Fluctuations and Concentration in Two-layer Neural Networks

Paul Stos

Early-Career Researchers' Day of the DATA Programme I-Site CAP 20-25

February 26, 2026

LMBP UMR 6620, Université Clermont-Auvergne

Plan

1. Setting
2. Mean-field limit
3. Fluctuation process
4. Concentration bounds

Setting

Supervised learning for regression

Data points: $(x_k, y_k) \overset{i.i.d.}{\sim} \pi \qquad x \in \mathbb{R}^d,\ y \in \mathbb{R}.$

Goal: Find $\hat{y}$ s.t. $\hat{y}(x) \approx y$ for previously unseen x.

Supervised learning for regression

Data points: $(x_k, y_k) \overset{i.i.d.}{\sim} \pi \qquad x \in \mathbb{R}^d,\ y \in \mathbb{R}.$

Goal: Find $\hat{y}$ s.t. $\hat{y}(x) \approx y$ for previously unseen x.

Two-layer neural network:

$$\hat{y}_\theta(x) = \frac{1}{N} \sum_{i=1}^{N} \sigma_*(\theta^i, x),$$

where

- $N \geq 1$ is the number of *neurons*;
- $\theta = (\theta^i)_{i=1}^N \in (\mathbb{R}^D)^N$ are the *parameters*;
- $\sigma_* : \mathbb{R}^D \times \mathbb{R}^d \to \mathbb{R}$ is the *activation function*, e.g.,

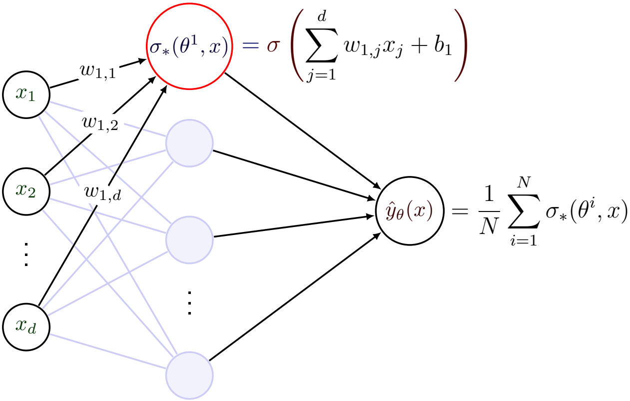
$$\theta^i = (w_i, b_i) \in \mathbb{R}^{d+1}, \quad \sigma_*(\theta^i, x) = \sigma(w_i \cdot x + b_i).$$

Supervised learning for regression

Data points: $(x_k, y_k) \overset{i.i.d.}{\sim} \pi \qquad x \in \mathbb{R}^d,\ y \in \mathbb{R}.$

Goal: Find $\hat{y}$ s.t. $\hat{y}(x) \approx y$ for previously unseen x.

Two-layer neural network:

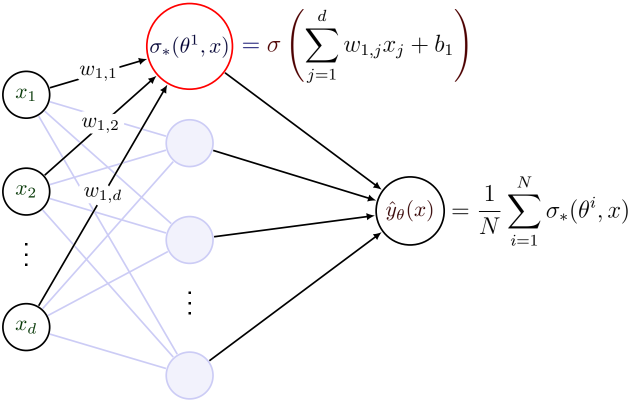


Supervised learning for regression

Data points: $(x_k, y_k) \overset{i.i.d.}{\sim} \pi \qquad x \in \mathbb{R}^d,\ y \in \mathbb{R}.$

Goal: Find $\theta \in (\mathbb{R}^D)^N$ s.t. $\hat{y}_\theta(x) \approx y$ for previously unseen x.

