National College of Ireland

Privacy-Preserving Logistic Regression for Cloud Environments

NCI Research Day 2024

Jorge M. Cortés Mendoza
IRC Postdoctoral Researcher
Cloud Competency Centre

June 2024, Dublin, Ireland

National College of Ireland

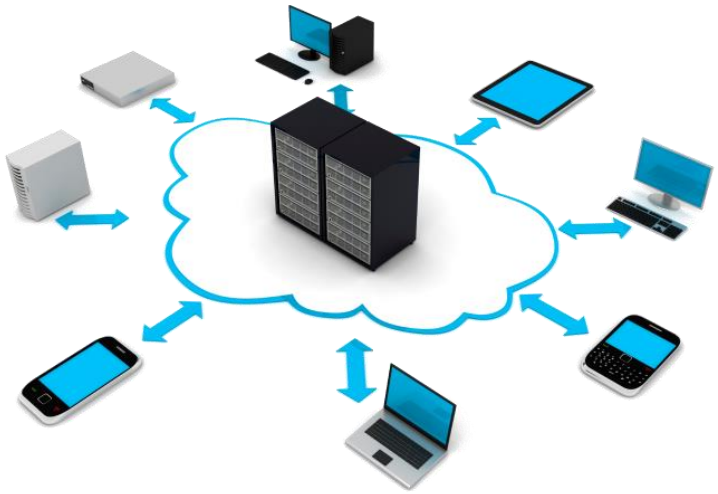
Content

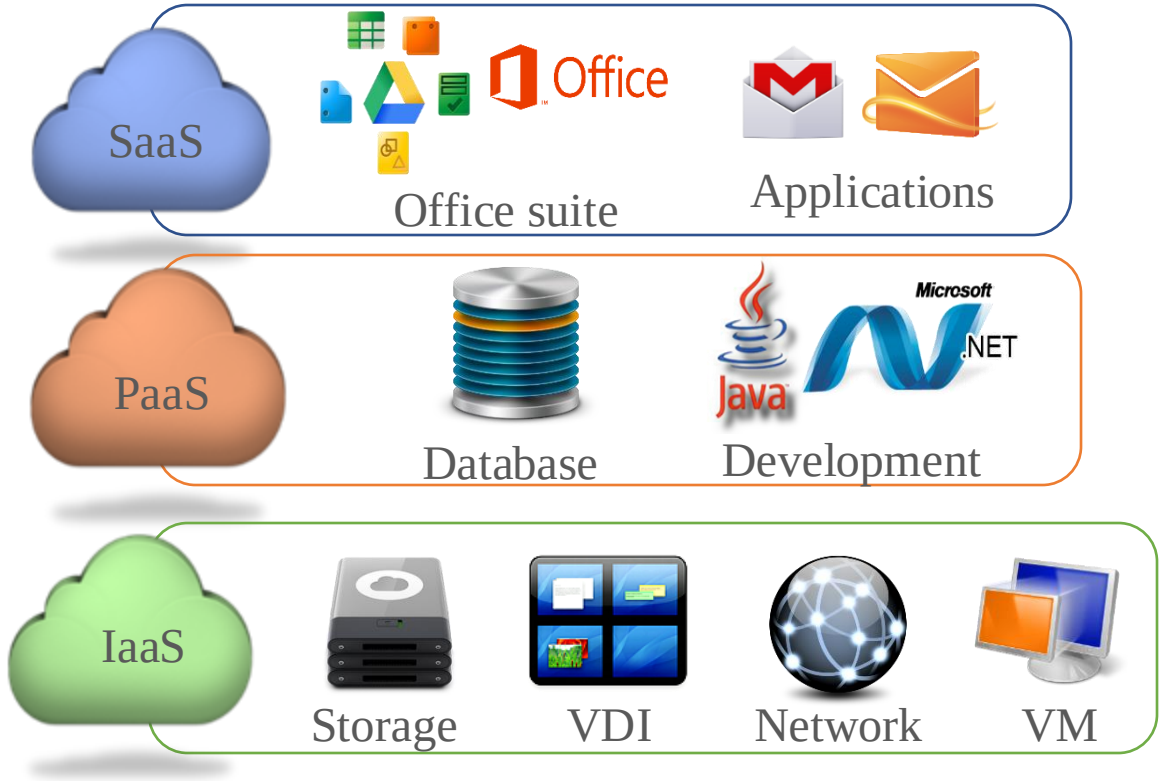
1. Machine Learning as a Service (MLaaS)
2. Privacy-preserving MLaaS
3. Homomorphic Encryption (HE)
4. HE limitations
5. Logistic Regression (LR)
6. Privacy-preserving LR
7. Latest advances in Privacy-preserving LR

