Time-Stepping Deep Gradient Flow Methods

Seminar National Technical University of Athens

Jasper Rou
joint work with Chenguang Liu & Antonis Papapantoleon

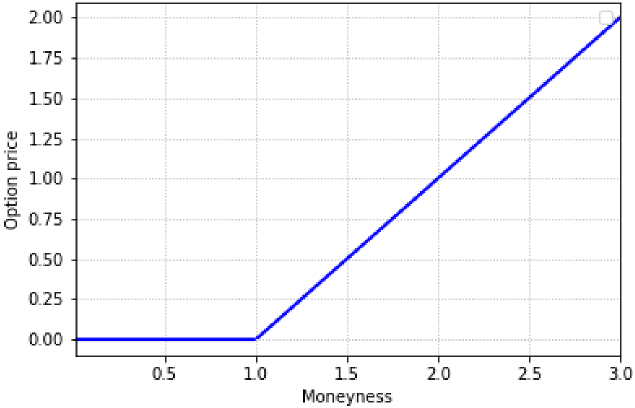
June 14, 2024

Options

A contract which gives the owner the right, but not the obligation, to buy a stock at a price K at a future time T

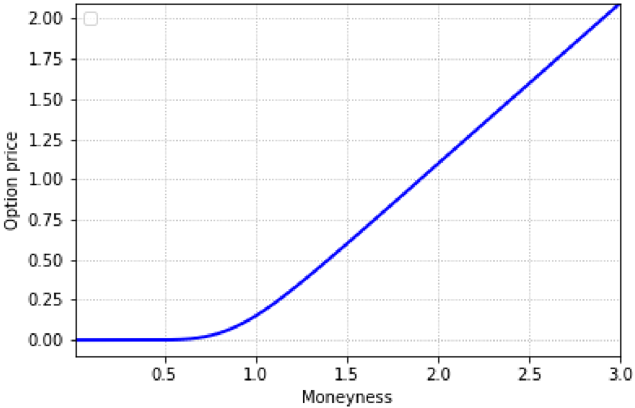
Pay-off

$$\Phi(S_T) = (S_T - K)^+$$



Pay-off

$$\Phi(S_T) = (S_T - K)^+$$

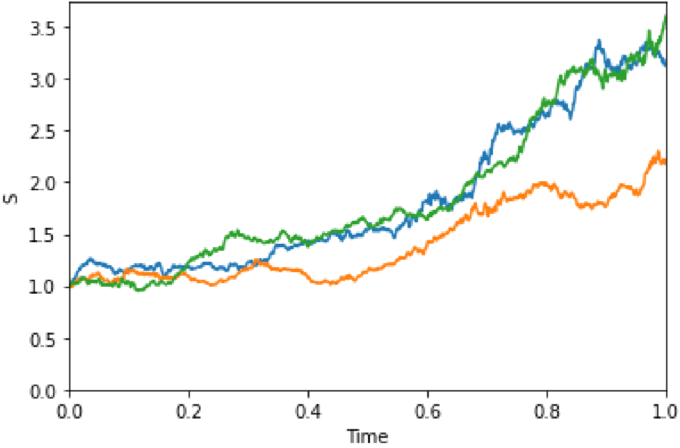


Stock price

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

Stock price

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



Pricing

Price of a derivative with pay-off $\Phi(S_T)$

$$u(t, S) = \mathbb{E}\left[\Phi(S_T) | S_t\right]$$

Pricing

Price of a derivative with pay-off $\Phi(S_T)$

$$u(t, S) = \mathbb{E}\left[\Phi(S_T)|S_t\right]$$

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

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

Pricing

Price of a derivative with pay-off $\Phi(S_T)$

$$u(t) = \mathbb{E}\left[e^{-r(T-t)}\Phi(S_T)|S_t\right]$$

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

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

Deep Galerkin Method

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

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

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

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

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

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

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

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

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

Example: Heston model

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

Example: Heston model

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

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

Example: Heston model

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

Example: Heston model

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

Example: Heston model

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

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

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

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.

Theorem

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

Proof.

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

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.

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

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

Time Deep **Gradient Flow**

Theorem

The minimizer $w_* \in \mathcal{H}_0^1\left(\mathbb{R}^d\right)$ 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.

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

$\square$

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 $\int_{\mathbb{R}^d} \psi\left(x\right) \mathrm{d}x \neq 0$.

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 $\int_{\mathbb{R}^d} \psi\left(x\right) \mathrm{d}x \neq 0$.

