Supplementary Information: A mean field view of the landscape of two-layer neural networks

Song Mei* Andrea Montanari† Phan-Minh Nguyen‡

July 25, 2018

Abstract

This document contains the Supplementary Information for the manuscript *'A mean field view of the landscape of two-layer neural networks'*. In particular, we present here proofs and additional technical details for our mathematical results, as well as additional information concerning the numerical experiments.

Contents

*Institute for Computational and Mathematical Engineering, Stanford University
†Department of Electrical Engineering and Department of Statistics, Stanford University
‡Department of Electrical Engineering, Stanford University

www.pnas.org/cgi/doi/10.1073/pnas.1806579115

1 Notations

We use lowercase bold for vectors (e.g. $\boldsymbol{u}, \boldsymbol{v}, \dots$), uppercase bold for matrices (e.g. $\boldsymbol{A}, \boldsymbol{B}, \dots$), and lowercase plain for scalar ($x, y, \dots$).

- Given a measurable space Ω, we denote by $\mathscr{P}(\Omega)$ the set of probability measures on Ω.
- $\mathsf{B}^d(\boldsymbol{x}; r)$ denotes the Euclidean ball with center $\boldsymbol{x}$ and radius r in $\mathbb{R}^d$. We will drop the dimension superscript whenever clear from the context.
- Given a measurable function f, and a measure μ, we denote by $\langle f, \mu\rangle = \langle \mu, f\rangle = \int f \, \mathrm{d}\mu$ the corresponding integral.
- For a univariate function $f : \mathbb{R} \to \mathbb{R}$, we denote by $f'(x)$ its derivative at x. If the argument is time, we will also use $\dot{f}(t)$.
- $\|f\|_{\text{Lip}} \equiv \sup_{x \neq y} |f(\boldsymbol{x}) - f(\boldsymbol{y})|/\|\boldsymbol{x} - \boldsymbol{y}\|_2$ denotes the Lipshitz constant of a function f.
- $d_{\text{BL}}(\,\cdot\,, \cdot\,)$ is the bounded Lipschitz distance between probability measures

$$d_{\text{BL}}(\mu, \nu) = \sup\left\{ \left| \int f(\boldsymbol{x})\, \mu(\mathrm{d}\boldsymbol{x}) - \int f(\boldsymbol{x})\, \nu(\mathrm{d}\boldsymbol{x}) \right| \;:\; \|f\|_\infty \le 1, \|f\|_{\text{Lip}} \le 1 \right\} \tag{1.1}$$

$$\le 2 \inf_{\gamma \in \mathcal{C}(\mu,\nu)} \int (\|\boldsymbol{x} - \boldsymbol{y}\|_2 \wedge 1) \gamma(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) \le 4\, d_{\text{BL}}(\mu, \nu)\,. \tag{1.2}$$

Here $\mathcal{C}(\mu,\nu)$ is the set of couplings of μ and ν.

- $W_p(\,\cdot\,,\,\cdot\,)$ is the Wasserstein distance between probability measures

$$W_p(\mu,\nu) = \Big(\inf_{\gamma\in\mathcal{C}(\mu,\nu)} \int \|\boldsymbol{x}-\boldsymbol{y}\|_2^p \gamma(\mathrm{d}\boldsymbol{x},\mathrm{d}\boldsymbol{y})\Big)^{1/p}. \tag{1.3}$$

For $p=1$, the Kantorovich-Rubinstein duality gives

$$W_1(\mu,\nu) = \sup\left\{\left|\int f(\boldsymbol{x})\,\mu(\mathrm{d}\boldsymbol{x}) - \int f(\boldsymbol{x})\,\nu(\mathrm{d}\boldsymbol{x})\right| \;:\; \|f\|_{\mathrm{Lip}}\le 1\right\}. \tag{1.4}$$

- K is a generic constant depending on K_0, K_1, K_2, K_3, where K_i's are constants which will be specified from the context.
- $\mathbb{N} = \{0,1,2,\dots\}$ denote the set of natural numbers.

2 General results: Statics

In this section, we discuss some properties of the population risk, $R_N(\boldsymbol{\theta})$, and its continuum counterpart $R(\rho)$. For future reference, we copy the key definitions from the main text:

$$R_N(\boldsymbol{\theta}) \equiv R_\# + \frac{2}{N}\sum_{i=1}^N V(\boldsymbol{\theta}_i) + \frac{1}{N^2}\sum_{i,j=1}^N U(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)\,, \tag{2.1}$$

$$R(\rho) \equiv R_\# + 2\int V(\boldsymbol{\theta})\;\rho(\mathrm{d}\boldsymbol{\theta}) + \int U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\;\rho(\mathrm{d}\boldsymbol{\theta}_1)\,\rho(\mathrm{d}\boldsymbol{\theta}_2)\,, \tag{2.2}$$

$$R_\# = \mathbb{E}\{y^2\}\,, \qquad V(\boldsymbol{\theta}) = -\mathbb{E}\{y\,\sigma_*(\boldsymbol{x};\boldsymbol{\theta})\}\,, \tag{2.3}$$

$$U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \mathbb{E}\{\sigma_*(\boldsymbol{x};\boldsymbol{\theta}_1)\sigma_*(\boldsymbol{x};\boldsymbol{\theta}_2)\}\,. \tag{2.4}$$

We further recall the notation

$$\Psi(\boldsymbol{\theta};\rho) = V(\boldsymbol{\theta}) + \int U(\boldsymbol{\theta},\boldsymbol{\theta}')\;\rho(\mathrm{d}\boldsymbol{\theta}')\,. \tag{2.5}$$

We will always assume that the expectations defining $V(\boldsymbol{\theta}), U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ exist finite for all $\boldsymbol{\theta},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\mathbb{R}^D$. A necessary and sufficient condition for this is that $\mathbb{E}\{\sigma_*(\boldsymbol{x};\boldsymbol{\theta})^2\}<\infty$ for all $\boldsymbol{\theta}$. Since in most cases of interest $|\sigma_*(\boldsymbol{x};\boldsymbol{\theta})|\le M(\boldsymbol{\theta})\|\boldsymbol{x}\|_2$, for this to happen, it is sufficient that $\boldsymbol{x}$ has a finite second moment.

Note that this $\rho\mapsto R(\rho)$ is a convex function on the set of probability measures on $\mathbb{R}^D$. We will denote by $\mathscr{P}_{V,U}$ the subset of probability measures ρ such that the expectations on the right-hand side are finite. We define $R(\rho)=\infty$ if $\rho\in\mathscr{P}(\mathbb{R}^D)\setminus\mathscr{P}_{V,U}$.

2.1 Proof of Proposition 1

The proof is divided in two parts:

1. We show that minimizing the population risk $R_N(\boldsymbol{\theta})$ yields similar results to minimizing its continuum counterpart $R(\rho)$:

$$\Big|\inf_{\boldsymbol{\theta}} R_N(\boldsymbol{\theta}) - \inf_{\rho} R(\rho)\Big| \le \frac{K}{N}\,. \tag{2.6}$$

2. We establish the condition for ρ_* to be a minimizer:

$$\mathrm{supp}(\rho_*) \subseteq \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^D} \Psi(\boldsymbol{\theta};\rho_*)\,. \tag{2.7}$$

First notice that, for any $\boldsymbol{\theta} = (\boldsymbol{\theta}_i)_{i\le N}$, we have

$$R_N(\boldsymbol{\theta}) \ge \inf_{\rho} R(\rho)\,. \tag{2.8}$$

Indeed, $R_N(\boldsymbol{\theta}) = R(\rho)$ for $\rho = (1/N)\sum_{i=1}^N \delta_{\boldsymbol{\theta}_i}$.

In order to prove Eq. (2.6), let $\rho_* \in \mathscr{P}(\mathbb{R}^D)$ be such that $R(\rho_*) = R_*$ under assumption (a), or $R(\rho_*) \le R_* + \varepsilon$ under assumption (b). Let $(\boldsymbol{\theta}_i)_{i\le N} \sim_{iid} \rho_*$. A simple calculation shows that

$$\mathbb{E}_{\boldsymbol{\theta}}[R_N(\boldsymbol{\theta})] - R(\rho_*) = \frac{1}{N}\left\{\int U(\boldsymbol{\theta},\boldsymbol{\theta})\,\rho_*(\mathrm{d}\boldsymbol{\theta}) - \int U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\,\rho_*(\mathrm{d}\boldsymbol{\theta}_1)\,\rho_*(\mathrm{d}\boldsymbol{\theta}_2)\right\} \tag{2.9}$$

$$\le \frac{1}{N}\int U(\boldsymbol{\theta},\boldsymbol{\theta})\,\rho_*(\mathrm{d}\boldsymbol{\theta}) \le \frac{K}{N}\,, \tag{2.10}$$

where the first inequality follows since $\int U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\,\rho_*(\mathrm{d}\boldsymbol{\theta}_1)\,\rho_*(\mathrm{d}\boldsymbol{\theta}_2) = \mathbb{E}\{y(\boldsymbol{x})^2\} \ge 0$ for $y(\boldsymbol{x}) = \int \sigma_*(\boldsymbol{x};\boldsymbol{\theta})\,\rho_*(\mathrm{d}\boldsymbol{\theta})$, and the second inequality follows by assumption. It follows that

$$\inf_{\boldsymbol{\theta}} R_N(\boldsymbol{\theta}) \le R_* + \frac{K}{N} + \varepsilon\,, \tag{2.11}$$

whence the claim (2.6) follows since ε is arbitrary.

We next establish the minimum condition (2.7). Notice that since $V(\,\cdot\,)$ is continuous, and $U(\,\cdot\,,\,\cdot\,)$ is bounded below, it follows from Fatou's lemma that, for any ρ, the function $\boldsymbol{\theta} \mapsto \Psi(\boldsymbol{\theta};\rho)$ is lower semicontinous and takes values in $(-\infty,\infty]$. In particular the set $S_0(\rho) \equiv \arg\min_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta};\rho)$ must be closed.

We first prove that any minimizer must satisfy (2.7). Let ρ_* be a minimizer and define $\Psi_* = \inf_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta};\rho_*)$. By rearranging terms, for any probability measure ρ, we have

$$R(\rho) - R(\rho_*) = 2\langle \Psi(\,\cdot\,;\rho_*), (\rho-\rho_*)\rangle + \langle U, (\rho-\rho_*)^{\otimes 2}\rangle\,. \tag{2.12}$$

First we will assume $\Psi_* > -\infty$ (whence, by lower semicontinuity, $S_0(\rho_*)$ must be a non-empty closed set). Let $\boldsymbol{\theta}_1 \in S_0(\rho_*)$, and assume by contradiction that there exist $\boldsymbol{\theta}_0 \in \mathrm{supp}(\rho_*)$, $\boldsymbol{\theta}_0 \notin S_0(\rho_*)$. Let $\mathsf{B}(\boldsymbol{\theta}_0;\varepsilon)$ be a ball of radius ε around $\boldsymbol{\theta}_0$. By lower semicontinuity, we can find $\varepsilon_0, \Delta > 0$ such that $\inf_{\boldsymbol{\theta}\in\mathsf{B}(\boldsymbol{\theta}_0;\varepsilon_0)} \Psi(\boldsymbol{\theta};\rho_*) = \Psi_* + \Delta > \Psi_*$. Further $t_0 \equiv \rho_*(\mathsf{B}(\boldsymbol{\theta}_0;\varepsilon_0)) > 0$ because $\boldsymbol{\theta}_0 \in \mathrm{supp}(\rho_*)$.

Let $\nu \equiv \mathbf{1}_{\mathsf{B}(\boldsymbol{\theta}_0;\varepsilon_0)}\rho_*/t_0$ (i.e. ν is the conditional distribution given $\boldsymbol{\theta} \in \mathsf{B}(\boldsymbol{\theta}_0;\varepsilon_0)$). Define, for $t \in [0,t_0]$, the probability measure

$$\rho_t = \rho_* - t\nu + t\delta_{\boldsymbol{\theta}_1}\,. \tag{2.13}$$

Using Eq. (2.12), we get

$$R(\rho_t) - R(\rho_*) = 2\langle \Psi(\,\cdot\,;\rho_*), (\delta_{\boldsymbol{\theta}_1} - \nu)\rangle\, t + \langle U, (\delta_{\boldsymbol{\theta}_1} - \nu)^{\otimes 2}\rangle t^2 \tag{2.14}$$

$$\le 2(\Psi_* - \Psi_* - \Delta)\, t + C_0\, t^2 = -2\Delta\, t + C_0\, t^2\,, \tag{2.15}$$

where the second inequality follows from the fact that U is continuous and $\delta_{\boldsymbol{\theta}_1}$, ν have bounded support. By taking t small enough, we get $R(\rho) < R(\rho_*)$ hence reaching a contradiction.

Next consider the case in which $\Psi_* \equiv \inf_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}; \rho_*) = -\infty$. For $M \in \mathbb{N}$, $M \geq 1$, let $\boldsymbol{\theta}_M \in \mathbb{R}^D$ be such that $\Psi(\boldsymbol{\theta}_M; \rho_*) \leq -M$. For $\boldsymbol{\theta}_0 \in \mathrm{supp}(\boldsymbol{\theta}_*)$, construct ν as before. Note that, and call $\inf_{\boldsymbol{\theta} \in \mathsf{B}(\boldsymbol{\theta}_0; \varepsilon_0)} \Psi(\boldsymbol{\theta}; \rho_*) = \Psi_0$. Define, for $t \in [0, t_0]$

$$\rho_{M,t} = \rho_* - t\nu + t\delta_{\boldsymbol{\theta}_M} \,. \tag{2.16}$$

By applying again Eq. (2.12), we get

$$R(\rho_{M,t}) - R(\rho_*) = 2\langle \Psi(\,\cdot\,; \rho_*), (\delta_{\boldsymbol{\theta}_M} - \nu)\rangle\, t + \langle U, (\delta_{\boldsymbol{\theta}_M} - \nu)^{\otimes 2}\rangle t^2 \tag{2.17}$$

$$\leq -2(M + \Psi_0)\, t + C_0(M)\, t^2 \,. \tag{2.18}$$

By selecting $t = t_M = \min(t_0, (M + \Psi_0)/C_0(M))$ (which is positive for all M large enough), we obtain $R(\rho_{M,t}) - R(\rho_*) < 0$ for all M large and hence reach a contradiction.

We finally prove that condition (2.7) is sufficient for ρ_* to be a minimizer. Indeed, for any non-negative measurable function $\mu : \mathbb{R}^D \to \mathbb{R}$, letting $\Psi_* = \min_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}; \rho_*)$,

$$R(\rho) \geq R_\# + 2\langle V, \rho\rangle + \langle U, \rho^{\otimes 2}\rangle - \langle \mu, \rho\rangle \tag{2.19}$$

$$= R(\rho_*) + 2\langle \Psi(\,\cdot\,; \rho_*), \rho - \rho_*\rangle + \langle U, (\rho - \rho_*)^{\otimes 2}\rangle - \langle \mu, \rho\rangle \tag{2.20}$$

$$= R(\rho_*) + 2\langle \Psi(\,\cdot\,; \rho_*) - \Psi_*, \rho - \rho_*\rangle + \langle U, (\rho - \rho_*)^{\otimes 2}\rangle - \langle \mu, \rho\rangle \,. \tag{2.21}$$

Setting $\mu = 2[\Psi(\,\cdot\,; \rho_*) - \Psi_*]$, and noticing that condition (2.7) implies $\langle \Psi(\,\cdot\,; \rho_*) - \Psi_*, \rho_*\rangle = 0$, we get $R(\rho) \geq R(\rho_*) + \langle U, (\rho - \rho_*)^{\otimes 2}\rangle \geq R(\rho_*)$.

2.2 Some additional results

We often find empirically that the optimal density ρ_* is supported on a set of Lebesgue measure 0 (sometimes on a finite set of points). The following consequence of the previous results partially explains these findings.

Corollary 2.1. *Assume $\boldsymbol{\theta} \mapsto V(\boldsymbol{\theta})$ to be an analytic function and $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \mapsto U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ to be analytic with respect to $\boldsymbol{\theta}_1$, uniformly in $\boldsymbol{\theta}_2$. Namely there exists a locally bounded function $\boldsymbol{\theta} \mapsto B(\boldsymbol{\theta})$ such that $\|\nabla^k_{\boldsymbol{\theta}_1} U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\|_2 \leq k! B(\boldsymbol{\theta}_1)^k$ for all k, $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$. If ρ_* is a minimizer of $R(\rho)$, then one of the following holds*

(a) *$\Psi(\boldsymbol{\theta}; \rho_*) = \Psi_*$ for some constant Ψ_* and all $\boldsymbol{\theta} \in \mathbb{R}^D$.*

(b) *The support of ρ_* has zero Lebesgue measure.*

If $D = 1$, then (b) can be replaced by: (b′) ρ_* is a convex combination of countably many point masses with no accumulation point (finitely many if $\Psi(\theta; \rho_*) \to \infty$ as $|\theta| \to \infty$).

Proof. Note that, under the stated conditions $f(\boldsymbol{\theta}) \equiv \int U(\boldsymbol{\theta}, \boldsymbol{\theta}')\, \rho_*(\mathrm{d}\boldsymbol{\theta}')$ is analytic. Indeed, by a standard dominated convergence argument, we have that $\nabla^k f$ is given by the integral of $\int \nabla^k U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\, \rho_*(\mathrm{d}\boldsymbol{\theta}_2)$ for any $k \geq 0$. Further, by an application of the intermediate value theorem there exists $t_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\delta}} \in [0, 1]$ such that

$$\left| f(\boldsymbol{\theta}_1 + \boldsymbol{\delta}) - \sum_{\ell=0}^{k-1} \frac{1}{\ell!} \langle \nabla^\ell f(\boldsymbol{\theta}_1), \boldsymbol{\delta}^{\otimes \ell}\rangle \right| \leq \frac{1}{k!} \left| \int \langle \nabla^k_{\boldsymbol{\theta}_1} U(\boldsymbol{\theta}_1 + t_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\delta}} \boldsymbol{\delta}, \boldsymbol{\theta}_2), \boldsymbol{\delta}^{\otimes k}\rangle\, \rho_*(\mathrm{d}\boldsymbol{\theta}_2) \right| \tag{2.22}$$

$$\leq \int B(\boldsymbol{\theta}_1 + t_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\delta}} \boldsymbol{\delta})^k \|\boldsymbol{\delta}\|_2^k\, \rho_*(\mathrm{d}\boldsymbol{\theta}_2) \tag{2.23}$$

$$\leq \sup_{\boldsymbol{\theta} \in \mathsf{B}(\boldsymbol{\theta}_1; \|\boldsymbol{\delta}\|_2)} B(\boldsymbol{\theta})^k\, \|\boldsymbol{\delta}\|_2^k \,, \tag{2.24}$$

which vanishes as $k \to \infty$ for uniformly over $\|\boldsymbol{\delta}\|_2 \le \delta_0$ for δ_0 small enough.

Let $\Psi_* = \min_{\boldsymbol{\theta}\in\mathbb{R}^D} \Psi(\boldsymbol{\theta};\rho_*)$. We thus have that $\boldsymbol{\theta} \mapsto \Psi(\boldsymbol{\theta};\rho_*)$ is also analytic and so is $\boldsymbol{\theta} \mapsto \Psi(\boldsymbol{\theta};\rho_*) - \Psi_*$. Since $\mathrm{supp}(\rho_*) \subseteq \{\boldsymbol{\theta} : \Psi(\boldsymbol{\theta};\rho_*) = \Psi_*\}$, the claim follows from the fact that the set of zeros of a non-trivial analytic function has vanishing Lebesgue measure [Mit15]. In the case $D = 1$, the set of zeros of an analytic function cannot have any accumulation point [Lan13], which therefore allows to replace (b) with (b'). □

3 General results: Dynamics

In this section we consider the SGD dynamics with step size $s_k = \varepsilon\xi(k\varepsilon)$, under the assumptions A1, A2, A3 stated in the main text. For the readers convenience, we reproduce here the form of the limiting PDE

$$\partial_t \rho_t(\boldsymbol{\theta}) = 2\xi(t)\nabla\cdot\left[\rho_t(\boldsymbol{\theta})\nabla\Psi(\boldsymbol{\theta};\rho_t)\right], \tag{3.1}$$

$$\Psi(\boldsymbol{\theta};\rho) = V(\boldsymbol{\theta}) + \int U(\boldsymbol{\theta},\boldsymbol{\theta}')\,\rho(\mathrm{d}\boldsymbol{\theta}')\,. \tag{3.2}$$

Recall that this is an evolution in the space of probability measures in $\mathbb{R}^D$, and is to be interpreted in weak sense. Namely ρ_t is a solution of Eq. (3.1), if, for any bounded differentiable function $\varphi : \mathbb{R}^D \to \mathbb{R}$ with bounded gradient:

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle\rho_t,\varphi\rangle = -2\xi(t)\int\langle\nabla\varphi(\boldsymbol{\theta}),\nabla\Psi(\boldsymbol{\theta};\rho_t)\rangle\,\rho_t(\mathrm{d}\boldsymbol{\theta})\,. \tag{3.3}$$

For background on this and similar PDEs (and the analogous ones at finite temperature, cf. Section 6), we refer to [MV00, CMV+03, CMV06, AGS08, CDF+11]. Our treatment will be mostly self-contained because of some differences between our setting and the one in these papers.

Remark 3.1. Recall assumptions A1, A2, A3 in the main text. By [Szn91, Theorem 1.1], assumptions A1 and A3 are sufficient for the existence and uniqueness of solution of PDE (3.1).

A very useful tool for the analysis of the PDE (3.1) is provided by the following *nonlinear dynamics*. We introduce trajectories $(\overline{\boldsymbol{\theta}}_i^t)_{1\le i\le N,\, t\in\mathbb{R}_{\ge 0}}$ by letting $\overline{\boldsymbol{\theta}}_i^0 = \boldsymbol{\theta}_i^0$ to be the same initialization as for SGD and, for $t \ge 0$ (here P_X denotes the law of the random variable X):

$$\overline{\boldsymbol{\theta}}_i^t = \boldsymbol{\theta}_i^0 - 2\int_0^t \xi(s)\,\nabla\Psi(\overline{\boldsymbol{\theta}}_i^s;\rho_s)\,\mathrm{d}s\,, \tag{3.4}$$

$$\rho_s = \mathrm{P}_{\overline{\boldsymbol{\theta}}_i^s}\,. \tag{3.5}$$

This should be regarded as an equation for the law of the trajectory $(\overline{\boldsymbol{\theta}}_i^t)_{t\in\mathbb{R}_{\ge 0}}$, with boundary condition determined by $\overline{\boldsymbol{\theta}}_i^0 \sim \rho_0$. As implied by [Szn91, Theorem 1.1], under the same assumptions A1 and A3, the nonlinear dynamics has a unique solution, with ρ_t satisfying Eq. (3.1).

Lemma 3.1. *Assume conditions* A1 *and* A3 *hold. Let $(\rho_t)_{t\ge 0}$ be the solution of the PDE (3.1). Let $(\overline{\boldsymbol{\theta}}_i^t)_{t\ge 0}$ be the solution of nonlinear dynamics (3.4). Then $t \mapsto \overline{\boldsymbol{\theta}}_i^t$ is K_1K_3-Lipschitz continuous, and $t \mapsto \rho_t$ is K_1K_3-Lipschitz continuous in W_2 Wasserstein distance, with K_1 and K_3 as per conditions* A1 *and* A3*. In particular, $t \mapsto \rho_t$ is continuous in the topology of weak convergence.*

Proof. Since ξ and $\nabla\Psi$ are K_1 and K_3 bounded respectively, $t \mapsto \overline{\boldsymbol{\theta}}_i^t$ is K_1K_3-Lipschitz continuous. Further, Eq. (1.2) implies that $t \mapsto \rho_t$ is Lipschitz continuous in W_2 Wasserstein distance, namely

$$d_{\mathrm{BL}}(\rho_t, \rho_s) \le W_2(\rho_t, \rho_s) \le (\mathbb{E}[\|\overline{\boldsymbol{\theta}}_i^t - \overline{\boldsymbol{\theta}}_i^s\|_2^2])^{1/2} \le K_1K_3|t-s|. \tag{3.6}$$

□

We notice that, under the nonlinear dynamics, the trajectories $(\overline{\boldsymbol{\theta}}_1^t)_{t\in\mathbb{R}_{\ge 0}}$, …, $(\overline{\boldsymbol{\theta}}_N^t)_{t\in\mathbb{R}_{\ge 0}}$ are independent and identically distributed. In particular, this implies that, almost surely,

$$\frac{1}{N}\sum_{i=1}^N \delta_{\overline{\boldsymbol{\theta}}_i^t} \stackrel{\mathrm{d}}{\Rightarrow} \rho_t\,. \tag{3.7}$$

3.1 Proof of Theorem 3: Convergence to the PDE

The proof follows a 'propagation of chaos' argument [Szn91]. Throughout this proof, we will use K to denote generic constant depending on the constants K_1, K_2, K_3 in conditions A1, A2, A3.

It is convenient to introduce the notations $\boldsymbol{z}_k = (\boldsymbol{x}_k, y_k)$ to denote the k-th example and define

$$\boldsymbol{F}_i(\boldsymbol{\theta}; \boldsymbol{z}_k) = (y_k - \hat{y}(\boldsymbol{x}_k; \boldsymbol{\theta}))\, \nabla_{\boldsymbol{\theta}_i}\sigma_*(\boldsymbol{x}_k; \boldsymbol{\theta}_i)\,, \qquad \boldsymbol{\theta} = (\boldsymbol{\theta}_i)_{i\le N} \in \mathbb{R}^{D\times N}, \tag{3.8}$$

$$\boldsymbol{G}(\boldsymbol{\theta}; \rho) = -\nabla\Psi(\boldsymbol{\theta}; \rho) = -\nabla V(\boldsymbol{\theta}) - \int \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}, \boldsymbol{\theta}')\, \rho(\mathrm{d}\boldsymbol{\theta}')\,, \quad \boldsymbol{\theta} \in \mathbb{R}^D. \tag{3.9}$$

Note that the assumption of bounded Lipschitz ∇V, $\nabla_1 U$ (here and below $\nabla_1 U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ denotes the gradient of U with respect to its first argument) implies $\|\boldsymbol{G}(\boldsymbol{\theta}; \rho)\|_2 \le K$ and $\|\boldsymbol{G}(\boldsymbol{\theta}_1; \rho) - \boldsymbol{G}(\boldsymbol{\theta}_2; \rho)\|_2 \le K\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$. Further

$$\|\boldsymbol{G}(\boldsymbol{\theta}; \rho_1) - \boldsymbol{G}(\boldsymbol{\theta}; \rho_2)\|_2 = \Big\| \int \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}; \boldsymbol{\theta}')(\rho_1 - \rho_2)(\mathrm{d}\boldsymbol{\theta}') \Big\|_2 \le K\, d_{\mathrm{BL}}(\rho_1, \rho_2)\,. \tag{3.10}$$

With these notations, we can rewrite the SGD dynamics [3] in the main text as

$$\boldsymbol{\theta}_i^{k+1} = \boldsymbol{\theta}_i^k + 2\varepsilon\, \xi(k\varepsilon)\, \boldsymbol{F}_i(\boldsymbol{\theta}_i^k; \boldsymbol{z}_{k+1})\,, \tag{3.11}$$

which yields

$$\boldsymbol{\theta}_i^k = \boldsymbol{\theta}_i^0 + 2\varepsilon \sum_{\ell=0}^{k-1} \xi(\ell\varepsilon)\, \boldsymbol{F}_i(\boldsymbol{\theta}_i^\ell; \boldsymbol{z}_{\ell+1})\,. \tag{3.12}$$

Recall $(\boldsymbol{\theta}_i^0)_{i\le N} \sim \rho_0$ independently.

For $t \in \mathbb{R}_{\ge 0}$ we will define $[t] = \varepsilon\lfloor t/\varepsilon \rfloor$. Eq. (3.12) should be compared with the nonlinear dynamics (3.4), which reads

$$\overline{\boldsymbol{\theta}}_i^t = \boldsymbol{\theta}_i^0 + 2\int_0^t \xi(s)\, \boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^s; \rho_s)\, \mathrm{d}s\,. \tag{3.13}$$

We next state and prove the key estimate controlling the difference between the original dynamics and the nonlinear dynamics.

Lemma 3.2. *Under the assumptions of Theorem 3, there exists a constant K depending uniquely on K_1, K_2, K_3 in conditions* A1, A2, *and* A3, *such that for any $T \geq 0$, we have*

$$\max_{i\leq N} \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} \|\boldsymbol{\theta}_i^k - \overline{\boldsymbol{\theta}}_i^{k\varepsilon}\|_2 \leq K e^{KT} \cdot \sqrt{1/N \vee \varepsilon} \cdot \left[\sqrt{D + \log(N(T/\varepsilon \vee 1))} + z\right] \tag{3.14}$$

with probability at least $1 - e^{-z^2}$.

Proof. Consider for simplicity of notation $t \in \mathbb{N}\varepsilon \cap [0,T]$. Taking the difference of Eqs. (3.12) and (3.13), we get

$$\begin{aligned}
\|\boldsymbol{\theta}_i^{t/\varepsilon} - \overline{\boldsymbol{\theta}}_i^t\|_2 =& 2\Big\| \int_0^t \xi(s)\, \boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^s; \rho_s)\, \mathrm{d}s - \varepsilon \sum_{k=0}^{t/\varepsilon-1} \xi(k\varepsilon)\, \boldsymbol{F}_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1})\, \mathrm{d}s \Big\|_2 \\
\leq& 2\int_0^t \Big\| \xi(s)\, \boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^s; \rho_s) - \xi([s])\, \boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^{[s]}; \rho_{[s]}) \Big\|_2 \mathrm{d}s \\
&+ 2\int_0^t \Big\| \xi([s])\, \boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^{[s]}; \rho_{[s]}) - \xi([s])\, \boldsymbol{G}(\boldsymbol{\theta}_i^{\lfloor s/\varepsilon \rfloor}; \rho_{[s]}) \Big\|_2 \mathrm{d}s \\
&+ 2\Big\| \varepsilon \sum_{k=0}^{t/\varepsilon-1} \xi(k\varepsilon) \Big\{ \boldsymbol{F}_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1}) - \boldsymbol{G}(\boldsymbol{\theta}_i^k; \rho_{k\varepsilon}) \Big\} \Big\|_2 \\
\equiv& 2E_1^i(t) + 2E_2^i(t) + 2E_3^i(t)\,.
\end{aligned} \tag{3.15}$$

We next consider the three terms above. Using the Lipschitz continuity of $\boldsymbol{G}(\boldsymbol{\theta}; \rho)$ with respect to $\boldsymbol{\theta}$ and ρ (see Eq. (3.10)), and due to condition A1 and Lemma 3.1 (implying that ξ, $\overline{\boldsymbol{\theta}}_i^t$, and ρ_s are Lipschitz continuous), we get

$$\begin{aligned}
E_1^i(t) \leq t \sup_{s\in[0,t]} \Big\{ &\|\xi(s)\, \boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^s; \rho_s) - \xi([s])\, \boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^s; \rho_s)\|_2 + \|\xi([s])\, \boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^s; \rho_s) - \xi([s])\, \boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^{[s]}; \rho_s)\|_2 \\
&+ \|\xi([s])\, \boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^{[s]}; \rho_s) - \xi([s])\, \boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^{[s]}; \rho_{[s]})\|_2 \Big\} \\
&\leq K\, t\, \varepsilon\,.
\end{aligned} \tag{3.16}$$

Bounding the second term yields (by using the Lipschitz continuity of $\boldsymbol{G}$ with respect to its first argument):

$$E_2^i(t) \leq K \int_0^t \|\boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^{[s]}; \rho_{[s]}) - \boldsymbol{G}(\boldsymbol{\theta}_i^{\lfloor s/\varepsilon \rfloor}; \rho_{[s]})\|_2 \mathrm{d}s \leq K^2 \int_0^t \|\overline{\boldsymbol{\theta}}_i^{[s]} - \boldsymbol{\theta}_i^{\lfloor s/\varepsilon \rfloor}\|_2 \mathrm{d}s\,. \tag{3.17}$$

In order to bound the last term we denote by $\mathcal{F}_k$, for $k \in \mathbb{N}$, the sigma-algebra generated by $(\boldsymbol{\theta}_i^0)_{i\leq N}$ and $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k$. Note that

$$\mathbb{E}\{\boldsymbol{F}_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1}) | \mathcal{F}_k\} = -\nabla V(\boldsymbol{\theta}_i^k) - \frac{1}{N}\sum_{j=1}^N \nabla_1 U(\boldsymbol{\theta}_i^k, \boldsymbol{\theta}_j^k) = \boldsymbol{G}(\boldsymbol{\theta}_i^k; \hat{\rho}_k^{(N)})\,, \tag{3.18}$$

where $\hat{\rho}_k^{(N)} \equiv (1/N)\sum_{i\leq N} \delta_{\boldsymbol{\theta}_i^k}$. Hence

$$E_3^i(t) \leq \Big\| \varepsilon \sum_{k=0}^{t/\varepsilon-1} \xi(k\varepsilon) \Big\{ \boldsymbol{G}(\boldsymbol{\theta}_i^k; \hat{\rho}_k^{(N)}) - \boldsymbol{G}(\boldsymbol{\theta}_i^k; \rho_{k\varepsilon}) \Big\} \Big\|_2 + \Big\| \varepsilon \sum_{k=0}^{t/\varepsilon-1} \xi(k\varepsilon) \boldsymbol{Z}_k^i \Big\|_2 \tag{3.19}$$

$$\equiv E_{3,0}^i(t) + Q_1^i(t)\,, \tag{3.20}$$

where we introduced the martingale differences $\boldsymbol{Z}_k^i \equiv \boldsymbol{F}_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1}) - \mathbb{E}\{\boldsymbol{F}_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1})|\mathcal{F}_k\}$. We can apply Azuma-Hoeffding inequality, cf. Lemma A.1. Indeed, condition (A.1) follows from the fact that $\sigma_*(\boldsymbol{x};\boldsymbol{\theta})$ is bounded and $\nabla_{\boldsymbol{\theta}}\sigma_*(\boldsymbol{x};\boldsymbol{\theta})$ is sub-Gaussian (the product of a sub-Gaussian random vector and a bounded random variable is sub-Gaussian, cf. for instance Lemma 1.(d) in [MBM16]), hence each $\xi(k\varepsilon)\boldsymbol{Z}_k^i$ are K^2-sub-Gaussian. We therefore get

$$\mathbb{P}\Big(\max_{k\in[0,t/\varepsilon]\cap\mathbb{N}} Q_1^i(k\varepsilon) \ge K\sqrt{t\varepsilon}(\sqrt{D}+u)\Big) \le e^{-u^2}\ , \tag{3.21}$$

and taking union bound over $i \le N$, we get

$$\mathbb{P}\Big(\max_{i\le N}\max_{k\in[0,t/\varepsilon]\cap\mathbb{N}} Q_1^i(k\varepsilon) \le K\sqrt{t\varepsilon}\,(\sqrt{D+\log N}+z)\Big) \ge 1-e^{-z^2}\ . \tag{3.22}$$

For the term $E_{3,0}^i(t)$, we use the Lipschitz continuity property (3.10), whence

$$\begin{aligned}
&\|\boldsymbol{G}(\boldsymbol{\theta}_i^k;\hat{\rho}_k^{(N)}) - \boldsymbol{G}(\boldsymbol{\theta}_i^k;\rho_{k\varepsilon})\|_2\\
\le&\Big\|\frac{1}{N}\sum_{j=1}^N[\nabla_1 U(\boldsymbol{\theta}_i^k,\boldsymbol{\theta}_j^k) - \nabla_1 U(\boldsymbol{\theta}_i^k,\overline{\boldsymbol{\theta}}_j^{k\varepsilon})]\Big\|_2 + \Big\|\frac{1}{N}\sum_{j=1}^N[\nabla_1 U(\boldsymbol{\theta}_i^k,\overline{\boldsymbol{\theta}}_j^{k\varepsilon}) - \mathbb{E}_{\overline{\boldsymbol{\theta}}}\nabla_1 U(\boldsymbol{\theta}_i^k,\overline{\boldsymbol{\theta}}_j^{k\varepsilon})]\Big\|_2\\
\le&\frac{K}{N}\sum_{j=1}^N\|\boldsymbol{\theta}_j^k - \overline{\boldsymbol{\theta}}_j^{k\varepsilon}\|_2 + Q_2^i(k\varepsilon) + \frac{K}{N}\,.
\end{aligned} \tag{3.23}$$

Here $Q_2^i(k\varepsilon)$ for $k\in\mathbb{N}$ is defined as

$$Q_2^i(k\varepsilon) = \Big\|\frac{1}{N}\sum_{j\le N, j\ne i}[\nabla_1 U(\boldsymbol{\theta}_i^k,\overline{\boldsymbol{\theta}}_j^{k\varepsilon}) - \mathbb{E}_{\overline{\boldsymbol{\theta}}}\nabla_1 U(\boldsymbol{\theta}_i^k,\overline{\boldsymbol{\theta}}_j^{k\varepsilon})]\Big\|_2\,.$$

Since for any fixed k, $(\overline{\boldsymbol{\theta}}_j^{k\varepsilon})_{j\le N, j\ne i}$ are i.i.d. and independent of $\boldsymbol{\theta}_i^k$, and $\nabla_1 U$ is bounded, we get by another application of Azuma-Hoeffding inequality, cf. Lemma A.1,

$$\mathbb{P}\Big(Q_2^i(k\varepsilon) \ge K\sqrt{1/N}(\sqrt{D}+u)\Big) \le\ e^{-u^2}\,. \tag{3.24}$$

Therefore, the union bound for $k\in[0,t/\varepsilon]\cap\mathbb{N}$, and $i\le N$ gives

$$\mathbb{P}\Big(\max_{i\le N}\max_{k\in[0,t/\varepsilon]\cap\mathbb{N}} Q_2^i(k\varepsilon) \le K\sqrt{1/N}\cdot\Big(\sqrt{D+\log(N(t/\varepsilon\vee 1))}+z\Big)\Big) \ge 1-e^{-z^2}\,. \tag{3.25}$$

Conditional on the good events in Eq. (3.22) and (3.25), Eq. (3.20) thus yields

$$E_3^i(t) \le \frac{K}{N}\sum_{j=1}^N\int_0^t\|\boldsymbol{\theta}_j^{\lfloor s/\varepsilon\rfloor} - \overline{\boldsymbol{\theta}}_j^{[s]}\|_2\,\mathrm{d}s + Q(t) + \frac{Kt}{N}\,, \tag{3.26}$$

where

$$\begin{aligned}
Q(t) \equiv&\max_{i\le N} Q_1^i(t) + t\cdot\max_{i\le N}\max_{k\in[0,t/\varepsilon]\cap\mathbb{N}} Q_2^i(k\varepsilon)\\
\le&K\sqrt{t\varepsilon}\Big(z+\sqrt{D+\log N}\Big) + tK\sqrt{1/N}\Big(\sqrt{D+\log(N(t/\varepsilon\vee 1))}+z\Big)\\
\le&K(\sqrt{t}\vee t)\cdot\sqrt{1/N\vee\varepsilon}\cdot\Big[\sqrt{D+\log(N(t/\varepsilon\vee 1))}+z\Big].
\end{aligned} \tag{3.27}$$

with probability at least $1 - e^{-z^2}$.

We finally define the random variable

$$\Delta(t; N, \varepsilon) \equiv \max_{i \le N} \sup_{k \in [0, t/\varepsilon] \cap \mathbb{N}} \|\boldsymbol{\theta}_i^k - \overline{\boldsymbol{\theta}}_i^{k\varepsilon}\|_2 \,. \tag{3.28}$$

Using the bounds (3.16), (3.17), (3.26) in Eq. (3.15), we get

$$\Delta(t; N, \varepsilon) \le K \int_0^t \Delta(s; N, \varepsilon) \mathrm{d}s + K\, t\varepsilon + \frac{Kt}{N} + Q(t) \,. \tag{3.29}$$

By Gronwall's inequality, we have

$$\Delta(t; N, \varepsilon) \le K\, e^{Kt} \Big\{ \varepsilon + \frac{1}{N} + Q(t) \Big\} \,. \tag{3.30}$$

Using the bound (3.27), the claim follows. □

Lemma 3.3. *Under the assumptions of Theorem 3, we have*

$$\max_{k \in [0, T/\varepsilon] \cap \mathbb{N}} \Big| R_N(\overline{\boldsymbol{\theta}}^{k\varepsilon}) - R_N(\boldsymbol{\theta}^k) \Big| \le K \cdot \max_{i \le N} \max_{k \in [0, T/\varepsilon] \cap \mathbb{N}} \big\| \boldsymbol{\theta}_i^k - \overline{\boldsymbol{\theta}}_i^{k\varepsilon} \big\|_2 . \tag{3.31}$$

Proof. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_i, \dots, \boldsymbol{\theta}_n)$ and $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}'_i, \dots, \boldsymbol{\theta}_n)$ be two configurations that differ only in position i. Then

$$\begin{aligned} &\big| R_N(\boldsymbol{\theta}) - R_N(\boldsymbol{\theta}') \big| \\ \le &\frac{1}{N} |V(\boldsymbol{\theta}_i) - V(\boldsymbol{\theta}'_i)| + \frac{1}{N^2} |U(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i) - U(\boldsymbol{\theta}'_i, \boldsymbol{\theta}'_i)| + \frac{2}{N^2} \sum_{j \le N, j \ne i} |U(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) - U(\boldsymbol{\theta}'_i, \boldsymbol{\theta}_j)| \\ \le &\frac{K}{N} (\|\boldsymbol{\theta}_i - \boldsymbol{\theta}'_i\|_2 \wedge 1). \end{aligned} \tag{3.32}$$

Then, Eq. (3.31) follows immediately.

□

Lemma 3.4. *Under the assumptions of Theorem 3, we have,*

$$\max_{k \in [0, T/\varepsilon] \cap \mathbb{N}} \Big| R_N(\overline{\boldsymbol{\theta}}^{k\varepsilon}) - R(\rho_{k\varepsilon}) \Big| \le K \sqrt{1/N} \cdot \Big(\sqrt{D + \log(N(T/\varepsilon \vee 1))} + z \Big) \tag{3.33}$$

with probability at least $1 - e^{-z^2}$.

Proof. By Eq. (3.32) and by Azuma-Höeffding inequality and union bound, we get

$$\max_{k \in [0, T/\varepsilon] \cap \mathbb{N}} \Big| R_N(\overline{\boldsymbol{\theta}}^{k\varepsilon}) - \mathbb{E} R_N(\overline{\boldsymbol{\theta}}^{k\varepsilon}) \Big| \le K \sqrt{1/N} \cdot \Big(\sqrt{D + \log(N(T/\varepsilon \vee 1))} + z \Big) \tag{3.34}$$

with probability at least $1 - e^{-z^2}$. The claim follows since

$$\Big| \mathbb{E} R_N(\overline{\boldsymbol{\theta}}^t) - R(\rho_t) \Big| = \frac{1}{N} \Big| \int U(\boldsymbol{\theta}, \boldsymbol{\theta})\, \rho_t(\mathrm{d}\boldsymbol{\theta}) - \int U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\, \rho_t(\mathrm{d}\boldsymbol{\theta}_1)\, \rho_t(\mathrm{d}\boldsymbol{\theta}_2) \Big| \le \frac{K}{N} \,. \tag{3.35}$$

□

The proof of the theorem follows from a straightforward application of Lemma 3.2, 3.3, 3.4. The proof for any bounded Lipschitz function f follows the same argument as Lemma 3.3, 3.4. As a result, for any sequence $(N, \varepsilon = \varepsilon_N)$ with $N/\log(1/\varepsilon_N) \to \infty$ and $\varepsilon_N \to 0$, we have $\hat{\rho}_k^{(N)}$ converges weakly to $\rho_{k\varepsilon}$ almost surely immediately.

3.2 Proof of Theorem 3: Generalization to $\beta < \infty$

Here we generalize the proof given in the previous section to noisy SGD at finite temperature $\beta < \infty$. Since the proof follows the same scheme as in the noiseless case, we will limit ourselves to describing the differences.

Throughout this section we assume that conditions A1, A2, A3 hold. We also let

$$\Psi_\lambda(\boldsymbol{\theta};\rho) = \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2 + V(\boldsymbol{\theta}) + \int U(\boldsymbol{\theta},\boldsymbol{\theta}')\rho(\boldsymbol{\theta}')\mathrm{d}\boldsymbol{\theta}' \tag{3.36}$$

for some $\lambda \le 1$. Further we assume ρ_0 is K_0^2-sub-Gaussian. Finally, we assume $1 \le \beta < \infty$.

For the reader's convenience, we reproduce here the form of the limiting PDE

$$\partial_t \rho_t(\boldsymbol{\theta}) = 2\xi(t)\nabla_{\boldsymbol{\theta}} \cdot \left[\rho_t(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\Psi_\lambda(\boldsymbol{\theta};\rho_t)\right] + 2\xi(t)/\beta \cdot \Delta_{\boldsymbol{\theta}}\rho_t(\boldsymbol{\theta})\,, \tag{3.37}$$

which again should be interpreted in weak sense.

Remark 3.2. Recall conditionss A1, A2, A3 in the main text. By a modified argument of [Szn91, Theorem 1.1], conditions A1 and A3 are sufficient for the existence and uniqueness of solution of PDE (3.37) in weak sense. Section 6 provides further information of this PDE, including a proof of existence and uniqueness.

As in the noiseless case, there is an equivalent formulation of this PDE as a fixed point distribution for the following nonlinear dynamics, which is an integration form of a stochastic differential equation,

$$\overline{\boldsymbol{\theta}}_i^t = \boldsymbol{\theta}_i^0 + 2\int_0^t \xi(s)\,\boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^s;\rho_s)\,\mathrm{d}s + \int_0^t \sqrt{2\xi(s)/\beta}\,\mathrm{d}\boldsymbol{W}_i(s)\,, \tag{3.38}$$

$$\rho_s = \mathrm{P}_{\overline{\boldsymbol{\theta}}_i^s}\,, \tag{3.39}$$

where $\{\boldsymbol{W}_i(s)\}_{s\ge 0}$ for $i \le N$ are independent D-dimensional Brownian motions, and $\boldsymbol{G}(\boldsymbol{\theta};\rho) \equiv -\nabla\Psi_\lambda(\boldsymbol{\theta};\rho)$. The assumptions on U, V, λ, and ξ ensures that this nonlinear dynamics has a unique continuous solution.

This nonlinear dynamics should be compared with the noisy SGD dynamics [11] in the main text that can be written as follows for $k \in \mathbb{N}$:

$$\boldsymbol{\theta}_i^k = \boldsymbol{\theta}_i^0 + 2\varepsilon\sum_{\ell=0}^{k-1}\xi(\ell\varepsilon)\,\boldsymbol{F}_i(\boldsymbol{\theta}^\ell;\boldsymbol{z}_\ell) + \int_0^{k\varepsilon}\sqrt{2\xi([s])/\beta}\,\mathrm{d}\boldsymbol{W}_i(s)\,, \tag{3.40}$$

where

$$\boldsymbol{F}_i(\boldsymbol{\theta};\boldsymbol{z}_k) = -\lambda\boldsymbol{\theta}_i + \big(y_k - \hat{y}(\boldsymbol{x}_k;\boldsymbol{\theta})\big)\,\nabla_{\boldsymbol{\theta}_i}\sigma_*(\boldsymbol{x}_k;\boldsymbol{\theta}_i), \qquad \boldsymbol{\theta} = (\boldsymbol{\theta}_i)_{i\le N} \in \mathbb{R}^{D\times N}. \tag{3.41}$$

It is convenient to collect some standard estimates about the solution of the stochastic differential equation (3.38).

Lemma 3.5. *Assume ρ_0 is K_0^2-sub-Gaussian, $\xi(s)$ and $\boldsymbol{G}(\mathbf{0};\rho_s)$ are K_0-bounded, $\boldsymbol{G}(\boldsymbol{\theta};\rho_s)$ is K_0-Lipschitz in $\boldsymbol{\theta}$, and $\beta \ge 1$. Let $(\overline{\boldsymbol{\theta}}_i^t)_{t\ge 0}$ for $i \le N$ be the solution of (3.38) with independent initialization $(\boldsymbol{\theta}_i^0)_{i\le N} \sim \rho_0$. Let $(\rho_t)_{t\ge 0}$ be the solution of PDE (3.37). Then there exists a constant K depending uniquely on K_0, such that*

$$\mathbb{P}\Big(\sup_{i\le N}\sup_{t\in[0,T]}\|\overline{\boldsymbol{\theta}}_i^t\|_2 \le Ke^{KT}[\sqrt{D+\log N}+z]\Big) \ge 1 - e^{-z^2}, \tag{3.42}$$

and

$$\mathbb{P}\Big(\sup_{i\le N}\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\sup_{u\in[0,\varepsilon]}\|\overline{\boldsymbol{\theta}}_i^{k\varepsilon+u}-\overline{\boldsymbol{\theta}}_i^{k\varepsilon}\|_2\le Ke^{KT}\Big[\sqrt{D+\log(N(T/\varepsilon\vee 1))}+z\Big]\sqrt{\varepsilon}\Big)\ge 1-e^{-z^2},\ (3.43)$$

and for any $t,h\ge 0$, $t+h\le T$,

$$d_{\mathrm{BL}}(\rho_t,\rho_{t+h})\le W_2(\rho_t,\rho_{t+h})\le Ke^{KT}\sqrt{D\,h}\,.\quad(3.44)$$

Proof. We decompose the proof into three parts.

Part (a). First, note that for any D-dimensional K_0^2-sub-Gaussian random vector $\boldsymbol{X}$, we have

$$\mathbb{E}_{\boldsymbol{X}}[\exp\{\tau\|\boldsymbol{X}\|_2^2/2\}]=\mathbb{E}_{\boldsymbol{X},\boldsymbol{G}}[\exp\{\tau\langle\boldsymbol{G},\boldsymbol{X}\rangle\}]\le\mathbb{E}_{\boldsymbol{G}}[\exp\{\tau K_0^2\|\boldsymbol{G}\|_2^2\}/2]=(1-\tau K_0^2)^{-D/2}.\quad(3.45)$$

Note that $(\boldsymbol{\theta}_i^0)_{i\le N}\sim\rho_0$ independently, and ρ_0 is K_0^2-sub-Gaussian. Therefore

$$\mathbb{P}(\|\boldsymbol{\theta}_i^0\|_2\ge u)\le\mathbb{E}[\exp(\tau\|\boldsymbol{\theta}_i\|_2^2/2)]/\exp\{\tau z^2/2\}\le(1-\tau K_0^2)^{-D/2}\exp\{-\tau u^2/2\}.$$

Taking union bound over $i\le N$ gives

$$\mathbb{P}\Big(\max_{i\le N}\|\boldsymbol{\theta}_i^0\|_2\ge u\Big)\le(1-\tau K_0^2)^{-D/2}\exp\{-\tau u^2/2+\log N\}.$$

Taking $\tau=1/(2K_0^2)$ and $u=2K_0(\sqrt{D+\log N}+z)$, we get

$$\mathbb{P}\Big(\max_{i\le N}\|\boldsymbol{\theta}_i^0\|_2\ge 2K_0(\sqrt{D+\log N}+z)\Big)\le\exp\{-z^2\}.\quad(3.46)$$

Then we define $\boldsymbol{W}_{\xi,i}(t)\equiv\int_0^t\sqrt{2\xi(s)}\,\mathrm{d}\boldsymbol{W}_i(s)$. We have $\mathrm{Var}(W_{\xi,i}^j(t))=\int_0^t 2\xi(s)\mathrm{d}s\le 2K_0t$ for $j\le D$. Note $\exp\{\tau\|\boldsymbol{W}_{\xi,i}(t)\|_2^2\}$ is a submartingale, due to Doob's martingale inequality, we have

$$\mathbb{P}\Big(\sup_{t\le T}\|\boldsymbol{W}_{\xi,i}(t)\|_2\ge u\Big)\le\mathbb{E}[\exp\{\tau\|\boldsymbol{W}_{\xi,i}(T)\|_2^2/2\}]\cdot\exp\{-\tau u^2/2\}\le(1-2K_0T\tau)^{-D/2}\exp\{-\tau u^2/2\}.$$

Taking union bound over $i\le N$ gives

$$\mathbb{P}\Big(\max_{i\le N}\sup_{t\le T}\|\boldsymbol{W}_{\xi,i}(t)\|_2\ge u\Big)\le(1-2K_0T\tau)^{-D/2}\exp\{-\tau u^2+\log N\}.$$

Taking $\tau=1/(4K_0T)$ and $u=4\sqrt{K_0T}(\sqrt{D+\log N}+z)$, we get

$$\mathbb{P}\Big(\max_{i\le N}\sup_{t\le T}\|\boldsymbol{W}_{\xi,i}(t)\|_2\ge 4\sqrt{K_0T}(\sqrt{D+\log N}+z)\Big)\le\exp\{-z^2\}.\quad(3.47)$$

By noting that $\xi(s)$, $\boldsymbol{G}(\mathbf{0};\rho_s)$ are K_0-bounded, and $\boldsymbol{G}(\boldsymbol{\theta};\rho_s)$ is K_0-Lipschitz in $\boldsymbol{\theta}$, according to Eq. (3.38), there exists some constant K depending on K_0, such that

$$\Delta_i(t)\le K\int_0^t\Delta_i(s)\mathrm{d}s+K[W/\sqrt{\beta}+\Theta],$$

where $\Delta_i(t)\equiv\sup_{s\le t}\|\overline{\boldsymbol{\theta}}_i^s\|_2$, $W\equiv\max_{i\le N}\sup_{t\le T}\|\boldsymbol{W}_{\xi,i}(t)\|_2$, and $\Theta\equiv\max_{i\le N}\|\boldsymbol{\theta}_i^0\|_2$. Due to Gronwall's inequality, we have

$$\Delta_i(T)\le K\exp(KT)[W/\sqrt{\beta}+\Theta].$$

The high probability bound (3.42) holds by noting the high probability bound for Θ and W in Eq. (3.46) and (3.47).

Part (b). Define $\Delta_i(h;k,\varepsilon) = \sup_{0\le u\le h}\|\overline{\boldsymbol{\theta}}_i^{k\varepsilon+u} - \overline{\boldsymbol{\theta}}_i^{k\varepsilon}\|_2$. By noting that $\xi(s)$, $\boldsymbol{G}(\mathbf{0};\rho_s)$ are K_0-bounded, and $\boldsymbol{G}(\boldsymbol{\theta};\rho_s)$ is K_0-Lipschitz in $\boldsymbol{\theta}$, according to Eq. (3.38), we have

$$\Delta_i(h;k,\varepsilon) \le K\Big[\sup_{s\le T}\|\overline{\boldsymbol{\theta}}_i^s\|_2 + 1\Big]h + \frac{1}{\sqrt{\beta}}\sup_{0\le u\le h}\big\|\boldsymbol{W}_{\xi,i,k}(u)\big\|_2, \tag{3.48}$$

where $\boldsymbol{W}_{\xi,i,k}(u) \equiv \int_{k\varepsilon}^{k\varepsilon+u}\sqrt{2\xi(s)}\,\mathrm{d}\boldsymbol{W}_i(s)$. Similar to the bound Eq. (3.47), we have

$$\mathbb{P}\Big(\max_{i\le N}\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\sup_{0\le u\le h}\|\boldsymbol{W}_{\xi,i,k}(u)\|_2 \le 4\sqrt{K_0 h}\Big(\sqrt{D+\log(N(T/\varepsilon\vee 1))}+z\Big)\Big) \ge 1-e^{-z^2}. \tag{3.49}$$

Plugging the bound Eq. (3.42) and Eq. (3.49) into Eq. (3.48), we have

$$\begin{aligned}\max_{i\le N}\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\Delta_i(h;k,\varepsilon) \le& Ke^{KT}[\sqrt{D+\log N}+z]h + K\Big(\sqrt{D+\log(N(T/\varepsilon\vee 1))}+z\Big)\sqrt{h}\\ \le& Ke^{KT}\Big[\sqrt{D+\log(N(T/\varepsilon\vee 1))}+z\Big]\sqrt{h}\end{aligned}$$

with probability at least $1-e^{-z^2}$.

Part (c). Equation (3.44) holds directly by noting that

$$W_2(\rho_t,\rho_{t+h})^2 \le \mathbb{E}\{\|\overline{\boldsymbol{\theta}}^t - \overline{\boldsymbol{\theta}}^{t+h}\|_2^2\}$$

and applying a integration over z in a modified version of Eq. (3.43) without union bound over $i\le N$ and $k\in[0,T/\varepsilon]\cap\mathbb{N}$.

□

As in the noiseless case, the key step consists in bounding the difference between the nonlinear dynamics and the SGD dynamics.

