Mean Field Analysis of Neural Networks

Konstantinos Spiliopoulos

Department of Mathematics & Statistics, Boston University

Partially supported by career award NSF-DMS 1550918

Joint work with Justin Sirignano (University of Illinois at Urbana-Champaign)

Outline

1. Neural networks and their mean field formulation
   - Introduction and mean field formulation
   - Main results
2. Real data analysis
3. Overview of the proofs
   - Proof for law of large numbers
   - Proof for fluctuations
4. Summary
5. References

Part I

Neural networks and their mean field formulation

Applications of Neural networks

- Neural networks and machine learning have revolutionized fields such as image, text and speech recognition.
- Growing interest in applying neural network techniques to engineering, robotics, medicine, finance, identify cancer and model protein folding.

Applications of Neural networks

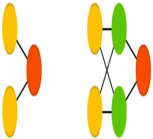
- Deep neural networks have the ability to accurately approximate high dimensional functions.
- In certain problems it has been shown that they can overcome the curse of dimensionality[1]
- New and exciting directions in applied mathematics!
- Need for mathematical understanding and mathematically appropriate framework.

[1] Han, Jentzen and E (2017), Sirignano and Spiliopoulos (2016,2017), Jentzen et all (2018)

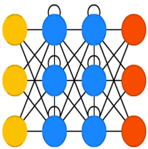
Immense success in applications but very limited mathematical understanding.

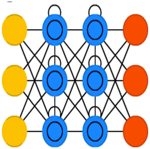
- P. Bartlett, D. Foster, and M. Telgarsky (margin bounds for neural networks)
- Mallat (understanding deep convolutional neural networks)
- Telgarsky (benefits of depth in neural networks)
- Hornik, Stinchcombe, White, Barron, Cybenko (universal approximation results and rates)

Some examples of neural networks:[2]



Feed forward NN





(a) Recurrent NN (RNN) (b) Long-Short-Term Memory Unit (LSTM)

[2] Source: http:// www.asimovinstitute.org/neural-network-zoo/

Mean field formulation of neural networks

Consider the one layer network

$$g_\theta^N(x) = \frac{1}{N}\sum_{i=1}^{N} c^i \sigma(w^i \cdot x), \tag{1}$$

where

- neural network parameters $\theta = (c^1, \dots, c^N, w^1, \dots, w^N) \in \mathbb{R}^{(1+d)N}$ which must be estimated from data.
- $\sigma(w^i \cdot x)$ is the i-th "hidden unit", and the vector $\big(\sigma(w^1 \cdot x), \dots, \sigma(w^N \cdot x)\big)$ is called the "hidden layer".
- The number of units in the hidden layer is N.

Mean field formulation of neural networks

The objective function, or loss function, is

$$L(\theta) = \mathbb{E}_{Y,X}[(Y - g^N_\theta(X))^2], \tag{2}$$

where the data (Y, X) is assumed to have a joint distribution $\pi(dx, dy)$. The parameters $\theta = (c, w)$ are estimated using stochastic gradient descent:

$$\begin{aligned} c^i_{k+1} &= c^i_k + \frac{\alpha}{N}(y_k - g^N_{\theta_k}(x_k))\sigma(w^i_k \cdot x_k), \\ w^{i,j}_{k+1} &= w^{i,j}_k + \frac{\alpha}{N}(y_k - g^N_{\theta_k}(x_k))c^i_k\sigma'(w^i_k \cdot x_k)x^j_k, \quad j = 1, \cdots, d, \end{aligned} \tag{3}$$

where α is the learning rate and $(x_k, y_k) \sim \pi(dx, dy)$.

- Stochastic gradient descent minimizes (2) using a sequence of noisy (but unbiased) gradient descent steps $\nabla_\theta[(y_k - g^N_{\theta_k}(x_k))^2]$.
- Typically $\nabla_\theta[(y - g^N_\theta(x))^2]$ is not a priori globally Lipschitz nor globally bounded as a function of θ.

