Quantitative Propagation of Chaos for SGD in Wide Neural Networks

Valentin De Bortoli⋆, Alain Durmus†, Xavier Fontaine†, Umut Simsekli□ ◦,

⋆University of Oxford, †Université Paris-Saclay, □LTCI, Télécom Paris, ◦INRIA Paris

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation & Contribution

- **Overparameterized** neural networks (*i.e.* with a very large number of neurons N) are highly efficient in practice. This seems in contradiction with classical statistical learning theory (overfitting phenomenon).
- Our contribution: theoretical analysis of overparameterized neural networks. We identify a **propagation of chaos phenomenon** [4, 3] and investigate the limiting dynamics of the Stochastic Gradient Descent (SGD) when $N \to +\infty$.
- We identify two regimes (**McKean-Vlasov processes**) depending on the scaling of the stepsize in SGD with the number of neurons.
- In the second regime, large stepsizes act as an **implicit regularizer**.
- We draw connections with the **Wasserstein gradient flow** approach [1, 2].

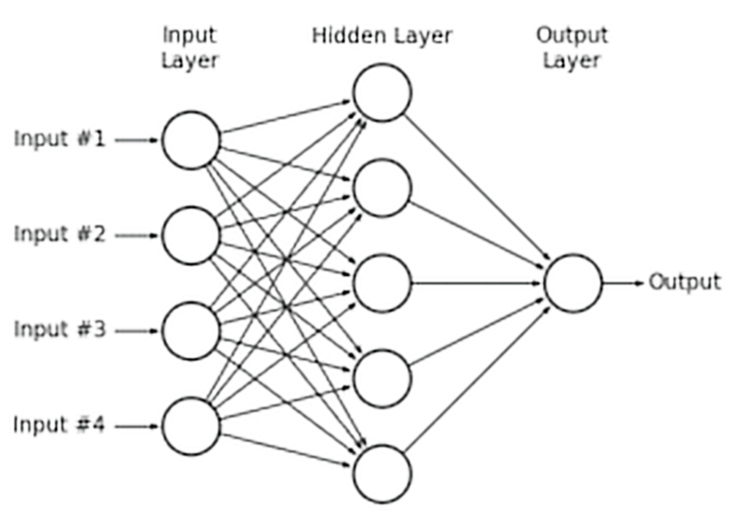
Mean-field formulation

We aim at minimizing the **structural risk**

$$\boxed{\mathscr{R}^N(w^{1:N}) = \int_{\mathsf{X}\times\mathsf{Y}} \ell\left(\tfrac{1}{N}\sum_{k=1}^N F(w^{k,N}, x), y\right) \mathrm{d}\pi(x,y) + \tfrac{1}{N}\sum_{k=1}^N V(w^{k,N})}\ ,$$

where

- $w^{1:N} = \{w^{k,N}\}_{k=1}^N$ are the weights of the neural network,
- F is a feature function, *e.g.* $F(w,x) = \mathrm{sigmoid}(\langle w, x\rangle)$,
- ℓ is a loss function,
- V is a regularizer (optional),
- π is the distribution of the pair data/label.



The **SGD recursion**:

$$W_{n+1}^{1:N} = W_n^{1:N} - \gamma N^{\beta} \nabla \hat{\mathscr{R}}^N(W_n^{1:N}, X_n, Y_n)\ ,$$

with $\hat{\mathscr{R}}^N$ the **empirical risk**

$$\hat{\mathscr{R}}^N(w^{1:N}, x, y) = \ell\left(\tfrac{1}{N}\sum_{k=1}^N F(w^{k,N}, x), y\right) + \tfrac{1}{N}\sum_{k=1}^N V(w^{k,N})\ .$$

Let

$$h(w,\mu) = -\int_{\mathsf{X}\times\mathsf{Y}} \partial_1 \ell\left(\mu[F(\cdot,x)], y\right) \nabla_w F(w,x)\, \mathrm{d}\pi(x,y) - \nabla V(w)\ ,$$
$$\xi(w,\mu,x,y) = -h(w,\mu) - \partial_1\ell(\mu[F(\cdot,x)], y)\nabla_w F(w,x) - \nabla V(w)\ .$$

Then the SGD recursion can be written in a **mean-field formulation**

$$\boxed{W_{n+1}^{k,N} = W_n^{k,N} + \gamma N^{\beta-1}\left\{h(W_n^{k,N}, \nu_n^N) + \xi(W_n^{k,N}, \nu_n^N, X_n, Y_n)\right\}}\ ,$$

where ν_n^N is the **empirical measure** $\nu_n^N = (1/N)\sum_{k=1}^N \updelta_{W_n^{k,N}}$.