Lemma 3.6. *Under the assumptions of Theorem 3, there exists a constant K depending uniquely on K_0, K_1, K_2, K_3, such that for any $T\ge 0$, we have*

$$\max_{i\le N}\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}}\|\boldsymbol{\theta}_i^k - \overline{\boldsymbol{\theta}}_i^{k\varepsilon}\|_2 \le Ke^{KT}\cdot\sqrt{1/N\vee\varepsilon}\cdot\Big[\sqrt{D+\log(N(T/\varepsilon\vee 1))}+z\Big] \tag{3.50}$$

with probability at least $1-e^{-z^2}$.

Proof. We take the difference of Eqs. (3.40) and (3.38), for $t\in\mathbb{N}\varepsilon\cap[0,T]$:

$$\begin{aligned}\|\boldsymbol{\theta}_i^{t/\varepsilon} - \overline{\boldsymbol{\theta}}_i^t\|_2 \le& 2\Big\|\int_0^t\Big[\xi(s)\,\boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^s;\rho_s) - \xi([s])\,\boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^{[s]};\rho_{[s]})\Big]\mathrm{d}s\Big\|_2\\ &+ 2\int_0^t\Big\|\xi([s])\,\boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^{[s]};\rho_{[s]}) - \xi([s])\,\boldsymbol{G}(\boldsymbol{\theta}_i^{\lfloor s/\varepsilon\rfloor};\rho_{[s]})\Big\|_2\mathrm{d}s\\ &+ 2\Big\|\varepsilon\sum_{k=0}^{t/\varepsilon-1}\xi(k\varepsilon)\Big\{\boldsymbol{F}_i(\boldsymbol{\theta}^k;\boldsymbol{z}_{k+1}) - \boldsymbol{G}(\boldsymbol{\theta}_i^k;\rho_{k\varepsilon})\Big\}\Big\|_2\\ &+ \Big\|\int_0^t(\sqrt{2\xi(s)/\beta} - \sqrt{2\xi([s])/\beta})\,\mathrm{d}\boldsymbol{W}_i(s)\Big\|_2\\ \equiv& 2E_1^i(t) + 2E_2^i(t) + 2E_3^i(t) + E_4^i(t)\,.\end{aligned} \tag{3.51}$$

Terms $E_2^i(t)$, $E_3^i(t)$ can be bounded the same as in Lemma 3.2, i.e., Eq. (3.17) and (3.26), by noting that the replacement of Ψ by Ψ_λ does not affect these estimates.

To bound $E_4^i(t)$, notice that $\boldsymbol{W}_{\xi,i} \equiv \int_0^T (\sqrt{2\xi(s)} - \sqrt{2\xi([s])})\,\mathrm{d}\boldsymbol{W}_i(s)$ is a Gaussian random vector, $\boldsymbol{W}_{\xi,i} \sim \mathsf{N}(\boldsymbol{0}, \tau^2 \boldsymbol{I}_D)$, where, using the Lipschitz continuity of ξ,

$$\tau^2 = \int_0^T \Big(\sqrt{2\xi(s)} - \sqrt{2\xi([s])}\Big)^2 \,\mathrm{d}s \le K\,T\varepsilon\,.$$

By Gaussian concentration

$$\mathbb{P}(\|\boldsymbol{W}_{\xi,i}\|_2 \ge (\sqrt{D} + z)\tau) \le e^{-z^2/2}\,,$$

and therefore by applying Doob's inequality to the submartingale $t \mapsto E_4^i(t)$, we get

$$\mathbb{P}\Big(\max_{s\le T} E_4^i(s) \ge K(\sqrt{D} + z)\sqrt{T\varepsilon}\Big) \le e^{-z^2/2},$$

and hence

$$\mathbb{P}\Big(\max_{i\le N}\max_{s\le T} E_4^i(s) \le K(\sqrt{D + \log N} + z)\sqrt{T\varepsilon}\Big) \ge 1 - e^{-z^2/2}. \tag{3.52}$$

We need to modify the proof of Lemma 3.2 to bound terms $E_1^i(t)$.

$$\begin{aligned} E_1^i(t) \le& \Big\| \int_0^t [\xi(s) - \xi([s])]\boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^s; \rho_s)\mathrm{d}s\Big\|_2 + \Big\| \int_0^t \xi([s])[\boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^s; \rho_s) - \boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^s; \rho_{[s]})]\mathrm{d}s\Big\|_2 \\ &+ \Big\| \int_0^t \xi([s])\,[\boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^s; \rho_{[s]}) - \boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^{[s]}; \rho_{[s]})]\mathrm{d}s\Big\|_2 \\ \equiv& E_{1,A}^i(t) + E_{1,B}^i(t) + E_{1,C}^i(t). \end{aligned} \tag{3.53}$$

To bound the first term $E_{1,A}^i(t)$, due to the Lipschitz property of $\boldsymbol{G}(\boldsymbol{\theta}; \rho)$ and the boundedness of $\boldsymbol{G}(\boldsymbol{0}; \rho)$, with probability at least $1 - e^{-z^2}$, we have for all $i \le N$ and $t \le T$,

$$\begin{aligned} E_{1,A}^i(t) \le& TK\varepsilon \cdot \sup_{s\in[0,T]} \|\boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^s; \rho_s)\|_2 \le TK\varepsilon \cdot \Big[K \sup_{s\in[0,T]} \|\overline{\boldsymbol{\theta}}_i^s\|_2 + K\Big] \\ \le& Ke^{KT}[\sqrt{D + \log N} + z]\varepsilon. \end{aligned} \tag{3.54}$$

Here the last inequality is due to Eq. (3.42) in Lemma 3.5.

To bound the second term $E_{1,B}^i(t)$, using the fact that $\nabla_1 U$ is bounded Lipschitz, we have for all $i \le N$ and $t \le T$,

$$E_{1,B}^i(t) \le TK \cdot \sup_{\boldsymbol{\theta}\in\mathbb{R}^D} \|\nabla_1 U(\boldsymbol{\theta}; \rho_s) - \nabla_1 U(\boldsymbol{\theta}; \rho_{[s]})\|_2 \le TK^2 \cdot d_{\mathrm{BL}}(\rho_s, \rho_{[s]}) \le Ke^{KT}\sqrt{D\varepsilon}. \tag{3.55}$$

Here the last inequality is due to Eq. (3.44) in Lemma 3.5.

To bound the third term $E_{1,C}^i(t)$, with probability at least $1 - e^{-z^2}$, we have for all $i \le N$ and $t \le T$,

$$\begin{aligned} E_{1,C}^i(t) \le& TK \cdot \sup_{s\in[0,T]} \big\|\boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^s; \rho_{[s]}) - \boldsymbol{G}(\overline{\boldsymbol{\theta}}_i^{[s]}; \rho_{[s]})\big\|_2 \\ \le& TK^2 \cdot \sup_{s\in[0,T]} \big\|\overline{\boldsymbol{\theta}}_i^s - \overline{\boldsymbol{\theta}}_i^{[s]}\big\|_2 \le Ke^{KT}\Big[\sqrt{D + \log(N(T/\varepsilon \vee 1))} + z\Big]\sqrt{\varepsilon}. \end{aligned} \tag{3.56}$$

Here the last inequality is due to Eq. (3.43) in Lemma 3.5.

As a result, combining Eq. (3.17), (3.26), (3.27), (3.51), (3.52), (3.54), (3.55), and (3.56), defining

$$\Delta(t; N, \varepsilon) \equiv \max_{i \le N} \sup_{k \in [0, T/\varepsilon] \cap \mathbb{N}} \|\boldsymbol{\theta}_i^k - \overline{\boldsymbol{\theta}}_i^{k\varepsilon}\|_2 \,, \tag{3.57}$$

we get

$$\Delta(t; N, \varepsilon) \le K \int_0^t \Delta(s; N, \varepsilon) \mathrm{d}s + \frac{Kt}{N} + E(T), \tag{3.58}$$

where

$$E(T) = K e^{KT} \cdot \sqrt{1/N \vee \varepsilon} \cdot \Big[\sqrt{D + \log(N(T/\varepsilon \vee 1))} + z\Big]. \tag{3.59}$$

Applying Gronwall's inequality gives the desired result.

□

The generalization of Theorem 3 to $\beta < \infty$ follows from this lemma exactly as in the previous section.

3.3 Proof of Proposition 2: Monotonicity of the risk

By simple algebra, we have

$$R(\rho_{t+h}) - R(\rho_t) = 2 \int \Psi(\boldsymbol{\theta}; \rho_t)(\rho_{t+h} - \rho_t)(\mathrm{d}\boldsymbol{\theta}) + \langle U, (\rho_{t+h} - \rho_t)^{\otimes 2} \rangle \,. \tag{3.60}$$

By Lemma 3.1, $t \mapsto \rho_t$ is Lipschitz continuous in Wasserstein distance $W_2(\rho_{t_1}, \rho_{t_2}) \le K|t_1 - t_2|$. Hence, we get

$$R(\rho_{t+h}) - R(\rho_t) = 2 \int \Psi(\boldsymbol{\theta}; \rho_t)(\rho_{t+h} - \rho_t)(\mathrm{d}\boldsymbol{\theta}) + O(h^2) \tag{3.61}$$

$$= -4\xi(t) \int \|\nabla \Psi(\boldsymbol{\theta}; \rho_t)\|_2^2 \, \rho_t(\mathrm{d}\boldsymbol{\theta})\, h + o(h) \,, \tag{3.62}$$

where in the second step we used Eq. (3.3). This immediately implies that $R(\rho_t)$ is non-increasing in t.

Let ρ be a fixed point of Eq. (3.1). Since $\partial_t R(\rho_t)|_{\rho_0 = \rho} = 0$, the above formula implies

$$\int \|\nabla \Psi(\boldsymbol{\theta}; \rho)\|_2^2 \, \rho(\mathrm{d}\boldsymbol{\theta}) = 0 \,, \tag{3.63}$$

and therefore ρ is supported in the set of $\boldsymbol{\theta}$'s such that $\nabla \Psi(\boldsymbol{\theta}; \rho) = \mathbf{0}$.

Vice versa, if this is the case, setting $\rho_0 = \rho$, Eq. (3.3) implies $\partial_t \langle \varphi, \rho_t \rangle = 0$, then $\rho_t \equiv \rho_0$ is a fixed point.

3.4 A general continuity result

It is useful to notice that the solution $(\rho_t)_{t \ge 0}$ of the PDE (3.1) is continuous with respect to changes in $V(\,\cdot\,)$, $U(\,\cdot\,,\,\cdot\,)$. Namely, we consider the following two PDEs:

$$\partial_t \rho_t(\boldsymbol{\theta}) = 2\xi(t) \nabla \cdot \big[\rho_t(\boldsymbol{\theta}) \nabla \Psi(\boldsymbol{\theta}; \rho_t)\big] \,, \tag{3.64}$$

$$\partial_t \tilde{\rho}_t(\boldsymbol{\theta}) = 2\xi(t) \nabla \cdot \big[\tilde{\rho}_t(\boldsymbol{\theta}) \nabla \tilde{\Psi}(\boldsymbol{\theta}; \tilde{\rho}_t)\big] \,, \tag{3.65}$$

where

$$\Psi(\boldsymbol{\theta};\rho) = V(\boldsymbol{\theta}) + \int U(\boldsymbol{\theta},\boldsymbol{\theta}')\ \rho(\mathrm{d}\boldsymbol{\theta}')\,, \tag{3.66}$$

$$\tilde{\Psi}(\boldsymbol{\theta};\tilde{\rho}) = \tilde{V}(\boldsymbol{\theta}) + \int \tilde{U}(\boldsymbol{\theta},\boldsymbol{\theta}')\ \tilde{\rho}(\mathrm{d}\boldsymbol{\theta}')\,. \tag{3.67}$$

Lemma 3.7. *Let assumptions* A1, A3 *hold both for V, U and $\tilde{V}, \tilde{U}$, and consider the solutions of Eqs. (3.64) and (3.65) with initial conditions ρ_0, $\tilde{\rho}_0$. Then there exists $K < \infty$ depending only on the constants K_1, K_3 in the assumptions (independent of D), such that*

$$\sup_{t\in[0,T]} d_{\mathrm{BL}}(\rho_t, \tilde{\rho}_t) \le K\, e^{KT}\ \cdot [d_{\mathrm{BL}}(\rho_0,\tilde{\rho}_0) + \varepsilon_0]\,, \tag{3.68}$$

where

$$\varepsilon_0 \equiv \sup_{\boldsymbol{\theta},\boldsymbol{\theta}'\in\mathbb{R}^D} \Big[\|\nabla V(\boldsymbol{\theta}) - \nabla\tilde{V}(\boldsymbol{\theta})\|_2 + \|\nabla_1 U(\boldsymbol{\theta},\boldsymbol{\theta}') - \nabla_1\tilde{U}(\boldsymbol{\theta},\boldsymbol{\theta}')\|_2\Big]. \tag{3.69}$$

Proof. The proof adapts the argument used to establish uniqueness in [Szn91]. Without loss of generality, we fix $\xi(t) \equiv 1/2$. We further denote by K generic constants depending on K_1, K_3.

The assumption of bounded Lipschitz ∇V and ∇U implies that $\nabla\Psi(\boldsymbol{\theta};\rho)$ is K-bounded Lipschitz with respect to argument $(\boldsymbol{\theta},\rho)$, that is,

$$\Big\|\nabla\Psi(\boldsymbol{\theta}_1;\rho_1) - \nabla\Psi(\boldsymbol{\theta}_2;\rho_2)\Big\|_2 \le K\Big[\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 \wedge 1 + d_{\mathrm{BL}}(\rho_1,\rho_2)\Big]. \tag{3.70}$$

The assumption of bounded Lipschitz $\nabla\tilde{V}$ and $\nabla\tilde{U}$ implies that $\nabla\tilde{\Psi}(\boldsymbol{\theta};\rho)$ is K-bounded Lipschitz. Under these conditions, according to [Szn91, Theorem 1.1], there is existence and uniqueness of PDE (3.64) and (3.65). We denote their solutions at time t to be $\rho_t, \tilde{\rho}_t \in \mathscr{P}(\mathbb{R}^D)$ respectively.

Let $\gamma_0 \in \mathscr{P}(\mathbb{R}^D\times\mathbb{R}^D)$ be a coupling of ρ_0 and $\tilde{\rho}_0$ that achieves $2d_{\mathrm{BL}}(\rho_0,\tilde{\rho}_0)$. Given these fixed $(\rho_t)_{t\ge0}$ and $(\tilde{\rho}_t)_{t\ge0}$, consider the nonlinear dynamics

$$\boldsymbol{\theta}^t = \boldsymbol{\theta}^0 - \int_0^t \nabla\Psi(\boldsymbol{\theta}^s;\rho_s)\mathrm{d}s, \tag{3.71}$$

$$\tilde{\boldsymbol{\theta}}^t = \tilde{\boldsymbol{\theta}}^0 - \int_0^t \nabla\tilde{\Psi}(\tilde{\boldsymbol{\theta}}^s;\tilde{\rho}_s)\mathrm{d}s, \tag{3.72}$$

with initialization $(\boldsymbol{\theta}^0,\tilde{\boldsymbol{\theta}}^0) \sim \gamma_0$. As implied by [Szn91, Theorem 1.1], since we have $\boldsymbol{\theta}^0 \sim \rho_0$, $\tilde{\boldsymbol{\theta}}^0 \sim \tilde{\rho}_0$, it follows that $\boldsymbol{\theta}_t \sim \rho_t$, $\tilde{\boldsymbol{\theta}}_t \sim \tilde{\rho}_t$, and therefore

$$d_{\mathrm{BL}}(\rho_t,\tilde{\rho}_t) \le 2\int \Big(\|\boldsymbol{\theta}^t - \tilde{\boldsymbol{\theta}}^t\|_2 \wedge 1\Big)\ \gamma_0(\mathrm{d}\boldsymbol{\theta}^0, \mathrm{d}\tilde{\boldsymbol{\theta}}^0)\,. \tag{3.73}$$

Taking the difference of Eqs. (3.71) and (3.72), for any $(\boldsymbol{\theta}^0,\tilde{\boldsymbol{\theta}}^0) \in \mathrm{supp}(\gamma_0)$,

$$\begin{aligned}
&\|\boldsymbol{\theta}^t - \tilde{\boldsymbol{\theta}}^t\|_2 \le \int_0^t \Big\|\nabla\Psi(\boldsymbol{\theta}^s;\rho_s) - \nabla\tilde{\Psi}(\tilde{\boldsymbol{\theta}}^s;\tilde{\rho}_s)\Big\|_2\mathrm{d}s + \|\boldsymbol{\theta}^0 - \tilde{\boldsymbol{\theta}}^0\|_2 \\
&\le \int_0^t \Big\|\nabla\Psi(\boldsymbol{\theta}^s;\rho_s) - \nabla\Psi(\tilde{\boldsymbol{\theta}}^s;\tilde{\rho}_s)\Big\|_2\mathrm{d}s + \int_0^t \Big\|\nabla\Psi(\tilde{\boldsymbol{\theta}}^s;\tilde{\rho}_s) - \nabla\tilde{\Psi}(\tilde{\boldsymbol{\theta}}^s;\tilde{\rho}_s)\Big\|_2\mathrm{d}s + \|\boldsymbol{\theta}^0 - \tilde{\boldsymbol{\theta}}^0\|_2 \\
&\equiv E_1(t) + E_2(t) + \|\boldsymbol{\theta}^0 - \tilde{\boldsymbol{\theta}}^0\|_2.
\end{aligned} \tag{3.74}$$

Using bound (3.70), the first term $E_1(t)$ is simply bounded by

$$E_1(t) \leq K \int_0^t \left[\|\boldsymbol{\theta}^s - \tilde{\boldsymbol{\theta}}^s\|_2 \wedge 1 + d_{\mathrm{BL}}(\rho_s, \tilde{\rho}_s) \right] \cdot \mathrm{d}s. \tag{3.75}$$

To bound the second term $E_2(t)$, we have

$$\begin{aligned} E_2(t) \leq & t \times \sup_{\boldsymbol{\theta} \in \mathbb{R}^D, \rho \in \mathscr{P}(\mathbb{R}^D)} \|\nabla \Psi(\boldsymbol{\theta}; \rho) - \nabla \tilde{\Psi}(\boldsymbol{\theta}; \rho)\|_2 \\ \leq & t \times \sup_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^D} \left[\|\nabla V(\boldsymbol{\theta}) - \nabla \tilde{V}(\boldsymbol{\theta})\|_2 + \|\nabla_1 U(\boldsymbol{\theta}, \boldsymbol{\theta}') - \nabla_1 \tilde{U}(\boldsymbol{\theta}, \boldsymbol{\theta}')\|_2 \right] = t \cdot \varepsilon_0, \end{aligned} \tag{3.76}$$

with the definition of ε_0 given by Equation (3.69).

Combining Equation (3.74), (3.75), and (3.76), we have

$$\|\boldsymbol{\theta}^t - \tilde{\boldsymbol{\theta}}^t\|_2 \wedge 1 \leq K \int_0^t \left[\|\boldsymbol{\theta}^s - \tilde{\boldsymbol{\theta}}^s\|_2 \wedge 1 + d_{\mathrm{BL}}(\rho_s, \tilde{\rho}_s) \right] \cdot \mathrm{d}s + t \cdot \varepsilon_0 + \|\boldsymbol{\theta}^0 - \tilde{\boldsymbol{\theta}}^0\|_2 \wedge 1. \tag{3.77}$$

Averaging the above inequality over $(\boldsymbol{\theta}^0, \tilde{\boldsymbol{\theta}}^0) \sim \gamma_0$, and using inequality (3.73), we have

$$\int \|\boldsymbol{\theta}^t - \tilde{\boldsymbol{\theta}}^t\|_2 \wedge 1 \cdot \mathrm{d}\gamma_0 \leq 2 d_{\mathrm{BL}}(\rho_0, \tilde{\rho}_0) + 3K \int_0^t \Big[\int \|\boldsymbol{\theta}^s - \tilde{\boldsymbol{\theta}}^s\|_2 \wedge 1 \cdot \mathrm{d}\gamma_0 \Big] \cdot \mathrm{d}s + t \cdot \varepsilon_0. \tag{3.78}$$

Using Gronwall's inequality, for any $t \in \mathbb{R}$, we have

$$\int \|\boldsymbol{\theta}^t - \tilde{\boldsymbol{\theta}}^t\|_2 \wedge 1 \cdot \gamma_0(\mathrm{d}\boldsymbol{\theta}^0, \mathrm{d}\tilde{\boldsymbol{\theta}}^0) \leq K \exp(Kt) \cdot [d_{\mathrm{BL}}(\rho_0, \tilde{\rho}_0) + \varepsilon_0].$$

Applying Equation (3.73), the result follows. □

3.5 Some properties of the solution of the PDE (3.1)

In this section we prove four lemmas on the properties of the solution of the PDE (3.1), under conditions A1 and A3. All of these facts are quite standard, but we provide complete proofs for them for reader's convenience.

We will use several times the following notations. Let ρ_t be a solution of the PDE (3.1) with initialization ρ_0. Let $(\boldsymbol{\theta}^t)_{t \geq 0}$ be the solution of the ordinary differential equation (ODE)

$$\dot{\boldsymbol{\theta}^t} = -2\xi(t) \nabla \Psi(\boldsymbol{\theta}^t; \rho_t)\,, \tag{3.79}$$

with initial condition $\boldsymbol{\theta}^0$. Without loss of generality, we will assume $\xi(t) = 1/2$ throughout this section. If $\boldsymbol{\theta}^0 \sim \rho_0$, then for any $t \geq 0$, we have $\boldsymbol{\theta}^t \sim \rho_t$. We will denote by $\boldsymbol{\varphi}^t : \mathbb{R}^D \mapsto \mathbb{R}^D$ the map between initial conditions of this ODE, and its state at time t (i.e. $\boldsymbol{\varphi}^t(\boldsymbol{\theta}^0) = \boldsymbol{\theta}^t$). Since $\nabla \Psi(\,\cdot\,; \rho_t)$ is bounded and Lipschitz continuous, it follows that $\boldsymbol{\varphi}^t$ is a homeomorphism on its image by Picard's theorem.

With these notations, ρ_t is the push forward of ρ_0 under $\boldsymbol{\varphi}^t$: $\rho_t = \boldsymbol{\varphi}^t_* \rho_0$. In other words, for any Borel set B, $\rho_t(\boldsymbol{\varphi}^t(B)) = \rho_0(B)$.

Lemma 3.8. *Assume conditions* A1, A3 *hold. Let $(\rho_t)_{t \geq 0}$ be the solution of the PDE (3.1) with initialization ρ_0. Let $\Omega \subseteq \mathbb{R}^D$ be a Borel set. Suppose $\boldsymbol{\varphi}^t(\Omega) \subseteq \Omega$, then we have $\rho_t(\Omega) \geq \rho_0(\Omega)$.*

Proof. The lemma holds immediately by noting that $\rho_t(\Omega) \geq \rho_t(\boldsymbol{\varphi}^t(\Omega)) = \rho_0(\Omega)$. □

Lemma 3.9. *Assume conditions* A1, A3 *hold. Further assume there exists a constant* $K < \infty$ *such that*

$$|\partial_i \Psi(\boldsymbol{\theta}; \rho)| \leq K \cdot \theta_i, \tag{3.80}$$

for any $\boldsymbol{\theta} \in (0, \infty)^D$ *and* $\rho \in \mathcal{P}([0, \infty]^D)$. *Let* $(\rho_t)_{t \geq 0}$ *be the solution of the PDE (3.1) with initial condition* ρ_0 *with* $\rho_0((0, \infty)^D) = 1$. *Then for any* $t < \infty$, $\rho_t((0, \infty)^D) = 1$.

Proof. According to Eqs. (3.80) and (3.79), we have for $i \in [d]$,

$$\theta_i^0 \cdot \exp\{-Kt\} \leq \theta_i^t \leq \theta_i^0 \cdot \exp\{Kt\}. \tag{3.81}$$

Denote

$$\Omega_k(t) = [1/k \cdot \exp\{-Kt\}, k \cdot \exp\{Kt\}]^D. \tag{3.82}$$

Then according to (3.81), we have $\boldsymbol{\varphi}^t(\Omega_k(0)) \subseteq \Omega_k(t)$. Note $\Omega_k(t)$ is increasing in k for fixed t, and $\cup_k \Omega_k(t) = \cup_k \Omega_k(0) = (0, \infty)^D$. Hence,

$$\rho_t((0, \infty)^D) = \lim_{k \to \infty} \rho_t(\Omega_k(t)) \geq \lim_{k \to \infty} \rho_t(\boldsymbol{\varphi}^t(\Omega_k(0))) = \lim_{k \to \infty} \rho_0(\Omega_k(0)) = \rho_0((0, \infty)^D) = 1. \tag{3.83}$$

□

Lemma 3.10. *Let* $(\rho_t)_{t \geq 0}$ *be a continuous curve in a compact metric space* (Ω, d). *Denoting*

$$\mathcal{S}_* \equiv \{\rho_* \in \Omega : \exists (t_k)_{k \geq 1}, \lim_{k \to \infty} t_k = \infty, s.t., \lim_{k \to \infty} d(\rho_{t_k}, \rho_*) = 0\}$$

to be the set of all limiting points of $(\rho_t)_{t \geq 0}$. *Then* $\mathcal{S}_*$ *is a connected compact set.*

Proof. First, it is easy to see that $\mathcal{S}_*$ should be closed. Note that Ω is a compact space, then $\mathcal{S}_*$ should be a compact set. If $\mathcal{S}_* = \{\rho_*\}$ is a singleton, this lemma holds automatically. Therefore, we would like to consider the case when $\mathcal{S}_*$ is not a singleton.

For any $\rho_1, \rho_2 \in \mathcal{S}_*$, and $d(\rho_1, \rho_2) > 0$. We would like to show ρ_1 and ρ_2 are connected in $\mathcal{S}_*$.

We use proof by contradiction. Now suppose ρ_1 and ρ_2 are not connected. Define $\mathcal{A} \subseteq \mathcal{S}_*$ to be the maximal connected subset of $\mathcal{S}_*$ containing ρ_1. It is easy to see that $\mathcal{A}$ is compact. It is also easy to see that its complement $\mathcal{B} \equiv \mathcal{S}_* \setminus \mathcal{A}$ is also a compact set, and $\rho_2 \in \mathcal{B}$. As a result, we have $\mathcal{A} \cup \mathcal{B} = \mathcal{S}_*$, $\mathcal{A} \cap \mathcal{B} = \emptyset$, and $\rho_1 \in \mathcal{A}$, $\rho_2 \in \mathcal{B}$.

Note that Ω is a metric space, so it satisfies T4 separation axiom. Since $\mathcal{A}$ and $\mathcal{B}$ are closed sets and $\mathcal{A} \cap \mathcal{B} = \emptyset$, there exists an open set $\mathcal{O}$, such that $\mathcal{A} \subseteq \mathcal{O}$, $\mathcal{O} \cap \mathcal{B} = \emptyset$. Hence, $\partial\mathcal{O} \subseteq \mathcal{S}_*^c$.

Note that ρ_1 and ρ_2 are limiting points of $(\rho_t)_{t \geq 0}$ which is a continuous curve in Ω. Therefore, it must cross the boundary $\partial\mathcal{O}$ infinite times. That is, there is a sequence $(t_k)_{k \geq 1}$ of time with $\lim_{k \to \infty} t_k = \infty$, such that $\rho_{t_k} \in \partial\mathcal{O}$. But since $\partial\mathcal{O}$ is compact, there exists a limiting point $\rho_* \in \partial\mathcal{O}$, so that a subsequence of sequence ρ_{t_k} converges to ρ_*. Therefore, ρ_* should be a limiting point of $(\rho_t)_{t \geq 0}$. But this contradict with $\partial\mathcal{O} \subseteq \mathcal{S}_*^c$. □

Lemma 3.11. *Under the assumptions of* A1 *and* A3, *further assume that* U, V *are twice continuous differentiable, and that* ρ_0 *has density with respect to the Lebesgue measure, bounded by* M_0. *Then* ρ_t *also has a density, bounded by* $M_t = K\, M_0 \exp\{KDt\}$ *(where* K *depends on the constants in the assumptions).*

Proof. Let $\boldsymbol{J}(\boldsymbol{\theta};t)$ for the Jacobian of $\boldsymbol{\varphi}^t(\,\cdot\,)$ at $\boldsymbol{\theta}^0=\boldsymbol{\theta}$. Then Eq. (3.79) implies that $\boldsymbol{J}(\boldsymbol{\theta};t)$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\boldsymbol{J}(\boldsymbol{\theta};t) = -\nabla^2\Psi(\boldsymbol{\varphi}^t(\boldsymbol{\theta});\rho_t)\,\boldsymbol{J}(\boldsymbol{\theta};t)\,, \tag{3.84}$$

with initial condition $\boldsymbol{J}(\boldsymbol{\theta};0)=\boldsymbol{I}_D$. This implies

$$\frac{\mathrm{d}}{\mathrm{d}t}\lambda_{\min}\big(\boldsymbol{J}(\boldsymbol{\theta};t)\big) \geq -\|\nabla^2\Psi(\boldsymbol{\varphi}^t(\boldsymbol{\theta});\rho_t)\|_{\mathrm{op}}\,\lambda_{\min}\big(\boldsymbol{J}(\boldsymbol{\theta};t)\big)\,. \tag{3.85}$$

Therefore, using the fact that $\|\nabla^2\Psi(\boldsymbol{\theta};\rho_t)\|_{\mathrm{op}}$ is K-bounded, we obtain $\lambda_{\min}\big(\boldsymbol{J}(\boldsymbol{\theta};t)\big)\geq\exp(-Kt)$. Finally, since $\boldsymbol{\varphi}^t$ is a diffeomorphism, we have

$$\rho_t(\boldsymbol{\theta}) = \rho_0\big((\boldsymbol{\varphi}^t)^{-1}(\boldsymbol{\theta})\big)\,\Big|\det\big(\boldsymbol{J}((\boldsymbol{\varphi}^t)^{-1}(\boldsymbol{\theta});t)\big)\Big|^{-1} \tag{3.86}$$

$$\leq \rho_0\big((\boldsymbol{\varphi}^t)^{-1}(\boldsymbol{\theta})\big)\,\exp(KDt)\,. \tag{3.87}$$

This completes the proof. □

3.6 Proof of Theorems 6: Stability conditions

In this section, we will prove the stability result in Theorem 6. Throughout the proof we can assume, without loss of generality, $\xi(t)=1/2$. Indeed $\xi(t)$ amounts just of a change of time. Further we introduce the matrix $\boldsymbol{H}_1=\boldsymbol{H}_1(\delta_{\boldsymbol{\theta}_*})\in\mathbb{R}^{D\times D}$ by

$$\boldsymbol{H}_1(\delta_{\boldsymbol{\theta}_*}) = \nabla^2V(\boldsymbol{\theta}_*)+\nabla^2_{1,1}U(\boldsymbol{\theta}_*,\boldsymbol{\theta}_*)+\nabla^2_{1,2}U(\boldsymbol{\theta}_*,\boldsymbol{\theta}_*)\,, \tag{3.88}$$

$$= \boldsymbol{H}_0(\delta_{\boldsymbol{\theta}_*})+\nabla^2_{1,2}U(\boldsymbol{\theta}_*,\boldsymbol{\theta}_*)\,, \tag{3.89}$$

where $\boldsymbol{H}_0\equiv\boldsymbol{H}_0(\delta_{\boldsymbol{\theta}_*})=\nabla^2V(\boldsymbol{\theta}_*)+\nabla^2_{1,1}U(\boldsymbol{\theta}_*,\boldsymbol{\theta}_*)$. Notice that

$$\langle\boldsymbol{u},\nabla^2_{1,2}U(\boldsymbol{\theta}_*,\boldsymbol{\theta}_*)\boldsymbol{u}\rangle = \mathbb{E}\{\langle\boldsymbol{u},\nabla_{\boldsymbol{\theta}}\sigma_*(\boldsymbol{x};\boldsymbol{\theta}_*)\rangle^2\}\,, \tag{3.90}$$

and therefore $\nabla^2_{1,2}U(\boldsymbol{\theta}_*,\boldsymbol{\theta}_*)\succeq\boldsymbol{0}$, whence $\boldsymbol{H}_1\succeq\boldsymbol{H}_0$.

We first establish the condition for $\rho_*=\delta_{\boldsymbol{\theta}_*}$ to be a fixed point. Note that $\Psi(\boldsymbol{\theta};\rho_*)=V(\boldsymbol{\theta})+U(\boldsymbol{\theta},\boldsymbol{\theta}_*)$ and $\mathrm{supp}(\rho_*)=\{\boldsymbol{\theta}_*\}$. Hence the condition [20] in the main text is satisfied if and only if $\nabla_{\boldsymbol{\theta}}\Psi(\boldsymbol{\theta};\rho_*)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*}=\boldsymbol{0}$, i.e. $\nabla V(\boldsymbol{\theta}_*)+\nabla_1U(\boldsymbol{\theta}_*,\boldsymbol{\theta}_*)=\boldsymbol{0}$.

To establish the stability result of Theorem 6, the following lemma provides a key estimate.

Lemma 3.12. *Under the assumptions of Theorem 6, let $\lambda\equiv\lambda_{\min}(\boldsymbol{H}_0)>0$. Then there exists $r_1,\varepsilon_1,\gamma>0$ such that the following hold*

(*i*) *If* $\mathrm{supp}(\rho)\subseteq\mathsf{B}(\boldsymbol{\theta}_*;r_1)\equiv\{\boldsymbol{\theta}:\ \|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2\leq r_1\}$, *then,*

$$\int\langle(\boldsymbol{\theta}-\boldsymbol{\theta}_*),\nabla\Psi(\boldsymbol{\theta};\rho)\rangle\,\rho(\mathrm{d}\boldsymbol{\theta})\geq\frac{\lambda}{2}\int\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2^2\,\rho(\mathrm{d}\boldsymbol{\theta})\,. \tag{3.91}$$

(*ii*) *If* $\int\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2^2\,\rho(\mathrm{d}\boldsymbol{\theta})\leq\varepsilon_1^2$ *and* $\mathrm{supp}(\rho)\subseteq\mathsf{B}(\boldsymbol{\theta}_*;r_1)$, *then for any* $\boldsymbol{\theta}\in\mathsf{B}(\boldsymbol{\theta}_*;r_1)\setminus\mathsf{B}(\boldsymbol{\theta}_*;r_1/2)$,

$$\langle(\boldsymbol{\theta}-\boldsymbol{\theta}_*),\nabla\Psi(\boldsymbol{\theta};\rho)\rangle\geq\gamma>0\,. \tag{3.92}$$

Proof. Note that

$$\nabla^2\Psi(\boldsymbol{\theta};\rho) = \nabla^2 V(\boldsymbol{\theta}) + \int \nabla_1^2 U(\boldsymbol{\theta},\boldsymbol{\theta}')\,\rho(\mathrm{d}\boldsymbol{\theta}')\,. \tag{3.93}$$

Since $\nabla^2 V(\boldsymbol{\theta})$ is continuous and $\nabla_1^2 U(\boldsymbol{\theta},\boldsymbol{\theta}')$ is bounded continuous, it follows that $\boldsymbol{\theta}\mapsto\nabla^2\Psi(\boldsymbol{\theta};\rho)$ is continuous, and $\rho\mapsto\nabla^2\Psi(\boldsymbol{\theta};\rho)$ is continuous in the weak topology, and in fact $(\boldsymbol{\theta},\rho)\mapsto\nabla^2\Psi(\boldsymbol{\theta};\rho)$ is continuous in the product topology.

Further, we have

$$\nabla^2\Psi(\boldsymbol{\theta}_*;\rho_*) = \nabla^2 V(\boldsymbol{\theta}_*) + \nabla_{11}^2 U(\boldsymbol{\theta}_*,\boldsymbol{\theta}_*) = \boldsymbol{H}_0\,. \tag{3.94}$$

Since $\boldsymbol{H}_0 \succ \boldsymbol{0}$ strictly, for any $\delta>0$ we can choose $r_1 = r_1(\delta) > 0$ such that

$$\nabla^2\Psi(\boldsymbol{\theta};\rho) \succeq (1-\delta)\,\boldsymbol{H}_0\,, \tag{3.95}$$

$$\|\nabla_{12}^2 U(\boldsymbol{\theta}_*,\boldsymbol{\theta}) - \nabla_{12}^2 U(\boldsymbol{\theta}_*,\boldsymbol{\theta}_*)\|_{\mathrm{op}} \le \delta\,, \tag{3.96}$$

for all $\boldsymbol{\theta}\in\mathsf{B}(\boldsymbol{\theta}_*;r_1)$, and ρ such that $\mathrm{supp}(\rho)\subseteq\mathsf{B}(\boldsymbol{\theta}_*;r_1)$. If these conditions hold

$$\langle(\boldsymbol{\theta}-\boldsymbol{\theta}_*),\nabla\Psi(\boldsymbol{\theta};\rho)\rangle = \langle(\boldsymbol{\theta}-\boldsymbol{\theta}_*),\nabla\Psi(\boldsymbol{\theta};\rho)-\nabla\Psi(\boldsymbol{\theta}_*;\rho)\rangle + \langle(\boldsymbol{\theta}-\boldsymbol{\theta}_*),\nabla\Psi(\boldsymbol{\theta}_*;\rho)\rangle \tag{3.97}$$

$$= \langle(\boldsymbol{\theta}-\boldsymbol{\theta}_*),\nabla^2\Psi(\tilde{\boldsymbol{\theta}};\rho)\,(\boldsymbol{\theta}-\boldsymbol{\theta}_*)\rangle + \langle(\boldsymbol{\theta}-\boldsymbol{\theta}_*),\nabla\Psi(\boldsymbol{\theta}_*;\rho)\rangle \tag{3.98}$$

$$\ge (1-\delta)\,\langle(\boldsymbol{\theta}-\boldsymbol{\theta}_*),\boldsymbol{H}_0\,(\boldsymbol{\theta}-\boldsymbol{\theta}_*)\rangle + \langle(\boldsymbol{\theta}-\boldsymbol{\theta}_*),\nabla\Psi(\boldsymbol{\theta}_*;\rho)\rangle\,. \tag{3.99}$$

In order to bound the second term, note that, since $\nabla\Psi(\boldsymbol{\theta}_*;\rho_*)=\boldsymbol{0}$,

$$\nabla\Psi(\boldsymbol{\theta}_*;\rho) = \int\left[\nabla_1 U(\boldsymbol{\theta}_*,\boldsymbol{\theta}') - \nabla_1 U(\boldsymbol{\theta}_*,\boldsymbol{\theta}_*)\right]\rho(\mathrm{d}\boldsymbol{\theta}') = \nabla_{12}^2 U(\boldsymbol{\theta}_*,\boldsymbol{\theta}_*)\boldsymbol{\mu} + \boldsymbol{\xi}\,, \tag{3.100}$$

$$\boldsymbol{\mu} = \int(\boldsymbol{\theta}-\boldsymbol{\theta}_*)\,\rho(\mathrm{d}\boldsymbol{\theta})\,, \tag{3.101}$$

$$\boldsymbol{\xi} = \int\left[\nabla_1 U(\boldsymbol{\theta}_*,\boldsymbol{\theta}') - \nabla_1 U(\boldsymbol{\theta}_*,\boldsymbol{\theta}_*) - \nabla_{12}^2 U(\boldsymbol{\theta}_*,\boldsymbol{\theta}_*)(\boldsymbol{\theta}'-\boldsymbol{\theta}_*)\right]\rho(\mathrm{d}\boldsymbol{\theta}')\,. \tag{3.102}$$

Substituting in Eq. (3.99), we obtain

$$\langle(\boldsymbol{\theta}-\boldsymbol{\theta}_*),\nabla\Psi(\boldsymbol{\theta};\rho)\rangle \ge (1-\delta)\langle(\boldsymbol{\theta}-\boldsymbol{\theta}_*),\boldsymbol{H}_0(\boldsymbol{\theta}-\boldsymbol{\theta}_*)\rangle + \langle(\boldsymbol{\theta}-\boldsymbol{\theta}_*),(\boldsymbol{H}_1-\boldsymbol{H}_0)\boldsymbol{\mu}\rangle + \langle(\boldsymbol{\theta}-\boldsymbol{\theta}_*),\boldsymbol{\xi}\rangle\,. \tag{3.103}$$

By the intermediate value theorem, for any $\boldsymbol{v}\in\mathbb{R}^D$, there exists $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{\theta}) \in [\boldsymbol{\theta}_*,\boldsymbol{\theta}]$ such that

$$\langle\boldsymbol{v},\boldsymbol{\xi}\rangle = \int\langle\boldsymbol{v},[\nabla_{12}^2 U(\tilde{\boldsymbol{\theta}},\boldsymbol{\theta}_*) - \nabla_{12}^2 U(\boldsymbol{\theta}_*,\boldsymbol{\theta}_*)](\boldsymbol{\theta}-\boldsymbol{\theta}_*)\rangle\,\rho(\mathrm{d}\boldsymbol{\theta}) \tag{3.104}$$

$$\ge -\int\|\boldsymbol{v}\|_2\big\|\nabla_{12}^2 U(\tilde{\boldsymbol{\theta}},\boldsymbol{\theta}_*) - \nabla_{12}^2 U(\boldsymbol{\theta}_*,\boldsymbol{\theta}_*)\big\|_{\mathrm{op}}\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2\,\rho(\mathrm{d}\boldsymbol{\theta}) \tag{3.105}$$

$$\ge -\delta\|\boldsymbol{v}\|_2\int\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2\,\rho(\mathrm{d}\boldsymbol{\theta}) \tag{3.106}$$

$$\ge -\delta\|\boldsymbol{v}\|_2\sqrt{\mathsf{Tr}(\boldsymbol{Q})+\|\boldsymbol{\mu}\|_2^2} \tag{3.107}$$

$$\ge -\delta\|\boldsymbol{v}\|_2\sqrt{\mathsf{Tr}(\boldsymbol{Q})} - \delta\|\boldsymbol{v}\|_2\|\boldsymbol{\mu}\|_2\,, \tag{3.108}$$

where $\boldsymbol{Q} = \int(\boldsymbol{\theta}-\mu)(\boldsymbol{\theta}-\mu)^{\mathsf{T}}\,\rho(\mathrm{d}\boldsymbol{\theta})$ is the covariance of $(\boldsymbol{\theta}-\boldsymbol{\theta}_*)$.

Let now consider the claim at point (i). Integrating Eq. (3.103) with respect to $\rho(\mathrm{d}\boldsymbol{\theta})$, we get

$$\int \langle (\boldsymbol{\theta}-\boldsymbol{\theta}_*), \nabla\Psi(\boldsymbol{\theta};\rho)\rangle\, \rho(\mathrm{d}\boldsymbol{\theta}) \geq (1-\delta)\langle \boldsymbol{H}_0, \boldsymbol{Q}+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}\rangle + \langle \boldsymbol{\mu}, (\boldsymbol{H}_1-\boldsymbol{H}_0)\boldsymbol{\mu}\rangle + \langle \boldsymbol{\mu}, \boldsymbol{\xi}\rangle \tag{3.109}$$

$$\geq (1-\delta)\langle \boldsymbol{H}_0, \boldsymbol{Q}\rangle + \langle \boldsymbol{\mu}, (\boldsymbol{H}_1-\delta\boldsymbol{H}_0)\boldsymbol{\mu}\rangle - \delta\|\boldsymbol{\mu}\|_2\sqrt{\mathsf{Tr}(\boldsymbol{Q})} - \delta\|\boldsymbol{\mu}\|_2^2 \tag{3.110}$$

$$\geq (1-\delta)\langle \boldsymbol{H}_0, \boldsymbol{Q}\rangle + \langle \boldsymbol{\mu}, (\boldsymbol{H}_1-\delta\boldsymbol{H}_0)\boldsymbol{\mu}\rangle - \frac{3\delta}{2}\|\boldsymbol{\mu}\|_2^2 - \frac{\delta}{2}\,\mathsf{Tr}(\boldsymbol{Q}) \tag{3.111}$$

$$= \langle (1-\delta)\boldsymbol{H}_0 - \frac{\delta}{2}\boldsymbol{I}, \boldsymbol{Q}\rangle + \langle \boldsymbol{\mu}, (\boldsymbol{H}_1-\delta\boldsymbol{H}_0-\frac{3\delta}{2}\boldsymbol{I})\boldsymbol{\mu}\rangle\,. \tag{3.112}$$

By choosing δ sufficiently small, we can ensure that $(1-\delta)\boldsymbol{H}_0 - (\delta/2)\boldsymbol{I} \succeq \lambda_{\min}(\boldsymbol{H}_0)\boldsymbol{I}/2$, $\boldsymbol{H}_1 - \delta\boldsymbol{H}_0 - (3\delta/2)\boldsymbol{I} \succeq \lambda_{\min}(\boldsymbol{H}_1)\boldsymbol{I}/2$, and therefore

$$\int \langle (\boldsymbol{\theta}-\boldsymbol{\theta}_*), \nabla\Psi(\boldsymbol{\theta};\rho)\rangle\, \rho(\mathrm{d}\boldsymbol{\theta}) \geq \frac{1}{2}\lambda_{\min}(\boldsymbol{H}_0)\,\mathsf{Tr}(\boldsymbol{Q}) + \frac{1}{2}\lambda_{\min}(\boldsymbol{H}_1)\,\|\boldsymbol{\mu}\|_2^2\,, \tag{3.113}$$

which yields the claim (3.91).

Next consider point (ii). In this case, Eq. (3.107) implies

$$\langle (\boldsymbol{\theta}-\boldsymbol{\theta}_*), \boldsymbol{\xi}\rangle \geq -\delta\varepsilon_1\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2\,. \tag{3.114}$$

Substituting in Eq. (3.103), and using $\|\boldsymbol{\mu}\|_2 \leq \varepsilon_1$, we get

$$\begin{aligned}\langle (\boldsymbol{\theta}-\boldsymbol{\theta}_*), \nabla\Psi(\boldsymbol{\theta};\rho)\rangle &\geq (1-\delta)\langle \boldsymbol{H}_0, (\boldsymbol{\theta}-\boldsymbol{\theta}_*)^{\otimes 2}\rangle - \varepsilon_1(\lambda_{\max}(\boldsymbol{H}_1)+\lambda_{\max}(\boldsymbol{H}_0)+\delta)\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2 \\ &\geq (1-\delta)\lambda\left(\frac{r_1}{2}\right)^2 - \varepsilon_1(\lambda_{\max}(\boldsymbol{H}_1)+\lambda_{\max}(\boldsymbol{H}_0)+\delta)r_1\,. \end{aligned} \tag{3.115}$$

This is strictly positive for all ε_1 small enough, hence implying the claim (3.92). □

We are now in position of proving Theorem 6.

Proof of Theorem 6. Let $r_0 = \min(r_1/2, \varepsilon_1/2)$ and assume, without loss of generality $t_0 = 0$, so that $\mathrm{supp}(\rho_0) \subseteq \mathsf{B}(\boldsymbol{\theta}_*; r_0)$. We also define

$$T_1 \equiv \inf\left\{t:\ \int \|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2^2\,\rho_t(\mathrm{d}\boldsymbol{\theta}) > \varepsilon_1^2\right\}, \tag{3.116}$$

$$T_2 \equiv \inf\left\{t:\ \mathrm{supp}(\rho_t) \not\subseteq \mathsf{B}(\boldsymbol{\theta}_*; r_1)\right\}, \tag{3.117}$$

$$T_* \equiv \min(T_1, T_2)\,. \tag{3.118}$$

As usual, we adopt the convention that the infimum of an empty set is equal to $+\infty$.

Define $\varphi_1(\boldsymbol{\theta}) = h(\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2)$, with h to be an non-decreasing function with

$$h(r) = \begin{cases} 0 & \text{if } r < r_1/2, \\ \text{smooth intropolation} & \text{if } r_1/2 \leq r < 5r_1/8, \\ 2r/r_1 - 1 & \text{if } 5r_1/8 \leq r < 7r_1/8, \\ \text{smooth intropolation} & \text{if } 7r_1/8 \leq r < r_1, \\ 1 & \text{if } r \geq r_1. \end{cases} \tag{3.119}$$

For any $t < T_*$, the PDE (3.1) implies

$$\partial_t \langle \varphi_1, \rho_t \rangle = -\int \langle \nabla \varphi_1(\boldsymbol{\theta}), \nabla \Psi(\boldsymbol{\theta}; \rho_t) \rangle \, \rho_t(\mathrm{d}\boldsymbol{\theta}) \tag{3.120}$$

$$= -\frac{2}{r_1} \int h'(\|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2) \langle \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}_*)}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2}, \nabla \Psi(\boldsymbol{\theta}; \rho_t) \rangle \, \rho_t(\mathrm{d}\boldsymbol{\theta}) \tag{3.121}$$

$$\le -\frac{4\gamma}{r_1^2} \, \rho_t\Big(\mathsf{B}(\boldsymbol{\theta}_*; 7r_1/8) \setminus \mathsf{B}(\boldsymbol{\theta}_*; 5r_1/8) \Big) \,, \tag{3.122}$$

where, in the last inequality, we used Lemma 3.12.(ii). Next, define

$$\varphi_2(\boldsymbol{\theta}) = \frac{1}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2^2 \,. \tag{3.123}$$

Applying again Eq. (3.1), we get, for $t \le T_*$,

$$\partial_t \langle \varphi_2, \rho_t \rangle = -\int \langle \nabla \varphi_2(\boldsymbol{\theta}), \nabla \Psi(\boldsymbol{\theta}; \rho_t) \rangle \, \rho_t(\mathrm{d}\boldsymbol{\theta}) \tag{3.124}$$

$$= -\int \langle (\boldsymbol{\theta} - \boldsymbol{\theta}_*), \nabla \Psi(\boldsymbol{\theta}; \rho_t) \rangle \, \rho_t(\mathrm{d}\boldsymbol{\theta}) \tag{3.125}$$

$$\le -\lambda \, \langle \varphi_2, \rho_t \rangle \,. \tag{3.126}$$

Together the last two bounds imply $T_* = \infty$. Indeed assume by contradiction $T_* < \infty$. Then either $T_1 \le T_2$, $T_1 < \infty$, or $T_2 < T_1$, $T_2 < \infty$.

Consider the first case: $T_1 \le T_2$, $T_1 < \infty$. Since $\langle \rho_{T_1}, \varphi_2 \rangle \ge \varepsilon_1^2$ but $\langle \rho_0, \varphi_2 \rangle \le r_0^2 \le \varepsilon_1^2/4$, there exists $t < T_*$ such that $\partial_t \langle \rho_0, \varphi_2 \rangle > 0$. However this contradicts Eq. (3.126). Consider then the second case: $T_2 < T_1$, $T_2 < \infty$. This implies $\langle \rho_{T_2}, \varphi_1 \rangle > 0$, but on the other hand $\langle \rho_0, \varphi_1 \rangle = 0$. Hence, there exists $t < T_*$ such that $\partial_t \langle \rho_0, \varphi_1 \rangle > 0$. However this contradicts Eq. (3.122).

We conclude that $T_* = \infty$ and hence we can apply Eq. (3.126) for any t, thus obtaining $\partial_t \langle \varphi_2, \rho_t \rangle \le -\lambda \, \langle \varphi_2, \rho_t \rangle$ and hence $\langle \varphi_2, \rho_t \rangle \le (r_0^2/2) e^{-\lambda t}$, which concludes the proof. □

3.7 Proof of Theorem 7: Instability conditions

In this section we will prove the instability result of Theorem 7. Throughout the section, we assume $\xi(t) \equiv 1/2$. We will use several times the nonlinear dynamics, defined for ρ_t a solution of Eq. (3.1) with initial condition ρ_0:

$$\dot{\boldsymbol{\theta}}^t = -\nabla \Psi(\boldsymbol{\theta}^t; \rho_t) \,. \tag{3.127}$$

Lemma 3.13. *Let ν be a probability measure on $\mathbb{R}^d$, absolutely continuous with respect to the Lebesgue measure, with density bounded by M, and let $\boldsymbol{u} \in \mathbb{R}^d$ be a unit vector. Further assume that, for some $\boldsymbol{x}_0 \in \mathbb{R}^d$, $r > 0$, we have $\nu(\mathsf{B}(\boldsymbol{x}_0; r)) \ge 1 - \varepsilon$, with $0 < \varepsilon < 1/20$. Then there exists a coupling γ of ν with itself (i.e. a probability distribution on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals $\int \gamma(\,\cdot\,, \mathrm{d}\boldsymbol{x}) = \int \gamma(\mathrm{d}\boldsymbol{x}, \,\cdot\,) = \nu(\,\cdot\,)$) and a constant $L = L(d, r, M)$ such that the following holds. If $(\boldsymbol{x}_1, \boldsymbol{x}_2) \sim \gamma$, then*

$$\gamma\left(\langle \boldsymbol{u}, \boldsymbol{x}_1 - \boldsymbol{x}_2 \rangle \ge \frac{1}{L}; \;\; \boldsymbol{P}_{\boldsymbol{u}}^{\perp}(\boldsymbol{x}_1 - \boldsymbol{x}_2) = \boldsymbol{0} \right) \ge \frac{9}{10} \,, \tag{3.128}$$

where $\boldsymbol{P}_{\boldsymbol{u}}^{\perp} = \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^{\mathsf{T}}$ is the projector orthogonal to vector $\boldsymbol{u}$.

Proof. First consider the case $d = 1$: in this case, the assumption $\nu(\mathsf{B}(\boldsymbol{x}_0; r)) \geq 1-\varepsilon$ is not required. Denote by F the distribution function associated to ν (i.e. $F(x) \equiv \nu((-\infty, x])$). By assumption F is differentiable with $F'(x) \leq M$. In order to construct the desired coupling, let Z be a random variable uniformly distributed in $[0, 1]$. For a small constant $\xi_0 > 0$, define the random variables (X_1, X_2) by letting

$$X_1 = F^{-1}(Z)\,, \tag{3.129}$$

$$X_2 = \begin{cases} F^{-1}(Z - \xi_0) & \text{if } Z > \xi_0, \\ F^{-1}(Z + 1 - \xi_0) & \text{if } Z < \xi_0. \end{cases} \tag{3.130}$$

(Note that X_2 is not defined for $Z = \xi_0$ but this is a zero-probability event.) On the event $\{Z > \xi_0\}$ (which has probability $1 - \xi_0$), we have, for some $W \in [X_1, X_2]$,

$$\xi_0 = F'(W)\,(X_1 - X_2) \leq M(X_1 - X_2)\,. \tag{3.131}$$

By choosing ξ_0 small enough, this proves the claim for $d = 1$.

Consider next $d > 1$ and assume without loss of generality $\boldsymbol{u} = \boldsymbol{e}_1$.

Let $\overline{\nu}(\,\cdot\,) = \nu(\,\cdot\,|\boldsymbol{X} \in \mathsf{B}(\boldsymbol{x}_0; r))$, $\boldsymbol{X}_a^b \equiv (X_a, \dots, X_b)$, and denote by $f_{1|[2,d]}$ the density of $\overline{\nu}(X_1 \in \,\cdot\,|\boldsymbol{X}_2^n)$, and by $f_{[a,b]}$ the density of $\overline{\nu}(\boldsymbol{X}_a^b \in \,\cdot\,)$. We then have

$$f_{1|[2,d]}(x_1|\boldsymbol{x}_2^d) = \frac{f_{[1,d]}(\boldsymbol{x}_1^d)}{f_{[2,d]}(\boldsymbol{x}_2^d)} \leq \frac{M}{f_{[2,d]}(\boldsymbol{x}_2^d)}\,. \tag{3.132}$$

Further, we have

$$\overline{\nu}(\{\boldsymbol{x}:\ f_{[2,d]}(\boldsymbol{x}_2^d) \leq \Delta\}) = \int \mathbf{1}_{f_{[2,d]}(\boldsymbol{x}_2^d) \leq \Delta}\, f_{[2,d]}(\boldsymbol{x}_2^d)\, \mathrm{d}\boldsymbol{x}_2^d \tag{3.133}$$

$$\leq \Delta \int_{\mathsf{B}((\boldsymbol{x}_0)_2^d; r)} \mathrm{d}\boldsymbol{x}_2^d \leq C_d \Delta\, r^{d-1}\,. \tag{3.134}$$

In order to construct the coupling, we sample $\boldsymbol{Z} \sim \nu$. If $\boldsymbol{Z} \notin \mathsf{B}(\boldsymbol{x}_0; r)$, then we take $\boldsymbol{X}_1 = \boldsymbol{X}_2 = \boldsymbol{Z}$. If $\boldsymbol{Z} \in \mathsf{B}(\boldsymbol{x}_0; r)$ and $\max_{x_1} f_{1|[2,d]}(x_1|\boldsymbol{Z}_2^d) > M/\Delta$, we also take $\boldsymbol{X}_1 = \boldsymbol{X}_2 = \boldsymbol{Z}$. Otherwise we have $\boldsymbol{Z} \in \mathsf{B}(\boldsymbol{x}_0; r)$ and $\max_{x_1} f_{1|[2,d]}(x_1|\boldsymbol{Z}_2^d) \leq M/\Delta$, then we sample $(X_{1,1}, X_{2,1})$ from the coupling developed in the case $d = 1$ applied to $f_{1|[2,d]}(\,\cdot\,|\boldsymbol{Z}_2^d)$, and set $\boldsymbol{X}_1 = (X_{1,1}, \boldsymbol{Z}_2^d)$, $\boldsymbol{X}_2 = (X_{2,1}, \boldsymbol{Z}_2^d)$. Now define γ to be the joint distribution of $\boldsymbol{X}_1, \boldsymbol{X}_2$. Then γ is a coupling of ν with itself.

