

Adaptive Nonlinear Approximation for Fast Privacy-Preserving Machine Learning

Emiliana Terron

Department of Computer Science, George Mason University, Fairfax, VA, USA.
emiliana.terron939@gmu.edu

Birgitt Mangold

Department of Computer Science, University of North Texas, Denton, TX, USA.
birgittm@unt.edu

Abstract

Privacy-preserving machine learning has matured from a collection of cryptographic demonstrations into a systems problem involving data governance, hardware acceleration, distributed trust, and service-level accountability. Nonlinear activation functions, particularly rectified linear units, remain among the most expensive operations in secure inference and training because they require comparisons, branching, or oblivious selection inside encrypted or secret-shared computation. This paper examines adaptive nonlinear approximation as a system-level strategy for fast privacy-preserving machine learning. Rather than treating approximation as a purely model-side compression technique, the discussion frames it as an architectural control surface that links cryptographic protocols, hardware accelerators, network topology, operational monitoring, and regulatory obligations. The paper synthesizes work on secure multi-party computation, homomorphic encryption, trusted execution, differential privacy, and federated learning to analyze how approximation decisions propagate through latency, energy use, robustness, fairness, and auditability. It argues that adaptive approximation can materially reduce communication and computation costs, but only when governed by explicit approximation budgets, continuous drift detection, fairness monitoring, and clear redress mechanisms. Cross-domain comparisons from cloud inference, edge deployments, healthcare, finance, and public-sector applications illustrate the structural trade-offs. The conclusion is that adaptive nonlinear approximation is a promising but contingent infrastructure strategy, one whose benefits depend less on a single algorithmic trick than on accountable deployment across heterogeneous cryptographic and organizational environments.

Keywords

privacy-preserving machine learning; secure multi-party computation; homomorphic encryption; adaptive approximation; nonlinear activation; trustworthy systems; governance.

1. Introduction

Privacy-preserving machine learning addresses a fundamental tension in contemporary data-intensive systems: models improve through access to large and diverse datasets, while legal, ethical, and commercial constraints increasingly limit the movement and disclosure of sensitive records. Approaches such as federated learning, secure aggregation, differential privacy, secure multi-party computation, and homomorphic encryption each address part of this tension [1], [2], [3], [19]. Yet the practical performance of these approaches is often determined not by the headline cryptographic primitive but by mundane computational details, especially nonlinear operations. In plaintext neural networks, rectified linear units and related

activation functions are inexpensive. Inside privacy-preserving computation, they become expensive because comparisons, branching, and data-dependent control flow must be expressed through cryptographic operations that are optimized for linear arithmetic or Boolean circuits. The result is a structural bottleneck that affects training, inference, hardware utilization, and even the economic viability of privacy-preserving services.

Adaptive nonlinear approximation responds to this bottleneck by replacing exact nonlinear functions with carefully bounded approximations whose form, degree, or placement depends on context. A static approximation applies a single polynomial or piecewise function everywhere. An adaptive approximation may vary by layer, input distribution, security parameter, latency budget, hardware platform, or regulatory domain. This makes approximation a cross-layer design choice rather than a standalone model compression method. It also introduces new governance questions because an approximation that improves speed may alter model behavior in ways that are difficult to audit, especially when decisions affect credit, health, employment, or public benefits. The core claim of this paper is that adaptive nonlinear approximation should be treated as an infrastructural capability, not merely an algorithmic optimization. Its value depends on how it is integrated with secure computation protocols, deployment pipelines, monitoring systems, and policy controls.

The paper proceeds through a system-level analysis. It first situates nonlinear approximation within the broader privacy-preserving machine learning stack. It then examines adaptive approximation as an architectural principle that coordinates model design, cryptographic protocol selection, and runtime scheduling. Subsequent sections discuss infrastructure and deployment, robustness and fairness, evaluation and sustainability, and cross-domain implications. Throughout, the focus remains on structural trade-offs rather than isolated accuracy metrics. The aim is to clarify when adaptive approximation is appropriate, what it costs, and how it should be governed when privacy-preserving machine learning becomes part of critical socio-technical infrastructure.

2. System Context and the Cryptographic Cost of Nonlinearity

The privacy-preserving machine learning landscape is heterogeneous. Federated learning keeps raw data on client devices but still requires secure aggregation and privacy accounting to prevent leakage from updates [1], [3], [19]. Differential privacy provides formal guarantees against inference from released statistics or gradients but imposes utility costs that must be managed across repeated queries [2], [18]. Secure multi-party computation distributes trust across parties and supports private inference or training through secret sharing, garbled circuits, and oblivious transfer [4], [5], [6], [7], [13], [14], [17]. Homomorphic encryption allows computation directly on encrypted data, particularly for linear operations, but nonlinear functions require approximation or interactive bootstrapping [8], [9], [10], [11]. Trusted execution environments offer another trust boundary, though they introduce hardware supply-chain and attestation assumptions. These approaches are often combined in layered architectures because no single primitive satisfies all requirements.

Within this landscape, nonlinear activation functions are expensive for different reasons across protocols. In secure multi-party computation, a rectified linear unit requires a comparison or oblivious selection, which can be implemented with garbled circuits, oblivious transfer, or specialized protocols. These operations increase communication rounds and bandwidth consumption. In homomorphic encryption, addition and multiplication are natural, but comparison and branching are not. A rectified linear unit must therefore be approximated by polynomials, rational functions, or discretized representations [8], [9], [10], [11]. In hybrid

systems, the compiler decides where to switch between arithmetic and Boolean domains, and each switch incurs cost [5], [12], [14]. Hardware acceleration can improve throughput, but it does not eliminate the underlying mismatch between neural network nonlinearity and cryptographic arithmetic [15]. Consequently, even modest reductions in the number or complexity of nonlinear operations can produce significant system-level gains.

The literature on private inference has repeatedly identified this bottleneck. Early systems demonstrated that neural networks could be evaluated securely, but with high latency or limited depth [4], [7], [8]. Subsequent work improved protocol efficiency, compiler optimization, and GPU acceleration [10], [12], [13], [15]. A parallel line of research has reduced the number of rectified linear units needed during private inference [16]. That direction is important because it reframes the problem from “make each nonlinearity cheaper” to “reduce the demand for nonlinearities in the first place.” Adaptive approximation extends this logic by allowing the approximation strategy to change according to operational conditions. The key systems question is not whether approximation works in principle, but whether it can be controlled without undermining privacy, fairness, or accountability.

3. Adaptive Approximation as an Architectural Principle

Adaptive nonlinear approximation operates at the intersection of model architecture, cryptographic protocol, and runtime orchestration. A model may use a low-degree polynomial approximation in early layers where feature extraction tolerates smoothing, while preserving more precise nonlinear behavior in later layers where decision boundaries are sensitive. Alternatively, the system may select approximation degree based on the available network bandwidth, the number of participating parties, the latency target, or the energy budget of an edge device. During training, approximation can be scheduled to reduce the cost of backpropagation through nonlinearities. During inference, it can be tuned to meet a service-level objective. This context sensitivity distinguishes adaptive approximation from fixed post-training quantization or pruning.

The architectural benefit is that approximation decisions can be made visible to the compiler and runtime. A privacy-preserving compiler can reserve expensive comparison operations for layers that require them and use encrypted arithmetic for layers that can tolerate polynomial approximations. The runtime can then negotiate approximation budgets with the cryptographic protocol, perhaps reducing degree when communication is constrained or increasing precision when regulatory audit requires stronger fidelity. This creates a feedback loop among model performance, cryptographic cost, and infrastructure conditions. However, the same feedback loop creates governance risk. If approximation changes dynamically, then the effective model may differ across users, regions, or time periods. A service that behaves accurately for one population may degrade for another because the approximation policy interacts with distributional differences. Therefore, adaptive approximation requires observability, versioning, and continuous evaluation.

From a security perspective, adaptivity introduces side-channel and leakage concerns. If approximation degree depends on input values or data distributions, then timing, memory access, or communication patterns may reveal information about sensitive inputs. A privacy-preserving system should therefore avoid data-dependent control flow unless it is explicitly protected. Approximation decisions should preferably depend on public or aggregate parameters rather than individual records. When individual-level adaptation is necessary, it must be blinded, padded, or otherwise made oblivious. These requirements increase system complexity, but they are consistent with established practices in secure computation, where

data-dependent branching is generally avoided [4], [6], [14]. The architectural challenge is to obtain the efficiency benefits of adaptivity while preserving the indistinguishability assumptions on which privacy guarantees rest.

A second structural trade-off concerns accuracy and robustness. Approximation can smooth sharp nonlinearities, which may improve generalization in noisy environments but also reduce sensitivity to rare or adversarial inputs. In safety-critical applications, this is not merely a performance issue; it is a question of whether the system can detect when approximation has pushed a decision beyond acceptable bounds. Adaptive approximation should therefore be paired with uncertainty estimation, out-of-distribution detection, and fallback mechanisms. If the approximation budget is exhausted or the input distribution shifts, the system may need to revert to a more expensive exact or higher-fidelity path. Such fallbacks must be designed into the infrastructure from the beginning, not added after deployment. In this sense, adaptive approximation is less like a single model optimization and more like a control system for privacy-preserving computation.

4. Infrastructure, Deployment, and Sustainability

Deploying adaptive nonlinear approximation at scale requires infrastructure that spans cloud data centers, edge nodes, and organizational boundaries. In a cloud inference service, the provider may operate secure computation accelerators, manage cryptographic key material, and enforce tenant isolation. The approximation policy must be compatible with multi-tenancy, because one tenant’s latency target may differ from another’s privacy requirements. In edge deployments, devices may have limited power, intermittent connectivity, and heterogeneous hardware, making approximation a way to reduce communication and energy use. In federated settings, clients may train local models with different approximation policies before secure aggregation [3], [19]. This heterogeneity complicates aggregation because the server must reconcile models that are not functionally identical. Without careful coordination, adaptive approximation can amplify inequities between clients with different computational resources.

Hardware acceleration is a critical enabler. Specialized platforms for secure computation can improve throughput for both linear and nonlinear operations, but they also shape which approximations are practical [15]. A GPU-accelerated secure computation stack may favor large batches and regular arithmetic, while a trusted execution environment may favor low communication and predictable memory access. Compiler infrastructure must therefore map approximation choices to hardware capabilities and protocol constraints [10], [12]. This is a systems integration problem rather than a pure machine learning problem. The most successful deployments are likely to be those that co-design model architecture, cryptographic protocol, compiler, and hardware rather than treating them as independent layers.

Sustainability is an often-overlooked dimension. Privacy-preserving computation is computationally intensive, and the energy cost of cryptographic operations can be substantial. Adaptive approximation can reduce energy use by lowering the number of expensive nonlinear operations, but it can also increase overhead if the system constantly switches between approximation modes or maintains multiple redundant paths. The net sustainability benefit depends on the workload, the hardware, and the frequency of adaptation. A well-designed system might amortize adaptation decisions across many inferences, while a poorly designed system might consume more energy through monitoring and fallback than it saves through approximation. Therefore, evaluation should include energy per inference, communication cost, and hardware utilization alongside accuracy. These metrics connect

technical design to organizational sustainability commitments and regulatory expectations around efficient use of computational resources.

Cross-domain comparisons clarify these trade-offs. In healthcare, privacy-preserving inference may need strong auditability and high fidelity for rare conditions, making aggressive approximation risky. In finance, fraud detection may benefit from fast approximation in high-volume screening but require exact or near-exact behavior for adverse action decisions. In public-sector applications, transparency and non-discrimination obligations may require that approximation policies be documented and contestable. In consumer applications, latency and energy may dominate, allowing more aggressive approximation if privacy remains intact. Adaptive approximation is therefore not a universal good. It is a contextual capability whose value depends on the domain's tolerance for approximation error, its governance capacity, and its ability to monitor outcomes over time.

5. Robustness, Fairness, and Governance

Robustness and fairness are central to the governance of adaptive approximation. Approximation error is not uniformly distributed across inputs. It may be larger in regions of the input space with sharp decision boundaries, rare classes, or underrepresented subgroups. If those regions correspond to protected or vulnerable populations, then approximation can produce discriminatory outcomes even when the underlying model was fair before approximation. This risk is compounded when approximation policies adapt to operational conditions, because different users may receive different effective models. Fairness auditing must therefore move beyond static model evaluation and consider the approximation policy as part of the deployed system. It should examine whether approximation error correlates with demographic or geographic variables, whether fallback mechanisms are equally available, and whether monitoring data are representative.

Security and privacy risks also require governance. Differential privacy provides a rigorous framework for quantifying privacy loss, but composition across adaptive queries and approximation modes must be managed carefully [2], [18]. If the system changes approximation degree based on aggregate statistics, those statistics may themselves leak information. Secure aggregation can protect individual updates, but it does not automatically protect against inference from repeated model releases [3], [19]. Adversarial actors may probe the system to learn decision boundaries, infer membership, or reconstruct training data [20]. Adaptive approximation may either mitigate or exacerbate these risks depending on whether it introduces additional randomness or creates detectable patterns. A governance framework should require threat modeling, red-team evaluation, and documented assumptions about the adversary's capabilities.

Policy implications follow from these technical realities. Regulators may need to treat approximation as a material design choice that affects model behavior, similar to quantization or pruning. Transparency obligations might require disclosure of approximation categories, error bounds, and fallback conditions, at least to auditors if not to end users. Liability regimes must address harms caused by approximation-induced errors, especially when the system is marketed as privacy-preserving and therefore trustworthy. Procurement standards could require vendors to provide evidence that adaptive approximation does not disproportionately harm protected groups. Because privacy-preserving machine learning often involves multiple organizations, contracts and data-sharing agreements should specify who is responsible for monitoring approximation drift and who has authority to adjust or suspend the system. In this sense, adaptive approximation is as much an institutional arrangement as a technical one.

6. Evaluation and Future Directions

Evaluating adaptive nonlinear approximation requires a multidimensional framework. Accuracy alone is insufficient because a system may achieve high average accuracy while degrading on critical subgroups or under distribution shift. Latency, throughput, communication rounds, bandwidth, energy consumption, and hardware utilization are equally important for system viability. Privacy guarantees must be evaluated against the chosen threat model, including honest-but-curious servers, malicious clients, and colluding parties. Fairness metrics should be disaggregated and tied to domain-specific harms. Auditability requires versioned approximation policies, reproducible evaluation pipelines, and clear documentation of fallback behavior. These dimensions interact: a lower-degree approximation may reduce latency but increase fairness risk; a higher-degree approximation may improve fidelity but consume more energy. Multi-objective evaluation is therefore necessary to avoid optimizing one dimension at the expense of another.

Future research should focus on several system-level directions. First, adaptive approximation needs formal methods for bounding error under distribution shift, not only on held-out test sets. Second, compilers and runtimes should expose approximation budgets as first-class parameters that can be audited and constrained by policy. Third, hardware accelerators should support efficient switching between exact and approximate nonlinear paths without leaking sensitive information. Fourth, federated and multi-party systems need protocols for reconciling heterogeneous approximation policies across clients and organizations. Fifth, governance frameworks should be developed in parallel with technical mechanisms, so that transparency, redress, and accountability are not afterthoughts. These directions connect to broader efforts in trustworthy machine learning, secure computation, and responsible infrastructure design [17], [20]. They also require interdisciplinary collaboration among cryptographers, systems researchers, legal scholars, and domain experts.

Cross-domain learning is especially valuable. Aviation safety, medical devices, and financial auditing have long histories of managing approximation, uncertainty, and fallback in high-stakes systems. Their practices, such as redundancy, conservative operating envelopes, and incident reporting, could inform privacy-preserving machine learning. Conversely, privacy-preserving machine learning introduces new constraints because the system may not be able to inspect raw data to diagnose failures. This creates a need for privacy-preserving observability: techniques that allow operators to detect drift, bias, and anomalies without violating privacy guarantees. Adaptive approximation could itself be used to support such observability, for example by allocating computational resources differently across monitoring tasks. The broader lesson is that privacy and performance are not the only objectives; robustness, fairness, and legitimacy are co-equal requirements in socio-technical infrastructure.

7. Conclusion

Adaptive nonlinear approximation offers a promising path toward faster privacy-preserving machine learning by reducing the cryptographic cost of activation functions and by coordinating model behavior with runtime conditions. Its benefits are real: lower latency, reduced communication, better hardware utilization, and potentially lower energy consumption. Yet those benefits are contingent on architectural discipline and governance. Approximation must be visible to compilers, bounded by policy, monitored for drift, and evaluated for fairness and robustness. It must not introduce data-dependent control flow that weakens privacy guarantees, nor should it create hidden variation in model behavior across

users or regions. When integrated into a layered infrastructure of secure computation, differential privacy, federated learning, and hardware acceleration, adaptive approximation can become a practical control surface for balancing privacy, performance, and accountability. When treated as a shortcut, it risks producing systems that are fast but opaque, efficient but inequitable, and private in name while fragile in practice. The next generation of privacy-preserving machine learning systems should therefore treat adaptive nonlinear approximation as a governed infrastructure capability rather than an isolated optimization, embedding it within evaluation pipelines, audit trails, and policy frameworks that make its trade-offs explicit and contestable.

References

1. Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1310–1321). Association for Computing Machinery.
2. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 308–318). Association for Computing Machinery.
3. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 1175–1191). Association for Computing Machinery.
4. Mohassel, P., & Zhang, