C. Ma, L. Ying *Res Math Sci* (2022)9:47
https://doi.org/10.1007/s40687-022-00344-0

Research in
the Mathematical Sciences

RESEARCH

A Riemannian mean field formulation for two-layer neural networks with batch normalization

Chao Ma* and Lexing Ying

*Correspondence:
chaoma@stanford.edu
Department of Mathematics,
Stanford University, Building 380,
Stanford, CA 94305, USA

Abstract

The training dynamics of two-layer neural networks with batch normalization (BN) is studied. It is written as the training dynamics of a neural network without BN on a Riemannian manifold. Therefore, we identify BN's effect of changing the metric in the parameter space. Later, the infinite-width limit of the two-layer neural networks with BN is considered, and a mean-field formulation is derived for the training dynamics. The training dynamics of the mean-field formulation is shown to be the Wasserstein gradient flow on the manifold. Theoretical analysis is provided on the well-posedness and convergence of the Wasserstein gradient flow.

Keywords: Neural networks, Batch normalization, Mean field formulation, Riemannian manifold

1 Introduction

Batch normalization (BN) [8] is a technique that greatly helps the training of deep neural networks. It is used in almost every neural network model in real-world applications. Given the benefit of batch normalization in practice, efforts are invested to explain the essential reason of its success. Existing works address this issue from many different perspectives, including covariate shift, landscape smoothing, length-direction decoupling, and learning rate adaptivity, etc. [2,8,10,16].

In this work, we provide a metric point of view to the effect of BN on the optimization dynamics of neural networks. We show that the gradient descent dynamics of a two-layer neural network with batch normalization can be written as the gradient descent of a two-layer neural network without BN (with the same width) on a Riemannian manifold. Hence, batch normalization changes the metric of the parameter space. We explicitly write down the manifold and the metric of the Riemannian manifold. Next, we consider the infinite-width limit, i.e., the continuous formulation, of the model, and derive a mean-field formulation for the training dynamics of the BN model by gradient flow. The mean-field dynamics is naturally a Wasserstein gradient flow on the Riemannian manifold. Adopting techniques from previous works [5], under appropriate conditions we show the existence of the solution and the global optimality of convergent solutions.

Finally, with the Riemannian manifold understanding of BN's effect, we identify several potential benefits of BN on the training of neural networks. First, models with batch normalization can adjust the speed of neurons according to the alignment of the neurons' directions with the data distribution, which helps neurons find significant directions quickly. Second, BN models assign different speeds to neurons with different magnitudes. Hence, BN models with diverse neuron length can explore an ensemble of learning rates. Finally, we also identify a first-step amplification effect that guarantees the well-behavedness of the BN model under discrete time gradient descent dynamics.

As a summary, our main contributions are as follows:

1. We show that the gradient flow dynamics of two-layer neural networks with BN is equivalent with the gradient flow dynamics of vanilla two-layer neural networks on a Riemannian manifold. Hence, batch normalization changes the metric of the parameter space.
2. In the infinite-width limit, we derive a mean-field formulation for the training dynamics of BN models. The mean-field formulation is a Wasserstein gradient flow on the Riemannian manifold. We analyze the existence and convergence of the dynamics.
3. With the Riemannian manifold understanding, we identify and discuss several special features of the training dynamics introduced by batch normalization.

2 Gradient flow on the Riemannian manifold

2.1 Preliminaries

In this paper, we consider two-layer neural networks with batch normalization before the activation function:

$$f(\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{m}\sum_{k=1}^{m} a_k \sigma(\mathrm{BN}(\mathbf{b}_k^T \mathbf{x})), \tag{1}$$

where we suppose $\mathbf{x}, \mathbf{b}_k \in \mathbb{R}^d$ and $a_k \in \mathbb{R}$, and $\boldsymbol{\theta}$ is a parameter vector containing all the entries in $(a_k, \mathbf{b}_k)_{k=1}^m$. Let μ be the probability distribution from which the data $\mathbf{x}$ are sampled. Then, for any $\mathbf{b} \in \mathbb{R}^d$ we consider the population batch normalization $\mathrm{BN}(\mathbf{b}^T\mathbf{x})$ defined as

$$\mathrm{BN}(\mathbf{b}^T\mathbf{x}) = \frac{\mathbf{b}^T\mathbf{x} - \mathbb{E}_{\mathbf{x}\sim\mu}\mathbf{b}^T\mathbf{x}}{\sqrt{\mathrm{Var}_{\mathbf{x}\sim\mu}[\mathbf{b}^T\mathbf{x}]}}, \tag{2}$$

For the ease of analysis, we make the following assumptions on the data distribution μ.

Assumption 1 Let $\Sigma = \mathbb{E}_{\mathbf{x}\sim\mu}\mathbf{x}\mathbf{x}^T$. Assume $\mathbb{E}_{\mathbf{x}\sim\mu}\mathbf{x} = 0$ and $\Sigma \succ 0$.

The assumption above is reasonable considering that data are usually normalized before being fed into the neural networks. Then, (2) can be written as

$$\mathrm{BN}(\mathbf{b}^T\mathbf{x}) = \frac{\mathbf{b}^T\mathbf{x}}{\sqrt{\mathbf{b}^T\Sigma\mathbf{b}^T}}. \tag{3}$$

Let $\|\mathbf{b}\|_\Sigma = \sqrt{\mathbf{b}^T\Sigma\mathbf{b}^T}$ and $\bar{\mathbf{b}} = \mathbf{b}/\|\mathbf{b}\|_\Sigma$, we have $\mathrm{BN}(\mathbf{b}^T\mathbf{x}) = \bar{\mathbf{b}}^T\mathbf{x}$, and hence we can write (1) as

$$f(\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{m}\sum_{k=1}^{m} a_k \sigma(\bar{\mathbf{b}}_k^T \mathbf{x}). \tag{4}$$

Next, consider a supervised learning problem with input-target pairs $(\mathbf{x}, y)$ sampled i.i.d. from probability distribution $\mathbb{P}$. Note that the marginal distribution of $\mathbb{P}$ on $\mathbf{x}$ is μ. Let

$l(\cdot,\cdot)$ be a risk function which is twice differentiable. Then, using the BN model (4) to learn the supervised learning problem asks to minimize the following loss function:

$$L(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}}\, l\left(f(\mathbf{x},\boldsymbol{\theta}), y\right). \tag{5}$$

The minimization is achieved by some optimization algorithms. In this paper, we study the gradient flow (the zero learning rate limit of the gradient descent algorithm), which is given by the following ODEs:

$$\begin{aligned}
\dot{a}_k &= -\frac{\partial L(\theta)}{\partial a_k} = \frac{1}{m}\mathbb{E}l'(f(\mathbf{x},\boldsymbol{\theta}), y)\sigma(\bar{\mathbf{b}}_k^T\mathbf{x}),\\
\dot{\mathbf{b}}_k &= -\frac{\partial L(\theta)}{\partial \mathbf{b}_k} = \frac{1}{m}\mathbb{E}l'(f(\mathbf{x},\boldsymbol{\theta}), y)a\sigma'(\bar{\mathbf{b}}_k^T\mathbf{x})\frac{1}{\|\mathbf{b}_k\|_\Sigma}\left(I - \frac{\Sigma\mathbf{b}_k\mathbf{b}_k^T}{\|\mathbf{b}_k\|_\Sigma^2}\right)\mathbf{x}.
\end{aligned} \tag{6}$$

In (6) and below, we drop the subscript of the expectation and use $\mathbb{E}$ to denote the expectation over $\mathbb{P}$. With the dynamics above, it is easy to show that $\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{b}_k\|^2 = 2\mathbf{b}_k^T\dot{\mathbf{b}}_k = 0$. Hence, the norm of $\mathbf{b}_k$ does not change during training. By the formulation of batch normalization, $\|\mathbf{b}_k\|$ does not influence the function implemented by the model, too. Therefore, without loss of generality, we can assume $\mathbf{b}_k \in \mathbb{S}^{d-1}$ for any $k = 1, 2, \ldots, m$ and at any time, where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$.