The above analyisis yields

$$\gamma\left(\langle \boldsymbol{u}, \boldsymbol{X}_1 - \boldsymbol{X}_2\rangle \geq \frac{\xi_0 \Delta}{M};\ \ \boldsymbol{P}_{\boldsymbol{u}}^{\perp}(\boldsymbol{X}_1 - \boldsymbol{X}_2) = \boldsymbol{0}\right) \geq 1 - \xi_0 - C_d \Delta\, r^{d-1} - \varepsilon\,. \tag{3.135}$$

Hence, we can choose Δ, ξ_0 small enough so that the claim (3.128) holds. □

For any $u \in \mathbb{R}$, define the level set $\tilde{\mathcal{L}}(u)$,

$$\tilde{\mathcal{L}}(u) \equiv \{\boldsymbol{\theta} \in \mathbb{R}^D:\ \Psi(\boldsymbol{\theta}; \rho_*) \leq u\}\,. \tag{3.136}$$

According to the notation of Theorem 7, we have $\mathcal{L}(\eta) = \tilde{\mathcal{L}}(\Psi(\boldsymbol{\theta}_*; \rho_*) - \eta)$ for any $\eta \in \mathbb{R}$.

Lemma 3.14. *For any $u \in \mathbb{R}$, $\Delta > 0$ such that $\partial\tilde{\mathcal{L}}(u_0)$ is compact for all $u_0 \in (u-\Delta, u)$, there exists $\varepsilon_{0,\#} > 0$ such that the following holds. Let $(\rho_t)_{t\geq t_0}$ be a solution of the PDE (3.1) such that $d_{\mathrm{BL}}(\rho_t, \rho_*) \leq \varepsilon_{0,\#}$ for all $t \geq t_0$. Let $(\boldsymbol{\theta}^t)_{t\geq t_0}$ be a solution of the ODE (3.127) with $\Psi(\boldsymbol{\theta}^{t_0}; \rho_*) \leq u - \Delta$. Then $\Psi(\boldsymbol{\theta}^t; \rho_*) \leq u$ for all $t \geq t_0$.*

Proof. By Sard's theorem [GP10], there exists $u_0 \in (u-\Delta, u)$ such that the boundary $\partial\tilde{\mathcal{L}}(u_0)$ contains no critical points of $\Psi(\,\cdot\,; \rho_*)$. If we define $g_0 = \min_{\boldsymbol{\theta}\in\partial\tilde{\mathcal{L}}(u_0)} \|\nabla\Psi(\boldsymbol{\theta}; \rho_*)\|_2$, the minimum is achieved by compactness, and therefore we have $g_0 > 0$ strictly. Notice that by the differentiability assumptions on V and U, $\partial\tilde{\mathcal{L}}(u_0)$ is a C^1 submanifold of $\mathbb{R}^D$, with $\nabla\Psi(\boldsymbol{\theta}; \rho_*)$ orthogonal to $\partial\tilde{\mathcal{L}}_0(u_0)$ and directed toward the exterior. Further, as observed already above,

$$\|\nabla\Psi(\boldsymbol{\theta}; \rho_t) - \nabla\Psi(\boldsymbol{\theta}; \rho_*)\|_2 = \left\| \int \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}; \boldsymbol{\theta}')(\rho_t - \rho_*)(\mathrm{d}\boldsymbol{\theta}') \right\|_2 \tag{3.137}$$

$$\leq K\, d_{\mathrm{BL}}(\rho_t, \rho_*) \leq K\, \varepsilon_{0,\#}\,. \tag{3.138}$$

By choosing $\varepsilon_{0,\#}$ small enough, we can ensure $\|\nabla\Psi(\boldsymbol{\theta}; \rho_t) - \nabla\Psi(\boldsymbol{\theta}; \rho_*)\|_2 \leq g_0/3$ for all $\boldsymbol{\theta}$ and all $t \geq t_0$.

Assume by contradiction that $\Psi(\boldsymbol{\theta}^{t_1}; \rho_*) > u$ for some $t_1 \geq t_0$, and let $t_* = \sup\{t \leq t_1 : \Psi(\boldsymbol{\theta}^t; \rho_*) \leq u_0\}$. Note that, by continuity of the trajectory, $\boldsymbol{\theta}^{t_*} \in \partial\tilde{\mathcal{L}}(u_0)$. We then must have

$$0 \leq \frac{\mathrm{d}}{\mathrm{d}t}\Psi(\boldsymbol{\theta}^{t_*}; \rho_*) = -\langle \nabla\Psi(\boldsymbol{\theta}^{t_*}; \rho_{t_*}), \nabla\Psi(\boldsymbol{\theta}^{t_*}; \rho_*)\rangle \tag{3.139}$$

$$\leq -\|\nabla\Psi(\boldsymbol{\theta}^{t_*}; \rho_*)\|_2^2 + \|\nabla\Psi(\boldsymbol{\theta}^{t_*}; \rho_*)\|_2 \|\nabla\Psi(\boldsymbol{\theta}^{t_*}; \rho_{t_*}) - \nabla\Psi(\boldsymbol{\theta}^{t_*}; \rho_*)\|_2 \tag{3.140}$$

$$\leq -\frac{2}{3} g_0\, \|\nabla\Psi(\boldsymbol{\theta}^{t_*}; \rho_*)\|_2\,, \tag{3.141}$$

which leads to a contradiction since $\boldsymbol{\theta}^{t_*} \in \partial\tilde{\mathcal{L}}(u_0)$ and hence $\|\nabla\Psi(\boldsymbol{\theta}^{t_*}; \rho_*)\|_2 > 0$. □

To prove Theorem 7, let now assume by contradiction that $\rho_t \Rightarrow \rho_* = p_*\delta_{\boldsymbol{\theta}_*} + (1-p_*)\tilde{\rho}_*$ weakly. Then for any $\varepsilon_0, r_0 > 0$ (to be chosen below), we can find $t_0 = t_0(\varepsilon_0, r_0)$ such that

$$d_{\mathrm{BL}}(\rho_t, \rho_*) \leq \varepsilon_0, \quad |\rho_t(\mathsf{B}(\boldsymbol{\theta}_*; r_0)) - p_*| \leq \varepsilon_0 \tag{3.142}$$

for all $t \geq t_0$. Let $\overline{\rho}_{t_0}$ be the conditional probability measure of ρ_{t_0} given $\boldsymbol{\theta} \in \mathsf{B}(\boldsymbol{\theta}_*; r_0)$. By Lemma 3.11, $\overline{\rho}_{t_0}$ has a density upper bounded by a constant $M = M(\varepsilon_0, t_0)$ (note that $\overline{\rho}_{t_0}(S) \leq \rho_{t_0}(S)/(p_* - \varepsilon_0)$).

Set $\boldsymbol{H}_0 = \boldsymbol{H}_0(\rho_*) = \nabla^2\Psi(\boldsymbol{\theta}_*; \rho_*)$. Since $\boldsymbol{\theta}_*$ is a critical point of $\boldsymbol{\theta} \mapsto \Psi(\boldsymbol{\theta}; \rho_*)$, for any $\delta > 0$, we can find $r_1(\delta) > 0$ such that

$$\boldsymbol{\theta} \in \mathsf{B}(\boldsymbol{\theta}_*; r_1) \;\;\Rightarrow\;\; \left\|\nabla^2\Psi(\boldsymbol{\theta}; \rho_*) - \boldsymbol{H}_0\right\|_{\mathrm{op}} \leq \frac{\delta}{2}\,, \qquad \|\nabla\Psi(\boldsymbol{\theta}_*; \rho_*)\|_2 = 0\,. \tag{3.143}$$

As shown in the proof of Theorem 6, the function $(\boldsymbol{\theta}, \rho) \mapsto \nabla^2\Psi(\boldsymbol{\theta}; \rho)$ is continuous when the space of probability distributions ρ is endowed with the weak topology. Analogously $\rho \mapsto \nabla\Psi(\boldsymbol{\theta}_*; \rho)$ is continuous in the weak topology. Hence for this $\delta > 0$ and $r_1(\delta) > 0$, there exists $\varepsilon_{0,*}(\delta, r_1) > 0$ small enough such that, the following inequalities hold

$$\boldsymbol{\theta} \in \mathsf{B}(\boldsymbol{\theta}_*; r_1),\;\; d_{\mathrm{BL}}(\rho, \rho_*) \leq \varepsilon_{0,*} \Rightarrow \;\; \left\|\nabla^2\Psi(\boldsymbol{\theta}; \rho) - \boldsymbol{H}_0\right\|_{\mathrm{op}} \leq \delta\,, \qquad \|\nabla\Psi(\boldsymbol{\theta}_*; \rho)\|_2 \leq \delta^2\, r_1/2\,. \tag{3.144}$$

Let us emphasize that r_1 depends on δ but can be taken to be independent of ε_0. Further, by an application of the intermediate value theorem, for all $\boldsymbol{\theta} \in \mathsf{B}(\boldsymbol{\theta}_*; r_1)$,

$$\left| \Psi(\boldsymbol{\theta}; \rho_*) - \Psi(\boldsymbol{\theta}_*; \rho_*) - \frac{1}{2} \langle (\boldsymbol{\theta} - \boldsymbol{\theta}_*), \boldsymbol{H}_0 (\boldsymbol{\theta} - \boldsymbol{\theta}_*) \rangle \right| \le \frac{1}{2} \delta \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2^2 \,. \tag{3.145}$$

For $r_0 < r_1$, $\boldsymbol{\theta}^{t_0} \in \mathsf{B}(\boldsymbol{\theta}_*; r_0)$, we let $(\boldsymbol{\theta}^t)_{t \ge t_0}$ be the solution of Eq. (3.127) with this initial condition. We then define

$$t_{\text{exit}}(\boldsymbol{\theta}^{t_0}, r_1) = \inf \left\{ t \ge t_0 \,:\; \boldsymbol{\theta}^t \notin \mathsf{B}(\boldsymbol{\theta}_*; r_1) \right\}, \tag{3.146}$$

$$t_{\text{return}}(\boldsymbol{\theta}^{t_0}, r_0, r_1) = \inf \left\{ t > t_{\text{exit}}(\boldsymbol{\theta}^{t_0}, r_1) \,:\; \boldsymbol{\theta}^t \in \mathsf{B}(\boldsymbol{\theta}_*; r_0) \right\}. \tag{3.147}$$

Lemma 3.15. *Under the conditions of Theorem 7, there exists $r_1 > 0$ and $\varepsilon_{0,*} > 0$ such that, for all $r_0 \le r_1$, $\varepsilon_0 \le \varepsilon_{0,*}$, there exists $T_{\text{UB}}(\varepsilon_0, r_0, r_1, t_0)$ such that the following happens. If $d_{\text{BL}}(\rho_t, \rho_*) \le \varepsilon_0$ and $|\rho_t(\mathsf{B}(\boldsymbol{\theta}_*; r_0)) - p_*| \le \varepsilon_0$ for all $t \ge t_0$ for some t_0, then*

$$\rho_{t_0}\Big(\{ \boldsymbol{\theta}^{t_0} \in \mathsf{B}(\boldsymbol{\theta}_*; r_0) \,:\; t_{\text{exit}}(\boldsymbol{\theta}^{t_0}, r_1) \le T_{\text{UB}}(\varepsilon_0, r_0, r_1, t_0) \} \Big) \ge \frac{1}{3} p_* \,. \tag{3.148}$$

Proof. Let $\boldsymbol{u}$ be an eigenvector of $\boldsymbol{H}_0$ corresponding to the eigenvalue $\lambda_{\min}(\boldsymbol{H}_0) = -\lambda_1$. By condition B1 of Theorem 7, we have $\lambda_1 > 0$. Let $-\lambda_2$ denote the second smallest eigenvalue (which can be positive). We further denote by $\boldsymbol{P} \in \mathbb{R}^{D \times D}$ the orthogonal projector onto the eigenspace corresponding to $\lambda_{\min}(\boldsymbol{H}_0)$ and by $\boldsymbol{P}_\perp = \boldsymbol{I} - \boldsymbol{P}$ the projector onto the orthogonal subspace.

We fix a $\delta \le (\lambda_1 - \lambda_2)/10$. Then we choose $r_1 > 0$ and $\varepsilon_{0,*} > 0$ such that Eq. (3.144) holds, with an additional requirement that $\varepsilon_{0,*} < p_*/10$. We will prove this lemma with this choice of r_1 and $\varepsilon_{0,*}$.

We always denote $(\boldsymbol{\theta}_i^t)_{t \ge t_0}$ to be the solution of Eq. (3.127) with initial condition $\boldsymbol{\theta}_i^{t_0}$, for $i = 1, 2$. First we claim that, for $0 < \delta \le (\lambda_1 - \lambda_2)/10$, assuming

$$\left\| \nabla^2 \Psi(\boldsymbol{\theta}; \rho_t) - \boldsymbol{H}_0 \right\|_{\text{op}} \le \delta \,, \quad \forall t \ge t_0, \quad \forall \boldsymbol{\theta} \in \mathsf{B}(\boldsymbol{\theta}_*; r_1), \tag{3.149}$$

then for any $\boldsymbol{\theta}_1^{t_0}, \boldsymbol{\theta}_2^{t_0} \in \mathsf{B}(\boldsymbol{\theta}_*; r_1)$ with $\boldsymbol{P}_\perp(\boldsymbol{\theta}_1^{t_0} - \boldsymbol{\theta}_2^{t_0}) = \boldsymbol{0}$, we have

$$\|\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t\|_2 \ge \|\boldsymbol{\theta}_1^{t_0} - \boldsymbol{\theta}_2^{t_0}\|_2 \, e^{\lambda_1 (t - t_0)/2} \tag{3.150}$$

for all $t \in [t_0, t_{\text{exit}}(\boldsymbol{\theta}_1^{t_0}, r_1) \wedge t_{\text{exit}}(\boldsymbol{\theta}_2^{t_0}, r_1)]$.

For now we assume this claim holds. Fix $r_0 \le r_1$ and $\varepsilon_0 \le \varepsilon_{0,*}$. Define γ to be the coupling of Lemma 3.13 corresponding to $\boldsymbol{u}$ which is the eigenvector corresponding to the least eigenvalue of $\boldsymbol{H}_0$, and $\nu = \overline{\rho}_{t_0}$ which is the conditional measure of ρ_{t_0} given $\boldsymbol{\theta}^{t_0} \in \mathsf{B}(\boldsymbol{\theta}_*; r_0)$. Note $\overline{\rho}_{t_0}$ has a density upper bounded by a constant $M = M(\varepsilon_0, t_0)$. By Lemma 3.13, we have $\gamma(\mathcal{E}) \ge 9/10$, where

$$\mathcal{E} \equiv \left\{ (\boldsymbol{\theta}_1^{t_0}, \boldsymbol{\theta}_2^{t_0}) \in \mathsf{B}(\boldsymbol{\theta}_*; r_0) \times \mathsf{B}(\boldsymbol{\theta}_*; r_0) : \; \langle \boldsymbol{u}, \boldsymbol{\theta}_1^{t_0} - \boldsymbol{\theta}_2^{t_0} \rangle \ge \frac{1}{Z}; \;\; \boldsymbol{P}_{\boldsymbol{u}}^{\perp}(\boldsymbol{\theta}_1^{t_0} - \boldsymbol{\theta}_2^{t_0}) = \boldsymbol{0} \right\} \tag{3.151}$$

for some $Z = Z(\varepsilon_0, r_0, t_0) > 0$. Now we take $(\boldsymbol{\theta}_1^{t_0}, \boldsymbol{\theta}_2^{t_0}) \in \mathcal{E}$. Note the assumption of this lemma gives $d_{\text{BL}}(\rho_t, \rho_*) \le \varepsilon_0 \le \varepsilon_{0,*}$ for all $t \ge t_0$. According to Eq. (3.144), we have Eq. (3.149) holds, and due to this claim, we have $\|\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t\|_2 \ge (1/Z) e^{\lambda_1 (t - t_0)/2}$ for all $t \in [t_0, t_{\text{exit}}(\boldsymbol{\theta}_1^{t_0}, r_1) \wedge t_{\text{exit}}(\boldsymbol{\theta}_2^{t_0}, r_1)]$.

Define $T_{\text{UB}}(\varepsilon_0, r_0, r_1, t_0) = (2/\lambda_1) \log(2Z\, r_1)$. Then for $t > T_{\text{UB}}$, we have $\|\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t\|_2 \ge 2 r_1$. This is impossible if $\boldsymbol{\theta}_1^t, \boldsymbol{\theta}_2^t \in \mathsf{B}(\boldsymbol{\theta}_*; r_1)$ and hence we deduce $(t_{\text{exit}}(\boldsymbol{\theta}_1^{t_0}, r_1) \wedge t_{\text{exit}}(\boldsymbol{\theta}_2^{t_0}, r_1)) \le T_{\text{UB}}$ for all $(\boldsymbol{\theta}_1^{t_0}, \boldsymbol{\theta}_2^{t_0}) \in \mathcal{E}$.

Therefore, we get

$$\begin{aligned}\frac{9}{10} &\le \gamma(\mathcal{E}) \le \gamma\Big(\Big\{(\boldsymbol{\theta}_1^{t_0}, \boldsymbol{\theta}_2^{t_0}) \in \mathsf{B}(\boldsymbol{\theta}_*; r_0) \times \mathsf{B}(\boldsymbol{\theta}_*; r_0) \ : \ t_{\text{exit}}(\boldsymbol{\theta}_1^{t_0}, r_1) \wedge t_{\text{exit}}(\boldsymbol{\theta}_2^{t_0}, r_1) \le T_{\text{UB}}\Big\}\Big) \\ &\le \gamma\Big(\Big\{\boldsymbol{\theta}_1^{t_0} \in \mathsf{B}(\boldsymbol{\theta}_*; r_0) \ : \ t_{\text{exit}}(\boldsymbol{\theta}_1^{t_0}, r_1) \le T_{\text{UB}}\Big\}\Big) + \gamma\Big(\Big\{\boldsymbol{\theta}_2^{t_0} \in \mathsf{B}(\boldsymbol{\theta}_*; r_0) \ : \ t_{\text{exit}}(\boldsymbol{\theta}_2^{t_0}, r_1) \le T_{\text{UB}}\Big\}\Big) \\ &= 2\,\overline{\rho}_{t_0}\Big(\Big\{\boldsymbol{\theta}^{t_0} \in \mathsf{B}(\boldsymbol{\theta}_*; r_0) \ : \ t_{\text{exit}}(\boldsymbol{\theta}^{t_0}, r_1) \le T_{\text{UB}}\Big\}\Big)\,.\end{aligned}$$

Denoting by S the event in the last expression, we obtain $\rho_{t_0}(S) \ge (p_* - \varepsilon_0)\overline{\rho}_{t_0}(S) \ge (9/20)(p_* - \varepsilon_0) \ge p_*/3$ by noting that $\varepsilon_0 < p_*/10$.

Proof of the claim. Define the quantities

$$x_{\parallel}(t) = \|\boldsymbol{P}(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t)\|_2^2\,, \tag{3.152}$$

$$x_{\perp}(t) = \|\boldsymbol{P}_{\perp}(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t)\|_2^2\,. \tag{3.153}$$

We then have, for $t \in [t_0, t_{\text{exit}}(\boldsymbol{\theta}_1^{t_0}, r_1) \wedge t_{\text{exit}}(\boldsymbol{\theta}_2^{t_0}, r_1)]$,

$$\begin{aligned}\dot{x}_{\parallel}(t) &= 2\langle \boldsymbol{P}(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t), -\nabla\Psi(\boldsymbol{\theta}_1^t; \rho_t) + \nabla\Psi(\boldsymbol{\theta}_2^t; \rho_t)\rangle \\ &\overset{(a)}{=} 2\langle \boldsymbol{P}(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t), -\nabla^2\Psi(\tilde{\boldsymbol{\theta}}^t; \rho_t)(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t)\rangle \\ &= -2\langle (\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t), \boldsymbol{P}\nabla^2\Psi(\tilde{\boldsymbol{\theta}}^t; \rho_t)\boldsymbol{P}(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t)\rangle - 2\langle (\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t), \boldsymbol{P}\nabla^2\Psi(\tilde{\boldsymbol{\theta}}^t; \rho_t)\boldsymbol{P}_{\perp}(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t)\rangle \\ &\overset{(b)}{\ge} 2(\lambda_1 - \delta)\|\boldsymbol{P}(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t)\|_2^2 - 2\delta\|\boldsymbol{P}(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t)\|_2\|\boldsymbol{P}_{\perp}(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t)\|_2 \\ &\ge 2(\lambda_1 - \delta)x_{\parallel}(t) - \delta(x_{\parallel}(t) + x_{\perp}(t))\,,\end{aligned}$$

where in (a) we used the intermediate value theorem (with $\tilde{\boldsymbol{\theta}}^t$ a point between $\boldsymbol{\theta}_1^t$ and $\boldsymbol{\theta}_2^t$), and in (b) we used Eq. (3.149).

Proceeding analogously for $x_{\perp}(t)$, we get (for a new choice of $\tilde{\boldsymbol{\theta}}^t$)

$$\begin{aligned}\dot{x}_{\perp}(t) &= 2\langle \boldsymbol{P}_{\perp}(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t), -\nabla\Psi(\boldsymbol{\theta}_1^t; \rho_t) + \nabla\Psi(\boldsymbol{\theta}_2^t; \rho_t)\rangle \\ &= 2\langle \boldsymbol{P}_{\perp}(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t), -\nabla^2\Psi(\tilde{\boldsymbol{\theta}}^t; \rho_t)(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t)\rangle \\ &= -2\langle (\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t), \boldsymbol{P}_{\perp}\nabla^2\Psi(\tilde{\boldsymbol{\theta}}^t; \rho_t)\boldsymbol{P}_{\perp}(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t)\rangle - 2\langle (\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t), \boldsymbol{P}_{\perp}\nabla^2\Psi(\tilde{\boldsymbol{\theta}}^t; \rho_t)\boldsymbol{P}(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t)\rangle \\ &\le 2(\lambda_2 + \delta)\|\boldsymbol{P}_{\perp}(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t)\|_2^2 + 2\delta\|\boldsymbol{P}(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t)\|_2\|\boldsymbol{P}_{\perp}(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t)\|_2 \\ &\le 2(\lambda_2 + \delta)x_{\parallel}(t) + \delta(x_{\parallel}(t) + x_{\perp}(t))\,.\end{aligned}$$

Summarizing, we obtained the inequalities

$$\dot{x}_{\parallel}(t) \ge (2\lambda_1 - 3\delta)x_{\parallel}(t) - \delta\, x_{\perp}(t)\,, \tag{3.154}$$

$$\dot{x}_{\perp}(t) \le \delta x_{\parallel}(t) + (2\lambda_2 + 3\delta)x_{\perp}(t)\,. \tag{3.155}$$

The matrix of coefficients on the right-hand side is

$$\boldsymbol{A} = \begin{pmatrix} 2\lambda_1 - 3\delta & -\delta \\ \delta & 2\lambda_2 + 3\delta \end{pmatrix}. \tag{3.156}$$

This has a (un-normalized) left eigenvectors $(1, -v)$, $(-v, 1)$ with eigenvalues $\xi_\pm$ given by:

$$v = \frac{1}{\delta}\Big[\lambda_1 - \lambda_2 - 3\delta - \sqrt{(\lambda_1 - \lambda_2 - 3\delta)^2 - \delta^2}\Big]\,, \tag{3.157}$$

$$\xi_\pm = \lambda_1 + \lambda_2 \pm \sqrt{(\lambda_1 - \lambda_2 - 3\delta)^2 - \delta^2}\,. \tag{3.158}$$

Note we took $\delta < (\lambda_1 - \lambda_2)/10$, we have $v > 0$, and $\xi_+ \geq \lambda_1$.

Multiplying the inequalities (3.154), (3.155) by $(1, -v)$, we thus obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(x_\parallel(t) - v\, x_\perp(t)\big) \geq \xi_+ \left(x_\parallel(t) - v\, x_\perp(t)\right). \tag{3.159}$$

Since we assumed $x_\perp(t_0) = 0$, whence, for all $t \in [t_0, t_{\rm exit}(\boldsymbol{\theta}_1^{t_0}, r_1) \wedge t_{\rm exit}(\boldsymbol{\theta}_2^{t_0}, r_1)]$, we have

$$x_\parallel(t) \geq x_\parallel(t) - v\, x_\perp(t) \geq x_\parallel(t_0)\, e^{\xi_+(t-t_0)} \geq x_\parallel(t_0)\, e^{\lambda_1(t-t_0)}. \tag{3.160}$$

□

We next strengthen the last lemma and prove that trajectories that exit $\mathsf{B}(\boldsymbol{\theta}_*; r_1)$ do not re-enter $\mathsf{B}(\boldsymbol{\theta}_*; r_0)$.

Lemma 3.16. *Under the conditions of Theorem 7, there exists $r_{0,*}, r_1 > 0$ (with $r_{0,*} < r_1$) and $\varepsilon_{0,*} > 0$ such that, for all $r_0 \leq r_{0,*}$, $\varepsilon_0 \leq \varepsilon_{0,*}$, there exists $T_{\rm UB}(\varepsilon_0, r_0, r_1, t_0)$ such that the following happens. If $d_{\rm BL}(\rho_t, \rho_*) \leq \varepsilon_0$ and $|\rho_t(\mathsf{B}(\boldsymbol{\theta}_*; r_0)) - p_*| \leq \varepsilon_0$ for all $t \geq t_0$ for some t_0, then*

$$\rho_{t_0}\Big(\big\{\boldsymbol{\theta}^{t_0} \in \mathsf{B}(\boldsymbol{\theta}_*; r_0)\,:\, t_{\rm exit}(\boldsymbol{\theta}^{t_0}, r_1) \leq T_{\rm UB}(\varepsilon_0, r_0, r_1, t_0),\ t_{\rm return}(\boldsymbol{\theta}^{t_0}, r_0, r_1) = \infty\big\}\Big) \geq \frac{1}{3}\, p_*\,. \tag{3.161}$$

Proof. Let $\boldsymbol{P}_+$ be the projector onto the eigenspace of $-\boldsymbol{H}_0$ corresponding to positive eigenvalues, and $\boldsymbol{P}_-$ the projector onto the subspace corresponding to negative eigenvalues, and let $\lambda_0 \equiv \min_{i \leq D} |\lambda_i(\boldsymbol{H}_0)|$ to be the least absolute value of eigenvalue of $\boldsymbol{H}_0$. By condition B1 of Theorem 7, we have $\lambda_0 > 0$. Let $\lambda_{\max}$ denote the largest absolute value of eigenvalue of $\boldsymbol{H}_0$.

Fix a δ such that $0 < \delta \leq \min\{\lambda_0/(1+\lambda_0+\lambda_{\max}), \sqrt{\lambda_0/\lambda_{\max}}, \lambda_1 - \lambda_2, 1\}/10$, where λ_1, λ_2 are as defined in Lemma 3.15. Next we choose r_1 as per Lemma 3.15, and we further require $\lambda_0 r_1^2 \leq \eta_0$, where η_0 is as per condition B3 in the statement of Theorem 7. We take $\varepsilon_{0,*}$ to be the minimum of the parameter $\varepsilon_{0,*}$ as per Lemma 3.15 and the parameter $\varepsilon_{0,\#}$ as per Lemma 3.14, where in Lemma 3.14, we choose $u = \Psi(\boldsymbol{\theta}_*; \rho_*) - \lambda_0 r_1^2/8$, and $\Delta = \lambda_0 r_1^2/8$. Then we will choose smaller r_1 and $\varepsilon_{0,*}$ so that Eq. (3.144) holds. Finally, we take $r_{0,*} = \delta r_1 < r_1$. We will prove this lemma with this choice of r_1, $\varepsilon_{0,*}$, and $r_{0,*}$, and with the same function $T_{\rm UB}$ as per Lemma 3.15.

Define

$$t_*(\boldsymbol{\theta}^{t_0}; r_1, \delta) \equiv \sup\left\{t \in (t_0, t_{\rm exit}(\boldsymbol{\theta}^{t_0}, r_1))\,:\, \|\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_*\|_2 < \delta r_1\right\}, \tag{3.162}$$

and define

$$z_+(t) = \|\boldsymbol{P}_+(\boldsymbol{\theta}^t - \boldsymbol{\theta}_*)\|_2^2\,, \tag{3.163}$$

$$z_-(t) = \|\boldsymbol{P}_-(\boldsymbol{\theta}^t - \boldsymbol{\theta}_*)\|_2^2\,. \tag{3.164}$$

We bound the evolution of these quantities following the same argument used above for $x_\parallel(t)$, $x_\perp(t)$. Namely

$$\begin{aligned}
\dot{z}_+(t) =& 2\langle \boldsymbol{P}_+(\boldsymbol{\theta}^t - \boldsymbol{\theta}_*), -\nabla\Psi(\boldsymbol{\theta}^t;\rho_t) + \nabla\Psi(\boldsymbol{\theta}_*;\rho_t)\rangle - 2\langle \boldsymbol{P}_+(\boldsymbol{\theta}^t - \boldsymbol{\theta}_*), \nabla\Psi(\boldsymbol{\theta}_*;\rho_t)\rangle \\
=& -2\langle \boldsymbol{P}_+(\boldsymbol{\theta}^t - \boldsymbol{\theta}_*), \nabla^2\Psi(\tilde{\boldsymbol{\theta}}^t;\rho_t)(\boldsymbol{\theta}^t - \boldsymbol{\theta}_*)\rangle - 2\langle \boldsymbol{P}_+(\boldsymbol{\theta}^t - \boldsymbol{\theta}_*), \nabla\Psi(\boldsymbol{\theta}_*;\rho_t)\rangle \\
=& -2\langle (\boldsymbol{\theta}^t - \boldsymbol{\theta}_*), \boldsymbol{P}_+\nabla^2\Psi(\tilde{\boldsymbol{\theta}}^t;\rho_t)\boldsymbol{P}_+(\boldsymbol{\theta}^t - \boldsymbol{\theta}_*)\rangle \\
& -2\langle (\boldsymbol{\theta}^t - \boldsymbol{\theta}_*), \boldsymbol{P}_+\nabla^2\Psi(\tilde{\boldsymbol{\theta}}^t;\rho_t)\boldsymbol{P}_-(\boldsymbol{\theta}^t - \boldsymbol{\theta}_*)\rangle - 2\langle \boldsymbol{P}_+(\boldsymbol{\theta}^t - \boldsymbol{\theta}_*), \nabla\Psi(\boldsymbol{\theta}_*;\rho_t)\rangle \\
\geq& 2(\lambda_0 - \delta)\|\boldsymbol{P}_+(\boldsymbol{\theta}^t - \boldsymbol{\theta}_*)\|_2^2 - 2\delta\|\boldsymbol{P}_+(\boldsymbol{\theta}^t - \boldsymbol{\theta}_*)\|_2\|\boldsymbol{P}_-(\boldsymbol{\theta}_1^t - \boldsymbol{\theta}_2^t)\|_2 - \delta^2 r_1\|\boldsymbol{P}_+(\boldsymbol{\theta}^t - \boldsymbol{\theta}_*)\|_2 \\
\geq& (2\lambda_0 - 3\delta)z_+(t) - \delta z_-(t) - \delta^2 r_1\sqrt{z_+(t)}\,.
\end{aligned}$$

For $t \in [t_*(\boldsymbol{\theta}^{t_0}; r_1, \delta), t_{\rm exit}(\boldsymbol{\theta}^{t_0}; r_1)]$, we have $\sqrt{z_+(t) + z_-(t)} \geq \delta r_1$. Using the inequality $\sqrt{a(a+b)} \leq a + b$ holding for non-negative a and b, we have

$$\begin{aligned}
\dot{z}_+(t) &\geq (2\lambda_0 - 3\delta)z_+(t) - \delta z_-(t) - \delta^2 r_1\sqrt{z_+(t)} && (3.165)\\
&\geq (2\lambda_0 - 3\delta)z_+(t) - \delta z_-(t) - \delta\sqrt{z_+(t)(z_+(t) + z_-(t))} && (3.166)\\
&\geq (2\lambda_0 - 3\delta)z_+(t) - \delta z_-(t) - \delta z_+(t) - \delta z_-(t) && (3.167)\\
&\geq (2\lambda_0 - 4\delta)z_+(t) - 2\delta z_-(t)\,. && (3.168)
\end{aligned}$$

Proceeding analogously for z_-, we arrive at the inequalities

$$\begin{aligned}
\dot{z}_+(t) &\geq (2\lambda_0 - 4\delta)z_+(t) - 2\delta z_-(t)\,, && (3.169)\\
\dot{z}_-(t) &\leq 2\delta z_+(t) - (2\lambda_0 - 4\delta)z_-(t)\,, && (3.170)
\end{aligned}$$

for $t \in [t_*(\boldsymbol{\theta}^{t_0}; r_1, \delta), t_{\rm exit}(\boldsymbol{\theta}^{t_0}; r_1)]$. The matrix of coefficients on the right-hand side has a left eigenvector of the form $(-w, 1)$ with corresponding eigenvalue $-\tilde{\xi}$, whereby $\tilde{\xi} = \sqrt{\lambda_0^2 - 4\delta^2}$ and $w = (\lambda_0 - \sqrt{\lambda_0^2 - 4\delta^2})/(2\delta)$. In particular, since $\delta < \lambda_0/10$, we have $\tilde{\xi} \geq \lambda_0/2 > 0$ and $w > 0$. Multiplying the above inequalities by $(-w, 1)$, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(-wz_+(t) + z_-(t)\big) \leq -\tilde{\xi}\big(-wz_+(t) + z_-(t)\big)\,, \tag{3.171}$$

and therefore, for all $t \in [t_*(\boldsymbol{\theta}^{t_0}; r_1, \delta), t_{\rm exit}(\boldsymbol{\theta}^{t_0}; r_1)]$, $z_-(t) \leq w\, z_+(t) + e^{-\tilde{\xi}t}\big(-w\, z_+(0) + z_-(0)\big) \leq wz_+(t) + \delta^2 r_1^2$. In particular, for $t = t_{\rm exit}(\boldsymbol{\theta}^{t_0}; r_1)$, using $z_+(t_{\rm exit}) + z_-(t_{\rm exit}) = r_1^2$, we finally obtain

$$\begin{aligned}
\big\|\boldsymbol{P}_+(\boldsymbol{\theta}^{t_{\rm exit}} - \boldsymbol{\theta}_*)\big\|_2^2 &\geq r_1^2\left(\frac{1 - \delta^2}{1 + w}\right) \geq r_1^2(1 - \delta)\,, && (3.172)\\
\big\|\boldsymbol{P}_-(\boldsymbol{\theta}^{t_{\rm exit}} - \boldsymbol{\theta}_*)\big\|_2^2 &\leq r_1^2\delta\,. && (3.173)
\end{aligned}$$

Using Eq. (3.145), we obtain

$$\begin{aligned}
\Psi(\boldsymbol{\theta}^{t_{\rm exit}};\rho_*) &\leq \Psi(\boldsymbol{\theta}_*;\rho_*) + \frac{1}{2}\langle(\boldsymbol{\theta} - \boldsymbol{\theta}_*), \boldsymbol{H}_0(\boldsymbol{\theta} - \boldsymbol{\theta}_*)\rangle + \frac{1}{2}\delta r_1^2 && (3.174)\\
&\leq \Psi(\boldsymbol{\theta}_*;\rho_*) - \frac{1}{2}\lambda_0\big\|\boldsymbol{P}_+(\boldsymbol{\theta}^{t_{\rm exit}} - \boldsymbol{\theta}_*)\big\|_2^2 + \frac{1}{2}\lambda_{\max}\big\|\boldsymbol{P}_-(\boldsymbol{\theta}^{t_{\rm exit}} - \boldsymbol{\theta}_*)\big\|_2^2 + \frac{1}{2}\delta r_1^2 && (3.175)\\
&\leq \Psi(\boldsymbol{\theta}_*;\rho_*) - \frac{1}{2}\lambda_0 r_1^2 + \frac{1}{2}(1 + \lambda_0 + \lambda_{\max})\delta r_1^2\,. && (3.176)
\end{aligned}$$

Since $\delta \le \lambda_0/(10(1+\lambda_0+\lambda_{\max}))$, we can ensure that $\Psi(\boldsymbol{\theta}^{t_{\text{exit}}};\rho_*) \le \Psi(\boldsymbol{\theta}_*;\rho_*) - \lambda_0 r_1^2/4$. By Lemma 3.14, since $d_{\text{BL}}(\rho_t,\rho_*) \le \varepsilon_{0,*} \le \varepsilon_{0,\#}$ for all $t \ge t_0$, we have $\Psi(\boldsymbol{\theta}^t;\rho_*) \le \Psi(\boldsymbol{\theta}_*;\rho_*) - \lambda_0 r_1^2/8$ for all $t \ge t_{\text{exit}}(\boldsymbol{\theta}^{t_0}; r_1)$. Note for all $\boldsymbol{\theta} \in \mathsf{B}(\boldsymbol{\theta}_*; \delta r_1)$, we have $\Psi(\boldsymbol{\theta};\rho_*) \ge \Psi(\boldsymbol{\theta}_*;\rho_*) - \lambda_{\max}\delta^2 r_1^2/2$. Since $\delta \le \sqrt{\lambda_0/\lambda_{\max}}/10$, we have $\boldsymbol{\theta}^t \notin \mathsf{B}(\boldsymbol{\theta}_*;\delta r_1)$ for all $t \ge t_{\text{exit}}(\boldsymbol{\theta}^{t_0}; r_1)$.

This implies that, for any $\boldsymbol{\theta}^{t_0} \in \mathsf{B}(\boldsymbol{\theta}_*; r_0)$ for $r_0 \le r_{0,*}$ with $t_{\text{exit}}(\boldsymbol{\theta}^{t_0}, r_1) \le T_{\text{UB}}(\varepsilon_0, r_0, r_1, t_0) < \infty$, it will never return to $\mathsf{B}(\boldsymbol{\theta}_*; r_0)$. This gives the desired result. □

Finally we upper bound the probability that $\boldsymbol{\theta}^t \in \mathsf{B}(\boldsymbol{\theta}_*; r_0)$ for some $t > t_0$, given that $\boldsymbol{\theta}^{t_0} \notin \mathsf{B}(\boldsymbol{\theta}_*; r_0)$. We define

$$t_{\text{enter}}(\boldsymbol{\theta}^{t_0}, r_0) = \inf\left\{t \ge t_0 \,:\; \boldsymbol{\theta}^t \in \mathsf{B}(\boldsymbol{\theta}_*; r_0)\right\}. \tag{3.177}$$

Lemma 3.17. *Under the conditions of Theorem 7, for any $\eta > 0$, there exists $r_{0,*} > 0$ and $\varepsilon_{0,*} > 0$ such that, for all $r_0 \le r_{0,*}$, $\varepsilon_0 \le \varepsilon_{0,*}$, the following happens. If $d_{\text{BL}}(\rho_t,\rho_*) \le \varepsilon_0$ and $|\rho_t(\mathsf{B}(\boldsymbol{\theta}_*; r_0)) - p_*| \le \varepsilon_0$ for all $t \ge t_0$ for some t_0, then*

$$\rho_{t_0}\Big(\big\{\boldsymbol{\theta}^{t_0} \notin \mathsf{B}(\boldsymbol{\theta}_*; r_0) \,:\; t_{\text{enter}}(\boldsymbol{\theta}^{t_0}, r_0) = \infty\big\}\Big) \ge 1 - p_* - \eta. \tag{3.178}$$

Proof. Due to condition B2 of Theorem 7, we can choose u_1 with $\Psi(\boldsymbol{\theta}_*;\rho_*) - \eta_0 < u_1 < \Psi(\boldsymbol{\theta}_*;\rho_*)$ (where η_0 is as per condition B3 of Theorem 7) such that $\rho_*(\tilde{\mathcal{L}}(u_1)) \ge 1 - p_* - \eta/2$ (recall the notation $\tilde{\mathcal{L}}$ defined as Eq. (3.136)). By taking $\varepsilon_{0,*}$ small enough, and since $\boldsymbol{\theta} \mapsto \Psi(\boldsymbol{\theta};\rho_*)$ is Lipschitz continuous, we can also choose $u_2 \in (u_1, \Psi(\boldsymbol{\theta}_*;\rho_*))$ such that $\rho_{t_0}(\tilde{\mathcal{L}}(u_2)) \ge 1 - p_* - \eta$. Fix $u_3 \in (u_2, \Psi(\boldsymbol{\theta}_*,\rho_*))$. Applying Lemma 3.14, we can further reduce $\varepsilon_{0,*}$, so that for any initialization $\boldsymbol{\theta}^{t_0} \in \tilde{\mathcal{L}}(u_2)$, we have $\boldsymbol{\theta}^t \in \tilde{\mathcal{L}}(u_3)$ for all t. Further, by continuity of $\Psi(\,\cdot\,;\rho_*)$, we can choose $r_{0,*}$ small enough so that $\mathsf{B}(\boldsymbol{\theta}_*; r_{0,*}) \cap \tilde{\mathcal{L}}(u_3) = \emptyset$, whence

$$\begin{aligned}
&\rho_{t_0}\Big(\big\{\boldsymbol{\theta}^{t_0} \notin \mathsf{B}(\boldsymbol{\theta}_*; r_0) \,:\; t_{\text{enter}}(\boldsymbol{\theta}^{t_0}, r_0) = \infty\big\}\Big) && (3.179)\\
\ge&\rho_{t_0}\Big(\big\{\boldsymbol{\theta}^{t_0} :\; \Psi(\boldsymbol{\theta}^{t_0};\rho_*) < u_2,\;\; t_{\text{enter}}(\boldsymbol{\theta}^{t_0}, r_0) = \infty\big\}\Big) && (3.180)\\
=&\rho_{t_0}\Big(\big\{\boldsymbol{\theta}^{t_0} :\; \Psi(\boldsymbol{\theta}^{t_0};\rho_*) < u_2\big\}) \ge 1 - p_* - \eta. && (3.181)
\end{aligned}$$

□

The proof of Theorem 7 follows immediately from Lemma 3.16 and Lemma 3.17. Indeed, let $\eta = p_*/10$. Take $\varepsilon_0 \le \min\{\varepsilon_{0,*}, p_*/10\}$ where $\varepsilon_{0,*}$ is the minimum of $\varepsilon_{0,*}$ as per Lemma 3.16 and 3.17. Take r_1 as per Lemma 3.16. Take $r_0 \le \min\{r_{0,*}, r_1\}$ where $r_{0,*}$ is the minimum of $r_{0,*}$ as per Lemma 3.16 and 3.17. With this choice of ε_0 and r_0, there exists $t_0 > 0$ such that Eq. (3.142) holds for all $t \ge t_0$. Setting $t_* = T_{\text{UB}}(\varepsilon_0, r_0, r_1, t_0) \ge t_0$ with T_{UB} given in Lemma 3.16. Denoting by $\mathbb{P}_{t_0,\rho_{t_0}}$ be the probability distribution over trajectories of (3.127) with $\boldsymbol{\theta}^{t_0} \sim \rho_{t_0}$, we have

$$\begin{aligned}
\rho_{t_*}(\mathsf{B}(\boldsymbol{\theta}_*; r_0)) =&\mathbb{P}_{t_0,\rho_{t_0}}(\boldsymbol{\theta}^{t_*} \in \mathsf{B}(\boldsymbol{\theta}_*; r_0))\\
=&\mathbb{P}_{t_0,\rho_{t_0}}(\boldsymbol{\theta}^{t_0} \in \mathsf{B}(\boldsymbol{\theta}_*; r_0);\; \boldsymbol{\theta}^{t_*} \in \mathsf{B}(\boldsymbol{\theta}_*; r_0)) + \mathbb{P}_{t_0,\rho_{t_0}}(\boldsymbol{\theta}^{t_0} \notin \mathsf{B}(\boldsymbol{\theta}_*; r_0);\; \boldsymbol{\theta}^{t_*} \in \mathsf{B}(\boldsymbol{\theta}_*; r_0))\\
\le&\mathbb{P}_{t_0,\rho_{t_0}}(\boldsymbol{\theta}^{t_0} \in \mathsf{B}(\boldsymbol{\theta}_*; r_0)) - \mathbb{P}_{t_0,\rho_{t_0}}(\boldsymbol{\theta}^{t_0} \in \mathsf{B}(\boldsymbol{\theta}_*; r_0)\,;\; t_{\text{exit}}(\boldsymbol{\theta}^{t_0}; r_1) < t_*, t_{\text{return}}(\boldsymbol{\theta}^{t_0}; r_0) = \infty)\\
&+ \mathbb{P}_{t_0,\rho_{t_0}}(\boldsymbol{\theta}^{t_0} \notin \mathsf{B}(\boldsymbol{\theta}_*; r_0)) - \mathbb{P}_{t_0,\rho_{t_0}}(\boldsymbol{\theta}^{t_0} \notin \mathsf{B}(\boldsymbol{\theta}_*; r_0);\; t_{\text{enter}}(\boldsymbol{\theta}^{t_0}, r_0) = \infty)\\
\le&1 - \frac{1}{3}p_* - (1 - p_* - \eta) = 2p_*/3 + \eta.
\end{aligned}$$

Since we also had $\rho_t(\mathsf{B}(\boldsymbol{\theta}_*; r_0)) \ge p_* - \varepsilon_0$ for all $t \ge t_0$, note $\eta, \varepsilon_0 \le p_*/10$, we reached a contradiction.

4 Centered isotropic Gaussians

In this section we consider the centered isotropic Gaussians example discussed in the main text. That is, we assume the joint law of $(y, \boldsymbol{x})$ to be as follows:

With probability 1/2: $y = +1$, $\boldsymbol{x} \sim \mathsf{N}(\boldsymbol{0}, (1+\Delta)^2 \boldsymbol{I}_d)$.

With probability 1/2: $y = -1$, $\boldsymbol{x} \sim \mathsf{N}(\boldsymbol{0}, (1-\Delta)^2 \boldsymbol{I}_d)$.

We assume $0 < \Delta < 1$, and choose $\sigma_*(\boldsymbol{x}; \boldsymbol{\theta}_i) = \sigma(\langle \boldsymbol{x}, \boldsymbol{w}_i \rangle)$ for some activation function σ. Define $q(r) \equiv \mathbb{E}\{\sigma(rG)\}$ for $G \sim \mathsf{N}(0,1)$. We assume $\sigma(\,\cdot\,)$ satisfies the following conditions S0 - S4:

- S0 $x \mapsto \sigma(x)$ is bounded, non-decreasing, Lipschitz continuous. Its weak derivative $x \mapsto \sigma'(x)$ is Lipschitz in a neighborhood of 0.
- S1 q is analytic on $(0, \infty)$ with $\sup_{r \in [0,\infty]} q''(r) < \infty$.
- S2 $q'(r) > 0$ for all $r \in (0, \infty)$, with $\sup_{r \in [0,\infty]} q'(r) < \infty$, and $\lim_{r \to 0} q'(r) = \lim_{r \to \infty} q'(r) = 0$.
- S3 $-\infty < q(0+) < -1$, $1 < q(+\infty) < \infty$, and $-1 < (q(0+) + q(+\infty))/2 < 1$.
- S4 Letting $Z(r) \equiv q'(\tau_- r)/q'(\tau_+ r)$ for some $\tau_+ > \tau_- > 0$ we have $Z'(r) > 0$ for all $r \in (0, \infty)$.

Note that condition S1 and part of S2 are implied by S0, but we list them here for conveniency. Some of these assumptions can be relaxed at the cost of extra technical work. In the interest of simplicity, we prefer to avoid being overly general.

As our running example we will use

$$\sigma(t) = \begin{cases} s_1 & \text{if } t \le t_1, \\ (s_2(t-t_1) + s_1(t_2 - t))/(t_2 - t_1) & \text{if } t \in (t_1, t_2), \\ s_2 & \text{if } t \ge t_2. \end{cases} \tag{4.1}$$

In particular, we choose $s_1 = -2.5$, $s_2 = 7.5$, $t_1 = 0.5$, $t_2 = 1.5$ in our simulations. In section 4.5, we check that this choice satisfies the above assumptions.

Throughout this section, we set $\tau_\pm = (1 \pm \Delta)$ and $q_+(r) = q(\tau_+ r)$, $q_-(r) = q(\tau_- r)$. Also, we will assume $\xi(t) = 1/2$, since other choices of $\xi(\,\cdot\,)$ merely amounts to a time reparametrization.

Before analyzing our model, we introduce the function space and space of probability measures we will work on. We equip the set $[0, \infty]$ with a metric $\bar{d}$, where $\bar{d}(x, y) = |1/(1+x) - 1/(1+y)|$ for any $x, y \in [0, \infty]$. Then $([0, \infty], \bar{d})$ is a compact metric space, and we will still denote it by $[0, \infty]$ for simplicity in notations. We denote $C_b([0, \infty])$ to be the set of bounded continuous functions on $[0, \infty]$, where continuity is defined using the topology generated by $\bar{d}$. More explicitly, we have isomorphism

$$C_b([0,\infty]) \simeq \{f \in C([0,\infty)) : \exists f(+\infty) \equiv \lim_{r \to +\infty} f(r), \sup_{r \in [0,\infty]} f(r) < \infty\}. \tag{4.2}$$

Because of condition S2 and S3, we have $q, q' \in C_b([0, \infty])$.

Let $\mathscr{P}([0, \infty])$ be the set of probability measures on $[0, \infty]$. Due to Prokhorov's theorem, there exists a complete metric $\bar{d}_{\mathscr{P}}$ on $\mathscr{P}([0, \infty])$ equivalent to the topology of weak convergence, so that $(\mathscr{P}([0, \infty]), \bar{d}_{\mathscr{P}})$ is a compact metric space. In this section, we will denote by $\overline{\mathscr{P}} = \mathscr{P}([0, \infty])$.

4.1 Statics

Since the distribution of $\boldsymbol{x}$ is invariant under rotations for each of the two classes, so are the functions

$$V(\boldsymbol{w}) = v(\|\boldsymbol{w}\|_2)\,, \qquad U(\boldsymbol{w}_1, \boldsymbol{w}_2) = u_0(\|\boldsymbol{w}_1\|_2, \|\boldsymbol{w}_2\|_2, \langle \boldsymbol{w}_1, \boldsymbol{w}_2\rangle)\,. \tag{4.3}$$

These take the form

$$v(r) = -\frac{1}{2}\, q(\tau_+ r) + \frac{1}{2}\, q(\tau_- r)\,, \qquad q(t) = \mathbb{E}\{\sigma(tG)\} \tag{4.4}$$

$$u_0(r_1, r_2, r_1 r_2 \cos\alpha) = \frac{1}{2}\mathbb{E}\{\sigma(\tau_+ r_1 G_1)\sigma(\tau_+ r_2 G_2)\} + \frac{1}{2}\mathbb{E}\{\sigma(\tau_- r_1 G_1)\sigma(\tau_+ r_2 G_2)\}\,, \tag{4.5}$$

where expectations are with respect to standard normals $G, G_1, G_2 \sim \mathsf{N}(0,1)$, with (G_1, G_2) jointly Gaussian and $\mathbb{E}\{G_1 G_2\} = \cos\alpha$.

In order to minimize $R(\rho)$, it is sufficient to restrict ourselves to distributions that are invariant under rotations. Indeed, for any probability distribution ρ on $\mathbb{R}^d$, we can define its symmetrization by letting, for any Borel set $Q \subseteq \mathbb{R}^d$,

$$\rho_s(Q) \equiv \int \rho(\boldsymbol{R}\, Q)\;\; \mu_{\text{Haar}}(\mathrm{d}\boldsymbol{R})\,, \tag{4.6}$$

where μ_{Haar} is the Haar measure over the group of orthogonal rotations. Since $\rho \mapsto R(\rho)$ is convex, $R(\rho_s) \le R(\rho)$.

We therefore restrict ourselves to ρ's that are invariant under rotations. In other words, under ρ, the vector $\boldsymbol{w}$ is uniformly random conditional on $\|\boldsymbol{w}\|_2$. We denote by $\overline{\rho}$ the probability distribution of $\|\boldsymbol{w}\|_2$ when $\boldsymbol{w} \sim \rho$ and we let $\overline{R}_d(\overline{\rho})$ denote the resulting risk. We then have

$$\overline{R}_d(\overline{\rho}) = 1 + 2\int v(r)\,\overline{\rho}(\mathrm{d}r) + \int u_d(r_1, r_2)\,\overline{\rho}(\mathrm{d}r_1)\,\overline{\rho}(\mathrm{d}r_2)\,, \tag{4.7}$$

$$u_d(r_1, r_2) = \mathbb{E}[u_0(r_1, r_2, r_1 r_2 \cos\Theta)]. \tag{4.8}$$

where $\Theta \sim (1/Z_d)\sin^{d-2}\theta \cdot \mathbf{1}\{\theta \in [0,\pi]\}\mathrm{d}\theta$.

As $d \to \infty$, we have $\lim_{d\to\infty} u_d(r_1, r_2) = u_\infty(r_1, r_2)$ (uniformly over compact sets), with

$$u_\infty(r_1, r_2) = \frac{1}{2}\Big[q(\tau_+ r_1)q(\tau_+ r_2) + q(\tau_- r_1)q(\tau_+ r_2)\Big], \tag{4.9}$$

and the risk function converges to

$$\overline{R}_\infty(\overline{\rho}) = \frac{1}{2}\left(1 - \int q(\tau_+ r)\,\overline{\rho}(\mathrm{d}r)\right)^2 + \frac{1}{2}\left(1 + \int q(\tau_- r)\,\overline{\rho}(\mathrm{d}r)\right)^2. \tag{4.10}$$

We also define

$$\psi_d(r; \overline{\rho}) = v(r) + \int u_d(r, r')\,\overline{\rho}(\mathrm{d}r')\,. \tag{4.11}$$

For $d = \infty$, we have the simpler expression

$$\psi_\infty(r; \overline{\rho}) = \lambda_+(\overline{\rho}) \cdot q_+(r) + \lambda_-(\overline{\rho}) \cdot q_-(r), \tag{4.12}$$

$$\lambda_+(\overline{\rho}) = \frac{1}{2}[\langle q_+, \overline{\rho}\rangle - 1], \tag{4.13}$$

$$\lambda_-(\overline{\rho}) = \frac{1}{2}[\langle q_-, \overline{\rho}\rangle + 1]. \tag{4.14}$$

The following theorem provides a characterization of global minimizers of $\overline{R}_d(\overline{\rho})$.

Proposition 4.1 (Lemma 1 in the main text)**.** *For any $d \le \infty$, define*

$$\psi_d(r;\overline{\rho}) \equiv v(r) + \int u_d(r,r')\,\overline{\rho}(\mathrm{d}r')\,. \tag{4.15}$$

Then

1. *$\overline{\rho}_*$ is a global minimizer of $\overline{R}_d(\overline{\rho})$ if and only if* $\mathrm{supp}(\overline{\rho}_*) \subseteq \arg\min_r \psi_d(r;\overline{\rho}_*)$.
2. *In particular, $\overline{\rho}_* = \delta_{r_*}$ is a global minimizer or $\overline{R}_d(\overline{\rho})$ if and only if $v(r) + u_d(r,r_*) \ge v(r_*) + u(r_*,r_*)$ for all r.*

Proof. Point 1 is essentially a special case of the second part of Proposition 1 in the main text (cf. Eq. (2.7)) and follows by the same argument. Point 2 is follows by taking $\overline{\rho}_* = \delta_{r_*}$. □

Given the last result, it is interesting to understand whether the optimal radial distribution $\overline{\rho}_*$ is a single point mass or not. Under the ansatz $\overline{\rho} = \delta_r$ (a single point mass at radius r) we obtain an effective risk $\overline{R}_d^{(1)}(r) \equiv \overline{R}_d(\delta_r)$ defined by $\overline{R}_d^{(1)}(r) = 1 + 2v(r) + u_d(r,r)$, which is plotted in Figure 7.6 for the case of our running example (4.1), and $\Delta = 0.4$.

Let $r_* = r_*(\Delta, d)$ be the minimizer of $\overline{R}_d^{(1)}(r)$, and define, for $d \le \infty$,

$$\Delta_d = \sup\left\{\Delta : \;\; v(r) + u_d(r,r_*) \ge v(r_*) + u_d(r_*,r_*),\ \forall r \ge 0\right\}. \tag{4.16}$$

In the case $d = \infty$, the minimization problem simplifies further. Either the minimum risk is 0, or it is achieved at a point mass $\overline{\rho}_* = \delta_{r_*}$.

