

# Convergence of Time-Stepping Deep Gradient Flow Methods

Bachelier World Congress

Jasper Rou

joint work with Chenguang Liu & Antonis Papapantoleon

July 9, 2024

# Option pricing

$$\frac{\partial u}{\partial t} - \sum_{i,j=0}^n a^{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} - \sum_{i=0}^n b^i \frac{\partial u}{\partial x_i} + ru = 0$$

$$u(0, x) = \Phi(x)$$

## Option pricing

$$\frac{\partial u}{\partial t} - \sum_{i,j=0}^n a^{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} - \sum_{i=0}^n b^i \frac{\partial u}{\partial x_i} + ru = 0$$
$$u(0, x) = \Phi(x)$$

Can we solve this PDE using a neural network?

# Deep Galerkin Method <sup>1</sup>

$$\frac{\partial u}{\partial t} - \sum_{i,j=0}^n a^{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} - \sum_{i=0}^n b^i \frac{\partial u}{\partial x_i} + ru = 0$$

$$u(0, x) = \Phi(x)$$

---

<sup>1</sup>Sirignano and Spiliopoulos 2018; Jiang, Sirignano, and Cohen 2023

# Deep Galerkin Method <sup>1</sup>

$$\frac{\partial u}{\partial t} - \sum_{i,j=0}^n a^{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} - \sum_{i=0}^n b^i \frac{\partial u}{\partial x_i} + ru = 0$$

$$u(0, x) = \Phi(x)$$

Minimize

$$\left\| \frac{\partial u}{\partial t} - \sum_{i,j=0}^n a^{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} - \sum_{i=0}^n b^i \frac{\partial u}{\partial x_i} + ru \right\|_{[0, T] \times \Omega}^2 + \|u(0, x) - \Phi(x)\|_{\Omega}^2$$

---

<sup>1</sup>Sirignano and Spiliopoulos 2018; Jiang, Sirignano, and Cohen 2023

# Deep Galerkin Method <sup>1</sup>

$$\frac{\partial u}{\partial t} - \sum_{i,j=0}^n a^{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} - \sum_{i=0}^n b^i \frac{\partial u}{\partial x_i} + ru = 0$$

$$u(0, x) = \Phi(x)$$

Minimize

$$\left\| \frac{\partial u}{\partial t} - \sum_{i,j=0}^n a^{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} - \sum_{i=0}^n b^i \frac{\partial u}{\partial x_i} + ru \right\|_{[0, T] \times \Omega}^2 + \|u(0, x) - \Phi(x)\|_{\Omega}^2$$

Issue: Taking second derivative makes training in high dimensions slow

---

<sup>1</sup>Sirignano and Spiliopoulos 2018; Jiang, Sirignano, and Cohen 2023

Rewrite PDE as energy minimization problem

Rewrite PDE as energy minimization problem

- Only first order derivative
- No norm



# Idea

Rewrite PDE as energy minimization problem

- Only first order derivative
- No norm

Split in symmetric and non-symmetric part

# Splitting method

$$\frac{\partial u}{\partial t} = \sum_{i,j=0}^n a^{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=0}^n b^i \frac{\partial u}{\partial x_i} - ru$$

# Splitting method

$$\begin{aligned}\frac{\partial u}{\partial t} &= \sum_{i,j=0}^n a^{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=0}^n b^i \frac{\partial u}{\partial x_i} - ru \\ &= \sum_{i,j=0}^n \frac{\partial}{\partial x_j} \left( a^{ij} \frac{\partial u}{\partial x_i} \right) - \sum_{i,j=0}^n \frac{\partial a^{ij}}{\partial x_j} \frac{\partial u}{\partial x_i} + \sum_{i=0}^n b^i \frac{\partial u}{\partial x_i} - ru\end{aligned}$$

# Splitting method

$$\begin{aligned}\frac{\partial u}{\partial t} &= \sum_{i,j=0}^n a^{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=0}^n b^i \frac{\partial u}{\partial x_i} - ru \\ &= \sum_{i,j=0}^n \frac{\partial}{\partial x_j} \left( a^{ij} \frac{\partial u}{\partial x_i} \right) - \sum_{i,j=0}^n \frac{\partial a^{ij}}{\partial x_j} \frac{\partial u}{\partial x_i} + \sum_{i=0}^n b^i \frac{\partial u}{\partial x_i} - ru \\ &= \sum_{i,j=0}^n \frac{\partial}{\partial x_j} \left( a^{ij} \frac{\partial u}{\partial x_i} \right) - \sum_{i=0}^n \left( \sum_{j=0}^n \frac{\partial a^{ij}}{\partial x_j} - b^i \right) \frac{\partial u}{\partial x_i} - ru\end{aligned}$$

