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

Song Mei[a], Andrea Montanari[b,c,1], and Phan-Minh Nguyen[b]

[a]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305; [b]Department of Electrical Engineering, Stanford University, Stanford, CA 94305; and [c]Department of Statistics, Stanford University, Stanford, CA 94305

UROP with Greg Pavliotis from Imperial College London

Pantelis Tassopoulos

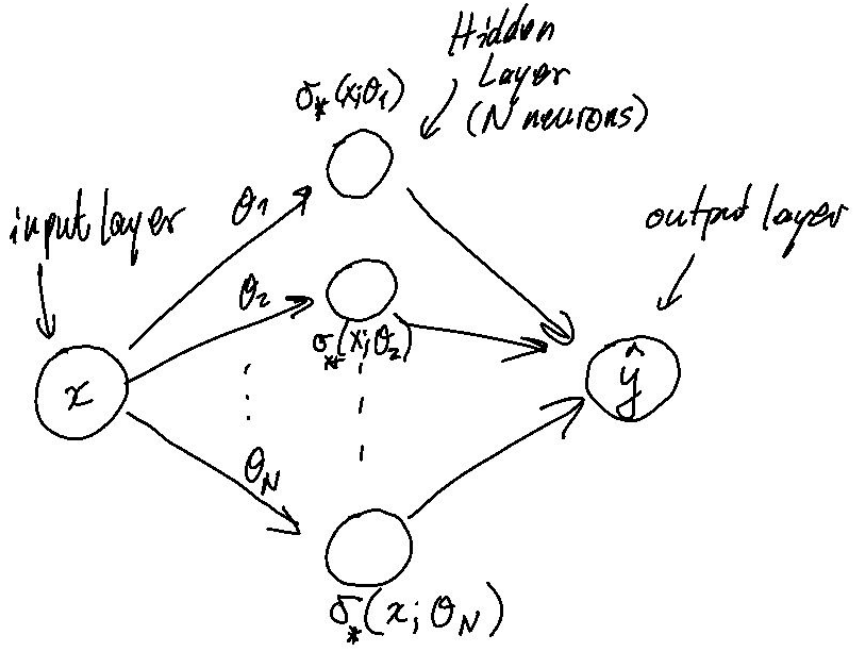
Significance

Learning a neural network from data requires solving a complex optimisation problem with millions of variables. This is done by stochastic gradient descent (SGD) algorithms.

One can study the case of two-layer networks and derive a compact description of the SGD dynamics in terms of a limiting partial differential equation.

Among other consequences, this shows that SGD dynamics do not become more complex when the network size increases.

Setup



- The setting of supervised learning
- Data points $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}, i \in \mathbb{N}$ iid.
- x -> feature vector, y -> label
- Model dependence of label on the feature vector
- In a two layer-network, this dependence is modelled by

$$\hat{y}(x;\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \sigma_*(x;\boldsymbol{\theta}_i)$$

- N is the number of hidden units (neurons)
- σ_*: $\mathbb{R}^d \times \mathbb{R}^D \to \mathbb{R}$ is an activation function
- $\boldsymbol{\theta} = (\boldsymbol{\theta}_i)_{i \le N}$, $\boldsymbol{\theta}_i \in \mathbb{R}^D$ are parameters, often $\boldsymbol{\theta}_i = (a_i, b_i, w_i)$ and

$$\sigma_*(x;\boldsymbol{\theta}_i) = a_i \sigma(\langle w_i, x\rangle + b_i)$$

for some σ: $\mathbb{R} \to \mathbb{R}$

- Ideally, parameters should be chosen to minimise the risk

$$R_N(\boldsymbol{\theta}) = \mathbb{E}\{\ell(y, \hat{y}(x;\boldsymbol{\theta}))\}$$

For a loss function ℓ: $\mathbb{R}\times\mathbb{R} \to \mathbb{R}$, in this case $\ell(y, \hat{y}) = (y - \hat{y})^2$