Two-layer neural network:



Supervised learning for regression

Data points: $(x_k, y_k) \overset{i.i.d.}{\sim} \pi \qquad x \in \mathbb{R}^d,\ y \in \mathbb{R}.$

Goal: Find $\theta \in (\mathbb{R}^D)^N$ s.t. $\hat{y}_\theta(x) \approx y$ for previously unseen x.

Two-layer neural network: $\hat{y}_\theta(x) = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \sigma_*(\theta^i, x).$

Supervised learning for regression

Data points: $(x_k, y_k) \overset{i.i.d.}{\sim} \pi \qquad x \in \mathbb{R}^d,\ y \in \mathbb{R}.$

Goal: Find $\theta \in (\mathbb{R}^D)^N$ s.t. $\hat{y}_\theta(x) \approx y$ for previously unseen x.

Two-layer neural network: $\hat{y}_\theta(x) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \sigma_*(\theta^i, x).$

Ideally, we choose θ so as so minimize the *population* risk

$$\mathcal{R}(\theta) = \mathbb{E}_\pi\left[\ell(y, \hat{y}_\theta(x))\right]$$

for some *loss function* $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, e.g., $\ell(y, y') = \frac{1}{2}(y - y')^2$.

Supervised learning for regression

Data points: $(x_k, y_k) \overset{i.i.d.}{\sim} \pi$ unknown $\quad x \in \mathbb{R}^d$, $y \in \mathbb{R}$.

Goal: Find $\theta \in (\mathbb{R}^D)^N$ s.t. $\hat{y}_\theta(x) \approx y$ for previously unseen x.

Two-layer neural network: $\hat{y}_\theta(x) = \frac{1}{N} \sum_{i=1}^{N} \sigma_*(\theta^i, x)$.

Ideally, we choose θ so as so minimize the *population* risk

$$\mathcal{R}(\theta) = \mathbb{E}_\pi \left[\ell(y, \hat{y}_\theta(x)) \right]$$

for some *loss function* $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, e.g., $\ell(y, y') = \frac{1}{2}(y - y')^2$.

Supervised learning for regression

Data points: $(x_k, y_k) \overset{i.i.d.}{\sim} \pi$ unknown $\quad x \in \mathbb{R}^d$, $y \in \mathbb{R}$.

Goal: Find $\theta \in (\mathbb{R}^D)^N$ s.t. $\hat{y}_\theta(x) \approx y$ for previously unseen x.

Two-layer neural network: $\hat{y}_\theta(x) = \frac{1}{N} \sum_{i=1}^{N} \sigma_*(\theta^i, x)$.

Ideally, we choose θ so as so minimize the *population* risk

$$\mathcal{R}(\theta) = \mathbb{E}_\pi \left[\ell(y, \hat{y}_\theta(x)) \right]$$

for some *loss function* $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, e.g., $\ell(y, y') = \frac{1}{2}(y - y')^2$.

In practice, we choose θ so as to minimise the *empirical* risk

$$\widehat{\mathcal{R}}(\theta, x_k, y_k) = \ell(y_k, \hat{y}_\theta(x_k)) + \Omega(\theta).$$

(Online) Stochastic Gradient Descent

Initialization: $\theta_0^i \overset{i.i.d.}{\sim} \mu_0 \in \mathcal{P}(\mathbb{R}^D)$.

Update: SGD with fixed *learning rate* $\alpha > 0$,

$$\theta_{k+1} = \theta_k - \alpha \nabla \widehat{\mathcal{R}}(\theta_k, x_k, y_k) \qquad \text{(square loss)}$$

(Online) Stochastic Gradient Descent

Initialization: $\theta_0^i \overset{i.i.d.}{\sim} \mu_0 \in \mathcal{P}(\mathbb{R}^D)$.

Update: SGD with fixed *learning rate* $\alpha > 0$,

$$\begin{aligned}\theta_{k+1} &= \theta_k - \alpha \nabla \widehat{\mathcal{R}}(\theta_k, x_k, y_k) \qquad \text{(square loss)} \\ &= \theta_k - \frac{\alpha}{N}\big(y_k - \hat{y}_{\theta_k}(x_k)\big) \nabla_{\theta^i} \sigma_*(\theta_k^i, x_k).\end{aligned}$$

(Online) Stochastic Gradient Descent

Initialization: $\theta_0^i \overset{i.i.d.}{\sim} \mu_0 \in \mathcal{P}(\mathbb{R}^D)$.