We study the **continuous-time counterpart** of SGD given by the following Stochastic Differential Equation (provably close to the original process for small values of $\gamma N^{\beta-1}$)

$$\boxed{\mathrm{d}\mathbf{W}_t^{k,N} = h(\mathbf{W}_t^{k,N}, \boldsymbol{\nu}_t^N)\mathrm{d}t + (\gamma N^{\beta-1})^{1/2}\Sigma^{1/2}(\mathbf{W}_t^{k,N}, \boldsymbol{\nu}_t^N)\mathrm{d}\mathbf{B}_t^k}\ , \qquad (1)$$

where $\Sigma(w,\mu) = \int_{\mathsf{X}\times\mathsf{Y}} \xi(w,\mu,x,y)\xi(w,\mu,x,y)^\top \mathrm{d}\pi(x,y)$ and $\boldsymbol{\nu}_t^N$ is the empirical measure $\boldsymbol{\nu}_t^N = (1/N)\sum_{k=1}^N \updelta_{\mathbf{W}_t^{k,N}}$.

Propagation of chaos

We identify **two regimes** depending on the value of $\beta \in [0,1]$.

■ **Deterministic** regime: $\beta \in [0,1)$.

Define the following **McKean-Vlasov** SDE

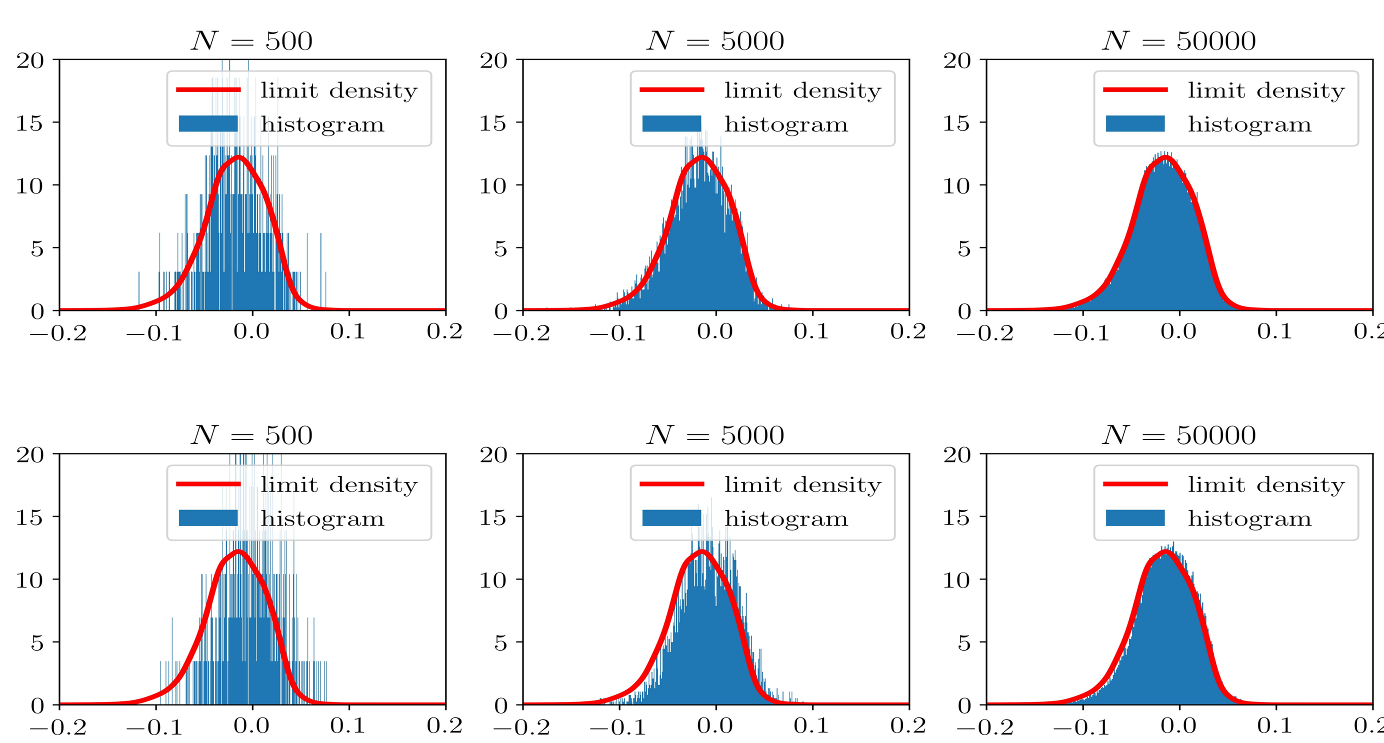
$$\boxed{\mathrm{d}\mathbf{W}_t^\star = h(\mathbf{W}_t^\star, \boldsymbol{\lambda}_t^\star)\mathrm{d}t\ , \quad \text{with } \boldsymbol{\lambda}_t^\star \text{ the distribution of } \mathbf{W}_t^\star\ .} \qquad (2)$$