Setup

- The parameters are learned by stochastic gradient Descent (SGD).
- In the present case, this amounts to the iteration

$$\boldsymbol{\theta}_i^{k+1} = \boldsymbol{\theta}_i^k + 2s_k(y_k - \hat{y}(x_k;\, \boldsymbol{\theta}^k))\nabla_{\boldsymbol{\theta}_i}\, \sigma_*(x_k; \boldsymbol{\theta}_i^k)$$

where $\boldsymbol{\theta}^k = \left(\boldsymbol{\theta}_i^k\right)_{i\leq N}$ are the parameters after k iterations, s_k is a step size and $(\boldsymbol{x}_k, y_k)$ is the kth sample (samples are iid. $\sim \mathbb{P}$).

- Can express the risk (generalisation error) as

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

where $V(\boldsymbol{\theta}) = -\mathbb{E}\{y\sigma_*(\boldsymbol{x}; \boldsymbol{\theta})\}, U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \mathbb{E}\{\sigma_*(\boldsymbol{x}; \boldsymbol{\theta}_1)\sigma_*(\boldsymbol{x}; \boldsymbol{\theta}_2)\}$ and $R_\# = \mathbb{E}\{y^2\}$ that is the risk of the trivial predictor $\hat{y} = 0$.

- The population risk depends on parameters through their empirical distribution

$$\hat{\rho}^{(N)} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\boldsymbol{\theta}_i}$$

- We can thus consider a risk function defined for $\rho \in \mathcal{P}(\mathbb{R}^D)$, the space of probability measures on $\mathbb{R}^D$

$$R(\rho) = R_\# + 2\int V(\boldsymbol{\theta})\rho(d\boldsymbol{\theta}) + \int U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\, \rho(d\boldsymbol{\theta}_1)\, \rho(d\boldsymbol{\theta}_2)$$

Informal Overview of Main Result

- The authors prove that the SGD is well approximated by a continuum dynamics described below.
- Suppose the step size in the SGD is given by $s_k = \varepsilon\xi(k\varepsilon)$, for $\xi: \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ a sufficiently regular function
- Denoting the empirical distribution of parameters after k SGD steps $\hat{\rho}_k^{(N)} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\boldsymbol{\theta}_i^k}$, it is shown that $\hat{\rho}_k^{(N)}$ converges in the weak sense to ρ_t, when $N \to \infty, \varepsilon \to 0$, where the asymptotic dynamics is the solution to the PDE

$$\partial_t \rho_t = 2\xi(t)\nabla_{\boldsymbol{\theta}} \cdot \left(\rho_t \nabla_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}; \rho_t)\right),$$

namely, the distributional dynamics (DD), where $\Psi(\boldsymbol{\theta};\ \rho) = V(\boldsymbol{\theta}) + \int U(\boldsymbol{\theta}, \boldsymbol{\theta}')\rho(d\boldsymbol{\theta}')$.

- The above PDE can be viewed as a gradient flow for the cost function $R(\rho)$ in the space $(\mathcal{P}(\mathbb{R}^D), W_2)$, where W_2 is the Wasserstein 2-metric.

The PDE formulation leads to several insights and simplifications. One can exploit symmetries in the data distribution
$\mathbb{P}$.

For instance, if $\mathbb{P}$ is invariant under rotations, one can look for a solution to the above PDE that has the same symmetry, thereby reducing the dimensionality of the problem, thereby making theoretical and numerical analysis easier. (This is indeed the case for the classification problem of two isotropic Gaussians mentioned later).

This is not possible for the finite$-N$ dynamics since no arrangement of the points $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N\} \subseteq \mathbb{R}^d$ is invariant under rotations, say.

Technical Assumptions

A1. $t \longmapsto \xi(t)$ is bounded Lipschitz, with $\int_0^\infty \xi(t)\, dt = \infty$

A2. The activation function $(x, \boldsymbol{\theta}) \longmapsto \sigma_*(x; \boldsymbol{\theta})$ is bounded, with a sub-Gaussian gradient. Labels y_k are bounded.

A3. The gradients $\boldsymbol{\theta} \longmapsto \nabla V(\boldsymbol{\theta}), (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \longmapsto \nabla_{\boldsymbol{\theta}_1} U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ are bounded, and Lipschitz continuous

Also define the following error term that quantifies in a non-asymptotic sense the accuracy of the PDE model:

$$\boldsymbol{err}_{N,D}(z) \equiv \sqrt{1/N \vee \varepsilon} \cdot \left[\sqrt{D + \log(N/\varepsilon)} + z\right]$$

Main Theorem

Assume that conditions **A1, A2, A3** hold. For $\rho_0 \in \mathcal{P}(\mathbb{R}^D)$, the SGD dynamics with initialisation $\left(\boldsymbol{\theta}_i^0\right)_{i\leq N} \sim \rho_0$ and step size $s_k = \varepsilon\xi(k\varepsilon)$. For $t \geq 0$, let ρ_t be the solution of the PDE (DD). Then, for any fixed k, $\hat{\rho}_k^{(N)}$ converges weakly to ρ_t almost surely along any sequence $(N, \varepsilon = \varepsilon_N)$ such that $N/\log(1/\varepsilon_N) \to \infty, \varepsilon_N \to 0$. Further, there exists a constant C (depending uniquely on the a priori bounds derived from conditions **A1, A2, A3**) such that, for any $f: \mathbb{R}^D \times \mathbb{R} \to \mathbb{R}$, with $\|f\|_\infty, \|f\|_{Lip} \leq 1, \varepsilon \leq 1$,

$$\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} \left| \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}_i^k) - \int f(\boldsymbol{\theta})\rho_{k\varepsilon}(\mathrm{d}\boldsymbol{\theta}) \right| \leq Ce^{CT}\,\mathbf{err}_{N,D}(z),$$

$$\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} \left| R_N(\boldsymbol{\theta}^k) - R(\rho_{k\varepsilon}) \right| \leq Ce^{CT}\,\mathbf{err}_{N,D}(z),$$

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

Proof Sketch

"Propagation of chaos argument”

The conditions **A1** and **A3** are sufficient for the existence and uniqueness of solutions to the PDE (DD) (interpreted in the weak sense).

Discrete SGD dynamics for $\boldsymbol{\theta}^k = \left(\theta_i^k\right)_{i \leq N}$, approximates some non-linear dynamics in continuous time.

The sub-gaussianity and Lipschitz continuity assuptions enables the use of concentration inequalities (Azuma-Hoeffding) to derive maximal inequalities for the deviation of the above discrete and non-linear dynamics that further controls the terms in the statement of the theorem.

Empirical Validation on Toy Example

Centred Isotropic Gaussians: classification of Gaussians with The same mean. That is, assume the joint law $\mathbb{P}$ of $(y, \boldsymbol{x})$ to be:
with probability $1/2: y = +1, \boldsymbol{x} \sim \mathbf{N}\left(\mathbf{0}, (\mathbf{1} + \boldsymbol{\Delta})^2 \mathbf{I}_d\right)$
with probability $1/2: y = -1, \boldsymbol{x} \sim \mathbf{N}\left(\mathbf{0}, (\mathbf{1} - \boldsymbol{\Delta})^2 \mathbf{I}_d\right)$

For the activation function set $\sigma_*(\boldsymbol{x}; \boldsymbol{\theta}_i) = \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle)$, where σ is a simple piecewise linear function.

Run SGD with $\left(\boldsymbol{w}_i^0\right)_{i \le N} \sim_{iid} \rho_0$, where ρ_0 is spherically symmetric.
Fig. 1 reports the result of such an experiment.