Machine Learning as a Service

Cloud computing has been widely adopted because it allows users to acquire on-demand computing resources

In recent years, cloud computing has emerged as a flexible and scalable solution for Machine Learning (ML)

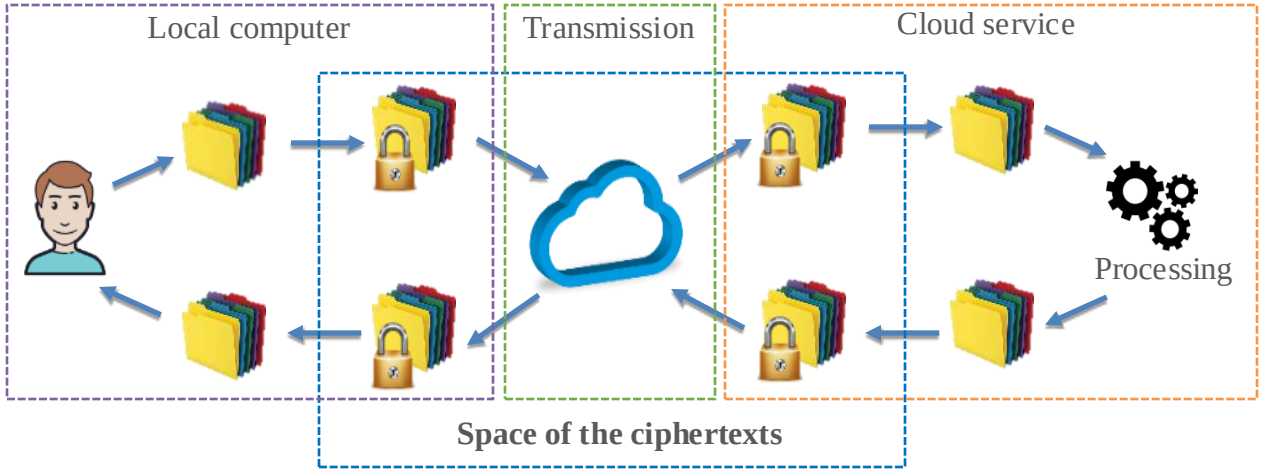




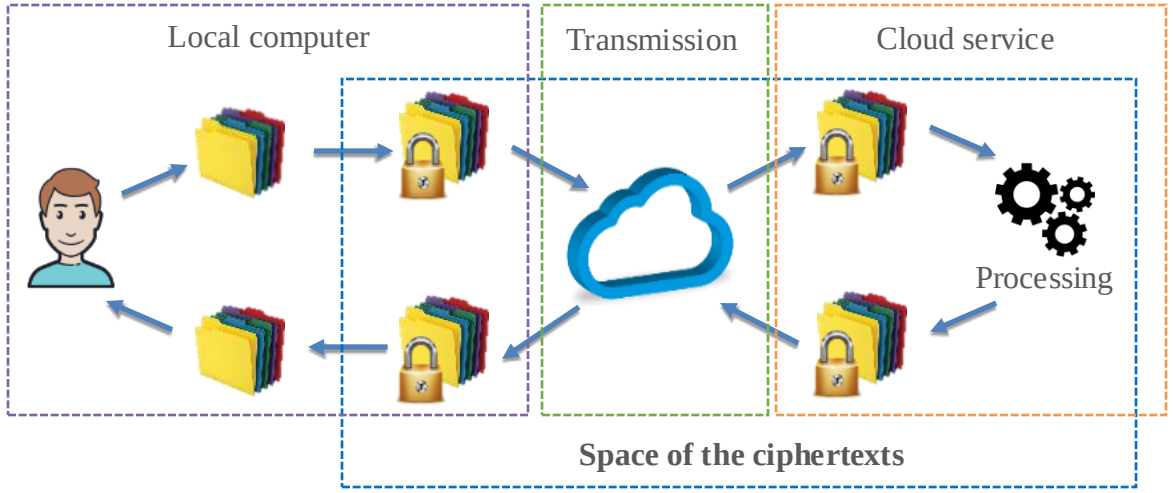
Privacy-Preserving Machine Learning as a Service

The use of a third-party provider can bring several cybersecurity issues because the data are processed on a shared infrastructure

Encrypting the data with conventional encryption does not solve the problem because data must be decrypted for statistical analysis