Mean field formulation of neural networks

- **Question:** Can we guarantee convergence of the algorithm? How does the distribution of the trained parameters evolve over time?

Mean field formulation of neural networks

- **Question:** Can we guarantee convergence of the algorithm? How does the distribution of the trained parameters evolve over time?

Define the empirical measure

$$\nu_k^N(dc, dw) = \frac{1}{N} \sum_{i=1}^{N} \delta_{c_k^i, w_k^i}(dc, dw). \tag{4}$$

The neural network's output can be re-written in terms of the empirical measure:

$$g_{\theta_k}^N(x) = \left\langle c\sigma(w \cdot x), \nu_k^N \right\rangle . \tag{5}$$

$\langle f, h \rangle$ denotes the inner product of f and h. The scaled empirical measure is

$$\mu_t^N = \nu_{\lfloor Nt \rfloor}^N. \tag{6}$$

Assumptions

At any time t, the scaled empirical measure μ_t^N is a random element of the Skorokhod space $D_E([0,T]) = D([0,T]; E)$ with with $E = \mathcal{M}(\mathbb{R}^{1+d})$. We study the convergence in distribution of μ_t^N in $D_E([0,T])$.

- The activation function $\sigma \in C_b^\infty(\mathbb{R})$, i.e. σ is continuously differentiable and bounded.
- The data $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ is compactly supported.
- The sequence of data samples (x_k, y_k) is i.i.d.
- The random initialization is such that (c_0^i, w_0^i) is i.i.d. generated from a distribution that has compact support.

Convergence result

Let numbers of neurons and number of iterations increase!

Theorem 1–Law of large numbers.

The scaled empirical measure μ_t^N converges in distribution to $\bar{\mu}_t$ in $D_E([0, T])$ as $N \to \infty$. For every $f \in C_b^2(\mathbb{R}^{1+d})$, $\bar{\mu}$ satisfies the measure evolution equation

$$\langle f, \bar{\mu}_t \rangle = \langle f, \bar{\mu}_0 \rangle + \int_0^t \left(\int_{\mathcal{X}\times\mathcal{Y}} \alpha \big(y - \big\langle c' \sigma(w' \cdot x), \bar{\mu}_s \big\rangle \big) \left\langle \nabla (c\sigma(w \cdot x)) \cdot \nabla f, \bar{\mu}_s \right\rangle \pi(dx, dy) \right) ds \tag{7}$$

Convergence result

Corollary 1.

Suppose that $\bar{\mu}_0$ admits a density $p_0(c,w)$ and there exists a solution to the nonlinear partial differential equation

$$\begin{aligned}\frac{\partial p(t,c,w)}{\partial t} &= -\alpha\int_{\mathcal{X}\times\mathcal{Y}}\left(\left(y-\left\langle c'\sigma(w'\cdot x), p(t,c',w')\right\rangle\right)\frac{\partial}{\partial c}\left[\sigma(w\cdot x)p(t,c,w)\right]\right)\pi(dx,dy)\\ &\quad -\alpha\int_{\mathcal{X}\times\mathcal{Y}}\left(\left(y-\left\langle c'\sigma(w'\cdot x), p(t,c',w')\right\rangle\right)x\cdot\nabla_w\left[c\sigma'(w\cdot x)p(t,c,w)\right]\right)\pi(dx,dy),\\ p(0,c,w) &= p_0(c,w). \end{aligned} \tag{8}$$

Then, we have that the solution to the measure evolution equation (7) is such that

$$\bar{\mu}_t(dc,dw) = p(t,c,w)dcdw.$$

Convergence result

Theorem 2-Propagation of chaos.