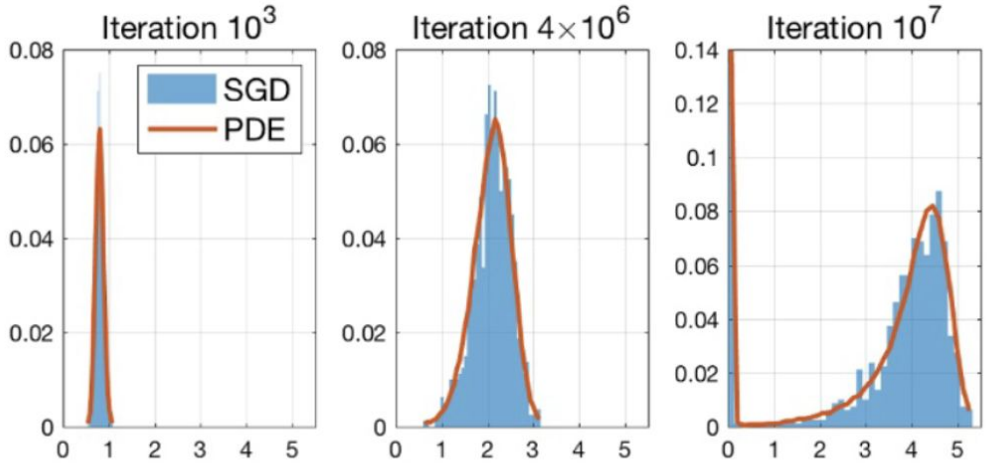
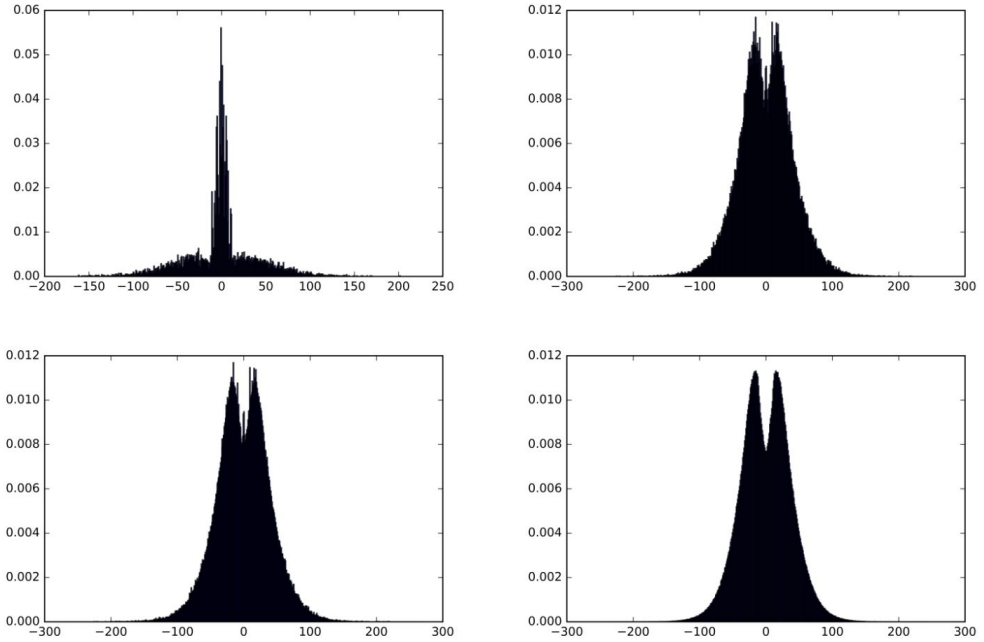


Fig. 1. Evolution of the radial distribution $\overline{\rho}_t$ for the isotropic Gaussian model, with $\Delta = 0.8$. Histograms are obtained from SGD experiments with $d = 40$, $N = 800$, initial weight distribution $\rho_0 = \mathrm{N}(\mathbf{0}, 0.8^2/d \cdot \boldsymbol{I}_d)$, and step size $\epsilon = 10^{-6}$ and $\xi(t) = 1$. Continuous lines correspond to a numerical solution of the DD

Due to the symmetry of the distribution $\mathbb{P}$, the distribution ρ_t remains spherical symmetric for all t and is hence completely determined by the distribution of the norm $\|\boldsymbol{w}\|_2$. This distribution satisfies a reduced, one-dimensional PDE.