2.2 Gradient flow on the Riemannian manifold

By Eq. (4), the BN model represents the same function as a vanilla two-layer neural networks with $(a_k, \bar{\mathbf{b}}_k)$ as parameters. Let $f_0(\mathbf{x},\boldsymbol{\theta})$ be the vanilla model

$$f_0(\mathbf{x},\boldsymbol{\theta}) = \frac{1}{m}\sum_{k=1}^m a_k\sigma(\mathbf{b}_k\mathbf{x}),$$

and let $\bar{\boldsymbol{\theta}}$ be the set of parameters containing $(a_k, \bar{\mathbf{b}}_k)_{k=1}^m$. Then, we have $f(\mathbf{x},\boldsymbol{\theta}) = f_0(\mathbf{x},\bar{\boldsymbol{\theta}})$. Yet, when training is considered, the dynamics of parameters are different. If we view the BN model as a vanilla model using the equivalence above, the dynamics of $\bar{\theta}$ is induced by the gradient flow for the BN model as described in (6). We have

$$\begin{aligned}
\dot{a}_k &= \frac{1}{m}\mathbb{E}l'(f_0(\mathbf{x},\bar{\boldsymbol{\theta}}), y)\sigma(\bar{\mathbf{b}}_k^T\mathbf{x}),\\
\dot{\bar{\mathbf{b}}}_k &= -\frac{1}{m}\|\bar{\mathbf{b}}_k\|^2(I - \bar{\mathbf{b}}_k\bar{\mathbf{b}}_k^T\Sigma)(I - \Sigma\bar{\mathbf{b}}_k\bar{\mathbf{b}}_k^T)\mathbb{E}l'(f_0(\mathbf{x},\bar{\boldsymbol{\theta}}), y)a_k\sigma'(\bar{\mathbf{b}}_k^T\mathbf{x})\mathbf{x}.
\end{aligned} \tag{7}$$

Recall that the gradient flow for the vanilla non-BN model at $\bar{\boldsymbol{\theta}}$ is

$$\begin{aligned}
\dot{a}_k &= \frac{1}{m}\mathbb{E}l'(f_0(\mathbf{x},\bar{\boldsymbol{\theta}}), y)\sigma(\bar{\mathbf{b}}_k^T\mathbf{x}),\\
\dot{\bar{\mathbf{b}}}_k &= -\frac{1}{m}\mathbb{E}l'(f_0(\mathbf{x},\bar{\boldsymbol{\theta}}), y)a_k\sigma'(\bar{\mathbf{b}}_k^T\mathbf{x})\mathbf{x}.
\end{aligned} \tag{8}$$

We see that the two dynamics (7) and (8) differ at the term $\|\bar{\mathbf{b}}_k\|^2(I - \bar{\mathbf{b}}_k\bar{\mathbf{b}}_k^T\Sigma)(I - \Sigma\bar{\mathbf{b}}_k\bar{\mathbf{b}}_k^T)$, which depends on the location of $\bar{\mathbf{b}}$ and the data covariance Σ. In the following, we show that this term appears if we consider the gradient flow of the non-BN model on a special Riemannian manifold. This means the GF dynamics of the BN model in the Euclidean space is equivalent with the GF dynamics of the non-BN model on a manifold.

To see this, first note that for any $k = 1, 2, \ldots, m$ we always have $\|\bar{\mathbf{b}}_k\|_\Sigma = 1$. Let $\Omega = \{\mathbf{b} : \|\mathbf{b}\|_\Sigma = 1\}$ and $\mathcal{M} = \mathbb{R} \times \Omega$. The matrix

$$G_{(a,\mathbf{b})} = \begin{bmatrix} 1 & 0 \\ 0 & G_{\mathbf{b}} \end{bmatrix} \text{ with } G_{\mathbf{b}} = \frac{1}{\|\mathbf{b}\|^2}\left(I - \frac{\Sigma\mathbf{b}\mathbf{b}^T\Sigma}{\mathbf{b}^T\Sigma^2\mathbf{b}}\right)[(I - \mathbf{b}\mathbf{b}^T\Sigma)(I - \Sigma\mathbf{b}\mathbf{b}^T)]^\dagger \in \mathbb{R}^{d\times d}$$

induces a Riemannian metric on $\mathcal{M}$.

Proposition 1 *Let $T_{(a,\mathbf{b})}\mathcal{M}$ be the tangent space of $\mathcal{M}$ at $(a, \mathbf{b})$. For any $(a, \mathbf{b})$, define function $h : (T_{(a,\mathbf{b})}\mathcal{M})^2 \to \mathbb{R} :\ h(\alpha, \beta) = \alpha^T G_{(a,\mathbf{b})}\beta$, where α and β are treated as vectors in $\mathbb{R}^{d+1}$. Then, h is an inner product on $T_{(a,\mathbf{b})}\mathcal{M}$.*

Proof Obviously, the metric along the direction of a is the standard metric. Hence, we only need to show the results for $\mathbf{b}$. Let $T_{\mathbf{b}}\Omega$ be the tangent space of Ω at $\mathbf{b}$. With an abuse of notation, let $\alpha, \beta \in T_{\mathbf{b}}\Omega$. Viewing α and β as vectors in $\mathbb{R}^d$, we have

$$\alpha^T G_{\mathbf{b}}\beta = \frac{1}{\|\mathbf{b}\|^2}\alpha^T\left(I - \frac{\Sigma\mathbf{b}\mathbf{b}^T\Sigma}{\mathbf{b}^T\Sigma^2\mathbf{b}}\right)[(I - \mathbf{b}\mathbf{b}^T\Sigma)(I - \Sigma\mathbf{b}\mathbf{b}^T)]^\dagger\beta.$$

Since the tangent space can be written as $T_{\mathbf{b}}\Omega = \{\alpha :\ \alpha^T\Sigma\mathbf{b} = 0\}$, we have

$$\alpha^T\left(I - \frac{\Sigma\mathbf{b}\mathbf{b}^T\Sigma}{\mathbf{b}^T\Sigma^2\mathbf{b}}\right) = \alpha.$$

Therefore,

$$\alpha^T G_{\mathbf{b}}\beta = \frac{1}{\|\mathbf{b}\|^2}\alpha^T[(I - \mathbf{b}\mathbf{b}^T\Sigma)(I - \Sigma\mathbf{b}\mathbf{b}^T)]^\dagger\beta,$$

and we directly have $\alpha^T G_{\mathbf{b}}\beta = \beta^T G_{\mathbf{b}}\alpha$.

Next, we show h is positive definite. First, it is easy to show that $[(I - \mathbf{b}\mathbf{b}^T\Sigma)(I - \Sigma\mathbf{b}\mathbf{b}^T)]^\dagger$ is positive semi-definite. Second, since the pseudo-inverse of a symmetric matrix has the same 0-eigenspace as the original matrix, if $\alpha^T[(I - \mathbf{b}\mathbf{b}^T\Sigma)(I - \Sigma\mathbf{b}\mathbf{b}^T)]^\dagger\alpha = 0$ holds for some α, then we must have $(I - \Sigma\mathbf{b}\mathbf{b}^T)\alpha = 0$. Recall that $\alpha^T\Sigma\mathbf{b} = 0$, we then have

$$0 = \alpha^T(I - \Sigma\mathbf{b}\mathbf{b}^T)\alpha = \|\alpha\|^2 + \alpha^T\Sigma\mathbf{b}\mathbf{b}^T\alpha = \|\alpha\|^2.$$

Therefore, h is strictly positive definite on $T_{\mathbf{b}}\Omega$, which completes the proof. □

In the following, $\mathcal{M}$ is always assumed to be the Riemannian manifold endowed with the metric $G_{(a,\mathbf{b})}$. The following theorem shows that, starting from the same initialization, the gradient flow dynamics of BN model in the Euclidean space equals to a corresponding non-BN model on $\mathcal{M}$.