Definition (Neural network)

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

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

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 $\int_{\mathbb{R}^d} \psi\left(x\right) \mathrm{d}x \neq 0$.

Definition (Neural network)

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

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

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.

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

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

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.

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

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

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.

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

Proof.

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

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

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.

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

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

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

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.

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

Proof.

$\Longrightarrow$ I^k is continuous.

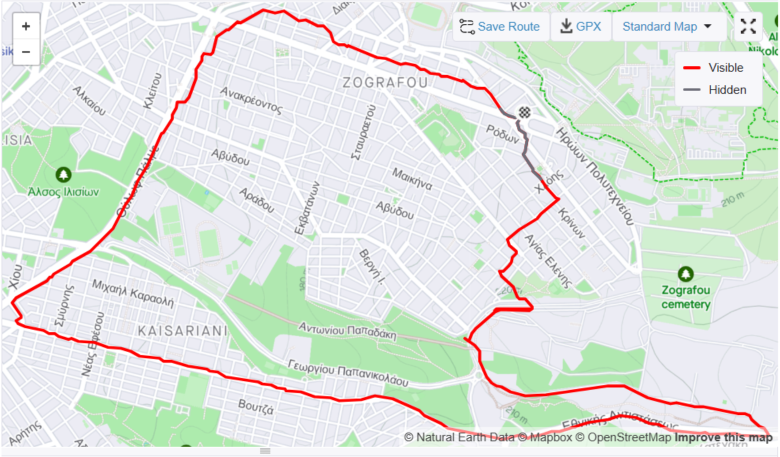
$\Longleftarrow$ $w_m \rightharpoonup w_*$. So $\mathcal{G}^n\left[w_m\right] \to \mathcal{G}^n\left[w_*\right]$.

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

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

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

Intermezzo: Gradient Descent

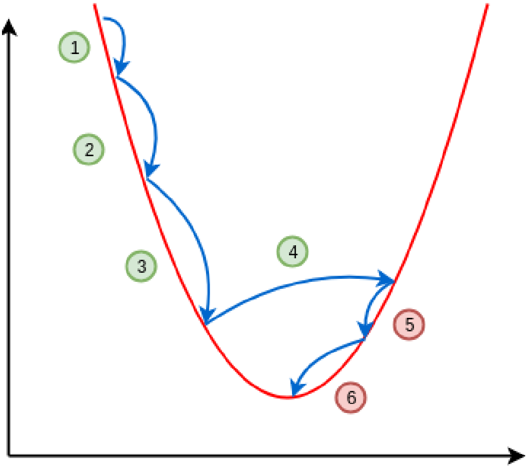


Intermezzo: Gradient Descent

Intermezzo: Gradient Descent

Intermezzo: Gradient Descent

Intermezzo: Gradient Descent



Intermezzo: Gradient Descent

$$\theta_{new} = \theta - \eta \nabla f(\theta)$$

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:

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

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

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

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

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

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

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

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

Wide network limit

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

Wide network limit

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

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

Wide network limit

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

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

Wide network limit

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

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

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

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

Wide network limit

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

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

Theorem

For any $T > 0$,

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

Convergence in time

Theorem

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

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

Convergence in time

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:

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

$\lambda_1 \geq \lambda_2 \geq ... > 0$, 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:

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

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

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

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:

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

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

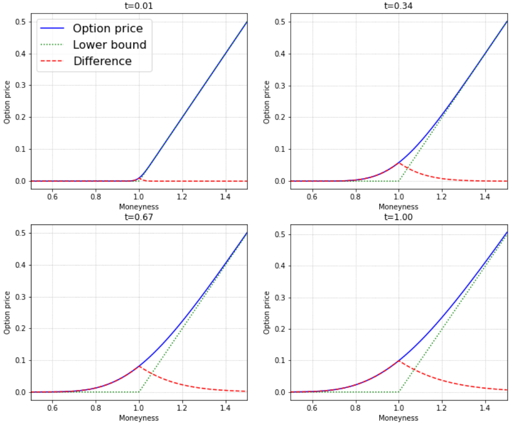
$$\begin{aligned}\frac{\mathrm{d}h_t^i}{\mathrm{d}t} :=& \frac{\langle \mathrm{d}(V_t - w_*), \tilde{e}_i \rangle}{\mathrm{d}t} = -\left\langle \tilde{\mathcal{T}}(V_t - w_*), \tilde{e}_i \right\rangle = -\left\langle V_t - w_*, \tilde{\mathcal{T}}\left(\tilde{e}_i\right)\right\rangle \\ =& -\lambda_i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

