Quantitative Propagation of Chaos for SGD in Wide Neural Networks

Valentin De Bortoli[1]
Joint work with: Alain Durmus[1], Xavier Fontaine[1], Umut Şimşekli [2].

[1]ENS Paris-Saclay [2]Telecom ParisTech

Motivation

Classification/regression problems

Classical machine learning problems:

- house pricing, stock exchange prediction... (regression problems)
- medical applications, astrophysics... (classification problems)

Main properties:

- Supervised setting (very large training dataset).
- High-dimensional.
- Structured data.

Motivation

- **Overparametrized** neural networks perform well in many experimental settings, why?
- Does the training of neural networks exhibit a limit behavior when the **number of neurons is large**?
- When it exists, can we use the **limiting dynamics** to gain insights on the optimization procedure and obtain theoretical results on the convergence of the training procedure?

Energy landscape

Overparametrization has been extensively studied... In overparametrized neural networks, landscapes are *simpler*.

- Soltanolkotabi et al. (2019): one hidden layer $\Rightarrow$ local minima are global minima if $N \geq 2d$.
- Choromanska et al. (2015): multiple hidden layers (spin-glass model) large $N \Rightarrow$ critical points with low "energy" are local minima.

See also Pascanu et al. (2014), Pennington and Bahri (2017), Venturi et al. (2018), Soudry and Hoffer (2018) for similar results.
What can we say about the gradient descent when N is large?

Gradient descent

In what follows we assume that $(W_n^{k,N})_{n\in\mathbb{N}}$ is given by a SGD procedure.

In many cases, we can infer a **limiting dynamics** for $(W_n^{k,N})_{n\in\mathbb{N}}$.

- Chizat and Bach (2018); Rotskoff and Vanden-Eijnden (2018); Chizat (2019) – analysis using *Wasserstein gradient flow*,
- Sirignano and Spiliopoulos (2018, 2020); Mei et al. (2018) – analysis using *probabilistic mean field approximations* and *McKean-Vlasov SDE*.

One-layer neural network

Our setting: **One hidden layer neural network**.

- a loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0, +\infty)$, e.g. $\ell(x, y) = (x - y)^2$.
- a feature function $F : \underbrace{\mathbb{R}^p}_{\text{weights}} \times \underbrace{\mathbb{R}^d}_{\text{data}} \to \mathbb{R}$, e.g. $F(w, x) = \sigma(\langle w, x \rangle)$

  (σ is the sigmoid function).

Given x, estimator $\hat{y}$ given by $\hat{y} = N^{-1} \sum_{k=1}^{N} F(w^{k,N}, x)$.

We want to minimize the following **population risk**

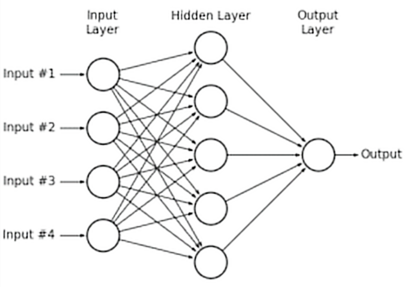
$$\boxed{\mathscr{R}^N(w^{1:N}) = \mathbb{E}_\pi\left[\ell(\hat{y}, y)\right] = \int_{(x,y)\in\mathbb{R}^d\times\mathbb{R}} \ell\left(\frac{1}{N}\sum_{k=1}^{N} F(w^{k,N}, x), y\right) \mathrm{d}\pi(x, y) \ ,}$$

where π is the distribution of the data and $w^{1:N} = (w^{k,N})_{k\in\{1,\ldots,N\}} \in (\mathbb{R}^d)^N$.

Question: what can we say when $N \gg d$, i.e. when the network is overparametrized?

A first example

$$\boxed{\mathscr{R}^N(w^{1:N}) = \int_{(x,y)\in\mathbb{R}^d\times\mathbb{R}} \ell\left(N^{-1}\sum_{k=1}^{N} F(w^{k,N},x), y\right) \mathrm{d}\pi(x,y)\ .}$$



In this case $d = 4$ and $N = 5$.

