Analysis of the Generalization Error: Empirical Risk Minimization Over Deep Neural Networks Overcomes the Curse of Dimensionality in the Numerical Approximation of Black-Scholes PDEs

Julius Berner[1], Philipp Grohs[1,2], Arnulf Jentzen[3]

[1]Faculty of Mathematics, University of Vienna
[2]Research Platform DataScience@UniVienna, University of Vienna
[3]Department of Mathematics, ETH Zürich

GAMM, 2019

Suitable Learning Problems for $d \in \mathbb{N}$

- input data: $X_d \sim U([u,v]^d)$
- label: random variable Y_d with $\|Y_d\|_{L^\infty} \leq D$

Definition (learning problem $\Rightarrow$ regression function)

$$\widehat{f}_d = \operatorname*{argmin}_{f\colon \mathbb{R}^d \to \mathbb{R}} \mathbb{E}\Big[\big(f(X_d) - Y_d\big)^2\Big]$$

Suitable Learning Problems for $d \in \mathbb{N}$

- input data: $X_d \sim U([u,v]^d)$
- label: random variable Y_d with $\|Y_d\|_{L^\infty} \leq D$

Definition (learning problem $\Rightarrow$ regression function)

$$\widehat{f}_d = \operatorname*{argmin}_{f\colon \mathbb{R}^d \to \mathbb{R}} \mathbb{E}\Big[\big(f(X_d) - Y_d\big)^2\Big]$$

- training data: $\big((X_d^{(i)}, Y_d^{(i)})\big)_{i\in\mathbb{N}}$ i.i.d. with $(X_d^{(1)}, Y_d^{(1)}) \sim (X_d, Y_d)$
- hypothesis class: $\mathcal{H}_d \subseteq C([u,v]^d)$ compact

Definition (ERM $\Rightarrow$ empirical target function)

$$\widehat{f}_{m,\mathcal{H}_d} \in \operatorname*{argmin}_{f\in\mathcal{H}_d} \frac{1}{m}\sum_{i=1}^m \Big(f(X_d^{(i)}) - Y_d^{(i)}\Big)^2$$

Neural Networks as Hypothesis Class

- affine linear mapping: $\mathcal{A}_{W,B}(x) := Wx + B$
- ReLU activation: $\varrho(x) := \max\{x, 0\}$
- clipping function: $\mathcal{C}_D(x) := \min\{|x|, D\}\,\mathrm{sgn}(x)$
- network architecture: $\mathbf{a} \in \mathbb{N}^{l+2}$

Definition (hypothesis class of clipped ReLU networks)

$$\mathcal{N}_{\mathbf{a},R} := \left\{ f \;\middle|\; \begin{array}{l} f = \mathcal{C}_D \circ \mathcal{A}_{W_l,B_l} \circ \varrho \circ \mathcal{A}_{W_{l-1},B_{l-1}} \circ \varrho \circ \cdots \circ \mathcal{A}_{W_0,B_0}, \\ W_i \in [-R,R]^{a_{i+1} \times a_i},\ B_i \in [-R,R]^{a_{i+1}},\ i = 0, \dots, l \end{array} \right\}$$

Neural Networks as Hypothesis Class

- affine linear mapping: $\mathcal{A}_{W,B}(x) := Wx + B$
- ReLU activation: $\varrho(x) := \max\{x, 0\}$
- clipping function: $\mathcal{C}_D(x) := \min\{|x|, D\}\operatorname{sgn}(x)$
- network architecture: $\mathbf{a} \in \mathbb{N}^{l+2}$

Definition (hypothesis class of clipped ReLU networks)

$$\mathcal{N}_{\mathbf{a},R} := \left\{ f \;\middle|\; \begin{array}{l} f = \mathcal{C}_D \circ \mathcal{A}_{W_l,B_l} \circ \varrho \circ \mathcal{A}_{W_{l-1},B_{l-1}} \circ \varrho \circ \cdots \circ \mathcal{A}_{W_0,B_0}, \\ W_i \in [-R,R]^{a_{i+1}\times a_i},\; B_i \in [-R,R]^{a_{i+1}},\; i = 0,\dots,l \end{array} \right\}$$