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

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

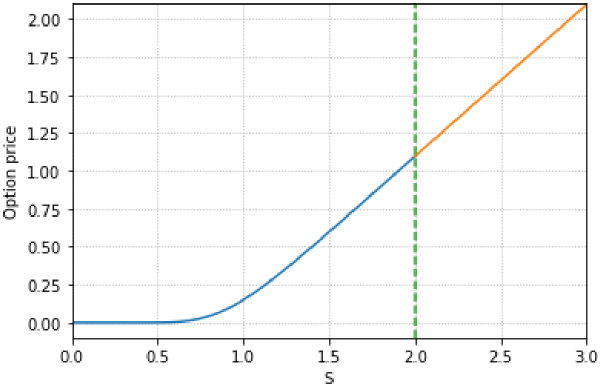
Architecture: base

No-arbitrage bound: $u(t, S) \geq S - Ke^{-rt}$



Architecture: linearization

$$u(x_p + y; \theta) = u(x_p; \theta) + y, \quad y > 0.$$



Architecture

$$
\begin{aligned}
S^1 &= \sigma_1\left(W^1\mathbf{x} + b^1\right), && \\
Z^l &= \sigma_1\left(U^{z,l}\mathbf{x} + W^{z,l}S^l + b^{z,l}\right), && l = 1, \ldots, L, \\
G^l &= \sigma_1\left(U^{g,l}\mathbf{x} + W^{g,l}S^1 + b^{g,l}\right), && l = 1, \ldots, L, \\
R^l &= \sigma_1\left(U^{r,l}\mathbf{x} + W^{r,l}S^l + b^{r,l}\right), && l = 1, \ldots, L, \\
H^l &= \sigma_1\left(U^{h,l}\mathbf{x} + W^{h,l}\left(S^l \odot R^l\right) + b^{h,l}\right), && l = 1, \ldots, L, \\
S^{l+1} &= \left(1 - G^l\right) \odot H^l + Z^l \odot S^l, && l = 1, \ldots, L, \\
f(\theta) &= \text{base} + \sigma_2\left(WS^{L+1} + b\right), && \sigma_2 > 0.
\end{aligned}
$$

Algorithm

1: Initialize $f(\theta^0; \mathbf{x}) = \Phi(S)$.

Algorithm

Initialize $f(\theta^0; \mathbf{x}) = \Phi(S)$.
for each time step $k = 1, ..., N_t$ **do**
    Initialize $\theta_0^k = \theta^{k-1}$.

Algorithm

Initialize $f(\theta^0; \mathbf{x}) = \Phi(S)$.
for each time step $k = 1, ..., N_t$ **do**
    Initialize $\theta_0^k = \theta^{k-1}$.
    for each sampling stage n **do**
        Generate random points $\mathbf{x}^i$ for training.

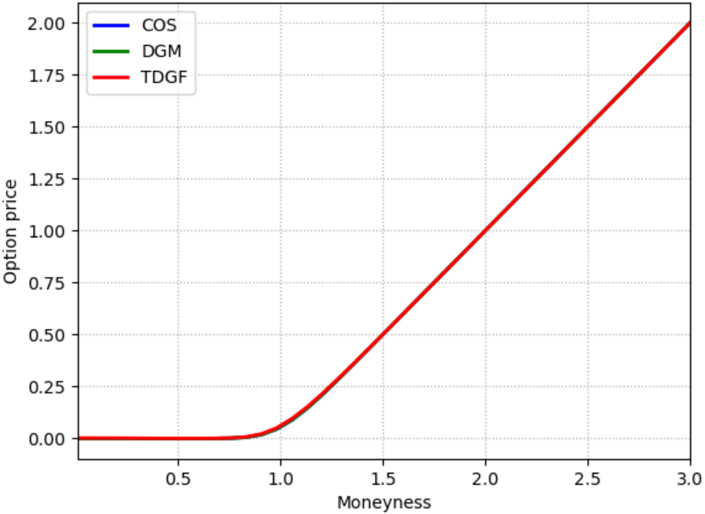
Algorithm

Initialize $f(\theta^0; \mathbf{x}) = \Phi(S)$.
for each time step $k = 1, ..., N_t$ **do**
  Initialize $\theta_0^k = \theta^{k-1}$.
  for each sampling stage n **do**
    Generate random points $\mathbf{x}^i$ for training.
    Calculate the cost functional $I^k(f(\theta_n^k; \mathbf{x}^i))$.

