solution, $k_1 = \kappa_E$, and a higher margin solution, $k_2 > \kappa_E$. One can show that, for any $(\kappa_E, \alpha)$ below the dAT transition, there exists a threshold $\kappa_{\rm krn}$ such that, with high probability, no energy barrier is encountered along the geodesic path between any solution with $k_2 \ge \kappa_{\rm krn}$ and any other typical solution with margin $k_1 \ge \kappa_E$. These findings imply that the solution space is *star-shaped*, and allow us to identify the *kernel* of solution space, i.e., a subset of solutions that are "visible," through geodesic paths, from any typical point of the solution manifold. In the bottom of Fig. 2(c), we show the decrease of the energy barrier with $k_2$.

In Fig. 3(a), we display the line $\kappa_{\rm krn}(\alpha)$ for a problem with margin $\kappa_E = -0.5$. Notice that its continuation above the dAT line (dashed), where the RS results are incorrect, would predict an intersection between the $\kappa_{\rm krn}$ and the $\kappa_{\rm max}$ lines, implying a breakdown of the star-shaped property. We revert to numerical experiments, in the next section, to understand what happens to the connectivity of typical solutions in this phase.

The analysis of the connectivity of solutions with different margins can be carried out also with $y > 2$. In Fig. 2(b), we display the $y = 3$ case, with the bottom vertices being typical solutions with margin $\kappa_E$ and the upper vertex having a margin larger than $\kappa_{\rm krn}$. As expected from the $y = 2$ analysis, by deviating through the core of the solution manifold, one can construct a piecewise

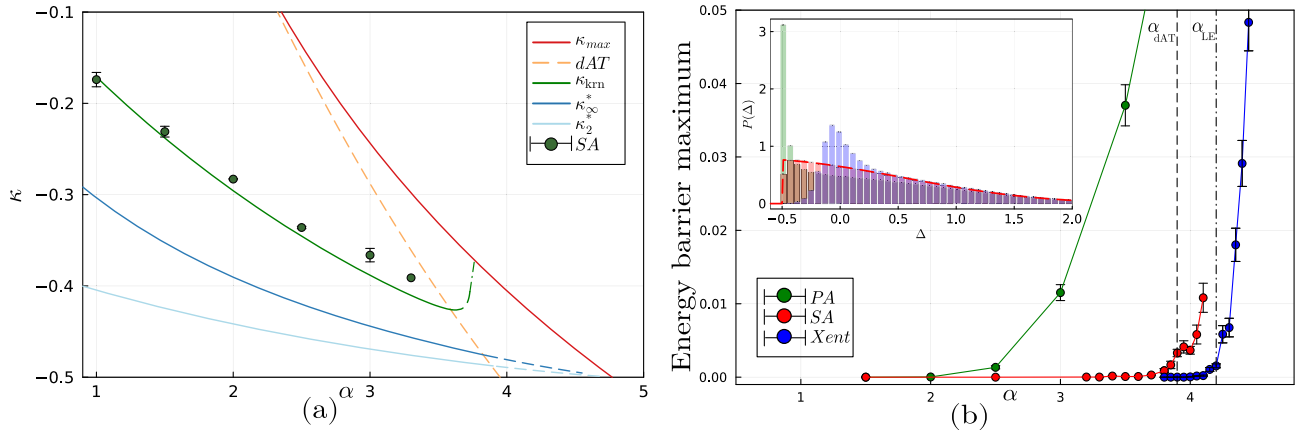


FIG. 3. (a) Coalescence threshold lines at $\kappa_E = -0.5$ (blue and cyan), and the $\kappa_{\text{krn}}$ threshold (green line) as a function of $\alpha$. In orange the dAT transition line, delimiting the RS-stable region; in red the RS estimate of the $\kappa_{\text{max}}$. Above the dAT line, dashed lines indicate the continuation of the (unstable) RS predictions. Points are numerical extrapolations from SA samples (see SM for details). (b) Maximum error along the geodesic path ($y = 2$) connecting numerical solutions found with different algorithms (PA, SA, and Xent) with the fBP max-margin solutions. Nonzero energies along the path indicate disconnection in the solution space. The vertical dashed lines denote the values of $\alpha_{dAT}$ and $\alpha_{LE}$ at $\kappa_E = -0.5$. The inset shows stability distributions for PA, SA, and Xent at $\alpha = 2$ compared with the theoretical stability distribution of typical solutions (red dashed line).