Update: SGD with fixed *learning rate* $\alpha > 0$,

$$\begin{aligned}\theta_{k+1} &= \theta_k - \alpha \nabla \widehat{\mathcal{R}}(\theta_k, x_k, y_k) \qquad \text{(square loss)} \\ &= \theta_k - \frac{\alpha}{N}\big(y_k - \hat{y}_{\theta_k}(x_k)\big) \nabla_{\theta^i} \sigma_*(\theta_k^i, x_k).\end{aligned}$$

Each data point is used only *once* and never revisited.

(Online) Stochastic Gradient Descent

Initialization: $\theta_0^i \overset{i.i.d.}{\sim} \mu_0 \in \mathcal{P}(\mathbb{R}^D)$.

Update: SGD with fixed *learning rate* $\alpha > 0$,

$$\begin{aligned}\theta_{k+1} &= \theta_k - \alpha \nabla \widehat{\mathcal{R}}(\theta_k, x_k, y_k) \qquad \text{(square loss)}\\ &= \theta_k - \frac{\alpha}{N}\big(y_k - \hat{y}_{\theta_k}(x_k)\big)\nabla_{\theta^i}\sigma_*(\theta_k^i, x_k).\end{aligned}$$

Each data point is used only *once* and never revisisted.

Empirical parameter measure: We track the evolution of the parameters through

$$\mu_k^N = \frac{1}{N}\sum_{i=1}^{N}\delta_{\theta_k^i} \in \mathcal{P}(\mathbb{R}^D).$$

(Online) Stochastic Gradient Descent

Initialization: $\theta_0^i \overset{i.i.d.}{\sim} \mu_0 \in \mathcal{P}(\mathbb{R}^D)$.

Update: SGD with fixed *learning rate* $\alpha > 0$,

$$\begin{aligned}\theta_{k+1} &= \theta_k - \alpha \nabla \widehat{\mathcal{R}}(\theta_k, x_k, y_k) \qquad \text{(square loss)}\\ &= \theta_k - \frac{\alpha}{N}\big(y_k - \hat{y}_{\theta_k}(x_k)\big)\nabla_{\theta^i}\sigma_*(\theta_k^i, x_k).\end{aligned}$$

Each data point is used only *once* and never revisited.

Empirical parameter measure: We track the evolution of the parameters through

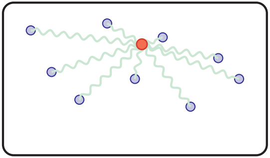
$$\mu_t^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta^i_{\lfloor Nt \rfloor}} \in \mathcal{P}(\mathbb{R}^D).$$

Mean-field limit

Mean-field approximation

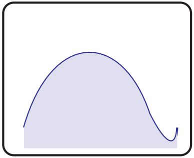
Microscopic (finite width)

N neurons $\equiv$ N interacting particles



Macroscopic (mean-field)

deterministic limit $\bar{\mu}_t$



$N \to \infty$

SGD: $\theta^i_{k+1} = \theta^i_k - \frac{\alpha}{N}\big(y_k - \hat{y}_{\theta_k}(x_k)\big)\nabla_{\theta^i}\sigma_*(\theta^i_k, x_k)$

Mean-field predictor:

$\bar{y}_t(x) = \langle \sigma_*(\cdot, x), \bar{\mu}_t \rangle$

Empirical measure: $\mu^N_t = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \delta_{\theta^i_{\lfloor Nt \rfloor}}$

A law of large numbers

Theorem (Descours et al.'22)

Under standard regularity assumptions on σ_*, μ_0 and π,

$$(\mu_t^N)_{t\geq 0} \xrightarrow[N\to\infty]{\mathbb{P}} (\bar{\mu}_t)_{t\geq 0} \in \mathcal{D}(\mathbb{R}+, \mathcal{P}_\gamma(\mathbb{R}^D)) \quad (\gamma > \tfrac{D}{2})$$

where $\bar{\mu}$ is characterized as the unique (deterministic) solution in $\mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^D))$ to the evolution equation:

$$\partial_t \bar{\mu}_t + \alpha \nabla \cdot \big(G(\cdot, \bar{\mu}_t)\, \bar{\mu}_t\big) = 0, \qquad \bar{\mu}_0 = \mu_0.$$

A law of large numbers

Theorem (Descours et al.'22)

Under standard regularity assumptions on σ_*, μ_0 and π,

$$(\mu_t^N)_{t\geq 0} \xrightarrow[N\to\infty]{\mathbb{P}} (\bar{\mu}_t)_{t\geq 0} \in \mathcal{D}(\mathbb{R}+, \mathcal{P}_\gamma(\mathbb{R}^D)) \quad (\gamma > \tfrac{D}{2})$$

where $\bar{\mu}$ is characterized as the unique (deterministic) solution in $\mathcal{C}(\mathbb{R}_+, \mathcal{P}_1(\mathbb{R}^D))$ to the evolution equation:

$$\partial_t \bar{\mu}_t + \alpha \nabla \cdot \big(G(\cdot, \bar{\mu}_t)\, \bar{\mu}_t\big) = 0, \qquad \bar{\mu}_0 = \mu_0.$$

Question

Can we quantify the deviations of μ_t^N from its mean-field limit $\bar{\mu}_t$?

Fluctuation process

A central limit theorem

Fluctuation process: $t \geq 0 \mapsto \eta_t^N = \sqrt{N}(\mu_t^N - \bar{\mu}_t)$
(signed measure) $\in \mathcal{D}(\mathbb{R}_+, H^{-J_0+1,j_0}(\mathbb{R}^D))$.

A central limit theorem

Fluctuation process: $t \geq 0 \mapsto \eta_t^N = \sqrt{N}(\mu_t^N - \bar{\mu}_t)$
(signed measure) $\in \mathcal{D}(\mathbb{R}_+, H^{-J_0+1,j_0}(\mathbb{R}^D))$.

Theorem (Descours et al.'22)

$$(\eta_t^N)_{t\geq 0} \xrightarrow[N\to\infty]{d} (\eta_t^*)_{t\geq 0} \in \mathcal{C}(\mathbb{R}+, H^{-J_0+1,j_0}(\mathbb{R}^D)).$$

The law of η^* is characterized as the unique (weak) solution to

$$\begin{aligned}
&\text{a.s.},\ \forall \varphi \in H^{J_0,j_0}(\mathbb{R}^D),\ \forall t \geq 0, \quad \langle \varphi, \eta_t^* \rangle = \langle \varphi, \eta_0^* \rangle + \langle \varphi, \mathcal{G}_t \rangle \\
&\quad + \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot, x), \bar{\mu}_s \rangle)\langle \nabla\varphi \cdot \nabla\sigma_*(\cdot, x), \eta_s^* \rangle \, \pi(dx, dy)\, ds \\
&\quad - \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha\langle \sigma_*(\cdot, x), \eta_s^* \rangle \langle \nabla\varphi \cdot \nabla\sigma_*(\cdot, x), \bar{\mu}_s \rangle \, \pi(dx, dy)\, ds,
\end{aligned}$$

with initial condition $\nu_0 \in H^{-J_0+1,j_0}(\mathbb{R}^D)$ s.t. for any $\varphi_1, \ldots, \varphi_k \in H^{J_0-1,j_0}(\mathbb{R}^D)$, $(\langle \varphi_1, \nu_0 \rangle, \ldots, \langle \varphi_k, \nu_0 \rangle)^\top \sim \mathcal{N}(0, \Gamma(\varphi_1, \ldots, \varphi_k))$, where $\Gamma(\varphi_1, \ldots, \varphi_k)$ is the covariance matrix of $(\varphi_1(\theta_1^0), \ldots, \varphi_k(\theta_1^0))^\top$.

A central limit theorem

Fluctuation process: $t \geq 0 \mapsto \eta_t^N = \sqrt{N}(\mu_t^N - \bar{\mu}_t)$
(signed measure) $\in \mathcal{D}(\mathbb{R}_+, H^{-J_0+1,j_0}(\mathbb{R}^D))$.

Theorem (Descours et al.'22)

$$(\eta_t^N)_{t\geq 0} \xrightarrow[N\to\infty]{d} (\eta_t^*)_{t\geq 0} \in \mathcal{C}(\mathbb{R}+, H^{-J_0+1,j_0}(\mathbb{R}^D)).$$

The law of η^* is characterized as the unique (weak) solution to

$$\begin{aligned}
&\text{a.s.}, \ \forall \varphi \in H^{J_0,j_0}(\mathbb{R}^D), \ \forall t \geq 0, \quad \langle \varphi, \eta_t^* \rangle = \langle \varphi, \eta_0^* \rangle + \langle \varphi, \mathcal{G}_t \rangle \\
&\quad + \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha(y - \langle \sigma_*(\cdot,x), \bar{\mu}_s \rangle) \langle \nabla\varphi \cdot \nabla\sigma_*(\cdot,x), \eta_s^* \rangle \, \pi(dx,dy)\, ds \\
&\quad - \int_0^t \int_{\mathcal{X}\times\mathcal{Y}} \alpha \langle \sigma_*(\cdot,x), \eta_s^* \rangle \langle \nabla\varphi \cdot \nabla\sigma_*(\cdot,x), \bar{\mu}_s \rangle \, \pi(dx,dy)\, ds,
\end{aligned}$$

Don't look!

with initial condition $\nu_0 \in H^{-J_0+1,j_0}(\mathbb{R}^D)$ *s.t. for any* $\varphi_1, \ldots, \varphi_k \in H^{J_0-1,j_0}(\mathbb{R}^D)$, $(\langle \varphi_1, \nu_0 \rangle, \ldots, \langle \varphi_k, \nu_0 \rangle)^\top \sim \mathcal{N}(0, \Gamma(\varphi_1, \ldots, \varphi_k))$, *where* $\Gamma(\varphi_1, \ldots, \varphi_k)$ *is the covariance matrix of* $(\varphi_1(\theta_1^0), \ldots, \varphi_k(\theta_1^0))^\top$.

Law of the asymptotic fluctuation process

The law of $\langle \varphi, \eta_t^* \rangle$ is *Gaussian* !

Law of the asymptotic fluctuation process

The law of $\langle \varphi, \eta_t^* \rangle$ is *Gaussian* !

Theorem (DGLMNS'26$^+$)

For any test function $\varphi \in \mathcal{C}_b^\infty(\mathbb{R}^D)$, $(\langle \varphi, \eta_t^* \rangle)_{t \geq 0}$ *is a centered Gaussian process, with variance*

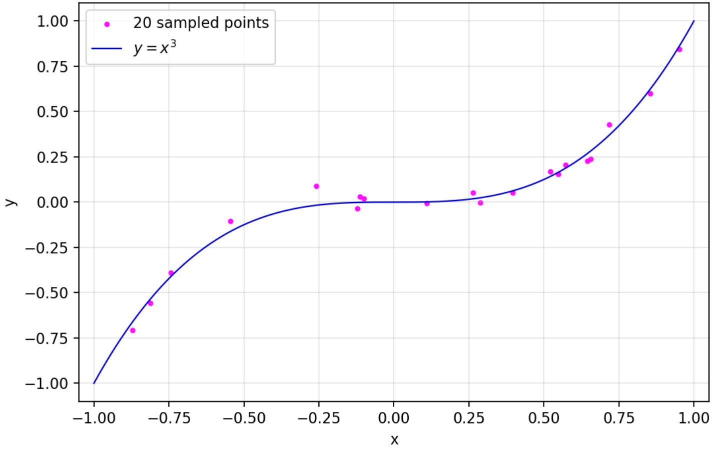
$$V_t = V_0 + \alpha^2 \int_0^t \mathrm{Var}_\pi(Q_s[f(s, \cdot)])\, ds.$$

where $f \in \mathcal{C}^1([0,t] \times \mathbb{R}^2)$ *is the solution of the following backward PDE, with final datum* $f(t, \cdot) = \varphi$*:*