Theorem 4.2. *Consider $d = \infty$. Recall $\overline{\mathscr{P}} = \mathscr{P}([0,\infty])$. In this case Δ_∞ defined as per Eq. (4.16) is such that $\Delta_\infty \in (0,1)$. Further*

1. *For $\Delta < \Delta_\infty$, $\inf_{\overline{\rho}\in\overline{\mathscr{P}}} \overline{R}_\infty(\overline{\rho}) > 0$ and the unique global minimizer of risk function $\overline{R}_\infty(\overline{\rho})$ is a point mass located at some $r_*(\Delta) \in (0,\infty)$.*
2. *For $\Delta \ge \Delta_\infty$, all global minimizers of risk function $\overline{R}_\infty(\overline{\rho})$ have risk zero, and there exists a global minimizer that has compact support bounded away from 0.*

Proof of Theorem 4.2. Recall the definitions $q_+(r) = q(\tau_+ r)$ and $q_-(r) = q(\tau_- r)$. Further, we define the set $\Gamma \subseteq [0,1]$ by

$$\Gamma = \{\Delta : \exists r \in (0,+\infty),\ \text{s.t.},\ q_+(r) \ge 1 \text{ and } q_-(r) \le -1\}. \tag{4.17}$$

According to condition S3, for $\Delta = 1$, we have $q_-(r) = q(0) < -1$ and $q_+(+\infty) = q(+\infty) > +1$. Since q is continuous, it is easy to see that there exists an $\varepsilon > 0$, such that $[1-\varepsilon, 1] \subseteq \Gamma$. Further, for $\Delta = 0$ we have $q_+(r) = q_-(r)$. By continuity, there exists an $\varepsilon > 0$, such that $[0,\varepsilon] \in [0,1] \setminus \Gamma$.

Since q is an increasing function, we have

$$\Gamma = [\Delta_\infty, 1]\,, \qquad \Delta_\infty = \inf_{\Delta\in\Gamma} \Delta. \tag{4.18}$$

By the remarks above, we have $0 < \Delta_\infty < 1$. Notice that this definition does not coincide with the one in Eq. (4.16). However, the proof below (together with Proposition 4.1) implies that the two definitions actually coincide.

Part (1): $\Delta < \Delta_\infty$.

Step 1. Prove that $\inf_{\overline{\rho}\in\overline{\mathscr{P}}} \overline{R}_\infty(\overline{\rho}) > 0$ as $\Delta < \Delta_\infty$.

First, we consider the optimization problem

$$f_* \equiv \sup_{\overline{\rho}\in\overline{\mathscr{P}}} \Big\{ \langle q_+, \overline{\rho}\rangle - 1 \quad \text{s.t.} \quad \langle q_-, \overline{\rho}\rangle \le -1 \Big\}. \tag{4.19}$$

We claim that, for $\Delta < \Delta_\infty$ we have $f_* < 0$. Indeed, for any $\lambda \in [0, +\infty)$, we have the following upper bound

$$f_* \le \sup_{\overline{\rho}\in\overline{\mathscr{P}}} \{ L(\overline{\rho}, \lambda) \equiv \langle q_+, \overline{\rho}\rangle - 1 - \lambda(\langle q_-, \overline{\rho}\rangle + 1)\}. \tag{4.20}$$

Since $q_+ - \lambda\, q_- \in C_b([0, +\infty])$, then $L(\,\cdot\,, \lambda)$ is continuous in $\overline{\rho}$ in weak topology. By the compactness of $\overline{\mathscr{P}}$, the supremum of $L(\,\cdot\,, \lambda)$ is attained by some $\overline{\rho}_\lambda \in \overline{\mathscr{P}}$. This $\overline{\rho}_\lambda$ should satisfy

$$\mathrm{supp}(\overline{\rho}_\lambda) \subseteq \mathrm{argmax}_{r\in[0,+\infty]} \{ q_+(r) - \lambda q_-(r)\}.$$

Let $h(r) \equiv q_+(r) - \lambda q_-(r)$. Note the supremum of h should either satisfy

$$h'(r) = q_+'(r) - \lambda q_-'(r) = 0, \tag{4.21}$$

for $r \in (0, \infty)$, or the supremum should be attained at the boundary 0 or $+\infty$. According to condition S4, $[q_-'(r)/q_+'(r)]' > 0$ for $r \in (0, \infty)$, the equation (4.21) has at most one solution $r_* \in (0, \infty)$.

Assume that there exists $r_* \in (0, \infty)$ such that $h'(r_*) = 0$. Then we have $h'(r) > 0$ for $0 < r < r_*$, and $h'(r) < 0$ for $r_* < r < +\infty$, whence $\mathrm{supp}(\overline{\rho}_\lambda) = \{r_*\}$. If $h'(r) = 0$ does not have a solution in $(0, \infty)$, the only supremum of $h(r)$ could be achieved at 0 or $+\infty$. Therefore, $\mathrm{supp}(\overline{\rho}_\lambda) = \{0\}$ or $\mathrm{supp}(\overline{\rho}_\lambda) = \{+\infty\}$. This concludes that, for any $\lambda \in [0, +\infty)$, $\sup_{\overline{\rho}\in\overline{\mathscr{P}}} L(\overline{\rho}, \lambda)$ is achieved by a point mass. Therefore, we have

$$f_* \le \inf_{\lambda\in[0,+\infty)} \sup_{r\in[0,+\infty]} \{ q_+(r) - 1 - \lambda(q_-(r) + 1)\} = q_+(q_-^{-1}(-1)) - 1.$$

For $\Delta < \Delta_\infty$, the right hand side of the above inequality is less than 0. Therefore, we cannot have a probability distribution $\overline{\rho}$ such that $\langle q_+, \overline{\rho}\rangle = 1$ and $\langle q_-, \overline{\rho}\rangle = -1$. The infimum of the risk cannot be 0.

Step 2. Show that the global minimizer should be a delta function for $\Delta < \Delta_\infty$.

According to Proposition 1, the global minimizer $\overline{\rho}_* \in \overline{\mathscr{P}}$ should satisfy

$$\mathrm{supp}(\overline{\rho}_*) \subseteq \arg\min_{r\in[0,+\infty]} \psi_\infty(r; \overline{\rho}_*)\,,$$

with ψ_∞ given in Eq. (4.12).

As proved in the last step, as $\Delta < \Delta_\infty$, we cannot have both $\lambda_+(\overline{\rho}_*) = 0$ and $\lambda_-(\overline{\rho}_*) = 0$. The argument given above also implies that $\psi_\infty(r; \overline{\rho}_*)$ is minimized at a unique point, and hence the support of $\overline{\rho}_*$ should be a single point. This proves the first part of the theorem.

Part (2): $\Delta \ge \Delta_\infty$.

For $\Delta \ge \Delta_\infty$, there exists $r > 0$, such that $q(\tau_+ r) \ge 1$, and $q(\tau_- r) \le -1$. Therefore, there exists $r_* > 0$ such that $q(\tau_+ r_*) - 1 = -1 - q(\tau_- r_*) = \varepsilon_* \ge 0$. Consider the following probability measure on $[0, +\infty]$,

$$\overline{\rho}_* = \frac{1}{1+\varepsilon_*}\delta_{r_*} + \frac{\varepsilon_*}{(1+\varepsilon_*)(q(+\infty) - q(0))}[q(+\infty)\delta_0 - q(0)\delta_{+\infty}].$$

It can be checked that $\overline{R}_\infty(\overline{\rho}_*) = 0$.

We would like to show further that there exists a global minimizer that is compactly supported. We construct this global minimizer as following. First, define

$$r_0 = \inf\{r : q_-(r) \geq -1\}.$$

Then we know that $q_-(r_0) = -1$ and $q_+(r_0) \geq 1$. Now for any $0 \leq r \leq r_0$, define $u(r) = q_-^{-1}(-2 - q_-(r))$. According to condition S3, we have $-1 < [q(0) + q(+\infty)]/2 < 1$, then $u(r)$ is well defined on $[0, r_0]$. It is easy to see that $u(r_0) = r_0$, and $[q_-(r) + q_-(u(r))]/2 = -1$ for any $0 \leq r \leq r_0$. Now we consider the function $z(r) = [q_+(r) + q_+(u(r))]/2 - 1$. Note that $z(r_0) > 0$, and $z(0) \leq [q(0) + q(\infty)]/2 - 1 < 0$. Therefore, there exists r_* satisfying $0 < r_* \leq r_0$ such that $z(r_*) = 0$. Consider the following probability measure on $(0, +\infty)$,

$$\overline{\rho}_* = \frac{1}{2}[\delta_{r_*} + \delta_{u(r_*)}].$$

It is easy to see that $\overline{R}_\infty(\overline{\rho}_*) = 0$. □

4.2 Dynamics: Fixed points

We specialize the general evolution (3.1) to the present case. Assuming ρ_0 to be spherically symmetric, then ρ_t is spherically symmetric for any $t \geq 0$. We let $\overline{\rho}_t$ denote the distribution of $\|\boldsymbol{w}\|_2$ when $\boldsymbol{w} \sim \rho_t$. This satisfies the following PDE:

$$\partial_t \overline{\rho}_t(r) = 2\xi(t)\partial_r\left[\overline{\rho}_t(r)\partial_r\psi_d(r; \overline{\rho}_t)\right]. \tag{4.22}$$

We will view this as an evolution in the space of probability distribution on the completed half-line $\mathscr{P}([0, \infty])$.

In analogy with Proposition 2, we can prove the following characterization of fixed points.

Proposition 4.3. *A distribution $\overline{\rho} \in \mathscr{P}([0, \infty])$ is a fixed point of the PDE (4.22) if and only if*

$$\mathrm{supp}(\overline{\rho}) \subseteq \{r \in [0, \infty] : \ \partial_r\psi_d(r; \overline{\rho}) = 0\}. \tag{4.23}$$

Notice, in particular, global minimizers of $\overline{R}_d(\overline{\rho})$ are fixed points of this evolution, but not vice-versa. The next result classifies fixed points.

Theorem 4.4. *Consider $d = \infty$ and recall the definition of $\lambda_+(\overline{\rho})$ and $\lambda_-(\overline{\rho})$ given by Eqs. (4.13) and (4.14). Then the fixed points of the PDE (4.22) (i.e. the probability measures $\overline{\rho} \in \mathscr{P}([0, \infty])$ satisfying (4.23)) are of one of the following types*

(a) *A fixed point with zero risk.*

(b) *A point mass $\overline{\rho}_{r_*} = \delta_{r_*}$ at some location $r_* \notin \{0, +\infty\}$, but not of type (a).*

(c) *A mixture of the type $\overline{\rho} = a_0\delta_0 + a_\infty\delta_{+\infty} + a\delta_{r_*}$, but not of type (a) or (b).*

For $\Delta < \Delta_\infty$, the PDE has a unique fixed point of type (b), with $\lambda_+(\overline{\rho}_*) < 0$ and $\lambda_-(\overline{\rho}_*) > 0$; it has no type-(a) fixed points; it has possibly fixed points of type (c).

For $\Delta > \Delta_\infty$, the PDE has some fixed points of type (b), with $\lambda_+(\overline{\rho}_*) > 0$ and $\lambda_-(\overline{\rho}_*) < 0$; it also has some type-(a) fixed points; it has possibly fixed points of type (c).

For $\Delta = \Delta_\infty$, the PDE has a unique fixed point of type (a) which is also a delta function at some location r_*, and no type (b) fixed points; it has possibly fixed points of type (c).

Proof. We use the characterization of fixed points in Proposition 4.3. Recall that $\psi_\infty(r;\overline{\rho}_*)$ is defined as in Equation (4.12). The derivative $\partial_r\psi_\infty(r;\overline{\rho})$ gives

$$\partial_r\psi_\infty(r;\overline{\rho}) = \lambda_+(\overline{\rho})q'_+(r) + \lambda_-(\overline{\rho})q'_-(r). \tag{4.24}$$

If a fixed point has $\lambda_+(\overline{\rho}_*) = \lambda_-(\overline{\rho}_*) = 0$, then $\overline{R}_\infty(\overline{\rho}_*) = 0$. This is type-$(a)$ fixed point. Consider then the case $(\lambda_+(\overline{\rho}_*), \lambda_-(\overline{\rho}_*)) \neq (0,0)$. For the same reason as in the proof of Theorem 4.2, we conclude that $\partial_r\psi_\infty(r;\overline{\rho}_*)$ has at most three zeros, two of which are located at 0 and $+\infty$. This proves that all fixed points are of type (a), (b) or (c).

We already proved in Theorem 4.2 that, for $\Delta < \Delta_\infty$, $\inf_{\overline{\rho}} \overline{R}_\infty(\overline{\rho}) > 0$. Therefore, for $\Delta < \Delta_\infty$, there is no type (a) fixed points.

We next prove that, as $\Delta < \Delta_\infty$, fixed point of type (b) is always unique. The location of the delta fixed point should satisfy

$$\partial_r\psi_\infty(r_*;\delta_{r_*}) = [q'_+(r_*)(q_+(r_*)-1) + q'_-(r_*)(q_-(r_*)+1)]/2 = 0. \tag{4.25}$$

Note that $\partial_r\psi_\infty(r_*;\delta_{r_*}) < 0$ for $r > 0$ small enough, and $\partial_r\psi_\infty(r_*;\delta_{r_*}) > 0$ for r large enough, whence this equation has at least one solution $r_* \in (0,\infty)$. In order to prove that it has a unique solution in $(0,+\infty)$, define $r_+ \equiv \inf\{r : q_+(r) \geq 1\}$ and $r_- \equiv \inf\{r : q_-(r) \geq -1\}$. Note that $q'_+(r_*), q'_-(r_*) > 0$ and that, in order to satisfy Eq. (4.25), the terms $\lambda_+(\delta_{r_*}) = 1/2 \cdot (q_+(r_*)-1)$ and $\lambda_-(\delta_{r_*}) = 1/2 \cdot (q_-(r_*)+1)$ must have opposite signs. For $\Delta < \Delta_\infty$, we must have $\lambda_+(\delta_{r_*}) < 0$ and $\lambda_-(\delta_{r_*}) > 0$, and all stationary points should be within $[r_-, r_+]$. Note that $q'_-(r)/q'_+(r)$ is strictly increasing, and $[1-q_+(r)]/[1+q_-(r)]$ is decreasing on $[r_-, r_+]$. Therefore, the fixed point of type δ_{r_*} with $r_* \in (0,\infty)$ is unique.

For $\Delta > \Delta_\infty$, we must have $\lambda_+(\overline{\rho}_*) > 0$ and $\lambda_-(\overline{\rho}_*) < 0$, and all solutions should be within $[r_+, r_-]$. There could possibly be multiple fixed points of type δ_{r_*} with $r_* \in [r_+, r_-]$.

If $\Delta = \Delta_\infty$, it is easy to see that, $\overline{\rho}_* = \delta_{r_*}$ at some $r_* \in (0,\infty)$ is the unique fixed point with zero risk, and the unique fixed point as a point mass. □

4.3 Dynamics: Convergence to global minimum for $d = \infty$

In this section, denote $\mathscr{P}_{\text{good}}$ to be

$$\mathscr{P}_{\text{good}} = \{\overline{\rho}_0 \in \mathscr{P}((0,\infty)) : \overline{R}_\infty(\overline{\rho}_0) < 1, \overline{\rho}_0 \text{ has bounded density on } (0,\infty)\}. \tag{4.26}$$

We then prove that the $d = \infty$ dynamics converges to a global minimizer from any initialization in $\mathscr{P}_{\text{good}}$.

Theorem 4.5. *Consider the PDE (4.22) for $d = \infty$, with initialization $\overline{\rho}_0 \in \mathscr{P}_{\text{good}}$. It has a unique solution $(\overline{\rho}_t)_{t\geq 0}$, such that*

$$\lim_{t\to+\infty} \overline{R}_\infty(\overline{\rho}_t) = \inf_{\overline{\rho}\in\overline{\mathscr{P}}} \overline{R}_\infty(\overline{\rho}).$$

Proof. Without loss of generality, we assume $\xi(t) = 1/2$. First we show the existence and uniqueness of solution of the PDE.

Step 1. Existence and uniqueness of solution. Mass $\overline{\rho}_t((0,\infty)) = 1$ for all t.

According to conditions S1 - S3, $q(r)$, $q'(r)$, and $q''(r)$ are uniformly bounded on $[0,\infty]$. Recall that

$$\begin{aligned} v(r) =& 1/2 \cdot [q_-(r) - q_+(r)], \\ u_\infty(r_1,r_2) =& 1/2 \cdot [q_+(r_1)q_+(r_2) + q_-(r_1)q_-(r_2)]. \end{aligned}$$

Hence $v'(r), \partial_1 u_\infty(r_1, r_2), v''(r), \partial_{11}^2 u_\infty(r_1, r_2), \partial_{12}^2 u_\infty(r_1, r_2)$ are uniformly bounded. Recall we further assumed $\xi(t) \equiv 1/2$. Therefore, conditions A1 and A3 are satisfied with $D = 1$, $V = v$, and $U = u$. By Remark 3.1, there is the existence and uniqueness of solution of PDE (4.22) for $d = \infty$. Denote this solution to be $(\overline{\rho}_t)_{t\geq 0}$.

Recall the formula of $\partial_r \psi_\infty(r; \overline{\rho})$ given in Equation (4.24), it is easy to see that the assumption of Lemma 3.9 is satisfied with $d = 1$ and $\Psi = \psi_\infty$. Hence, we have $\overline{\rho}_t((0, \infty)) = 1$ for any $t < \infty$.

Step 2. Classify the limiting set $\mathcal{S}_*$.

Recall the definition of $(\mathscr{P}([0, +\infty]), \bar{d}_{\mathscr{P}})$ at the beginning of Section 4. Since $(\mathscr{P}([0, +\infty]), \bar{d}_{\mathscr{P}})$ is a compact metric space, and $(\overline{\rho}_t)_{t\geq 0}$ is a continuous curve in this space, then there exists a subsequence $(t_k)_{k\geq 1}$ of times, such that $(\overline{\rho}_{t_k})_{k\geq 1}$ converges in metric $\bar{d}_{\mathscr{P}}$ to a probability distribution $\overline{\rho}_* \in \mathscr{P}([0, +\infty])$.

Analogously to Proposition 2 (using Eq. (4.22)), we have

$$\partial_t \overline{R}_\infty(\overline{\rho}_t) = -\int [\partial_r \psi_\infty(r; \overline{\rho}_t)]^2 \, \overline{\rho}_t(\mathrm{d}r) \, .$$

Since $\overline{R}_\infty(\overline{\rho}_t) \geq 0$, we have

$$\lim_{t\to+\infty} \int [\partial_r \psi_\infty(r; \overline{\rho}_t)]^2 \, \overline{\rho}_t(\mathrm{d}r) = 0.$$

Recall the definition of $\lambda_+(\overline{\rho})$ and $\lambda_-(\overline{\rho})$ given by Eq. (4.13) and (4.14). Since $q \in C_b([0, \infty])$, we have

$$\lim_{k\to\infty} \lambda_+(\overline{\rho}_{t_k}) = \lambda_+(\overline{\rho}_*), \qquad \lim_{k\to\infty} \lambda_-(\overline{\rho}_{t_k}) = \lambda_-(\overline{\rho}_*). \tag{4.27}$$

Note $\partial_r \psi_\infty(r; \overline{\rho})$ is given by Eq. (4.24), and $q' \in C_b([0, +\infty])$, hence

$$\lim_{k\to+\infty} \langle [\partial_r \psi_\infty(\,\cdot\,; \overline{\rho}_{t_k})]^2, \overline{\rho}_{t_k} \rangle = \langle [\partial_r \psi_\infty(\,\cdot\,; \overline{\rho}_*)]^2, \overline{\rho}_* \rangle,$$

which implies

$$\langle [\partial_r \psi_\infty(\,\cdot\,; \overline{\rho}_*)]^2, \overline{\rho}_* \rangle = 0.$$

In other words, any limiting point $\overline{\rho}_*$ of the PDE is a fixed point of the PDE (4.22).

Note $\overline{R}_\infty(\overline{\rho}) = 1/2 \cdot [\lambda_+(\overline{\rho})^2 + \lambda_-(\overline{\rho})^2]$, we have

$$\lim_{k\to+\infty} \overline{R}_\infty(\overline{\rho}_{t_k}) = \overline{R}_\infty(\overline{\rho}_*).$$

Note $\overline{R}_\infty(\overline{\rho}_t)$ is decreasing with t, hence

$$\lim_{t\to+\infty} \overline{R}_\infty(\overline{\rho}_t) = \overline{R}_\infty(\overline{\rho}_*).$$

Let $\mathcal{S}_* = \mathcal{S}_*(\overline{\rho}_0)$ be the set of all limiting points of the $(\overline{\rho}_t)_{t\geq 0}$,

$$\mathcal{S}_* = \{\overline{\rho}_* \in \mathscr{P}([0, \infty]) : \exists (t_k)_{k\geq 1}, \lim_{k\to\infty} t_k = +\infty, s.t., \lim_{k\to\infty} \bar{d}_{\mathscr{P}}(\overline{\rho}_*, \overline{\rho}_{t_k}) = 0\}.$$

Due to Lemma 3.10, $\mathcal{S}_*$ is a connected compact set. Since $\overline{R}_\infty(\overline{\rho}_t)$ is decreasing as t increases, we have $\overline{R}_\infty(\overline{\rho}_*) \equiv \overline{R}_*$ is a constant for all $\overline{\rho}_* \in \mathcal{S}_*$. Since we assumed $\overline{R}_\infty(\overline{\rho}_0) < 1$, and $\overline{R}_\infty(\overline{\rho}_t)$ is decreasing in t, we have $\overline{R}_* < 1$.

Let $\overline{\rho}_*$ be a fixed point of PDE such that $\lambda_+(\overline{\rho}_*) \geq 0, \lambda_-(\overline{\rho}_*) \geq 0$ or $\lambda_+(\overline{\rho}_*) \leq 0, \lambda_-(\overline{\rho}_*) \leq 0$ but not both $\lambda_+(\overline{\rho}_*)$ and $\lambda_-(\overline{\rho}_*)$ equal 0. In this case, according to Eq. (4.24), $\partial_r \psi_\infty(r; \overline{\rho}_*)$ must be strictly increasing or strictly decreasing in r. Since $\mathrm{supp}(\overline{\rho}_*) \subseteq \{r \in [0, \infty] : \partial_r \psi_\infty(r; \overline{\rho}_*) = 0\}$, $\overline{\rho}_*$ must be a combination of two delta functions located at 0 and $+\infty$, i.e., $\overline{\rho}_* = a_0\delta_0 + (1 - a_0)\delta_\infty$. But for a fixed point of this type, it is easy to see that $\overline{R}_\infty(\overline{\rho}_*) \geq 1$. Such fixed points $\overline{\rho}_*$ cannot be one of the limiting points of the PDE since $\overline{R}_\infty(\overline{\rho}_0) < 1$.

Let L be a mapping $L : \mathscr{P}([0, +\infty]) \to \mathbb{R}^2$, $\overline{\rho} \mapsto (\lambda_+(\overline{\rho}), \lambda_-(\overline{\rho}))$. The above argument implies that for any $\overline{\rho}_0 \in \mathscr{P}_{\mathrm{good}}$, we have

$$L(\mathcal{S}_*(\overline{\rho}_0)) \cap (\{(\lambda_+, \lambda_-) : \lambda_+ \geq 0, \lambda_- \geq 0, \text{ or } \lambda_+ \leq 0, \lambda_- \leq 0\} \setminus \{(0, 0)\}) = \emptyset.$$

Since $\mathcal{S}_*$ is a connected set, $L(\mathcal{S}_*)$ should also be a connected set. Further notice that $\overline{R}_\infty(\overline{\rho}_*) = 1/2 \cdot [\lambda_+(\overline{\rho}_*)^2 + \lambda_-(\overline{\rho}_*)^2]$, and $\overline{R}_\infty(\overline{\rho}_1) = \overline{R}_\infty(\overline{\rho}_2)$ for any $\overline{\rho}_1, \overline{\rho}_2 \in \mathcal{S}_*$. Therefore, we can only have $L(\mathcal{S}_*) \subseteq \mathcal{P}_2 \equiv \{(\lambda_+, \lambda_-) : \lambda_+ > 0, \lambda_- < 0\}$, or $L(\mathcal{S}_*) \subseteq \mathcal{P}_1 \equiv \{(\lambda_+, \lambda_-) : \lambda_+ < 0, \lambda_- > 0\}$, or $L(\mathcal{S}_*) = \{(0, 0)\}$.

Step 3. Finish the proof using two claims.

We make the following two claims.

Claim (1). If $L(\mathcal{S}_*) \subseteq \mathcal{P}_1$, then for any $\overline{\rho}_* \in \mathcal{S}_*$, we have $\overline{\rho}_*((0, \infty)) = 1$.

Claim (2). We cannot have $L(\mathcal{S}_*) \subseteq \mathcal{P}_2$.

Here we assume these two claims hold, and use them to prove our results. For $\Delta < \Delta_\infty$, we proved in Theorem 4.4 that, there is not a fixed point such that $L(\overline{\rho}_*) = (0, 0)$. Therefore, we cannot have $L(\mathcal{S}_*) = \{(0, 0)\}$. Due to Claim (2), we cannot have $L(\mathcal{S}_*) \subseteq \mathcal{P}_2$. Hence, we must have $L(\mathcal{S}_*) \subseteq \mathcal{P}_1$. According to Theorem 4.4, for $\Delta < \Delta_\infty$, the only fixed point of PDE with $\overline{\rho}_*((0, \infty)) = 1$ is a point mass at some location r_*. Furthermore, this delta function fixed point is unique and is also the global minimizer of the risk. Therefore, we conclude that, as $\Delta < \Delta_\infty$, the PDE will converge to this global minimizer.

For $\Delta \geq \Delta_\infty$, according to Claim (1), if $\overline{\rho}_*$ is a limiting point such that $L(\overline{\rho}_*) \in \mathcal{P}_1$, then $\overline{\rho}_*((0, \infty)) = 1$. According to Theorem 4.4, a fixed point $\overline{\rho}_*$ with $\overline{\rho}_*((0, \infty)) = 1$ and $L(\overline{\rho}_*) \neq (0, 0)$ must be a point mass at some location r_*, with $L(\overline{\rho}_*) \in \mathcal{P}_2$. Therefore, we cannot have $L(\mathcal{S}_*) \subseteq \mathcal{P}_1$. Claim (2) also tells us that we cannot have $L(\mathcal{S}_*) \subseteq \mathcal{P}_2$. Hence, we must have $L(\mathcal{S}_*) = \{(0, 0)\}$. In this case, all the points in the set $\mathcal{S}_*$ have risk 0. Therefore, we conclude that, as $\Delta \geq \Delta_\infty$, the PDE will converge to some limiting set with risk 0.

Step 4. Proof of the two claims.

We are left with the task of proving the two claims above. Before that, we introduce some useful notations. Recall $Z(r) = q'_-(r)/q'_+(r)$ for $r \in (0, +\infty)$. According to condition S4, $Z'(r) > 0$ for $r \in (0, +\infty)$. This implies that $Z(0+) \equiv Z_0 \geq 0$ and $Z(+\infty) \equiv Z_\infty \leq \infty$ exist. We rewrite $\partial_r \psi_\infty(r; \overline{\rho})$ as

$$\partial_r \psi_\infty(r; \overline{\rho}) = \lambda_+(\overline{\rho}) q'_+(r) + \lambda_-(\overline{\rho}) q'_-(r) = \lambda_-(\overline{\rho}) q'_+(r)[\lambda_+(\overline{\rho})/\lambda_-(\overline{\rho}) + Z(r)]. \tag{4.28}$$

Proof of Claim (1). If $L(\mathcal{S}_*) \subseteq \mathcal{P}_1$, then for any $\overline{\rho}_* \in \mathcal{S}_*$, we have $\overline{\rho}_*(\{0, \infty\}) = 0$.

Assume $L(\mathcal{S}_*) \subseteq \mathcal{P}_1$. Then, we must have $L(\mathcal{S}_*) \subseteq \mathcal{P}_1 \cap \{(\lambda_+, \lambda_-) : Z_0 < -\lambda_+/\lambda_- < Z_\infty\}$. Otherwise suppose there exists $\overline{\rho}_* \in \mathcal{S}_*$, such that $-\lambda_+(\overline{\rho}_*)/\lambda_-(\overline{\rho}_*) \geq Z_\infty$ or $-\lambda_+(\overline{\rho}_*)/\lambda_-(\overline{\rho}_*) \leq Z_0$, according to Eq. (4.28), $\psi_\infty(r; \overline{\rho}_*)$ must be strictly increasing or strictly decreasing in r. Since

$\mathrm{supp}(\overline{\rho}_*) \subseteq \{r \in [0,\infty] : \partial_r \psi_\infty(r;\overline{\rho}_*) = 0\}$, then $\overline{\rho}_*$ must be a combination of two delta functions located at 0 and $+\infty$. But such $\overline{\rho}_*$ must have $\overline{R}_\infty(\overline{\rho}_*) \geq 1$, and thus $\overline{\rho}_*$ cannot be a limiting point of the PDE. Hence the claim that $L(\mathcal{S}_*) \subseteq \mathcal{P}_1 \cap \{(\lambda_+,\lambda_-) : Z_0 < -\lambda_+/\lambda_- < Z_\infty\}$ holds.

Since $\mathcal{S}_*$ is a compact set, and L is a continuous map, then $L(\mathcal{S}_*)$ is a compact set. Therefore, there must exist $\varepsilon_0 > 0$, so that for any $\overline{\rho}_* \in \mathcal{S}_*$, we have $Z_0 + 3\varepsilon_0 < -\lambda_+(\overline{\rho}_*)/\lambda_-(\overline{\rho}_*) < Z_\infty - 3\varepsilon_0$. For this $\varepsilon_0 > 0$, since $\mathcal{S}_*$ contains all the limiting points of PDE starting from $\overline{\rho}_0$, there exists t_0 large enough, so that as $t \geq t_0$, we have $Z_0 + 2\varepsilon_0 < -\lambda_+(\overline{\rho}_t)/\lambda_-(\overline{\rho}_t) < Z_\infty - 2\varepsilon_0$, and $\lambda_+(\overline{\rho}_t) < 0$, $\lambda_-(\overline{\rho}_t) > 0$. For the same ε_0, since $Z(r)$ is continuous at 0 and $+\infty$, there exists $0 < r_0 < r_\infty < \infty$, so that $Z(r) < Z_0 + \varepsilon_0$ for $r \in (0, r_0)$, and $Z(r) > Z_\infty - \varepsilon_0$ for $r \in (r_\infty, \infty)$. Therefore, for any $t \geq t_0$, $\partial_r \psi_\infty(r;\overline{\rho}_t) < 0$ for any $r \in (0, r_0)$, and $\partial_r \psi_\infty(r;\overline{\rho}_t) > 0$ for any $r \in (r_\infty, +\infty)$.

As a result, according to the equation (4.28), we must have $\partial_r \psi_\infty(r;\overline{\rho}_t) < 0$ for any $r \in (0, r_0)$ and $t \geq t_0$, and $\partial_r \psi_\infty(r;\overline{\rho}_t) > 0$ for any $r \in (r_\infty, \infty)$ and $t \geq t_0$.

Due to Lemma 3.9, $\overline{\rho}_{t_0}((0,\infty)) = 1$. Denoting $\Omega_k = [1/k, k]$, then $\lim_{k\to\infty} \overline{\rho}_{t_0}(\Omega_k) = 1$. With this choice of Ω_k, for any $k \geq \{r_\infty, 1/r_0\}$, and for any $t \geq t_0$, we have $\langle \partial_r \psi_\infty(r;\overline{\rho}_t), \boldsymbol{n}(r)\rangle > 0$ for $r \in \partial\Omega_k$ where $\boldsymbol{n}(r)$ is the normal vector point outside Ω_k. Therefore, if we consider the ODE

$$\dot{r}(t) = -\partial \psi_\infty(r(t);\overline{\rho}_t). \tag{4.29}$$

starting with $r(t_0) \in \Omega_k$, $r(t)$ cannot leak outside Ω_k from either boundaries of Ω_k, and we must have $r(t) \in \Omega_k$ for any $t \geq t_0$. Due to Lemma 3.8, $\overline{\rho}_t(\Omega_k) \geq \overline{\rho}_{t_0}(\Omega_k)$ for any $t \geq t_0$. As a result, we conclude that for any $\overline{\rho}_* \in \mathcal{S}_*$,

$$\overline{\rho}_*(\cup_k \Omega_k) \geq \lim_{k\to\infty} \overline{\rho}_*(\Omega_k) \geq \lim_{k\to\infty} \overline{\rho}_{t_0}(\Omega_k) = 1. \tag{4.30}$$

Note $\cup_k \Omega_k = (0,\infty)$. This gives $\overline{\rho}_*(\{0,\infty\}) = 0$, which proves Claim (1).

Proof of Claim (2), step (1). If $L(\mathcal{S}_*) \subseteq \mathcal{P}_2$, then $\mathcal{S}_*$ must be a singleton.

In the case $L(\mathcal{S}_*) \subseteq \mathcal{P}_2$, the argument is similar to the proof of Claim (1), and hence will be presented in a synthetic form. First, we must have $L(\mathcal{S}_*) \subseteq \mathcal{P}_2 \cap \{(\lambda_+,\lambda_-) : Z_0 < -\lambda_+/\lambda_- < Z_\infty\}$. Therefore, there must exist $\varepsilon_0 > 0$, so that for any $\overline{\rho}_* \in \mathcal{S}_*$, we have $Z_0 + 3\varepsilon_0 < -\lambda_+(\overline{\rho}_*)/\lambda_-(\overline{\rho}_*) < Z_\infty - 3\varepsilon_0$. For this $\varepsilon_0 > 0$, there exists t_0 large enough, so that as $t \geq t_0$, we have $Z_0 + 2\varepsilon_0 < -\lambda_+(\overline{\rho}_t)/\lambda_-(\overline{\rho}_t) < Z_\infty - 2\varepsilon_0$, and $\lambda_+(\overline{\rho}_t) > 0$, $\lambda_-(\overline{\rho}_t) < 0$. Further, there exists $0 < r_0 < r_\infty < \infty$, so that $\partial_r \psi_\infty(r;\overline{\rho}_t) > 0$ for any $r \in (0, r_0)$ and $t \geq t_0$, and $\partial_r \psi_\infty(r;\overline{\rho}_t) < 0$ for any $r \in (r_\infty, \infty)$ and $t \geq t_0$.

Therefore, if we consider the ODE (4.29) starting with $r(t_0) \in [0, r_0)$, we must have $r(t) \in [0, r_0)$ for any $t \geq t_0$; if we start with $r(t_0) \in (r_\infty, \infty]$, we must have $r(t) \in (r_\infty, \infty]$ for any $t \geq t_0$. Due to Lemma 3.8, $\{\overline{\rho}_t([0,r))\}_{t\geq t_0}$ for $0 < r \leq r_0$ and $\{\overline{\rho}_t((r,+\infty])\}_{t\geq t_0}$ for $r \geq r_\infty$ must be non-decreasing in t. According to Theorem 4.4, we can express $\overline{\rho}_* \in \mathcal{S}_*$ in the form $\overline{\rho}_* = a_0(\overline{\rho}_*)\delta_0 + a_\infty(\overline{\rho}_*)\delta_\infty + a(\overline{\rho}_*)\delta_{r_*}$. By the stated monotonicity property, for any $\overline{\rho}_1, \overline{\rho}_2 \in \mathcal{S}_*$, it holds that $a_0(\overline{\rho}_1) = a_0(\overline{\rho}_2)$, $a_\infty(\overline{\rho}_1) = a_\infty(\overline{\rho}_2)$, and hence $a(\overline{\rho}_1) = a(\overline{\rho}_2)$. We denote them in short as a_0, a_∞, and a.

For any such fixed point $\overline{\rho}_* \in \mathcal{S}_*$, since we must have $\mathrm{supp}(\overline{\rho}_*) \subseteq \{r : \partial_r \psi_\infty(r;\overline{\rho}_*) = 0\}$, $r_* \in (0, +\infty)$ should be a solution of $\phi(r) = 0$ where

$$\phi(r) = (a_0 q(0) + a_\infty q_\infty + a q_+(r) - 1) q'_+(r) + (a_0 q(0) + a_\infty q_\infty + a q_-(r) + 1) q'_-(r).$$

By condition S1, the function $\phi(r)$ is analytic, and it is not constant. Therefore, the set of all its zeros $\{r_*^i\}_{i\in\mathbb{N}} \subseteq (0, +\infty)$ is a countable set, and it does not have accumulation points in $(0, +\infty)$.

Furthermore, according to Lemma 3.10, the limiting set $\mathcal{S}_*$ should be a connected compact set with respect to the metric $\bar{d}_{\mathscr{P}}$. Therefore, the limiting set could only be a singleton. That is, $\mathcal{S}_* = \{a_0\delta_0 + a_\infty\delta_\infty + a\delta_{r_*}\}$ for some r_*.

Proof of Claim (2)**, step** (2)**. If $\overline{\rho}_*$ is a fixed point with $L(\overline{\rho}_*) \in \mathcal{P}_2$, then $\overline{\rho}_*$ is unstable.**

We apply Theorem 7 to $\overline{\rho}_* = a_0\delta_0 + a_\infty\delta_\infty + a\delta_{r_*}$. We will check the conditions of Theorem 7 to show that this type of fixed point is unstable.

First we check condition B1. Since $[q'_-(r)/q'_+(r)]' > 0$ and $q'_+(r) > 0$ for $r \in (0, +\infty)$, we have

$$q''_-(r_*)q'_+(r_*) - q''_+(r_*)q'_-(r_*) > 0. \tag{4.31}$$

Note the stationary condition of the PDE implies

$$\partial_r\psi(r_*; \overline{\rho}_*) = \lambda_+(\overline{\rho}_*)q'_+(r_*) + \lambda_-(\overline{\rho}_*)q'_-(r_*) = 0, \tag{4.32}$$

and $\lambda_+(\overline{\rho}_*) > 0$, $\lambda_-(\overline{\rho}_*) < 0$. Combined with the equation above, we have

$$\begin{aligned}\partial_r^2\psi_\infty(r_*; \overline{\rho}_*) =& \lambda_+(\overline{\rho}_*)q''_+(r_*) + \lambda_-(\overline{\rho}_*)q''_-(r_*)\\ =& [q'_+(r_*)q''_-(r_*) - q'_-(r_*)q''_+(r_*)] \cdot \lambda_-(\overline{\rho}_*)/q'_+(r_*) < 0.\end{aligned} \tag{4.33}$$

This verifies condition B1 of Theorem 7.

Second, since $\lambda_+(\overline{\rho}_*) > 0$ and $\lambda_-(\overline{\rho}_*) < 0$, according to Equation (4.28), we must have $\partial_r\psi_\infty(r; \overline{\rho}_*) > 0$ for $r \in (0, r_*)$, and $\partial_r\psi_\infty(r; \overline{\rho}_*) < 0$ for $r \in (r_*, \infty)$. Therefore, we have $\psi_\infty(0; \overline{\rho}_*) < \psi_\infty(r_*; \overline{\rho}_*)$ and $\psi_\infty(+\infty; \overline{\rho}_*) < \psi_\infty(r_*; \overline{\rho}_*)$. Note $\mathcal{L}(\eta) \equiv \{r : \psi_\infty(r; \overline{\rho}_*) \le \psi_\infty(r_*; \overline{\rho}_*) - \eta\}$. For any $\eta > 0$ small enough, $\overline{\rho}_*(\mathcal{L}(\eta)) = 1 - a$, which verifies condition B2. It is also easy to see that, for any $\eta > 0$, $\partial\mathcal{L}(\eta)$ is a compact set, hence condition B3 holds. Note that we assumed further that $\overline{\rho}_0$ has a bounded density with respect to Lebesgue measure, all the assumptions of Theorem 7 are satisfied. Theorem 7 implies that the PDE cannot converge to $\overline{\rho}_*$. As a result, we conclude that we cannot have $L(\mathcal{S}_*(\overline{\rho}_0)) \subseteq \mathcal{P}_2$ for $\overline{\rho}_0 \in \mathscr{P}_{\text{good}}$. This proves Claim (2).

□

4.4 Proof of Theorem 1

The key step consists in proving that the dynamics for large but finite d is well approximated by the dynamics at $d = \infty$. The key estimate is provided by the next lemma.

Lemma 4.6. *Assume σ satisfies condition* S0*, recall the definition of u_d and u_∞ given by Equation (4.8) and (4.9). Then we have*

$$\lim_{d\to\infty} \sup_{r_1, r_2 \in [0,\infty)} |u_d(r_1, r_2) - u_\infty(r_1, r_2)| = 0,$$

and

$$\lim_{d\to\infty} \sup_{r_1, r_2 \in [0,\infty)} |\partial_1 u_d(r_1, r_2) - \partial_1 u_\infty(r_1, r_2)| = 0.$$

Proof. Recall that u_d is given by

$$\begin{aligned}u_d(r_1, r_2) =& 1/2 \cdot [u_{d,1}(r_1, r_2) + u_{d,2}(r_1, r_2)],\\ u_{d,1}(r_1, r_2) =& \mathbb{E}[\sigma(r_1(1+\Delta)G_1)\sigma(r_2(1+\Delta)(G_1\cos\Theta + G_2\sin\Theta))],\\ u_{d,2}(r_1, r_2) =& \mathbb{E}[\sigma(r_1(1-\Delta)G_1)\sigma(r_2(1-\Delta)(G_1\cos\Theta + G_2\sin\Theta))],\end{aligned}$$

where $(G_1, G_2) \sim \mathsf{N}(0, \boldsymbol{I}_2)$, and $\Theta \sim (1/Z_d)\sin(\theta)^{d-2} \cdot \mathbf{1}\{\theta \in [0,\pi]\}\mathrm{d}\theta$ are mutually independent.

Define $G_3 = G_1\cos\Theta + G_2\sin\Theta$, then

$$\begin{aligned}
&|u_{d,1}(r_1,r_2) - u_{\infty,1}(r_1,r_2)| \\
=&|\mathbb{E}[\sigma(r_1(1+\Delta)G_1)[\sigma(r_2(1+\Delta)G_3) - \sigma(r_2(1+\Delta)G_2)]]| \\
\le&\|\sigma\|_\infty \mathbb{E}[|\sigma(r_2(1+\Delta)G_3) - \sigma(r_2(1+\Delta)G_2)|],
\end{aligned} \tag{4.34}$$

and

$$\begin{aligned}
&|\partial_1 u_{d,1}(r_1,r_2) - \partial_1 u_{\infty,1}(r_1,r_2)| \\
=&|\mathbb{E}[(1+\Delta)G_1 \cdot \sigma'(r_1(1+\Delta)G_1)[\sigma(r_2(1+\Delta)G_3) - \sigma(r_2(1+\Delta)G_2)]]| \\
\le&(1+\Delta)\|\sigma'\|_\infty \mathbb{E}[G_1^2]^{1/2}\mathbb{E}[[\sigma(r_2(1+\Delta)G_3) - \sigma(r_2(1+\Delta)G_2)]^2]^{1/2} \\
\le&(1+\Delta)\|\sigma'\|_\infty (2\|\sigma\|_\infty^{1/2}) \cdot \mathbb{E}[|\sigma(r_2(1+\Delta)G_3) - \sigma(r_2(1+\Delta)G_2)|]^{1/2}.
\end{aligned} \tag{4.35}$$

According to condition $\mathsf{S0}$, $\|\sigma'\|_\infty$ and $\|\sigma\|_\infty$ are bounded, it is sufficient to bound the following quantity uniformly for $r \in [0,\infty)$

$$T(r) \equiv 1/2 \cdot \mathbb{E}\{|\sigma(rG_2) - \sigma(rG_3)]|\} = \mathbb{E}\{[\sigma(rG_2) - \sigma(rG_3)]\,\mathbf{1}_{G_2 > G_3}\}\,. \tag{4.36}$$

We claim that, for any $a \in \mathbb{R}$,

$$\mathbb{P}(G_3 \le a, G_2 \ge a) \le \mathbb{P}(G_3 \le 0, G_2 \ge 0) = \mathbb{E}[|\pi/2 - \Theta|/(2\pi)]. \tag{4.37}$$

Assuming this claim holds, let us show that it implies the desired bound on $T(r)$. We have

$$\begin{aligned}
T(r) =&\mathbb{E}\left\{\int_{\mathbb{R}} \sigma'(t)\,\mathbf{1}_{rG_2 \ge t \ge rG_3}\,\mathrm{d}t\right\} = \int_{\mathbb{R}} \sigma'(t)\,\mathbb{P}\{G_2 \ge t/r \ge G_3\}\,\mathrm{d}t \\
\le&\sup_{a\in\mathbb{R}} \mathbb{P}(G_3 \le a, G_2 \ge a) \int_{\mathbb{R}} \sigma'(t)\,\mathrm{d}t \le 2\|\sigma\|_\infty \cdot \mathbb{E}[|\pi/2 - \Theta|/(2\pi)]\,.
\end{aligned}$$

Note that $\cos(\Theta) \overset{\mathrm{d}}{=} Z_1/\|\boldsymbol{Z}\|_2$ for $\boldsymbol{Z} \sim \mathsf{N}(0, \boldsymbol{I}_d)$ and hence $\mathbb{E}\{|\Theta - \pi/2|\} \le K/\sqrt{d}$ for a universal constant K. We therefore obtain

$$\sup_r |T(r)| \le (K/\pi)\|\sigma\|_\infty/\sqrt{d}. \tag{4.38}$$

We are left with the task of proving Eq. (4.37).

Denote $X = G_2$ and $Y = G_3$ for simplicity in notations. Note that $(X,Y) \overset{\mathrm{d}}{=} (Y,X) \overset{\mathrm{d}}{=} (-X,-Y)$. It follows that we can assume, without loss of generality, $a > 0$. We have

$$\begin{aligned}
\mathbb{P}(Y \le a, X \ge a) =&\mathbb{P}(Y \le 0, X \ge a) + \mathbb{P}(0 \le Y \le a, X \ge a), \\
\mathbb{P}(Y \le 0, X \ge 0) =&\mathbb{P}(Y \le 0, X \ge a) + \mathbb{P}(Y \le 0, 0 \le X \le a),
\end{aligned}$$

suffice to prove that

$$\mathbb{P}(0 \le Y \le a, X \ge a) \le \mathbb{P}(Y \le 0, 0 \le X \le a).$$

Define $U = (X-Y)/2$, $V = (X+Y)/2$, and $\mathcal{A}_1 = \{0 \le Y \le a, X \ge a\}$, $\mathcal{A}_2 = \{Y \le 0, 0 \le X \le a\}$. It is easy to see that $[U|\Theta = \theta]$ and $[V|\Theta = \theta]$ are independent normal random variables. Therefore, it is sufficient to show $\mathbb{P}(\mathcal{A}_1|U = u, \Theta = \theta) \le \mathbb{P}(\mathcal{A}_2|U = u, \Theta = \theta)$ for $u \ge 0$ and $\theta \in [0,\pi]$ (as $u < 0$, both conditional probability equal 0).

Fix an $u \geq 0$ and $\theta \in [0, \pi]$. Consider the closed interval $\mathcal{I}_i = \mathcal{I}_i(u) \subseteq \mathbb{R}$ for $i = 1, 2$, with definition $\mathcal{I}_i(u) \equiv \{v : \{U = u, V = v, \Theta = \theta\} \subseteq \mathcal{A}_i\}$. Then $\mathbb{P}(\mathcal{A}_i | U = u, \Theta = \theta) = \int_{\mathcal{I}_i(u)} p_{V|\Theta}(v|\theta) \mathrm{d}v$, where $p_{V|\Theta}(v|\theta)$ is the density of $[V|\Theta = \theta]$ at v. It is not hard to see that every element in $\mathcal{I}_1$ is greater or equal to $a/2$, and every element in $\mathcal{I}_2$ is less or equal to $a/2$; in the meanwhile, $\mathcal{I}_1$ and $\mathcal{I}_2$ are symmetric with respect to $a/2$. Note that $[V|\Theta = \theta]$ is a Gaussian random variable with zero mean, therefore $p_{V|\Theta}(a/2 + s|\theta) \leq p_{V|\Theta}(a/2 - s|\theta)$ for any $s \geq 0$ and $\theta \in [0, \pi]$. This implies that $\mathbb{P}(\mathcal{A}_1 | U = u, \Theta = \theta) \leq \mathbb{P}(\mathcal{A}_2 | U = u, \Theta = \theta)$, for any $u \geq 0$ and $\theta \in [0, \pi]$. □

Lemma 4.7. *Let* $y \sim \mathrm{Unif}(\{-1, +1\})$, $[\boldsymbol{x}|y = +1] \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma}_+)$, $[\boldsymbol{x}|y = -1] \sim \mathsf{N}(0, \boldsymbol{\Sigma}_-)$ *with* $\tau_-^2 \boldsymbol{I}_D \preceq \boldsymbol{\Sigma}_+, \boldsymbol{\Sigma}_- \preceq \tau_+^2 \boldsymbol{I}_D$ *for some* $0 < \tau_- < \tau_+ < \infty$. *Assume that the activation function* σ *satisfies condition* $\mathsf{S0}$. *Define*

$$\begin{aligned} V(\boldsymbol{\theta}) =& - \mathbb{E}[y\, \sigma(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle)], \\ U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) =& \mathbb{E}[\sigma(\langle \boldsymbol{x}, \boldsymbol{\theta}_1 \rangle) \sigma(\langle \boldsymbol{x}, \boldsymbol{\theta}_2 \rangle)]. \end{aligned} \tag{4.39}$$

Then assumptions $\mathsf{A2}$ *and* $\mathsf{A3}$ *are satisfied.*

Proof. Note that $\boldsymbol{x}$ is sub-Gaussian, and by condition $\mathsf{S0}$ we have σ' is bounded, then $\nabla_{\boldsymbol{\theta}} \sigma(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle) = \sigma'(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle) \boldsymbol{x}$ is also sub-Gaussian (with sub-Gaussian parameter independent of D). Condition $\mathsf{S0}$ also gives that σ is bounded, therefore assumption $\mathsf{A2}$ is satisfied.

To verify assumption $\mathsf{A3}$, it is sufficient to check that ∇V, $\nabla_1 U$, $\nabla_{12}^2 U$, $\nabla^2 V$, and $\nabla_{11}^2 U$ are uniformly bounded in ℓ_2 norm (for the gradients) or operator norm (for the Hessians). For any unit vector $\boldsymbol{n}$, we have

$$\langle \nabla V(\boldsymbol{\theta}), \boldsymbol{n} \rangle = - \mathbb{E}[y \sigma'(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle) \langle \boldsymbol{x}, \boldsymbol{n} \rangle], \tag{4.40}$$

$$\langle \nabla_1 U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{n} \rangle = \mathbb{E}[\sigma'(\langle \boldsymbol{x}, \boldsymbol{\theta}_1 \rangle) \langle \boldsymbol{x}, \boldsymbol{n} \rangle \sigma(\langle \boldsymbol{x}, \boldsymbol{\theta}_2 \rangle)], \tag{4.41}$$

$$\langle \nabla_{12}^2 U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{n}^{\otimes 2} \rangle = \mathbb{E}[\sigma'(\langle \boldsymbol{x}, \boldsymbol{\theta}_1 \rangle) \langle \boldsymbol{x}, \boldsymbol{n} \rangle^2 \sigma'(\langle \boldsymbol{x}, \boldsymbol{\theta}_2 \rangle)]. \tag{4.42}$$

Since $\|\sigma\|_\infty, \|\sigma'\|_\infty < \infty$, applying Cauchy-Schwarz inequality, we have $\nabla V, \nabla_1 U, \nabla_{12}^2 U$ are uniformly bounded.

It is difficult to bound $\nabla^2 V$ and $\nabla_1^2 U$ directly because σ' may not be differentiable. We will use a longer argument to bound them.

First, for a bounded-Lipschitz function f, and for $g \in \{1, \sigma\}$, define

$$W_{f,g}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \mathbb{E}_{\boldsymbol{G}}[f(\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle) g(\langle \boldsymbol{\theta}_2, \boldsymbol{G} \rangle)], \tag{4.43}$$

where $\boldsymbol{G} \sim \mathsf{N}(0, \boldsymbol{I}_d)$. Since we have $\tau_-^2 \boldsymbol{I}_D \preceq \boldsymbol{\Sigma}_+, \boldsymbol{\Sigma}_- \preceq \tau_+^2 \boldsymbol{I}_D$ for some $0 < \tau_- < \tau_+ < \infty$, in order to bound $\nabla^2 V$ and $\nabla_1^2 U$, it is sufficient to bound $\nabla_1^2 W_{\sigma,1}$ and $\nabla_1^2 W_{\sigma,\sigma}$.

Since σ' is K_0-Lipschitz on $[-2\delta_0, 2\delta_0]$ for some $\delta_0 > 0$ and $K_0 < \infty$, then, there exists a function $\sigma_0 : \mathbb{R} \to \mathbb{R}$, so that σ_0 is non-decreasing and K-bounded-Lipschitz, σ_0' is K-bounded-Lipschitz, and $\sigma_0(r) = \sigma(r)$ for $r \in [-\delta_0, \delta_0]$. For this σ_0, a second weak derivative exists and $|\sigma_0''| \leq K$. Hence

$$\langle \nabla_1^2 W_{\sigma_0,g}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{n}^{\otimes 2} \rangle = \mathbb{E}[\sigma_0''(\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle) \langle \boldsymbol{G}, \boldsymbol{n} \rangle^2 g(\langle \boldsymbol{\theta}_2, \boldsymbol{G} \rangle)] \tag{4.44}$$

is uniformly bounded for $g = 1$ or $g = \sigma$. Let $h = \sigma - \sigma_0$, then $h = 0$ for $r \in [-\delta_0, \delta_0]$, and h is K-bounded-Lipschitz for some constant K. It is sufficient to bound $\nabla_1^2 W_{h,g}$ for $g \in \{1, \sigma\}$.

Since $\boldsymbol{G}$ is Gaussian, using Stein's formula, for any unit vector $\boldsymbol{n}$, we have

$$\begin{aligned}
&\langle \nabla_1 W_{h,g}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{n} \rangle = \mathbb{E}[h'(\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle) \langle \boldsymbol{n}, \boldsymbol{G} \rangle g(\langle \boldsymbol{\theta}_2, \boldsymbol{G} \rangle)] \\
&= \underbrace{\frac{1}{\|\boldsymbol{\theta}_1\|_2^2} \mathbb{E}[h(\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle) \langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle \langle \boldsymbol{n}, \boldsymbol{G} \rangle g(\langle \boldsymbol{\theta}_2, \boldsymbol{G} \rangle)]}_{E_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{n})} - \underbrace{\frac{1}{\|\boldsymbol{\theta}_1\|_2^2} \mathbb{E}[h(\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle) \langle \boldsymbol{\theta}_1, \boldsymbol{n} \rangle g(\langle \boldsymbol{\theta}_2, \boldsymbol{G} \rangle)]}_{E_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{n})} \\
&\quad - \underbrace{\frac{1}{\|\boldsymbol{\theta}_1\|_2^2} \mathbb{E}[h(\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle) \langle \boldsymbol{n}, \boldsymbol{G} \rangle g'(\langle \boldsymbol{\theta}_2, \boldsymbol{G} \rangle) \langle \boldsymbol{\theta}_2, \boldsymbol{\theta}_1 \rangle]}_{E_3(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{n})}.
\end{aligned} \tag{4.45}$$

Taking directional derivatives of E_1 and E_2, we have

$$\begin{aligned}
\langle \nabla_1 E_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{n}), \boldsymbol{n} \rangle &= \underbrace{\frac{1}{\|\boldsymbol{\theta}_1\|_2^2} \mathbb{E}[h'(\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle) \langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle \langle \boldsymbol{n}, \boldsymbol{G} \rangle^2 g(\langle \boldsymbol{\theta}_2, \boldsymbol{G} \rangle)]}_{E_{11}} \\
&+ \underbrace{\frac{1}{\|\boldsymbol{\theta}_1\|_2^2} \mathbb{E}[h(\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle) \langle \boldsymbol{n}, \boldsymbol{G} \rangle^2 g(\langle \boldsymbol{\theta}_2, \boldsymbol{G} \rangle)]}_{E_{12}} - \underbrace{\frac{2\langle \boldsymbol{\theta}_1, \boldsymbol{n} \rangle}{\|\boldsymbol{\theta}_1\|_2^4} \mathbb{E}[h(\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle) \langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle \langle \boldsymbol{n}, \boldsymbol{G} \rangle g(\langle \boldsymbol{\theta}_2, \boldsymbol{G} \rangle)]}_{E_{13}},
\end{aligned} \tag{4.46}$$

and

$$\begin{aligned}
\langle \nabla_1 E_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{n}), \boldsymbol{n} \rangle &= \underbrace{\frac{1}{\|\boldsymbol{\theta}_1\|_2^2} \mathbb{E}[h'(\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle) \langle \boldsymbol{\theta}_1, \boldsymbol{n} \rangle \langle \boldsymbol{G}, \boldsymbol{n} \rangle g(\langle \boldsymbol{\theta}_2, \boldsymbol{G} \rangle)]}_{E_{21}} \\
&+ \underbrace{\frac{1}{\|\boldsymbol{\theta}_1\|_2^2} \mathbb{E}[h(\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle) g(\langle \boldsymbol{\theta}_2, \boldsymbol{G} \rangle)]}_{E_{22}} - \underbrace{\frac{2\langle \boldsymbol{\theta}_1, \boldsymbol{n} \rangle}{\|\boldsymbol{\theta}_1\|_2^4} \mathbb{E}[h(\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle) \langle \boldsymbol{\theta}_1, \boldsymbol{n} \rangle g(\langle \boldsymbol{\theta}_2, \boldsymbol{G} \rangle)]}_{E_{23}}.
\end{aligned} \tag{4.47}$$

To bound E_{11}, note $h'(r) = 0$ for $r \in (-\delta_0, \delta_0)$, and $|h'(r)| \le K$ for $r \in \mathbb{R}$, we have

$$\begin{aligned}
E_{11} \le& \frac{K}{\|\boldsymbol{\theta}_1\|_2} \mathbb{E}\Big[\mathbf{1}\{|\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle| \ge \delta_0\} \cdot |\langle \boldsymbol{\theta}_1 / \|\boldsymbol{\theta}_1\|_2, \boldsymbol{G} \rangle| \cdot \langle \boldsymbol{n}, \boldsymbol{G} \rangle^2 |g(\langle \boldsymbol{\theta}_2, \boldsymbol{G} \rangle)|\Big] \\
\le& \frac{K}{\|\boldsymbol{\theta}_1\|_2} \cdot \mathbb{P}(|\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle| \ge \delta_0)^{1/2} \cdot \{\mathbb{E}[(\langle \boldsymbol{\theta}_1 / \|\boldsymbol{\theta}_1\|_2, \boldsymbol{G} \rangle^2 \langle \boldsymbol{n}, \boldsymbol{G} \rangle^4 g(\langle \boldsymbol{\theta}_2, \boldsymbol{G} \rangle))^2]\}^{1/2}.
\end{aligned} \tag{4.48}$$

Take $r = \|\boldsymbol{\theta}_1\|_2$, then

$$1/\|\boldsymbol{\theta}_1\|_2 \cdot \mathbb{P}(|\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle| \ge \delta_0)^{1/2} \le 1/r \cdot \exp\{-\delta_0^2/(4r^2)\} \tag{4.49}$$

is uniformly bounded for $r \in [0, \infty]$. Hence E_{11} is uniformly bounded. Using a similar argument, we can show that each terms E_{12}, E_{13}, E_{21}, E_{22}, and E_{23} are uniformly bounded.