# Splitting method

$$\begin{aligned}\frac{\partial u}{\partial t} &= \sum_{i,j=0}^n a^{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=0}^n b^i \frac{\partial u}{\partial x_i} - ru \\ &= \sum_{i,j=0}^n \frac{\partial}{\partial x_j} \left( a^{ij} \frac{\partial u}{\partial x_i} \right) - \sum_{i,j=0}^n \frac{\partial a^{ij}}{\partial x_j} \frac{\partial u}{\partial x_i} + \sum_{i=0}^n b^i \frac{\partial u}{\partial x_i} - ru \\ &= \sum_{i,j=0}^n \frac{\partial}{\partial x_j} \left( a^{ij} \frac{\partial u}{\partial x_i} \right) - \sum_{i=0}^n \left( \sum_{j=0}^n \frac{\partial a^{ij}}{\partial x_j} - b^i \right) \frac{\partial u}{\partial x_i} - ru \\ &= \nabla \cdot (A \nabla u) - ru - F(u)\end{aligned}$$

$$F(u) = \mathbf{b} \cdot \nabla u$$

## Example: Heston model

$$\begin{aligned}dS_t &= rS_t dt + \sqrt{V_t} S_t dW_t & S_0 > 0 \\dV_t &= \kappa(\theta - V_t) dt + \eta \sqrt{V_t} dB_t & V_0 > 0\end{aligned}$$

## Example: Heston model

$$dS_t = rS_t dt + \sqrt{V_t} S_t dW_t \quad S_0 > 0$$

$$dV_t = \kappa(\theta - V_t) dt + \eta \sqrt{V_t} dB_t \quad V_0 > 0$$

$$\frac{\partial u}{\partial t} = rS \frac{\partial u}{\partial S} + \kappa(\theta - V) \frac{\partial u}{\partial V} + \frac{1}{2} S^2 V \frac{\partial^2 u}{\partial S^2} + \frac{1}{2} \eta^2 V \frac{\partial^2 u}{\partial V^2} + \rho \eta S V \frac{\partial^2 u}{\partial S \partial V} - ru$$

## Example: Heston model

$$\frac{\partial u}{\partial t} = rS \frac{\partial u}{\partial S} + \kappa(\theta - V) \frac{\partial u}{\partial V} + \frac{1}{2} S^2 V \frac{\partial^2 u}{\partial S^2} + \frac{1}{2} \eta^2 V \frac{\partial^2 u}{\partial V^2} + \rho \eta S V \frac{\partial^2 u}{\partial S \partial V} - ru$$



## Example: Heston model

$$\begin{aligned}\frac{\partial u}{\partial t} &= rS \frac{\partial u}{\partial S} + \kappa(\theta - V) \frac{\partial u}{\partial V} + \frac{1}{2} S^2 V \frac{\partial^2 u}{\partial S^2} + \frac{1}{2} \eta^2 V \frac{\partial^2 u}{\partial V^2} + \rho \eta S V \frac{\partial^2 u}{\partial S \partial V} - ru \\ &= rS \frac{\partial u}{\partial S} + \kappa(\theta - V) \frac{\partial u}{\partial V} + \frac{\partial}{\partial S} \left( \frac{1}{2} S^2 V \frac{\partial u}{\partial S} \right) - S V \frac{\partial u}{\partial S} \\ &\quad + \frac{\partial}{\partial V} \left( \frac{1}{2} \eta^2 V \frac{\partial u}{\partial V} \right) - \frac{1}{2} \eta^2 \frac{\partial u}{\partial V} + \frac{\partial}{\partial S} \left( \frac{1}{2} \rho \eta S V \frac{\partial u}{\partial V} \right) - \frac{1}{2} \rho \eta V \frac{\partial u}{\partial V} \\ &\quad + \frac{\partial}{\partial V} \left( \frac{1}{2} \rho \eta S V \frac{\partial u}{\partial S} \right) - \frac{1}{2} \rho \eta S \frac{\partial u}{\partial S} - ru\end{aligned}$$

## Example: Heston model

$$\begin{aligned}\frac{\partial u}{\partial t} &= rS \frac{\partial u}{\partial S} + \kappa(\theta - V) \frac{\partial u}{\partial V} + \frac{1}{2} S^2 V \frac{\partial^2 u}{\partial S^2} + \frac{1}{2} \eta^2 V \frac{\partial^2 u}{\partial V^2} + \rho \eta S V \frac{\partial^2 u}{\partial S \partial V} - ru \\ &= rS \frac{\partial u}{\partial S} + \kappa(\theta - V) \frac{\partial u}{\partial V} + \frac{\partial}{\partial S} \left( \frac{1}{2} S^2 V \frac{\partial u}{\partial S} \right) - SV \frac{\partial u}{\partial S} \\ &\quad + \frac{\partial}{\partial V} \left( \frac{1}{2} \eta^2 V \frac{\partial u}{\partial V} \right) - \frac{1}{2} \eta^2 \frac{\partial u}{\partial V} + \frac{\partial}{\partial S} \left( \frac{1}{2} \rho \eta S V \frac{\partial u}{\partial V} \right) - \frac{1}{2} \rho \eta V \frac{\partial u}{\partial V} \\ &\quad + \frac{\partial}{\partial V} \left( \frac{1}{2} \rho \eta S V \frac{\partial u}{\partial S} \right) - \frac{1}{2} \rho \eta S \frac{\partial u}{\partial S} - ru \\ &= \nabla \cdot \left( \frac{1}{2} \begin{bmatrix} S^2 V & \rho \eta S V \\ \rho \eta S V & \eta^2 V \end{bmatrix} \nabla u \right) - ru \\ &\quad - \begin{bmatrix} SV - rS + \frac{1}{2} \rho \eta S \\ \kappa(V - \theta) + \frac{1}{2} \rho \eta V + \frac{1}{2} \eta^2 \end{bmatrix} \cdot \nabla u\end{aligned}$$

# Time Deep Gradient Flow

$$\begin{cases} u_t - \nabla \cdot (A \nabla u) + ru + F(u) = 0 & (t, \mathbf{x}) \in [0, T] \times \Omega \\ u(0, \mathbf{x}) = \Phi(\mathbf{x}) & \mathbf{x} \in \Omega \end{cases}$$

# Time Deep Gradient Flow

$$\begin{cases} u_t - \nabla \cdot (A \nabla u) + ru + F(u) = 0 & (t, \mathbf{x}) \in [0, T] \times \Omega \\ u(0, \mathbf{x}) = \Phi(\mathbf{x}) & \mathbf{x} \in \Omega \end{cases}$$

- Divide  $[0, T]$  in intervals  $(t_{k-1}, t_k]$  with  $h = t_k - t_{k-1}$

$$\begin{aligned} \frac{U^k - U^{k-1}}{h} - \nabla \cdot (A \nabla U^k) + rU^k + F(U^{k-1}) &= 0 \\ U^0 &= \Phi \end{aligned}$$

# Time Deep Gradient Flow

$$\begin{cases} u_t - \nabla \cdot (A \nabla u) + ru + F(u) = 0 & (t, \mathbf{x}) \in [0, T] \times \Omega \\ u(0, \mathbf{x}) = \Phi(\mathbf{x}) & \mathbf{x} \in \Omega \end{cases}$$

- Divide  $[0, T]$  in intervals  $(t_{k-1}, t_k]$  with  $h = t_k - t_{k-1}$

$$\begin{aligned} \frac{U^k - U^{k-1}}{h} - \nabla \cdot (A \nabla U^k) + rU^k + F(U^{k-1}) &= 0 \\ U^0 &= \Phi \end{aligned}$$

Theorem (Akrivis and Crouzeix 2004)

*There exists a constant  $C$  independent of  $h$  and  $k$  such that*

$$\max_{0 \leq k \leq N} \|u(t_k) - U^k\| \leq Ch$$

# Time Deep Gradient Flow

$$\frac{U^k - U^{k-1}}{h} - \nabla \cdot (A \nabla U^k) + rU^k + F(U^{k-1}) = 0$$

# Time Deep Gradient Flow

$$\frac{U^k - U^{k-1}}{h} - \nabla \cdot (A \nabla U^k) + rU^k + F(U^{k-1}) = 0$$

$$I^k(u) = \frac{1}{2} \|u - U^{k-1}\|^2 + h \int_{\Omega} \frac{1}{2} \left( (\nabla u)^T A \nabla u + ru^2 \right) + F(U^{k-1}) u dx$$

# Time Deep Gradient Flow

$$\frac{U^k - U^{k-1}}{h} - \nabla \cdot (A \nabla U^k) + rU^k + F(U^{k-1}) = 0$$

$$I^k(u) = \frac{1}{2} \|u - U^{k-1}\|^2 + h \int_{\Omega} \frac{1}{2} \left( (\nabla u)^T A \nabla u + ru^2 \right) + F(U^{k-1}) \, u \, dx$$

## Theorem

*The minimizer  $w_* \in \mathcal{H}_0^1(\mathbb{R}^d)$  of  $I^k$  is the unique solution  $U^k$ .*



# Time Deep Gradient Flow

## Theorem

*The minimizer  $w_* \in \mathcal{H}_0^1(\mathbb{R}^d)$  of  $I^k$  is the unique solution  $U^k$ .*

## Proof.

# Time Deep Gradient Flow

## Theorem

*The minimizer  $w_* \in \mathcal{H}_0^1(\mathbb{R}^d)$  of  $I^k$  is the unique solution  $U^k$ .*

## Proof.

$$i^k(\tau) = I^k(w_* + \tau v)$$

# Time Deep Gradient Flow

## Theorem

*The minimizer  $w_* \in \mathcal{H}_0^1(\mathbb{R}^d)$  of  $I^k$  is the unique solution  $U^k$ .*

## Proof.

$$i^k(\tau) = I^k(w_* + \tau v)$$

Since  $w_*$  minimizes  $I^k$ ,  $\tau = 0$  minimizes  $i^k$ .

# Time Deep Gradient Flow

## Theorem

The minimizer  $w_* \in \mathcal{H}_0^1(\mathbb{R}^d)$  of  $I^k$  is the unique solution  $U^k$ .

## Proof.

$$i^k(\tau) = I^k(w_* + \tau v)$$

Since  $w_*$  minimizes  $I^k$ ,  $\tau = 0$  minimizes  $i^k$ .

$$\begin{aligned} 0 &= (i^k)'(0) \\ &= \int_{\mathbb{R}^d} \left( (w_* - U^{k-1}) + h(-\nabla \cdot (A \nabla w_*) + r w_* + F(U^{k-1})) \right) v dx. \end{aligned}$$



# Time **Deep** Gradient Flow

## Definition (Activation function)

An activation function is a function  $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$  such that  $\psi \in C_c^\infty(\mathbb{R}^d)$  and  $\int_{\mathbb{R}^d} \psi(x) dx \neq 0$ .

# Time **Deep** Gradient Flow

## Definition (Activation function)

An activation function is a function  $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$  such that  $\psi \in C_c^\infty(\mathbb{R}^d)$  and  $\int_{\mathbb{R}^d} \psi(x) dx \neq 0$ .

## Definition (Neural network)

$$\mathcal{C}^n(\psi) = \left\{ \zeta(x) : \mathbb{R}^d \rightarrow \mathbb{R} : \zeta(x) = \sum_{i=1}^n \beta_i \psi(\alpha_i x + c_i) \right\},$$
$$\mathcal{C}(\psi) = \bigcup_{n \geq 1} \mathcal{C}^n(\psi)$$

# Time **Deep** Gradient Flow

## Definition (Activation function)

An activation function is a function  $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$  such that  $\psi \in C_c^\infty(\mathbb{R}^d)$  and  $\int_{\mathbb{R}^d} \psi(x) dx \neq 0$ .

## Definition (Neural network)

$$\mathcal{C}^n(\psi) = \left\{ \zeta(x) : \mathbb{R}^d \rightarrow \mathbb{R} : \zeta(x) = \sum_{i=1}^n \beta_i \psi(\alpha_i x + c_i) \right\},$$
$$\mathcal{C}(\psi) = \bigcup_{n \geq 1} \mathcal{C}^n(\psi)$$

## Theorem

$\mathcal{C}(\psi)$  is dense in  $\mathcal{H}_0^1(\mathbb{R}^d)$ .

# Convergence of the minimizer

## Theorem

Let  $w_m$  be a sequence in  $\mathcal{H}_0^1(\mathbb{R}^d)$  and  $w_*$  the minimizer of  $I^k$ .

$$\lim_{m \rightarrow \infty} \|w_m - w_*\|_{\mathcal{H}_0^1} = 0 \quad \Longleftrightarrow \quad \lim_{m \rightarrow \infty} I^k(w_m) = I^k(w_*)$$

$$I^k(u) = \frac{1}{2} \|u - U^{k-1}\|^2 + h \int_{\mathbb{R}^d} \frac{1}{2} \left( (\nabla u)^T A \nabla u + ru^2 \right) + F(U^{k-1}) u dx$$



# Convergence of the minimizer

## Theorem

Let  $w_m$  be a sequence in  $\mathcal{H}_0^1(\mathbb{R}^d)$  and  $w_*$  the minimizer of  $I^k$ .

$$\lim_{m \rightarrow \infty} \|w_m - w_*\|_{\mathcal{H}_0^1} = 0 \quad \iff \quad \lim_{m \rightarrow \infty} I^k(w_m) = I^k(w_*)$$

$$\begin{aligned} I^k(u) &= \frac{1}{2} \|u - U^{k-1}\|^2 + h \int_{\mathbb{R}^d} \frac{1}{2} \left( (\nabla u)^T A \nabla u + ru^2 \right) + F(U^{k-1}) u dx \\ &=: \mathcal{L}^k(u) + \mathcal{G}^k(u), \end{aligned}$$

$$\mathcal{L}^k(u) = \frac{1}{2} \|u\|^2 + \frac{h}{2} \int_{\mathbb{R}^d} (\nabla u)^T A \nabla u + ru^2 dx,$$

$$\mathcal{G}^k(u) = -\langle u, U^{k-1} \rangle + \frac{1}{2} \|U^{k-1}\|^2 + h \int_{\mathbb{R}^d} F(U^{k-1}) u dx$$

# Convergence of the minimizer

## Theorem

Let  $w_m$  be a sequence in  $\mathcal{H}_0^1(\mathbb{R}^d)$  and  $w_*$  the minimizer of  $I^k$ .

$$\lim_{m \rightarrow \infty} \|w_m - w_*\|_{\mathcal{H}_0^1} = 0 \quad \iff \quad \lim_{m \rightarrow \infty} I^k(w_m) = I^k(w_*)$$

# Convergence of the minimizer

## Theorem

Let  $w_m$  be a sequence in  $\mathcal{H}_0^1(\mathbb{R}^d)$  and  $w_*$  the minimizer of  $I^k$ .

$$\lim_{m \rightarrow \infty} \|w_m - w_*\|_{\mathcal{H}_0^1} = 0 \quad \Longleftrightarrow \quad \lim_{m \rightarrow \infty} I^k(w_m) = I^k(w_*)$$

## Proof.

$\implies$   $I^k$  is continuous.

# Convergence of the minimizer

## Theorem

Let  $w_m$  be a sequence in  $\mathcal{H}_0^1(\mathbb{R}^d)$  and  $w_*$  the minimizer of  $I^k$ .

$$\lim_{m \rightarrow \infty} \|w_m - w_*\|_{\mathcal{H}_0^1} = 0 \quad \iff \quad \lim_{m \rightarrow \infty} I^k(w_m) = I^k(w_*)$$

## Proof.

$\implies$   $I^k$  is continuous.

$\impliedby$   $w_m \rightharpoonup w_*$ .

# Convergence of the minimizer

## Theorem

Let  $w_m$  be a sequence in  $\mathcal{H}_0^1(\mathbb{R}^d)$  and  $w_*$  the minimizer of  $I^k$ .

$$\lim_{m \rightarrow \infty} \|w_m - w_*\|_{\mathcal{H}_0^1} = 0 \quad \iff \quad \lim_{m \rightarrow \infty} I^k(w_m) = I^k(w_*)$$

## Proof.

$\implies$   $I^k$  is continuous.

$\impliedby$   $w_m \rightharpoonup w_*$ . So  $\mathcal{G}^n[w_m] \rightarrow \mathcal{G}^n[w_*]$ .

# Convergence of the minimizer

## Theorem

Let  $w_m$  be a sequence in  $\mathcal{H}_0^1(\mathbb{R}^d)$  and  $w_*$  the minimizer of  $I^k$ .

$$\lim_{m \rightarrow \infty} \|w_m - w_*\|_{\mathcal{H}_0^1} = 0 \quad \iff \quad \lim_{m \rightarrow \infty} I^k(w_m) = I^k(w_*)$$

## Proof.

$\implies$   $I^k$  is continuous.

$\impliedby$   $w_m \rightharpoonup w_*$ . So  $\mathcal{G}^n[w_m] \rightarrow \mathcal{G}^n[w_*]$ .

Since  $I^k(w_m) \rightarrow I^k(w_*)$ ,  $\mathcal{L}^k(w_m) \rightarrow \mathcal{L}^k(w_*)$ .

# Convergence of the minimizer

## Theorem

Let  $w_m$  be a sequence in  $\mathcal{H}_0^1(\mathbb{R}^d)$  and  $w_*$  the minimizer of  $I^k$ .

$$\lim_{m \rightarrow \infty} \|w_m - w_*\|_{\mathcal{H}_0^1} = 0 \quad \iff \quad \lim_{m \rightarrow \infty} I^k(w_m) = I^k(w_*)$$

## Proof.

$\implies$   $I^k$  is continuous.

$\impliedby$   $w_m \rightharpoonup w_*$ . So  $\mathcal{G}^n[w_m] \rightarrow \mathcal{G}^n[w_*]$ .

Since  $I^k(w_m) \rightarrow I^k(w_*)$ ,  $\mathcal{L}^k(w_m) \rightarrow \mathcal{L}^k(w_*)$ .

$$\frac{1+hr}{2} \|w_m - w_*\|^2 + \frac{h}{2} \left\| \sqrt{A} \nabla (w_m - w_*) \right\|^2 \rightarrow 0.$$

# Convergence of the minimizer

## Theorem

Let  $w_m$  be a sequence in  $\mathcal{H}_0^1(\mathbb{R}^d)$  and  $w_*$  the minimizer of  $I^k$ .

$$\lim_{m \rightarrow \infty} \|w_m - w_*\|_{\mathcal{H}_0^1} = 0 \quad \iff \quad \lim_{m \rightarrow \infty} I^k(w_m) = I^k(w_*)$$

## Proof.

$\implies$   $I^k$  is continuous.

$\impliedby$   $w_m \rightharpoonup w_*$ . So  $\mathcal{G}^n[w_m] \rightarrow \mathcal{G}^n[w_*]$ .

Since  $I^k(w_m) \rightarrow I^k(w_*)$ ,  $\mathcal{L}^k(w_m) \rightarrow \mathcal{L}^k(w_*)$ .

$$\frac{1+hr}{2} \|w_m - w_*\|^2 + \frac{h}{2} \left\| \sqrt{A} \nabla (w_m - w_*) \right\|^2 \rightarrow 0.$$

$$\|w_m - w_*\|_{\mathcal{H}_0^1} \rightarrow 0. \quad \square$$



# Convergence when training

Neural network:

$$V_t^N(\theta^N; x) = V^N(\theta_t^N; x) = N^{-\delta} \sum_{i=1}^N \beta^i \psi(\alpha^i x + c^i),$$

$$\theta^N = (\beta^i, \alpha^i, c^i)_{i=1}^N, \quad \frac{1}{2} < \delta < 1.$$

# Convergence when training

Neural network:

$$V_t^N(\theta^N; x) = V^N(\theta_t^N; x) = N^{-\delta} \sum_{i=1}^N \beta^i \psi(\alpha^i x + c^i),$$

$$\theta^N = (\beta^i, \alpha^i, c^i)_{i=1}^N, \quad \frac{1}{2} < \delta < 1.$$

$$V_t^N \xrightarrow{N \rightarrow \infty} V_t \xrightarrow{t \rightarrow \infty} w_*$$

# Gradient Descent

$$V^N(\theta^N; \mathbf{x}) = N^{-\delta} \sum_{i=1}^N \beta^i \psi(\alpha^i \mathbf{x} + \mathbf{c}^i),$$

$$\theta^N = (\beta^i, \alpha^i, \mathbf{c}^i)_{i=1}^N, \quad \frac{1}{2} < \delta < 1.$$

# Gradient Descent

$$V^N(\theta^N; \mathbf{x}) = N^{-\delta} \sum_{i=1}^N \beta^i \psi(\alpha^i \mathbf{x} + \mathbf{c}^i),$$

$$\theta^N = (\beta^i, \alpha^i, \mathbf{c}^i)_{i=1}^N, \quad \frac{1}{2} < \delta < 1. \quad \eta_N = N^{2\delta-1}$$

$$\frac{d\theta_t^N}{dt} = -\eta_N \nabla_{\theta} l^k(V^N(\theta_t^N; \mathbf{x}))$$

# Gradient Descent

$$V^N(\theta^N; \mathbf{x}) = N^{-\delta} \sum_{i=1}^N \beta^i \psi(\alpha^i \mathbf{x} + \mathbf{c}^i),$$

$$\theta^N = (\beta^i, \alpha^i, \mathbf{c}^i)_{i=1}^N, \quad \frac{1}{2} < \delta < 1. \quad \eta_N = N^{2\delta-1}$$

$$\frac{d\theta_t^N}{dt} = -\eta_N \nabla_{\theta} l^k(V^N(\theta_t^N; \mathbf{x}))$$

$$\begin{aligned} \frac{dV_t^N(\mathbf{x})}{dt} &= \nabla_{\theta} V^N(\theta_t^N; \mathbf{x}) \cdot \frac{d\theta_t^N}{dt} \\ &= -\eta_N \nabla_{\theta} V^N(\theta_t^N; \mathbf{x}) \cdot \nabla_{\theta} l^k(V^N(\theta_t^N; \mathbf{x})) \end{aligned}$$

## Convergence in neurons

$$\frac{dV_t^N(x)}{dt} = -\eta_N \nabla_{\theta} V^N(\theta_t^N; x) \cdot \nabla_{\theta} J^k(V^N(\theta_t^N; x))$$

## Convergence in neurons

$$\begin{aligned}\frac{dV_t^N(x)}{dt} &= -\eta_N \nabla_{\theta} V^N(\theta_t^N; x) \cdot \nabla_{\theta} I^k(V^N(\theta_t^N; x)) \\ &= -\left\langle \mathcal{D}I^k(V_t^N), Z_t^N(x, \cdot) \right\rangle_{\mathcal{H}_0^1}\end{aligned}$$

$$Z_t^N(x, y) = N^{-1} \sum_{i=1}^N \nabla_{\beta, \alpha, c} \beta_t^i \psi(\alpha_t^i x + c_t^i) \cdot \nabla_{\beta, \alpha, c} \beta_t^i \psi(\alpha_t^i y + c_t^i)$$

## Convergence in neurons

$$\begin{aligned}\frac{dV_t^N(x)}{dt} &= -\eta_N \nabla_{\theta} V^N(\theta_t^N; x) \cdot \nabla_{\theta} I^k(V^N(\theta_t^N; x)) \\ &= -\left\langle \mathcal{D}I^k(V_t^N), Z_t^N(x, \cdot) \right\rangle_{\mathcal{H}_0^1}\end{aligned}$$

$$\frac{dV_t(x)}{dt} = -\left\langle \mathcal{D}I^k(V_t), Z(x, \cdot) \right\rangle_{\mathcal{H}_0^1}$$

$$Z_t^N(x, y) = N^{-1} \sum_{i=1}^N \nabla_{\beta, \alpha, c} \beta_t^i \psi(\alpha_t^i x + c_t^i) \cdot \nabla_{\beta, \alpha, c} \beta_t^i \psi(\alpha_t^i y + c_t^i)$$

$$Z(x, y) = \mathbb{E} \left[ \nabla_{\beta, \alpha, c} \beta_0^1 \psi(\alpha_0^1 x + c_0^1) \cdot \nabla_{\beta, \alpha, c} \beta_0^1 \psi(\alpha_0^1 y + c_0^1) \right]$$



## Wide network limit

$$\frac{dV_t^N(x)}{dt} = - \left\langle \mathcal{D}I^k(V_t^N), Z_t^N(x, \cdot) \right\rangle_{\mathcal{H}_0^1}$$

$$\frac{dV_t(x)}{dt} = - \left\langle \mathcal{D}I^k(V_t), Z(x, \cdot) \right\rangle_{\mathcal{H}_0^1}$$

$$Z_t^N(x, y) = N^{-1} \sum_{i=1}^N \nabla_{\beta, \alpha, c} \beta_t^i \psi(\alpha_t^{i,N} x + c_t^{i,N}) \cdot \nabla_{\beta, \alpha, c} \beta_t^i \psi(\alpha_t^{i,N} y + c_t^{i,N})$$

$$Z(x, y) = \mathbb{E} \left[ \nabla_{\beta, \alpha, c} \beta_0^1 \psi(\alpha_0^1 x + c_0^1) \cdot \nabla_{\beta, \alpha, c} \beta_0^1 \psi(\alpha_0^1 y + c_0^1) \right]$$

## Wide network limit

$$\frac{dV_t^N(x)}{dt} = - \left\langle \mathcal{D}I^k(V_t^N), Z_t^N(x, \cdot) \right\rangle_{\mathcal{H}_0^1}$$

$$\frac{dV_t(x)}{dt} = - \left\langle \mathcal{D}I^k(V_t), Z(x, \cdot) \right\rangle_{\mathcal{H}_0^1}$$

$$Z_t^N(x, y) = N^{-1} \sum_{i=1}^N \nabla_{\beta, \alpha, c} \beta_t^i \psi(\alpha_t^{i,N} x + c_t^{i,N}) \cdot \nabla_{\beta, \alpha, c} \beta_t^i \psi(\alpha_t^{i,N} y + c_t^{i,N})$$

$$Z(x, y) = \mathbb{E} \left[ \nabla_{\beta, \alpha, c} \beta_0^1 \psi(\alpha_0^1 x + c_0^1) \cdot \nabla_{\beta, \alpha, c} \beta_0^1 \psi(\alpha_0^1 y + c_0^1) \right]$$

### Theorem

For any  $T > 0$ ,

$$\sup_{0 \leq t \leq T} \mathbb{E} \left[ \left\| V_t^N - V_t \right\|_{\mathcal{H}_0^1} \right] \xrightarrow{N \rightarrow \infty} 0.$$

# Wide network limit

## Theorem

$$\lim_{t \rightarrow \infty} \|V_t - w_*\|_{\mathcal{H}_0^1} = 0.$$

$$\frac{dV_t(x)}{dt} = - \left\langle \mathcal{D}I^k(V_t), Z(x, \cdot) \right\rangle_{\mathcal{H}_0^1}$$

## Theorem

$$\lim_{t \rightarrow \infty} \|V_t - w_*\|_{\mathcal{H}_0^1} = 0.$$

$$\begin{aligned}\frac{dV_t(x)}{dt} &= - \left\langle \mathcal{D}I^k(V_t), Z(x, \cdot) \right\rangle_{\mathcal{H}_0^1} \\ \frac{d(V_t - w_*)(x)}{dt} &= - \left\langle \mathcal{D}I^k(V_t - w_* + w_*), Z(x, \cdot) \right\rangle_{\mathcal{H}_0^1} \\ &= - \tilde{\mathcal{T}}(V_t - w_*)(x)\end{aligned}$$

## Convergence in time

Proof:  $\lim_{t \rightarrow \infty} \|V_t - w_*\|_{\mathcal{H}_0^1} = 0$ .

$\tilde{\mathcal{T}}$  is a self-adjoint, positive definite trace class operator. Spectral decomposition:

$$\tilde{\mathcal{T}}(\tilde{e}_i) = \lambda_i \tilde{e}_i,$$

$\lambda_1 \geq \lambda_2 \geq \dots > 0$ , orthogonal basis  $\{\tilde{e}_i\}_{i=1}^{\infty}$ .

## Convergence in time

Proof:  $\lim_{t \rightarrow \infty} \|V_t - w_*\|_{\mathcal{H}_0^1} = 0$ .

$\tilde{\mathcal{T}}$  is a self-adjoint, positive definite trace class operator. Spectral decomposition:

$$\tilde{\mathcal{T}}(\tilde{e}_i) = \lambda_i \tilde{e}_i,$$

$\lambda_1 \geq \lambda_2 \geq \dots > 0$ , orthogonal basis  $\{\tilde{e}_i\}_{i=1}^{\infty}$ .

$$\begin{aligned} \frac{dh_t^i}{dt} &:= \frac{\langle d(V_t - w_*), \tilde{e}_i \rangle}{dt} = -\langle \tilde{\mathcal{T}}(V_t - w_*), \tilde{e}_i \rangle = -\langle V_t - w_*, \tilde{\mathcal{T}}(\tilde{e}_i) \rangle \\ &= -\lambda_i h_t^i. \end{aligned}$$

## Convergence in time

Proof:  $\lim_{t \rightarrow \infty} \|V_t - w_*\|_{\mathcal{H}_0^1} = 0$ .

$\tilde{\mathcal{T}}$  is a self-adjoint, positive definite trace class operator. Spectral decomposition:

$$\tilde{\mathcal{T}}(\tilde{e}_i) = \lambda_i \tilde{e}_i,$$

$\lambda_1 \geq \lambda_2 \geq \dots > 0$ , orthogonal basis  $\{\tilde{e}_i\}_{i=1}^{\infty}$ .

$$\begin{aligned} \frac{dh_t^i}{dt} &:= \frac{\langle d(V_t - w_*), \tilde{e}_i \rangle}{dt} = -\langle \tilde{\mathcal{T}}(V_t - w_*), \tilde{e}_i \rangle = -\langle V_t - w_*, \tilde{\mathcal{T}}(\tilde{e}_i) \rangle \\ &= -\lambda_i h_t^i. \end{aligned}$$

$$h_t^i = e^{-\lambda_i t} h_0^i.$$

## Convergence in time

Proof:  $\lim_{t \rightarrow \infty} \|V_t - w_*\|_{\mathcal{H}_0^1} = 0$ .

$\tilde{\mathcal{T}}$  is a self-adjoint, positive definite trace class operator. Spectral decomposition:

$$\tilde{\mathcal{T}}(\tilde{e}_i) = \lambda_i \tilde{e}_i,$$

$\lambda_1 \geq \lambda_2 \geq \dots > 0$ , orthogonal basis  $\{\tilde{e}_i\}_{i=1}^{\infty}$ .

$$\begin{aligned} \frac{dh_t^i}{dt} &:= \frac{\langle d(V_t - w_*), \tilde{e}_i \rangle}{dt} = -\langle \tilde{\mathcal{T}}(V_t - w_*), \tilde{e}_i \rangle = -\langle V_t - w_*, \tilde{\mathcal{T}}(\tilde{e}_i) \rangle \\ &= -\lambda_i h_t^i. \end{aligned}$$

$h_t^i = e^{-\lambda_i t} h_0^i$ . Parseval's identity:

$$\|V_t - w_*\|^2 = \sum_{i=1}^{\infty} (h_t^i)^2 = \sum_{i=1}^{\infty} e^{-2\lambda_i t} (h_0^i)^2 \xrightarrow{t \rightarrow \infty} 0. \quad \square$$



# Convergence of Time-Stepping Deep Gradient Flow Methods

Bachelier World Congress

Jasper Rou

July 9, 2024

[j.g.rou@tudelft.nl](mailto:j.g.rou@tudelft.nl)

[www.jasperrou.nl](http://www.jasperrou.nl)