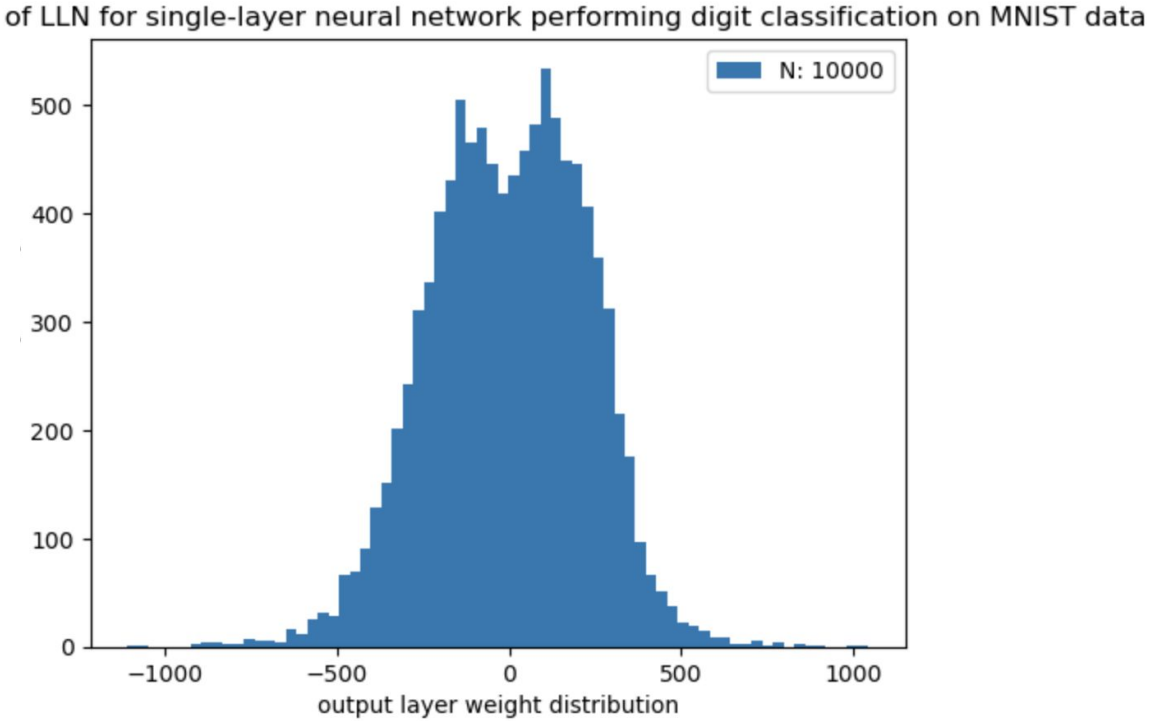


Figure 1: Distribution of parameters for a neural network trained on MNIST dataset. Clockwise: $N = 1,000$, $N = 10,000$, $N = 100,000$, and $N = 250,000$ hidden units.

References

- Mei, Song, Andrea Montanari, and Phan-Minh Nguyen. "A mean field view of the landscape of two-layer neural networks." *Proceedings of the National Academy of Sciences* 115.33 (2018): E7665-E7671.
- Sirignano, Justin, and Konstantinos Spiliopoulos. "Mean field analysis of neural networks: A law of large numbers." *SIAM Journal on Applied Mathematics* 80.2 (2020): 725-752.