Algorithm

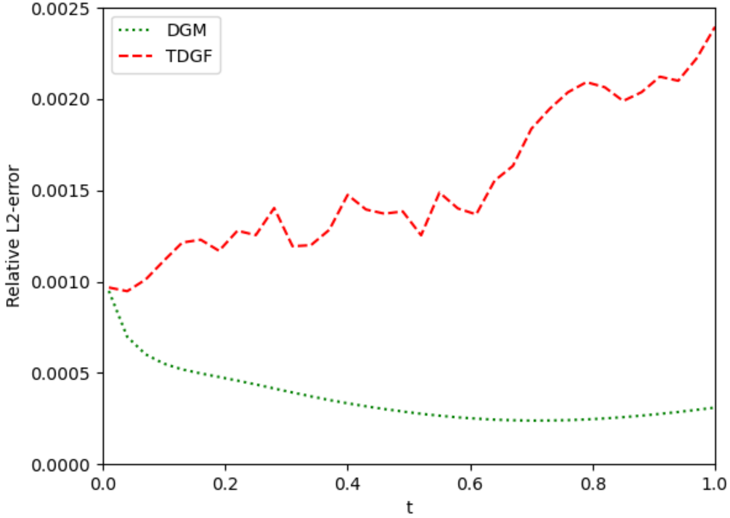
Initialize $f(\theta^0; \mathbf{x}) = \Phi(S)$.
for each time step $k = 1, ..., N_t$ **do**
    Initialize $\theta_0^k = \theta^{k-1}$.
    for each sampling stage n **do**
        Generate random points $\mathbf{x}^i$ for training.
        Calculate the cost functional $I^k(f(\theta_n^k; \mathbf{x}^i))$.
        Take a descent step $\theta_{n+1}^k = \theta_n^k - \alpha_n \nabla_\theta I^k(f(\theta_n^k; \mathbf{x}^i))$.
    end for
end for

Heston



COS
DGM
TDGF
Option price
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.5
1.0
1.5
2.0
2.5
3.0
Moneyness

Heston

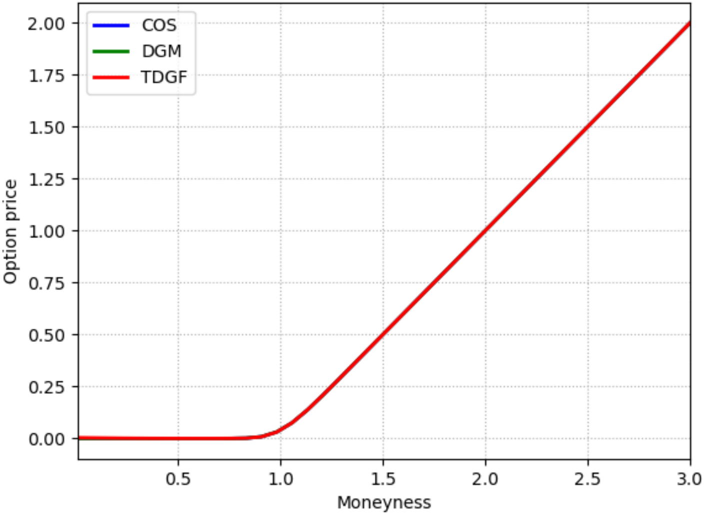


DGM
TDGF
Relative L2-error
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0.0
0.2
0.4
0.6
0.8
1.0
t