geodesic path between *any* pair of solutions lying exactly at zero energy.

*Sampling bias and disconnection.*—We compare the properties of solutions found with different solvers on instances of the negative perceptron with $\kappa_E = -0.5$. In particular, in Fig. 3(b), we characterize their geodesic connectivity to solutions located in the kernel region, as a function of $\alpha$. Note that, because of the nested overlap structure, we expect the maximum-margin solutions of the problem to be located in the kernel. Therefore, for obtaining them we employ the focusing Belief Propagation (fBP) algorithm, which was shown in [41] to yield good proxies of the $\kappa_{\text{max}}$ solutions.

Typical solutions instead are approximated by carefully applying simulated annealing (SA) on the square hinge loss with margin $\kappa_E$ and are found to be in good agreement with the theory [cf. the points in Figs. 2(a), 2(b), and 3(a), and further experiments in the SM]. Nonzero energy barriers with the fBP solutions seem to appear in close proximity of the dAT transition line, confirming the star-shapedness of typical solutions in the RS stable region.

SGD on the cross-entropy loss—the most common optimization objective for this class of problems—yields *robust* solutions [14] with higher average stability than typical, as shown in the inset in Fig. 3(b). The geodesic path between independent optimization trajectories, starting from random initialization, shows no energy barriers as soon as the zero-energy region is accessed, revealing an algorithmic bias toward the geodesically convex component of the solution manifold (details in Appendix C and in SM). The disconnection transition with the core solutions [Xent in Fig. 3(b)] is delayed with respect to SA, and seems to happen in close proximity of the $\alpha_{LE}$ transition characterized in [41].

Finally, we implement the classic perceptron algorithm (PA). When the learning rate is sufficiently small, this algorithm is able to sample solutions with a large mass of stabilities at threshold [inset of Fig. 3(b)], and therefore less robust than typical ones. The disconnection with the core region of the solution manifold is in this case anticipated before the dAT line. Notice that this result is not incompatible with our predictions, since these solutions seem to be subdominant in the flat measure over solutions, and cannot be seen through an equilibrium analysis.

These numerical results are consistent with our theoretical picture of a star-shaped space of solutions in the overparametrized regime, and reveal a progressive disconnection transition that affects different types of solutions according to their degree of robustness.

*Discussion and conclusions.*—In this Letter, we characterized the connectivity properties of a prototypical model of nonconvex neural networks. The theoretical analysis unveiled the presence, in the overparametrized regime, of a connected manifold of solutions organized in a star-shaped structure. Similar types of structures have been shown to appear in completely unrelated high-dimensional problems [47,48]. We conjecture that simple mode connectivity may be a universal property of nonconvex optimization problems in the overparametrized regime. A promising future research direction would be to investigate analytically whether the star-shaped geometry, or a generalization thereof, holds in more complex [49] and more realistic models of neural networks [50,51].

With a precise picture in hand, we were also able to characterize where different solvers end up in the solution space. Understanding how algorithmic bias can be exploited to enhance learning performance is a central question in the field of deep learning. At the same time,

probing the landscape with different dynamics could help characterize the solution space in more complex settings, following up on [26,27].