Definition (size of the hypothesis class)

$$\operatorname{size}(\mathcal{N}_{\mathbf{a},R}) := \max\left\{ R, \sum_{i=0}^{l} a_{i+1}a_i + a_{i+1} \right\}$$

Generalization without Curse

Assumption (approximation without Curse)

Assume there are $\mathcal{N}_{\mathbf{a}_{d,\varepsilon},R_{d,\varepsilon}}$ with $\text{size}(\mathcal{N}_{\mathbf{a}_{d,\varepsilon},R_{d,\varepsilon}}) \in \mathcal{O}(poly(d,\varepsilon^{-1}))$ and

$$\min_{f \in \mathcal{N}_{\mathbf{a}_{d,\varepsilon},R_{d,\varepsilon}}} \frac{1}{(v-u)^d} \left\| f - \widehat{f}_d \right\|^2_{L^2[u,v]^d} \leq \varepsilon.$$

Generalization without Curse

Assumption (approximation without Curse)

Assume there are $\mathcal{N}_{\mathbf{a}_{d,\varepsilon},R_{d,\varepsilon}}$ with $\text{size}(\mathcal{N}_{\mathbf{a}_{d,\varepsilon},R_{d,\varepsilon}}) \in \mathcal{O}(poly(d,\varepsilon^{-1}))$ and

$$\min_{f\in\mathcal{N}_{\mathbf{a}_{d,\varepsilon},R_{d,\varepsilon}}} \frac{1}{(v-u)^d} \left\| f - \widehat{f}_d \right\|^2_{L^2[u,v]^d} \leq \varepsilon.$$

Theorem (generalization without curse)

Then there exists $m \in \mathcal{O}(poly(d,\varepsilon^{-1})\ln(\varrho^{-1}))$ with

$$\mathbb{P}\left[\frac{1}{(v-u)^d} \left\| \widehat{f}_{m,\mathcal{N}_{\mathbf{a}_{d,\varepsilon},R_{d,\varepsilon}}} - \widehat{f}_d \right\|^2_{L^2[u,v]^d} \leq \varepsilon\right] \geq 1-\varrho.$$

Generalization without Curse

Assumption (approximation without Curse)

Assume there are $\mathcal{N}_{\mathbf{a}_{d,\varepsilon},R_{d,\varepsilon}}$ with $\text{size}(\mathcal{N}_{\mathbf{a}_{d,\varepsilon},R_{d,\varepsilon}}) \in \mathcal{O}(poly(d,\varepsilon^{-1}))$ and

$$\min_{f\in\mathcal{N}_{\mathbf{a}_{d,\varepsilon},R_{d,\varepsilon}}} \frac{1}{(v-u)^d} \left\| f - \widehat{f}_d \right\|^2_{L^2[u,v]^d} \leq \varepsilon.$$

Theorem (generalization without curse)

Then there exists $m \in \mathcal{O}(poly(d,\varepsilon^{-1})\ln(\varrho^{-1}))$ with

$$\mathbb{P}\left[\frac{1}{(v-u)^d}\left\|\widehat{f}_{m,\mathcal{N}_{\mathbf{a}_{d,\varepsilon},R_{d,\varepsilon}}} - \widehat{f}_d\right\|^2_{L^2[u,v]^d} \leq \varepsilon\right] \geq 1-\varrho.$$

Proof: covering number of $\mathcal{N}_{\mathbf{a},R}$ and Hoeffding's inequality

Partial Summary

Assume

- X_d uniformly distributed
- Y_d uniformly bounded
- i.i.d. training data
- $\widehat{f}_d$ can be approximated by ReLU networks without curse

Partial Summary

Assume

- X_d uniformly distributed
- Y_d uniformly bounded
- i.i.d. training data
- $\widehat{f}_d$ can be approximated by ReLU networks without curse

Then $\widehat{f}_{m,\mathcal{N}_{\mathbf{a},R}}$ (solution of to the ERM problem)

- approximates $\widehat{f}_d$ within accuracy ε with high probability
- with size$(\mathcal{N}_{\mathbf{a},R})$ and m scaling polynomially in d and ε^{-1}

Partial Summary

Assume

- X_d uniformly distributed
- Y_d uniformly bounded
- i.i.d. training data
- $\widehat{f}_d$ can be approximated by ReLU networks without curse

Then $\widehat{f}_{m,\mathcal{N}_{\mathbf{a},R}}$ (solution of to the ERM problem)

- approximates $\widehat{f}_d$ within accuracy ε with high probability
- with size($\mathcal{N}_{\mathbf{a},R}$) and m scaling polynomially in d and ε^{-1}

Can the assumptions be satisfied?