Now we look at $\nabla_1 E_3(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{n})$. We have

$$\begin{aligned}
\langle \nabla_1 E_3(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{n}), \boldsymbol{n} \rangle &= \underbrace{\frac{1}{\|\boldsymbol{\theta}_1\|_2^2} \mathbb{E}[h'(\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle) \langle \boldsymbol{n}, \boldsymbol{G} \rangle^2 g'(\langle \boldsymbol{\theta}_2, \boldsymbol{G} \rangle) \langle \boldsymbol{\theta}_2, \boldsymbol{\theta}_1 \rangle]}_{E_{31}} \\
&+ \underbrace{\frac{1}{\|\boldsymbol{\theta}_1\|_2^2} \mathbb{E}[h(\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle) \langle \boldsymbol{n}, \boldsymbol{G} \rangle g'(\langle \boldsymbol{\theta}_2, \boldsymbol{G} \rangle) \langle \boldsymbol{\theta}_2, \boldsymbol{n} \rangle]}_{E_{32}} - \underbrace{\frac{2\langle \boldsymbol{\theta}_1, \boldsymbol{n} \rangle}{\|\boldsymbol{\theta}_1\|_2^4} \mathbb{E}[h(\langle \boldsymbol{\theta}_1, \boldsymbol{G} \rangle) \langle \boldsymbol{n}, \boldsymbol{G} \rangle g'(\langle \boldsymbol{\theta}_2, \boldsymbol{G} \rangle) \langle \boldsymbol{\theta}_2, \boldsymbol{\theta}_1 \rangle]}_{E_{33}}.
\end{aligned} \tag{4.50}$$

In order to bound E_{32}, we apply Stein's formula to get

$$\begin{aligned} E_{32} =& \frac{\langle \boldsymbol{\theta}_2, \boldsymbol{n}\rangle}{\|\boldsymbol{\theta}_1\|_2^2\|\boldsymbol{\theta}_2\|_2^2}\Big\{\mathbb{E}[h(\langle \boldsymbol{\theta}_1, \boldsymbol{G}\rangle)\langle \boldsymbol{n}, \boldsymbol{G}\rangle g(\langle \boldsymbol{\theta}_2, \boldsymbol{G}\rangle)\langle \boldsymbol{\theta}_2, \boldsymbol{G}\rangle] \\ & - \mathbb{E}[h(\langle \boldsymbol{\theta}_1, \boldsymbol{G}\rangle)\langle \boldsymbol{n}, \boldsymbol{\theta}_2\rangle g(\langle \boldsymbol{\theta}_2, \boldsymbol{G}\rangle)] - \mathbb{E}[h'(\langle \boldsymbol{\theta}_1, \boldsymbol{G}\rangle)\langle \boldsymbol{\theta}_1, \boldsymbol{\theta}_2\rangle\langle \boldsymbol{n}, \boldsymbol{G}\rangle g(\langle \boldsymbol{\theta}_2, \boldsymbol{G}\rangle)]\Big\}. \end{aligned} \tag{4.51}$$

For each terms above, we can bound them using the same argument as for bounding E_{11}. Similarly, we can bound E_{33}. We cannot apply directly Stein's formula to E_{31} similar to what we did for E_{32}, because $h' = \sigma' - \sigma_0'$ may not have weak derivative. However, recall that $h'(r) = 0$ for $r \in [-\delta_0, \delta_0]$ and h' is K-bounded. Therefore, we can find a function $h_0 : \mathbb{R} \to \mathbb{R}$, such that $|h'(r)| \le h_0(r)$ for $r \in \mathbb{R}$, $h_0(r) = 0$ for $r \in [-\delta_0/2, \delta_0/2]$, and h_0 is K-bounded-Lipschitz (for some larger constant K). Hence, recalling that $g'(r) \ge 0$, we get

$$E_{31} \le \frac{1}{\|\boldsymbol{\theta}_1\|_2}\mathbb{E}[h_0(\langle \boldsymbol{\theta}_1, \boldsymbol{G}\rangle)\langle \boldsymbol{n}, \boldsymbol{G}\rangle^2 g'(\langle \boldsymbol{\theta}_2, \boldsymbol{G}\rangle)\|\boldsymbol{\theta}_2\|_2]. \tag{4.52}$$

We can apply Stein's formula to the right hand side of the last equation. Using the same argument as above, we obtain that E_{31} is uniformly bounded.

As a result, $\nabla^2 V$ and $\nabla_1^2 U$ are uniformly bounded. Therefore, assumption A3 is satisfied.

□

We are now in position to prove Theorem 1.

Proof of Theorem 1. First we consider PDE (4.22) for $d = \infty$. We fix an initial radial density $\overline{\rho}_0 \in \mathscr{P}_{\text{good}}$. Due to Theorem 4.5, for any $\eta > 0$, there exists $T = T(\eta, \overline{\rho}_0, \Delta) > 0$, so that the solution $(\overline{\rho}_t^\infty)_{t\ge 0}$ of PDE (4.22) for $d = \infty$ with initialization $\overline{\rho}_0$ satisfies

$$\overline{R}_\infty(\overline{\rho}_t^\infty) \le \inf_{\overline{\rho}\in\overline{\mathscr{P}}} \overline{R}_\infty(\overline{\rho}) + \eta/5$$

for any $t \ge T$.

Then we consider the general PDE

$$\partial_t \rho_t(\boldsymbol{\theta}) = 2\xi(t)\nabla\cdot\left[\rho_t(\boldsymbol{\theta})\nabla\Psi(\boldsymbol{\theta};\rho_t)\right], \tag{4.53}$$

with initialization ρ_0 the distribution of $r\boldsymbol{n}$, where $(r, \boldsymbol{n}) \sim \overline{\rho}_0 \times \mathrm{Unif}(\mathbb{S}^{d-1})$. Due to Lemma 4.7, we have the existence and uniqueness of the solution of PDE (4.53), and let $(\rho_t)_{t\ge 0}$ be the solution. Let $\overline{\rho}_t^d$ be the radial marginal distribution of ρ_t. It is easy to see that $(\overline{\rho}_t^d)_{t\ge 0}$ is the unique solution of (4.22) for d finite.

Now, we would like to bound the distance of $\overline{\rho}_t^d$ and $\overline{\rho}_t^\infty$ using Lemma 3.7. We take $D = 1$, $V = v$, $U = u_d$, $\tilde{V} = v$, $\tilde{U} = u_\infty$ in Lemma 3.7. Let $\varepsilon_0(d)$ be defined as in Eq. (3.69). Due to Lemma 4.6, we have $\varepsilon_0(d) \to 0$ as $d \to \infty$. Therefore, according to Lemma 3.7, we have $\lim_{d\to\infty} \sup_{t\le 10T} d_{\mathrm{BL}}(\overline{\rho}_t^d, \overline{\rho}_t^\infty) = 0$. Further note that $\overline{R}_\infty$ is uniformly continuous with respect to $\overline{\rho}$ in bounded-Lipschitz distance. Therefore, there exists $d_0 = d_0(\eta, \overline{\rho}_0, \Delta)$ large enough, so that for $d \ge d_0$ we have

$$|\overline{R}_\infty(\overline{\rho}_t^d) - \overline{R}_\infty(\overline{\rho}_t^\infty)| \le \eta/5.$$

for any $t \le 10T$.

Next we would like to bound the difference of $\overline{R}_\infty(\overline{\rho})$ and $\overline{R}_d(\overline{\rho})$ for any $\overline{\rho}$. Note

$$|\overline{R}_\infty(\overline{\rho}) - \overline{R}_d(\overline{\rho})| \le \int |u_d(r_1, r_2) - u_\infty(r_1, r_2)|\, \overline{\rho}(\mathrm{d}r_1)\overline{\rho}(\mathrm{d}r_2). \tag{4.54}$$

By Lemma 4.6, there exists $d_0 = d_0(\eta, \Delta)$ large enough, so that for $d \ge d_0$, we have

$$\sup_{\overline{\rho}\in\overline{\mathscr{P}}} |\overline{R}_\infty(\overline{\rho}) - \overline{R}_d(\overline{\rho})| \le \eta/5. \tag{4.55}$$

Finally, let $(\boldsymbol{\theta}^k)_{k\ge1}$ be the trajectory of SGD, with step size $s_k = \varepsilon\xi(k\varepsilon)$, and initialization $\boldsymbol{w}_i^0 \sim_{iid} \rho_0$ for $i \le N$. We apply Theorem 3 to bound the difference of the law of trajectory of SGD and the solution of PDE (4.53). The assumptions of Theorem 3 are verified by Lemma 4.7. As a consequence, there exists constant K (which depend uniquely on the constants in assumptions A1 A2 A3), such that for any $t \le 10T$, we have

$$R_N(\boldsymbol{\theta}^{\lfloor t/\varepsilon\rfloor}) - \overline{R}_d(\overline{\rho}_t^d) \le Ke^{10KT} \cdot \mathsf{err}_{N,d}(z).$$

with probability $1 - e^{-z^2}$, where

$$\mathsf{err}_{N,d}(z) = \sqrt{1/N \vee \varepsilon} \cdot \Big[\sqrt{d + \log(N(1/\varepsilon \vee 1))} + z\Big].$$

As a consequence, for any $\delta > 0$, there exists $C_0 = C_0(\delta, \eta, \overline{\rho}_0, \Delta)$, so that as $N, 1/\varepsilon \ge C_0 d$ and $\varepsilon \ge 1/N^{10}$, for any $t \le 10T$, we have

$$R_N(\boldsymbol{\theta}^{\lfloor t/\varepsilon\rfloor}) - \overline{R}_d(\overline{\rho}_t^d) \le \eta/5$$

with probability at least $1 - \delta$.

Therefore, the trajectory $\boldsymbol{\theta}^{\lfloor t/\varepsilon\rfloor}$ of SGD as $t \in [T, 10T]$ satisfies

$$\begin{aligned} R_N(\boldsymbol{\theta}^{\lfloor t/\varepsilon\rfloor}) \le& \overline{R}_d(\overline{\rho}_t^d) + \eta/5 \le \overline{R}_\infty(\overline{\rho}_t^d) + 2\eta/5 \le \overline{R}_\infty(\overline{\rho}_t^\infty) + 3\eta/5 \\ \le& \inf_{\overline{\rho}\in\overline{\mathscr{P}}} \overline{R}_\infty(\overline{\rho}) + 4\eta/5 \le \inf_{\overline{\rho}\in\overline{\mathscr{P}}} \overline{R}_d(\overline{\rho}) + \eta = \inf_{\rho\in\mathscr{P}(\mathbb{R}^d)} R(\rho) + \eta \\ \le& \inf_{\boldsymbol{\theta}\in\mathbb{R}^{d\times N}} R_N(\boldsymbol{\theta}) + \eta \end{aligned}$$

with probability at least $1 - \delta$. This gives the desired result.

□

4.5 Checking conditions S0–S4 for the running example

Lemma 4.8. *Consider the activation function σ with definition in Equation (4.1), with $s_1 < s_2$, $s_1 < -1$, $(s_1 + s_2)/2 > 1$, $(3s_1 + s_2)/4 \in (-1, 1)$, $0 < t_1 < t_2$. For $r \in (0, +\infty)$, define $q(r) = \mathbb{E}_G[\sigma(rG)]$ where $G \sim \mathsf{N}(0, 1)$. Then conditions S0–S4 hold.*

Remark 4.1. The requirements of Lemma 4.8 are not restrictive. An example of parameters that satisfies all conditions gives $s_1 = -2.5$, $s_2 = 7.5$, $t_1 = 0.5$, $t_2 = 1.5$.

Proof. It is straightforward to see that condition S0 holds. To show condition S1, denote by $\sigma'(r)$ the weak derivative of $\sigma(r)$, we calculate the function $q'(r)$ for $r > 0$ explicitly,

$$\begin{aligned} q'(r) =& \mathbb{E}[\sigma'(rG)G] = \frac{s_2 - s_1}{t_2 - t_1} \int_{\mathbb{R}} \mathbf{1}\{rx \in [t_1, t_2]\} \cdot \frac{1}{\sqrt{2\pi}} \exp\Big\{ -\frac{x^2}{2} \Big\} \cdot x \cdot \mathrm{d}x \\ =& \frac{s_2 - s_1}{\sqrt{2\pi}(t_2 - t_1)} \Big\{ \exp\Big[-\frac{t_1^2}{2r^2} \Big] - \exp\Big[-\frac{t_2^2}{2r^2} \Big] \Big\}. \end{aligned} \tag{4.56}$$

Since $s_1 < s_2$ and $0 < t_1 < t_2$, it is easy to see that $q'(r)$ is analytic on $(0, \infty)$, and hence $q(r)$ is analytic on $(0, \infty)$. Differentiating $q'(r)$ in Eq. (4.56), it is easy to see that $\lim_{r\to\infty} q''(r) = 0$, and $q''(0+) = 0$. Hence, we have $\sup_{r\in[0,+\infty]} q''(r) < \infty$. Then condition S1 holds.

Since $s_2 > s_1$, $0 < t_1 < t_2$, we have $q'(r) > 0$ for $r \in (0, +\infty)$, $\lim_{r\to\infty} q'(r) = 0$, and $q'(0+) = 0$. Hence, we have $\sup_{r\in[0,+\infty]} q'(r) < \infty$. Then condition S2 holds. Note that $q(0) = \sigma(0) = s_1 < -1$, and $q(+\infty) = (s_1 + s_2)/2 > 1$. In addition, $[q(0) + q(+\infty)]/2 = (3s_1 + s_2)/4 \in (-1, 1)$. Therefore, condition S3 holds.

Finally, we show that condition S4 holds. Define $p(r) = \exp[-t_1^2/(2r^2)] - \exp[-t_2^2/(2r^2)]$, which is a positively scaled version of $q'(r)$. To show that for $r \in (0, \infty)$,

$$[q'(\tau_- r)/q'(\tau_+ r)]' = [\tau_- \cdot q''(\tau_- r) q'(\tau_+ r) - \tau_+ \cdot q'(\tau_- r) q''(\tau_+ r)]/[q'(\tau_+ r)]^2 > 0,$$

we only need to show that for $r \in (0, \infty)$

$$F_1(r) \equiv \tau_- \cdot p'(\tau_- r) p(\tau_+ r) - \tau_+ \cdot p'(\tau_+ r) p(\tau_- r) > 0.$$

We have

$$\begin{aligned} F_1(r) =& + 1/(\tau_-^2 r^3) \cdot \{t_1^2 \exp[-t_1^2/(2\tau_-^2 r^2)] - t_2^2 \exp[-t_2^2/(2\tau_-^2 r^2)]\} \\ & \times \{\exp[-t_1^2/(2\tau_+^2 r^2)] - \exp[-t_2^2/(2\tau_+^2 r^2)]\} \\ & - 1/(\tau_+^2 r^3) \cdot \{t_1^2 \exp[-t_1^2/(2\tau_+^2 r^2)] - t_2^2 \exp[-t_2^2/(2\tau_+^2 r^2)]\} \\ & \times \{\exp[-t_1^2/(2\tau_-^2 r^2)] - \exp[-t_2^2/(2\tau_-^2 r^2)]\}. \end{aligned}$$

Define $x \equiv t_2^2/(2\tau_+^2 r^2) > 0$, $s \equiv \tau_+^2/\tau_-^2 > 1$, $0 < c \equiv t_1^2/t_2^2 < 1$, we have

$$\begin{aligned} F_1(r) =& + t_2^2/(\tau_+^2 r^3) \cdot \{cs \cdot \exp[-xsc] - s \exp[-xs]\} \cdot \{\exp[-xc] - \exp[-x]\} \\ & - t_2^2/(\tau_+^2 r^3) \cdot \{c \cdot \exp[-xc] - \exp[-x]\} \cdot \{\exp[-xsc] - \exp[-xs]\} \\ =& t_2^2/(\tau_+^2 r^3) \{(cs - c) \exp[-xc - xsc] + (c - s) \exp[-xs - xc] \\ & + (1 - cs) \exp[-x - xsc] + (s - 1) \exp[-x - xs]\} \\ =& t_2^2/(\tau_+^2 r^3) \exp\{-x - xsc\} \{(cs - c) \exp[x - xc] \\ & + (c - s) \exp[x - xs - xc + xsc] + (1 - cs) + (s - 1) \exp[xsc - xs]\}. \end{aligned}$$

Define

$$F_2(x; s, c) = (cs - c) \exp[x - xc] + (c - s) \exp[x - xs - xc + xsc] + (1 - cs) + (s - 1) \exp[xsc - xs].$$

It is sufficient to show that $F_2(x; s, c) > 0$ for $x > 0$, $s > 1$, and $0 < c < 1$. Note that $F_2(0+; s, c) = 0$. Hence it is sufficient to show that $\partial_x F_2(x; s, c) > 0$ for $x > 0$.

We have

$$\begin{aligned}\partial_x F_2(x;s,c) =& c(s-1)(1-c)\exp[x-xc] + (s-c)(s-1)(1-c)\exp[x-xs-xc+xsc] \\ &+ (s-1)s(c-1)\exp[xsc-xs] \\ =& (s-1)(1-c)\exp[xsc-xs]\{c\cdot\exp[x-xc-xsc+xs] + (s-c)\exp[x-xc] - s\}.\end{aligned}$$

Define

$$F_3(x;s,c) = c\cdot\exp[x-xc-xsc+xs] + (s-c)\exp[x-xc] - s.$$

Note that $s>1$ and $0\le c<1$, $F_3(0+;s,c)=0$. It is therefore sufficient to show that $\partial_x F_3(x;s,c) > 0$ for $x>0$.

We have

$$\partial_x F_3(x;s,c) = c(1-c)(1+s)\exp[x-xc-xsc+xs] + (s-c)(1-c)\exp[x-xc].$$

Since $0<c<1$, $s>1$, and $x>0$, we have $\partial_x F_3(x;s,c)>0$, and hence condition S4 holds. □

5 Centered anisotropic Gaussians

In this section we consider the centered anisotropic Gaussian example discussed in the main text. That is, we assume the joint law of $(y,\boldsymbol{x})$ to be as follows:

With probability 1/2: $y=+1$, $\boldsymbol{x}\sim\mathsf{N}(\boldsymbol{0},\boldsymbol{\Sigma}_+)$.

With probability 1/2: $y=-1$, $\boldsymbol{x}\sim\mathsf{N}(\boldsymbol{0},\boldsymbol{\Sigma}_-)$.

We will assume $\boldsymbol{\Sigma}_+,\boldsymbol{\Sigma}_+$ to be diagonalizable in the same orthonormal basis, and to differ only on a subspace of dimension s_0. We want to study whether and how the neural network will identify this subspace of relevant features. Without loss of generality, we can assume that the eigenvalues correspond to the standard basis. In order to focus on the simplest possible model of this type, we will choose:

$$\boldsymbol{\Sigma}_+ = \mathrm{Diag}\big(\underbrace{(1+\Delta)^2,\dots,(1+\Delta)^2}_{s_0},\underbrace{1,\dots,1}_{d-s_0}\big)\,, \tag{5.1}$$

$$\boldsymbol{\Sigma}_- = \mathrm{Diag}\big(\underbrace{(1-\Delta)^2,\dots,(1-\Delta)^2}_{s_0},\underbrace{1,\dots,1}_{d-s_0}\big)\,. \tag{5.2}$$

We assume $0<\Delta<1$. As in the previous section, we choose $\sigma_*(\boldsymbol{x};\boldsymbol{\theta}_i)=\sigma(\langle\boldsymbol{x},\boldsymbol{w}_i\rangle)$ for some activation function σ. Define $q(r)\equiv\mathbb{E}\{\sigma(rG)\}$ for $G\sim\mathsf{N}(0,1)$. We assume $\sigma(\,\cdot\,)$ satisfies conditions S0 - S4 stated at the beginning of Section 4. We will still use the specific σ in Eq. (4.1) as our running example.

Throughout this section, we assume $s_0=\gamma\cdot d$ for some fixed $0<\gamma<1$. Therefore, as $d\to\infty$, we have $s_0=\gamma\cdot d\to\infty$ and $d-s_0=(1-\gamma)\cdot d\to\infty$. For any $\boldsymbol{w}\in\mathbb{R}^d$, we denote $\boldsymbol{w}_1\in\mathbb{R}^{s_0}$ and $\boldsymbol{w}_2\in\mathbb{R}^{d-s_0}$ by writing $\boldsymbol{w}=(\boldsymbol{w}_1,\boldsymbol{w}_2)$. We denote $\tau_+=1+\Delta$ and $\tau_-=1-\Delta$. Then we have $0<\tau_-<1<\tau_+<2$. Denote $q_+(r)=q(\tau_+ r)$ and $q_-(r)=q(\tau_- r)$. For any $\boldsymbol{a}=(a_1,a_2)\in\mathbb{R}^2$, denote

$$r_+(\boldsymbol{a}) = (\tau_+^2 a_1^2 + a_2^2)^{1/2},\quad r_-(\boldsymbol{a}) = (\tau_-^2 a_1^2 + a_2^2)^{1/2}. \tag{5.3}$$

Before analyzing our model, we introduce the function space and space of probability measures we will work on. Let $E_2 \equiv [0,+\infty)^2 \cup \{\infty\}$. Note there is a bijection ι between E_2 and $\mathbb{S}^2 \cap \{(x,y,x) \in \mathbb{R}^3 : x,y \geq 0\}$. Indeed, for any $\boldsymbol{r} = (r_1, r_2) \in [0,+\infty)^2$, consider the line crossing $(r_1, r_2, 0)$ and $(0,0,1)$. This line intersects with $\mathbb{S}^2$ at two points. One intersection point is $(0,0,1)$, and we denote the other intersection point as $\iota(\boldsymbol{r})$. Moreover, let $\iota(\infty) = (0,0,1)$. With this bijection ι, we equip E_2 with a metric $\bar{d}$ induced by the usual round metric on $\mathbb{S}^2$. Then $(E_2, \bar{d})$ is a compact metric space, and we will still denote it as E_2 for simplicity in notations. We denote $C_b(E_2)$ to be the set of bounded continuous functions on E_2, where continuity is defined using the topology generated by $\bar{d}$. More explicitly, we have isomorphism

$$C_b(E_2) \simeq \{f \in C([0,\infty)^2) : \exists f(\infty) \equiv \lim_{\|\boldsymbol{r}\|_2 \to \infty} f(\boldsymbol{r}), \sup_{\boldsymbol{r} \in E_2} f(\boldsymbol{r}) < \infty\}. \tag{5.4}$$

Because of condition S2 and S3, we have $q \circ r_+, q \circ r_-, q' \circ r_+, q' \circ r_- \in C_b(E_2)$.

Let $\mathscr{P}(E_2)$ be the set of probability measures on E_2. Due to Prokhorov's theorem, there exists a complete metric $\bar{d}_{\mathscr{P}}$ on $\mathscr{P}(E_2)$ equivalent to the topology of weak convergence, so that $(\mathscr{P}(E_2), \bar{d}_{\mathscr{P}})$ is a compact metric space. In this section, we will denote by $\overline{\mathscr{P}} = \mathscr{P}(E_2)$.

5.1 Statics

Since the distribution of $\boldsymbol{x}$ is invariant under rotations in first s_0 coordinates, and invariant under rotations in last $d - s_0$ coordinates, so are the functions

$$V(\boldsymbol{a}) = v(\|\boldsymbol{a}_1\|_2, \|\boldsymbol{a}_2\|_2), \tag{5.5}$$

$$U(\boldsymbol{a}, \boldsymbol{b}) = u_0(\|\boldsymbol{a}_1\|_2, \|\boldsymbol{b}_1\|_2, \langle \boldsymbol{a}_1, \boldsymbol{b}_1 \rangle, \|\boldsymbol{a}_2\|_2, \|\boldsymbol{b}_2\|_2, \langle \boldsymbol{a}_2, \boldsymbol{b}_2 \rangle). \tag{5.6}$$

These take the form

$$v(a_1, a_2) = -\frac{1}{2} q(r_+(a_1, a_2)) + \frac{1}{2} q(r_-(a_1, a_2)), \qquad q(t) = \mathbb{E}\{\sigma(tG)\}$$

and

$$\begin{aligned} &u_0(a_1, b_1, a_1 b_1 \cos\alpha, a_2, b_2, a_2 b_2 \cos\beta) \\ =&\frac{1}{2}\mathbb{E}\{\sigma(\tau_+ a_1 F_1 + a_2 G_1)\sigma(\tau_+ b_1 F_2 + b_2 G_2)\} + \frac{1}{2}\mathbb{E}\{\sigma(\tau_- a_1 F_1 + a_2 G_1)\sigma(\tau_- b_1 F_2 + b_2 G_2)\}, \end{aligned}$$

where expectations are with respect to standard normals $G, F_1, F_2, G_1, G_2 \sim \mathsf{N}(0,1)$, with (F_1, F_2) independent of (G_1, G_2). Moreover, (F_1, F_2) are jointly Gaussian, (G_1, G_2) are jointly Gaussian, and covariance $\mathbb{E}\{F_1 F_2\} = \cos\alpha$, $\mathbb{E}\{G_1 G_2\} = \cos\beta$.

In order to minimize $R(\rho)$, it is sufficient to restrict ourselves to distributions that are invariant under product of rotations. Indeed, for any probability distribution ρ on $\mathbb{R}^d$, we can define its symmetrization by letting, for any Borel set $Q_1 \subseteq \mathbb{R}^{s_0}$, $Q_2 \subseteq \mathbb{R}^{d-s_0}$,

$$\rho_s(Q_1 \times Q_2) \equiv \int \rho((\boldsymbol{R}_1 Q_1) \times (\boldsymbol{R}_2 Q_2)) \ \mu_{\text{Haar}}(\mathrm{d}\boldsymbol{R}_1)\mu_{\text{Haar}}(\mathrm{d}\boldsymbol{R}_2), \tag{5.7}$$

where μ_{Haar} is the Haar measure over the group of orthogonal rotations. Since $\rho \mapsto R(\rho)$ is convex, $R(\rho_s) \leq R(\rho)$.

We therefore restrict ourselves to ρ's that are invariant under product of rotations. In other words, under ρ, the vector $\boldsymbol{w} = (\boldsymbol{w}_1, \boldsymbol{w}_2) \in \mathbb{R}^d$ is sampled as following: $\boldsymbol{w}_1 \in \mathbb{R}^{s_0}$ is uniformly

random conditional on $\|\boldsymbol{w}_1\|_2$, and $\boldsymbol{w}_2 \in \mathbb{R}^{d-s_0}$ is uniformly random conditional on $\|\boldsymbol{w}_2\|_2$. We denote by $\overline{\rho} \in \mathscr{P}(E_2)$ the probability distribution of $(\|\boldsymbol{w}_1\|_2, \|\boldsymbol{w}_2\|_2)$ when $\boldsymbol{w} \sim \rho$ and we let $\overline{R}_d(\overline{\rho})$ denote the corresponding risk. We then have

$$\overline{R}_d(\overline{\rho}) = 1 + 2\int v(r_1, r_2)\,\overline{\rho}(\mathrm{d}\boldsymbol{r}) + \int u_d(a_1, a_2, b_1, b_2)\,\overline{\rho}(\mathrm{d}\boldsymbol{a})\,\overline{\rho}(\mathrm{d}\boldsymbol{b})\,, \tag{5.8}$$

and

$$u_d(a_1, a_2, b_1, b_2) = \mathbb{E}_{\Theta_1,\Theta_2}[u_0(a_1, b_1, a_1 b_1 \cos\Theta_1, a_2, b_2, a_2 b_2 \cos\Theta_2)], \tag{5.9}$$

where $\Theta_1 \sim (1/Z_{s_0}) \sin^{s_0-2}\theta \cdot \mathbf{1}\{\theta \in [0,\pi]\}\mathrm{d}\theta$ and $\Theta_2 \sim (1/Z_{d-s_0}) \sin^{d-s_0-2}\theta \cdot \mathbf{1}\{\theta \in [0,\pi]\}\mathrm{d}\theta$ are independent.

As $d \to \infty$, we have $\lim_{d\to\infty} u_d(a_1, a_2, b_1, b_2) = u_\infty(a_1, a_2, b_1, b_2)$, with

$$u_\infty(a_1, a_2, b_1, b_2) = \frac{1}{2}\Big[q(r_+(a_1, a_2))q(r_+(b_1, b_2)) + q(r_-(a_1, a_2))q(r_-(b_1, b_2))\Big], \tag{5.10}$$

and the risk function converges to (for $\boldsymbol{a} = (a_1, a_2)$)

$$\overline{R}_\infty(\overline{\rho}) = \frac{1}{2}\left(1 - \int q(r_+(\boldsymbol{a}))\,\overline{\rho}(\mathrm{d}\boldsymbol{a})\right)^2 + \frac{1}{2}\left(1 + \int q(r_-(\boldsymbol{a}))\,\overline{\rho}(\mathrm{d}\boldsymbol{a})\right)^2. \tag{5.11}$$

We also define

$$\psi_d(\boldsymbol{a}; \overline{\rho}) = v(\boldsymbol{a}) + \int u_d(\boldsymbol{a}, \boldsymbol{b})\,\overline{\rho}(\mathrm{d}\boldsymbol{b})\,. \tag{5.12}$$

For $s_0 = \gamma \cdot d$ with $0 < \gamma < 1$ and $d \to \infty$, we have the simpler expression

$$\psi_\infty(\boldsymbol{a}; \overline{\rho}) = \lambda_+(\overline{\rho}) \cdot q(r_+(\boldsymbol{a})) + \lambda_-(\overline{\rho}) \cdot q(r_-(\boldsymbol{a})), \tag{5.13}$$

$$\lambda_+(\overline{\rho}) = \frac{1}{2}[\langle q \circ r_+, \overline{\rho}\rangle - 1], \tag{5.14}$$

$$\lambda_-(\overline{\rho}) = \frac{1}{2}[\langle q \circ r_-, \overline{\rho}\rangle + 1]. \tag{5.15}$$

The following theorem provides a characterization of the global minimizers of $\overline{R}_\infty(\overline{\rho})$.

Theorem 5.1. *Consider $d = \infty$. Recall $\overline{\mathscr{P}} = \mathscr{P}(E_2)$ where $E_2 \equiv [0, +\infty)^2 \cup \{\infty\}$. Then there exists $\Delta_\infty \in (0, 1)$, such that*

1. *For $\Delta < \Delta_\infty$, $\inf_{\overline{\rho}\in\overline{\mathscr{P}}} \overline{R}_\infty(\overline{\rho}) > 0$ and the unique global minimizer of risk function $\overline{R}_\infty(\overline{\rho})$ is a point mass located at $(r_*, 0)$ for some $r_* = r_*(\Delta) \in (0, \infty)$.*
2. *For $\Delta \geq \Delta_\infty$, all global minimizers of risk function $\overline{R}_\infty(\overline{\rho})$ have risk zero, and there exists a global minimizer that has finite support.*

Proof. Throughout the proof, we will denote $\overline{R}^{(1)}_\infty : \mathscr{P}([0,\infty]) \to \mathbb{R}$ as the risk function defined as in Eq. (4.10), and $\overline{R}^{(2)}_\infty : \mathscr{P}(E_2) \to \mathbb{R}$ as the risk function defined as in Eq. (5.11). Recall the definition $\tau_+ = 1 + \Delta$, $\tau_- = 1 - \Delta$, $q_+(r) = q(\tau_+ r)$, $q_-(r) = q(\tau_- r)$, $r_+(\boldsymbol{a}) = (\tau_+^2 a_1^2 + a_2^2)^{1/2}$, and $r_-(\boldsymbol{a}) = (\tau_-^2 a_1^2 + a_2^2)^{1/2}$ for $\boldsymbol{a} = (a_1, a_2) \in E_2$.

Suppose $\overline{\rho}_2^* \in \arg\min_{\overline{\rho}_2 \in \mathscr{P}(E_2)} \overline{R}_\infty^{(2)}(\overline{\rho}_2)$. Then we must have $\langle q \circ r_+, \overline{\rho}_2^* \rangle \le 1$ and $\langle q \circ r_-, \overline{\rho}_2^* \rangle \ge -1$. Indeed, if either $\langle q \circ r_+, \overline{\rho}_2^* \rangle > 1$ or $\langle q \circ r_-, \overline{\rho}_2^* \rangle < -1$, since $q(+\infty) > 1$ and $q(0) < -1$, the distribution $\overline{\rho}_2' = a_0 \delta_{\mathbf{0}} + a_\infty \delta_\infty + (1 - a_0 - a_\infty)\overline{\rho}_2^*$ with appropriate choice of a_0 and a_∞ will give a lower risk.

This $\overline{\rho}_2^* \in \mathscr{P}(E_2)$ induces a $\overline{\rho}_1 \in \mathscr{P}([0,\infty])$ as follows: for any Borel set $B \subseteq [0,\infty]$, $\overline{\rho}_1(B) = \overline{\rho}_2^*(\{\boldsymbol{r} \in E_2 : \|\boldsymbol{r}\|_2 \in B\})$. For this $\overline{\rho}_1$, it is easy to see that $\langle q_-, \overline{\rho}_1 \rangle \le \langle q \circ r_-, \overline{\rho}_2^* \rangle$ and $\langle q_+, \overline{\rho}_1 \rangle \ge \langle q \circ r_+, \overline{\rho}_2^* \rangle$, and the equalities hold if and only if $\overline{\rho}_2^*(E_1) = 1$, where $E_1 \equiv ([0,+\infty) \times \{0\}) \cup \{\infty\}$. Since $q(+\infty) > 1$ and $q(0) < -1$, we can take $\overline{\rho}_1^* = a_0 \delta_0 + a_\infty \delta_\infty + (1 - a_0 - a_\infty)\overline{\rho}_1$ with appropriate choice of a_0 and a_∞, so that $\langle q \circ r_+, \overline{\rho}_2^* \rangle \le \langle q_+, \overline{\rho}_1^* \rangle \le 1$ and $\langle q \circ r_-, \overline{\rho}_2^* \rangle \ge \langle q_-, \overline{\rho}_1^* \rangle \ge -1$. Therefore, we always have $\inf_{\overline{\rho}_1 \in \mathscr{P}([0,\infty])} \overline{R}_\infty^{(1)}(\overline{\rho}_1) \le \inf_{\overline{\rho}_2 \in \mathscr{P}(E_2)} \overline{R}_\infty^{(2)}(\overline{\rho}_2)$, and $\overline{\rho}_2^*(E_1) = 1$ for any $\overline{\rho}_2^* \in \arg\min_{\overline{\rho}_2 \in \mathscr{P}(E_2)} \overline{R}_\infty^{(2)}(\overline{\rho}_2)$. Note that $\overline{R}_\infty^{(2)}(\overline{\rho}_1 \times \delta_0) = \overline{R}_\infty^{(1)}(\overline{\rho}_1)$ for any $\overline{\rho}_1 \in \mathscr{P}([0,\infty])$. Hence, we must have $\inf_{\overline{\rho}_1 \in \mathscr{P}([0,\infty])} \overline{R}_\infty^{(1)}(\overline{\rho}_1) = \inf_{\overline{\rho}_2 \in \mathscr{P}(E_2)} \overline{R}_\infty^{(2)}(\overline{\rho}_2)$.

Due to the above argument, we reduced our analysis to the centered isotropic Gaussians case. All the conclusions can be proved using the same argument as in the proof of Theorem 4.2.

□

5.2 Dynamics: Fixed points

We specialize the general evolution (3.1) to the present case. Assuming ρ_0 to be invariant with respect to products of orthogonal transformations, the same happens for ρ_t. We let $\overline{\rho}_t \in \mathscr{P}(E_2)$ denote the distribution of $(\|\boldsymbol{w}_1\|_2, \|\boldsymbol{w}_2\|_2)$ when $\boldsymbol{w} \sim \rho_t$. Then $\overline{\rho}_t$ satisfies the following PDE:

$$\partial_t \overline{\rho}_t(\boldsymbol{r}) = 2\xi(t) \nabla \cdot \left[\overline{\rho}_t(\boldsymbol{r}) \nabla \psi_d(\boldsymbol{r}; \overline{\rho}_t)\right]. \tag{5.16}$$

We will view this as an evolution in the space of probability distribution on $\overline{\mathscr{P}} = \mathscr{P}(E_2)$.

In analogy with Proposition 2, we can prove the following characterization of fixed points.

Proposition 5.2. *A distribution $\overline{\rho} \in \overline{\mathscr{P}}$ is a fixed point of the PDE (5.16) if and only if*

$$\mathrm{supp}(\overline{\rho}) \subseteq \{\boldsymbol{r} \in E_2 : \ \nabla_{\boldsymbol{r}} \psi_d(\boldsymbol{r}; \overline{\rho}) = \mathbf{0}\}. \tag{5.17}$$

Notice, in particular, global minimizers of $\overline{R}_d(\overline{\rho})$ are fixed points of this evolution, but not vice-versa. The next result classifies fixed points.

Theorem 5.3. *Consider $d = \infty$, and recall the definition of $\lambda_+(\overline{\rho})$ and $\lambda_-(\overline{\rho})$ given by Eq. (5.15) and (5.14). Then the fixed points of the PDE (5.16) (i.e. the probability measures $\overline{\rho} \in \overline{\mathscr{P}}$ satisfying (5.17)) must be of one of the following types*

(a) *A fixed point with zero risk.*

(b) *A point mass $\overline{\rho}_{r_*} = \delta_{(r_*,0)}$ at some location $(r_*, 0)$ with $r_* \notin \{0, +\infty\}$, but not of type (a).*

(c) *A mixture of the type $\overline{\rho} = a_0 \delta_{\mathbf{0}} + a_\infty \delta_\infty + a_1 \delta_{(r_{*1},0)} + a_2 \overline{\rho}_2$ with $\mathrm{supp}(\overline{\rho}_2) \subseteq \{0\} \times (0,\infty)$, but not of type (b) and (a).*

For $\Delta < \Delta_\infty$, the PDE has a unique fixed point of type (b), with $\lambda_+(\overline{\rho}_*) < 0$ and $\lambda_-(\overline{\rho}_*) > 0$; it has no type-(a) fixed points; it has possibly fixed points of type (c).

For $\Delta > \Delta_\infty$, the PDE has some fixed points of type (b), with $\lambda_+(\overline{\rho}_*) > 0$ and $\lambda_-(\overline{\rho}_*) < 0$; it also has some type-(a) fixed points; it has possibly fixed points of type (c).

For $\Delta = \Delta_\infty$, the PDE has a unique fixed point of type (a) which is also a delta function at some location $(r_{*1}, 0)$, and no type (b) fixed points; it has possibly fixed points of type (c).

Proof. We use the characterization of fixed points in Proposition 5.2. Recall that $\psi_\infty(\boldsymbol{r};\overline{\rho}_*)$ is defined as in Eq. (5.13). The gradient $\nabla\psi_\infty(\boldsymbol{r};\overline{\rho})$ is given by

$$\begin{aligned}\partial_{r_1}\psi_\infty(\boldsymbol{r};\overline{\rho}) =&\lambda_+(\overline{\rho})q'(r_+(\boldsymbol{r}))\tau_+^2 r_1/r_+(\boldsymbol{r}) + \lambda_-(\overline{\rho})q'(r_-(\boldsymbol{r}))\tau_-^2 r_1/r_-(\boldsymbol{r}),\\ \partial_{r_2}\psi_\infty(\boldsymbol{r};\overline{\rho}) =&\lambda_+(\overline{\rho})q'(r_+(\boldsymbol{r}))r_2/r_+(\boldsymbol{r}) + \lambda_-(\overline{\rho})q'(r_-(\boldsymbol{r}))r_2/r_-(\boldsymbol{r}).\end{aligned} \tag{5.18}$$

If a fixed point $\overline{\rho}_*$ gives $\lambda_+(\overline{\rho}_*) = \lambda_-(\overline{\rho}_*) = 0$, then $\overline{R}_\infty(\overline{\rho}_*) = 0$. This is type-$(a)$ fixed point. Consider then the case $(\lambda_+(\overline{\rho}_*), \lambda_-(\overline{\rho}_*)) \neq (0,0)$.

Suppose $\overline{\rho}_*((0,+\infty)^2) > 0$. Since $q'(r) > 0$ and $\tau_+ > 1 > \tau_-$, in order for $\nabla\psi_\infty(\boldsymbol{r};\overline{\rho}_*) = \mathbf{0}$ for some $\boldsymbol{r} \in (0,+\infty)^2$, we must have $(\lambda_+(\overline{\rho}_*), \lambda_-(\overline{\rho}_*)) = (0,0)$. Therefore, as $\overline{\rho}_*$ is a fixed point with $(\lambda_+(\overline{\rho}_*), \lambda_-(\overline{\rho}_*)) \neq (0,0)$, we must have $\overline{\rho}_*((0,+\infty)^2) = 0$. That is, we can write $\overline{\rho}_* = a_0\delta_{\mathbf{0}} + a_\infty\delta_\infty + a_1\overline{\rho}_1 + a_2\overline{\rho}_2$, with $\mathrm{supp}(\overline{\rho}_1) \in (0,\infty)\times\{0\}$, and $\mathrm{supp}(\overline{\rho}_2) \in \{0\}\times(0,\infty)$.

The solutions of $\nabla\psi_\infty((r_1,r_2);\overline{\rho}_*) = 0$ with $r_2 = 0$ are of the form $\mathbf{0}$, $(r_{*1},0)$, and ∞. Therefore, $\overline{\rho}_1 = \delta_{(r_{*1},0)}$ for some $r_{*1} \in (0,\infty)$. Hence, as $\overline{\rho}_*$ is not a type-(a) stationary point, it must be a type-(b) or type-(c) stationary point.

This proves that all fixed points are of type (a), (b), or (c). The remaining claims follows the same argument as the proof of Theorem 4.4.

□

5.3 Dynamics: Convergence to global minimum for $d = \infty$

In this section, denote $\mathscr{P}_{\mathrm{good}}$ to be

$$\mathscr{P}_{\mathrm{good}} = \{\overline{\rho}_0 \in \mathscr{P}((0,\infty)^2) : \overline{R}_\infty(\overline{\rho}_0) < 1\}. \tag{5.19}$$

We then prove that the $d = \infty$ dynamics converges to a global minimizer from any initialization $\overline{\rho}_0 \in \mathscr{P}_{\mathrm{good}}$.

Theorem 5.4. *Consider the PDE (5.16) for $d = \infty$, with initialization $\overline{\rho}_0 \in \mathscr{P}_{\mathrm{good}}$. It has a unique solution $(\overline{\rho}_t)_{t\geq 0}$, such that*

$$\lim_{t\to+\infty} \overline{R}_\infty(\overline{\rho}_t) = \inf_{\overline{\rho}\in\overline{\mathscr{P}}} \overline{R}_\infty(\overline{\rho})\,.$$

Proof. Without loss of generality, we assume $\xi(t) = 1/2$. First we show the existence and uniqueness of solution of the PDE.

Step 1. Existence and uniqueness of solution. Mass $\overline{\rho}_t((0,\infty)^2) = 1$ for all t.

According to conditions S1 - S3, $q(r)$, $q'(r)$, and $q''(r)$ are uniformly bounded on $[0,\infty]$. Note

$$\begin{aligned}v(\boldsymbol{r}) =&1/2\cdot[q(r_-(\boldsymbol{r})) - q(r_+(\boldsymbol{r}))],\\ u_\infty(\boldsymbol{r}_1,\boldsymbol{r}_2) =&1/2\cdot[q(r_+(\boldsymbol{r}_1))q(r_+(\boldsymbol{r}_2)) + q(r_-(\boldsymbol{r}_1))q(r_-(\boldsymbol{r}_2))].\end{aligned}$$

Then $\nabla v(\boldsymbol{r}), \nabla_1 u_\infty(\boldsymbol{r}_1,\boldsymbol{r}_2), \nabla^2 v(\boldsymbol{r}), \nabla_{11}^2 u_\infty(\boldsymbol{r}_1,\boldsymbol{r}_2), \nabla_{12}^2 u_\infty(\boldsymbol{r}_1,\boldsymbol{r}_2)$ are uniformly bounded. Therefore, conditions A1 and A3 are satisfied with $D = 2$, $V = v$, and $U = u$. Then, there is the existence and uniqueness of solution of PDE (5.16) for $d = \infty$. Denote this solution to be $(\overline{\rho}_t)_{t\geq 0}$.

Recall the expression for $\nabla\psi_\infty(\boldsymbol{r};\overline{\rho})$ in Eq. (5.18). It is easy to see that the assumption of Lemma 3.9 is satisfied with $d = 2$ and $\Psi = \psi_\infty$. Hence, we have $\overline{\rho}_t((0,\infty)^2) = 1$ for any fixed $t < \infty$.

Step 2. Classify the limiting set $\mathcal{S}_*$.

Recall the definition of $(\mathscr{P}(E_2), \bar{d}_{\mathscr{P}})$ at the beginning of Section 5. Since $(\mathscr{P}(E_2), \bar{d}_{\mathscr{P}})$ is a compact metric space, and $(\overline{\rho}_t)_{t\geq 0}$ is a continuous curve in this space, then there exists a subsequence $(t_k)_{k\geq 1}$ of times, such that $(\overline{\rho}_{t_k})_{k\geq 1}$ converges in metric $\bar{d}_{\mathscr{P}}$ to a probability distribution $\overline{\rho}_* \in \mathscr{P}(E_2)$.

For any $\overline{\rho}_0 \in \mathscr{P}_{\text{good}}$, let $\mathcal{S}_* = \mathcal{S}_*(\overline{\rho}_0)$ be the set of limiting points of the PDE,

$$\mathcal{S}_* = \{\overline{\rho}_* \in \mathscr{P}(E_2) : \exists (t_k)_{k\geq 1}, \lim_{k\to\infty} t_k = +\infty, s.t., \lim_{k\to\infty} \bar{d}_{\mathscr{P}}(\overline{\rho}_*, \overline{\rho}_{t_k}) = 0\}.$$

Analogous to the proof of Theorem 4.5, we have the following properties for $\mathcal{S}_*$:

1. $\mathcal{S}_*$ is connected and compact.
2. For any $\overline{\rho}_* \in \mathcal{S}_*$, $\overline{\rho}_*$ is a fixed point of PDE.
3. For any $\overline{\rho}_* \in \mathcal{S}_*$, $\overline{R}_\infty(\overline{\rho}_*) = \overline{R}_* < 1$.

Recall the definition of $\lambda_+(\overline{\rho}_*)$ and $\lambda_-(\overline{\rho}_*)$ given by Equation (5.14) and (5.15). Let $\overline{\rho}_*$ be a fixed point of PDE such that $\lambda_+(\overline{\rho}_*) \geq 0, \lambda_-(\overline{\rho}_*) \geq 0$ or $\lambda_+(\overline{\rho}_*) \leq 0, \lambda_-(\overline{\rho}_*) \leq 0$ but not both $\lambda_+(\overline{\rho}_*)$ and $\lambda_-(\overline{\rho}_*)$ equal 0. In this case, according to Eq. (5.18), both $\partial_{r_1}\psi_\infty(\boldsymbol{r}; \overline{\rho}_*)$ and $\partial_{r_2}\psi_\infty(\boldsymbol{r}; \overline{\rho}_*)$ must be strictly positive or strictly negative. Since $\text{supp}(\overline{\rho}_*) \subseteq \{\boldsymbol{r} \in E_2 : \nabla_{\boldsymbol{r}}\psi_\infty(\boldsymbol{r}; \overline{\rho}_*) = \mathbf{0}\}$, $\overline{\rho}_*$ must be a combination of two delta functions located at $\mathbf{0}$ and ∞, i.e., $\overline{\rho}_* = a_0\delta_{\mathbf{0}} + (1-a_0)\delta_\infty$. But for a fixed point like this, it is easy to see that $\overline{R}_\infty(\overline{\rho}_*) \geq 1$. Such fixed points $\overline{\rho}_*$ cannot be one of the limiting points of the PDE since $\overline{R}_\infty(\overline{\rho}_0) < 1$.

Let L be a mapping $L : \mathscr{P}(E_2) \to \mathbb{R}^2$, $\overline{\rho} \mapsto (\lambda_+(\overline{\rho}), \lambda_-(\overline{\rho}))$. The above argument concludes that for any $\overline{\rho}_0 \in \mathscr{P}_{\text{good}}$, we have

$$L(\mathcal{S}_*(\overline{\rho}_0)) \cap (\{(\lambda_+, \lambda_-) : \lambda_+ \geq 0, \lambda_- \geq 0, \text{ or } \lambda_+ \leq 0, \lambda_- \leq 0\} \setminus \{(0,0)\}) = \emptyset.$$

Since $\mathcal{S}_*$ is a connected set, $L(\mathcal{S}_*)$ should also be a connected set. Further notice that $\overline{R}_\infty(\overline{\rho}_*) = 1/2 \cdot [\lambda_+(\overline{\rho}_*)^2 + \lambda_-(\overline{\rho}_*)^2]$, and $\overline{R}_\infty(\overline{\rho}_1) = \overline{R}_\infty(\overline{\rho}_2)$ for any $\overline{\rho}_1, \overline{\rho}_2 \in \mathcal{S}_*$. Therefore, we can only have $L(\mathcal{S}_*) \subseteq \mathcal{P}_2 \equiv \{(\lambda_+, \lambda_-) : \lambda_+ > 0, \lambda_- < 0\}$, or $L(\mathcal{S}_*) \subseteq \mathcal{P}_1 \equiv \{(\lambda_+, \lambda_-) : \lambda_+ < 0, \lambda_- > 0\}$, or $L(\mathcal{S}_*) = \{(0,0)\}$.

Step 3. Finish the proof using two claims.

We make the following two claims.

Claim (1). If $L(\mathcal{S}_*) \subseteq \mathcal{P}_1$, then for any $\overline{\rho}_* \in \mathcal{S}_*$, we have $\overline{\rho}_*((0,\infty) \times \{0\}) = 1$.

Claim (2). We cannot have $L(\mathcal{S}_*) \subseteq \mathcal{P}_2$.

Here we assume these two claims holds, and use it to prove our results. For $\Delta < \Delta_\infty$, we proved in Theorem 5.3 that, there is no fixed point such that $L(\overline{\rho}_*) = (0,0)$. Therefore, we cannot have $L(\mathcal{S}_*) = \{(0,0)\}$. Due to Claim (2), we cannot have $L(\mathcal{S}_*) \subseteq \mathcal{P}_2$. Hence, we must have $L(\mathcal{S}_*) \subseteq \mathcal{P}_1$. According to Theorem 5.3, for $\Delta < \Delta_\infty$, the only fixed point of PDE with $\overline{\rho}_*((0,\infty) \times \{0\}) = 1$ is a point mass at some location $\boldsymbol{r}_* = (r_{*1}, 0)$. Furthermore, this delta function fixed point is unique and is also the global minimizer of the risk. Therefore, we conclude that, for $\Delta < \Delta_\infty$, the PDE will converge to this global minimizer.

For $\Delta \geq \Delta_\infty$, according to Claim (1), if $\overline{\rho}_*$ is a limiting point such that $L(\overline{\rho}_*) \in \mathcal{P}_1$, then $\overline{\rho}_*((0,\infty) \times \{0\}) = 1$. According to Theorem 5.3, a fixed point $\overline{\rho}_*$ with $\overline{\rho}_*((0,\infty) \times \{0\}) = 1$ and $L(\overline{\rho}_*) \neq (0,0)$ must be a point mass at some location $\boldsymbol{r}_* = (r_{*1}, 0)$, with $L(\overline{\rho}_*) \in \mathcal{P}_2$. Therefore, we cannot have $L(\mathcal{S}_*) \subseteq \mathcal{P}_1$. Claim (2) also tells us that we cannot have $L(\mathcal{S}_*) \subseteq \mathcal{P}_2$. Hence, we must

have $L(\mathcal{S}_*) = \{(0,0)\}$. In this case, all the points in the set $\mathcal{S}_*$ have risk 0. Therefore, we conclude that, as $\Delta \geq \Delta_\infty$, the PDE will converge to some limiting set with risk 0.

Step 4. Proof of the two claims.

We are left with the task of proving the two claims above. Before that, we introduce some useful notions used in the proof. Define $Z(\boldsymbol{r})$ for $\boldsymbol{r} \in E_2$,

$$Z(\boldsymbol{r}) \equiv [q'(r_-(\boldsymbol{r}))r_-(\boldsymbol{r})]/[q'(r_+(\boldsymbol{r}))r_+(\boldsymbol{r})]. \tag{5.20}$$

Define $Z_l(r) \equiv Z((r, lr))$ for $r, l \in [0, \infty]$. Then we have

$$Z_l(r) = [q'((\tau_-^2 + l^2)^{1/2}r)/q'((\tau_+^2 + l^2)^{1/2}r)] \cdot [(\tau_-^2 + l^2)^{1/2}/(\tau_+^2 + l^2)^{1/2}]. \tag{5.21}$$

According to condition S4, for any fixed $l \in [0, \infty]$, $Z_l(r)$ is increasing in r.