Theorem 2 *Let $f_0(\mathbf{x}, \bar{\boldsymbol{\theta}}) = \frac{1}{m}\sum_{k=1}^{m} a_k\sigma(\bar{\mathbf{b}}_k\mathbf{x})$. The manifold gradient of $L_0(\bar{\theta}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}} l(f_0(\mathbf{x}, \bar{\boldsymbol{\theta}}), y)$ on $\mathcal{M}$ with respect $\bar{\boldsymbol{\theta}}$ is*

$$\frac{\partial L_0(\bar{\boldsymbol{\theta}})}{\partial a_k} = -\frac{1}{m}\mathbb{E}l'(f_0(\mathbf{x}, \bar{\boldsymbol{\theta}}), y)\sigma(\bar{\mathbf{b}}_k^T\mathbf{x}),$$

$$\frac{\partial L_0(\bar{\boldsymbol{\theta}})}{\partial \bar{\mathbf{b}}_k} = -\frac{1}{m}\|\bar{\mathbf{b}}_k\|^2(I - \bar{\mathbf{b}}_k\bar{\mathbf{b}}_k^T\Sigma)(I - \Sigma\bar{\mathbf{b}}_k\bar{\mathbf{b}}_k^T)\mathbb{E}l'(f_0(\mathbf{x}, \bar{\boldsymbol{\theta}}), y)a_k\sigma'(\bar{\mathbf{b}}_k^T\mathbf{x})\mathbf{x}.$$

Proof Again, we focus on the $\bar{\mathbf{b}}$ part. For any $k = 1, 2, \ldots, m$, the gradient of L_0 with respect to $\bar{\mathbf{b}}_k$ under Euclidean metric is

$$\partial_{\bar{\mathbf{b}}_k} L_0(\bar{\boldsymbol{\theta}}) = \mathbb{E}l'(f_0(\mathbf{x}, \bar{\boldsymbol{\theta}}), y) a_k \sigma'(\bar{\mathbf{b}}_k^T \mathbf{x})\mathbf{x}.$$

Let $\mathcal{P}_{\bar{\mathbf{b}}}$ be the orthogonal projection matrix onto $T_{\bar{\mathbf{b}}}\Omega$, and $\partial_{m,\bar{\mathbf{b}}_k} L_0(\bar{\boldsymbol{\theta}})$ be the manifold gradient. Then, we have

$$\mathcal{P}_{\bar{\mathbf{b}}} = I - \frac{\Sigma \bar{\mathbf{b}}\bar{\mathbf{b}}^T \Sigma}{\bar{\mathbf{b}}^T \Sigma^2 \bar{\mathbf{b}}},$$

and the condition for the manifold gradient:

$$\mathbf{v}^T G_{\bar{\mathbf{b}}_k} \partial_{m,\bar{\mathbf{b}}_k} L_0(\bar{\boldsymbol{\theta}}) = \mathbf{v}^T P_{\bar{\mathbf{b}}_k} \partial_{\bar{\mathbf{b}}_k} L_0(\bar{\boldsymbol{\theta}}),$$

where $\mathbf{v} \in \mathbb{R}^d$ is an arbitrary vector. Therefore, we have

$$\partial_{m,\bar{\mathbf{b}}_k} L_0(\bar{\boldsymbol{\theta}}) = G_{\bar{\mathbf{b}}_k}^{\dagger} P_{\bar{\mathbf{b}}_k} \partial_{\bar{\mathbf{b}}_k} L_0(\bar{\boldsymbol{\theta}}).$$

Substituting

$$G_{\bar{\mathbf{b}}_k} = \frac{1}{\|\bar{\mathbf{b}}_k\|^2} \left(I - \frac{\Sigma \bar{\mathbf{b}}_k \bar{\mathbf{b}}_k^T \Sigma}{\bar{\mathbf{b}}_k^T \Sigma^2 \bar{\mathbf{b}}_k} \right) [(I - \bar{\mathbf{b}}_k \bar{\mathbf{b}}_k^T \Sigma)(I - \Sigma \bar{\mathbf{b}}_k \bar{\mathbf{b}}_k^T)]^{\dagger} \in \mathbb{R}^{d \times d}$$

into the above equation gives

$$\partial_{m,\bar{\mathbf{b}}_k} L_0(\bar{\boldsymbol{\theta}}) = \|\bar{\mathbf{b}}_k\|^2 (I - \bar{\mathbf{b}}_k \bar{\mathbf{b}}_k^T \Sigma)(I - \Sigma \bar{\mathbf{b}}_k \bar{\mathbf{b}}_k^T) \mathbb{E}l'(f_0(\mathbf{x}, \bar{\boldsymbol{\theta}}), y) a_k \sigma'(\bar{\mathbf{b}}_k^T \mathbf{x})\mathbf{x}.$$

□

3 The continuous formulation and the large width limit

3.1 The continuous formulation

In this section, we consider the limiting model when the width of the two-layer neural networks tends to infinity. In this case, a conventional way to represent the model is by integral transformation [5,12,18], and the parameters become a probability distribution on the parameter space. Specifically, we consider the model

$$f(\mathbf{x}, \rho) = \int_{\mathbb{R}\times\mathbb{R}^d} a\sigma(\mathrm{BN}(\mathbf{b}^T\mathbf{x}))\rho(\mathrm{d}a, \mathrm{d}\mathbf{b}) = \int_{\mathbb{R}\times\mathbb{R}^d} a\sigma(\bar{\mathbf{b}}^T\mathbf{x})\rho(\mathrm{d}a, \mathrm{d}\mathbf{b}), \tag{9}$$

where ρ is a probability distribution on $\mathbb{R} \times \mathbb{S}^{d-1}$. Note that this formulation can also represent networks with finite width by taking ρ as empirical distributions.

Still consider the data distribution $\mathbb{P}$, the loss function is

$$L(\rho) = \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}} l(f(\mathbf{x}, \rho), y).$$

By (6), following the gradient flow, the velocity field of any particle $(a, \mathbf{b})$ is

$$v_t(a, \mathbf{b}) = - \begin{bmatrix} \mathbb{E}l'(f(\mathbf{x}, \rho), y)\sigma(\bar{\mathbf{b}}^T\mathbf{x}) \\ \mathbb{E}l'(f(\mathbf{x}, \rho), y) a\sigma'(\bar{\mathbf{b}}^T\mathbf{x}) \frac{1}{\|\mathbf{b}\|_\Sigma} \left(I - \frac{\Sigma\mathbf{b}\mathbf{b}^T}{\|\mathbf{b}\|_\Sigma^2} \right) \mathbf{x} \end{bmatrix}, \tag{10}$$

and thus $(\rho_t)_{t\geq 0}$ as a time series of probability distributions satisfies the following continuity equation:

$$\partial_t \rho_t = -\mathrm{div}(\rho_t v_t). \tag{11}$$

Because of the batch normalization, any solution $(\rho_t)_{t\geq 0}$ is always supported on $\mathbb{R} \times \mathbb{S}^{d-1}$.

Similar to the finite width case, we want to write the above dynamics for the BN model as a Wasserstein gradient flow on a Riemannian manifold of a non-BN model. To achieve this, we still consider the manifold $\mathcal{M}$ defined in Sect. 2. Let $T : \mathbb{S}^{d-1} \to \Omega$ be the "normalization map" to Ω defined as $T\mathbf{b} = \bar{\mathbf{b}} = \frac{\mathbf{b}}{\|\mathbf{b}\|_\Sigma}$. Let $\bar{\rho}$ be the pushforward of ρ by $\mathrm{id} \times T$, defined as

$$\bar{\rho}(A \times B) = \rho(A \times T^{-1}B) \tag{12}$$

for any Borel measurable set $A \in \mathbb{R}$ and $B \in \Omega$. Then, we easily have

$$f(\mathbf{x}, \rho) = f_0(\mathbf{x}, \bar{\rho}),$$

where $f_0(\mathbf{x}, \bar{\rho})$ is the non-BN infinite-width neural network model with parameter distribution $\bar{\rho}$:

$$f_0(\mathbf{x}, \bar{\rho}) = \int_{\mathcal{M}} a\sigma(\bar{\mathbf{b}}^T\mathbf{x})\mathrm{d}\bar{\rho}(a, \bar{\mathbf{b}}).$$