Traditional cloud storage and processing



Cloud environment with homomorphic encryption

Homomorphic Encryption (HE) is an alternative to address vulnerabilities and compute encrypted data

Homomorphic Encryption (HE)

The jewelry store problem illustrates the HE concept

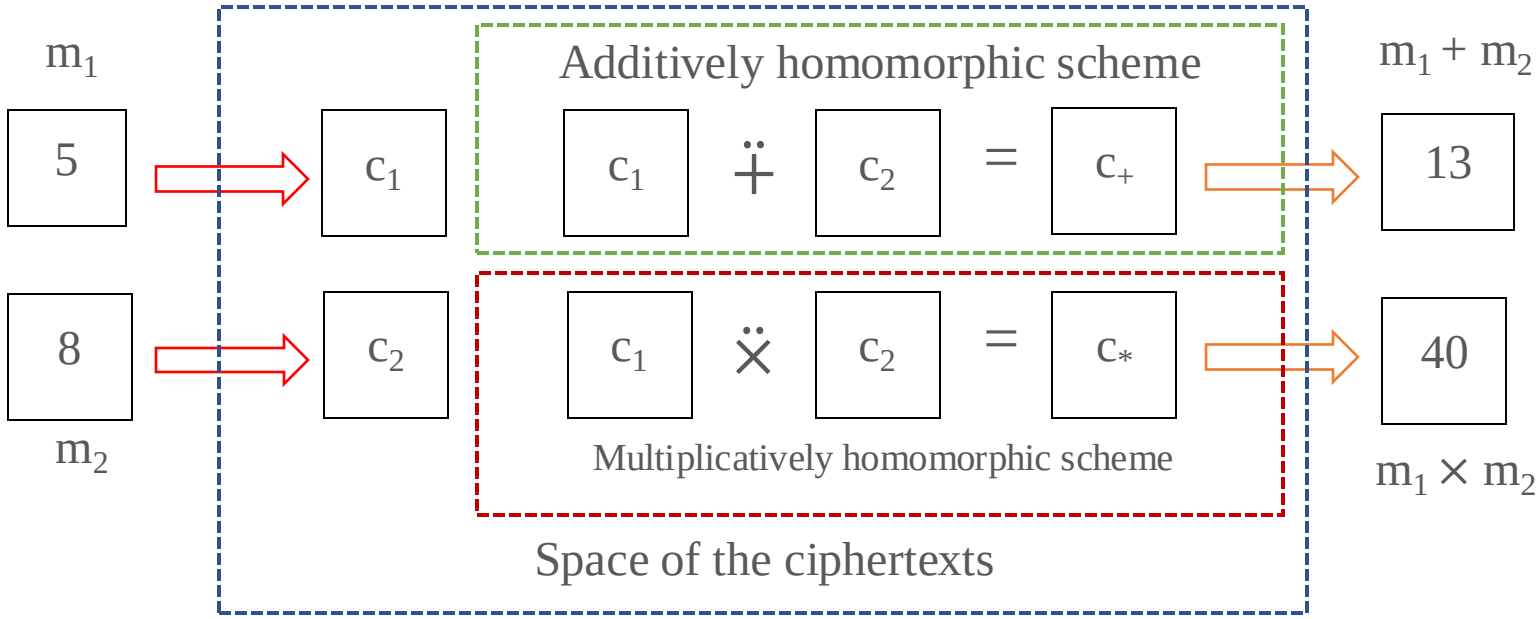
- Alice, a shop owner, wants her workers to assemble precious materials, such as gold and diamonds, into intricately designed rings and necklaces
- She does not want her workers to come in direct contact with the materials since she is afraid that they might steal the material