Recall the formula of $\nabla_{\boldsymbol{r}}\psi_\infty(\boldsymbol{r}; \overline{\rho})$ given by Equation (5.18). Define

$$\chi_{\rm nm}(\boldsymbol{r}; \overline{\rho}) \equiv \langle \nabla_{\boldsymbol{r}}\psi_\infty(\boldsymbol{r}; \overline{\rho}), \boldsymbol{r}/\|\boldsymbol{r}\|_2 \rangle, \tag{5.22}$$

$$\chi_{\rm tg}(\boldsymbol{r}; \overline{\rho}) \equiv \langle \nabla_{\boldsymbol{r}}\psi_\infty(\boldsymbol{r}; \overline{\rho}), (-r_2, r_1)/\|\boldsymbol{r}\|_2 \rangle. \tag{5.23}$$

Then we have

$$\begin{aligned} \chi_{\rm nm}(\boldsymbol{r}; \overline{\rho}) =& \lambda_+(\overline{\rho})q'(r_+(\boldsymbol{r}))r_+(\boldsymbol{r})/\|\boldsymbol{r}\|_2 + \lambda_-(\overline{\rho})q'(r_-(\boldsymbol{r}))r_-(\boldsymbol{r})/\|\boldsymbol{r}\|_2, \\ =& \lambda_-(\overline{\rho})q'(r_+(\boldsymbol{r}))r_+(\boldsymbol{r})/\|\boldsymbol{r}\|_2 \cdot [\lambda_+(\overline{\rho})/\lambda_-(\overline{\rho}) + Z(\boldsymbol{r})], \end{aligned} \tag{5.24}$$

and

$$\begin{aligned} \chi_{\rm tg}(\boldsymbol{r}; \overline{\rho}) =& [+\lambda_+(\overline{\rho})(1 - \tau_+^2)q'(r_+(\boldsymbol{r}))/r_+(\boldsymbol{r}) \\ & + \lambda_-(\overline{\rho})(1 - \tau_-^2)q'(r_-(\boldsymbol{r}))/r_-(\boldsymbol{r})] \times r_1 r_2/\|\boldsymbol{r}\|_2. \end{aligned} \tag{5.25}$$

Proof of Claim (1)**. If $L(\mathcal{S}_*) \subseteq \mathcal{P}_1$, then for any $\overline{\rho}_* \in \mathcal{S}_*$, we have $\overline{\rho}_*((0,\infty) \times \{0\}) = 1$.**

Assume $L(\mathcal{S}_*) \subseteq \mathcal{P}_1$. There must exist t_0 large enough, so that as $t \geq t_0$, we have $\lambda_+(\overline{\rho}_t) < 0$, and $\lambda_-(\overline{\rho}_t) > 0$. Therefore, we must have $\chi_{\rm tg}(\boldsymbol{r}; \overline{\rho}_t) > 0$ for any $\boldsymbol{r} \in (0,\infty)^2$. We denote

$$\Gamma_k \equiv \{\boldsymbol{r} \in [0,\infty)^2 : r_2 \leq k \cdot r_1\}. \tag{5.26}$$

Consider the ODE

$$\dot{\boldsymbol{r}}(t) = -\nabla_{\boldsymbol{r}}\psi_\infty(\boldsymbol{r}(t); \overline{\rho}_t), \tag{5.27}$$

starting with $\boldsymbol{r}(t_0) \in \Gamma_k$ for some $k \in (0,\infty)$, we claim $\boldsymbol{r}(t) \in \Gamma_k$ for any $t \geq t_0$. Indeed, for any $\boldsymbol{r} \in \partial\Gamma_k \cap \{\boldsymbol{r} : r_2 = kr_1 > 0\}$, its normal vector pointing outside Γ_k gives $\boldsymbol{n}(\boldsymbol{r}) = (-r_2, r_1)/\|\boldsymbol{r}\|_2$, and hence $\langle \nabla_{\boldsymbol{r}}\psi_\infty(\boldsymbol{r}; \overline{\rho}), \boldsymbol{n}(\boldsymbol{r}) \rangle = \chi_{\rm tg}(\boldsymbol{r}; \overline{\rho}_t) > 0$. Therefore, $\boldsymbol{r}(t)$ cannot leak outside Γ_k from this boundary. Further note that $\boldsymbol{r}(t)$ cannot reach the boundary $([0,\infty) \times \{0\}) \cup \{\infty\}$ for any finite time t. This proves the claim that $\boldsymbol{r}(t) \in \Gamma_k$ for any $t \geq t_0$.

According to Lemma 3.8, we have $\rho_t(\Gamma_k) \geq \rho_{t_0}(\Gamma_k)$ for any $k \in (0,\infty)$. Furthermore, according to Lemma 3.9, $\overline{\rho}_{t_0}((0,\infty)^2) = 1$, hence $\lim_{k\to\infty} \overline{\rho}_{t_0}(\Gamma_k) = 1$. Therefore, for any $\overline{\rho}_* \in \mathcal{S}_*$, we must have

$$\overline{\rho}_*(\{0\} \times (0,\infty)) \leq \lim_{k\to\infty} \overline{\rho}_*([0,\infty)^2 \setminus \Gamma_k) \leq \lim_{k\to\infty} \overline{\rho}_{t_0}([0,\infty)^2 \setminus \Gamma_k) = 0. \tag{5.28}$$

Theorem 5.3 implies that for any such fixed point $\overline{\rho}_*$, we have $\mathrm{supp}(\overline{\rho}_*) \subseteq ([0,\infty) \times \{0\}) \cup \{\infty\}$.

In this case, we claim $L(\mathcal{S}_*) \subseteq \mathcal{P}_1 \cap \{(\lambda_+, \lambda_-) : Z_0(0) < -\lambda_+/\lambda_- < Z_0(\infty)\}$. Indeed, suppose there exists $\overline{\rho}_* \in \mathcal{S}_*$, such that $-\lambda_+(\overline{\rho}_*)/\lambda_-(\overline{\rho}_*) \geq Z_0(\infty)$ or $-\lambda_-(\overline{\rho}_*)/\lambda_-(\overline{\rho}_*) \leq Z_0(0)$, according

to Equation (5.24), $\chi_{\rm nm}((r,0);\overline{\rho}_*)$ must be strictly positive or strictly negative. However, we know $\mathrm{supp}(\overline{\rho}_*) \in \{\boldsymbol{r} : \nabla\psi_\infty(\boldsymbol{r};\overline{\rho}_*) = \mathbf{0}\}$. Hence, $\overline{\rho}_*$ should be a combination of two delta functions located at $\mathbf{0}$ and ∞. Such fixed point $\overline{\rho}_*$ has risk $\overline{R}_\infty(\overline{\rho}_*) \geq 1$, hence $\overline{\rho}_*$ cannot be a limiting point of the PDE. Hence the claim holds.

Since $\mathcal{S}_*$ is a compact set, and L is a continuous map, then $L(\mathcal{S}_*)$ is a compact set. Therefore, there must exist $\varepsilon_0 > 0$, so that for any $\overline{\rho}_* \in \mathcal{S}_*$, we have $Z_0(0) + 3\varepsilon_0 < -\lambda_+(\overline{\rho}_*)/\lambda_-(\overline{\rho}_*) < Z_0(\infty) - 3\varepsilon_0$. For this $\varepsilon_0 > 0$, we take t_0 large enough, so that for $t \geq t_0$, we have $Z_0(0) + 2\varepsilon_0 < -\lambda_+(\overline{\rho}_t)/\lambda_-(\overline{\rho}_t) < Z_0(\infty) - 2\varepsilon_0$, and $\lambda_+(\overline{\rho}_t) < 0$, $\lambda_-(\overline{\rho}_t) > 0$.

According to the conditions S0 - S4 on $q(r)$, for any fixed l, $Z_l(r)$ is an increasing function of r, and for any fixed r, $Z_l(r)$ is continuous in l. Therefore, for the fixed $\varepsilon_0 > 0$, there exists $0 < r_0 < r_\infty < \infty$ and $b > 0$, such that

$$\sup_{r\in[0,r_0]} \sup_{l\in[0,b]} Z_l(r) < Z_0(0) + \varepsilon_0, \tag{5.29}$$

$$\inf_{r\in[r_\infty,\infty]} \inf_{l\in[0,b]} Z_l(r) > Z_0(\infty) - \varepsilon_0. \tag{5.30}$$

As a result, for any $t \geq t_0$, we have

$$\begin{aligned} \chi_{\rm nm}(\boldsymbol{r};\overline{\rho}_t) <& 0, \quad \forall \boldsymbol{r} \in \mathsf{B}(\mathbf{0}; r_0) \cap \Gamma_b, \\ \chi_{\rm nm}(\boldsymbol{r};\overline{\rho}_t) >& 0, \quad \forall \boldsymbol{r} \in \mathsf{B}(\mathbf{0}; r_\infty)^c \cap \Gamma_b, \end{aligned} \tag{5.31}$$

where $\Gamma_{(\cdot)}$ is defined as in Equation (5.26).

According to Lemma 3.9, $\overline{\rho}_{t_0}((0,\infty)^2) = 1$. Define

$$O_k = \Gamma_k \cap \mathsf{B}(\mathbf{0}; k) \cap \mathsf{B}(\mathbf{0}; 1/k)^c. \tag{5.32}$$

We have O_k is increasing in k, and $\cup_k O_k \supset (0,\infty)^2$. Hence $\lim_{k\to\infty} \overline{\rho}_{t_0}(O_k) = 1$. Now we fix a parameter k.

Recall the formula for $\chi_{\rm nm}$ and $\chi_{\rm tg}$ given by Equation (5.24) and (5.25). It is easy to see that, there exists $0 < u_{k1}, u_{k2} < \infty$ depending on $(b, k, \tau_+, \tau_-, Z_0(0), Z_0(\infty), \varepsilon_0)$, such that for any $\boldsymbol{r} \in (0,\infty)^2$ with $b \cdot r_1 \leq r_2 \leq k \cdot r_1$, and $t \geq t_0$, we have

$$\chi_{\rm tg}(\boldsymbol{r};\overline{\rho}_t) \geq u_{k1}|\lambda_+(\overline{\rho}_t)|q'(r_+(\boldsymbol{r})) > 0, \tag{5.33}$$

$$|\chi_{\rm nm}(\boldsymbol{r};\overline{\rho}_t)| \leq u_{k2}|\lambda_+(\overline{\rho}_t)|q'(r_+(\boldsymbol{r})) < \infty, \tag{5.34}$$

and hence

$$|\chi_{\rm nm}(\boldsymbol{r};\overline{\rho}_t)|/\chi_{\rm tg}(\boldsymbol{r};\overline{\rho}_t) \leq u_{k2}/u_{k1} \equiv u_k < \infty. \tag{5.35}$$

Consider the following spiral curve $\boldsymbol{r}_k^\infty(s) = (r_{k1}^\infty(s), r_{k2}^\infty(s))$, with

$$\begin{aligned} r_{k1}^\infty(s) =& k \cdot \cos(\arctan(k) - s)\exp\{2u_k s\}, \\ r_{k2}^\infty(s) =& k \cdot \sin(\arctan(k) - s)\exp\{2u_k s\}, \end{aligned} \tag{5.36}$$

and another spiral curve $\boldsymbol{r}_k^0(s) = (r_{k1}^0(s), r_{k2}^0(s))$, with

$$\begin{aligned} r_{k1}^0(s) =& 1/k \cdot \cos(\arctan(k) - s)\exp\{-2u_k s\}, \\ r_{k2}^0(s) =& 1/k \cdot \sin(\arctan(k) - s)\exp\{-2u_k s\}, \end{aligned} \tag{5.37}$$

for $s \in [0, s_{k*}]$ with $s_{k*} = \arctan(k) - \arctan(b)$.

Because of inequality (5.35), along the curve $\boldsymbol{r}_k^\infty(s)$, denoting $\boldsymbol{n}(\boldsymbol{r}_k^\infty(s))$ to be its normal vector with $[\boldsymbol{n}(\boldsymbol{r}_k^\infty(s))]_2 > 0$, we have for any $t \geq t_0$ and $s \in [0, s_{k*}]$,

$$\langle \nabla \psi_\infty(\boldsymbol{r}_k^\infty(s); \overline{\rho}_t), \boldsymbol{n}(\boldsymbol{r}_k^\infty(s)) \rangle > 0. \tag{5.38}$$

Along the curve $\boldsymbol{r}_k^0(s)$, denoting $\boldsymbol{n}(\boldsymbol{r}_k^0(s))$ to be its normal vector with $[\boldsymbol{n}(\boldsymbol{r}_k^0(s))]_2 > 0$, we have for any $t \geq t_0$ and $s \in [0, s_{k*}]$,

$$\langle \nabla \psi_\infty(\boldsymbol{r}_k^0(s); \overline{\rho}_t), \boldsymbol{n}(\boldsymbol{r}_k^0(s)) \rangle > 0, \tag{5.39}$$

Define the set Ω_k to be

$$\begin{aligned} \Omega_k =& \Gamma_k \cap \mathsf{B}(\mathbf{0}; k \cdot \exp\{2u_k s_{k*}\}) \cap \mathsf{B}(\mathbf{0}; 1/k \cdot \exp\{-2u_k s_{k*}\})^c \\ & \cap \{\boldsymbol{r} : \exists s \in [0, s_{k*}], s.t., r_1 = r_{k1}^\infty(s), r_2 \geq r_{k2}^\infty(s)\}^c \\ & \cap \{\boldsymbol{r} : \exists s \in [0, s_{k*}], s.t., r_1 = r_{k1}^0(s), r_2 \geq r_{k2}^0(s)\}^c. \end{aligned} \tag{5.40}$$

Consider the ODE (5.27) starting with $\boldsymbol{r}(t_0) \in \Omega_k$ for any $k \geq \{r_\infty, 1/r_0\}$, we claim $\boldsymbol{r}(t) \in \Omega_k$ for any $t \geq t_0$. Indeed, combining Eq. (5.31), (5.33), (5.39), and (5.38), for any $\boldsymbol{r} \in \partial\Omega_k \setminus ((([0,\infty) \times \{0\}) \cup \{\infty\})$ and $t \geq t_0$, the gradient $\nabla \psi_\infty(\boldsymbol{r}; \overline{\rho}_t)$ pointing outside Ω_k. Therefore, $\boldsymbol{r}(t)$ cannot leak outside Γ_k from this boundary. Further note that $\boldsymbol{r}(t)$ cannot reach the boundary $([0,\infty) \times \{0\}) \cup \{\infty\}$ for any finite time t. This proves the claim that $\boldsymbol{r}(t) \in \Omega_k$ for any $t \geq t_0$. According to Lemma 3.8, $\overline{\rho}_t(\overline{\Omega}_k) \geq \overline{\rho}_{t_0}(\overline{\Omega}_k)$ for any $k \geq \{r_\infty, 1/r_0\}$ and $t \geq t_0$.

Recall the definition of O_k given by Equation (5.32). Note that $O_k \subseteq \Omega_k$, and $\lim_{k\to\infty} \overline{\rho}_{t_0}(\overline{O}_k) = 1$, which implies $\lim_{k\to\infty} \overline{\rho}_{t_0}(\overline{\Omega}_k) = 1$. Hence, for any $\overline{\rho}_* \in \mathcal{S}_*$,

$$\overline{\rho}_*(\cup_k \overline{\Omega}_k) \geq \lim_{k\to\infty} \overline{\rho}_*(\overline{\Omega}_k) \geq \lim_{k\to\infty} \overline{\rho}_{t_0}(\overline{\Omega}_k) = 1. \tag{5.41}$$

It is easy to see that $\cup_k \overline{\Omega}_k = (0,\infty) \times [0,\infty)$. Combining with the fact that $\overline{\rho}_*((0,\infty)^2) = 0$ for any $\overline{\rho}_* \in \mathcal{S}_*$, claim (1) holds.

Proof of Claim (2). We cannot have $L(\mathcal{S}_*) \subseteq \mathcal{P}_2$.

In the case $L(\mathcal{S}_*) \subseteq \mathcal{P}_2$, the argument is similar to the proof of Claim (1), and hence will be presented in a synthetic form. First, there exists t_0 large enough, so that as $t \geq t_0$, we have $\lambda_+(\overline{\rho}_t) > 0$, and $\lambda_-(\overline{\rho}_t) < 0$. Then $\chi_{\mathrm{tg}}(\boldsymbol{r}; \overline{\rho}_t) < 0$ for any $\boldsymbol{r} \in (0,\infty)^2$. Letting

$$\Gamma_k \equiv \{\boldsymbol{r} \in [0,\infty)^2 : r_1 \leq k \cdot r_2\}, \tag{5.42}$$

According to the same argument as in the proof of Claim (1), we have $\rho_t(\Gamma_k) \geq \rho_{t_0}(\Gamma_k)$ for any $k \in (0,\infty)$ and $t \geq t_0$. As a result, we have $\mathrm{supp}(\overline{\rho}_*) \subseteq (\{0\} \times [0,\infty)) \cup \{\infty\}$.

However, the fixed point $\overline{\rho}_*$ with support on $(\{0\}\times[0,\infty))\cup\{\infty\}$ has risk $\overline{R}_\infty(\overline{\rho}_*) \geq 1$. Therefore, we cannot have $L(\mathcal{S}_*) \subseteq \mathcal{P}_2$. This proves claim (2).

□

5.4 Dynamics: Proof of Theorem 2

We will prove that the dynamics for large but finite d is well approximated by the dynamics at $d = \infty$. The key estimate is provided by the next lemma.

Lemma 5.5. *Assume σ satisfies condition $\mathsf{S0}$, recall the definition of u_d and u_∞ given by Equation (5.9) and (5.10). Assuming $k = \gamma \cdot d$ for some $\gamma \in (0,1)$, then we have*

$$\lim_{d\to\infty} \sup_{\boldsymbol{a},\boldsymbol{b}\in[0,\infty)^2} |u_d(\boldsymbol{a},\boldsymbol{b}) - u_\infty(\boldsymbol{a},\boldsymbol{b})| = 0.$$

and

$$\lim_{d\to\infty}\sup_{\boldsymbol{a},\boldsymbol{b}\in[0,\infty)^2}\|\nabla_{\boldsymbol{a}}u_d(\boldsymbol{a},\boldsymbol{b})-\nabla_{\boldsymbol{a}}u_\infty(\boldsymbol{a},\boldsymbol{b})\|_2=0.$$

Proof. We rewrite u_d here as

$$\begin{aligned}
u_d(\boldsymbol{a},\boldsymbol{b}) =&1/2\cdot[u_{d,1}(\boldsymbol{a},\boldsymbol{b})+u_{d,2}(\boldsymbol{a},\boldsymbol{b})],\\
u_{d,1}(\boldsymbol{a},\boldsymbol{b}) =&\mathbb{E}[\sigma(\tau_+a_1F_1+a_2G_1)\sigma(\tau_+b_1(F_1\cos\Theta_1+F_2\sin\Theta_1)+b_2(G_1\cos\Theta_2+G_2\sin\Theta_2))],\\
u_{d,2}(\boldsymbol{a},\boldsymbol{b}) =&\mathbb{E}[\sigma(\tau_-a_1F_1+a_2G_1)\sigma(\tau_-b_1(F_1\cos\Theta_1+F_2\sin\Theta_1)+b_2(G_1\cos\Theta_2+G_2\sin\Theta_2))],
\end{aligned}$$

where

$$\begin{aligned}
(F_1,F_2,G_1,G_2)\sim&\mathsf{N}(0,\boldsymbol{I}_4),&(5.43)\\
\Theta_1\sim&(1/Z_{s_0})\sin(\theta)^{s_0-2}\mathbf{1}\{\theta\in[0,\pi]\}\mathrm{d}\theta,&(5.44)\\
\Theta_2\sim&(1/Z_{d-s_0})\sin(\theta)^{d-s_0-2}\mathbf{1}\{\theta\in[0,\pi]\}\mathrm{d}\theta,&(5.45)
\end{aligned}$$

are mutually independent.

Define $F_3=F_1\cos\Theta_1+F_2\sin\Theta_1$, $G_3=G_1\cos\Theta_2+G_2\sin\Theta_2$, then

$$\begin{aligned}
&|u_{d,1}(\boldsymbol{a},\boldsymbol{b})-u_{\infty,1}(\boldsymbol{a},\boldsymbol{b})|\\
=&|\mathbb{E}\{\sigma(\tau_+a_1F_1+a_2G_1)[\sigma(\tau_+b_1F_3+b_2G_3)-\sigma(\tau_+b_1F_2+b_2G_2)]\}|\\
\le&\|\sigma\|_\infty\cdot\mathbb{E}\{|\sigma(\tau_+b_1F_3+b_2G_3)-\sigma(\tau_+b_1F_2+b_2G_2)|\},
\end{aligned}\tag{5.46}$$

and

$$\begin{aligned}
&|\partial_{a_1}u_{d,1}(\boldsymbol{a},\boldsymbol{b})-\partial_{a_1}u_{\infty,1}(\boldsymbol{a},\boldsymbol{b})|\\
=&|\mathbb{E}\{\tau_+F_1\cdot\sigma'(\tau_+a_1F_1+a_2G_1)[\sigma(\tau_+b_1F_3+b_2G_3)-\sigma(\tau_+b_1F_2+b_2G_2)]\}|\\
\le&\tau_+\|\sigma'\|_\infty\mathbb{E}[F_1^2]^{1/2}\mathbb{E}\{[\sigma(\tau_+b_1F_3+b_2G_3)-\sigma(\tau_+b_1F_2+b_2G_2)]^2\}^{1/2}\\
\le&\tau_+\|\sigma'\|_\infty(2\|\sigma\|_\infty^{1/2})\cdot\mathbb{E}\{|\sigma(\tau_+b_1F_3+b_2G_3)-\sigma(\tau_+b_1F_2+b_2G_2)|\}^{1/2}.
\end{aligned}\tag{5.47}$$

We have similar bounds for $|\partial_{a_2}u_{d,1}(\boldsymbol{a},\boldsymbol{b})-\partial_{a_2}u_{\infty,1}(\boldsymbol{a},\boldsymbol{b})|$.

According to condition $\mathsf{S0}$, $\|\sigma'\|_\infty$ and $\|\sigma\|_\infty$ are bounded, it is sufficient to bound the following quantity uniformly for $r\in[0,\infty)$ and $\boldsymbol{a}\in\mathbb{S}^1$,

$$T(r,\boldsymbol{a})\equiv1/2\cdot\mathbb{E}\{|\sigma(rH_2)-\sigma(rH_3)|\}=\mathbb{E}\{[\sigma(rH_2)-\sigma(rH_3)]\,\mathbf{1}_{H_2>H_3}\},\tag{5.48}$$

where

$$\begin{aligned}
H_2=H_2(\boldsymbol{a})=&[\tau_+a_1F_2+a_2G_2]/[\tau_+^2a_1^2+a_2^2]^{1/2},&(5.49)\\
H_3=H_3(\boldsymbol{a})=&[\tau_+a_1F_3+a_2G_3]/[\tau_+^2a_1^2+a_2^2]^{1/2}.&(5.50)
\end{aligned}$$

We denote $\Theta_3=\Theta_3(\boldsymbol{a})=\arcsin\{\mathbb{E}[H_2H_3|\Theta_1,\Theta_2]\}$. It is easy to see that $H_2,H_3\sim\mathsf{N}(0,1)$ with

$$\sin(\Theta_3)=\mathbb{E}[H_2H_3|\Theta_1,\Theta_2]=[\tau_+^2a_1^2\sin\Theta_1+a_2^2\sin\Theta_2]/[\tau_+^2a_1^2+a_2^2].\tag{5.51}$$

Using the same argument as in the proof of Theorem 4.6, we have for any $z\in\mathbb{R}$,

$$\mathbb{P}(H_3\le z,H_2\ge z)\le\mathbb{P}(H_3\le0,H_2\ge0)=\mathbb{E}[|\pi/2-\Theta_3|/(2\pi)].\tag{5.52}$$

Hence, we have

$$\begin{aligned}T(r,\boldsymbol{a}) =&\mathbb{E}\left\{\int_{\mathbb{R}}\sigma'(t)\,\mathbf{1}_{rH_2\geq t\geq rH_3}\,\mathrm{d}t\right\}=\int_{\mathbb{R}}\sigma'(t)\,\mathbb{P}\{H_2\geq t/r\geq H_3\}\,\mathrm{d}t\\ \leq&\sup_{z\in\mathbb{R}}\mathbb{P}(H_3\leq z,H_2\geq z)\int_{\mathbb{R}}\sigma'(t)\,\mathrm{d}t\leq 2\|\sigma\|_\infty\cdot\mathbb{E}[|\pi/2-\Theta_3|/(2\pi)]\,.\end{aligned}$$

Note that $\cos(\Theta_1)\overset{\mathrm{d}}{=}Y_1/\|\boldsymbol{Y}\|_2$, for $\boldsymbol{Y}\sim\mathsf{N}(0,\boldsymbol{I}_{s_0})$, and $\cos(\Theta_2)\overset{\mathrm{d}}{=}Z_1/\|\boldsymbol{Z}\|_2$, for $\boldsymbol{Z}\sim\mathsf{N}(0,\boldsymbol{I}_{d-s_0})$. Hence, there exists a universal constant K, such that $\mathbb{E}\{|\Theta_1-\pi/2|\}\leq K/\sqrt{s_0}$, $\mathbb{E}\{|\Theta_2-\pi/2|\}\leq K/\sqrt{d-s_0}$.

Note the relationship of $\Theta_3=\Theta_3(\boldsymbol{a})$ with (Θ_1,Θ_2) is given by Eq. (5.51), which yields

$$\sin(\Theta_3(\boldsymbol{a}))\geq\min\{\sin\Theta_1,\sin\Theta_2\},\tag{5.53}$$

hence

$$|\pi/2-\Theta_3(\boldsymbol{a})|\leq\max\{|\pi/2-\Theta_1|,|\pi/2-\Theta_2|\}.\tag{5.54}$$

As a result,

$$\sup_{\boldsymbol{a}\in\mathbb{S}^1}\mathbb{E}\{|\Theta_3(\boldsymbol{a})-\pi/2|\}\leq K\cdot\max\{1/\sqrt{s_0},1/\sqrt{d-s_0}\}.\tag{5.55}$$

We therefore obtain

$$\sup_{r\in\mathbb{R},\boldsymbol{a}\in\mathbb{S}^1}|T(r,\boldsymbol{a})|\leq K/\pi\cdot\|\sigma\|_\infty\cdot\max\{1/\sqrt{s_0},1/\sqrt{d-s_0}\}.\tag{5.56}$$

The lemma holds by noting that as $d\to\infty$, we have $s_0\to\infty$ and $d-s_0\to\infty$. □

Proof of Theorem 2. Recall the definition of $\overline{R}_\infty$ given by Eq. (5.11), and R given by Eq. (2.2). Recall the set of good initialization given by

$$\mathscr{P}_{\mathrm{good}}=\{\overline{\rho}_0\in\mathscr{P}((0,\infty)):\lim_{d\to\infty}R(\overline{\rho}\times\mathrm{Unif}(\mathbb{S}^{d-1}))<1\}.$$

Define $\mathscr{P}^1_{\mathrm{good}}$ and $\mathscr{P}^2_{\mathrm{good}}$ to be

$$\mathscr{P}^1_{\mathrm{good}}=\{\overline{\rho}^1_0\in\mathscr{P}((0,\infty)):\overline{R}_\infty(\overline{\rho}^2_0)<1,\ \text{where }\overline{\rho}^2_0\sim(\gamma^{1/2}u,(1-\gamma)^{1/2}u)\text{ with }u\sim\overline{\rho}^1_0\},\tag{5.57}$$

$$\mathscr{P}^2_{\mathrm{good}}=\{\overline{\rho}^2_0\in\mathscr{P}((0,\infty)^2):\overline{R}_\infty(\overline{\rho}^2_0)<1\}.\tag{5.58}$$

With this definition, it is easy to see that $\mathscr{P}^1_{\mathrm{good}}=\mathscr{P}_{\mathrm{good}}$.

For any $\overline{\rho}^1_0\in\mathscr{P}^1_{\mathrm{good}}$, let $u\sim\overline{\rho}^1_0$, $Y_1\sim\chi^2(\gamma\cdot d)$, and $Y_2\sim\chi^2((1-\gamma)\cdot d)$ be independent. We take $u_{d1}=u\cdot[Y_1/(Y_1+Y_2)]^{1/2}$, $u_{d2}=u\cdot[Y_2/(Y_1+Y_2)]^{1/2}$, $\boldsymbol{u}_d=(u_{d1},u_{d2})$, $u_{\infty1}=u\cdot[s_0/d]^{1/2}=u\cdot\gamma^{1/2}$, $u_{\infty2}=u\cdot[(d-s_0)/d]^{1/2}=u\cdot(1-\gamma)^{1/2}$, and $\boldsymbol{u}_\infty=(u_{\infty1},u_{\infty2})$. Denote $\overline{\rho}^{2,d}_0$ to be the distribution of $\boldsymbol{u}_d$, and $\overline{\rho}^{2,\infty}_0$ to be the distribution of $\boldsymbol{u}_\infty$. Then we have $\overline{\rho}^{2,\infty}_0\in\mathscr{P}^2_{\mathrm{good}}$. Further, if we sample $(r,\boldsymbol{n})\sim\overline{\rho}^1_0\times\mathrm{Unif}(\mathbb{S}^{d-1})$ and $(\boldsymbol{r},\boldsymbol{n}_1,\boldsymbol{n}_2)\sim\overline{\rho}^{2,d}_0\times\mathrm{Unif}(\mathbb{S}^{k-1})\times\mathrm{Unif}(\mathbb{S}^{d-k-1})$, then $r\boldsymbol{n}\overset{\mathrm{d}}{=}(r_1\boldsymbol{n}_1,r_2\boldsymbol{n}_2)$.

Here we bound $d_{\mathrm{BL}}(\overline{\rho}^{2,d}_0,\overline{\rho}^{2,\infty}_0)$. Note the joint distribution of $\boldsymbol{u}_d$ and $\boldsymbol{u}_\infty$ is a coupling of $\overline{\rho}^{2,d}_0$ and $\overline{\rho}^{2,\infty}_0$, hence

$$\begin{aligned}d_{\mathrm{BL}}(\overline{\rho}^{2,d}_0,\overline{\rho}^{2,\infty}_0)\leq&\mathbb{E}[\|\boldsymbol{u}_d-\boldsymbol{u}_\infty\|_2\wedge1]\\ =&\mathbb{E}[\{u[((Y_1/(Y_1+Y_2))^{1/2}-\gamma^{1/2})^2+((Y_2/(Y_1+Y_2))^{1/2}-(1-\gamma)^{1/2})^2]^{1/2}\}\wedge1].\end{aligned}\tag{5.59}$$

It is easy to see that $\lim_{d\to\infty} Y_1/(Y_1+Y_2) = \gamma$ almost surely. Bounded convergence theorem implies that $\lim_{d\to\infty} d_{\mathrm{BL}}(\overline{\rho}_0^{2,d}, \overline{\rho}_0^{2,\infty}) = 0$.

Now we consider the PDE (5.16) for $d = \infty$. We fix its initialization $\overline{\rho}_0^{2,\infty} \in \mathscr{P}^2_{\mathrm{good}}$ induced by $\overline{\rho}_0^1 \in \mathscr{P}^1_{\mathrm{good}}$. Denote the solution of PDE (5.16) to be $(\overline{\rho}_t^\infty)_{t\geq 0}$. Due to Theorem 5.4, for any $\eta > 0$, there exists $T = T(\eta, \overline{\rho}_0^1, \gamma, \Delta) > 0$, so that its solution $(\overline{\rho}_t^\infty)_{t\geq 0}$ satisfies

$$\overline{R}_\infty(\overline{\rho}_t^\infty) \leq \inf_{\overline{\rho}\in\mathscr{P}(E_2)} \overline{R}_\infty(\overline{\rho}) + \eta/5$$

for any $t \geq T$.

Then we consider the general PDE

$$\partial_t \rho_t(\boldsymbol{\theta}) = 2\xi(t)\nabla\cdot\left[\rho_t(\boldsymbol{\theta})\nabla\Psi(\boldsymbol{\theta};\rho_t)\right], \tag{5.60}$$

with initialization ρ_0 the distribution of $r\boldsymbol{n}$, where $(r, \boldsymbol{n}) \sim \overline{\rho}_0^1 \times \mathrm{Unif}(\mathbb{S}^{d-1})$. Due to Lemma 4.7 and Remark 3.1, we have the existence and uniqueness of the solution of PDE (5.60). We denote its solution to be $(\rho_t)_{t\geq 0}$. Let $\overline{\rho}_t^d$ be the distribution of $(\|\boldsymbol{w}_1\|_2, \|\boldsymbol{w}_2\|_2)$ with $\boldsymbol{w} = (\boldsymbol{w}_1, \boldsymbol{w}_2) \sim \rho_t$, $\boldsymbol{w}_1 \in \mathbb{R}^{s_0}$ and $\boldsymbol{w}_2 \in \mathbb{R}^{d-s_0}$. It is easy to see that $(\overline{\rho}_t^d)_{t\geq 0}$ is the unique solution of (5.16) with initialization $\overline{\rho}_0^{2,d}$.

Now, we would like to bound the distance of $\overline{\rho}_t^d$ and $\overline{\rho}_t^\infty$ using Lemma 3.7. We take $D = 2$, $V = v$, $U = u_d$, $\tilde{V} = v$, $\tilde{U} = u_\infty$ in Lemma 3.7. Let $\varepsilon_0(d)$ be as defined in Eq. (3.69). Due to Lemma 5.5, we have $\lim_{d\to\infty} \varepsilon_0(d) = 0$. We also showed that $\lim_{d\to\infty} d_{\mathrm{BL}}(\overline{\rho}_0^{2,d}, \overline{\rho}_0^{2,\infty}) = 0$. Therefore, according to Lemma 3.7, we have $\lim_{d\to\infty} \sup_{t\leq 10T} d_{\mathrm{BL}}(\overline{\rho}_t^{2,d}, \overline{\rho}_t^{2,\infty}) = 0$. Further note $\overline{R}_\infty$ is uniformly continuous with respect to $\overline{\rho}$ in bounded-Lipschitz distance. Therefore, there exists $d_0 = d_0(\eta, \overline{\rho}_0^1, \gamma, \Delta)$ large enough, so that for $d \geq d_0$ we have

$$|\overline{R}_\infty(\overline{\rho}_t^d) - \overline{R}_\infty(\overline{\rho}_t^\infty)| \leq \eta/5.$$

for any $t \leq 10T$.

Then we would like to bound the difference of $\overline{R}_\infty(\overline{\rho})$ and $\overline{R}_d(\overline{\rho})$ for any $\overline{\rho}$. Note

$$|\overline{R}_\infty(\overline{\rho}) - \overline{R}_d(\overline{\rho})| \leq \int |u_d(\boldsymbol{a},\boldsymbol{b}) - u_\infty(\boldsymbol{a},\boldsymbol{b})|\overline{\rho}(\mathrm{d}\boldsymbol{a})\overline{\rho}(\mathrm{d}\boldsymbol{b}). \tag{5.61}$$

By Lemma 5.5, there exists $d_0 = d_0(\eta, \Delta)$ large enough, so that for $d \geq d_0$, we have

$$\sup_{\overline{\rho}\in\mathscr{P}(E_2)} |\overline{R}_\infty(\overline{\rho}) - \overline{R}_d(\overline{\rho})| \leq \eta/5. \tag{5.62}$$

Finally, let $(\boldsymbol{\theta}^k)_{k\geq 1}$ be the trajectory of SGD, with step size $s_k = \varepsilon\xi(k\varepsilon)$, and initialization $\boldsymbol{w}_i^0 \sim_{iid} \rho_0$ for $i \leq N$. We apply Theorem 3 to bound the difference of the law of trajectory of SGD and the solution of PDE (5.60). The assumptions of Theorem 3 are verified by Lemma 4.7. As a consequence, there exists constant K (which depend uniquely on the constants in assumptions A1 A2 A3), such that

$$R_N(\boldsymbol{\theta}^{\lfloor t/\varepsilon\rfloor}) - \overline{R}_d(\overline{\rho}_t^d) \leq Ke^{10KT}\cdot \mathsf{err}_{N,d}(z).$$

with probability $1 - e^{-z^2}$ for any $t \leq 10T$, where

$$\mathsf{err}_{N,d}(z) = \sqrt{1/N \vee \varepsilon}\cdot\Big[\sqrt{D + \log(N(1/\varepsilon \vee 1))} + z\Big].$$

As a consequence, for any $\delta > 0$, there exists $C_0 = C_0(\delta, \eta, \overline{\rho}_0^1, \gamma, \Delta)$, so that as $N, 1/\varepsilon \geq C_0 d$ and $\varepsilon \geq 1/N^{10}$, for $t \leq 10T$, we have

$$R_N(\boldsymbol{\theta}^{\lfloor t/\varepsilon \rfloor}) - \overline{R}_d(\overline{\rho}_t^d) \leq \eta/5$$

with probability at least $1 - \delta$.

Therefore, the trajectory $\boldsymbol{\theta}^{\lfloor t/\varepsilon \rfloor}$ of SGD as $t \in [T, 10T]$ satisfies

$$\begin{aligned} R_N(\boldsymbol{\theta}^{\lfloor t/\varepsilon \rfloor}) \leq & \overline{R}_d(\overline{\rho}_t^d) + \eta/5 \leq \overline{R}_\infty(\overline{\rho}_t^d) + 2\eta/5 \leq \overline{R}_\infty(\overline{\rho}_t^\infty) + 3\eta/5 \\ \leq & \inf_{\overline{\rho} \in \overline{\mathscr{P}}} \overline{R}_\infty(\overline{\rho}) + 4\eta/5 \leq \inf_{\overline{\rho} \in \overline{\mathscr{P}}} \overline{R}_d(\overline{\rho}) + \eta = \inf_{\rho \in \mathscr{P}(\mathbb{R}^d)} R(\rho) + \eta \\ \leq & \inf_{\boldsymbol{\theta} \in \mathbb{R}^{d \times N}} R_N(\boldsymbol{\theta}) + \eta \end{aligned}$$

with probability at least $1 - \delta$. This gives the desired result. □

6 Finite temperature

We will states the lemma regarding statics properties of the finite temperature free energy in Section 6.1, and regarding dynamics properties in Section 6.2. We will prove Proposition 3, Theorem 4, and Theorem 5 in Section 6.3. Throughout Section 6.1 and 6.2, to distinguish the dimension of parameters with the generalized differential operator, we will denote the dimension of parameters by d instead of D. This should not be confused with the dimension of feature vectors, which never appears throughout this section.

We introduce the set $\mathcal{K}$ of admissible probability densities,

$$\mathcal{K} = \Big\{ \rho : \mathbb{R}^d \to [0, +\infty) \text{ measurable } : \int_{\mathbb{R}^d} \rho(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = 1, M(\rho) < \infty \Big\}, \tag{6.1}$$

where

$$M(\rho) \equiv \int_{\mathbb{R}^d} \|\boldsymbol{\theta}\|_2^2 \rho(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}. \tag{6.2}$$

Recall

$$R(\rho) = R_\# + 2 \int_{\mathbb{R}^d} V(\boldsymbol{\theta}) \rho(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} + \int_{\mathbb{R}^d \times \mathbb{R}^d} U(\boldsymbol{\theta}, \boldsymbol{\theta}') \rho(\boldsymbol{\theta}) \rho(\boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\theta}', \tag{6.3}$$

$$R_\# = \mathbb{E}\{y^2\}, \qquad V(\boldsymbol{\theta}) = -\mathbb{E}\{y\, \sigma_*(\boldsymbol{x}; \boldsymbol{\theta})\}, \tag{6.4}$$

$$U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \mathbb{E}\{\sigma_*(\boldsymbol{x}; \boldsymbol{\theta}_1) \sigma_*(\boldsymbol{x}; \boldsymbol{\theta}_2)\}, \tag{6.5}$$

$$\Psi(\boldsymbol{\theta}; \rho) = V(\boldsymbol{\theta}) + \int_{\mathbb{R}^d} U(\boldsymbol{\theta}, \boldsymbol{\theta}')\ \rho(\boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta}'. \tag{6.6}$$

Let

$$R_\lambda(\rho) = \lambda M(\rho) + R(\rho), \tag{6.7}$$

$$\Psi_\lambda(\boldsymbol{\theta}; \rho) = \lambda/2 \cdot \|\boldsymbol{\theta}\|_2^2 + V(\boldsymbol{\theta}) + \int_{\mathbb{R}^d} U(\boldsymbol{\theta}, \boldsymbol{\theta}') \rho(\boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta}', \tag{6.8}$$

$$\mathrm{Ent}(\rho) = -\int_{\mathbb{R}^d} \rho(\boldsymbol{\theta}) \log \rho(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \tag{6.9}$$

$$F_{\beta,\lambda}(\rho) = 1/2 \cdot [\lambda M(\rho) + R(\rho)] - 1/\beta \cdot \mathrm{Ent}(\rho). \tag{6.10}$$

6.1 Statics

Lemma 6.1. *For any $\rho \in \mathcal{K}$, we have*

$$\mathrm{Ent}(\rho) \le \int_{\mathbb{R}^d} \rho(\boldsymbol{\theta}) \cdot |\min(\log \rho(\boldsymbol{\theta}), 0)| \cdot \mathrm{d}\boldsymbol{\theta} \le 1 + M(\rho)/\sigma^2 + d \cdot \log(2\pi\sigma^2) \tag{6.11}$$

for any $\sigma^2 > 0$.

Proof. Define $\Omega_0 = \{\boldsymbol{\theta} : 1/(\sqrt{2\pi}\sigma)^d \cdot \exp\{-\|\boldsymbol{\theta}\|_2^2/(2\sigma^2)\} \le \rho(\boldsymbol{\theta})^{1/2} \le 1\}$. Then we have

$$\begin{aligned}\mathrm{Ent}(\rho) =& -\int_{\mathbb{R}^d} \rho(\boldsymbol{\theta}) \log \rho(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \le \int_{\mathbb{R}^d} \rho(\boldsymbol{\theta}) \cdot |\min(\log \rho(\boldsymbol{\theta}), 0)| \cdot \mathrm{d}\boldsymbol{\theta} \\ \le& \int_{\Omega_0} \rho(\boldsymbol{\theta}) \cdot |\min(\log \rho(\boldsymbol{\theta}), 0)| \cdot \mathrm{d}\boldsymbol{\theta} + \int_{\Omega_0^c} \rho(\boldsymbol{\theta}) \cdot |\min(\log \rho(\boldsymbol{\theta}), 0)| \cdot \mathrm{d}\boldsymbol{\theta}.\end{aligned}$$

The first term is bounded by

$$\int_{\Omega_0} \rho(\boldsymbol{\theta}) \cdot |\min(\log \rho(\boldsymbol{\theta}), 0)| \cdot \mathrm{d}\boldsymbol{\theta} \le \int_{\mathbb{R}^d} \rho(\boldsymbol{\theta})[\|\boldsymbol{\theta}\|_2^2/\sigma^2 + d \cdot \log(2\pi\sigma^2)]\mathrm{d}\boldsymbol{\theta} = M(\rho)/\sigma^2 + d \cdot \log(2\pi\sigma^2).$$

Noting that $|\rho \log \rho| \le \sqrt{\rho}$ for any $\rho \in [0, 1]$, the second term is bounded by

$$\begin{aligned}&\int_{\Omega_0^c} \rho(\boldsymbol{\theta}) \cdot |\min(\log \rho(\boldsymbol{\theta}), 0)| \cdot \mathrm{d}\boldsymbol{\theta} \le \int_{\Omega_0^c} \rho(\boldsymbol{\theta})^{1/2} \mathbf{1}\{\rho(\boldsymbol{\theta}) \le 1\} \mathrm{d}\boldsymbol{\theta} \\ \le& \int_{\mathbb{R}^d} 1/(\sqrt{2\pi}\sigma)^d \cdot \exp\{-\|\boldsymbol{\theta}\|_2^2/(2\sigma^2)\} \mathrm{d}\boldsymbol{\theta} = 1.\end{aligned}$$

□

Lemma 6.2. *Assume U and V are bounded-Lipschitz. Then for any $\lambda > 0$ and $0 < \beta < \infty$, $F_{\beta,\lambda}(\rho)$ has a unique minimizer $\rho_* \in \mathcal{K}$. Moreover, we have*

$$F_{\beta,\lambda}(\rho) \ge 1/2 \cdot R(\rho) + \lambda/4 \cdot M(\rho) - 1/\beta \cdot [1 + d \cdot \log(8\pi/(\beta\lambda))]. \tag{6.12}$$

Proof. First, by Lemma 6.1, we have

$$\begin{aligned}F_{\beta,\lambda}(\rho) =& 1/2 \cdot R(\rho) + \lambda/2 \cdot M(\rho) - 1/\beta \cdot \mathrm{Ent}(\rho) \\ \ge& 1/2 \cdot R(\rho) + \lambda/2 \cdot M(\rho) - 1/\beta \cdot [1 + M(\rho)/\sigma^2 + d \cdot \log(2\pi\sigma^2)].\end{aligned}$$

Taking $\sigma^2 = 4/(\beta\lambda)$ gives Eq. (6.12) .

The argument to show the existence and uniqueness of minimizer of $F_{\beta,\lambda}$ is similar to the proof of [JKO98, Proposition 4.1], and we will just give a sketch here. Since U, V are bounded-Lipschitz, it follows that $\rho \mapsto R(\rho)$ is continuous with respect to the topology of weak convergence in $L^1(\mathbb{R}^d)$. Fatou's lemma implies that M is lower semi-continuous. [JKO98, Proposition 4.1] shows the upper semi-continuity of Ent. Hence $F_{\beta,\lambda}$ is lower semi-continuous. Note (as just shown) $F_{\beta,\lambda}$ is lower bounded, there exists a sequence $(\rho_k)_{k\ge1} \subset \mathcal{K}$ such that $\lim_{k\to\infty} F_{\beta,\lambda}(\rho_k) = \inf_{\rho\in\mathcal{K}} F_{\beta,\lambda}(\rho) > -\infty$. By the same argument as [JKO98, Proposition 4.1], we can see that $\{\int \max\{\rho_k \log \rho_k, 0)\}\mathrm{d}\boldsymbol{\theta}\}_{k\ge1}$ and $\{M(\rho_k)\}_{k\ge1}$ are uniformly upper bounded, and by de la Vallée-Poussin criterion, there exists $\rho_* \in \mathcal{K}$ such that there is a subsequence of $(\rho_k)_{k\ge1}$ converges weakly to ρ_* in $L^1(\mathbb{R}^d)$. The lower semi-continuity of $F_{\beta,\lambda}$ implies that ρ_* is the minimizer of $F_{\beta,\lambda}$. Uniqueness follows by noting that U is positive semi-definite, Ent is strongly concave, and $\langle V, \rho\rangle$ and M are linear in ρ, so that $F_{\beta,\lambda}$ is a strongly convex functional.

□

For any $\rho \in \mathcal{K}$, we call the following equation the Boltzmann fixed point condition

$$\begin{aligned} \rho(\boldsymbol{\theta}) =& 1/Z(\beta, \lambda; \rho) \, \exp\{-\beta \Psi_\lambda(\boldsymbol{\theta}; \rho)\}, \\ Z(\beta, \lambda; \rho) =& \int \exp\{-\beta \Psi_\lambda(\boldsymbol{\theta}; \rho)\} \mathrm{d}\boldsymbol{\theta}. \end{aligned} \tag{6.13}$$

Lemma 6.3. *Under the assumption of Lemma 6.2, the minimizer $\rho_* \in \mathcal{K}$ of $F_{\beta,\lambda}(\rho)$ satisfies the Boltzmann fixed point condition.*

Proof. We denote μ_0 to be the Lebesgue measure on $\mathbb{R}^d$.

First, we show that ρ is positive almost everywhere. Let $\rho_* \in \mathcal{K}$ be a minimizer of $F(\rho)$, and assume by contradiction that there exists a measurable set $\Omega_0 \subset \mathbb{R}^d$, such that $\mu_0(\Omega_0) > 0$, and $\rho_*(\Omega_0) = 0$. Without loss of generality, we assume that the support of Ω_0 is compact so that $\mu_0(\Omega_0) < \infty$, otherwise we can always consider the intersection of Ω_0 with a large ball. Define $\rho_\varepsilon = (1-\varepsilon)\rho_* + \varepsilon/\mu_0(\Omega_0) \cdot \mathbf{1}_{\Omega_0} \in \mathcal{K}$. It is easy to see that there exists $\varepsilon_0 > 0$ and $C < \infty$, such that $|R_\lambda(\rho_*) - R_\lambda(\rho_\varepsilon)| \le C \cdot \varepsilon$, and

$$\begin{aligned} \mathrm{Ent}(\rho_\varepsilon) =& (1-\varepsilon)\mathrm{Ent}(\rho_*) - (1-\varepsilon)\log(1-\varepsilon) + \varepsilon \log(\mu_0(\Omega_0)/\varepsilon) \\ \ge& \mathrm{Ent}(\rho_*) - C \cdot \varepsilon + \varepsilon \log(\mu_0(\Omega_0)/\varepsilon) \end{aligned}$$

for any $\varepsilon < \varepsilon_0$. As ε is sufficiently small, we have $F_{\beta,\lambda}(\rho_\varepsilon) < F_{\beta,\lambda}(\rho_*)$. This contradict with the fact that $\rho_* \in \mathcal{K}$ is the minimizer of $F_{\beta,\lambda}(\rho)$.

Next we show that, for all $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\Psi_\lambda(\boldsymbol{\theta}; \rho_*) + 1/\beta \cdot \log \rho_*(\boldsymbol{\theta}) \equiv \gamma(\beta, \lambda; \rho_*) \tag{6.14}$$

for some constant $\gamma(\beta, \lambda; \rho_*)$.

Let $\rho_* \in \mathcal{K}$ be the minimizer of $F_{\beta,\lambda}(\rho)$. Fix $\varepsilon_0 > 0$ and define $\Gamma_{\varepsilon_0} \equiv \{\boldsymbol{\theta} \in \mathbb{R}^d : \rho_*(\boldsymbol{\theta}) \ge \varepsilon_0\} \cap \mathsf{B}(\mathbf{0}; 1/\varepsilon_0)$, and $\mathcal{A}_{\varepsilon_0} \equiv \{v \in C^\infty(\mathbb{R}^d) : \|v\|_\infty \le 1, \mathrm{supp}(v) \subseteq \Gamma_{\varepsilon_0}, \int_{\mathbb{R}^d} v(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = 0\}$. For any $v \in \mathcal{A}_{\varepsilon_0}$, define $\rho_{\varepsilon,v} = \rho + \varepsilon v$. Note that, for $-\varepsilon_0 < \varepsilon < \varepsilon_0$, we have $\rho_{\varepsilon,v} \in \mathcal{K}$. Since ρ_* is the minimizer of $F_{\beta,\lambda}(\rho)$, we must have $\lim_{\varepsilon \to 0+}[F_{\beta,\lambda}(\rho_{\varepsilon,v}) - F_{\beta,\lambda}(\rho_*)]/\varepsilon \ge 0$. It can be easily verified that

$$\lim_{\varepsilon \to 0}[F_{\beta,\lambda}(\rho_{\varepsilon,v}) - F_{\beta,\lambda}(\rho_*)]/\varepsilon = \int_{\mathbb{R}^d} [\Psi_\lambda(\boldsymbol{\theta}; \rho_*) + 1/\beta \cdot \log \rho_*(\boldsymbol{\theta})] v(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta},$$

which implies

$$\int_{\mathbb{R}^d} [\Psi_\lambda(\boldsymbol{\theta}; \rho_*) + 1/\beta \cdot \log \rho_*(\boldsymbol{\theta})] v(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = 0 \tag{6.15}$$

for any $v \in \mathcal{A}_{\varepsilon_0}$. This implies that Eq. (6.14) holds for any $\boldsymbol{\theta} \in \Gamma_{\varepsilon_0}$. But note that $\mu_0(\mathbb{R}^d \setminus (\cup_{\varepsilon_0 > 0} \Gamma_{\varepsilon_0})) = 0$. This implies that Eq. (6.14) holds almost surely.

Note we have $\int \rho_*(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = 1$. Therefore, we must have $\gamma(\beta, \lambda; \rho_*) = -1/\beta \cdot \log Z(\beta, \lambda; \rho_*)$. This proves that ρ_* satisfies the Boltzmann fixed point condition.

□

Lemma 6.4. *Under the assumption of Lemma 6.2, the Boltzmann fixed point condition has a unique solution in $\mathcal{K}$.*

Proof. The last two lemmas already imply that the Boltzmann fixed point condition has at least one solution. Assume $\rho_1, \rho_2 \in K$ to be two such solutions. Then ρ_i is positive, and

$$\log Z(\beta, \lambda; \rho_i) = -\beta \Psi_\lambda(\boldsymbol{\theta}; \rho_i) - \log \rho_i(\boldsymbol{\theta}).$$

Therefore

$$\begin{aligned} 0 =& \int_{\mathbb{R}^d} [\log Z(\beta, \lambda; \rho_1) - \log Z(\beta, \lambda; \rho_2)] \cdot [\rho_1(\boldsymbol{\theta}) - \rho_2(\boldsymbol{\theta})] \mathrm{d}\boldsymbol{\theta} \\ =& -\beta \langle U, (\rho_1 - \rho_2)^{\otimes 2} \rangle - \int_{\mathbb{R}^d} \log(\rho_1(\boldsymbol{\theta})/\rho_2(\boldsymbol{\theta}))[\rho_1(\boldsymbol{\theta}) - \rho_2(\boldsymbol{\theta})] \mathrm{d}\boldsymbol{\theta}. \end{aligned}$$

Note the right hand side does not equal 0 unless $\rho_1 = \rho_2$.

□

Lemma 6.5. *Under the assumption of Lemma 6.2, and further assume condition* A3 *holds. Let* $\rho_*^{\beta,\lambda}$ *be the minimizer of* $F_{\beta,\lambda}(\rho)$*. Then there is a constant* K *depending on the parameter* K_3 *in condition* A3*, such that for any* $\beta \geq 1$*, we have*

$$R(\rho_*^{\beta,\lambda}) \leq \inf_{\rho \in \mathscr{P}(\mathbb{R}^d)} R_\lambda(\rho) + K(1+\lambda)[d \log(2 + 1/\lambda)]/\beta. \tag{6.16}$$

Proof. Fix a $\rho \in \mathscr{P}(\mathbb{R}^d)$. Let $g_\tau(\boldsymbol{\theta})$ be the density for $\mathsf{N}(\mathbf{0}, \tau^2 \boldsymbol{I}_d)$. Denote $\rho * g_\tau$ to be the convolution of ρ and g_τ. Now we derive the formula for $F_{\beta,\lambda}(\rho * g_\tau)$.

Let $\boldsymbol{G}, \boldsymbol{G}_1, \boldsymbol{G}_2 \sim \mathsf{N}(\mathbf{0}, \boldsymbol{I}_d)$ be independent, we have

$$\begin{aligned} R(\rho * g_\tau) =& R(\rho) + 2 \int \{\mathbb{E}[V(\boldsymbol{\theta} + \tau \boldsymbol{G})] - V(\boldsymbol{\theta})\} \rho(\mathrm{d}\boldsymbol{\theta}) \\ &+ \int \{\mathbb{E}[U(\boldsymbol{\theta}_1 + \tau \boldsymbol{G}_1, \boldsymbol{\theta}_2 + \tau \boldsymbol{G}_2)] - U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\} \rho(\mathrm{d}\boldsymbol{\theta}_1) \rho(\mathrm{d}\boldsymbol{\theta}_2). \end{aligned}$$

Using the intermediate value theorem and Cauchy-Schwarz inequality, and noting that $\nabla^2 V$ is K_3-bounded by condition A3, we have

$$\begin{aligned} & \int \{V(\boldsymbol{\theta}) - \mathbb{E}[V(\boldsymbol{\theta} + \tau \boldsymbol{G})]\} \rho(\mathrm{d}\boldsymbol{\theta}) \\ =& \tau \int \mathbb{E}[\langle \nabla V(\boldsymbol{\theta}), \boldsymbol{G} \rangle] \rho(\mathrm{d}\boldsymbol{\theta}) + \frac{\tau^2}{2} \int \mathbb{E}[\langle \nabla^2 V(\tilde{\boldsymbol{\theta}}), \boldsymbol{G}^{\otimes 2} \rangle] \rho(\mathrm{d}\boldsymbol{\theta}) \leq \frac{\tau^2}{2} K_3 d, \end{aligned}$$

We have similar bound for the U term. Therefore,

$$R(\rho * g_\tau) \leq R(\rho) + 2\tau^2 K_3 d. \tag{6.17}$$

For the term $M(\rho * g_\tau)$, we have

$$M(\rho * g_\tau) = \int \mathbb{E}[\|\boldsymbol{\theta} + \tau \boldsymbol{G}\|_2^2] \rho(\mathrm{d}\boldsymbol{\theta}) = M(\rho) + \tau^2 d. \tag{6.18}$$

Next we give a lower bound for $\mathrm{Ent}(\rho * g_\tau)$:

$$\mathrm{Ent}(\rho * g_\tau) \geq \mathrm{Ent}(g_\tau) = (d/2) \log(2\pi e \tau^2). \tag{6.19}$$

As a result, taking $\tau = 1/\beta$, we have

$$F_{\beta,\lambda}(\rho_*^{\beta,\lambda}) \leq (1/2) R_\lambda(\rho) + (2K_3 + \lambda) d/(2\beta^2) + d \cdot \log(2\pi e \beta^2)/(2\beta). \tag{6.20}$$

Combining with Eq. (6.12), we have

$$R(\rho_*^{\beta,\lambda}) \leq R_\lambda(\rho) + \frac{(2K_3 + \lambda) d}{\beta^2} + \frac{2}{\beta} + \frac{d \cdot \log(2\pi e \beta^2)}{\beta} - \frac{2d \cdot \log(\lambda \beta/(8\pi))}{\beta} \tag{6.21}$$

for any $\rho \in \mathscr{P}(\mathbb{R}^d)$. Hence, the theorem holds by taking infimum over $\rho \in \mathscr{P}(\mathbb{R}^d)$. □

6.2 Dynamics

Recall that the finite-temperature distributional dynamics reads:

$$\partial_t \rho_t(\boldsymbol{\theta}) = 2\xi(t)\nabla_{\boldsymbol{\theta}} \cdot (\nabla_{\boldsymbol{\theta}} \Psi_\lambda(\boldsymbol{\theta}; \rho_t)\rho_t(\boldsymbol{\theta})) + 2\xi(t)/\beta \cdot \Delta_{\boldsymbol{\theta}} \rho_t(\boldsymbol{\theta}). \tag{6.22}$$

We say $(\rho_t)_{t\geq 0} \subseteq \mathscr{P}(\mathbb{R}^d)$ is a weak solution of (6.22), if for any $\zeta \in C_0^\infty(\mathbb{R} \times \mathbb{R}^d)$ (the space of smooth functions, decaying to 0 at infinity), we have

$$\begin{aligned} &\int_{\mathbb{R}^d} \rho_0(\boldsymbol{\theta})\zeta_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ =& -\int_{(0,\infty)\times\mathbb{R}^d} [\partial_t \zeta_t(\boldsymbol{\theta}) - 2\xi(t)\langle \nabla_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}; \rho_t), \nabla_{\boldsymbol{\theta}} \zeta_t(\boldsymbol{\theta})\rangle + 2\xi(t)\Delta_{\boldsymbol{\theta}} \zeta_t(\boldsymbol{\theta})]\rho_t(\mathrm{d}\boldsymbol{\theta})\,\mathrm{d}t \end{aligned} \tag{6.23}$$

Notice that this notion of weak solution is equivalent to the one introduced earlier in Eq. (3.3), see for instance [San15, Proposition 4.2].