*Appendix A: Definitions.*—Typical solutions: In high dimensions, due to entropic factors, independent sampling from a given measure will return with high probability configurations with specific shared properties. For instance, almost all the exponentially many solutions of the perceptron problem have the same stability distribution profile. In this work, we define independent identical distributed samples from measure (2) as *typical* solutions. The measure also contains different types of solutions that we call *atypical* and that are exponentially subdominant in number and therefore become statistically irrelevant when considering the high-dimensional limit. Some atypical solutions satisfy a higher margin constraint than the one of the problem, or more generally have a different stability profile (e.g., *robust* solutions sampled with SGD have higher average stability than typical).

Projection of the Euclidean $y$ simplex on the $N$-dimensional sphere—geodesic paths and polytopes: All the results presented in the Letter for the interpolation energy Eq. (3) are obtained first on the Euclidean simplex by interpolating configurations among $y$ normalized solutions, and then by normalizing all interpolated configurations, i.e., projecting them on the $N$-dimensional sphere. This gives rise, on the $N$-dimensional sphere, to geodesic paths for $y = 2$ and geodesic polytopes of dimension $y-1$ for $y > 2$ (their edges are themselves geodesic polytopes of dimension $y-2$). See Fig. 4 for an illustration of the projection of the straight path onto the geodesic for $y = 2$.

Geodesic connectivity and geodesic convexity: Given two points $\boldsymbol{x}_1, \boldsymbol{x}_2$ of a set $S$ in a Euclidean space, we say that $\boldsymbol{x}_1$ is connected to $\boldsymbol{x}_2$ via $S$ if the straight path $[\boldsymbol{x}_1, \boldsymbol{x}_2] = \{\gamma \boldsymbol{x}_1 + (1-\gamma)\boldsymbol{x}_2 : \gamma \in [0,1]\}$ is such that $[\boldsymbol{x}_1, \boldsymbol{x}_2] \subset S$. A set $C \subset \mathbb{R}^d$ is convex if all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in C$ are connected via $C$ [40]. Keeping in mind the projection operation of the $y-1$ simplex on the $N$-dimensional sphere illustrated in Fig. 4 for a geodesic path ($y = 2$), we can notice that *all notions of connectivity and convexity usually defined in the Euclidean space straightforwardly generalize to the N-sphere by substituting the concept of straight path with the one of geodesic*. Since our model is defined on the $N$ sphere, the concepts of connectivity and convexity that appear in the main text of the Letter are to be intended in the geodesic sense. Figure 4 also illustrates the notion of geodesic connectivity.

Star-shaped set, its kernel and the geodesically convex component: As a natural relaxation of the convexity notion, a set $S \subset \mathbb{R}^d$ is star-shaped if $\exists \boldsymbol{x}_1 \in S$ such that $\forall\, \boldsymbol{x}_2 \in S$ we have $[\boldsymbol{x}_1, \boldsymbol{x}_2] \subset S$. The *kernel* of a star-shaped set $S$ is defined as the set of all $\boldsymbol{x}_1 \in S$ such that $[\boldsymbol{x}_1, \boldsymbol{x}_2] \subset S\ \forall\, \boldsymbol{x}_2 \in S$. Its elements are called the star centers of $S$. The kernel of a star-shaped set is a convex set.

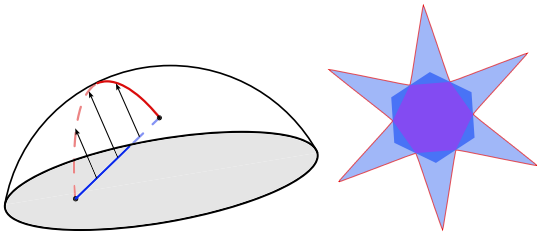

FIG. 4. Left: illustrative example of the projection of a straight path (blue) onto the geodesic (red) in the case of a one-dimensional path, i.e., $y = 2$. All points along the geodesic (or that can be connected by a geodesic) are said to be *geodesically connected*. Right: simple 2-dimensional visualization of the shape of the manifold of solutions. The large geodesically convex component, (darker blue) and its kernel (purple) are also shown.

The kernel of the star-shaped set of solutions of the negative perceptron considered in this Letter is represented by the set of solutions with $\kappa > \kappa_{\mathrm{krn}}$; see Fig. 3. In Fig. 4 we represent a simple two-dimensional star-shaped set, its kernel, the possible existence inside the star-shaped set of a convex set containing the kernel and distinct from it. This is called geodesically convex component in the main text.

*Appendix B: Details on the analytical computation for the connectivity thresholds.*—To compute the asymptotic training error $E_\gamma$ in Eq. (4) on the manifold spanned by $y$ independently sampled solutions, we resort to the replica trick [5]. The computation for the general case in which each sampled solution has a distinct margin $\{k_r\}_{r=1}^{y} \geq \kappa_E$ involves standard steps (see SM for details). The final expression of $E_\gamma$ can be factored into the product of two terms:

$$E_\gamma = \Theta(f_\gamma(\kappa_E, \{k_r\}_r)) I_\gamma(\kappa_E, \{k_r\}_r), \tag{B1}$$

where $f_\gamma(\kappa_E, \{k_r\}_r) = \kappa_E c_\gamma - \sum_r \gamma_r k_r$ is a linear function of the margin, with $c_\gamma = \sqrt{\sum_{rs} q_{rs}\gamma_r\gamma_s}$, and $I_\gamma$ is a non-negative function involving $2y$-dimensional Gaussian integral, whose expression is fully reported in SM. Since the expression evaluates to zero only if the argument of the Heaviside $\Theta$ function is negative, the sign of $f_\gamma$ is sufficient to analytically derive the connectivity thresholds reported in Fig. 3. It is useful to distinguish between cases. (1) All sampled solutions corresponding to the vertices of the $y-1$-dimensional simplex have the same margin $k_r = k \geq \kappa_E$, giving $f_\gamma = \kappa_E c_\gamma - k$. In this case, since all solutions are statistically equivalent, the only order parameter is the typical overlap between them $q = (1/N)\mathbb{E}\langle \boldsymbol{W}^1 \cdot \boldsymbol{W}^2 \rangle_{k,k}$. Consequently the norm of the interpolating vector is $c_\gamma = (1-q)\sum_r \gamma_r^2 + q \leq 1$. We consider two subcases. (a) In the case where all vertices are typical solutions with $\kappa \equiv \kappa_E$, since $c_\gamma \leq 1$ then $f_\gamma = \kappa_E c_\gamma - \kappa_E > 0$ when

$\kappa_E < 0$ and therefore every interpolated configuration has nonvanishing training error (Fig. 2). When $\kappa_E \geq 0$ instead, we have that $f_{\boldsymbol{\gamma}} \leq 0$ and therefore $E_{\boldsymbol{\gamma}} = 0$ for all $\boldsymbol{\gamma}$, consistently with the convexity of the solution space. (b) When $\kappa_E < \kappa < 0$, the minimum value of $c_{\boldsymbol{\gamma}}$ is attained on the barycenter $\gamma_r = (1/y)\ \forall\ r$. Hence, if the inequality $\kappa_E c_{\boldsymbol{\gamma}} - k < 0$ holds for the barycenter, all interpolated configurations have zero training error. The condition $\kappa > \kappa_E c_{\text{barycenter}} = \kappa_E \sqrt{(1-q)(1/y)+q} \equiv \kappa_y^\star$ defines the coalescence thresholds $\kappa_2^\star < \kappa_3^\star < \ldots < \kappa_\infty^\star = \kappa_E\sqrt{q}$ (see Fig. 3) indicating that for $\kappa > \kappa_y^\star$ the normalized $y-1$-simplex lies at zero error. Their ordering in $y$ is due to the fact that the overlap $q$ is an increasing monotonic function of the margin $\kappa$. (2) Solutions sampled with different margins. If $y-1$ vertices have margins $k_1$ and one vertex has margin $k_2$, with $k_2 > k_1 \geq \kappa_E$, then the overlap matrix will depend on $q_1 = (1/N)\mathbb{E}\langle \boldsymbol{W}^1 \cdot \boldsymbol{W}^2 \rangle_{k_1,k_1}$, $q_2 = (1/N)\mathbb{E}\langle \boldsymbol{W}^1 \cdot \boldsymbol{W}^2 \rangle_{k_2,k_2}$ and $p = (1/N)\mathbb{E}\langle \boldsymbol{W}^1 \cdot \boldsymbol{W}^2 \rangle_{k_1,k_2}$. In this case we can rewrite $f_{\boldsymbol{\gamma}} = \kappa_E c_{\boldsymbol{\gamma}} - k_1 \sum_{r=1}^{y-1} \gamma_r - k_2 \gamma_y$ and study it analytically when an explicit algebraic relation can be derived. For $y = 2$ (and by using the explicit expression of the norm $c_{\boldsymbol{\gamma}}$, see SM for details) we can find the minimum margin $k_2$ that should be imposed on $\boldsymbol{W}^2$ given the margin $k_1$ on $\boldsymbol{W}^1$ such that the two solutions are geodesically connected. We call it $\kappa_{\text{krn}}(k_1)$ and define it as $\kappa_{\text{krn}}(k_1) = k_1 p - \sqrt{(1-p^2)(\kappa_E^2 - k_1^2)}$. The maximum value of $\kappa_{\text{krn}}$ is obtained for $k_1 = \kappa_E$, for which $\kappa_{\text{krn}} = k_1 p$, which is reported in Fig. 3.

*Appendix C: Numerical simulations on the attractiveness of the convex core for SGD.*—We study the inductive bias of SGD dynamics on the cross-entropy loss by evaluating the average maximum energy barrier's height between pairs of independent solutions (means are over different initial conditions).

We observe that randomly initialized trajectories on the $N$ sphere access the zero-energy star-shaped manifold of solutions by entering directly in the geodesically convex component: as soon as the solutions are obtained, they are not only connected to fBP solutions in the core of the solution space [see Xent curve in Fig. 3(b)], but they are also connected between each other with zero-energy geodesic paths, as shown in the upper panel of Fig. 5, where the optimization is stopped as soon as the solutions are found. The disconnection transition of the SGD-SGD barrier is located close to the $\alpha_{LE}$ transition [41].

To further investigate the dynamical bias of SGD toward the convex manifold, we consider pairs of trajectories initialized inside the zero-energy region in correspondence to typical solutions (provided by SA). We find that the initial nonzero energy barrier between them quickly drops to zero in a few epochs, showing that SGD naturally drifts to the inner and convex region of the solution space; see lower panel of Fig. 5.