Consider $T < \infty$ and let $t \in (0, T]$. Define the probability measure $\rho^N_t \in \mathcal{M}(\mathbb{R}^{(1+d)N})$ where

$$\rho^N_t(dx^1, \dots, dx^N) = \mathbb{P}[(c^1_{\lfloor Nt \rfloor}, w^1_{\lfloor Nt \rfloor}) \in dx^1, \dots, (c^N_{\lfloor Nt \rfloor}, w^N_{\lfloor Nt \rfloor}) \in dx^N].$$

Then, the sequence of probability measures $\rho^N_\cdot$ is $\bar{\mu}_\cdot$-chaotic. That is, for $k \in \mathbb{N}$

$$\lim_{N\to\infty} \left\langle f_1(x^1) \times \cdots \times f_k(x^k), \rho^N_\cdot(dx^1, \dots, dx^N) \right\rangle = \prod_{i=1}^{k} \langle f_i, \bar{\mu}_\cdot \rangle, \quad \forall f_1, \dots, f_k \in C^2_b(\mathbb{R}^{1+d}).$$

Convergence result-fluctuations

Define the fluctuation process

$$\eta_t^N = \sqrt{N}(\mu_t^N - \bar{\mu}_t). \tag{9}$$

Theorem 3-Fluctuations.

let $J \geq 3\left\lceil \frac{d+1}{2} \right\rceil + 7$. Let $T > 0$ be given. The sequence $\{\eta_t^N, t \in [0,T]\}_{N \in \mathbb{N}}$ is relatively compact in $D_{W^{-J,2}}([0,T])$ and $\{\eta_t^N, t \in [0,T]\}_{N \in \mathbb{N}}$ converges in distribution in $D_{W^{-J,2}}([0,T])$ to the process $\{\bar{\eta}_t, t \in [0,T]\}$ where

$$\begin{aligned} \langle f, \bar{\eta}_t \rangle &= \langle f, \bar{\eta}_0 \rangle + \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha\big(y - \langle c\sigma(w\cdot x), \bar{\mu}_s\rangle\big) \langle \nabla(c\sigma(w\cdot x))\nabla f, \bar{\eta}_s\rangle \pi(dx,dy)ds \\ &- \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \langle c\sigma(w\cdot x), \bar{\eta}_s\rangle \langle \nabla(c\sigma(w\cdot x))\nabla f, \bar{\mu}_s\rangle \pi(dx,dy)ds + \langle f, \bar{M}_t\rangle, \end{aligned} \tag{10}$$

for every $f \in W_0^{J,2}(\Omega)$. $\bar{M}_t$ is a mean-zero distribution valued Gaussian process. Finally, the stochastic evolution equation (10) has a unique solution in $W^{-J,2}$.

Variance-covariance structure

Define the operator

$$\mathcal{R}_{x,y,\mu}[h] = (y - \langle c\sigma(w \cdot x), \mu\rangle) \langle \nabla(c\sigma(w \cdot x)) \cdot \nabla h, \mu\rangle .$$

Then, for every $f, g \in W_0^{J,2}(\Theta)$, $(\sqrt{N}\left\langle f, M_t^N\right\rangle, \sqrt{N}\left\langle g, M_t^N\right\rangle) \in D_{\mathbb{R}^2}([0,T])$ converges to a distribution valued mean-zero Gaussian martingale with covariance function

$$\begin{aligned}\mathrm{Cov}\left[\left\langle f, \bar{M}_t\right\rangle, \left\langle g, \bar{M}_t\right\rangle\right] &= \alpha^2 \int_0^t \left[\int_{\mathcal{X}\times\mathcal{Y}} \left(\mathcal{R}_{x,y,\bar{\mu}_s}[f] - \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{R}_{x,y,\bar{\mu}_s}[f]\pi(dx,dy)\right) \times \right.\\ &\left. \times \left(\mathcal{R}_{x,y,\bar{\mu}_s}[g] - \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{R}_{x,y,\bar{\mu}_s}[g]\pi(dx,dy)\right)\pi(dx,dy)\right] ds.\end{aligned}$$

Insights from convergence results

- As $N \to \infty$, the neural network converges (in probability) to a deterministic model. This is despite the fact that the neural network is randomly initialized and it is trained on a random sequence of data samples via stochastic gradient descent.
- The learning rate α was assumed to be constant and to not decay with time. For finite N, the α must decay with the number of iterations in order for stochastic gradient descent to converge. Despite this, the noise disappears and the neural network's parameter distribution converges to a deterministic evolution equation. This is due to the normalization of $\frac{1}{N}$ in the hidden layer replacing the role of the learning rate decay.