Lemma 6.6. *Assume conditions* A1, A2 *and* A3 *hold. Let initialization $\rho_0 \in \mathcal{K}$ so that $F_{\beta,\lambda}(\rho_0) < \infty$. Then, the weak solution $(\rho_t)_{t\geq 0} \subseteq \mathscr{P}(\mathbb{R}^d)$ of PDE (6.23) exists and is unique. Moreover, for any fixed t, $\rho_t \in \mathcal{K}$ is absolutely continuous with respect to the Lebesgue measure, and* $\mathrm{Ent}(\rho_t)$ *and $M(\rho_t)$ are uniformly bounded in t.*

Proof. Without loss of generality, we assume $\xi(t) \equiv 1/2$.

We use the JKO scheme of [JKO98, Theorem 5.1] to show the existence, uniqueness, and absolute continuousness of solution of PDE (6.22). Since the proof is basically the same as the proof of [JKO98, Theorem 5.1], we will skip several details.

First, we consider the following discrete scheme. Let $\overline{\rho}_0^h = \rho_0$, and define $\{\overline{\rho}_k^h\}_{k\in\mathbb{N}}$ recursively by

$$\overline{\rho}_{k+1}^h \in \arg\min_{\rho\in\mathcal{K}} \{hF(\rho) + (1/2)W_2^2(\rho, \overline{\rho}_k^h)\}, \tag{6.24}$$

where $W_2(\mu,\nu)$ is the Wasserstein distance between $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$, with definition

$$W_2^2(\mu,\nu) = \inf\Big\{ \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2 \gamma(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2) : \gamma \text{ is a coupling of } \mu,\nu \Big\}.$$

For any $\overline{\rho}_{k-1}^h$, the optimization problem (6.24) has a unique minimizer $\overline{\rho}_k^h \in \mathcal{K}$, where the proof is basically the same as Lemma 6.2, by additionally noting that $W_2^2(\rho, \overline{\rho}_{k-1}^h)$ as a function of ρ is lower bounded, lower semi-continuous, and convex over $\rho \in \mathcal{K}$.

Hence, we have a sequence of probability densities $(\overline{\rho}_k^h)_{k\geq 0}$ with each $\overline{\rho}_k^h \in \mathcal{K}$. Now we define its interpolation $\rho^h : (0,\infty)\times\mathbb{R}^d \to [0,\infty)$ by

$$\rho^h(t,\,\cdot\,) = \overline{\rho}_k^h \quad \text{for} \quad t \in [kh, (k+1)h) \quad \text{and} \quad k \in \mathbb{N}.$$

In the following, we will show that this ρ^h approximately satisfies PDE (6.23) in the weak form.

Let $\boldsymbol{\xi} \in C_0^\infty(\mathbb{R}^d, \mathbb{R}^d)$ be a smooth vector field with bounded support, and define the corresponding flux $\{\Phi_\tau\}_{\tau\in\mathbb{R}}$ by

$$\partial_\tau \Phi_\tau = \boldsymbol{\xi} \circ \Phi_\tau \text{ for all } \tau \in \mathbb{R} \quad \text{ and } \quad \Phi_0 = \mathrm{id}. \tag{6.25}$$

For any $\tau \in \mathbb{R}$, let the measure ν_τ to be the push forward of $\overline{\rho}_k^h$ under Φ_τ. This means that

$$\int_{\mathbb{R}^d} \nu_\tau(\boldsymbol{\theta})\zeta(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \int_{\mathbb{R}^d} \overline{\rho}_k^h(\boldsymbol{\theta})\zeta(\Phi_\tau(\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta}, \quad \text{for all} \quad \zeta \in C(\mathbb{R}^d). \tag{6.26}$$

Since $\overline{\rho}_k^h$ is the minimizer of optimization problem (6.24), we have for each $\tau > 0$,

$$\Big(\frac{1}{2}W_2^2(\overline{\rho}_{k-1}^h, \nu_\tau) + hF(\nu_\tau)\Big) - \Big(\frac{1}{2}W_2^2(\overline{\rho}_{k-1}^h, \overline{\rho}_k^h) + hF(\overline{\rho}_k^h)\Big) \geq 0. \tag{6.27}$$

Using the result in the proof of [JKO98, Theorem 5.1], and noting ∇V is bounded Lipschitz, we have

$$\frac{\mathrm{d}}{\mathrm{d}\tau}[\langle V, \nu_\tau\rangle]_{\tau=0} = \int_{\mathbb{R}^d} \langle \nabla V(\boldsymbol{\theta}), \boldsymbol{\xi}(\boldsymbol{\theta})\rangle\, \overline{\rho}_k^h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{6.28}$$

$$\frac{\mathrm{d}}{\mathrm{d}\tau}[\mathrm{Ent}(\nu_\tau)]_{\tau=0} = \int_{\mathbb{R}^d} \overline{\rho}_k^h(\boldsymbol{\theta}) \cdot \mathrm{div}(\boldsymbol{\xi}(\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta}, \tag{6.29}$$

$$\limsup_{\tau\to 0+} \frac{1}{\tau}[M(\nu_\tau) - M(\overline{\rho}_k^h)] \leq \int_{\mathbb{R}^d} 2\langle \boldsymbol{\theta}, \boldsymbol{\xi}(\boldsymbol{\theta})\rangle\, \overline{\rho}_k^h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{6.30}$$

$$\limsup_{\tau\to 0+} \frac{1}{\tau}[W_2^2(\overline{\rho}_{k-1}^h, \nu_\tau) - W_2^2(\overline{\rho}_{k-1}^h, \overline{\rho}_k^h)] \leq \int_{\mathbb{R}^d} 2\langle (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_1)\rangle\, p(\mathrm{d}\boldsymbol{\theta}_1, \mathrm{d}\boldsymbol{\theta}_2), \tag{6.31}$$

where p is an optimal coupling of ρ_k^h and ρ_{k-1}^h in Wasserstein metric. Further we have for any $\zeta \in C_0^\infty(\mathbb{R}^d)$,

$$\Big|\int_{\mathbb{R}^d} (\overline{\rho}_k^h - \overline{\rho}_{k-1}^h)\zeta \mathrm{d}\boldsymbol{\theta} - \int_{\mathbb{R}\times\mathbb{R}} \langle \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \nabla\zeta(\boldsymbol{\theta}_1)\rangle \mathrm{d}p\Big| \leq \frac{1}{2}\sup_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\nabla^2\zeta(\boldsymbol{\theta})\|_{\mathrm{op}} W_2^2(\overline{\rho}_k^h, \overline{\rho}_{k-1}^h)\,. \tag{6.32}$$

We need to further calculate the derivative of $\langle U, \nu_\tau^{\otimes 2}\rangle$ with respect to τ. Note U is symmetric, we have

$$\begin{aligned}
&\frac{1}{\tau}[\langle U, \nu_\tau^{\otimes 2}\rangle - \langle U, (\overline{\rho}_k^h)^{\otimes 2}\rangle] - 2\int_{\mathbb{R}^d\times\mathbb{R}^d} \langle \nabla_{\boldsymbol{\theta}_1} U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_1)\rangle \overline{\rho}_k^h(\boldsymbol{\theta}_1)\overline{\rho}_k^h(\boldsymbol{\theta}_2)\mathrm{d}\boldsymbol{\theta}_1\mathrm{d}\boldsymbol{\theta}_2 \\
=& \int_{\mathbb{R}^d\times\mathbb{R}^d} \{\frac{1}{\tau}[U(\Phi_\tau(\boldsymbol{\theta}_1), \Phi_\tau(\boldsymbol{\theta}_2)) - U(\Phi_\tau(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2)] - \langle \nabla_{\boldsymbol{\theta}_2} U(\Phi_\tau(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_2)\rangle\} \overline{\rho}_k^h(\boldsymbol{\theta}_1)\overline{\rho}_k^h(\boldsymbol{\theta}_2)\mathrm{d}\boldsymbol{\theta}_1\mathrm{d}\boldsymbol{\theta}_2 \\
&+ \int_{\mathbb{R}^d\times\mathbb{R}^d} \{\frac{1}{\tau}[U(\Phi_\tau(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2) - U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)] - \langle \nabla_{\boldsymbol{\theta}_1} U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_1)\rangle\} \overline{\rho}_k^h(\boldsymbol{\theta}_1)\overline{\rho}_k^h(\boldsymbol{\theta}_2)\mathrm{d}\boldsymbol{\theta}_1\mathrm{d}\boldsymbol{\theta}_2 \\
&+ \int_{\mathbb{R}^d\times\mathbb{R}^d} [\langle \nabla_{\boldsymbol{\theta}_2} U(\Phi_\tau(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_2)\rangle - \langle \nabla_{\boldsymbol{\theta}_2} U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_2)\rangle] \overline{\rho}_k^h(\boldsymbol{\theta}_1)\overline{\rho}_k^h(\boldsymbol{\theta}_2)\mathrm{d}\boldsymbol{\theta}_1\mathrm{d}\boldsymbol{\theta}_2.
\end{aligned}$$

According to condition A3, $\nabla_{\boldsymbol{\theta}_1} U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is Lipschitz in $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, and note $\boldsymbol{\xi}(\boldsymbol{\theta}) \in C_0^\infty(\mathbb{R}^d)$ is uniformly bounded, hence $1/\tau\cdot[U(\Phi_\tau(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2) - U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)] - \langle \nabla_{\boldsymbol{\theta}_1} U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_1)\rangle$, $1/\tau[U(\Phi_\tau(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2) - U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)] - \langle \nabla_{\boldsymbol{\theta}_1} U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_1)\rangle$, and $[\langle \nabla_{\boldsymbol{\theta}_2} U(\Phi_\tau(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_2)\rangle - \langle \nabla_{\boldsymbol{\theta}_2} U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_2)\rangle]$ converges to 0 for $\tau \to 0+$, uniformly over $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathbb{R}^d \times \mathbb{R}^d$. Therefore, we have

$$\frac{\mathrm{d}}{\mathrm{d}\tau}[\langle U, \nu_\tau^{\otimes 2}\rangle]_{\tau=0} = 2\int_{\mathbb{R}^d\times\mathbb{R}^d} \langle \nabla_{\boldsymbol{\theta}_1} U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \boldsymbol{\xi}(\boldsymbol{\theta}_1)\rangle \cdot \overline{\rho}_k^h(\boldsymbol{\theta}_1)\overline{\rho}_k^h(\boldsymbol{\theta}_2)\mathrm{d}\boldsymbol{\theta}_1\mathrm{d}\boldsymbol{\theta}_2. \tag{6.33}$$

Combining Eq. (6.28) to (6.33), choosing $\boldsymbol{\xi} = \nabla\zeta$ and $\boldsymbol{\xi} = -\nabla\zeta$, we have for any $\zeta \in C_0^\infty(\mathbb{R})$,

$$\Big|\int_{\mathbb{R}^d} \Big\{\frac{1}{h}(\overline{\rho}_k^h - \overline{\rho}_{k-1}^h)\zeta + (\langle \nabla_{\boldsymbol{\theta}}\Psi_\lambda(\boldsymbol{\theta}; \overline{\rho}_k^h), \nabla\zeta\rangle - \Delta\zeta)\overline{\rho}_k^h\Big\}\mathrm{d}\boldsymbol{\theta}\Big| \leq \frac{1}{2}\sup_{\mathbb{R}^d} \|\nabla^2\zeta\|_{\mathrm{op}} \cdot \frac{1}{h}W_2^2(\overline{\rho}_{k-1}^h, \overline{\rho}_k^h). \tag{6.34}$$

According to the estimates in [JKO98, Theorem 5.1], for any $T < \infty$, there exists a constant $C < \infty$ such that for all $N \in \mathbb{N}$ and all $h \in (0, 1]$ with $Nh \le T$, there holds

$$\max\Big\{M(\overline{\rho}_N^h), \quad \int_{\mathbb{R}^d} \max\{\overline{\rho}_N^h \log(\overline{\rho}_N^h), 0\}\mathrm{d}\boldsymbol{\theta}, \quad R(\overline{\rho}_N^h), \quad \frac{1}{h}\sum_{k=1}^{N} W_2^2(\overline{\rho}_k^h, \overline{\rho}_{k-1}^h)\Big\} \le C. \tag{6.35}$$

As in [JKO98, Theorem 5.1], by de la Vallée-Poussin criterion, the second condition in Eq. (6.35) implies that there exists a measurable function $(t, \boldsymbol{\theta}) \mapsto \rho(t, \boldsymbol{\theta})$ and a sequence $(h_s)_{s\ge1}$ with $\lim_{s\to\infty} h_s = 0$, such that $(t, \boldsymbol{\theta}) \mapsto \rho^{h_s}(t, \boldsymbol{\theta})$ converges to ρ weakly in $L^1((0, T) \times \mathbb{R}^d)$ for all $T < \infty$. Eq. (6.35) also guarantees that $\rho(t, \cdot) \in \mathcal{K}$ for almost every $t \in (0, \infty)$, and $M(\rho), R(\rho) \in L^\infty((0, T))$ for all $T < \infty$. By Eq. (6.34) and (6.35), we have that ρ satisfies Eq. (6.23). Since this equation is not affected by changing $\rho(t, \cdot)$ for a set of values of t with measure 0, we can ensure that the $\rho(t, \cdot) \in \mathcal{K}$ for all t. Therefore, ρ is a solution of the weak form of PDE (6.23).

The uniqueness of solution of Eq. (6.23) can be proved using standard method from theory of elliptic-parabolic equations (see, for instance, [JKO98, Theorem 5.1]). In the proof of uniqueness we need the smoothness property of the solution, which is proved by Lemma 6.7.

□

Lemma 6.7. *Assume conditions* A1 - A4 *hold. Let initialization $\rho_0 \in \mathcal{K}$ with $F_{\beta,\lambda}(\rho_0) < \infty$. Denote the solution of PDE (6.22) to be $(\rho_t)_{t\ge0}$. Then $\rho_t(\boldsymbol{\theta})$ as a function of $(t, \boldsymbol{\theta})$ is in $C^{1,2}((0, \infty) \times \mathbb{R}^d)$, where $C^{1,2}((0, \infty) \times \mathbb{R}^d)$ is the function space of continuous function with continuous derivative in time, and second order continuous derivative in space.*

Before proving this lemma, we give some notations in the following.

For any open set $\Omega \subseteq \mathbb{R}^d$, and $1 \le p \le \infty$, define $L^p(\Omega)$ to be the Banach space consisting of all measurable functions on Ω with a finite norm

$$\|u\|_{L^p(\Omega)} \equiv \Big(\int_\Omega |u(\boldsymbol{\theta})|^p \mathrm{d}\boldsymbol{\theta}\Big)^{1/p}. \tag{6.36}$$

We say $u \in L^p_{\mathrm{loc}}(\Omega)$ if for any compact subset $\Omega' \subset \Omega$, we have $u \in L^p(\Omega')$. We denote $\|\cdot\|_{L^p(\mathbb{R}^d)}$ simply by $\|\cdot\|_{L^p}$.

For any nonnegative integer l and $1 \le p \le \infty$, we denote $W_p^l(\Omega)$ to be the Banach space (Sobolev space) consisting of the elements of $L^p(S)$ having generalized derivatives of all forms up to order l included, that are p'th power integrable on Ω. The norm in $W_p^l(\Omega)$ is defined by the equality

$$\|u\|^{(l)}_{L^p(\Omega)} = \sum_{j=0}^{l} \langle\langle u\rangle\rangle^{(j)}_{L^p(\Omega)}, \quad \langle\langle u\rangle\rangle^{(j)}_{L^p(\Omega)} = \sum_{|\boldsymbol{\alpha}|=j} \|D^{\boldsymbol{\alpha}}_{\boldsymbol{\theta}} u\|_{L^p(\Omega)}, \tag{6.37}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$ is a multi-index with $|\boldsymbol{\alpha}| = \sum_{i=1}^d \alpha_i$, and $D^{\boldsymbol{\alpha}}_{\boldsymbol{\theta}} u = \partial^{|\boldsymbol{\alpha}|} u / \partial\theta_1^{\alpha_1} \cdots \partial\theta_d^{\alpha_d}$.

Let $(t_1, t_2) \subseteq (0, T)$ be an open interval and $\Omega \subseteq \mathbb{R}^d$ be an open set, in these three paragraphs we temporarily denote $S = (t_1, t_2) \times \Omega$. For any $1 \le r, p \le \infty$, define $L^{r,p}(S)$ to be the Banach space consisting of all measurable functions on S with a finite norm

$$\|u\|_{L^{r,p}(S)} \equiv \Big(\int_{t_1}^{t_2} \Big(\int_\Omega |u(t, \boldsymbol{\theta})|^p \mathrm{d}\boldsymbol{\theta}\Big)^{r/p} \mathrm{d}t\Big)^{1/r}. \tag{6.38}$$

We say $u \in L^{r,p}_{\mathrm{loc}}(S)$ if for any compact subset $[t_1', t_2'] \subset (t_1, t_2)$ and compact subset $\Omega' \subset \Omega$, we have $u \in L^{r,p}([t_1', t_2'] \times \Omega')$. We will denote $L^{p,p}(S)$ by $L^p(S)$, and $L^{p,p}_{\mathrm{loc}}(S)$ by $L^p_{\mathrm{loc}}(S)$.

For nonnegative integer l and $1 \le p \le \infty$, we denote $W_p^{2l,l}(S)$ to be the Banach space consisting of the elements of $L^p(S)$ having generalized derivatives of the form $D_t^r D_{\boldsymbol{\theta}}^{\boldsymbol{\alpha}}$ with r and $\boldsymbol{\alpha}$ satisfying the inequality $2r + |\boldsymbol{\alpha}| \le 2l$. The corresponding norm is defined by

$$\|u\|_{L^p(S)}^{(2l)} = \sum_{j=0}^{2l} \langle\langle u \rangle\rangle_{L^p(S)}^{(j)}, \quad \langle\langle u \rangle\rangle_{L^p(S)}^{(j)} = \sum_{|\boldsymbol{\alpha}|+2r=j} \|D_t^r D_{\boldsymbol{\theta}}^{\boldsymbol{\alpha}} u\|_{L^p(S)}. \tag{6.39}$$

We denote $C^{m,n}(S)$ to be the function space of continuous function with m continuous derivative in time, and n continuous derivatives in space. For example, $u \in C^{1,2}(S)$ if and only if $u, \partial_t u, \nabla_{\boldsymbol{\theta}} u, \nabla_{\boldsymbol{\theta}}^2 u \in C^{0,0}(S) \equiv C(S)$. We say $u \in C_c^{m,n}(S)$ if $u \in C^{m,n}(S)$ and the support of u is compact. We will denote $C^{n,n}(S)$ by $C^n(S)$, and $C_c^{n,n}(S)$ by $C_c^n(S)$.

For any measurable functions f, g defined on $\mathbb{R}^d$, we denote $f * g$ to be their space convolution, which is a measurable function on $\mathbb{R}^d$, with

$$(f * g)(\boldsymbol{\theta}) = \int_{\mathbb{R}^d} f(\boldsymbol{\theta}') g(\boldsymbol{\theta} - \boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta}'. \tag{6.40}$$

For any measurable function u, v defined on $\mathbb{R} \times \mathbb{R}^d$, we denote $u *_2 v$ to be their space and time convolution, which is a measurable function on $\mathbb{R} \times \mathbb{R}^d$, with

$$(u *_2 v)(t, \boldsymbol{\theta}) = \int_{\mathbb{R}} \mathrm{d}t' \int_{\mathbb{R}^d} u(t', \boldsymbol{\theta}') v(t - t', \boldsymbol{\theta} - \boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta}'. \tag{6.41}$$

If u, v are defined on a subset of $\mathbb{R} \times \mathbb{R}^d$, we define $u *_2 v$ using their zero extensions.

We denote G to be the heat kernel, where for $t > 0$, we have

$$G(t, \boldsymbol{\theta}) = t^{-d/2} g(t^{-1/2}\boldsymbol{\theta}), \quad g(\boldsymbol{\theta}) = (2\pi)^{-d/2} \exp\{-1/2 \cdot \|\boldsymbol{\theta}\|_2^2\}. \tag{6.42}$$

Proof. The proof is similar to the one of [JKO98, Theorem 5.1], so we will skip some details. Without loss of generality we can set $\beta = 1$, and $\xi(t) = 1/2$ (different choices can be obtained by rescaling $\Psi(\boldsymbol{\theta}; \rho)$ and reparametrizing time).

Let $E = (0, \infty) \times \mathbb{R}^d$. With a slight abuse of notations, we denote $\Psi(t, \boldsymbol{\theta}) = \Psi_\lambda(\boldsymbol{\theta}; \rho_t)$. Since $V \in C^4(\mathbb{R}^d)$, and $\nabla_1^k U$ are uniformly bounded for $0 \le k \le 4$, we have $\nabla_{\boldsymbol{\theta}}^k \Psi \in L_{\mathrm{loc}}^\infty(E)$ for $0 \le k \le 4$.

In the following, we will write $\rho(t, \boldsymbol{\theta}) = \rho_t(\boldsymbol{\theta})$ for clarity. When we write $\rho(t)$, we regard it as a function in $L^1(\mathbb{R}^d)$ at any fixed t. For other functions, we also use this convention.

Step 1. Show that $\rho \in L_{\mathrm{loc}}^{\infty,p}(E)$.

Taking G to be the heat kernel, it is easy to see that

$$\|G(t)\|_{L^p} = t^{(\frac{1}{p}-1)\frac{d}{2}} \|g\|_{L^p}, \quad \|\nabla G(t)\|_{L^p} = t^{\frac{1}{p}\frac{d}{2} - \frac{d+1}{2}} \|\nabla g\|_{L^p}.$$

Then for any $\eta \in C_c^\infty(\mathbb{R}^d)$, Duhamel's principle gives

$$\begin{aligned} \rho(t)\eta = &\int_\varepsilon^t [\rho(s)(\Delta\eta - \langle \nabla\Psi(s), \nabla\eta \rangle)] * G(t-s) \mathrm{d}s \\ &+ \int_\varepsilon^t [\rho(s)(2\nabla\eta - \eta\nabla\Psi(s))] * \nabla G(t-s) \mathrm{d}s + (\rho(\varepsilon)\eta) * G_\varepsilon(t) \end{aligned} \tag{6.43}$$

for almost every $0 \le \varepsilon < t < \infty$, where $*$ denotes convolution in the $\boldsymbol{\theta}$-variables, and $G_\varepsilon(t, \boldsymbol{\theta}) \equiv G(t - \varepsilon, \boldsymbol{\theta})$. By Young's convolution inequality, we have $\|f * g\|_{L^r} \le C\|f\|_{L^p}\|g\|_{L_q}$ for $1/p + 1/q =$

$1/r+1$ and $p,q,r\geq 1$. For fixed t, we estimate the $L^p(\mathbb{R}^d)$ norm of $\rho(t)\eta$, which gives

$$\begin{aligned}\|\rho(t)\eta\|_{L^p}\leq&\int_\varepsilon^t\|\rho(s)(\Delta\eta-\langle\nabla\Psi(s),\nabla\eta\rangle)\|_{L^1}\|G(t-s)\|_{L^p}\mathrm{d}s\\&+\int_\varepsilon^t\|\rho(t)(2\nabla\eta-\eta\nabla\Psi(t))\|_{L^1}\|\nabla G(t-s)\|_{L^p}\mathrm{d}s+\|\rho(\varepsilon)\eta\|_{L^1}\|G(t-\varepsilon)\|_{L^p}\\\leq&\operatorname*{ess\,sup}_{s\in[\varepsilon,t]}\|\rho(s)(\Delta\eta-\langle\nabla\Psi(s),\nabla\eta\rangle)\|_{L^1}\|g\|_{L^p}\int_0^{t-\varepsilon}s^{(\frac{1}{p}-1)\frac{d}{2}}\mathrm{d}s\\&+\operatorname*{ess\,sup}_{s\in[\varepsilon,t]}\|\rho(s)(2\nabla\eta-\eta\nabla\Psi(s))\|_{L^1}\|\nabla g\|_{L^p}\int_0^{t-\varepsilon}s^{\frac{1}{p}\frac{d}{2}-\frac{d+1}{2}}\mathrm{d}s\\&+\|\rho(\varepsilon)\eta\|_{L^1}\|g\|_{L^p}(t-\varepsilon)^{(\frac{1}{p}-1)\frac{d}{2}}\end{aligned}$$

for almost every $0\leq\varepsilon<t<\infty$. For $p<d/(d-1)$, the s-integrals are finite. Therefore, we have $\rho\eta\in L^{\infty,p}((\delta,T)\times\mathbb{R}^d)$ for any δ,T such that $\varepsilon<\delta<T<\infty$. Hence we have $\rho\in L^{\infty,p}_{\mathrm{loc}}((0,\infty)\times\mathbb{R}^d)$.

Step 2. Show that $\rho\in L^\infty_{\mathrm{loc}}((0,\infty)\times\mathbb{R}^d)$ using bootstrap.

In what follows, we let $E\equiv(0,\infty)\times\mathbb{R}^d$.

We can iteratively use the strategy in step 1 to show that $\rho\in L^\infty_{\mathrm{loc}}(E)$. We will summarize our key estimates in Step 1 as follows. For any measurable function u defined on $S=(\delta,T)\times\mathbb{R}^d$ for some $0\leq\delta<T<\infty$, we have

$$\|u*_2G\|_{L^{\infty,p_o}(S)}\leq C\|u\|_{L^{\infty,p_i}(S)},\tag{6.44}$$

$$\|u*_2\nabla G\|_{L^{\infty,p_o}(S)}\leq C\|u\|_{L^{\infty,p_i}(S)},\tag{6.45}$$

provided that the p_o,p_i satisfy the relations

$$1\leq p_i\leq p_o,\quad d\cdot(1/p_i-1/p_o)<1.\tag{6.46}$$

Here, C is a constant depends only on T,δ and on p_i,p_o.

Define $\varphi_1\equiv\rho(\Delta\eta-\langle\nabla\Psi,\nabla\eta\rangle)\mathbf{1}\{t>\varepsilon\}$, $\varphi_2\equiv\rho(2\nabla\eta-\eta\nabla\Psi)\mathbf{1}\{t>\varepsilon\}$, and $\psi\equiv\rho(\varepsilon)\eta$. Then Eq. (6.43) reads

$$\rho\eta=\varphi_1*_2G+\varphi_2*_2\nabla G+\psi*G_\varepsilon.\tag{6.47}$$

Since $\psi=\rho(\varepsilon)\eta\in L^1(\mathbb{R}^d)$, the behavior of $\psi*G_\varepsilon$ on $S=(\delta,T)\times\mathbb{R}^d$ for $\varepsilon<\delta<T<\infty$ will be extremely nice: for any generalized gradient $D_t^rD^{\boldsymbol{\alpha}}[\psi*G_\varepsilon]$,

$$\|D_t^rD^{\boldsymbol{\alpha}}[\psi*G_\varepsilon]\|_{L^\infty(S)}\leq\|\psi\|_{L^1(\mathbb{R}^d)}\|D_t^rD^{\boldsymbol{\alpha}}G_\varepsilon\|_{L^\infty(S)}<\infty.\tag{6.48}$$

Hence $D_t^rD^{\boldsymbol{\alpha}}[\psi*G_\varepsilon]\in L^\infty(S)$. From now on, we fix $0<\varepsilon<\delta<T<\infty$ and take $S\equiv(\delta,T)\times\mathbb{R}^d$.

According to Eq. (6.47) we have

$$\begin{aligned}\|\rho\eta\|_{L^{\infty,p_o}(S)}\leq&\|\varphi_1*_2G\|_{L^{\infty,p_o}(S)}+\|\varphi_2*_2\nabla G\|_{L^{\infty,p_o}(S)}+\|\psi*G_\varepsilon\|_{L^{\infty,p_o}(S)}\\\leq&C\{\|\varphi_1\|_{L^{\infty,p_i}(S)}+\|\varphi_2\|_{L^{\infty,p_i}(S)}+\|\psi\|_{L^1(\mathbb{R}^d)}\}\end{aligned}\tag{6.49}$$

Now we assume $\rho\in L^{\infty,p_i}_{\mathrm{loc}}(E)$ for some p_i. Note $\nabla\Psi\in L^\infty_{\mathrm{loc}}(E)$ so that $\max\{\|\varphi_1\|_{L^{\infty,p_i}(S)},\|\varphi_2\|_{L^{\infty,p_i}(S)}\}\leq C_\eta\|\rho\|_{L^{\infty,p_i}((\delta,T)\times\Omega_2)}$, where $\Omega_2\supseteq\mathrm{supp}(\eta)$ is a compact set. As a result, for any $\eta\in C_c^\infty(\mathbb{R}^d)$, we have

$$\|\rho\|_{L^{\infty,p_o}((\delta,T)\times\Omega_1)}\leq C_\eta(\|\rho\|_{L^{\infty,p_i}((\delta,T)\times\Omega_2)}+1),\tag{6.50}$$

where $\Omega_1 \subseteq \text{supp}(\eta) \subseteq \Omega_2$. Therefore, $\rho \in L^{\infty,p_o}_{\text{loc}}(E)$, where p_i, p_o satisfy Eq. (6.46).

Note there exists a sequence $p_{i,l}, p_{o,l}$ for $1 \le l \le k$ and $k < \infty$, so that $p_{i,l+1} = p_{o,l}$, $p_{i,1} = p < d/(d-1)$, $p_{i,k} = \infty$, and $p_{i,l}, p_{o,l}$ for fixed l satisfies Eq. (6.46). Since we have $\rho \in L^{\infty,p}_{\text{loc}}(E)$, using Eq. (6.50) iteratively, we have $\rho \in L^{\infty,p_{o,l}}_{\text{loc}}(E)$ for any $1 \le l \le k$. As a result, we have $\rho \in L^{\infty}_{\text{loc}}(E)$.

Step 3. Derivatives, $D\rho$, $D^2\rho$, and $D^3\rho$.

By [LSU88, Chapter IV, section 3, (3.1)], for any function u defined on $E = (0,\infty) \times \mathbb{R}^d$, we have

$$\langle\langle G *_2 u\rangle\rangle^{(2m+2)}_{L^p(E)} \le C\langle\langle u\rangle\rangle^{(2m)}_{L^p(E)}, \tag{6.51}$$

where $1 < p \le \infty$ and m is a nonnegative integer.

First, we show the regularity of $D\rho$. Note that $\rho \in L^{\infty}_{\text{loc}}(E)$, $\eta \in C^{\infty}_c(\mathbb{R}^d)$, $\nabla\Psi \in L^{\infty}_{\text{loc}}(E)$, we have $\varphi_1, \varphi_2 \in L^{\infty}(E)$. Due to Eq. (6.51), we have $D^2\{\varphi_1 *_2 G\}, D^2\{\varphi_2 *_2 G\} \in L^{\infty}(E)$, which also implies $D\{\varphi_1 *_2 G\} \in L^{\infty}_{\text{loc}}(E)$. Hence we have $D(\rho\eta) = D\{\varphi_1 *_2 G\} + D^2\{\varphi_2 *_2 G\} + D[\psi * G_\varepsilon] \in L^{\infty}(S)$, which gives $D\rho \in L^{\infty}_{\text{loc}}(E)$.

Then we show the regularity of $D^2\rho$. Note that $\nabla^2\Psi \in L^{\infty}_{\text{loc}}(E)$, we have $D\varphi_1, D\varphi_2 \in L^{\infty}(E)$. Due to Eq. (6.51), we have $D^3\{\varphi_1 *_2 G\}, D^3\{\varphi_2 *_2 G\} \in L^{\infty}(E)$, which also implies $D^2\{\varphi_1 *_2 G\} \in L^{\infty}_{\text{loc}}(E)$. Hence we have $D^2(\rho\eta) = D^2\{\varphi_1 *_2 G\} + D^3\{\varphi_2 *_2 G\} + D^2[\psi * G_\varepsilon] \in L^{\infty}(S)$, which gives $D^2\rho \in L^{\infty}_{\text{loc}}(E)$.

Next we show the regularity of $D^3\rho$. Note that $\nabla^3\Psi \in L^{\infty}_{\text{loc}}(E)$, we have $D^2\varphi_1, D^2\varphi_2 \in L^{\infty}(E)$. Due to Eq. (6.51), we have $D^4\{\varphi_1 *_2 G\}, D^4\{\varphi_2 *_2 G\} \in L^{\infty}(E)$, which also implies $D^3\{\varphi_1 *_2 G\} \in L^{\infty}_{\text{loc}}(E)$. Hence we have $D^3(\rho\eta) = D^3\{\varphi_1 *_2 G\} + D^4\{\varphi_2 *_2 G\} + D^3[\psi * G] \in L^{\infty}(S)$, which gives $D^3\rho \in L^{\infty}_{\text{loc}}(E)$.

Step 4. Derivatives, $D_t\rho$, $D_tD\rho$, and $D_tD^2\rho$.

Now we study the regularity of $D_t\rho, D_tD\rho, D_tD^2\rho$. Note we have $D_t(\rho\eta) = D_t\{\varphi_1 *_2 G\} - D_t\{D\varphi_1 *_2 G\} + D_t\{\psi * G_\varepsilon\}$. Due to Eq. (6.51), $\varphi_1, D\varphi_2 \in L^{\infty}(E)$ implies that $D_t\{\varphi_1 *_2 G\}, D_t\{D\varphi_1 *_2 G\} \in L^{\infty}(E)$ and hence $D_t[\rho\eta] \in L^{\infty}(S)$, $D_t\rho \in L^{\infty}_{\text{loc}}(E)$.

Note we have $D_tD(\rho\eta) = D_t\{D\varphi_1 *_2 G\} + D_t\{D^2\varphi_1 *_2 G\} + D_t\{D\psi * G_\varepsilon\}$. The fact that $D\varphi_1, D^2\varphi_2 \in L^{\infty}(E)$ implies that $D_t\{D\varphi_1 *_2 G\}, D_t\{D^2\varphi_1 *_2 G\} \in L^{\infty}(E)$ and hence $D_tD\rho \in L^{\infty}_{\text{loc}}(E)$.

Note we have $D_tD^2(\rho\eta) = D_t\{D^2\varphi_1 *_2 G\} - D_t\{D^3\varphi_1 *_2 G\} + D_t\{D^2\psi * G_\varepsilon\}$. Note that $\nabla^4\Psi \in L^{\infty}_{\text{loc}}(E)$, hence $D^3\varphi_2 \in L^{\infty}(E)$. Combining with the fact that $D^2\varphi_1 \in L^{\infty}(E)$, we have $D_t\{D^2\varphi_1 *_2 G\}, D_t\{D^3\varphi_1 *_2 G\} \in L^{\infty}(E)$ and hence $D_tD^2\rho \in L^{\infty}_{\text{loc}}(E)$.

Step 5. Derivatives, $D_t^2\rho$.

Finally we show the regularity of $D_t^2\rho$. We have $D_t^2(\rho\eta) = D_t\{D_t[\varphi_1 *_2 G] - D_t[D\varphi_1 *_2 G] + D_t[\psi * G_\varepsilon]\}$, and

$$D_t[\varphi_1 *_2 G] = [\Delta\varphi_1] *_2 G + \varphi_1(\varepsilon) * G_\varepsilon, \tag{6.52}$$

$$D_t[D\varphi_2 *_2 G] = [D\Delta\varphi_2] *_2 G + [D\varphi_2(\varepsilon)] * G_\varepsilon. \tag{6.53}$$

Note that $\nabla^4\Psi \in L^{\infty}_{\text{loc}}(E)$, we have $\Delta\varphi_1, D\Delta\varphi_1 \in L^{\infty}_{\text{loc}}(E)$, and $\varphi_1(\varepsilon), D\varphi_2(\varepsilon) \in L^1(\mathbb{R}^d)$. Hence according to Eq. (6.51), we have $D_t\{[\Delta\varphi_1] *_2 G\}, D_t\{[D\Delta\varphi_2] *_2 G\}$. In addition $D_t\{\varphi_1(\varepsilon) * G_\varepsilon\}, D_t\{[D\varphi_2(\varepsilon)] * G_\varepsilon\} \in L^{\infty}(S)$. As a result, we have $D_t^2\rho \in L^{\infty}_{\text{loc}}(E)$.

Step 6. Finish the proof.

As a result, we have $\rho, D\rho, D^2\rho, D^3\rho, D_t\rho, D_tD\rho, D_tD^2\rho, D_t^2\rho \in L^{\infty}_{\text{loc}}(E)$. Sobolev embedding theorem implies that $\rho, \partial_t\rho, \nabla_{\boldsymbol{\theta}}\rho, \nabla^2_{\boldsymbol{\theta}}\rho \in C^{0,0}(\mathbb{R}^d)$. In other words, $\rho \in C^{1,2}(E)$, which is the desired result.

□

Lemma 6.8. *Assume conditions* A1 - A4 *hold. Let initialization $\rho_0 \in \mathcal{K}$ with $F_{\beta,\lambda}(\rho_0) < \infty$. Denote the solution of PDE (6.22) to be $(\rho_t)_{t\geq 0}$. Then $\rho_t(\boldsymbol{\theta}) > 0$ for any $(t, \boldsymbol{\theta}) \in (0, \infty) \times \mathbb{R}^d$.*

Proof. Note that $\rho_t \in C^{1,2}((0, \infty) \times \mathbb{R}^d)$. By the Harnack's inequality [Eva09], we immediately have $\rho_t(\boldsymbol{\theta}) > 0$ for any $(t, \boldsymbol{\theta}) \in (0, \infty) \times \mathbb{R}^d$.

□

We say ρ_* is a fixed point of PDE (6.22), if its solution $(\rho_t)_{t\geq 0}$ starting from ρ_* satisfies $\rho_t \equiv \rho_*$ for any $t \geq 0$.

Lemma 6.9. *Assume conditions* A1 - A3 *hold. Then any fixed point ρ_* of PDE (6.22) with $\rho_* \in \mathcal{K}$ must satisfy the Boltzmann fixed point condition (6.13).*

Proof. Suppose $\rho_* \in \mathcal{K}$ is a fixed point of PDE (6.22), taking $W(\boldsymbol{\theta}) \equiv \Psi_\lambda(\boldsymbol{\theta}; \rho_*)$, then $\rho_* \in \mathcal{K}$ is a fixed point of the Fokker-Planck equation (6.54).

$$\partial_t \rho_t(\boldsymbol{\theta}) = 2\xi(t)\nabla \cdot (\nabla W(\boldsymbol{\theta})\rho_t(\boldsymbol{\theta})) + 2\xi(t)/\beta \cdot \Delta_{\boldsymbol{\theta}} \rho_t(\boldsymbol{\theta}). \tag{6.54}$$

Since $\lambda/2 \cdot \|\boldsymbol{\theta}\|_2^2 - 2K_3 \leq \Psi_\lambda(\boldsymbol{\theta}; \rho_*) \leq \lambda/2 \cdot \|\boldsymbol{\theta}\|_2^2 + 2K_3$, the Fokker-Planck equation has a unique fixed point [MV00], which solves

$$\rho_*(\boldsymbol{\theta}) = \frac{1}{Z_\beta} \exp\{-\beta W(\boldsymbol{\theta})\}, \quad Z_\beta = \int_{\mathbb{R}^d} \exp\{-\beta W(\boldsymbol{\theta})\} \mathrm{d}\boldsymbol{\theta}.$$

This is exactly the Boltzmann fixed point condition.

□

Lemma 6.10. *Assume conditions* A1 - A4 *hold. Let $(\rho_t)_{t\geq 0}$ be the solution of PDE (6.22) for an initialization $\rho_0 \in \mathcal{K}$. Then the free energy $F_{\beta,\lambda}(\rho_t)$ is differentiable with respect to t, with*

$$\partial_t F_{\beta,\lambda}(\rho_t) = -2\xi(t) \int_{\mathbb{R}^d} \|\nabla_{\boldsymbol{\theta}}(\Psi_\lambda(\boldsymbol{\theta}; \rho_t) + 1/\beta \cdot \log \rho_t(\boldsymbol{\theta}))\|_2^2 \rho_t(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}. \tag{6.55}$$

Therefore, $F_{\beta,\lambda}(\rho_t)$ is non-increasing in t.

Proof. Calculate the differential of the free energy along the curve ρ_t, we have

$$\begin{aligned}
\partial_t F_{\beta,\lambda}(\rho_t) =& \int_{\mathbb{R}^d} \Psi_\lambda(\boldsymbol{\theta}; \rho_t) \partial_t \rho_t(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} + 1/\beta \cdot \int \log(\rho_t(\boldsymbol{\theta})) \partial_t \rho_t(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
=& -\xi(t) \int_{\mathbb{R}^d} \|\nabla_{\boldsymbol{\theta}}(\Psi_\lambda(\boldsymbol{\theta}; \rho_t) + 1/\beta \cdot \log \rho_t(\boldsymbol{\theta}))\|_2^2 \rho_t(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.
\end{aligned}$$

□

Lemma 6.11. *Assume $K_0\|\boldsymbol{\theta}\|_2^2 - K_1 \leq \Phi(\boldsymbol{\theta}) \leq K_0\|\boldsymbol{\theta}\|_2^2 + K_1$ for some positive constant K_0, K_1. Define*

$$\mu_*(\mathrm{d}\boldsymbol{\theta}) = \frac{1}{Z_*} \exp\{-\Phi(\boldsymbol{\theta})\} \mathrm{d}\boldsymbol{\theta}, \quad Z_* = \int_{\mathbb{R}^d} \exp\{-\Phi(\boldsymbol{\theta})\} \mathrm{d}\boldsymbol{\theta} \tag{6.56}$$

Let $\mathcal{D} \equiv \{f \in L^2(\mathbb{R}^d, \mu_*) \cap C^1(\mathbb{R}^d) : \|\nabla f\|_2 \in L^2(\mathbb{R}^d, \mu_*)\}$. For any $f \in \mathcal{D}$, define

$$I(f) \equiv \int_{\mathbb{R}^d} \|\nabla f(\boldsymbol{\theta})\|_2^2 \cdot \mu_*(\mathrm{d}\boldsymbol{\theta}) < \infty. \tag{6.57}$$

Assume $(f_n)_{n\geq 1} \subseteq \mathcal{D}$, with $\lim_{n\to\infty} I(f_n) = 0$, and f_n converges weakly to f_* in $L^2(\mathbb{R}^d, \mu_*)$. Then $f_*(\boldsymbol{\theta}) \equiv F_*$ for some constant F_*.

Proof. First we show that the measure μ_* satisfies the Poincare inequality: for any $f \in \mathcal{D}$,

$$\mu_*((f - \mu_*(f))^2) \leq K \cdot I(f), \tag{6.58}$$

for some constant K.

Let μ be the Gaussian distribution $\mathsf{N}(\mathbf{0}, 1/(2K_0) \cdot \boldsymbol{I}_d)$. Then for any $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\mu(\boldsymbol{\theta}) \cdot \exp\{-2K_1\} \leq \mu_*(\boldsymbol{\theta}) \leq \mu(\boldsymbol{\theta}) \cdot \exp\{2K_1\}. \tag{6.59}$$

Therefore, for any nonnegative measurable function $f : \mathbb{R}^d \to [0, \infty)$ and $g : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$, letting $(G, G') \sim \mu \times \mu$ and $(X, X') \sim \mu_* \times \mu_*$, we have

$$\begin{aligned}
\mathbb{E}[f(G)] \cdot \exp\{-2K_1\} \leq & \mathbb{E}[f(X)] \leq \mathbb{E}[f(G)] \cdot \exp\{2K_1\}, \\
\mathbb{E}[g(G, G')] \cdot \exp\{-4K_1\} \leq & \mathbb{E}[g(X, X')] \leq \mathbb{E}[g(G, G')] \cdot \exp\{4K_1\}.
\end{aligned}$$

Note we have the Poincare inequality for the Gaussian distribution μ,

$$\mathrm{Var}[f(G)] \leq 1/(2K_0) \cdot \mathbb{E}[\|\nabla f(G)\|_2^2] \tag{6.60}$$

for any differentiable f. Therefore, we have

$$\begin{aligned}
\mathrm{Var}[f(X)] =& \frac{1}{2}\mathbb{E}[(f(X) - f(X'))^2] \leq \frac{1}{2}\exp\{4K_1\} \cdot \mathbb{E}[(f(G) - f(G'))^2] \\
=& \exp\{4K_1\} \cdot \mathrm{Var}[f(G)] \leq 1/(2K_0) \cdot \exp\{4K_1\} \cdot \mathbb{E}[\|\nabla f(G)\|_2^2] \\
\leq& 1/(2K_0) \cdot \exp\{6K_1\} \cdot \mathbb{E}[\|\nabla f(X)\|_2^2].
\end{aligned}$$

This proves the Poincare inequality (6.58) for μ_*.

Since $\lim_{n\to\infty} I(f_n) = 0$, due to (6.58), we immediately have $f_n - \mu_*(f_n)$ converges to 0 in $L^2(\mathbb{R}^d, \mu_*)$. Note we assumed f_n converges weakly to f_* in $L^2(\mathbb{R}^d, \mu_*)$, and $1 \in L^2(\mathbb{R}^d, \mu_*)$, we have

$$\lim_{n\to\infty} \mu_*(f_n) = \mu_*(f).$$

Therefore, $f_n - \mu_*(f_n)$ converges weakly to $f_* - \mu_*(f_*)$ in $L^2(\mathbb{R}^d, \mu_*)$. Hence $f_*(\boldsymbol{\theta}) \equiv \mu_*(f_*)$.

□

Lemma 6.12. *Assume conditions* A1 - A4 *hold. Then the solution $(\rho_t)_{t\geq 0}$ of PDE (6.22) for any initialization $\rho_0 \in \mathcal{K}$ converges weakly to $\rho_* \in \mathcal{K}$ as $t \to \infty$, where ρ_* is the unique solution of the Boltzmann fixed point condition, which is the global minimizer of $F_{\beta,\lambda}$.*

Proof. According to Lemma 6.10, $F_{\beta,\lambda}$ is non-increasing along the solution path. According to Lemma 6.2, $F_{\beta,\lambda}(\rho_t)$ is lower bounded. Therefore, we have

$$\lim_{t\to\infty} \int_{\mathbb{R}^d} \|\nabla_{\boldsymbol{\theta}}(\Psi_\lambda(\boldsymbol{\theta}; \rho_t) + 1/\beta \cdot \log \rho_t(\boldsymbol{\theta}))\|_2^2 \rho_t(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = 0. \tag{6.61}$$

Since $M(\rho_t)$ is uniformly bounded, by Lemma 6.6, $(\rho_t)_{t\geq 0}$ as a sequence of probability distribution in $\mathscr{P}(\mathbb{R}^d)$ is uniformly tight. Hence there exists $\rho_* \in \mathscr{P}(\mathbb{R}^d)$ and a subsequence $(\rho_{t_k})_{k\geq 1}$ with $\lim_{k\to\infty} t_k = \infty$ such that $(\rho_{t_k})_{k\geq 1}$ converges weakly to ρ_*. By Lemma 6.6 and Lemma 6.1, $\{\int \max\{\rho_{t_k} \log \rho_{t_k}, 0)\} \mathrm{d}\boldsymbol{\theta}\}_{k\geq 1}$ is uniformly bounded. Using de la Vallée-Poussin's criteria, we can show that $(\rho_{t_k})_{k\geq 1}$ is uniformly integrable, and hence ρ_* is absolute continuous with respect to Lebesgue measure, which means ρ_* has a density.

Note we have

$$\nabla_{\boldsymbol{\theta}}\Psi_\lambda(\boldsymbol{\theta};\rho_t)-\nabla_{\boldsymbol{\theta}}\Psi_\lambda(\boldsymbol{\theta};\rho_*)=\int_{\mathbb{R}^d}\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta},\boldsymbol{\theta}')(\rho_t(\boldsymbol{\theta}')-\rho_*(\boldsymbol{\theta}'))\mathrm{d}\boldsymbol{\theta}'.$$

According to condition A3, $\nabla_{\boldsymbol{\theta}}U$ is K_3-bounded-Lipschitz with respect to $(\boldsymbol{\theta},\boldsymbol{\theta}')$. Therefore,

$$\sup_{\boldsymbol{\theta}\in\mathbb{R}^d}\|\nabla_{\boldsymbol{\theta}}\Psi_\lambda(\boldsymbol{\theta};\rho_t)-\nabla_{\boldsymbol{\theta}}\Psi_\lambda(\boldsymbol{\theta};\rho_*)\|_2\le K_3\cdot d_{\mathrm{BL}}(\rho_t,\rho_*)\to 0, \tag{6.62}$$

as $d_{\mathrm{BL}}(\rho_t,\rho_*)\to 0$. Accordingly, we have

$$\lim_{k\to\infty}\int_{\mathbb{R}^d}\|\nabla_{\boldsymbol{\theta}}(\Psi_\lambda(\boldsymbol{\theta};\rho_{t_k})-\Psi_\lambda(\boldsymbol{\theta};\rho_*))\|_2^2\rho_{t_k}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\le K_3^2\cdot\lim_{k\to\infty}d_{\mathrm{BL}}(\rho_{t_k},\rho_*)^2=0. \tag{6.63}$$

Combining Eq. (6.63) with Eq. (6.61), we have

$$\lim_{k\to\infty}\int_{\mathbb{R}^d}\|\nabla_{\boldsymbol{\theta}}(\Psi_\lambda(\boldsymbol{\theta};\rho_*)+1/\beta\cdot\log\rho_{t_k}(\boldsymbol{\theta}))\|_2^2\rho_{t_k}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}=0. \tag{6.64}$$

Note we have

$$\begin{aligned}
&\int_{\mathbb{R}^d}\|\nabla_{\boldsymbol{\theta}}(\Psi_\lambda(\boldsymbol{\theta};\rho_*)+1/\beta\cdot\log\rho_{t_k}(\boldsymbol{\theta}))\|_2^2\rho_{t_k}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\\
=&\frac{1}{\beta^2}\int_{\mathbb{R}^d}\|\nabla_{\boldsymbol{\theta}}(\rho_{t_k}(\boldsymbol{\theta})\exp\{\beta\Psi_\lambda(\boldsymbol{\theta};\rho_*)\})\|_2^2\cdot\rho_{t_k}(\boldsymbol{\theta})^{-1}\exp\{-2\beta\Psi_\lambda(\boldsymbol{\theta};\rho_*)\}\mathrm{d}\boldsymbol{\theta}\\
=&\frac{1}{\beta^2}\int_{\mathbb{R}^d}\|\nabla_{\boldsymbol{\theta}}[(\rho_{t_k}(\boldsymbol{\theta})\exp\{\beta\Psi_\lambda(\boldsymbol{\theta};\rho_*)\})^{1/2}]\|_2^2\cdot\exp\{-\beta\Psi_\lambda(\boldsymbol{\theta};\rho_*)\}\mathrm{d}\boldsymbol{\theta}.
\end{aligned} \tag{6.65}$$

Define

$$\mu_*(\mathrm{d}\boldsymbol{\theta})=1/Z_*\cdot\exp\{-\beta\Psi_\lambda(\boldsymbol{\theta};\rho_*)\}\mu_0(\mathrm{d}\boldsymbol{\theta}),\quad Z_*=\int_{\mathbb{R}^d}\exp\{-\beta\Psi_\lambda(\boldsymbol{\theta};\rho_*)\}\mu_0(\mathrm{d}\boldsymbol{\theta}), \tag{6.66}$$

$f_k(\boldsymbol{\theta})=[\exp(\beta\Psi_\lambda(\boldsymbol{\theta};\rho_*))\rho_{t_k}(\boldsymbol{\theta})]^{1/2}\in\mathcal{D}\equiv\{f\in L^2(\mathbb{R}^d,\mu_*)\cap C^1(\mathbb{R}^d):\|\nabla f\|_2\in L^2(\mathbb{R}^d,\mu_*)\}$ ($f_k\in C^1(\mathbb{R}^d)$ because $\rho_t(\boldsymbol{\theta})>0$ for any $\boldsymbol{\theta}\in\mathbb{R}^d$ and $\rho_t(\boldsymbol{\theta})\in C^1(\mathbb{R}^d)$ for fixed t), and $f_*(\boldsymbol{\theta})=[\exp(\beta\Psi_\lambda(\boldsymbol{\theta};\rho_*))\rho_*(\boldsymbol{\theta})]^{1/2}\in L^2(\mathbb{R}^d,\mu_*)$. Since we have ρ_{t_k} converges to ρ_* weakly in $L^1(\mathbb{R}^d,\mu_0)$, then f_k converges weakly to f_* in $L^2(\mathbb{R}^d,\mu_*)$. Define $I(f)\equiv\int_{\mathbb{R}^d}\|\nabla f(\boldsymbol{\theta})\|_2^2\cdot\mu_*(\mathrm{d}\boldsymbol{\theta})$. Eq. (6.64) and (6.65) give $\lim_{k\to\infty}I(f_k)=0$. Now we apply Lemma 6.11 with $\Phi(\boldsymbol{\theta})=\beta\Psi_\lambda(\boldsymbol{\theta};\rho_*)$. This Φ satisfies $\beta\lambda/2\cdot\|\boldsymbol{\theta}\|_2^2-2\beta K_2\le\Phi(\boldsymbol{\theta})\le\beta\lambda/2\cdot\|\boldsymbol{\theta}\|_2^2+2\beta K_2$, where K_2 is the constant in Assumption A2. Lemma 6.11 implies $f_*(\boldsymbol{\theta})\equiv F_*$ for some constant F_*.

This proves that $\rho_*(\boldsymbol{\theta})=F_*\cdot\exp\{-\beta\Psi_\lambda(\boldsymbol{\theta};\rho_*)\}$. Combining with the fact that $\int_{\mathbb{R}^d}\rho_*(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}=1$, ρ_* satisfies the Boltzmann fixed point condition. According to Lemma 6.4, the Boltzmann fixed point condition has a unique solution $\rho_*^{\beta,\lambda}$. Therefore, all the converging weak limit of subsequence of ρ_t converges to the same point $\rho_*^{\beta,\lambda}$. As a result, ρ_t converges to $\rho_*^{\beta,\lambda}$ weakly in $L^1(\mathbb{R}^d)$.

□

6.3 Proof of Proposition 3, Theorem 4, and Theorem 5

Proposition 3 is given by Lemma 6.6, 6.4, and Lemma 6.9. Theorem 4 is given by Lemma 6.2, 6.4, 6.5, and 6.12.

Now we prove Theorem 5. First, according to Lemma 6.5, for any $\eta > 0$, there exists constant K depending on η, K_0, K_1, K_2, K_3, such that as we take $\beta \geq KD$, we have

$$R(\rho_*^{\beta,\lambda}) \leq \inf_{\rho \in \mathscr{P}(\mathbb{R}^D)} R_\lambda(\rho) + \eta/3. \tag{6.67}$$

According to Lemma 6.12, we have ρ_t converges to $\rho_*^{\beta,\lambda}$ weakly. Therefore, there exists $T = T(\eta, V, U, \{K_i\}, D, \lambda, \beta) < \infty$, so that $d_{\mathrm{BL}}(\rho_t, \rho_*^{\beta,\lambda}) \leq \eta/(3Z)$ for any $t \geq T$, where $Z = Z(\{K_i\})$ is the bounded-Lipschitz constant of R with respect to ρ. Hence, we have

$$R(\rho_t) \leq R(\rho_*^{\beta,\lambda}) + \eta/3 \tag{6.68}$$

for any $t \geq T$.

Finally, according to Theorem 3, there exists K' depending on K_i's, so that for all $k \leq 10T/\varepsilon$, we have

$$|R_N(\boldsymbol{\theta}^k) - R_{\rho_{k\varepsilon}}| \leq K' e^{K'T} \sqrt{1/N \vee \varepsilon} \cdot \Big[\sqrt{D + \log(N(1/\varepsilon \vee 1))} + z\Big],$$

with probability at least $1 - e^{-z^2}$. Hence there exists $C_0 = C_0(\eta, \{K_i\}, \delta)$, so that as $N, 1/\varepsilon \geq C_0 \exp\{C_0 T\} D$ and $\varepsilon \geq 1/N^{10}$, we have

$$|R_N(\boldsymbol{\theta}^k) - R(\rho_{k\varepsilon})| \leq \eta/3, \tag{6.69}$$

with probability at least $1 - \delta$.

Combining Eq. (6.67), (6.68), and (6.69) we get the desired result.