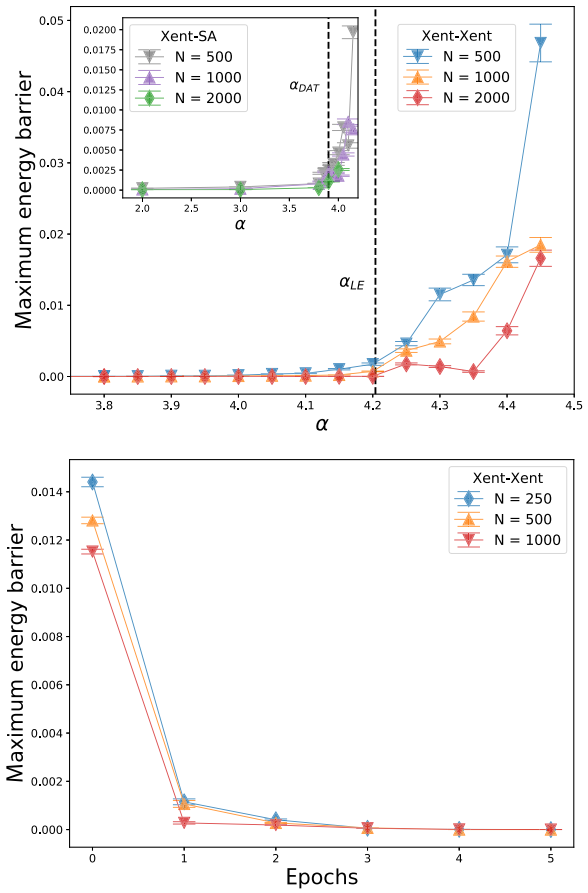


FIG. 5. Top: maximum barrier height along the geodesic path connecting two SGD solutions obtained with the cross-entropy loss in function of $\alpha$. The training dynamics is stopped as soon as a solution is found. Inset: maximum barrier height along the geodesic path connecting SGD solutions with SA solutions. Bottom: maximum barrier height between two configurations initialized on typical solutions (obtained with SA) as a function of the number of SGD epochs on the cross-entropy loss. After a few epochs, the gradient drives the solutions toward the convex core.


[1] M. Mézard, G. Parisi, and R. Zecchina, Science **297**, 812 (2002).

[2] O. C. Martin, R. Monasson, and R. Zecchina, Theor. Comput. Sci. **265**, 3 (2001).

[3] M. Mézard, T. Mora, and R. Zecchina, Phys. Rev. Lett. **94**, 197205 (2005).

[4] L. Zdeborová and F. Krząkała, Phys. Rev. E **76**, 031131 (2007).

[5] M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* (World Scientific Publishing Company, Singapore, 1987), Vol. 9.