Mean field approximation

A mean-field formulation

We recall that we want to minimize

$$\begin{aligned}\mathscr{R}^N(w^{1:N}) &= \int_{(x,y)\in\mathbb{R}^d\times\mathbb{R}} \ell\left(N^{-1}\sum_{k=1}^N F(w^{k,N},x),y\right)\mathrm{d}\pi(x,y)\\ &= \int_{(x,y)\in\mathbb{R}^d\times\mathbb{R}} \hat{\mathscr{R}}^N(w^{1:N},x,y)\mathrm{d}\pi(x,y)\ ,\end{aligned}$$

Stochastic **G**radient **D**escent (SGD):

$$\begin{aligned}&W_{n+1}^{k,N} - W_n^{k,N} = -\gamma_N \partial_{w^{k,N}}\hat{\mathscr{R}}^N(W_n^{1:N},X_n,Y_n)\\ &= -\frac{\gamma_N}{N}\left\{\mathbb{E}_\pi\left[N\partial_{w^{k,N}}\hat{\mathscr{R}}^N(W_n^{1:N},\cdot,\cdot)\right] + N\partial_{w^{k,N}}\hat{\mathscr{R}}^N(W_n^{1:N},X_n,Y_n) - \mathbb{E}_\pi\left[(N\partial_{w^{k,N}}\hat{\mathscr{R}}^N(W_n^{1:N},\cdot,\cdot)\right]\right\}\\ &= -\frac{\gamma_N}{N}\left\{h(W_n^{k,N},\nu_n) + \eta_n(W_n^{k,N},\nu_n)\right\}\ ,\end{aligned}$$

where γ_N is a step-size, $\nu_n = N^{-1}\sum_{k=1}^N \delta_{W_n^{k,N}}$ **(empirical measure)**, (X_n, Y_n) i.i.d. and

$$\begin{cases} H(w,\nu,x,y) = \partial_1\ell\left(\int_{\mathbb{R}^p} F(w,x)\mathrm{d}\nu(w),y\right)\nabla F(w,x)\ ,\\ h(w,\nu) = \int_{(x,y)\in\mathbb{R}^d\times\mathbb{R}} H(w,\nu,x,y)\mathrm{d}\pi(x,y)\ ,\\ \eta_n(w,\nu) = H(w,\nu,x_n,y_n) - \int_{(x,y)\in\mathbb{R}^d\times\mathbb{R}} H(w,\nu,x,y)\mathrm{d}\pi(x,y)\ .\end{cases}$$

A continuous-time approximation

Stochastic **G**radient **D**escent (SGD):

$$W_{n+1}^{k,N} = W_n^{k,N} - (\gamma_N/N)\left\{h(W_n^{k,N},\nu_n) + \eta_n(W_n^{k,N},\nu_n)\right\} .$$

h is called the **mean field** approximation. We will work with the **continous-time** version of SGD

$$\mathrm{dW}_t^{k,N} = h(\mathrm{W}_t^{k,N},\boldsymbol{\nu}_t^N)\mathrm{d}t + (\gamma_N/N)^{1/2}\Sigma^{1/2}(\mathrm{W}_t^{k,N},\boldsymbol{\nu}_t^N)\mathrm{dB}_t^k ,$$

with $\Sigma(w,\nu) = \mathrm{Cov}_\pi[H(w,\nu,\cdot,\cdot)]$ and $(\mathrm{B}_t)_{t\geq 0}$ Brownian motion.

Approximation results

If F is regular enough with bounded derivatives and bounded and if ℓ is regular enough with bounded second-order derivatives then for any $T \geq 0$, there exists $C \geq 0$ such that for any $t \in [0,T]$, $N \in \mathbb{N}$ and $k \in \{1,\dots,N\}$

$$\boxed{\sup_{t\in[0,T]} \mathbb{E}^{1/2}\left[\left\|\mathrm{W}_t^{k,N} - W_{\lfloor Nt/\gamma_N\rfloor}^{k,N}\right\|^2\right] \leq C(\gamma_N/N)^{1/2}\log(1+(\gamma_N/N)^{-1}) .}$$

From deterministic to stochastic

Approximation: We only need to consider

$$\mathrm{d}\mathrm{W}_t^{k,N} = h(\mathrm{W}_t^{k,N}, \boldsymbol{\nu}_t^N)\mathrm{d}t + (\gamma_N/N)^{1/2}\Sigma^{1/2}(\mathrm{W}_t^{k,N}, \boldsymbol{\nu}_t^N)\mathrm{d}\mathrm{B}_t^k \ .$$

Until now $\rightsquigarrow$ γ_N does not depend on N, see Mei et al. (2019, 2018); Sirignano and Spiliopoulos (2020, 2018); Chizat (2019); Chizat and Bach (2018); Rotskoff and Vanden-Eijnden (2018).