Kolmogorov PDE

- initial condition: $\varphi_d \in C(\mathbb{R}^d, [-D, D])$
- coefficient functions: $\sigma_d \colon \mathbb{R}^d \to R^{d \times d}$, $\mu_d \colon \mathbb{R}^d \to \mathbb{R}^d$ affine linear

Definition (Kolmogorov equation)

$$\begin{cases} \frac{\partial F_d}{\partial t}(t,x) = \frac{1}{2}\mathrm{Trace}\big(\sigma_d(x)\sigma_d^T(x)\mathrm{Hess}_x F_d(t,x)\big) + \big\langle \mu_d(x), \nabla_x F_d(t,x)\big\rangle \\ F_d(0,x) = \varphi_d(x) \end{cases}$$

for every $t \in [0, T]$, $x \in \mathbb{R}^d$

Kolmogorov PDE

- initial condition: $\varphi_d \in C(\mathbb{R}^d, [-D, D])$
- coefficient functions: $\sigma_d \colon \mathbb{R}^d \to R^{d \times d}$, $\mu_d \colon \mathbb{R}^d \to \mathbb{R}^d$ affine linear

Definition (Kolmogorov equation)

$$\begin{cases} \frac{\partial F_d}{\partial t}(t,x) = \frac{1}{2}\mathrm{Trace}\big(\sigma_d(x)\sigma_d^T(x)\mathrm{Hess}_x F_d(t,x)\big) + \big\langle \mu_d(x), \nabla_x F_d(t,x)\big\rangle \\ F_d(0,x) = \varphi_d(x) \end{cases}$$

for every $t \in [0, T]$, $x \in \mathbb{R}^d$

$\Rightarrow$ goal: approximately compute the function (end value)

$$[u,v]^d \ni x \mapsto F_d(T,x)$$

Learning Problem

- let $Y_d := \varphi_d(S_T^{X_d})$ where S^{X_d} is the solution processes to the stochastic differential equation (SDE)

$$\begin{cases} dS_t^{X_d} = \sigma_d(S_t^{X_d})dB_t^d + \mu_d(S_t^{X_d})dt \\ S_0^{X_d} = X_d \end{cases}$$

Learning Problem

- let $Y_d := \varphi_d(S_T^{X_d})$ where S^{X_d} is the solution processes to the stochastic differential equation (SDE)

$$\begin{cases} dS_t^{X_d} = \sigma_d(S_t^{X_d})dB_t^d + \mu_d(S_t^{X_d})dt \\ S_0^{X_d} = X_d \end{cases}$$

Theorem (regression function is solution to PDE - Beck et al. '18)

For a.e. $x \in [u, v]^d$ it holds that

$$F_d(T, x) = \widehat{f}_d(x).$$

Learning Problem

- let $Y_d := \varphi_d(S_T^{X_d})$ where S^{X_d} is the solution processes to the stochastic differential equation (SDE)

$$\begin{cases} dS_t^{X_d} = \sigma_d(S_t^{X_d})dB_t^d + \mu_d(S_t^{X_d})dt \\ S_0^{X_d} = X_d \end{cases}$$

Theorem (regression function is solution to PDE - Beck et al. '18)

For a.e. $x \in [u, v]^d$ it holds that

$$F_d(T, x) = \widehat{f}_d(x).$$

Proof: Feynman-Kac formula $F_d(T, x) = \mathbb{E}[\varphi_d(S_T^x)]$ and representation of regression function $\widehat{f}_d(x) = \mathbb{E}[Y_d | X_d = x]$

Approximation without Curse

- assume φ_d can be approximated by ReLU networks without curse of dimensionality

$\Rightarrow$ satisfied for applications in financial engineering

Approximation without Curse

- assume φ_d can be approximated by ReLU networks without curse of dimensionality

$\Rightarrow$ satisfied for applications in financial engineering

Theorem (approximation without curse - Grohs et al. '18)

Then there are $\mathcal{N}_{\mathbf{a},R}$ with size$(\mathcal{N}_{\mathbf{a},R}) \in \mathcal{O}(poly(d, \varepsilon^{-1}))$ and

$$\min_{f \in \mathcal{N}_{\mathbf{a},R}} \frac{1}{(v-u)^d} \|f - F_d(T, \cdot)\|^2_{L^2[u,v]^d} \leq \varepsilon.$$

Approximation without Curse

- assume φ_d can be approximated by ReLU networks without curse of dimensionality

$\Rightarrow$ satisfied for applications in financial engineering

Theorem (approximation without curse - Grohs et al. '18)

Then there are $\mathcal{N}_{\mathbf{a},R}$ with size$(\mathcal{N}_{\mathbf{a},R}) \in \mathcal{O}(poly(d, \varepsilon^{-1}))$ and

$$\min_{f \in \mathcal{N}_{\mathbf{a},R}} \frac{1}{(v-u)^d} \|f - F_d(T, \cdot)\|^2_{L^2[u,v]^d} \leq \varepsilon.$$

Proof: Monte-Carlo mean squared error and representation of SDE solution

ERM without Curse

Our assumptions are satisfied!

ERM without Curse

Our assumptions are satisfied!

Corollary (ERM solves the Kolmogorov PDE without curse)

There exists $\mathcal{N}_{\mathbf{a},R}$ and m with

- $\text{size}(\mathcal{N}_{\mathbf{a},R}) \in \mathcal{O}(poly(d,\varepsilon^{-1}))$
- $m \in \mathcal{O}(poly(d,\varepsilon^{-1})\ln(\varrho^{-1}))$
- $\mathbb{P}\left[\frac{1}{(v-u)^d}\left\|\widehat{f}_{m,\mathcal{N}_{\mathbf{a},R}} - \widehat{F}_d(T,\cdot)\right\|^2_{L^2[u,v]^d} \leq \varepsilon\right] \geq 1-\varrho.$

Proof: approximation without curse implies generalization without curse

Pricing of European Options without Curse

- capped European put option:
  $\varphi_d(x) = \min\left\{\max\left\{D - \sum_{i=1}^{d} c_{d,i} x_i, 0\right\}, D\right\}$

$\Rightarrow$ exactly representable by a neural network with size scaling linearly in d

Pricing of European Options without Curse

- capped European put option:
  $\varphi_d(x) = \min\left\{\max\left\{D - \sum_{i=1}^{d} c_{d,i}x_i, 0\right\}, D\right\}$

$\Rightarrow$ exactly representable by a neural network with size scaling linearly in d

$\Rightarrow$ quantitative version: there exist $\mathbf{a} = (d, a_1, a_2, 1)$, R and m with

  - $\mathrm{size}(\mathcal{N}_{\mathbf{a},R}) \in \mathcal{O}(d^2\varepsilon^{-2})$
  - $m \in \mathcal{O}(d^2\varepsilon^{-4}\ln(d\varepsilon^{-1}\varrho^{-1}))$
  - $\mathbb{P}\left[\frac{1}{(v-u)^d}\left\|\widehat{f}_{m,\mathcal{N}_{\mathbf{a},R}} - F_d(T,\cdot)\right\|^2_{L^2[u,v]^d} \leq \varepsilon\right] \geq 1 - \varrho.$

Numerical Experiments (Beck et al. '18)

Number of descent steps	Relative L^1-error	Relative L^2-error	Relative L^∞-error	Runtime in seconds
0	1.004285	1.004286	1.009524	1
100000	0.371515	0.371551	0.387978	437.9
250000	0.001220	0.001538	0.010039	1092.6
500000	0.000949	0.001187	0.005105	2183.8

Table: ERM with $\mathbf{a} = (100, 200, 200, 1)$ for a Black-Scholes PDE on $[90, 110]^{100}$

Thank you for your Attention!

Preprint available on arXiv:

Berner, J., Grohs, P., and Jentzen, A.
Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations.
arXiv:1809.03062 (9 2018).