Convergence of Time-Stepping Deep Gradient Flow Methods

Jasper Rou[†,*] Chenguang Liu[†] Antonis Papapantoleon[†]

[*]j.g.rou@tudelft.nl, [†]Delft Institute of Applied Mathematics, TU Delft

Introduction

Neural networks are becoming increasingly popular in finance. However, good theoretical results are often lacking. In this research we prove convergence of the time deep gradient flow method, a neural network method which shows good numerical performance in solving partial differential equations (PDEs).

The price of an option can be written as the solution to the PDE:

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

We split the PDE into a symmetric and an asymmetric part:

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

Time Deep Gradient Flow

First, we discretize the PDE in time. Divide $[0, T]$ in intervals $(t_{k-1}, t_k]$ with $h = t_k - t_{k-1}$ and seek approximations U^k such that

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

Theorem 1: Akrivis and Crouzeix 2004

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

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

Second, we rewrite the solution of the discretized PDE as the minimizer of the functional

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

Theorem 2:

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

Proof

Define

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

Since w_* minimizes I^k, $\tau = 0$ minimizes i^k. So using integration by parts,

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

Theorem 3:

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

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

Third, we approximate this minimizer by a neural network.

Definition 1: 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 2: Neural network

Denote the parameters of the neural network by $\theta^N = \left(\beta^i, \alpha^i, c^i\right)_{i=1}^N$ and let $\frac{1}{2} < \delta < 1$. The space of neural networks with one layer and N neurons is

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

Theorem 4:

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

Training

We train the network using gradient descent with learning rate η_N:

$$\begin{aligned}
\frac{\mathrm{d}\theta_t^N}{\mathrm{d}t} &= -\eta_N \nabla_\theta I^k\left(V^N\left(\theta_t^N; x\right)\right), \quad \eta_N = N^{2\delta - 1}, \\
\frac{\mathrm{d}V_t^N(x)}{\mathrm{d}t} &= \nabla_\theta V^N\left(\theta_t^N; x\right) \cdot \frac{\mathrm{d}\theta_t^N}{\mathrm{d}t} \\
&= -\eta_N \nabla_\theta V^N\left(\theta_t^N; x\right) \cdot \nabla_\theta I^k\left(V^N\left(\theta_t^N; x\right)\right) \\
&= -\left\langle \mathcal{D} I^k\left(V_t^N\right), Z_t^N(x, .) \right\rangle_{\mathcal{H}_0^1}, \\
Z_t^N(x, y) &= N^{-1} \sum_{i=1}^{N} \nabla_{\beta,\alpha,c} \beta_t^i \psi\left(\alpha_t^i x + c_t^i\right) \cdot \nabla_{\beta,\alpha,c} \beta_t^i \psi\left(\alpha_t^i y + c_t^i\right).
\end{aligned}$$

We can prove that as the number of neurons goes to infinity, the neural network converges to the following gradient flow:

$$\begin{aligned}
\frac{\mathrm{d}V_t(x)}{\mathrm{d}t} &= -\left\langle \mathcal{D} I^k\left(V_t\right), Z(x, .) \right\rangle_{\mathcal{H}_0^1}, \\
Z(x, y) &= \mathbb{E}\left[\nabla_{\beta,\alpha,c} \beta_0^1 \psi\left(\alpha_0^1 x + c_0^1\right) \cdot \nabla_{\beta,\alpha,c} \beta_0^1 \psi\left(\alpha_0^1 y + c_0^1\right) \right].
\end{aligned}$$

Theorem 5:

For any $T > 0$,

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

Furthermore, we can prove that as the training time goes to infinity, this gradient flow converges to the minimizer of the functional.

Theorem 6:

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

Proof

We define the operator $\tilde{\mathcal{T}}$ by

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

Then $\tilde{\mathcal{T}}$ is a self-adjoint, positive definite trace class operator. So we can apply a spectral decomposition:

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

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

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

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

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