Alice uses a transparent, impenetrable glovebox to solve this problem

The gloves portray the homomorphism of the encryption scheme

Homomorphic Encryption (HE)

In additive and multiplicative homomorphic encryption scheme, operations on ciphertext space are mirrored in the plaintext space after decryption



Ciphertexts c_1 and c_2 encrypt the content of messages m_1 and m_2

- c_+ is created using c_1 and c_2, and its decryption produces $m_1 + m_2$
- $c_\times$ encrypts $m_1 \times m_2$

National College of Ireland

HE limitations

1) ***Limited number of operations***. Current HE schemes support only additions and multiplications
   - For ML, it becomes necessary to implement the comparison and division operations or determine the sign of a number
2) ***Noisy ciphertexts***. Noise growth limits the number of operations that can be accomplished. Each ciphertext has some noise that hides the message
   - If the noise is small, noise can be corrected
   - If the noise is large, decryption is hopeless

   Each homomorphic operation increases the underlying noises

3) ***Bootstrapping***. It reduces the noise in a ciphertext by generating a refreshed ciphertext from its equivalent exhausted one
   - A sophisticated and compute-intensive component

Logistic Regression (LR)

Logistic Regression (LR) is a statistical method for analyzing information where:

- A dataset $X \in \mathbb{R}^d$ and their labels $Y \in \{0,1\}$ are used to model a binary dependent variable
- The predict of a binary outcome considers the logistic function

The inference of LR considers the hypothesis $h_\theta(x^{(i)}) = g(\theta^T x^{(i)})$ where

- Logistic function: $g(z) = \frac{1}{1+e^{-z}}$
- Weights: $\theta^T = [\theta_0, \theta_1, \dots, \theta_d]^T$
- Data: $x^{(i)} = [1, {x_1}^{(i)}, {x_2}^{(i)}, \dots, {x_d}^{(i)}]^T$

The training phase of LR focuses on finding θ^*, the values of θ that minimizes the number of errors in the prediction

- θ^* is used to estimate the binary classification of new data

National College of Ireland

Logistic Regression (LR)

For $x' = [1, x_1, \ldots, x_d] \in \mathbb{R}^{d+1}$ is possible to guess its binary value $y' \in \{0,1\}$ by

$$y' = \begin{cases} 1 & if\ h_{\theta^*}(x') \geq \tau \\ 0 & if\ h_{\theta^*}(x') < \tau \end{cases}$$

- τ defines a variable threshold in $0 < \tau < 1$, typically with a value equal to 0.5

Gradient Descent (GD) is the optimization process to find θ^* according to the partial derivate of the cost function $J(\theta)$, represented by $\nabla_\theta J(\theta)$

Algorithm 1. Batch Gradient Descent

Input: $X, Y, \theta, \alpha, and\ nIter$.

Output: θ.

1 For $i \leftarrow 1$ to $nIter$

2 $\theta \leftarrow \theta - \alpha \times \nabla_\theta J(\theta, X, Y)$

Algorithm 2. Gradient Descent ($\nabla_\theta J$)

Input: $X, Y, and\ \theta$.