Insights from convergence results

- Under the setup of (1), (2) and (3), the limiting equation characterizing the evolution of the distribution of parameters is a first-order PDE. Therefore, the asymptotic dynamics are of a "transport" instead of a "diffusive" nature.
- The propagation of chaos result (9) indicates that, as $N \to \infty$, the dynamics of the weights (c_k^i, w_k^i) will become independent of the dynamics of the weights (c_k^j, w_k^j) for any $i \neq j$. Note that the dynamics (c_k^i, w_k^i) are still random due to the random initialization. However, the dynamics of the i-th set of weights will be uncorrelated with the dynamics of the j-th set of weights in the limit as $N \to \infty$.

Insights from convergence results

- The fluctuations theorem indicates that for large N the empirical distribution of the neural network's parameters behaves as

  $$\mu^N \approx \bar{\mu} + \frac{1}{\sqrt{N}}\bar{\eta}, \tag{11}$$

  where $\bar{\eta}$ has a Gaussian distribution.
- The relation between the number of particles ("hidden units" in the language of neural networks) and the number of stochastic gradient steps should be of the same order to have convergence and statistically good behavior.

Related Literature

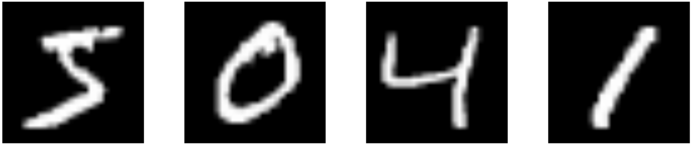
- Extensive research on stochastic gradient descent in discrete time.
- Relatively little mathematical work of convergence properties of neural networks and machine learning algorithms.
- Mei and Montanari and Nguyen (2018), Rotskoff and Vanden-Eijnden (2018), Wang and Mattingly and Lu (2017)

Part II

Real data analysis

Real data analysis

- MNIST dataset, which is a standard image dataset in machine learning. The dataset includes $60,000$ images of handwritten numbers $\{0, 1, 2, \ldots, 9\}$.
- The neural network is trained to identify the handwritten numbers using only the image pixels as an input (i.e., it learns to recognize images as a human would).
- In the MNIST dataset, each image has 784 pixels. A pixel takes values in $\{0, 1, \ldots, 255\}$. Neural networks can achieve 98-99% out-of-sample accuracy on the MNIST dataset.



Real data analysis

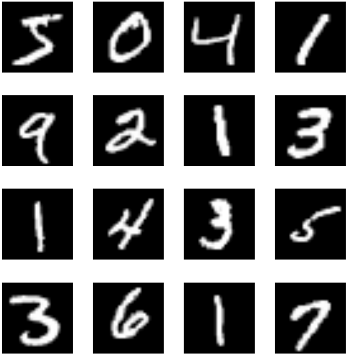


Figure 1: Examples of images from the MNIST dataset. Each image is described by a 28×28 array of pixels, which can be re-arranged into a vector $x \in \mathbb{R}^{784}$. The vector x containing the pixel values is the input to the neural network, which attempts to correctly predict the handwritten number in the image.

Real data analysis

- The neural network has a single hidden layer followed by a softmax function. Figure 2 reports the distribution of the parameters connecting the hidden layer to the softmax function.
- The distributions are presented as histograms.
- The neural network is trained on the MNIST dataset.
- Figure 2 shows that the distribution of parameters converges to a fixed distribution as $N \to \infty$.

Real data analysis

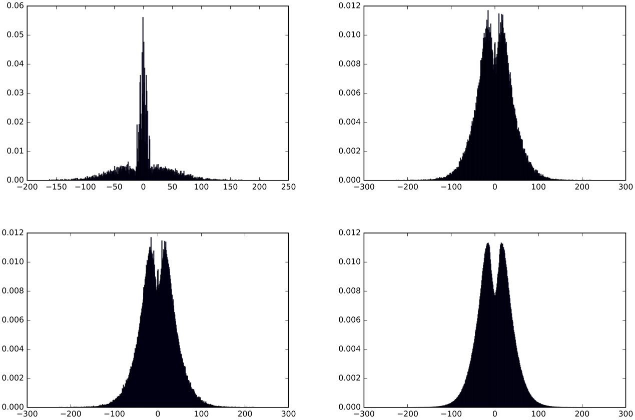


Figure 2: Clockwise: $N = 1,000$, $N = 10,000$, $N = 100,000$, and $N = 250,000$ hidden units.

Part III

Overview of the proofs

In general...

- Tightness of the involved measure valued processes
- Identification of the limit
- Uniqueness to the solution of the limiting equation

Relative compactness (tightness)

Lemma (compact containment).

For each $\eta > 0$ and $t \geq 0$, there is a compact subset $\mathcal{K}$ of E such that

$$\sup_{N \in \mathbb{N}, 0 \leq t \leq T} \mathbb{P}[\mu_t^N \notin \mathcal{K}] < \eta.$$

In fact, there is a uniform constant C (which does not depend on k nor N, but can depend on T) such that for all $k < TN$

$$|c_k^i| + \| w_k^i \| \leq C.$$

This uniform bound actually implies the stronger statement of compact support. Define

$$\mathcal{K} = \left\{ \omega \in M(\mathbb{R}^{1+d}) : \omega \left([-C, C]^{1+d} \right) = 1 \right\}.$$

Then $\mathcal{K} \subset\subset M(\mathbb{R}^{1+d})$, and $\mathbb{P}$-a.s. $\mu_t^N \in \mathcal{K}$ for all $N \in \mathbb{N}$ and $t \in [0, T]$.

Relative compactness (tightness)

Lemma (regularity).

Define the function $q(z_1, z_2) = \min\{|z_1 - z_2|, 1\}$ where $z_1, z_2 \in \mathbb{R}$. For any $p \in (0,1)$, there is a constant $C < \infty$ such that for $0 \leq u \leq \delta$, $0 \leq v \leq \delta \wedge t$, $t \in [0, T]$,

$$\mathbb{E}\left[q(\left\langle f, \mu^N_{t+u}\right\rangle, \left\langle f, \mu^N_t\right\rangle)q(\left\langle f, \mu^N_t\right\rangle, \left\langle f, \mu^N_{t-v}\right\rangle)\Big|\mathcal{F}^N_t\right] \leq C\delta^p + O_N(1).$$

These two lemmas then imply relative compactness of $\{\mu^N\}_{N\in\mathbb{N}}$ in $D_E([0, T])$ (see for example Theorem 8.6 of Chapter 3 of Ethier and Kurtz).

Limit identification

Recall that

$$\begin{aligned} c^i_{k+1} &= c^i_k + \frac{\alpha}{N}(y_k - g^N_{\theta_k}(x_k))\sigma(w^i_k \cdot x_k), \\ w^{i,j}_{k+1} &= w^{i,j}_k + \frac{\alpha}{N}(y_k - g^N_{\theta_k}(x_k))c^i_k\sigma'(w^i_k \cdot x_k)x^j_k, \quad j = 1, \cdots, d, \end{aligned}$$

where α is the learning rate and $(x_k, y_k) \sim \pi(dx, dy)$.
Using Taylor expansion and the equations evolving c^i_k and w^i_k we can write

$$\begin{aligned} \left\langle f, \nu^N_{k+1} \right\rangle - \left\langle f, \nu^N_k \right\rangle &= \frac{1}{N^2}\sum_{i=1}^{N} \partial_c f(c^i_k, w^i_k)\alpha(y_k - g^N_{\theta_k}(x_k))\sigma(w^i_k \cdot x_k) \\ &\quad + \frac{1}{N^2}\sum_{i=1}^{N} \alpha(y_k - g^N_{\theta_k}(x_k))c^i_k\sigma'(w^i_k \cdot x_k)\nabla_w f(c^i_k, w^i_k) \cdot x_k + O\left(N^{-2}\right). \end{aligned}$$

Limit identification

Decomposing into drift and martingale components we then obtain for the scaled empirical measure satisfies, as N grows,

$$\begin{aligned}\left\langle f,\mu_t^N\right\rangle-\left\langle f,\mu_0^N\right\rangle&=\int_0^t\left(\int_{\mathcal{X}\times\mathcal{Y}}\alpha\big(y-\left\langle c\sigma(w\cdot x),\mu_s^N\right\rangle\big)\left\langle\sigma(w\cdot x)\nabla_c f,\mu_s^N\right\rangle\pi(dx,dy)\right)ds\\&+\int_0^t\left(\int_{\mathcal{X}\times\mathcal{Y}}\alpha\big(y-\left\langle c\sigma(w\cdot x),\mu_s^N\right\rangle\big)\left\langle c\sigma'(w\cdot x)x\cdot\nabla_w f,\mu_s^N\right\rangle\pi(dx,dy)\right)ds\\&+M^{1,N}(t)+M^{2,N}(t)+O(N^{-1}).\end{aligned}$$

such that

$$\lim_{N\to\infty}\mathbb{E}\left[\left(M^{1,N}(t)\right)^2\right]=\lim_{N\to\infty}\mathbb{E}\left[\left(M^{2,N}(t)\right)^2\right]=0.$$

Uniqueness

Set up a Picard type of iteration and prove that it has a unique fixed point through a contraction mapping. Notice that

$$\langle f, \bar{\mu}_t \rangle = \langle f, \bar{\mu}_0 \rangle + \int_0^t \langle G(z, Q(\bar{\mu}_s, \cdot)) \cdot \nabla f, \bar{\mu}_s \rangle \, ds, \tag{12}$$

where for $z = (c, w_1, \cdots, w_d) \in \mathbb{R}^{1+d}$, $Q(\bar{\mu}, x) = \langle c\sigma(w \cdot x), \bar{\mu} \rangle$ we have

$$G(z, Q(\bar{\mu}, \cdot)) = (G_1(z, Q(\bar{\mu}, \cdot)), G_2(z, Q(\bar{\mu}, \cdot))) \in \mathbb{R}^{1+d}$$

with

$$G_1(z, Q(\bar{\mu}, \cdot)) = \int_{\mathcal{X} \times \mathcal{Y}} \alpha(y - Q(\bar{\mu}, x))\sigma(w \cdot x)\pi(dx, dy) \in \mathbb{R}$$
$$G_2(z, Q(\bar{\mu}, \cdot)) = \int_{\mathcal{X} \times \mathcal{Y}} \alpha(y - Q(\bar{\mu}, x))c\sigma'(w \cdot x)x\pi(dx, dy) \in \mathbb{R}^d.$$

Uniqueness

- Let $F : D([0,T];\mathbb{R}) \mapsto D([0,T];M(\mathbb{R}^{1+d}))$ be such that for a path $(R_t)_{t\in[0,T]} \in D([0,T];\mathbb{R})$, we have that $F(R_\cdot) = \text{Law}(Y_\cdot)$ where $Y_\cdot$ is given by

$$Y_t = Y_0 + \int_0^t G(Y_s, R_s)ds, \quad Y_0 \sim \bar{\mu}(0,c,w).$$

- Let us also define the map $L : D([0,T];M(\mathbb{R}^{1+d})) \mapsto D([0,T];\mathbb{R})$ taking a measure valued process μ_t and mapping it to $Q(\mu_t, x) = L(\mu)$ where

$$Q(\mu_t, x) = \langle c\sigma(w\cdot x), \mu_t \rangle .$$

- Then, we consider the mapping $H : D([0,T];M(\mathbb{R}^{1+d})) \mapsto D([0,T];M(\mathbb{R}^{1+d}))$ defined via the composition of the mappings F and L, we set $H = F \circ L$.