Theorem 1. *Let $(\mathbf{W}_0^k)_{k\in\mathbb{N}}$ be a sequence of i.i.d. $\mathbb{R}^p$-valued random variables with distribution $\mu_0 \in \mathscr{P}_2(\mathbb{R}^p)$ and set for any $N \in \mathbb{N}^\star$, $\mathbf{W}_0^{1:N} = \{\mathbf{W}_0^k\}_{k=1}^N$. Then, for any $m \in \mathbb{N}^\star$ and $T \geq 0$, there exists $C_{m,T} \geq 0$ such that for any $\beta \in [0,1)$ and $N \in \mathbb{N}^\star$ with $N \geq m$*

$$\boxed{\mathbb{E}\left[\sup_{t\in[0,T]} \|\mathbf{W}_t^{1:m,N} - \mathbf{W}_t^{1:m,\star}\|^2\right] \leq C_{m,T} N^{-(1-\beta)}}\ ,$$

with $(\mathbf{W}_t^{1:m,N}, \mathbf{W}_t^{1:m,\star}) = \{(\mathbf{W}_t^{k,N}, \mathbf{W}_t^{k,\star})\}_{k=1}^m$, $(\mathbf{W}_t^{1:N})$ solution of (1) starting from $\mathbf{W}_0^{1:N}$, and for any $k \in \mathbb{N}^\star$, $\mathbf{W}_t^{k,\star}$ solution of (2) starting from $\mathbf{W}_0^k$.

- $(\mathbf{X}_t^{k,\star})_{k\in\mathbb{N}}$ are i.i.d. (**propagation of chaos** result),
- the limiting McKean-Vlasov SDE is deterministic (no Brownian motion),
- same limiting dynamic for any $\beta \in [0,1)$ (first row, empirical measure $\beta = .5$, second row, empirical measure $\beta = .75$)



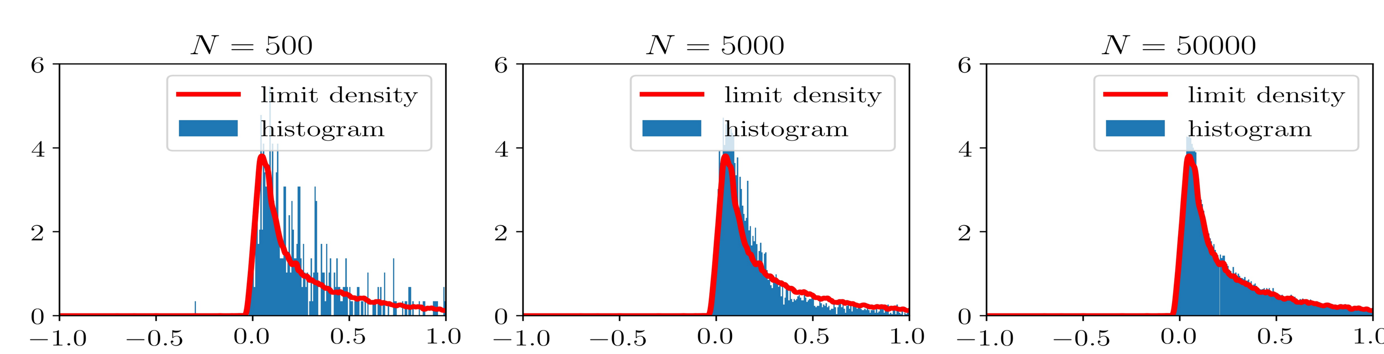
■ **Stochastic** regime: $\beta = 1$.

$$\boxed{\mathrm{d}\mathbf{W}_t^\star = h(\mathbf{W}_t^\star, \boldsymbol{\lambda}_t^\star)\mathrm{d}t + (\gamma\Sigma(\mathbf{W}_t^\star, \boldsymbol{\lambda}_t^\star))^{1/2}\mathrm{d}\mathbf{B}_t}\ . \qquad (3)$$

Theorem 2. *For any $m \in \mathbb{N}^\star$ and $T \geq 0$, there exists $C_{m,T} \geq 0$ such that for any $N \in \mathbb{N}^\star$ with $N \geq m$*

