Convergence of Time-Stepping Deep Gradient Flow Methods

Dutch Math Finance Afternoon

Jasper Rou
joint work with Chenguang Liu & Antonis Papapantoleon

June 7, 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

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

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

Deep Galerkin Method

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

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

Minimize

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

Deep Galerkin Method

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

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

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

Idea

Rewrite PDE as energy minimization problem

Idea

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

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

Splitting method

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

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

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

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

Example: Heston model

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

Example: Heston model

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

Example: Heston model

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

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 $h = t_k - t_{k-1}$

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

$$U^0 = \Phi$$

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 $h = t_k - t_{k-1}$

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

Theorem (Akrivis and Crouzeix 2004)

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

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

Time Deep **Gradient Flow**

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

Time Deep **Gradient Flow**

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

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

Time Deep **Gradient Flow**

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

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

Theorem

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

Time Deep **Gradient Flow**

Theorem

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

Proof.

Time Deep **Gradient Flow**

Theorem

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

Proof.

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

Time Deep **Gradient Flow**

Theorem

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

Proof.

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

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

Time Deep **Gradient Flow**

Theorem

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

Proof.

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

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

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

☐

Time **Deep** Gradient Flow

Definition (Activation function)

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

Time **Deep** Gradient Flow

Definition (Activation function)

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

Definition (Neural network)

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

$$\mathcal{C}(\psi) = \cup_{n \geq 1} \mathcal{C}^n(\psi)$$

Time **Deep** Gradient Flow

Definition (Activation function)

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

Definition (Neural network)

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

$$\mathcal{C}(\psi) = \cup_{n \geq 1} \mathcal{C}^n(\psi)$$

Theorem

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

Convergence of the minimizer

Theorem

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

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

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

Convergence of the minimizer

Theorem

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

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

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

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

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

Convergence of the minimizer

Theorem

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

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

Convergence of the minimizer

Theorem

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

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

Proof.

$\Longrightarrow$ $\quad I^k$ is continuous.

Convergence of the minimizer

Theorem

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

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

Proof.

$\Longrightarrow$ I^k is continuous.

$\Longleftarrow$ $w_m \rightharpoonup w_*$.

Convergence of the minimizer

Theorem

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

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

Proof.

$\Longrightarrow$ I^k is continuous.

$\Longleftarrow$ $w_m \rightharpoonup w_*$. So $\mathcal{G}^n\left[w_m\right] \to \mathcal{G}^n\left[w_*\right]$.

Convergence of the minimizer

Theorem

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

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

Proof.

$\Longrightarrow$ I^k is continuous.

$\Longleftarrow$ $w_m \rightharpoonup w_*$. So $\mathcal{G}^n\left[w_m\right] \to \mathcal{G}^n\left[w_*\right]$.
Since $I^k\left(w_m\right) \to I^k\left(w_*\right)$, $\mathcal{L}^k\left(w_m\right) \to \mathcal{L}^k\left(w_*\right)$.

Convergence of the minimizer

Theorem

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

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

Proof.

$\Longrightarrow$ I^k is continuous.

$\Longleftarrow$ $w_m \rightharpoonup w_*$. So $\mathcal{G}^n\left[w_m\right] \to \mathcal{G}^n\left[w_*\right]$.
Since $I^k\left(w_m\right) \to I^k\left(w_*\right)$, $\mathcal{L}^k\left(w_m\right) \to \mathcal{L}^k\left(w_*\right)$.

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

Convergence of the minimizer

Theorem

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

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

Proof.

$\Longrightarrow$ I^k is continuous.

$\Longleftarrow$ $w_m \rightharpoonup w_*$. So $\mathcal{G}^n\left[w_m\right] \to \mathcal{G}^n\left[w_*\right]$.

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

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

$\|w_m - w_*\|_{\mathcal{H}_0^1} \to 0.$ □

Convergence when training

Neural network:

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

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

Convergence when training

Neural network:

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

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

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

Gradient Descent

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

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

Gradient Descent

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

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

$$\frac{\mathrm{d}\theta_t^N}{\mathrm{d}t} = -\eta_N \nabla_\theta I^k\left(V^N\left(\theta_t^N; x\right)\right)$$

Gradient Descent

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

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

$$\frac{\mathrm{d}\theta_t^N}{\mathrm{d}t} = -\eta_N \nabla_\theta I^k\left(V^N\left(\theta_t^N; x\right)\right)$$

$$\begin{aligned}\frac{\mathrm{d}V_t^N(x)}{\mathrm{d}t} &= \nabla_\theta V^N\left(\theta_t^N; x\right) \cdot \frac{\mathrm{d}\theta_t^N}{\mathrm{d}t} \\ &= -\eta_N \nabla_\theta V^N\left(\theta_t^N; x\right) \cdot \nabla_\theta I^k\left(V^N\left(\theta_t^N; x\right)\right)\end{aligned}$$

Convergence in neurons

$$\frac{\mathrm{d}V_t^N(x)}{\mathrm{d}t} = -\eta_N \nabla_\theta V^N\left(\theta_t^N; x\right) \cdot \nabla_\theta I^k\left(V^N\left(\theta_t^N; x\right)\right)$$

Convergence in neurons

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

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

Convergence in neurons

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

$$\begin{aligned}Z_t^N(x, y) =& N^{-1}\sum_{i=1}^{N} \nabla_{\beta,\alpha,c}\beta_t^i\psi\left(\alpha_t^i x + c_t^i\right) \cdot \nabla_{\beta,\alpha,c}\beta_t^i\psi\left(\alpha_t^i y + c_t^i\right) \\ Z(x, y) =& \mathbb{E}\left[\nabla_{\beta,\alpha,c}\beta_0^1\psi\left(\alpha_0^1 x + c_0^1\right) \cdot \nabla_{\beta,\alpha,c}\beta_0^1\psi\left(\alpha_0^1 y + c_0^1\right)\right]\end{aligned}$$

Wide network limit

$$\frac{\mathrm{d}V_t^N(x)}{\mathrm{d}t} = -\left\langle \mathcal{D}I^k\left(V_t^N\right), Z_t^N(x,.)\right\rangle_{\mathcal{H}_0^1}$$

$$\frac{\mathrm{d}V_t(x)}{\mathrm{d}t} = -\left\langle \mathcal{D}I^k\left(V_t\right), Z(x,.)\right\rangle_{\mathcal{H}_0^1}$$

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

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

Wide network limit

$$\frac{\mathrm{d}V_t^N(x)}{\mathrm{d}t} = -\left\langle \mathcal{D}I^k\left(V_t^N\right), Z_t^N(x,.)\right\rangle_{\mathcal{H}_0^1}$$

$$\frac{\mathrm{d}V_t(x)}{\mathrm{d}t} = -\left\langle \mathcal{D}I^k\left(V_t\right), Z(x,.)\right\rangle_{\mathcal{H}_0^1}$$

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

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

Theorem

For any $T > 0$,

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

Wide network limit

Theorem

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

$$\frac{\mathrm{d} V_t(x)}{\mathrm{d} t} = - \left\langle \mathcal{D} I^k (V_t), Z(x, .) \right\rangle_{\mathcal{H}_0^1}$$

Wide network limit

Theorem

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

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

Convergence in time

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

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

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

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

Convergence in time

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

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

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

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

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

Convergence in time

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

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

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

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

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

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

Convergence in time

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

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

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

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

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

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

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

☐

Convergence of Time-Stepping Deep Gradient Flow Methods

Dutch Math Finance Afternoon

Jasper Rou

June 7, 2024

j.g.rou@tudelft.nl www.jasperrou.nl