Uniqueness

Let us define for notational convenience $C_T = C([0,T], \mathbb{R}^{1+d})$ and let M_T be the set of probability measures on C_T. For $m, m' \in M_T$ and $p \geq 1$ define the metric

$$D_{T,p}(m,m') = \inf\left\{ \left(\int_{C_T \times C_T} \sup_{s \leq T} \|x_s - y_s\|_p^p \wedge 1 d\nu(x,y) \right)^{1/p}, \nu \in P(m,m') \right\},$$

Lemma

Let $m^1, m^2 \in M_T$ and $T < \infty$. Then, there exists a constant $C < \infty$ that may depend on T such that

$$D_{t,1}(H(m^1), H(m^2)) \leq C \int_0^t D_{u,1}(m^1, m^2) du,$$

for any $0 < t < T$.

Uniqueness

The previous lemma immediately proves there is a contraction on the interval $[0, T_0]$.

$$\begin{aligned} D_{t,1}(H(m^1), H(m^2)) &\leq C \int_0^t D_{u,1}(m^1, m^2) du \\ &\leq C \int_0^t D_{t,1}(m^1, m^2) du \\ &\leq C t D_{t,1}(m^1, m^2). \end{aligned}$$

Then, choose T_0 such that $CT_0 < 1$. In fact we have:

Lemma

Let $T < \infty$. The mapping $H_T = (F \circ F)_T$ has a unique fixed point.

Main ideas for the proof of fluctuations...[3]

The analysis of the limiting behavior of the fluctuation process involves issues that do not occur in the treatment of the LLN. It is considerably more complicated.

- Even though the fluctuation process $\eta_t^N = \sqrt{N}(\mu_t^N - \bar{\mu}_t)$ is a signed-measure-valued process, its limit process is distribution-valued in an appropriate space.
- In general, the space of signed measures endowed with the weak topology is not metrizable.
- The difficulty is then to identify a rich enough space, where tightness and uniqueness can be proven.
- It turns out that we have to consider the convergence in Sobolev spaces $W_o^J(\Omega)$ with "enough" weak derivatives $J \geq 3\lceil \frac{d+1}{2} \rceil + 7$.

[3] Kurtz and Xiong (2004), Fernandez and Meleard (1997), S. and Sirignano and Giesecke (2014)

Representation for fluctuation process

Let $\eta_t^N = \sqrt{N}(\mu_t^N - \bar{\mu}_t)$. We can write

$$\begin{aligned}\left\langle f, \eta_t^N \right\rangle - \left\langle f, \eta_0^N \right\rangle = & \int_0^t \left(\int_{\mathcal{X}\times\mathcal{Y}} \alpha\big(y - \langle c\sigma(w\cdot x), \bar{\mu}_s\rangle\big) \left\langle \nabla(c\sigma(w\cdot x)) \cdot \nabla f, \eta_s^N \right\rangle \pi(dx,dy) \right) ds \\ & - \int_0^t \left(\int_{\mathcal{X}\times\mathcal{Y}} \alpha \left\langle c\sigma(w\cdot x), \eta_s^N \right\rangle \langle \nabla(c\sigma(w\cdot x)) \cdot \nabla f, \bar{\mu}_s \rangle \, \pi(dx,dy) \right) ds \\ & + \sqrt{N} \left\langle f, M_t^N \right\rangle + R_t^N\end{aligned}$$

where the remainder term

$$\lim_{N\to\infty} \mathbb{E}\left[\sup_{t\in[0,T]} |R_t^N| \right] = 0.$$

Relative compactness for fluctuation process

Lemma

Let $J_2 = 3\left\lceil \frac{D}{2} \right\rceil + 6$, $T < \infty$ and $r, t \in [0, T]$ with $(t - r) < \delta$. Then there are unimportant constants $C_0, C_1, C_2 < \infty$ such that

$$\begin{aligned} \sup_{N\in\mathbb{N}} \mathbb{E} \sup_{t\in[0,T]} \left\| \eta^N_t \right\|^2_{-J_2} &< C_0. \\ \mathbb{E}\left[\left\| \eta^N_t - \eta^N_r \right\|^2_{-J_2} \right] &\leq C_1\delta + C_2\frac{1}{N}. \end{aligned}$$