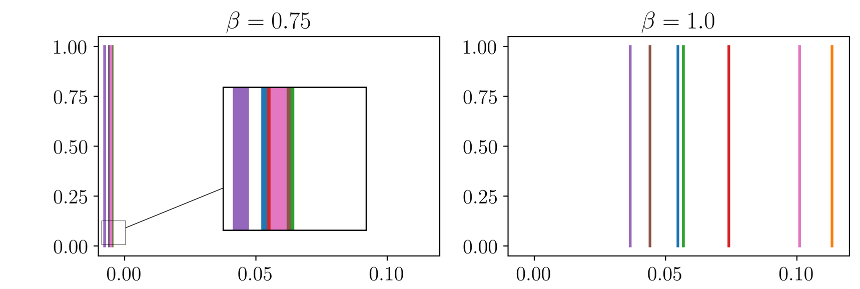
$$\boxed{\mathbb{E}\left[\sup_{t\in[0,T]} \|\mathbf{W}_t^{1:m,N} - \mathbf{W}_t^{1:m,\star}\|^2\right] \leq C_{m,T} N^{-1}}\ ,$$

with $(\mathbf{W}_t^{1:m,N}, \mathbf{W}_t^{1:m,\star}) = \{(\mathbf{W}_t^{k,N}, \mathbf{W}_t^{k,\star})\}_{k=1}^m$, $(\mathbf{W}_t^{1:N})$ solution of (1) starting from $\mathbf{W}_0^{1:N}$, and for any $k \in \mathbb{N}^\star$, $\mathbf{W}_t^{k,\star}$ solution of (3) starting from $\mathbf{W}_0^k$ and Brownian motion $(\mathbf{B}_t^k)_{t\geq 0}$.



Deterministic VS Stochastic

For $\beta \in [0,1)$, SGD converges towards a deterministic dynamics (same as GD). For $\beta = 1$ the limiting dynamics remains stochastic.



Each bar corresponds to the position of $W_T^{1,N}$ for large $N \in \mathbb{N}$ and $T \geq 0$ and different random seeds but same initialization.

Connection with Wasserstein gradient flows

Links with Wasserstein gradient flow approaches [1].

■ **Deterministic** regime: $(\boldsymbol{\lambda}_t^\star)_{t\geq 0}$ satisfies the Partial Differential Equation (PDE)

$$\partial_t \boldsymbol{\lambda}_t^\star(w) = -\mathrm{div}(h(\cdot, \boldsymbol{\lambda}_t^\star)\boldsymbol{\lambda}_t^\star)(w)\ ,$$

This is the gradient flow associated with

$$\mathscr{R}^\star(\rho) = \int_{\mathsf{X}\times\mathsf{Y}} \ell\left(\int_{\mathbb{R}^p} F(\tilde{w}, x)\mathrm{d}\rho(\tilde{w}), y\right) \mathrm{d}\pi(x,y)\ ,$$

■ **Stochastic** regime: $(\boldsymbol{\lambda}_t^\star)_{t\geq 0}$ satisfies the PDE

$$\partial_t \boldsymbol{\lambda}_t^\star(w) = -\mathrm{div}(h(\cdot, \boldsymbol{\lambda}_t^\star)\boldsymbol{\lambda}_t^\star)(w) + (\gamma/2)\sum_{i,j}\partial_{i,j}(\Sigma_{i,j}(\cdot, \boldsymbol{\lambda}_t^\star)\boldsymbol{\lambda}_t^\star)(w)\ .$$

If $\Sigma = \theta\,\mathrm{Id}$, this is the gradient flow associated with $\mathscr{R}^\star + (\gamma\theta/2)\mathrm{Ent}$, where

$$\mathrm{Ent}(\rho) = -\int_{\mathbb{R}^p} \rho(x)\log(\rho(x))\mathrm{d}x\ .$$

Hence large stepsizes correspond to an **implicit regularization** of the risk $\mathscr{R}^\star$. Better **generalization properties** (MNIST classification task)

Values of N and β	$N = 5000$ $\beta = 0.75$	$N = 5000$ $\beta = 1.0$	$N = 10000$ $\beta = 0.75$	$N = 10000$ $\beta = 1.0$	$N = 50000$ $\beta = 0.75$	$N = 50000$ $\beta = 1.0$
Train acc.	**100%**	97.2%	**100%**	97.2%	**100%**	99%
Test acc.	55.5%	**56.5%**	56.0%	**56.5%**	56.7%	**57.7%**

Acknowledgements

V. De Bortoli was partially supported by EPSRC grant EP/R034710/1. X. Fontaine was supported by grants from Région Ile-de-France. The contribution of U. Simsekli to this work is partly supported by the French National Research Agency (ANR) as a part of the FBIMATRIX (ANR-16-CE23-0014) project.

References

References

[1] L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Neurips*, 2018.

[2] S. Mei, A. Montanari, and P. Nguyen. A mean field view of the landscape of two-layer neural networks. *PNAS*, 2018.

[3] J. Sirignano and K. Spiliopoulos. Mean field analysis of deep neural networks. *arXiv preprint arXiv:1903.04440*, 2019.

[4] A. Sznitman. Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*. 1991.