Note that the above integral is evaluated on $\mathcal{M}$. Later when $\bar{\rho}$ is understood as a density function (e.g. in (13)), the measure is evaluated by integrating the density function with the volume form on $\mathcal{M}$. Now, let $(\bar{\rho}_t)_{t\geq 0}$ be the series of distributions induced by $(\rho_t)_{t\geq 0}$ that satisfies $f(\mathbf{x}, \rho_t) = f_0(\mathbf{x}, \bar{\rho}_t)$ for any $t \geq 0$. Then, using the dynamics of $(a, \bar{\mathbf{b}})$ derive in (7), $(\bar{\rho}_t)$ satisfies the following continuity equation:

$$\partial_t \bar{\rho}_t = -\mathrm{div}_m(\bar{\rho}_t \bar{v}_t), \tag{13}$$

where div_m is the divergence operator on the manifold, and the velocity field $\bar{v}_t$ is given by

$$\bar{v}_t(a, \bar{\mathbf{b}}) = -\begin{bmatrix} \mathbb{E}l'(f_0(\mathbf{x}, \bar{\rho}_t), y)\sigma(\bar{\mathbf{b}}^T\mathbf{x}) \\ \|\bar{\mathbf{b}}\|^2(I - \bar{\mathbf{b}}\bar{\mathbf{b}}^T\Sigma)(I - \Sigma\bar{\mathbf{b}}\bar{\mathbf{b}}^T)\mathbb{E}l'(f_0(\mathbf{x}, \bar{\rho}_t), y)a\sigma'(\bar{\mathbf{b}}^T\mathbf{x})\mathbf{x} \end{bmatrix}. \tag{14}$$

Similar to the derivations in Sect. 2, the velocity field above is the manifold gradient of $L_0(\bar{\rho}_t) := \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}}l(f_0(\mathbf{x}, \bar{\rho}_t), y)$ (which is the same as $L(\rho_t)$) on $\mathcal{M}$, at the point $(a, \bar{\mathbf{b}})$. Hence, the continuity equation (13) can be written as

$$\partial_t \bar{\rho}_t = \nabla_m \cdot \left(\bar{\rho}_t \nabla_m \frac{\delta L_0(\bar{\rho}_t)}{\delta \bar{\rho}_t}\right). \tag{15}$$

Let $\mathrm{dist}(\cdot, \cdot) : \mathcal{M}^2 \to \mathbb{R}_+$ be the metric function on $\mathcal{M}$, and $\mathcal{P}(\mathcal{M})$ be the set of probability distributions on $\mathcal{M}$ with finite second moment. The 2-Wasserstein distance between any pair of $\bar{\rho}_1, \bar{\rho}_2 \in \mathcal{P}(\mathcal{M})$ is defined as

$$W_2(\rho_1, \rho_2) = \left(\inf_{\gamma\in\Gamma(\rho_1,\rho_2)} \int_{\mathcal{M}\times\mathcal{M}} \mathrm{dist}(\mathbf{x}_1, \mathbf{x}_2)^2 \mathrm{d}\gamma(\mathbf{x}_1, \mathbf{x}_2)\right)^{\frac{1}{2}}, \tag{16}$$

where $\Gamma(\rho_1, \rho_2)$ contains all measures on $\mathcal{M} \times \mathcal{M}$ whose marginals with respect to $\mathbf{x}_1$ and $\mathbf{x}_2$ are ρ_1 and ρ_2, respectively. Then, $\nabla_m \frac{\delta L_0(\bar{\rho}_t)}{\delta \bar{\rho}_t}$ is the Wasserstein differential of L_0 at $\bar{\rho}_t$. Therefore, (15) is the Wasserstein gradient flow with respect to the Wasserstein metric on the Riemannian manifold $\mathcal{M}$. We summarize the result as the following theorem:

Theorem 3 *Let $(\rho_t)_{t\geq 0}$ be the trajectory of probability distributions obtained by learning the BN model (9) (minimizing the loss $L(\rho)$) using gradient flow in the Euclidean space. Let $(\bar{\rho}_t)_{t\geq 0}$ be the "normalized" trajectories of (ρ_t) supported on $\mathcal{M}$. Then, $(\bar{\rho}_t)$ satisfies the Wasserstein gradient flow of $L_0(\bar{\rho})$ on $\mathcal{M}$.*

3.2 Convergence to the infinite-width limit

In this part, we show that when the width tends to infinity, the empirical distribution given by the finite parameters tends to a solution of (15). This also shows the existence of the solution of (15). Specifically, a model with m neurons with parameters on $\mathcal{M}$ is

$$\bar{f}(\mathbf{x}, \bar{\rho}_m) = \int_{\mathcal{M}} a\sigma(\bar{\mathbf{b}}^T\mathbf{x})\mathrm{d}\bar{\rho}_m(a, \bar{\mathbf{b}}) = \frac{1}{m}\sum_{k=1}^{m} a_k\sigma(\bar{\mathbf{b}}_k^T\mathbf{x}), \tag{17}$$

where $\bar{\rho}_m$ is the empirical distribution $\frac{1}{m}\sum_{k=1}^{m}\delta_{a_k}(a)\delta_{\bar{\mathbf{b}}_k}(\bar{\mathbf{b}})$. Naturally, the dynamics of the $\bar{\rho}_{m,t}$ follows the same continuity equation (15) as the infinite width case, i.e., $(\bar{\rho}_{m,t})$ solves the Wasserstein gradient flow of $L_0(\bar{\rho})$ starting from $\bar{\rho}_{m,0}$.

Under appropriate assumptions, as $m \to \infty$, if $\bar{\rho}_{m,0}$ tends to some limiting distribution $\bar{\rho}_0$, then we can find a subsequence of trajectories $\bar{\rho}_{m,t}$ converging to the solution of (15) initialized from $\bar{\rho}_0$. The result is stated in Theorem 4. First, we make the following assumptions:

Assumption 2 Assume:

- There exists a constant $C_{\mathbf{x}}$ such that for any input data $\mathbf{x}$ we have $\|\mathbf{x}\| \le C_{\mathbf{x}}$.
- The activation function is L_σ-Lipschitz.
- The derivative of the loss function, l', is $L_{l'}$-Lipschitz.

Theorem 4 *Let $(\bar{\rho}_{m,t})_{m=1}^{\infty}$ be the trajectories of empirical distributions generated by the gradient flow of parameters of models with different widths m. Let $\mathcal{M}_r = [-r, r] \times \Omega$, and assume that any $\bar{\rho}_{m,0}$ is supported in $\mathcal{M}_{r_0}$ for some $r_0 > 0$. If there exists $\bar{\rho}_0 \in \mathcal{P}(\mathcal{M})$ such that $\bar{\rho}_{m,0} \to \bar{\rho}_0$ weakly as $m \to \infty$, then there exists a subsequence of $(\bar{\rho}_{m,t})_{m=1}^{\infty}$, denoted still by $(\bar{\rho}_{m,t})_{m=1}^{\infty}$, and a trajectory $\bar{\rho}_t \in [0, \infty) \times \mathcal{P}(\mathcal{M})$ which solves (15) starting from $\bar{\rho}_0$, that satisfies $\bar{\rho}_{m,t}$ converges weakly to $\bar{\rho}_t$ for any $t > 0$.*

Proof First, by the definition of $\mathcal{M}$, we know that $\|\bar{\mathbf{b}}\|$ is always bounded. Let $C_{\mathbf{b}}$ be an upper bound for $\mathbf{b}$ that satisfies $\|\bar{\mathbf{b}}\| \le C_{\mathbf{b}}$.

Our proof takes similar path like [5]. For any $r > 0$, let t_r be the first time that some particles represented by some $\bar{\rho}_{m,t}$ goes out of $\mathcal{M}_r$, i.e.,

$$t_r := \inf\{t > 0, \exists m \in \mathbb{N}, \bar{\rho}_{m,t}(\mathcal{M}_r) < 1\}.$$

We first show that $t_r > 0$ for any $r > r_0$, and $\lim_{r\to\infty} t_r = \infty$.