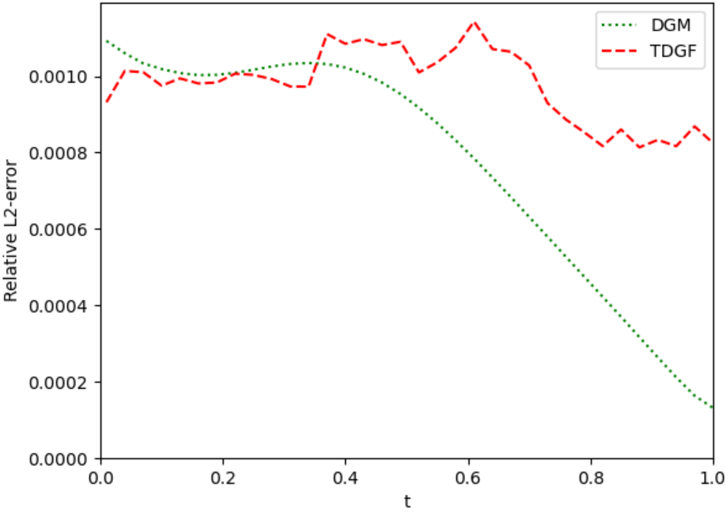
Lifted Heston

$$
\begin{aligned}
dS_t &= rS_t dt + \sqrt{V_t^n} S_t dW_t, & S_0 > 0, \\
V_t^n &= g^n(t) + \sum_{i=1}^{n} c_i^n V_t^{n,i}, & \\
dV_t^{n,i} &= -\left(\gamma_i^n V_t^{n,i} + \lambda V_t^n\right) dt + \eta \sqrt{V_t^n} dB_t, & V_0^{n,i} = 0, \\
g^n(t) &= V_0 + \lambda\theta \sum_{i=1}^{n} c_i^n \int_0^t e^{-\gamma_i^n (t-s)} ds. &
\end{aligned}
$$

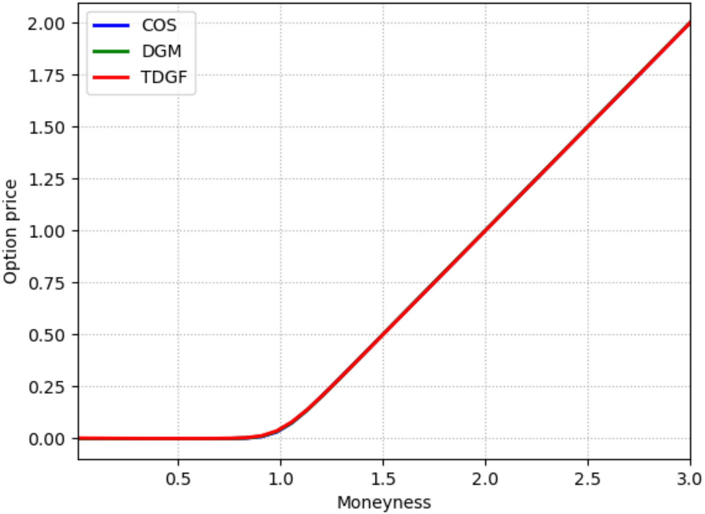
Lifted Heston, $n = 1$



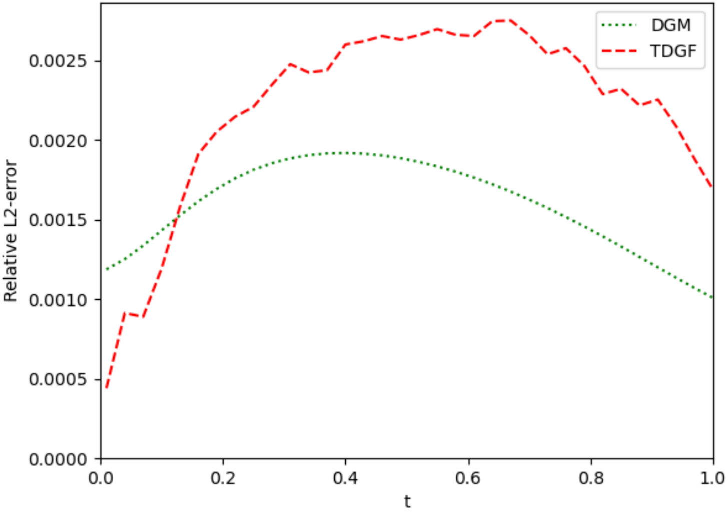
Lifted Heston, $n = 1$



Lifted Heston, $n = 20$



Lifted Heston, $n = 20$



Running times

Model	Heston	LH, n=1	LH, n=20
DGM	12.5×10^3	13.3×10^3	56.1×10^3
TDGF	6.0×10^3	6.4×10^3	7.6×10^3

Table: Training time

Running times

Model	Heston	LH, n=1	LH, n=20
DGM	12.5×10^3	13.3×10^3	56.1×10^3
TDGF	6.0×10^3	6.4×10^3	7.6×10^3

Table: Training time

Model	Heston	LH, n=1	LH, n=20
COS	0.018	8.9	10.4
DGM	0.0016	0.0034	0.0053
TDGF	0.0060	0.020	0.025

Table: Computing time

Time-Stepping Deep Gradient Flow Methods

Seminar National Technical University of Athens

Jasper Rou

June 14, 2024

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