[6] R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman, and L. Troyansky, Nature (London) **400**, 133 (1999).
[7] A. G. Cavaliere and F. Ricci-Tersenghi, arXiv:2303.14879.
[8] Loss Landscape of Neural Networks: theoretical insights and practical implications, EPFL Virtual Symposium.
[9] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, arXiv:1609.04836.
[10] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., Montréal, 2018), Vol. 31.
[11] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Phys. Rev. Lett. **115**, 128101 (2015).
[12] C. Baldassi, C. Borgs, J. T. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Proc. Natl. Acad. Sci. U.S.A. **113**, E7655 (2016).
[13] C. Baldassi, E. M. Malatesta, and R. Zecchina, Phys. Rev. Lett. **123**, 170602 (2019).
[14] C. Baldassi, F. Pittorino, and R. Zecchina, Proc. Natl. Acad. Sci. U.S.A. **117**, 161 (2020).
[15] C. Baldassi, E. M. Malatesta, M. Negri, and R. Zecchina, J. Stat. Mech. (2020) 124012.
[16] C. Baldassi, C. Lauditi, E. M. Malatesta, R. Pacelli, G. Perugini, and R. Zecchina, Phys. Rev. E **106**, 014116 (2022).
[17] L. Sagun, L. Bottou, and Y. LeCun, arXiv:1611.07476.
[18] L. Sagun, U. Evci, V. U. Guney, Y. Dauphin, and L. Bottou, arXiv:1706.04454.
[19] F. Pittorino, C. Lucibello, C. Feinauer, G. Perugini, C. Baldassi, E. Demyanenko, and R. Zecchina, in Entropic gradient descent algorithms and wide flat minima, J. Stat. Mech. 124015 (2021).
[20] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, in *International Conference on Learning Representations* (2021), https://openreview.net/forum?id=6Tm1mposlrM.
[21] Y. Feng and Y. Tu, Proc. Natl. Acad. Sci. U.S.A. **118**, e2015617118 (2021).
[22] G. Chen, C. K. Qu, and P. Gong, Neural Netw. **149**, 18 (2022).
[23] D. Kunin, J. Sagastuy-Brena, L. Gillespie, E. Margalit, H. Tanaka, S. Ganguli, and D. L. Yamins, arXiv:2107.09133.
[24] F. Draxler, K. Veschgini, M. Salmhofer, and F. Hamprecht, in *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 80, edited by J. Dy and A. Krause (PMLR, Stockholm, 2018), pp. 1309–1318.
[25] T. Garipov, P. Izmailov, D. Podoprikhin, D. P. Vetrov, and A. G. Wilson, in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., Montréal, 2018), Vol. 31.
[26] R. Entezari, H. Sedghi, O. Saukh, and B. Neyshabur, in *International Conference on Learning Representations* (2022), https://openreview.net/forum?id=dNigytemkL.
[27] F. Pittorino, A. Ferraro, G. Perugini, C. Feinauer, C. Baldassi, and R. Zecchina, in *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 162, edited by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato (PMLR, Baltimore, 2022), pp. 17759–17781.
[28] K. Jordan, H. Sedghi, O. Saukh, R. Entezari, and B. Neyshabur, in *The Eleventh International Conference on Learning Representations* (2023), https://openreview.net/forum?id=gU5sJ6ZggcX.
[29] J. Frankle and M. Carbin, in *International Conference on Learning Representations* (2019), https://openreview.net/forum?id=rJl-b3RcF7.
[30] J. Frankle, G. K. Dziugaite, D. Roy, and M. Carbin, in *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 119, edited by H. D. III and A. Singh (PMLR, 2020), pp. 3259–3269.
[31] S. I. Mirzadeh, M. Farajtabar, D. Gorur, R. Pascanu, and H. Ghasemzadeh, in *International Conference on Learning Representations* (2021), https://proceedings.mlr.press/v139/benton21a.html.
[32] G. Benton, W. Maddox, S. Lotfi, and A. G. G. Wilson, in *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 139, edited by M. Meila and T. Zhang (PMLR, 2021), pp. 769–779.
[33] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).
[34] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, England, 2001).
[35] S. Franz, G. Parisi, M. Sevelev, P. Urbani, and F. Zamponi, SciPost Phys. **2**, 019 (2017).
[36] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.131.227301 for a discussion of its phase diagram and related references on the connection of the model with hard spheres [37–39].
[37] G. Parisi and F. Zamponi, Rev. Mod. Phys. **82**, 789 (2010).
[38] L. Berthier and G. Biroli, Rev. Mod. Phys. **83**, 587 (2011).
[39] T. R. Kirkpatrick and D. Thirumalai, Rev. Mod. Phys. **87**, 183 (2015).
[40] G. Hansen, I. Herburt, H. Martini, and M. Moszyńska, Aequ. Math. **94**, 1001 (2020).
[41] C. Baldassi, E. M. Malatesta, G. Perugini, and R. Zecchina, Phys. Rev. E **108**, 024310 (2023).
[42] S. Franz and G. Parisi, J. Phys. A **49**, 145001 (2016).
[43] A. El Alaoui and M. Sellke, J. Stat. Phys. **189**, 27 (2022).
[44] A. Montanari, Y. Zhong, and K. Zhou, arXiv:2110.15824.
[45] C. Baldassi, C. Lauditi, E. M. Malatesta, G. Perugini, and R. Zecchina, Phys. Rev. Lett. **127**, 278301 (2021).
[46] H. Huang, K. Y. M. Wong, and Y. Kabashima, J. Phys. A **46**, 375002 (2013).
[47] Y. Zhang and S. H. Strogatz, Phys. Rev. Lett. **127**, 194101 (2021).
[48] S. Martiniani and M. Casiulis, Papers Phys. **15**, 150001 (2023).
[49] S. Franz, S. Hwang, and P. Urbani, Phys. Rev. Lett. **123**, 160602 (2019).
[50] S. Spigler, M. Geiger, S. d. Ascoli, L. Sagun, G. Biroli, and M. Wyart, J. Phys. A **52**, 474001 (2019).
[51] A. Jacot, F. Ged, B. Simsek, C. Hongler, and F. Gabriel, arXiv:2106.15933.