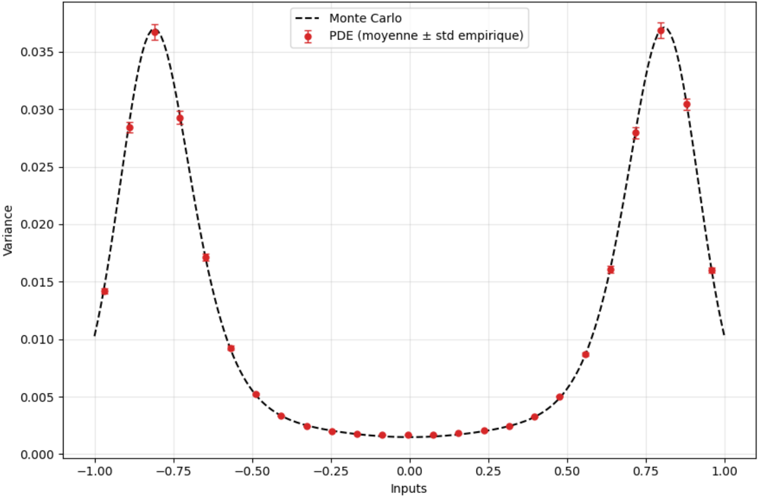
$$\partial_s f + F \cdot \nabla_\theta f + A f = 0.$$

Numerical simulations on a toy dataset

Toy Dataset: $y = x^3 + \varepsilon, \quad x \sim \mathcal{U}([-1,1]) \perp\!\!\!\perp \varepsilon \sim \mathcal{N}(0, 0.05^2).$



Numerical simulations on toy dataset



Concentration bounds

Finite-width approximation bounds

Theorem (MMN'18)

For any bounded-Lipschitz $f : \mathbb{R}^D \to \mathbb{R}$*, for all* $T > 0$ *and* $z > 0$*, with probability* $1 - e^{-z^2}$*,*

$$\sup_{t \in [0,T]} \left| \langle f, \mu_t^N \rangle - \langle f, \bar{\mu}_t \rangle \right| \leq C e^{CT} \delta_{N,d}(z)$$

Finite-width approximation bounds

Theorem (MMN'18)

For any bounded-Lipschitz $f : \mathbb{R}^D \to \mathbb{R}$, for all $T > 0$ and $z > 0$, with probability $1 - e^{-z^2}$,

$$\sup_{t \in [0,T]} \left| \langle f, \mu_t^N \rangle - \langle f, \bar{\mu}_t \rangle \right| \le C e^{CT} \delta_{N,d}(z)$$

Finite-width approximation bounds

Theorem (MMN'18)

For any bounded-Lipschitz $f : \mathbb{R}^D \to \mathbb{R}$*, for all* $T > 0$ *and* $z > 0$*, with probability* $1 - e^{-z^2}$*,*

$$\sup_{t \in [0,T]} \left| \langle f, \mu_t^N \rangle - \langle f, \bar{\mu}_t \rangle \right| \leq C e^{CT} \delta_{N,d}(z)$$

For a *ridge*-regularized version of SGD:

Theorem (Guillin, Nectoux, S.'26)

For any $t \geq 0$ *and* $\delta \in (0,1)$*, with probability* $1 - \delta$*,*

$$\Delta(\mu_t^N, \bar{\mu}_t) \leq C\,\kappa_N + C\sqrt{\frac{\log(2/\delta)}{N}}.$$

with $\kappa_N \to 0$ *as* $N \to \infty$ *and* C *independent of* N *and* t*.*

Finite-width approximation bounds

Theorem (MMN'18)

For any bounded-Lipschitz $f : \mathbb{R}^D \to \mathbb{R}$, for all $T > 0$ and $z > 0$, with probability $1 - e^{-z^2}$,

$$\sup_{t\in[0,T]} \left|\langle f, \mu_t^N\rangle - \langle f, \bar{\mu}_t\rangle\right| \leq C e^{CT} \delta_{N,d}(z)$$

For a *ridge*-regularized version of SGD:

Theorem (Guillin, Nectoux, S.'26)

For any $t \geq 0$ and $\delta \in (0,1)$, with probability $1 - \delta$,

$$\Delta(\mu_t^N, \bar{\mu}_t) \leq C\,\kappa_N + C\sqrt{\frac{\log(2/\delta)}{N}}.$$

with $\kappa_N \to 0$ as $N \to \infty$ and C independent of N and t.

Thank you for your attention!