To show $t_r > 0$, note that the $\bar{\mathbf{b}}$ of any particle always moves on Ω. Hence, we only need to consider a. For fixed m and some $1 \le k \le m$, let $(a_k, \bar{\mathbf{b}}_k)$ be the k-th particle of $\bar{\rho}_{m,t}$. Recall the dynamics of a_k:

$$\dot{a}_k = \mathbb{E}_{\mathbf{x}\sim\mu} l'(\bar{f}(\mathbf{x}, \bar{\rho}_{m,t}))\sigma(\bar{\mathbf{b}}_k^T\mathbf{x}).$$

By Assumption 2, for any $t < t_r$, we have

$$\begin{aligned}
|\dot{a}_k| &\le \mathbb{E}_{\mathbf{x}\sim\mu}|l'(\bar{f}(\mathbf{x}, \bar{\rho}_{m,t}))\sigma(\bar{\mathbf{b}}_k^T\mathbf{x})| \\
&\le \mathbb{E}_{\mathbf{x}\sim\mu}(|l'(0)| + L_{l'}|\bar{f}(\mathbf{x}, \bar{\rho}_{m,t})|)(|\sigma(0)| + L_\sigma C_{\mathbf{b}} C_{\mathbf{x}}) \\
&\le \left(|l'(0)| + L_{l'}\max_{\mathbf{x}\sim\mu}|\bar{f}(\mathbf{x}, \bar{\rho}_{m,t})|\right)(|\sigma(0)| + L_\sigma C_{\mathbf{b}} C_{\mathbf{x}}).
\end{aligned}$$

Considering

$$\bar{f}(\mathbf{x}, \bar{\rho}_{m,t}) = \int a\sigma(\bar{\mathbf{b}}^T\mathbf{x})\mathrm{d}\bar{\rho}_{m,t},$$

we have for any $\mathbf{x} \sim \mu$,

$$|\bar{f}(\mathbf{x}, \bar{\rho}_{m,t})| \leq r\left(|\sigma(0)| + L_\sigma C_{\mathbf{b}} C_{\mathbf{x}}\right).$$

Therefore, come back to the dynamics of a, we have

$$|\dot{a}_k| \leq \left(|l'(0)| + L_{l'} r\left(|\sigma(0)| + L_\sigma C_{\mathbf{b}} C_{\mathbf{x}}\right)\right)\left(|\sigma(0)| + L_\sigma C_{\mathbf{b}} C_{\mathbf{x}}\right).$$

This mean there exist two constants A, B that satisfies

$$|\dot{a}_k| \leq A + Br.$$

Consequently, as long as $\bar{\rho}_{m,t}$ is supported on $\mathcal{M}_r$, the speed of a_k is bounded. This indicates

$$t_r \geq \frac{r - r_0}{A + Br} > 0. \tag{18}$$

To show $\lim_{r\to\infty} t_r = \infty$, we let $r_i = i \cdot r_0$ for $i \in \mathbb{N}$. Then, by similar arguments as (18), from t_{r_i} to $t_{r_{i+1}}$, we have $|\dot{a}_k| \leq A + B(i+1)r_0$, and thus

$$t_{r_{i+1}} - t_{r_i} \geq \frac{(i+1)r_0 - ir_0}{A + B(i+1)r_0} = \frac{r_0}{A + B(i+1)r_0}.$$

Hence,

$$\lim_{r\to\infty} t_r = \lim_{i\to\infty} t_{r_i} \geq \sum_{i=1}^{\infty} \frac{r_0}{A + B(i+1)r_0} = \infty. \tag{19}$$

Next, we show the existence of convergent subsequence of $(\bar{\rho}_{m,t})$. We first show the result in $t \in [0, t_r]$ for finite r by the Arzela–Ascoli theorem, i.e., we show that $(t \to \bar{\rho}_{m,t})$ is equicontinuous and pointwise bounded.

For equicontinuity, we take the proof from [5]. For any $0 \leq t < t' \leq t_r$ and any $m \in \mathbb{N}$, we have

$$\begin{aligned}
W_2(\bar{\rho}_{m,t}, \bar{\rho}_{m,t'})^2 &\leq \frac{1}{m}\sum_{k=1}^{m}\left(|a_k(t') - a_k(t)|^2 + \|\bar{\mathbf{b}}_k(t') - \bar{\mathbf{b}}_k(t)\|^2\right)\\
&\leq \frac{(t'-t)}{m}\int_t^{t'}\sum_{k=1}^{m}\left(|\dot{a}_k|^2 + \|\dot{\bar{\mathbf{b}}}_k\|^2\right)\mathrm{d}t\\
&= (t'-t)\left(\bar{L}(\bar{\rho}_{m,t'}) - \bar{L}(\bar{\rho}_{m,t})\right)\\
&\leq (t'-t)\bar{L}(\bar{\rho}_{m,t'}).
\end{aligned}$$

Since $\bar{\rho}_{m,t'}$ is supported on $\mathcal{M}_r$, the function represented by $\bar{\rho}_{m,t'}$, $f(\mathbf{x}, \bar{\rho}_{m,t'})$ is bounded by $r(\sigma(0) + L_\sigma C_{\mathbf{b}} C_{\mathbf{x}})$. Hence, $\bar{L}(\bar{\rho}_{m,t'})$ is upper bounded by a constant C_r depending on r but independent with m, i.e.,

$$W_2(\bar{\rho}_{m,t}, \bar{\rho}_{m,t'}) \leq \sqrt{C_r(t'-t)}.$$

This shows $(t \to \bar{\rho}_{m,t})$ is equicontinuous (in W_2 metric).

On the other hand, pointwise boundedness follows naturally since all a_k and $\bar{\mathbf{b}}_k$ are uniformly bounded. Therefore, by the Arzela–Ascoli theorem, there exists a subsequence $(t \to \bar{\rho}_{m_j,t})_{j=1}^{\infty}$ and a trajectory $\bar{\rho}_t$ such that $\bar{\rho}_{m_j,t} \to \bar{\rho}_t$ weakly and uniformly for $t \in [0, t_r]$ as $j \to \infty$.

Then, we extend the convergence from $[0, t_r]$ to $[0, \infty)$. Let $r_i = ir_0$. By the analysis above, we can find a subsequence of $(\bar{\rho})_{m,t}$ that converges weakly and uniformly in $t \in$

$[0, t_{2r_0}]$. Denote this subsequence by $(\bar{\rho}_{m_j^2,t})_{j=1}^{\infty}$. Then, still by the same argument, we can find a subsequence in $(\bar{\rho}_{m_j^2,t})_{j=1}^{\infty}$, denoted by $(\bar{\rho}_{m_j^3,t})_{j=1}^{\infty}$, that converges in $[0, t_{3r_0}]$. Repeat this process, for any $i \geq 2$, we can find a sequence $(\bar{\rho}_{m_j^i,t})_{j=1}^{\infty}$ that converges on $[0, t_{ir_0}]$. Moreover, $(\bar{\rho}_{m_j^i,t})_{j=1}^{\infty}$ is a subsequence of $(\bar{\rho}_{m_j^{i-1},t})_{j=1}^{\infty}$. Eventually, by the diagonal trick, it is easy to show that the sequence $(\bar{\rho}_{m_j^{j+1},t})_{j=1}^{\infty}$ converges weakly for any $t \in [0, \infty)$.

Denote the limit obtained above by $\bar{\rho}_t$. For the last step of the proof, we show that $\bar{\rho}_t$ is a solution of the continuity equation (15) starting from $\bar{\rho}_0$. Since the initial condition holds by definition, we only need to check $\bar{\rho}_t$ satisfies the equation weakly, i.e., for any $r > r_0$ and any bounded, Lipschitz test function $\phi(a, \bar{\mathbf{b}}, t)$ defined on $\mathcal{M} \times [0, t_r]$ we have

$$\int_0^{t_r} \int_{\mathcal{M}} (\partial_t \phi + (\nabla_m \phi)^T G \bar{v}_t) \mathrm{d}\bar{\rho}_t \mathrm{d}t = 0,$$

where $\bar{v}_t$ is the velocity field defined in (14) with $\bar{\rho}_t$ and G is the metric matrix $G_{(a,\bar{\mathbf{b}})}$. With an abuse of notation, we use $(\bar{\rho}_{m,t})_{m=1}^{\infty}$ to denote the subsequence that converges to $\bar{\rho}_t$. Let $v_{m,t}$ be the velocity field given by $\bar{\rho}_{m,t}$. Then, for any m, $\bar{\rho}_{m,t}$ satisfies the continuity equation, thus we have

$$\int_0^{t_r} \int_{\mathcal{M}} (\partial_t \phi + (\nabla_m \phi)^T G \bar{v}_{m,t}) \mathrm{d}\bar{\rho}_{m,t} \mathrm{d}t = 0.$$

By the uniform convergence of $(\bar{\rho}_{m,t})$ on time, we have

$$\int_0^{t_r} \int_{\mathcal{M}} \partial_t \phi d\bar{\rho}_{m,t} \mathrm{d}t \to \int_0^{t_r} \int_{\mathcal{M}} \partial_t \phi \mathrm{d}\bar{\rho}_t \mathrm{d}t, \qquad m \to \infty.$$

Therefore, we only need to show

$$\int_0^{t_r} \int_{\mathcal{M}} (\nabla_m \phi)^T G \bar{v}_{m,t} \mathrm{d}\bar{\rho}_{m,t} \mathrm{d}t \to \int_0^{t_r} \int_{\mathcal{M}} (\nabla_m \phi)^T G \bar{v}_t \mathrm{d}\bar{\rho}_t \mathrm{d}t. \tag{20}$$

To proof (20), first notice that for any bounded ψ we have

$$\begin{aligned}
&\left| \int_0^{t_r} \int_{\mathcal{M}} \psi^T G \bar{v}_{m,t} \mathrm{d}\bar{\rho}_{m,t} \mathrm{d}t - \int_0^{t_r} \int_{\mathcal{M}} \psi^T G \bar{v}_t \mathrm{d}\bar{\rho}_t \mathrm{d}t \right| \\
&\quad \leq \left| \int_0^{t_r} \int_{\mathcal{M}} \psi^T G (\bar{v}_{m,t} - \bar{v}_t) \mathrm{d}\bar{\rho}_{m,t} \mathrm{d}t \right| + \left| \int_0^{t_r} \int_{\mathcal{M}} \psi^T G \bar{v}_t (\mathrm{d}\bar{\rho}_{m,t} - \mathrm{d}\bar{\rho}_t) \mathrm{d}t \right| \\
&\quad := I + II.
\end{aligned}$$

For I, since ψ and G are bounded, we show $\bar{v}_{m,t} \to \bar{v}_t$ uniformly for any $(a, \mathbf{b}, t)$. Let $\bar{v}_{t,a}$ and $\bar{v}_{t,\bar{\mathbf{b}}}$ be the a and $\bar{\mathbf{b}}$ component of the velocity field, respectively. $\bar{v}_{m,t,a}$ and $\bar{v}_{m,t,\bar{\mathbf{b}}}$ are similarly defined. Recall the definition of $\bar{v}$, for the a component, we have

$$\begin{aligned}
|\bar{v}_{m,t,a} - \bar{v}_{t,a}| &= \left| \mathbb{E}_{\mathbf{x} \sim \mu} (l'(\bar{f}(\mathbf{x}, \bar{\rho}_{m,t})) - l'(\bar{f}(\mathbf{x}, \bar{\rho}_t))) \sigma(\bar{\mathbf{b}}^T \mathbf{x}) \right| \\
&\leq L_{l'}(\sigma(0) + L_\sigma C_{\mathbf{b}} C_{\mathbf{x}}) \mathbb{E}_{\mathbf{x} \sim \mu} |\bar{f}(\mathbf{x}, \bar{\rho}_{m,t}) - \bar{f}(\mathbf{x}, \bar{\rho}_t)| \\
&\leq L_{l'}(\sigma(0) + L_\sigma C_{\mathbf{b}} C_{\mathbf{x}}) \mathbb{E}_{\mathbf{x} \sim \mu} \left| \int a\sigma(\bar{\mathbf{b}}^T \mathbf{x})(\mathrm{d}\bar{\rho}_{m,t} - \mathrm{d}\bar{\rho}_t) \right| \\
&\leq L_{l'}(\sigma(0) + L_\sigma C_{\mathbf{b}} C_{\mathbf{x}})^2 r \|\bar{\rho}_{m,t} - \bar{\rho}_t\|_{\mathrm{BL}},
\end{aligned}$$

where the last inequality follows from the bound of $a\sigma(\bar{\mathbf{b}}^T \mathbf{x})$ by $r(\sigma(0) + L_\sigma C_{\mathbf{b}} C_{\mathbf{x}})$. Therefore, by the uniform convergence of $\bar{\rho}_{m,t}$ to $\bar{\rho}_t$ for $t \in [0, t_r]$, we know $\bar{v}_{m,t,a}$ converges uniformly to $\bar{v}_{t,a}$. For the $\bar{\mathbf{b}}$ component, similarly we have

$$\begin{aligned}\|\bar{v}_{m,t,\bar{\mathbf{b}}} - \bar{v}_{t,\bar{\mathbf{b}}}\| &\le \|\bar{\mathbf{b}}\|^2 \|(I - \bar{\mathbf{b}}\bar{\mathbf{b}}^T\Sigma)(I - \Sigma\bar{\mathbf{b}}\bar{\mathbf{b}}^T)\| \\ &\quad \cdot \|\mathbb{E}_{\mathbf{x}\sim\mu}(l'(\bar{f}(\mathbf{x},\bar{\rho}_{m,t})) - l'(\bar{f}(\mathbf{x},\bar{\rho}_t)))a\sigma'(\bar{\mathbf{b}}^T\mathbf{x})\mathbf{x}\| \\ &\le C_{\bar{\mathbf{b}}}^2 \|(I - \bar{\mathbf{b}}\bar{\mathbf{b}}^T\Sigma)(I - \Sigma\bar{\mathbf{b}}\bar{\mathbf{b}}^T)\| \\ &\quad \cdot \|L_{l'}(\sigma(0) + L_\sigma C_{\mathbf{b}} C_{\mathbf{x}}) r L_\sigma C_{\mathbf{x}}\| \bar{\rho}_{m,t} - \bar{\rho}_t\|_{\mathrm{BL}}.\end{aligned}$$

Since both $\|\bar{\mathbf{b}}\bar{\mathbf{b}}^T\|$ and $\|\Sigma\|$ are upper bounded, we have $\|(I - \bar{\mathbf{b}}\bar{\mathbf{b}}^T\Sigma)(I - \Sigma\bar{\mathbf{b}}\bar{\mathbf{b}}^T)\|$ is upper bounded. Hence, the $\bar{\mathbf{b}}$ component of the velocity field converges uniformly, too. Combining the results above, we show $\bar{v}_{m,t} \to \bar{v}_t$ uniformly, and thus $I \to 0$.

For II, note that $\bar{v}_t$ is continuous and bounded. Therefore, there exists a constant C such that

$$II \le C \int_0^{t_r} \|\bar{\rho}_{m,t} - \bar{\rho}_t\|_{\mathrm{BL}} \mathrm{d}t \to 0.$$

Combining the results for I and II, we finish the proof of (20) and thus finish the whole proof. □

The theorem above shows the existence of the solution for the continuity equation for any initial distribution with bounded support. If the solution is furthermore unique, then we can show that the sequence $\bar{\rho}_{m,t}$ converges to the unique solution. Here we did not establish uniqueness result. Hence, our theorem only shows the existence of convergent subsequence, and also the limit of any convergent subsequent is a solution of the continuity equation.

3.3 Convergence of the Wasserstein gradient flow

Next, we study the limit of the Wasserstein gradient flow as $t \to \infty$. We show a similar results as in [5], i.e., as long as $\bar{\rho}_t$ given by the Wasserstein gradient flow from $\bar{\rho}_0$ converges in W_2 to a distribution $\bar{\rho}_\infty$, we have $\bar{\rho}_\infty$ is the global minimum of $\bar{L}(\rho)$. Our proof follows the proof in Appendix C of [5], with several differences concerning the Riemannian manifold and the manifold gradient.

Recall that the loss function is $\bar{L}(\bar{\rho}) = \mathbb{E}_{\mathbf{x}\sim\mu} l(\bar{f}(\mathbf{x},\bar{\rho}))$ and $\bar{\rho}_t$, $t \in [0,\infty)$ satisfies the continuity equation (15). Before stating the theorem, we make the following technical assumptions:

Assumption 3 Assume

1. The activation function $\sigma(\cdot)$ is differentiable and $\sigma'(\cdot)$ is $L_{\sigma'}$-Lipschitz continuous.
2. The loss $l(\cdot)$ is convex and differentiable, and $l'(\cdot)$ is bounded on sublevel sets of l. (Recall that l' is also $L_{l'}$-Lipschitz).
3. For any ρ, the set of regular values of $g_\rho(\bar{\mathbf{b}}) := \mathbb{E}_{\mathbf{x}\sim\mu} l'(\bar{f}(\mathbf{x},\rho))\sigma(\bar{\mathbf{b}}^T\mathbf{x})$ as a function $\Omega \to \mathbb{R}$ on the standard metric is dense in its range.

We also assume the initial distribution $\bar{\rho}_0$ is supported on some $\mathcal{M}_{r_0}$ and satisfies the "separation condition" which is also used in [5]: any curve connecting $\{r_0\}\times\Omega$ and $\{-r_0\}\times\Omega$ intersects with the support of $\bar{\rho}_0$. Under these assumptions, we have the following theorem.