6.4 Dependence of convergence time on D and η

Theorem 5 does not provide any estimate for the dependence of the convergence time on the problem dimensions D and on the accuracy η. However the proof suggests the following heuristic. When ρ_t is sufficiently close to the minimizer ρ_*, we heuristically can approximate the free energy dissipation formula (6.2) as

$$\partial_t F_{\beta,\lambda}(\rho_t) \approx -\int_{\mathbb{R}^d} \|\nabla_{\boldsymbol{\theta}}(\Psi_\lambda(\boldsymbol{\theta}; \rho_*) + 1/\beta \cdot \log \rho_t(\boldsymbol{\theta}))\|_2^2 \rho_t(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}\,. \tag{6.70}$$

This is the same as the free energy dissipation for the Fokker-Planck equation with potential $\Psi_\lambda(\boldsymbol{\theta}; \rho_*)$. This suggests that, close to ρ_*, convergence should be dominated by the speed of convergence in this Fokker-Plank equation, which is controlled by the log-Sobolev constant of the potential $\Psi_\lambda(\boldsymbol{\theta}; \rho_*)$, to be denote by c_* [MV00]:

$$F_{\beta,\lambda}(\rho_t) \lesssim F_{\beta,\lambda}(\rho_{t_0})\, e^{-c_*(t-t_0)}\,. \tag{6.71}$$

Note that the log-Sobolev constant can be exponentially small in D. We expect this heuristic to capture the rough dependence of the convergence time T on η and D, hence suggesting $T = e^{O(D)} \log(1/\eta)$.

7 Numerical Experiments

In this section, we discuss numerical experiments whose results were presented in the main text, as well as some additional ones. Some technical details of the figures in the main text are also presented here; in particular, Section 7.1.1 for Figure 1, Section 7.1.2 for Figure 2, Section 7.2 for Figure 3, and Section 7.3 for Figure 4.

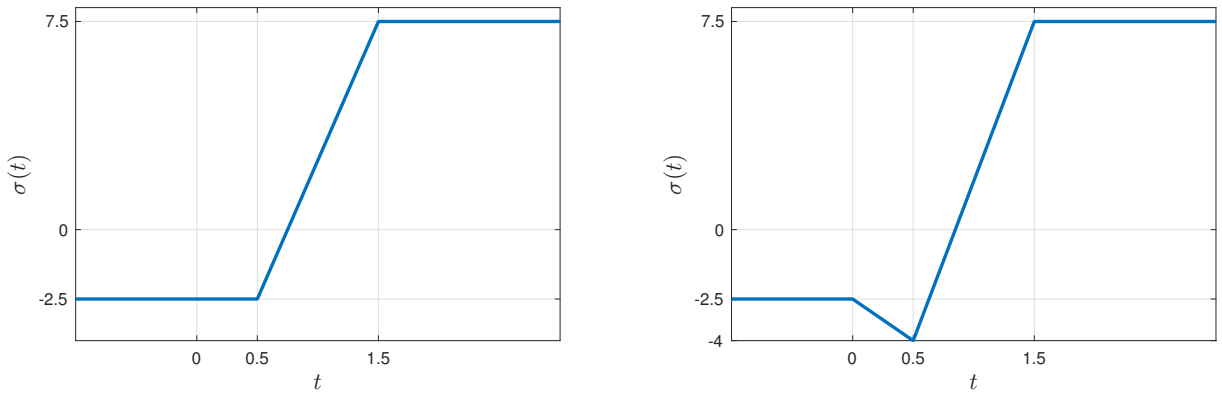

Figure 7.1: The activation functions $\sigma(t)$ used in Section 7.1 (left plot) and Section 7.3 (right plot).

7.1 Isotropic Gaussians

In this section, we present details of the numerical experiments pertaining to the example of centered isotropic Gaussians:

With probability 1/2: $y = +1$, $\boldsymbol{x} \sim \mathsf{N}(0, (1+\Delta)^2 \boldsymbol{I}_d)$.

With probability 1/2: $y = -1$, $\boldsymbol{x} \sim \mathsf{N}(0, (1+\Delta)^2 \boldsymbol{I}_d)$.

In all numerical examples in this section, we use the activation $\sigma_*(\boldsymbol{x}; \boldsymbol{\theta}_i) = \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle)$, where $\sigma(t) = s_1$ if $t \le t_1$, $\sigma(t) = s_2$ if $t \ge t_2$, and $\sigma(t)$ interpolated linearly for $t \in (t_1, t_2)$. In simulations we use $t_1 = 0.5$, $t_2 = 1.5$, $s_1 = -2.5$, $s_2 = 7.5$. This is also used for examples with centered Gaussians in the main text, cf. Figures 1 and 2, and Section 4 in the supplemental information. This activation is plotted in Figure 7.1.

7.1.1 Empirical validation of distributional dynamics

Here we discuss empirical validation for the dynamics in the isotropic Gaussian example.

PDE simulation. Simulating the PDE (Eq. [13] of the main text) for general d is computationally intensive. In order to simplify the problem, we only consider $d = \infty$. In that case, we recall that the risk is given by Eq. (4.10), which we copy here for ease of reference:

$$\overline{R}_\infty(\overline{\rho}) = \frac{1}{2}\left(1 - \int q_+(r)\,\overline{\rho}(\mathrm{d}r)\right)^2 + \frac{1}{2}\left(1 + \int q_-(r)\,\overline{\rho}(\mathrm{d}r)\right)^2, \tag{7.1}$$

where $q_\pm(t) = \mathbb{E}\{\sigma((1 \pm \Delta)tG)\}$, $G \sim \mathsf{N}(0,1)$. In addition, from Eq. (4.12),

$$\psi_\infty(r; \overline{\rho}) = \frac{1}{2}[\langle q_+, \overline{\rho}\rangle - 1]q_+(r) + \frac{1}{2}[\langle q_-, \overline{\rho}\rangle + 1]q_-(r). \tag{7.2}$$

The PDE is then $\partial_t \overline{\rho}_t = 2\xi(t)\partial_r[\overline{\rho}_t \partial_r \psi_\infty(r; \overline{\rho}_t)]$.

The solution to the PDE is approximated, at all time t, by the following multiple-deltas ansatz:

$$\overline{\rho}_t = \frac{1}{J}\sum_{i=1}^{J} \delta_{r_i(t)}\,, \tag{7.3}$$

where $J \in \mathbb{N}$ is a pre-chosen parameter. Note that for a fixed J, if the PDE is initialized at $\overline{\rho}_0$ taking the above form, then for any $t \geq 0$, $\overline{\rho}_t$ remains in the above form. Then for any smooth test function $f : \mathbb{R} \mapsto \mathbb{R}$ with compact support,

$$\frac{1}{J}\sum_{i=1}^{J} f'(r_i(t))r_i'(t) = \partial_t\langle f, \overline{\rho}_t\rangle = -2\xi(t)\langle f', \overline{\rho}_t \partial_r \psi_\infty(r; \overline{\rho}_t)\rangle \tag{7.4}$$

$$= -2\xi(t)\frac{1}{J}\sum_{i=1}^{J} f'(r_i(t))\partial_r\psi_\infty(r_i(t); \overline{\rho}_t). \tag{7.5}$$

Under this ansatz, let us write $\overline{R}_\infty(\overline{\rho}_t) = \overline{R}_{\infty,J}(\mathbf{r}(t))$, where $\mathbf{r}(t) = (r_1(t), ..., r_J(t))^\top$, and

$$\overline{R}_{\infty,J}(\mathbf{r}) = \frac{1}{2}\left(1 - \frac{1}{J}\sum_{i=1}^{J} q_+(r_i)\right)^2 + \frac{1}{2}\left(1 + \frac{1}{J}\sum_{i=1}^{J} q_-(r_i)\right)^2. \tag{7.6}$$

Notice that $\partial_r\psi_\infty(r_i(t); \overline{\rho}_t) = (J/2)(\nabla\overline{R}_{\infty,J}(\mathbf{r}(t)))_i$. Therefore we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{r}(t) = -J\xi(t)\nabla\overline{R}_{\infty,J}(\mathbf{r}(t)). \tag{7.7}$$

Hence under the multiple-deltas ansatz, one can simulate numerically the PDE via the above evolution equation of $\mathbf{r}(t)$. In particular, given $\mathbf{r}(t)$, one approximates $\mathbf{r}(t + \delta t)$ for some small displacement δt by

$$\mathbf{r}(t + \delta t) \approx \mathbf{r}(t) - J\xi(t)\nabla\overline{R}_{\infty,J}(\mathbf{r}(t))\delta t. \tag{7.8}$$

In general, one would want to take a large J to obtain a more accurate approximation. There are certain cases where one can take small J (even $J = 1$). An example of such case is given in the following.

Details of Figure 1 of the main text. For the data generation, we set $\Delta = 0.8$. For the SGD simulation, we take $d = 40$, $N = 800$, with $\varepsilon = 10^{-6}$ and $\xi(t) = 1$. The weights are initialized as $(\boldsymbol{w}_i)_{i\leq N} \sim_{iid} \mathsf{N}(0, 0.8^2/d \cdot \mathbf{I}_d)$. We take a single SGD run. At iteration $10^3, 4\times10^6, 10^7$, we plot the histogram of $(\|\boldsymbol{w}_i\|_2)_{i\leq N}$. This produces the results of the SGD in Figure 1 of the main text.

To obtain results from the PDE, we take $J = 400$, and generate $r_i(0) = \|Z_i\|_2$, where $(Z_i)_{i\leq J} \sim_{iid} \mathsf{N}(0, 0.8^2/d \cdot \mathbf{I}_d)$. We obtain $\mathbf{r}(t)$ from $t = 0$ until $t = 10^7\varepsilon$, by discretizing this interval with 10^5 points equally spaced on the $\log_{10}$ scale and sequentially computing $\mathbf{r}(t)$ at each point using Eq. (7.8). Note that the SGD result at iteration k corresponds to $\mathbf{r}(\varepsilon k)$. We re-simulate the PDE for 100 times, each with an independently generated initialization. The obtained histogram for the PDE, as shown in the figure, is the aggregation of these 100 runs.

Further numerical simulations. Figure 7.2 plots the evolution of $\overline{\rho}_t$ for $\Delta = 0.2$. The setting is identical to the one in Figure 1 of the main text, described in the previous paragraphs.

In Figure 7.3, we plot the evolution of the population risk for the SGD and its PDE prediction counterpart, for $\Delta = 0.2$ and $\Delta = 0.8$. The setting for the SGD plots is the same as described in the previous paragraphs. We compute the risk attained by the SGD by Monte Carlo averaging over 10^4 samples. The setting for the PDE plots tagged "$J = 400$" is almost the same as in the previous paragraphs, except that we take only 1 run. For the PDE plot tagged "$J = 1$", we take $J = 1$ and $r(0) = 0.8$ instead. In the inset plot, we also show the evolution of $(1/N)\sum_{i=1}^{N}\|\boldsymbol{w}_i\|_2$ of the SGD, and $(1/J)\sum_{i=1}^{J} r_i(t)$ of the PDE.

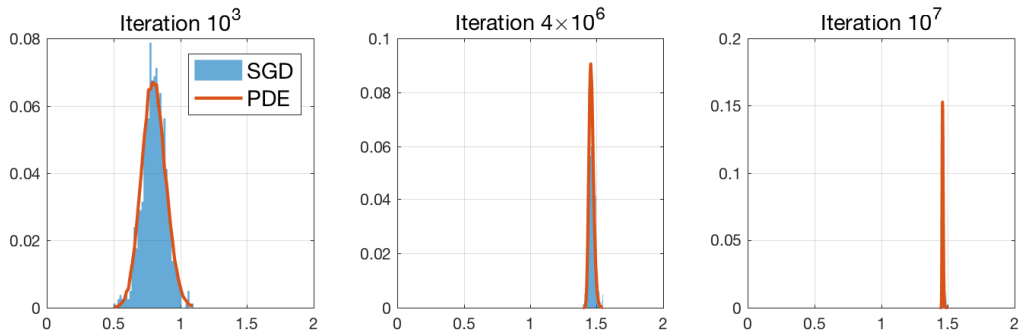

Figure 7.2: Evolution of the reduced distribution $\overline{\rho}_t$ for $\Delta = 0.2$, in the isotropic Gaussians example of Section 7.1.

In Figure 7.4, we plot the function $\overline{R}_d^{(1)}(r)$, for $d = 40$ and $\Delta = 0.2$. (Recall $\overline{R}_d^{(1)}(r)$ from Eq. [14] of the main text, and see also Section 7.1.3.) On this landscape, we also plot the evolution of the corresponding SGD and PDE, as described in the last paragraph.

Comments. We observe in Figure 7.3 a good match between the SGD and the PDE, even when $J = 1$, for $\Delta = 0.2$. This can be explained with our theory, which predicts that at $\Delta = 0.2$, the minimum risk is achieved by the uniform distribution over a sphere of radius $\|\boldsymbol{w}\|_2 = r_*$ (see also Section 7.1.3). This corresponds to $\overline{\rho}_t$, as $t \to \infty$, being a delta function and placing probability 1 at r_*. Furthermore due to the way we initialize the SGD, $\overline{\rho}_0$ is well concentrated. One can then expect that $\overline{\rho}_t$ is also well concentrated at all time t, in which case $J = 1$ is sufficient. This claim is reflected in our numerical experiments, shown in Figure 7.2.

We also observe in Figure 7.3 that the case $\Delta = 0.2$ has a rapid transition from a high risk to a lower risk, unlike the case $\Delta = 0.8$. This is also expected from our theory. As said above, $\overline{\rho}_t$ is approximately a delta function at all time t, and the position $r(t)$ evolves by gradient flow in the landscape of $\overline{R}_d^{(1)}(r)$. This latter claim is well supported by Figure 7.4. As observed in Figure 7.4, $\overline{R}_d^{(1)}(r)$ is rather benign, and hence the transition of the population risk should be smooth. However the case for $\Delta = 0.8$ is different: $\overline{\rho}_t$ is not concentrating at large t, as evident in Figure 1 of the main text, even though $\overline{R}_d^{(1)}(r)$ is generally benign for a vast variety of values of d and Δ (see Figure 7.6 and Section 7.1.3).

Note that the computation of the PDE assumes $d = \infty$. Furthermore it also requires $N = \infty$ (recalling Theorem 3 of the main text). The discrepancy to the SGD is due to the fact that d and N are finite in the SGD simulations. Nevertheless in our numerical examples, such discrepancy is insignificant.

7.1.2 Empirical validation of the statics

Here we discuss numerical verification for the statics in the isotropic Gaussian example.

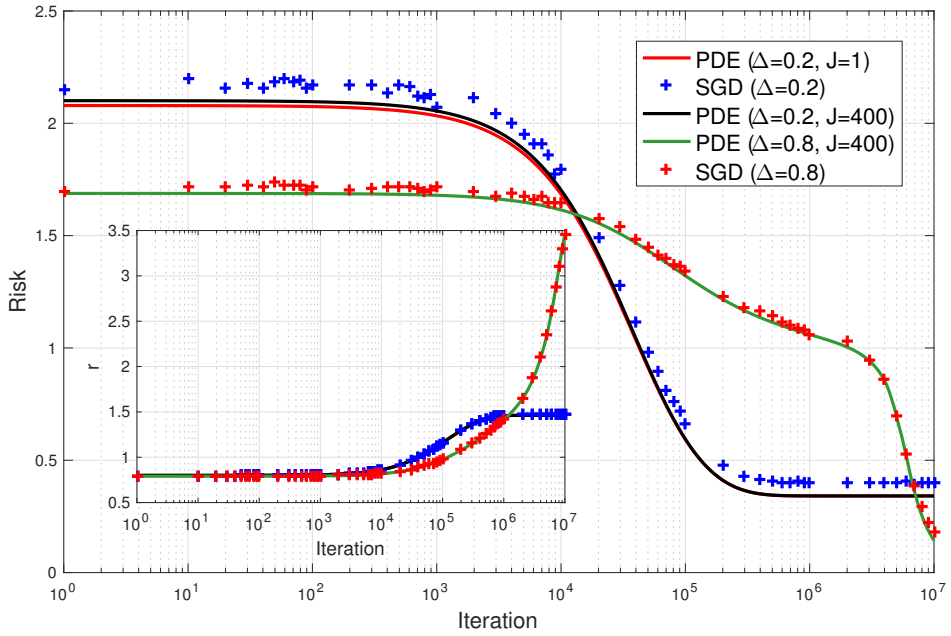

Figure 7.3: The evolution of the population risk and the parameter r of the reduced distribution $\overline{\rho}_t$, in the isotropic Gaussians example of Section 7.1.

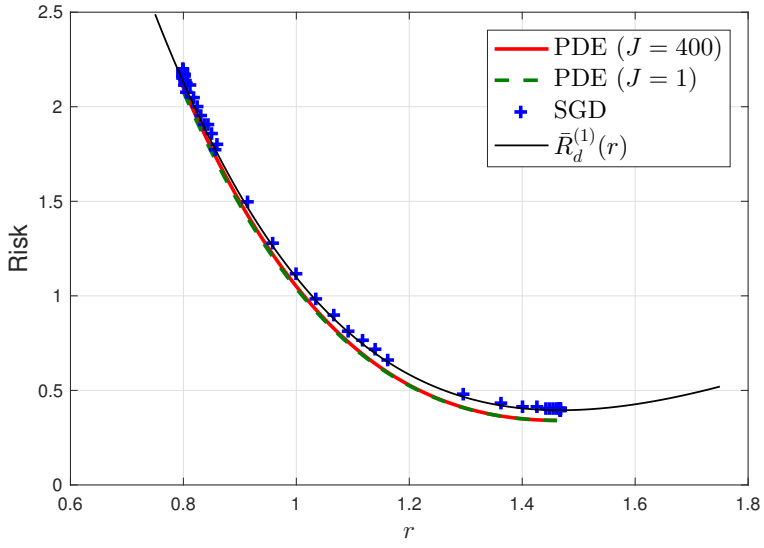

Figure 7.4: The function $\overline{R}_d^{(1)}(r)$ vs r, as well as the evolution of the SGD and the PDE on this landscape, for $\Delta = 0.2$ and $d = 40$, in the isotropic Gaussians example of Section 7.1. Here the SGD and the PDE evolve from the leftmost point to the rightmost point.

Optimizing $\overline{R}_d(\overline{\rho})$. For the chosen activation, we have from Eq. (4.8) that

$$\overline{R}_d(\overline{\rho}) = 1 + 2\int v(r)\,\overline{\rho}(\mathrm{d}r) + \int u_d(r_1, r_2)\,\overline{\rho}(\mathrm{d}r_1)\,\overline{\rho}(\mathrm{d}r_2)\,, \tag{7.9}$$

$$v(r) = -\frac{1}{2}g(0, (1+\Delta)r) + \frac{1}{2}g(0, (1-\Delta)r)\,, \tag{7.10}$$

$$u_d(r_1, r_2) = \frac{\Gamma(d/2)}{\Gamma(1/2)\Gamma((d-1)/2)}\int_{\theta=0}^{\pi} \hat{u}(r_1, r_2, \theta)\sin^{d-2}\theta \mathrm{d}\theta\,, \tag{7.11}$$

$$\hat{u}(r_1, r_2, \theta) = \frac{1}{2}f((1+\Delta)r_1, (1+\Delta)r_2, \theta) + \frac{1}{2}f((1-\Delta)r_1, (1-\Delta)r_2, \theta)\,, \tag{7.12}$$

$$f(r_1, r_2, \theta) = \int_{x=-\infty}^{+\infty} \sigma(r_1 x) g(r_2 x\cos\theta, r_2\sin\theta)\phi(x)\mathrm{d}x\,, \tag{7.13}$$

$$\begin{aligned} g(a, b) = {} & s_2 + (s_1 - \sigma_{\mathrm{itc}} - \sigma_{\mathrm{sl}}a)\Phi\left(\frac{t_1 - a}{b}\right) + (\sigma_{\mathrm{sl}}a + \sigma_{\mathrm{itc}} - s_2)\Phi\left(\frac{t_2 - a}{b}\right) \\ & + \sigma_{\mathrm{sl}} b\left[\phi\left(\frac{t_1 - a}{b}\right) - \phi\left(\frac{t_2 - a}{b}\right)\right]. \end{aligned} \tag{7.14}$$

where $\sigma_{\mathrm{sl}} = (s_2 - s_1)/(t_2 - t_1)$, $\sigma_{\mathrm{itc}} = s_1 - \sigma_{\mathrm{sl}}t_1$, $\phi(x) = \exp(-x^2/2)/\sqrt{2\pi}$, $\Phi(x) = \int_{-\infty}^{x}\phi(t)\mathrm{d}t$, and Γ is the Gamma function. To numerically optimize $\overline{R}_d(\overline{\rho})$, we perform the following approximation:

$$\inf_{\overline{\rho}} \overline{R}_d(\overline{\rho}) \approx \inf_{p_i \geq 0,\, \sum_{i=1}^K p_i = 1} \overline{R}_d\left(\sum_{i=1}^K p_i \delta_{o_i}\right). \tag{7.15}$$

Here $o_i \in \mathbb{R}$, $i = 1, ..., K$, are K pre-chosen points. Let $\mathbf{v} = (v(o_1), ..., v(o_K))^\top$ and $\mathbf{U} = (u_d(o_i, o_j))_{1\leq i,j\leq K}$. Then the approximation becomes

$$\inf_{\overline{\rho}} \overline{R}_d(\overline{\rho}) \approx \inf_{p_i \geq 0,\, \sum_{i=1}^K p_i = 1} \left\{1 + 2\mathbf{v}^\top \mathbf{p} + \mathbf{p}^\top \mathbf{U}\mathbf{p}\right\}, \tag{7.16}$$

which is a quadratic programming problem and can be solved numerically. Here $\mathbf{v}$ can be computed easily with the explicit formula, and the computation of $\mathbf{U}$ amounts to numerically evaluating double integrals. In the case $d = \infty$, the computation of $\mathbf{U}$ is much easier, since

$$u_\infty(r_1, r_2) = \frac{1}{2}g(0, (1+\Delta)r_1)g(0, (1+\Delta)r_2) + \frac{1}{2}g(0, (1-\Delta)r_1)g(0, (1-\Delta)r_2). \tag{7.17}$$

Details of Figure 2 of the main text. For the SGD simulation, we take $N = 800$, with $\varepsilon = 3 \times 10^{-3}$ and $\xi(t) = t^{-1/4}$. The weights are initialized as $(\boldsymbol{w}_i)_{i\leq N} \sim_{iid} \mathsf{N}(0, 0.4^2/d \cdot \mathbf{I}_d)$. We compute the risk attained by the SGD by Monte Carlo averaging over 10^4 samples. We take a single SGD run per Δ, per d, and report the risk at iteration 10^7.

For the approximate optimization of $\overline{R}_d(\overline{\rho})$, we choose $K = 100$, and o_i, $i = 1, ..., K$, being equally spaced on the interval [0.01, 10].

For the optimization of $\overline{R}_d^{(1)}(r)$ (recalling Eq. [14] in the main text), we approximate it with $\min_{i=1,...,K} \overline{R}_d^{(1)}(o_i)$, for the above chosen o_i and K.

We find that in general, one needs higher $\max_{i=1,...,K} o_i$ to produce accurate results for higher Δ. For the chosen set of o_i's, we choose to plot up until $\Delta = 0.8$.

Further numerical simulations. In Figure 7.5, we extend Figure 2 of the main text to include results for additional values of d. The setting remains the same.

This figure provides further support to the respective discussion in the main text. For the threshold values of Δ for which the minimum risk is achieved by a uniform distribution $\rho_{r_*}^{\mathrm{unif}}$ over a sphere of radius $\|\boldsymbol{w}\|_2 = r_*$ (see the main text around Eq. [14], and Section 7.1.3).

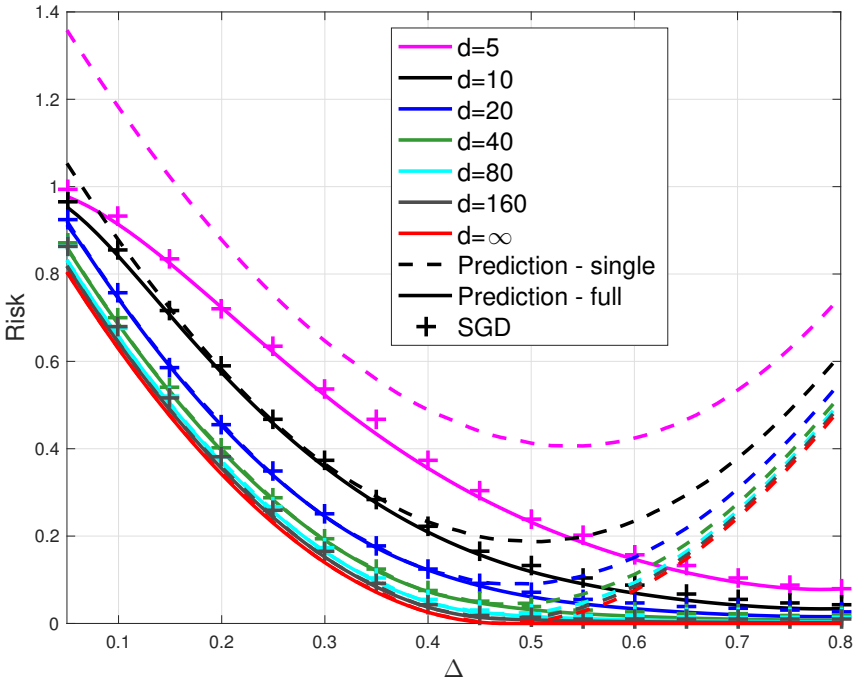

Figure 7.5: The population risk as a function of Δ for different values of d, in the isotropic Gaussians example of Section 7.1. Here "Prediction - single" refers to $\min_{r \geq 0} \overline{R}_d^{(1)}(r)$, "Prediction - full" refers to the optimized $R(\rho)$ as described in Section 7.1.2, and "SGD" refers to the risk attained by the SGD.

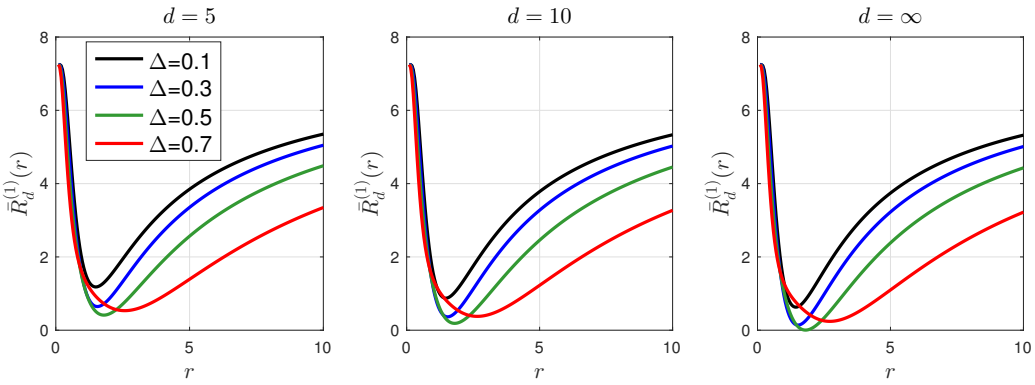

Figure 7.6: The function $\overline{R}_d^{(1)}(r)$ for different values of d and Δ, in the isotropic Gaussians example of Section 7.1.

d	Δ_d^{l}	Δ_d^{h}
5	N/A	N/A
10	N/A	N/A
20	0.08	0.38
40	0.03	0.42
80	0.02	0.45
160	0.0	0.46
∞	0	0.47

Table 1: Δ_d^{l} and Δ_d^{h} for different values of d, in the isotropic Gaussians example of Section 7.1. Here "N/A" refers to that no values of Δ are found to satisfy the condition of Lemma 1 in the main text. Note that for $d = \infty$, the value $\Delta_\infty^{\mathrm{l}} = 0$ is exact, according to Theorem 4.2.

7.1.3 Checking the condition of Lemma 1 in the main text

We check of the condition of Lemma 1 in the main text. This has two steps: (1) we solve for the minimizer r_* of $\overline{R}_d^{(1)}(r) = 1 + 2v(r) + u_d(r, r)$, where $v(r)$ and $u_d(r_1, r_2)$ are given by Eq. (7.10) and (7.11) respectively, and (2) we check whether $v(r) + u_d(r, r_*) \geq v(r_*) + u_d(r_*, r_*)$ for all $r \geq 0$. Figure 7.6 suggests that the behavior of $\overline{R}_d^{(1)}(r)$ is rather benign and hence r_* can be solved easily by searching for a local minimum. For the second step, we check the condition on a grid of values of r from 0.1 to 10 with a spacing of 0.1, for each value of Δ on a grid from 0.01 to 0.99 with a spacing of 0.01. In general, we find that the conditioned is satisfied for $\Delta \in [\Delta_d^{\mathrm{l}}, \Delta_d^{\mathrm{h}}]$. Table 1 reports Δ_d^{l} and Δ_d^{h} for a number of values of d for the isotropic Gaussians example with the given activation function.

7.2 Centered anisotropic Gaussians with ReLU Activation

In this section, we present details of the numerical experiments pertaining to the example of anisotropic Gaussians with ReLU activation. In particular, we use the activation $\sigma_*(\boldsymbol{x}; \boldsymbol{\theta}) = a \max(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b, 0)$, with $\boldsymbol{\theta} = (\boldsymbol{w}, a, b) \in \mathbb{R}^{d+2}$. We consider the centered anisotropic Gaussian case:

With probability 1/2: $y = +1$, $\boldsymbol{x} \sim \mathsf{N}(0, \boldsymbol{\Sigma}_+)$.

With probability 1/2: $y = -1$, $\boldsymbol{x} \sim \mathsf{N}(0, \boldsymbol{\Sigma}_-)$.

More specifically, we opt for

$$\boldsymbol{\Sigma}_+ = \mathrm{Diag}\big(\underbrace{(1+\Delta)^2, \ldots, (1+\Delta)^2}_{s_0}, \underbrace{1, \ldots, 1}_{d-s_0}\big)\,, \tag{7.18}$$

$$\boldsymbol{\Sigma}_- = \mathrm{Diag}\big(\underbrace{(1-\Delta)^2, \ldots, (1-\Delta)^2}_{s_0}, \underbrace{1, \ldots, 1}_{d-s_0}\big)\,. \tag{7.19}$$

This setting is used in Figure 3 in the main text.

We consider $s_0 = \gamma d$ for some $\gamma \in (0,1)$. For simplicity, we consider the limit $d \to \infty$. For $\boldsymbol{\theta} \sim \rho$, let $\overline{\rho}$ be the joint distribution of four parameters $\mathbf{r} = (a, b, r_1 = \|\boldsymbol{w}_{1:s_0}\|_2, r_2 = \|\boldsymbol{w}_{(s_0+1):d}\|_2)$, where $\boldsymbol{w}_{i:j} = (w_i, ..., w_j)^\top$. Using a similar argument to Section 4, we have, in the limit $d \to \infty$, the risk $R(\rho) = \overline{R}_\infty(\overline{\rho})$ for

$$\overline{R}_\infty(\overline{\rho}) = \frac{1}{2}\left(1 - \int a q_+(r_1, r_2, b)\overline{\rho}(\mathrm{d}\mathbf{r})\right)^2 + \frac{1}{2}\left(1 + \int a q_-(r_1, r_2, b)\overline{\rho}(\mathrm{d}\mathbf{r})\right)^2\,, \tag{7.20}$$

$$q_\pm(r_1, r_2, b) = b\,\Phi\left(\frac{b}{\sqrt{(1\pm\Delta)^2 r_1^2 + r_2^2}}\right) + \sqrt{(1\pm\Delta)^2 r_1^2 + r_2^2}\,\phi\left(\frac{b}{\sqrt{(1\pm\Delta)^2 r_1^2 + r_2^2}}\right)\,, \tag{7.21}$$

where $\phi(x) = \exp(-x^2/2)/\sqrt{2\pi}$ and $\Phi(x) = \int_{-\infty}^x \phi(t)\mathrm{d}t$. Furthermore, letting $\overline{\rho}_t$ denote the corresponding distribution at time t, the PDE [7] in the main text can be reduced to the following PDE of $\overline{\rho}_t$:

$$\partial_t \overline{\rho}_t = 2\xi(t)\nabla_{\mathbf{r}} \cdot (\overline{\rho}_t \nabla_{\mathbf{r}} \psi_\infty(\mathbf{r}; \overline{\rho}_t))\,, \tag{7.22}$$

$$\begin{aligned}\psi_\infty(\mathbf{r}; \overline{\rho}) = {} & \frac{1}{2}\left[\int a' q_+(r_1', r_2', b')\mathrm{d}\overline{\rho}(a', b', r_1', r_2') - 1\right] a q_+(r_1, r_2, b) \\ & + \frac{1}{2}\left[\int a' q_-(r_1', r_2', b')\mathrm{d}\overline{\rho}(a', b', r_1', r_2') + 1\right] a q_-(r_1, r_2, b).\end{aligned} \tag{7.23}$$

PDE simulation. As in Section 7.1.1, we posit that the solution to the PDE can be approximated, at all time t, by the multiple-deltas ansatz:

$$\overline{\rho}_t = \frac{1}{J}\sum_{i=1}^J \delta_{\mathbf{r}_i(t)}\,, \tag{7.24}$$

where $J \in \mathbb{N}$ is a pre-chosen parameter, and $\mathbf{r}_i(t) = (a_i(t), b_i(t), r_{1,i}(t), r_{2,i}(t))$. Following the same argument as in Section 7.1.1, we obtain the following evolution equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{r}_i(t) = -J\xi(t)\nabla_i \overline{R}_{\infty,J}(\mathbf{r}_1(t), ..., \mathbf{r}_J(t)), \tag{7.25}$$

for $i = 1, ..., J$, where $\overline{R}_{\infty,J}(\mathbf{r}_1(t), \ldots, \mathbf{r}_J(t)) = \overline{R}_\infty(\overline{\rho}_t)$ under the ansatz, and ∇_i denotes the gradient of $\overline{R}_{\infty,J}(\mathbf{r}_1, ..., \mathbf{r}_J)$ w.r.t. $\mathbf{r}_i$. More explicitly,

$$\overline{R}_{\infty,J}(\mathbf{r}_1, \ldots, \mathbf{r}_J) = \frac{1}{2}\left(1 - \frac{1}{J}\sum_{i=1}^J a_i q_+(r_{1,i}, r_{2,i}, b_i)\right)^2 + \frac{1}{2}\left(1 + \frac{1}{J}\sum_{i=1}^J a_i q_-(r_{1,i}, r_{2,i}, b_i)\right)^2\,. \tag{7.26}$$

Again, given $\mathbf{r}_i(t)$, one approximates $\mathbf{r}_i(t+\delta t)$ for some small displacement δt by

$$\mathbf{r}_i(t+\delta t) \approx \mathbf{r}_i(t) - J\xi(t)\nabla_i \overline{R}_{\infty,J}(\mathbf{r}_1,\dots,\mathbf{r}_J)\delta t. \tag{7.27}$$

Details of Figure 3 of the main text. For the SGD simulation, we take $d = 320$, $s_0 = 60$, $N = 800$, with $\varepsilon = 2\times 10^{-4}$ and $\xi(t) = t^{-1/4}$. The weights are initialized as $(\boldsymbol{w}_i)_{i\le N} \sim_{iid} \mathsf{N}(0, 0.8^2/d\cdot \mathbf{I}_d)$, $(a_i)_{i\le N} = 1$ and $(b_i)_{i\le N} = 1$. We take a single SGD run. We compute the risk attained by the SGD by Monte Carlo averaging over 10^4 samples.

To obtain results from the PDE, we take $J = 400$. We initialize $r_{1,i}(0) = \|Z_{1,i}\|_2$ and $r_{2,i}(0) = \|Z_{2,i}\|_2$, where $(Z_{1,i})_{i\le N} \sim_{iid} \mathsf{N}(0, 0.8^2/d\cdot \mathbf{I}_{s_0})$ and $(Z_{2,i})_{i\le N} \sim_{iid} \mathsf{N}(0, 0.8^2/d\cdot \mathbf{I}_{d-s_0})$ independently, along with $a_i(0) = 1$, $b_i(0) = 1$. We obtain $\mathbf{r}_i(t)$ from $t = 0$ until $t = 10^7\varepsilon$, by discretizing this interval with 10^5 points equally spaced on the $\log_{10}$ scale and sequentially computing $\mathbf{r}_i(t)$ at each point using Eq. (7.27). Note that the SGD result at iteration ℓ corresponds to $\mathbf{r}_i(\varepsilon^{4/3}\ell)$. We take a single run of the PDE.

To produce the inset plot in Figure 3 of the main text, for the "a (mean)" axis, we compute $\frac{1}{N}\sum_{i=1}^N a_i$ for the SGD and $\frac{1}{J}\sum_{i=1}^J a_i(t)$ for the PDE. Similarly, for the "b (mean)" axis, we compute $\frac{1}{N}\sum_{i=1}^N b_i$ for the SGD and $\frac{1}{J}\sum_{i=1}^J b_i(t)$ for the PDE, and for the "r_1 (mean)" axis, we compute $\frac{1}{N}\sum_{i=1}^N \|\boldsymbol{w}_{i,1:s_0}\|_2$ for the SGD and $\frac{1}{J}\sum_{i=1}^J r_{1,i}(t)$ for the PDE.

Further numerical simulations. In Figure 7.7, we plot the evolution of the four parameters, for the same setting as Figure 3 of the main text. Here "a (mean)", "b (mean)" and "r_1 (mean)" hold the same meanings, and "r_2 (mean)" refers to $\frac{1}{N}\sum_{i=1}^N \|\boldsymbol{w}_{i,(s_0+1):d}\|_2$ for the SGD and $\frac{1}{J}\sum_{i=1}^J r_{2,i}(t)$ for the PDE.

In Figure 7.8, we plot the population risk's evolution for the same setting as Figure 3 of the main text, apart from that $\Delta = 0.6$ and s_0 varies.

Comments. We observe a good match between the SGD and the PDE in Figure 3 of the main text as well as Figure 7.7, up until iteration 10^6. In general there is less discrepancy with larger s_0, d and N, recalling that the PDE is computed assuming infinite s_0, d and N. This is evident from Figure 7.8.

As a note, in Figure 7.8, the PDE evolves differently for different s_0. This is because the ratio s_0/d is used to determine the initialization of the PDE.

7.3 Isotropic Gaussians: Predictable Failure of SGD

In this section, we consider the isotropic Gaussians example (see Section 7.1 for the setting and notations), with the following activation function: $\sigma_*(\boldsymbol{x};\boldsymbol{\theta}_i) = \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle)$, where $\sigma(t) = -2.5$ for $t \le 0$, $\sigma(t) = 7.5$ for $t \ge 1.5$, and $\sigma(t)$ linearly interpolates from the knot $(0, -2.5)$ to $(0.5, -4)$, and from $(0.5, -4)$ to $(1.5, 7.5)$. This activation is plotted in Figure 7.1. This corresponds to Section "Predicting failure" and Figure 4 in the main text. The simulation of the PDE can be done in the same way as in Section 7.1.1.

Rationale of the activation choice. We give an explanation for the choice of the above activation based on our theory. We aim to find an activation $\sigma_*(\boldsymbol{x};\boldsymbol{\theta}_i) = \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle)$ in which there exists a local minimum that does not generalize well. To simplify the task, we wish for such minimum to be attained at $\rho_* = \delta_{\mathbf{0}}$. This minimum does not generalize well, since it implies all the weights are zero and the neuron outputs are constant, rendering the network unable to perform classification. Theorem 6 of the main text suggests taking $\sigma(t)$ such that

$$\nabla^2 V(\mathbf{0}) + \nabla^2_{1,1} U(\mathbf{0},\mathbf{0}) \succ 0\,. \tag{7.28}$$

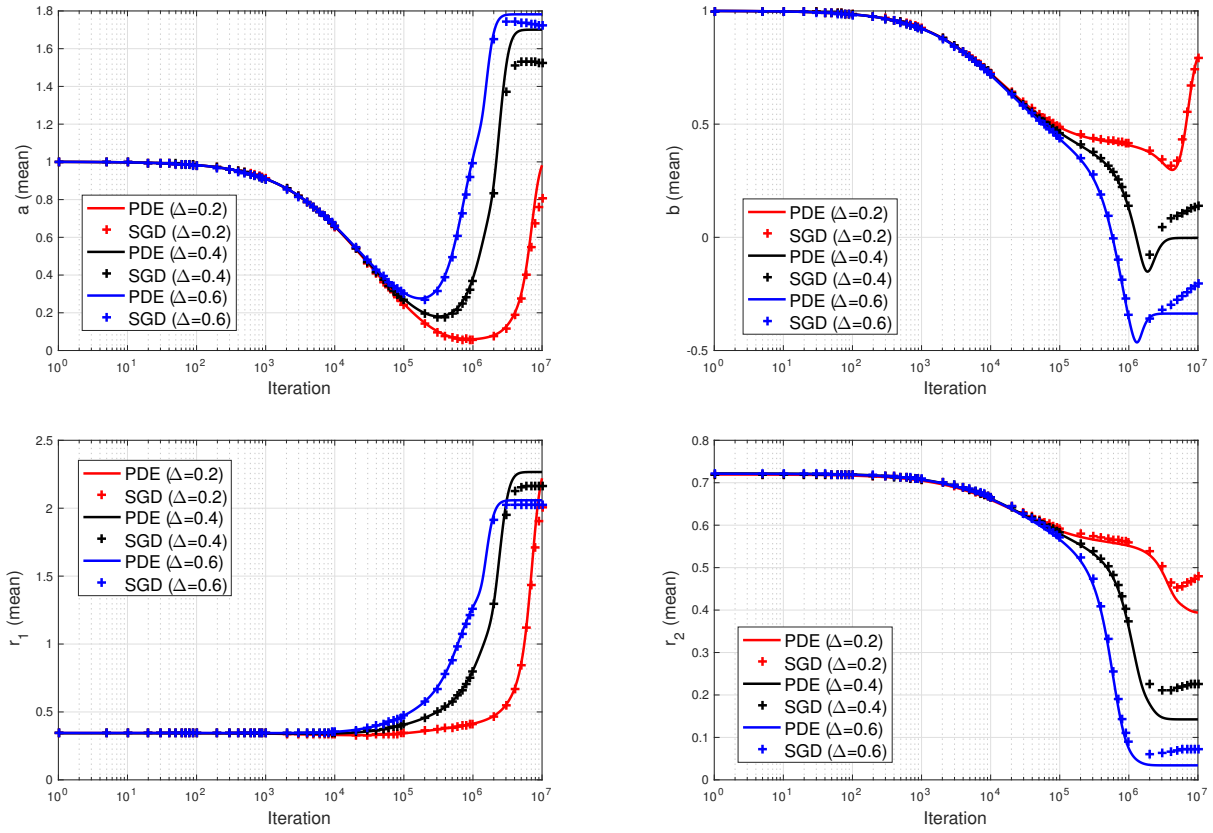

Figure 7.7: The evolution of the four parameters in the anisotropic Gaussians example of Section 7.2.

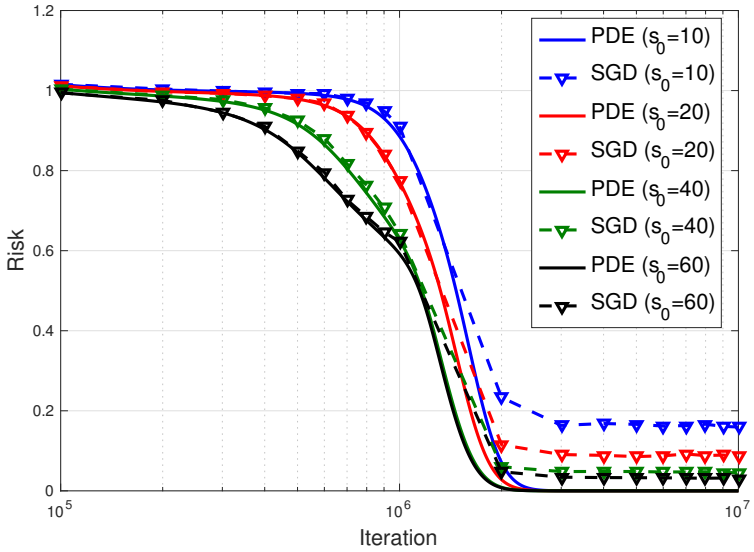

Figure 7.8: The evolution of the population risk for $\Delta = 0.6$, $d = 320$, $N = 800$ in the anisotropic Gaussians example of Section 7.2.

In the isotropic Gaussians case, this becomes

$$\sigma''(0)\left\{(1-\Delta)^2-(1+\Delta)^2+\sigma(0)[(1-\Delta)^2+(1+\Delta)^2]\right\}>0\,. \tag{7.29}$$

(Note that the condition $\nabla V(\mathbf{0})+\nabla_1 U(\mathbf{0},\mathbf{0})=\mathbf{0}$ in Theorem 6 of the main text is trivially satisfied.) Another requirement is that there should still be a minimum whose risk is nearly zero. Hence we do not wish for a dramatic change in the choice of the activation function, as compared to the one used in Section 7.1. That is, we leave $\sigma(0)<0$ unchanged. Hence we would want $\sigma''(0)<0$, which is accomplished by our aforementioned choice.

Note that Theorem 6 of the main text also suggests that if the SGD is initialized sufficiently close to this local minimum, the SGD trajectory should converge to it.

Details of Figure 4 of the main text. For the data generation, we set $\Delta = 0.5$. For the SGD simulation, we take $d = 320$, $N = 800$, with $\varepsilon = 10^{-5}$ and $\xi(t) = t^{-1/4}$. We take a single SGD run each for two different initializations: the weights are initialized as $(\boldsymbol{w}_i)_{i\le N} \sim_{iid} \mathsf{N}(0, \kappa^2/d \cdot \mathbf{I}_d)$ for either $\kappa = 0.1$ or $\kappa = 0.4$. We compute the risk attained by the SGD by Monte Carlo averaging over 10^4 samples.

To obtain results from the PDE, we take $J = 400$, and generate $r_i(0) = \|Z_i\|_2$, where $(Z_i)_{i\le N} \sim_{iid} \mathsf{N}(0, \kappa^2/d \cdot \mathbf{I}_d)$. We obtain $\boldsymbol{r}(t)$ from $t = 0$ until $t = 10^7\varepsilon$, by discretizing this interval with 10^5 points equally spaced on the $\log_{10}$ scale and sequentially computing $\mathbf{r}(t)$ at each point using Eq. (7.8). Note that the SGD result at iteration k corresponds to $\mathbf{r}(\varepsilon^{4/3}k)$. We take a single run of the PDE.

To produce the inset plot, we compute $\frac{1}{N}\sum_{i=1}^N \|\boldsymbol{w}_i\|_2$ for the SGD, and $\frac{1}{J}\sum_{i=1}^J r_i(t)$ for the PDE.

As observed from Figure 4 of the main text, the SGD trajectory with $\kappa = 0.1$ converges to a point where $\|\boldsymbol{w}_i\|_2$ is nearly zero and the risk is very high, in stark contrast to the SGD trajectory with $\kappa = 0.4$, as expected.

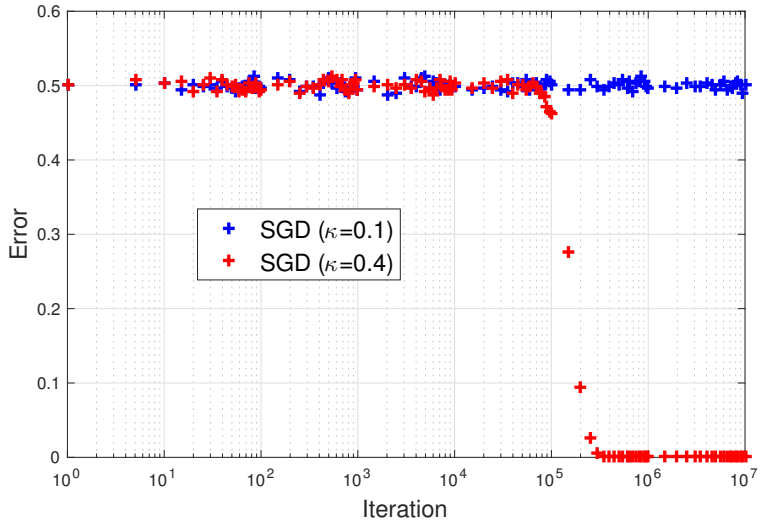

Figure 7.9: The error rate attained by the SGD in the example of Figure 4 of the main text.

Error plot. In Figure 7.9, we plot the empirical error rate attained by the SGD in the above example for the two initializations. Here the error rate is defined as the misclassification probability $\mathbb{P}\{\mathrm{sign}(\hat{y}(\boldsymbol{x};\boldsymbol{\theta})) \neq y\}$, and is computed with Monte Carlo averaging over 10^4 samples. This validates the claim that, in this example, there exists a local minimum which the SGD can converge to, yet has bad generalization (i.e. attains the trivial misclassification rate of 0.5), whereas there is a global minimum which the SGD can also find and yet generalizes well.

A Concentration inequalities

Lemma A.1 (Azuma-Höeffding bound)**.** *Let $(\boldsymbol{X}_k)_{k\geq 0}$, be a martingale taking values in $\mathbb{R}^d$ with respect to the filtration $(\mathcal{F}_k)_{k\geq 0}$, with $\boldsymbol{X}_0 = \boldsymbol{0}$. Assume that the following holds almost surely for all $k \geq 1$:*

$$\mathbb{E}\{e^{\langle \boldsymbol{\lambda}, \boldsymbol{X}_k - \boldsymbol{X}_{k-1}\rangle} | \mathcal{F}_{k-1}\} \leq e^{L^2\|\boldsymbol{\lambda}\|^2/2}\,. \tag{A.1}$$

Then we have

$$\mathbb{P}\Big(\max_{k\leq n} \|\boldsymbol{X}_k\|_2 \geq 2L\sqrt{n}(\sqrt{d}+t)\Big) \leq e^{-t^2}\,. \tag{A.2}$$

Proof. Let $\boldsymbol{Z}_k = \boldsymbol{X}_k - \boldsymbol{X}_{k-1}$ be the martingale differences. By the subgaussian condition (A.1), we get

$$\mathbb{E}\{e^{\langle \boldsymbol{\lambda}, \boldsymbol{X}_n\rangle}\} \leq \mathbb{E}\Big\{\mathbb{E}\{e^{\langle \boldsymbol{\lambda}, \boldsymbol{Z}_n\rangle}|\mathcal{F}_{n-1}\}\, e^{\langle \boldsymbol{\lambda}, \boldsymbol{X}_{n-1}\rangle}\Big\} \tag{A.3}$$

$$\leq e^{L^2\|\boldsymbol{\lambda}\|^2/2}\,\mathbb{E}\{e^{\langle \boldsymbol{\lambda}, \boldsymbol{X}_{n-1}\rangle}\} \leq e^{nL^2\|\boldsymbol{\lambda}\|_2^2/2}\,. \tag{A.4}$$

Letting $\boldsymbol{G} \sim \mathsf{N}(0, \boldsymbol{I}_d)$ a standard Gaussian vector and $\xi \geq 0$,

$$\mathbb{E}\{e^{\xi\|\boldsymbol{X}_n\|_2^2/2}\} = \mathbb{E}_{\boldsymbol{G}}\mathbb{E}\{e^{\sqrt{\xi}\langle \boldsymbol{G}, \boldsymbol{X}_n\rangle}\} \leq \mathbb{E}_{\boldsymbol{G}} e^{nL^2\xi\|\boldsymbol{G}\|_2^2/2} \tag{A.5}$$

$$= (1 - nL^2\xi)^{-d/2}\,. \tag{A.6}$$

By Markov inequality, setting $\xi = 1/(2nL^2)$, we get, for all $t \geq 0$,

$$\mathbb{P}\Big(\|\boldsymbol{X}_n\|_2 \geq 2L\sqrt{n}(\sqrt{d}+t)\Big) \leq e^{d/2-(\sqrt{d}+t)^2} \leq e^{-t^2}. \tag{A.7}$$

Finally, to upper bound $\max_{k\leq n}\|\boldsymbol{X}_k\|_2$, we define the stopping time $\tau \equiv \min\{k : \|\boldsymbol{X}_k\|_2 \geq 2L\sqrt{n}(\sqrt{d}+t)\}$, and the martingale $\overline{\boldsymbol{X}}_k = \boldsymbol{X}_{k\wedge\tau}$. Since $\{\max_{k\leq n}\|\boldsymbol{X}_n\|_2 \geq 2L\sqrt{n}(\sqrt{d}+t)\} = \{\|\overline{\boldsymbol{X}}_n\|_2 \geq 2L\sqrt{n}(\sqrt{d}+t)\}$, the claim follows by applying the previous inequality to $\overline{\boldsymbol{X}}_n$. □

B On the generalization to other loss functions

The objective of this section is to show that the framework of this paper can be formally extended to other loss functions $\ell : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$. All arguments here will be heuristic, and we defer a rigorous study of this problem to future work.

First of all, we note that the population risk reads

$$R_N(\boldsymbol{\theta}) = \mathbb{E}\left\{\ell\left(y, \frac{1}{N}\sum_{i=1}^{N}\sigma_*(\boldsymbol{x};\boldsymbol{\theta}_i)\right)\right\}, \tag{B.1}$$

which naturally leads to the following mean field risk $R : \mathscr{P}(\mathbb{R}^D)\to\mathbb{R}$:

$$R(\rho) = \mathbb{E}\left\{\ell\left(y, \int \sigma_*(\boldsymbol{x};\boldsymbol{\theta})\,\rho(\mathrm{d}\boldsymbol{\theta})\right)\right\}. \tag{B.2}$$

The corresponding distributional dynamics is formally identical to the one for quadratic loss, cf. Eq. (3.1). The only change is in the definition of $\Psi(\boldsymbol{\theta};\rho)$:

$$\partial_t\rho_t(\boldsymbol{\theta}) = 2\xi(t)\nabla\cdot\left[\rho_t(\boldsymbol{\theta})\nabla\Psi(\boldsymbol{\theta};\rho_t)\right], \tag{B.3}$$

$$\Psi(\boldsymbol{\theta};\rho) = \frac{\delta R(\rho)}{\delta\rho(\boldsymbol{\theta})} = \mathbb{E}\left\{\partial_2\ell\left(y, \int \sigma_*(\boldsymbol{x};\bar{\boldsymbol{\theta}})\,\rho(\mathrm{d}\bar{\boldsymbol{\theta}})\right)\,\sigma_*(\boldsymbol{x};\boldsymbol{\theta})\right\}, \tag{B.4}$$

where $\partial_2\ell$ denotes the derivative of ℓ with respect to its second argument. It is immediate to see that, for the quadratic loss $\ell(y,\hat{y}) = (y-\hat{y})^2$, we recover the expressions used in the rest of the paper.

References

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré, *Gradient flows: in metric spaces and in the space of probability measures*, Springer Science & Business Media, 2008.

[CDF+11] José A Carrillo, Marco DiFrancesco, Alessio Figalli, Thomas Laurent, Dejan Slepčev, et al., *Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations*, Duke Mathematical Journal **156** (2011), no. 2, 229–271.

[CMV+03] José A Carrillo, Robert J McCann, Cédric Villani, et al., *Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates*, Revista Matematica Iberoamericana **19** (2003), no. 3, 971–1018.

[CMV06] José A Carrillo, Robert J McCann, and Cédric Villani, *Contractions in the 2-wasserstein length space and thermalization of granular media*, Archive for Rational Mechanics and Analysis **179** (2006), no. 2, 217–263.

[Eva09] Lawrence C. Evans, *Partial differential equations*, Springer, 2009.

[GP10] Victor Guillemin and Alan Pollack, *Differential topology*, vol. 370, American Mathematical Soc., 2010.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto, *The variational formulation of the fokker–planck equation*, SIAM journal on mathematical analysis **29** (1998), no. 1, 1–17.

[Lan13] Serge Lang, *Complex analysis*, vol. 103, Springer Science & Business Media, 2013.

[LSU88] Olga Aleksandrovna Ladyzhenskaia, Vsevolod Alekseevich Solonnikov, and Nina N Ural'tseva, *Linear and quasi-linear equations of parabolic type*, vol. 23, American Mathematical Soc., 1988.

[MBM16] Song Mei, Yu Bai, and Andrea Montanari, *The landscape of empirical risk for non-convex losses*, arXiv:1607.06534 (2016).

[Mit15] Boris Mityagin, *The zero set of a real analytic function*, arXiv:1512.07276 (2015).

[MV00] Peter A Markowich and Cédric Villani, *On the trend to equilibrium for the fokker-planck equation: an interplay between physics and functional analysis*, Mat. Contemp **19** (2000), 1–29.

[San15] Filippo Santambrogio, *Optimal transport for applied mathematicians: Calculus of variations, pdes, and modeling*, vol. 87, Birkhäuser, 2015.

[Szn91] Alain-Sol Sznitman, *Topics in propagation of chaos*, Ecole d'été de probabilités de Saint-Flour XIX—1989, Springer, 1991, pp. 165–251.