Due to the fact that the set $\left\{\phi \in W^{-(J_2+1),2} : \|\phi\|_{-J_2} \leq C_\epsilon\right\}$ is a compact subset of $W^{-(J_2+1),2}$, we obtain the process $\{\eta^N_\cdot\}_{N\in\mathbb{N}}$ is relatively compact in $W^{-J,2}(\Theta)$ with $J \geq J_2 + 1 = 3\left\lceil \frac{1+d}{2} \right\rceil + 7$.

Uniqueness and conclusion of the proof

- Similarly $\{\sqrt{N} M^N_\cdot\}_{N\in\mathbb{N}}$ is relatively compact in $W^{-J,2}(\Theta)$ with $J \geq 2\lceil\frac{1+d}{2}\rceil + 5$.
- The limit of $\sqrt{N} M^N_\cdot$ is a distribution valued Gaussian martingale with the appropriate covariance structure.
- The solution $\bar{\eta}$ to the limiting stochastic evolution equation is unique in $W^{-J,2}$ (assume two solutions, subtract them and using a-priori bounds show that the $W^{-J,2}$ norm of their difference is zero).

Part IV

Summary

Summary

- Mean field formulation of single layer neural networks.
- Rigorously proved convergence of the empirical measures of the parameters to the solution to a specific PDE.
- Rigorously proved convergence of the fluctuations to the empirical measures of the parameters to the solution to a SPDE.

Summary

- Mean field formulation of single layer neural networks.
- Rigorously proved convergence of the empirical measures of the parameters to the solution to a specific PDE.
- Rigorously proved convergence of the fluctuations to the empirical measures of the parameters to the solution to a SPDE.

- This is just the beginning of the story!
- Mean field formulation appears to be the way to go for quantitative results!
- Study of the limiting PDEs and SPDEs; properties etc.
- Other related limiting results (e.g. effect of initialization)

Part V

References

References. I

1. J. Sirignano and K. Spiliopoulos, *Stochastic gradient descent in continuous time*, SIAM Financial Mathematics, (2017), Vol. 8, Issue 1, pp. 933–961.
2. J. Sirignano and K. Spiliopoulos, *DGM: A deep learning algorithm for solving partial differential equations*, Journal of Computational Physics, (2018), to appear.
3. J. Inglis and D. Talay., *Mean-field limit of a stochastic particle system smoothly interacting through threshold hitting-times and applications to neural networks with dendritic component*, SIAM Journal on Mathematical Analysis, (2015), 47(5), pp. 3884-3916.
4. S. Mallat, *Understanding deep convolutional neural networks.*, Philosophical Transactions of the Royal Society A., (2016), 374.2065, 20150203.

References. II

5. C. Wang, J. Mattingly, and Y. Lu, *Scaling limit: Exact and tractable analysis of online learning algorithms with applications to regularized regression and PCA*, (2017), preprint.
6. S. Mei, A. Montanari, and P. Nguyen., *A mean field view of the landscape of two-layer neural networks*, (2018), April 18 arXiv preprint: 1804.06561.
7. Grant M. Rotskoff, Eric Vanden-Eijnden., *Neural Networks as Interacting Particle Systems: Asymptotic Convexity of the Loss Landscape and Universal Scaling of the Approximation Error, (2018), May 2 arXiv preprint: 1805.00915*
8. J. Sirignano and K. Spiliopoulos, *Mean Field Analysis of Neural Networks*, (2018), May 2 arXiv Preprint: 1805.01053.
9. J. Sirignano and K. Spiliopoulos, *Mean Field Analysis of Neural Networks: Central Limit Theorem*, (2018), Aug 28 arXiv Preprint: 1808.09372.

Thank You!!!!!

More Questions?