Theorem 5 *Let $(\bar{\rho}_t)_{t\geq 0}$ be a solution of the continuity equation (15). Assume Assumption 3 holds. Assume there exists $r_0 > 0$ such that $\bar{\rho}_0$ is supported on $\mathcal{M}_{r_0}$, and separates $\{r_0\} \times \Omega$ and $\{-r_0\} \times \Omega$. If there exists $\bar{\rho}_\infty$ such that*

$$\lim_{t\to\infty} W_2(\bar{\rho}_t, \bar{\rho}_\infty) = 0,$$

then $\bar{\rho}_\infty$ is a global minimizer of $\bar{L}(\rho)$ over all probability distributions on $\mathcal{M}$.

Proof Our model falls into the "partial 1-homogeneous" case in [5]. Hence, Theorem 5 holds mostly following the proof therein. We only need to show that the regular value theorem still holds if the g_ρ in Assumption 3 is treated as a function on Ω with metric $G_{\bar{\mathbf{b}}}$. Equivalently, we show a regular value of g_ρ as a function under standard metric is still a regular value g_ρ as a function under the metric $G_{\bar{\mathbf{b}}}$. Let $\nabla g_\rho(\bar{\mathbf{b}})$ be the nonzero gradient of g_ρ at some regular point $\bar{\mathbf{b}}$. Since $\nabla g_\rho(\bar{\mathbf{b}})$ is tangent to Ω, we have $\bar{\mathbf{b}}^T\Sigma\nabla g_\rho(\bar{\mathbf{b}}) = 0$. Therefore,

$$\begin{aligned}
\nabla g_\rho(\bar{\mathbf{b}})^T(I - \Sigma\bar{\mathbf{b}}\bar{\mathbf{b}}^T)\nabla g_\rho(\bar{\mathbf{b}}) &= \nabla g_\rho(\bar{\mathbf{b}})^T\nabla g_\rho(\bar{\mathbf{b}}) - \nabla g_\rho(\bar{\mathbf{b}})^T\Sigma\bar{\mathbf{b}}\bar{\mathbf{b}}^T\nabla g_\rho(\bar{\mathbf{b}}) \\
&= \nabla g_\rho(\bar{\mathbf{b}})^T\nabla g_\rho(\bar{\mathbf{b}}) \\
&> 0.
\end{aligned}$$

This means $(I - \Sigma\bar{\mathbf{b}}\bar{\mathbf{b}}^T)\nabla g_\rho(\bar{\mathbf{b}})$ is nonzero. Hence,

$$\begin{aligned}
\nabla g_\rho(\bar{\mathbf{b}})^T\nabla_m g_\rho(\bar{\mathbf{b}})^T &= \nabla g_\rho(\bar{\mathbf{b}})^T\|\bar{\mathbf{b}}\|^2(I - \bar{\mathbf{b}}\bar{\mathbf{b}}^T\Sigma)(I - \Sigma\bar{\mathbf{b}}\bar{\mathbf{b}}^T)\nabla g_\rho(\bar{\mathbf{b}})^T \\
&= \|\bar{\mathbf{b}}\|^2\|(I - \Sigma\bar{\mathbf{b}}\bar{\mathbf{b}}^T)\nabla g_\rho(\bar{\mathbf{b}})^T\|^2 \\
&> 0.
\end{aligned}$$

This shows $\nabla_m g_\rho(\bar{\mathbf{b}}) \neq 0$, which means $\bar{\mathbf{b}}$ is also a regular point for g_ρ under metric $G_{\bar{\mathbf{b}}}$. □

4 Discussion

In this section, we discuss the potential benefits brought by batch normalization to the training dynamics of neural networks. We compare the dynamics of the models with BN

$$\begin{aligned}
\dot{a} &= -\mathbb{E}_{\mathbf{x}\sim\mu}l'(f(\mathbf{x},\rho))\sigma(\bar{\mathbf{b}}^T\mathbf{x}), \\
\dot{\mathbf{b}} &= -\mathbb{E}_{\mathbf{x}\sim\mu}l'(f(\mathbf{x},\rho))a\sigma'(\bar{\mathbf{b}}^T\mathbf{x})\frac{1}{\|\mathbf{b}\|_\Sigma}\left(I - \frac{\Sigma\mathbf{b}\mathbf{b}^T}{\|\mathbf{b}\|_\Sigma^2}\right)\mathbf{x},
\end{aligned} \tag{21}$$

and without BN

$$\begin{aligned}
\dot{a} &= -\mathbb{E}_{\mathbf{x}\sim\mu}l'(f(\mathbf{x},\rho))\sigma(\mathbf{b}^T\mathbf{x}), \\
\dot{\mathbf{b}} &= -\mathbb{E}_{\mathbf{x}\sim\mu}l'(f(\mathbf{x},\rho))a\sigma'(\mathbf{b}^T\mathbf{x})\mathbf{x}.
\end{aligned} \tag{22}$$

We focus on the speed of parameters $\mathbf{b}$, i.e., the speed of the changing of the neurons' directions.

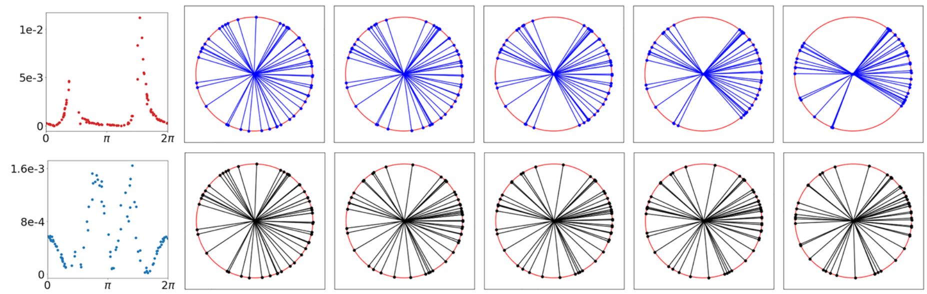

Fig. 1 The speed of neurons and distribution of neurons at different time of models with and without BN. The data $\mathbf{x} \in \mathbb{R}^2$ with $\Sigma = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}$. The leftmost column shows the speed of the neuron ($\mathbf{b}$ along the tangent direction of the unit circle) with respect to directions. The speed is calculated at the 500-th iteration. The top panel is for BN model, while the bottom panel is for non-BN model. Note that the scale of the top panel's vertical axis is one order of magnitude bigger than that of the bottom panel. The other five columns show the directions of the neurons at different training times (0, 2000, 4000, 6000, 8000 iterations from left to right). The first row shows results for BN model, and the second row shows results for non-BN model. The red circle is the unit circle. The $\mathbf{b}$'s for non-BN model are normalized to the unit circle

4.1 Speed and data distribution

By (21), for the model with batch normalization the speed of $\mathbf{b}$ depends on $\|\mathbf{b}\|_\Sigma$, which further depends on the distribution of the input data. Thus, when the data distribution is not isotropic, the speed of the neuron will be influenced by the relation between its direction and the data distribution. Specifically, when $\mathbf{b}$ points to the direction in which $\mathbf{x}$ is small, i.e., $\mathbb{E}\mathbf{b}^T\mathbf{x}$ is small, its moving speed will get bigger because $\|\mathbf{b}\|_\Sigma$ is small. On the other hand, when $\mathbf{b}$ points to the direction in which $\mathbf{x}$ is big, the moving speed will get smaller. This anisotropic speed effect helps neurons escape the nonsignificant directions where data are small and concentrate to significant directions where data are large, and thus speeds up the learning of the features along these significant directions. In Fig. 1, we show this effect with a comparison with vanilla models.

4.2 Speed and parameter magnitude

The existence of the term $\frac{1}{\|\mathbf{b}\|_\Sigma}$ also produces a connection between the speed of the neuron with its magnitude. To see this, consider two neurons $(a, \mathbf{b}_1)$ and $(a, \mathbf{b}_2)$ with $\mathbf{b}_2 = c\mathbf{b}_1$ for a positive constant c. Assume $\sigma'(\cdot)$ satisfies $\sigma'(cw) = \sigma'(w)$ for any input $w \in \mathbb{R}$ and $c > 0$. (This is true is σ is ReLU or leaky ReLU). Then, for models without BN, we have

$$\dot{\mathbf{b}}_2 = \dot{\mathbf{b}}_1,$$

while for models with BN, we have

$$\dot{\mathbf{b}}_2 = \frac{1}{c}\dot{\mathbf{b}}_1.$$

Hence, for the BN model, smaller neurons move faster. This is verified numerically in Fig. 2.