Our observation: with $\gamma_N = \gamma$ for all $N \in \mathbb{N}^*$ the obtained limiting dynamics **is not stochastic anymore**.

$\rightsquigarrow$ In what follows, we consider $\gamma_N = \gamma N^{\beta}$.

Propagation of chaos

Question: what can we say about the law of **a fixed number** of particles when the total number of particles grow towards $+\infty$?

$\rightarrow$ propagation of chaos, see Sznitman (1991); Gottlieb (2000); Jourdain and Méléard (1998)...

Propagation of chaos

The **chaos propagates** if for any $t \geq 0$ and $j \in \mathbb{N}$

$$\lim_{N \to +\infty} \mathcal{L}((W_t^{1,N}, \ldots, W_t^{j,N})) = (\lambda_t^\star)^{\otimes j} ,$$

for some distribution $\lambda_t^\star$.

- independence between the particles when $N \to +\infty$.
- the particles have identical laws, $\lambda_t^\star$.

A basic result

Adapted from Sznitman (1991)

If for any $N \in \mathbb{N}$, $\mathcal{L}(W_0^N) = \rho^{\otimes N}$ with $\int_{\mathbb{R}^d} \|x\|^2 \mathrm{d}\rho(x) < +\infty$ and

$$\mathrm{dW}_t^{k,N} = b(\mathrm{W}_t^{k,N}, \boldsymbol{\nu}_t^N)\mathrm{d}t + \Sigma(\mathrm{W}_t^{k,N}, \boldsymbol{\nu}_t^N)\mathrm{dB}_t^k \ ,$$

with for any $w_1, w_2 \in \mathbb{R}^d$ and $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$

$$\|b(w_1, \mu_1) - b(w_2, \mu_2)\| + \|\Sigma(w_1, \mu_1) - \Sigma(w_2, \mu_2)\| \\ \leq \mathrm{L}\left\{\|w_1 - w_2\| + \|\mu_1[f] - \mu_2[f]\|\right\} \ , \quad (1)$$

with $f : \mathbb{R}^d \to \mathbb{R}^d$ Lipschitz. Then for any $T \geq 0$ and $j, N \in \mathbb{N}$ with $N \geq j$

$$\mathbb{E}[\sup_{t\in[0,T]} \|\mathrm{W}_t^{1:j,N} - \mathrm{W}_t^{1:j,\star}\|^2] \leq C_{T,j} N^{-1} \ .$$

with

$$\mathrm{dW}_t^{k,\star} = b(\mathrm{W}_t^{k,\star}, \lambda_t^\star)\mathrm{d}t + \Sigma(\mathrm{W}_t^{k,\star}, \lambda_t^\star)\mathrm{dB}_t^k \ .$$

(McKean-Vlasov process)

The deterministic regime

First case: $\gamma_N = \gamma N^\beta$ with $\beta \in [0,1)$.

Convergence result (I)

For any $\ell \in \mathbb{N}^*$, $T \geq 0$, there exists $C_T \geq 0$ such that for any $\beta \in [0,1)$,

$$\boxed{\mathbb{E}\left[\sup_{t\in[0,T]} \|W_t^{\ell,N} - W_t^{\ell,\star}\|^2\right] \leq C_T N^{-(1-\beta)}\ ,}$$

with

$$\mathrm{d}W_t^{\ell,\star} = h(W_t^{\ell,\star}, \lambda_t^{\ell,\star})\mathrm{d}t\ , \qquad \text{with } \lambda_t^{\ell,\star} \text{ the distribution of } W_t^{\ell,\star}\ .$$

- Deterministic McKean-Vlasov limit (ODE mean-field).
- Rate of convergence $N^{1-\beta}$.
- For any $\ell_1, \ell_2 \in \mathbb{N}$, $\lambda_t^{\ell_1,\star} = \lambda_t^{\ell_2,\star}$.

The stochastic regime

Second case: $\gamma_N = \gamma N^{\beta}$ with $\beta = 1$.