Output: $grad$.

1 For $i \leftarrow 1$ to $size(X)$

2 For $j \leftarrow 1$ to $size(\theta)$

3 $grad_j \leftarrow grad_j + (g(x^{(i)} \times \theta) - y^{(i)}) \times x_j{}^{(i)}$

4 For $j \leftarrow 1$ to $size(\theta)$

5 $grad_j \leftarrow grad_j / size(X)$

6 return $grad$

National College of Ireland

Privacy-preserving LR

The HE version of LR (HE-LR) substitutes +, -, ×, and g for their homomorphic versions

- $\bar{X}, \bar{Y}, \bar{\theta}$, and $\bar{\alpha}$ define the corresponding ciphertexts of X, Y, θ, and α,
- The homomorphic version of g ($\ddot{g}$) is a polynomial approximation with only $\ddot{+}$ and $\ddot{\times}$

The sigmoid function approximation is critical to the Privacy-preserving LR performance

- A ***higher degree*** approximation provides more ***accurate*** results but with time increase. Meanwhile, a lower degree approximation is less accurate but faster

Algorithm 1. Batch Gradient Descent

Input: $\bar{X}, \bar{Y}, \bar{\theta}, \bar{\alpha}, and\ nIter.$

Output: θ.

1 For $i \leftarrow 1$ to $nIter$

2 $\bar{\theta} \leftarrow \bar{\theta} \ddot{-} \bar{\alpha} \ddot{\times} \nabla_{\theta} J(\bar{\theta}, \bar{X}, \bar{Y})$

Algorithm 2. Gradient Descent ($\nabla_{\theta} J$)

Input: $\bar{X}, \bar{Y}, \bar{\theta}$, and $\overline{av}$

Output: $\overline{grad}$

1 For $i \leftarrow 1$ to $size(X)$

2 For $j \leftarrow 1$ to $size(\theta)$

3 $\overline{grad_j} \leftarrow \overline{grad_j} \ddot{+} (\ddot{g}(\overline{x^{(i)}} \ddot{\times} \theta) \ddot{-} y^{(i)}) \ddot{\times} \overline{x_j^{(i)}}$

4 For $j \leftarrow 1$ to $size(\theta)$

5 $\overline{grad_j} \leftarrow \overline{grad_j} \ddot{\times} \overline{av}$

6 return $\overline{grad}$

National College of Ireland

Latest advances in Privacy-preserving LR

Table I. Main characteristics of HE and MPC approaches for privacy-preserving LR

Approach	PAD	Algorithm	Metric	Method	Dataset	System	Ref
HE, MPC	-	SGD	A	Paillier	MNIST, notMNIST, CIFAR-10	Simulation	[1]
MPC	1	BGD	A	Additive SSS	iDASH (BC-TCGA, GSE2034)	AWS	[2]
MPC	1	BGD	A	Shamir's SSS	CIFAR-10, GISETTE	Amazon EC2	[3]
MPC	1	LiR, SGD, NN	Throughput	Semantic SSS	Superconductivity, FMA, Parkinson	Google Cloud	[4]
MPC	-	NRGD	A	Local training	UPHS fetal loss	Simulation	[5]
MPC	-	NRGD	AUC	Local training	Head and Neck Cancer (HNC)	Local system	[6]
MPC	3,5,7	NRGD	A, AUC	Additive SSS	Synthetic, Lbw, Pcs, Pima, Uis	Simulation	[7]
MPC	7	SGD	A	Shamir's SSS	CST, ACA	Simulation	[8]
MPC	1	BGD, SGD, MGD, NGD	A, AUC	RNS	Lbw, Mi, Nhanes3, Pcs, Pima, Uis	Simulation	[9]
HE	1	NRGD	AUC	FV	iDASH (Genomic), financial	Simulation	[10]
HE	7	BGD	p-values, F1	CKKS	iDASH	Simulation	[11]
HE	3	NGD	A, AUC, K-S values	CKKS	Korea Credit Bureau (KCB), MNIST	Simulation	[12]
HE	3	NGD, NRGD	A, AUC	CKKS	iDASH, Lbw, Mi, Nhanes3, Pcs, Uis	Public cloud	[13]
HE, MPC	-	SGD	Overhead	PHE	Not described	Simulation	[14]
HE	7	BGD	A, AUC, F1, P, R	CKKS	Mi, Nhanes3, Uis	Simulation	[15]-[17]
HE	-	BGD	A	CKKS	Digits (scikit-learn library)	Local system	[18]
FL	-	SGD	A, AUC	Symmetric	Pima, BCWD, BDM	Local system	[19]
FL	1	SGD	Time	Paillier	MNIST	Local system	[20]
FL	-	LiD, RR, BGD	MAE	Paillier	BCD, Diabetes Dataset (DD), UCID	Local system	[21]
FL	-	BGD	P, R	SS	DD, WIBC, HDD, ACAD	Local system	[22]

National College of Ireland

Latest advances in Privacy-preserving LR

Table II. State-of-the-art logistic function approximations.

Approximation function	Method
$g_{1a}(x) = 0.5 + 0.25x$	Taylor series
$g_{3a}(x) = 0.5 + 1.20096(x/8) - 0.81562(x/8)^3$	Least squares
$g_{3b}(x) = 0.5 + 0.15x - 0.0015x^3$	Least squares
$g_{3c}(x) = 0.5 + x/4 - x^3/48$	Taylor expansion
$g_{3d}(x) = 0.499999999992724 + 0.139786538317376x - 1.455183675923 46e-13x^2 - 0.001003773735684 84x^3$	Chebyshev
$g_{5a}(x) = 0.5 + (1.53048)(x/8) - (2.3533056)(x/8)^3 + (1.3511295)(x/8)^5$	Least squares
$g_{5b}(x) = 0.500000000006453 + 0.187819515164365x - 5.65619279205865e-13x^2 - 0.00336794817488311x^3 + 5.82078097744257e-15x^4 + 2.04674243327 92e-5x^5$	Chebyshev
$g_{7a}(x) = 0.5 + 1.73496(x/8) - 4.19407(x/8)^3 + 5.43402(x/8)^5 - 2.50739(x/8)^7$	Least squares
$g_{7b}(x) = 0.5 + 1.735(x/8) - 4.194(x/8)^3 + 5.434(x/8)^5 - 2.507(x/8)^7$	Least squares
$g_{7c}(x) = 0.5 + 0.249995x - 0.0207869x^3 + 0.00198305x^5 - 0.000135007x^7$	Lagrange interpolation
$g_{7d}(x) = 0.500000000015461 + 0.216030242339756x - 3.00134166245124e-12x^2 - 0.006526130098898 38x^3 + 8.44014964905896e-14x^4 + 9.18419138902492e-5x^5 - 5.82079696725621e-16x^6 - 4.34913635838155e-7x^7$	Chebyshev
$g_{9a}(x) = 0.500000000005353 + 0.231624826001611x - 2.49775180627496e-12\,x^2 - 0.009784870092723 3x^3 + 1.47390762630842e-13x^4 + 0.000229352354062705x^5 - 2.60854055994519e-15x^6 - 2.42773327147286e-6x^7 + 1.39698499452418e-17x^8 + 9.32721914680041e-9x^9$	Chebyshev

National College of Ireland

Latest advances in Privacy-preserving LR

Figure 1. Logistic function approximations in the literature for the interval [-10, 10].

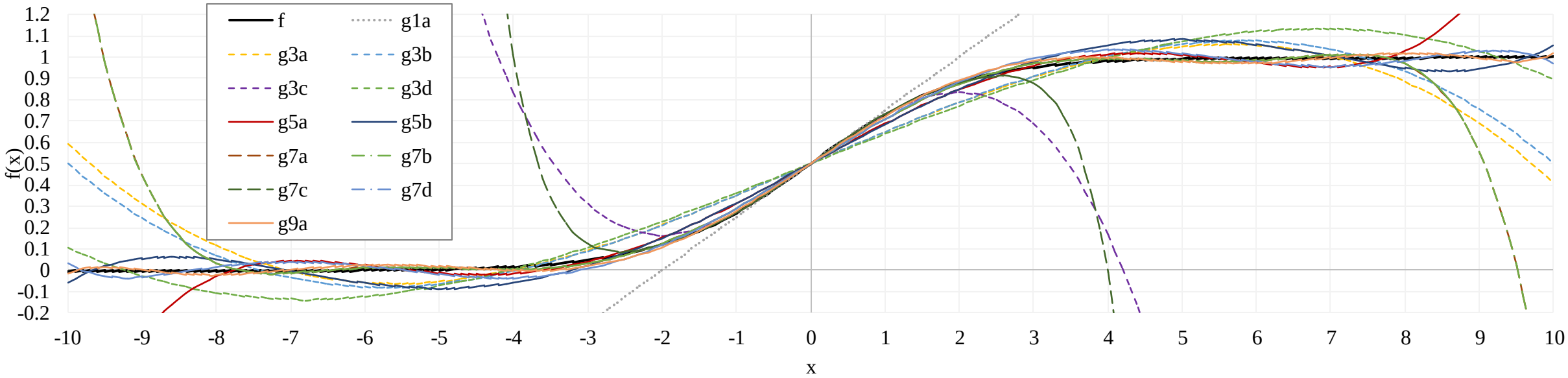


Table III. Characteristics of the 12 state-of-the-art logistic function approximations.

Name	$g_{1a}(x)$	$g_{3a}(x)$	$g_{3b}(x)$	$g_{3c}(x)$	$g_{3d}(x)$	$g_{5a}(x)$	$g_{5b}(x)$	$g_{7a}(x)$	$g_{7b}(x)$	$g_{7c}(x)$	$g_{7d}(x)$	$g_{9a}(x)$
Evaluation interval	[-2,2]	[-8,8]	[-8,8]	[-2,2]	[-10,10]	[-8,8]	[-10,10]	[-8,8]	[-8,8]	[-1.6,1.6]	[-10,10]	[-10,10]
L_1	13.330	83.976	88.635	3.4619	158.72	41.124	90.473	17.772	17.783	**0.0011**	50.127	27.362
L_2	0.9725	2.3854	2.4841	0.3036	3.973	1.1599	2.3147	0.5058	0.5058	**9.79E−5**	1.2979	0.7107
L_∞	0.1192	0.1143	0.0982	0.0474	0.136	0.0471	0.0894	0.0321	0.0317	**1.46E−5**	0.0525	0.0289
Evaluation time (ms)	6.8973	6.2826	6.6863	**5.7622**	24.548	20.857	41.155	32.865	32.214	32.4511	59.89	87.933

Conclusion

We analyze the latest advances in privacy-preserving logistic regression solutions for processing confidential data using HE

We present the characteristics of the most recent advances in the field: algorithms, evaluation metrics, used datasets, approximation functions, implementation characteristics, etc.

We study the accuracy and execution time of the state-of-the-art polynomial approximations for the sigmoid function using CKKS with a security level of 128 bits

National College of Ireland

Questions?

National College of Ireland

References

1. Edemacu, K., Kim, J. W.: Multi-party privacy-preserving Logistic Regression with poor quality data filtering for iot contributors, Electronics, vol. 10 (17), p. 2049, (2021)
2. De Cock, M., Dowsley, R., Nascimento, A. C., et al.: High-performance logistic regression for privacy-preserving genome analysis, BMC Medical Genomics, vol. 14, pp. 1–18, (2021)
3. So, J., Güler, B., Avestimehr, A. S.: Codedprivateml: A fast and privacy-preserving framework for distributed machine learning, Journal on Selected Areas in Information Theory, vol. 2 (1), pp. 441– 451, (2021)
4. Patra, A., Suresh, A.: BLAZE: blazing fast privacy-preserving machine learning, in NDSS, The Internet Society, (2020)
5. Duan, R., Boland, M. R., Liu, Z., et al.: Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm, Journal of the American Medical Informatics Association, vol. 27 (3), pp. 376–385, (2020)
6. Bogowicz, M., Jochems, A., Deist, T. M., et al.: Privacy-preserving distributed learning of radiomics to predict overall survival and hpv status in head and neck cancer, Scientific reports, vol. 10 (1), p. 4542, (2020)
7. Ghavamipour, A. R., Turkmen, F., and Jiang, X.: Privacy-preserving logistic Regression with secret sharing, Medical Informatics and Decision Making, vol. 22 (1), pp. 1–11, (2022)
8. Zhou, H.: Information-theoretically secure multi-party linear Regression and logistic Regression, in CCGrid, pp. 192–199, IEEE, July (2023)
9. Cortés-Mendoza, J. M., Tchernykh, A. Babenko, M., et al.: Multi-cloud privacy-preserving logistic Regression, in RRuSCDays, vol. 1510 of CCIS, pp. 457–471, Springer, (2021)

National College of Ireland

References

10. Bonte, C., Vercauteren, F.: Privacy-preserving logistic regression training, BMC medical genomics, vol. 11, pp. 13–21, (2018)
11. Kim, D., Son, Y., Kim, D., et al.: Privacy-preserving approximate GWAS computation based on homomorphic encryption," BMC Medical Genomics, vol. 13, no. 7, pp. 1–12, (2020)
12. Han, K., Hong, S., Cheon, J. H., Park, D.: Logistic Regression on homomorphic encrypted data at scale, in AAAI, vol. 33, pp. 9466–9471, (2019)
13. Chiang, J.: Privacy-preserving logistic regression training with a faster gradient variant, arXiv preprint arXiv:2201.10838, (2022)
14. Zhou, Y., Song, L., Liu, Y., et al.: A privacy-preserving logistic regression-based diagnosis scheme for digital healthcare, Future Generation Computer Sys., vol. 144, pp. 63–73, (2023)
15. Yu, X., Zhao, W., Huang, Y., et al.: Privacy-preserving outsourced logistic Regression on encrypted data from homomorphic encryption, Security and Communication Networks, ID 1321198, (2022)
16. Yu, X., Zhao, W., Tang, D., Liang, K.: Privacy-preserving vertical collaborative logistic Regression without a trusted third-party coordinator, Security and Communication Networks, ID 5094830, (2022).
17. Yu, X., Tang, D. Zhao, W.: Privacy-preserving cloud-edge collaborative learning without a trusted third-party coordinator, Journal of Cloud Computing, vol. 12 (1), pp. 1–11, (2023)
18. Liu, C., et al.: Efficient and privacy-preserving logistic regression scheme based on leveled fully homomorphic encryption, in INFOCOM, pp. 1–6, IEEE, (2022)

National College of Ireland

References

19. Zhao, J., Zhu, H., Wang, F., et al.: VFLR: An efficient and privacy-preserving vertical federated framework for logistic Regression, Transactions on Cloud Comp., vol. 11 (4), (2023)
20. He, D., Du, R., Zhu, S., Zhang, M., Liang, K., and Chan, S. Secure logistic regression for vertical federated learning. IEEE Internet Computing, vol. 26 (2), pp. 61-68, (2021).
21. Wang, F., Zhu, H., Lu, R., Zheng, Y., Li, H.: A privacy-preserving and non-interactive federated learning scheme for regression training with gradient descent. Information Sciences, 552, 183-200, (2021).
22. Zhang, Y., Tang, M.: VPPLR: Privacy-preserving logistic Regression on vertically partitioned data using vectorization sharing. Information Security and Applications, 82, (2024).