Considering that for the BN model neurons with different magnitude (of $\mathbf{b}$) actually express the same function, the influence of neuron magnitude on its speed allows the model to explore different learning rates–those neurons with small $\mathbf{b}$ are learning with big learning rates while the neurons with large $\mathbf{b}$ learn with small learning rates. This effect

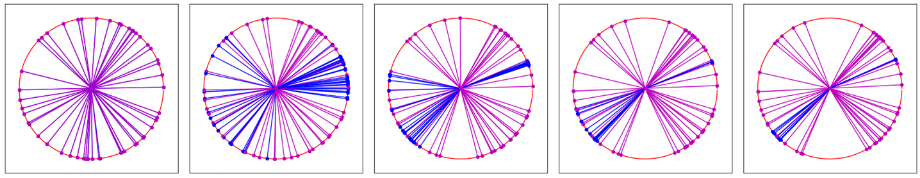

Fig. 2 The directions of the neurons for a model with multiscale initialization. Neurons shown in purple lines have 10× bigger initialization than neurons shown in blue lines. From left to right, the figures show neuron directions at iteration 0, 2000, 4000, 6000, 8000, 10,000. The figures show that blue neurons converge faster than purple neurons, i.e., shorter neurons moves faster than long neurons

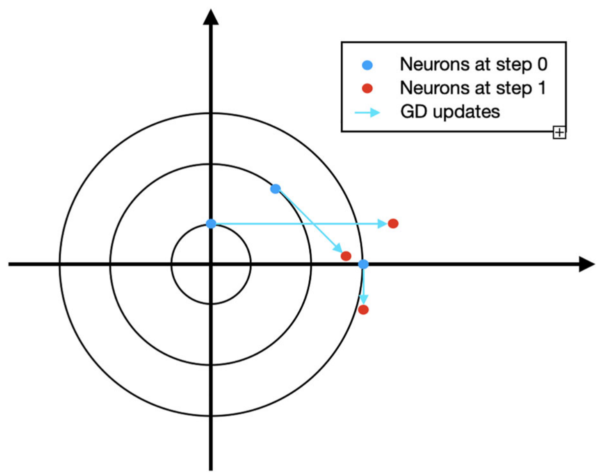

Fig. 3 The first-step amplification effect: The gradients of small neurons are larger than that of big neurons. Hence, after the first iteration, very small neurons become large and thus will not suffer from unstable time-discrete dynamics

gives the BN model an "adaptivity" to select the right learning rate itself and hence less sensitivity to the choice of learning rate.

4.3 The "first-step amplification"

By the above discussion, for the BN model, neurons with smaller magnitude move faster. At a first glance, this speed effect may cause problem when very small neurons exist, whose speeds are very large. However, for the discrete dynamics, i.e., gradient descent algorithm, this would not be a problem, because very small neurons will get big after the first iteration. This phenomenon is illustrated in Fig. 3. At the first iteration, the gradient for a small $\mathbf{b}$ is very large. Moreover, by the nature of batch normalization, the gradient lies in the tangent space of the sphere with radius $\|\mathbf{b}\|$. Hence, the first iteration will only make the magnitude of $\|\mathbf{b}\|$ bigger. The smaller the neuron is initially, the bigger it becomes after one iteration. Therefore, starting from the second iteration, all neurons are large enough to avoid unstable dynamics.

5 Related work

The good performance of batch normalization was initially believed to be caused by the prevention of internal covariance shift [8]. Later work challenged this point of view by numerical observations [16] and connected the benefit of BN with optimization landscape smoothness. Many other attempts are made to understand batch normalization. In [4], it is observed that BN can enable larger learning rate, which leads to faster convergence. In [17], it is shown that BN can flatten the optimization landscape. The authors of [11] analyzed the regularization effect of BN. A Riemannian manifold understanding was provided in [6], in which the authors proposed a manifold optimization framework based on the scale invariant property of BN. However, in [6] a Grassmann manifold with standard metric is considered, and thus normalization is still needed on the manifold. As a comparison, in this paper we find a special metric to eliminate the normalization step. In practice, a series of other normalization methods are proposed as substitutes of batch normalization [3,9,15,19].

On the other hand, the mean-field formulation for neural networks is a helpful tool to analyze and understand the training dynamics of neural networks, especially those with infinite neurons. The mean-field formulation was first studied for two-layer neural networks [5,12,14], and then extended to deep networks [1,7,13].

6 Conclusion

In this paper, we study the infinite-width limit of two-layer neural networks with batch normalization. We show that the dynamics of the model with batch normalization is equivalent with the dynamics of the original model on a Riemannian manifold. Then, we derive a mean-field formulation for the training dynamics with GD. The mean-field dynamics is a Wasserstein gradient flow on a Riemannian manifold. Based on existing results, we prove the existence of the Wasserstein gradient flow and show that once it converges, the limit is a global minimum. The Riemannian manifold understanding for batch normalization model provides us a new perspective to study the benefit of batch normalization.

Funding
Not applicable.

Availability of data and materials
Not applicable.

Declarations

Code availability
Will be uploaded for download.

Received: 21 September 2021 Accepted: 10 June 2022
Published online: 20 July 2022

References

1. Araújo, D., Oliveira, R.I., Yukimura, D.: A mean-field limit for certain deep neural networks. arXiv preprint arXiv:1906.00193 (2019)
2. Arora, S., Li, Z., Lyu, K.: Theoretical analysis of auto rate-tuning by batch normalization. arXiv preprint arXiv:1812.03981 (2018)
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
4. Bjorck, J., Gomes, C., Selman, B., Weinberger, K.Q.: Understanding batch normalization. arXiv preprint arXiv:1806.02375 (2018)
5. Chizat, L., Bach, F.: On the global convergence of gradient descent for over-parameterized models using optimal transport. arXiv preprint arXiv:1805.09545 (2018)
6. Cho, M., Lee, J.: Riemannian approach to batch normalization. arXiv preprint arXiv:1709.09603 (2017)
7. Fang, C., Lee, J., Yang, P., Zhang, T.: Modeling from features: a mean-field framework for over-parameterized deep neural networks. In: Conference on Learning Theory, pp. 1887–1936. PMLR (2021)

8. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456. PMLR (2015)
9. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 972–981 (2017)
10. Kohler, J., Daneshmand, H., Lucchi, A., Zhou, M., Neymeyr, K., Hofmann, T.: Towards a theoretical understanding of batch normalization. stat **1050**, 27 (2018)
11. Luo, P., Wang, X., Shao, W., Peng, Z.: Towards understanding regularization in batch normalization. arXiv preprint arXiv:1809.00846 (2018)
12. Mei, S., Montanari, A., Nguyen, P.-M.: A mean field view of the landscape of two-layer neural networks. Proc. Natl. Acad. Sci. **115**(33), 7665–7671 (2018)
13. Nguyen, P.-M.: Mean field limit of the learning dynamics of multilayer neural networks. arXiv preprint arXiv:1902.02880 (2019)
14. Rotskoff, G.M., Vanden-Eijnden, E.: Neural networks as interacting particle systems: asymptotic convexity of the loss landscape and universal scaling of the approximation error. stat **1050**, 22 (2018)
15. Salimans, T., Kingma, D.P.: Weight normalization: a simple reparameterization to accelerate training of deep neural networks. Adv. Neural. Inf. Process. Syst. **29**, 901–909 (2016)
16. Santurkar, S., Tsipras, D., Ilyas, A., Madry, A.: How does batch normalization help optimization? In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 2488–2498 (2018)
17. Wei, M., Stokes, J., Schwab, D.J.: Mean-field analysis of batch normalization. arXiv preprint arXiv:1903.02606 (2019)
18. Weinan, E., Ma, C., Wu, L.: Machine learning from a continuous viewpoint, I. Sci. China Math. **63**(11), 2233–2266 (2020)
19. Wu, Y., He, K.: Group normalization. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)

Publisher's Note