Convergence result (II)

For any $\ell \in \mathbb{N}^*$, $T \geq 0$, there exists $C_T \geq 0$ such that

$$\boxed{\mathbb{E}\left[\sup_{t\in[0,T]} \|\mathrm{W}_t^{\ell,N} - \mathrm{W}_t^{\ell,\star}\|^2\right] \leq C_T N^{-1}\ .}$$

with

$$\mathrm{dW}_t^{\ell,\star} = h(\mathrm{W}_t^{\ell,\star}, \lambda_t^{\ell,\star})\mathrm{d}t + \gamma^{1/2}\Sigma(\mathrm{W}_t^{\ell,\star}, \lambda_t^{\ell,\star})\mathrm{dB}_t^{\ell}\ .$$

- Stochastic McKean-Vlasov limit (SDE mean-field).
- Convergence rate N^{-1}.
- For any $\ell_1, \ell_2 \in \mathbb{N}$, $\lambda_t^{\ell_1,\star} = \lambda_t^{\ell_2,\star}$.

Stochastic/Deterministic

- depending on the scaling $\gamma_N \sim \gamma N^\beta$ we obtain two different **regimes**.
- For $\beta \in [0, 1)$, the SDE is an **ODE** and $\lambda_t^\star$ satisfies the following **Fokker-Planck equation**.

$$\boxed{\partial_t \lambda_t^\star = -\sum_{i=1}^{d} \partial_i(\lambda_t^\star h_i)\ .}$$

- For $\beta = 1$, the SDE is an **SDE** and $\lambda_t^\star$ also satisfies a Fokker-Planck equation

$$\boxed{\partial_t \lambda_t^\star = -\sum_{i=1}^{N} \partial_i(\lambda_t^\star h_i) + (\gamma/2)\sum_{i=1}^{d}\sum_{j=1}^{d} \partial_{i,j}(\lambda_t^\star \Sigma_{i,j})\ .}$$

- Larger stepsizes enforce some **entropic regularization** of the model.

Experiments

A toy experiment

- MNIST dataset $\rightarrow$ classification task between ten digits.
- Fully connected, one hidden layer.
- ReLU activation function.
- Cross-entropy loss.

Question : what happens when we train SGD for N large with stepsize $\gamma_N = \gamma N^{\beta}$ and $\beta \in [0, 1]$?

Different regimes

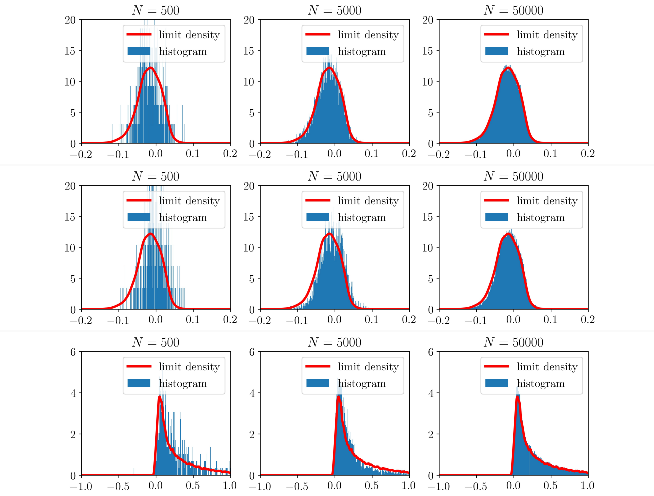


Figure 1: First line $\beta = 0.5$, second $\beta = 0.75$, third $\beta = 1$ (recall $\gamma_N = \gamma N^{\beta}$)

From stochastic to deterministic

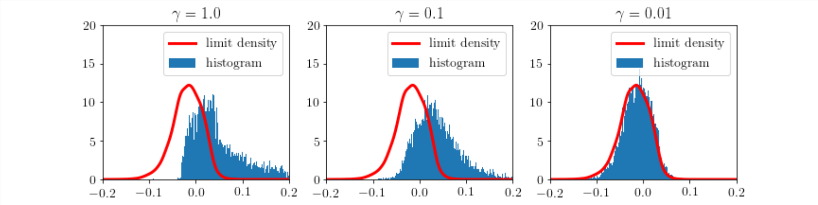


Figure 2: as $\gamma \to 0$ we converge towards the deterministic model.

Regularization effect

Values of N and β	$N = 5000$, $\beta = 0.75$	$N = 5000$, $\beta = 1.0$	$N = 10000$, $\beta = 0.75$	$N = 10000$, $\beta = 1.0$
Train acc.	**100%**	97.2%	**100%**	97.2%
Test acc.	55.5%	**56.5%**	56.0%	**56.5%**

Table 1: $\beta = 1$ setting exhibits better regularization properties.

Conclusion

The study of overparametrized (wide) neural networks gives some insights on what happens when we optimize neural networks...

- Limiting dynamics
- Independence of weights
- Equivalence with PDE evolutions

Ongoing work:

- Propagation of chaos = Law of large numbers, how about a CLT? Sirignano and Spiliopoulos (2020),
- Extension to deep networks, is the analysis still valid? What kind of behavior is specific to the **depth** of the network?
- Stationary solutions of the PDE are not easy to compute $\rightarrow$ fixed point equations. Properties of these solutions?

Thank your for your attention!

Our paper: `https://arxiv.org/abs/2007.06352`

Bibliography i

References

L. Chizat. Sparse optimization on measures with over-parameterized gradient descent. *arXiv preprint arXiv:1907.10300*, 2019.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3036–3046, 2018.

L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2933–2943, 2019.

Anna Choromanska, Mikael Henaff, Michaël Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, 2015.

B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems*, pages 9108–9118, 2019.

A. Gottlieb. Markov transitions and the propagation of chaos. *arXiv preprint math/0001076*, 2000.

B. Jourdain and S. Méléard. Propagation of chaos and fluctuations for a moderate model with smooth initial data. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 34, pages 727–766. Elsevier, 1998.

S. Mei, A. Montanari, and P. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. *arXiv preprint arXiv:1902.06015*, 2019.

R. Pascanu, Y. N Dauphin, S. Ganguli, and Y. Bengio. On the saddle point problem for non-convex optimization. *arXiv preprint arXiv:1405.4604*, 2014.

Bibliography ii

J. Pennington and Y. Bahri. Geometry of neural network loss surfaces via random matrix theory. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2798–2806. JMLR. org, 2017.

G. M. Rotskoff and E. Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach, 2018.

J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks. *arXiv preprint arXiv:1805.01053*, 2018.

J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

M. Soltanolkotabi, A. Javanmard, and J. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Trans. Information Theory*, 65(2):742–769, 2019. doi: 10.1109/TIT.2018.2854560. URL https://doi.org/10.1109/TIT.2018.2854560.

Daniel Soudry and Elad Hoffer. Exponentially vanishing sub-optimal local minima in multilayer neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*, 2018.

A. Sznitman. Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*, pages 165–251. Springer, 1991.

L. Venturi, A. Bandeira, and J. Bruna. Neural networks with finite intrinsic dimension have no spurious valleys. *CoRR*, abs/1802.06384, 2018. URL http://arxiv.org/abs/1802.06384.

A first model

$$\boxed{\mathscr{R}^N(w^{1:N}) = \int_{(x,y)\in\mathbb{R}^d\times\mathbb{R}} \ell\left(N^{-\delta}\sum_{k=1}^{N} F(w^{k,N}, x), y\right) \mathrm{d}\pi(x, y)\ ,}$$

with $\delta \in [0, 1)$. (Recall that in the previous setting $\delta = 1$)

In this case, **lazy training** occurs, see Chizat et al. (2019). Why **lazy**? $\to$ weights don't move a lot, see `https://rajatvd.github.io/NTK/`.

SGD is provably close to a **linear model**, i.e. **N**eural **T**angent **K**ernel (NTK) gradient descent.

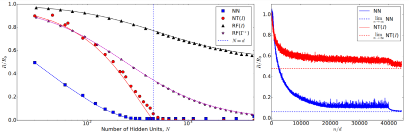


Figure 3: Figure extracted from Ghorbani et al. (2019) $\to$ poor performance of NTK.

Comparison

- Only the case $\beta = 0$ has been previously studied: Sirignano and Spiliopoulos (2018); Mei et al. (2018); Chizat and Bach (2018); Rotskoff and Vanden-Eijnden (2018); Sirignano and Spiliopoulos (2020).
- **Weak convergence** of SGD (Sirignano and Spiliopoulos, 2018, Theorem 1.6), (Mei et al., 2018, Theorem 3)(high probability)
- **Central limit theorem** (Sirignano and Spiliopoulos, 2020, Theorem 1.5);
- (Chizat and Bach, 2018, Theorem 2.6) and (Rotskoff and Vanden-Eijnden, 2018, Proposition 3.2) $\rightarrow$ gradient flows techniques + convergence if strongly convex.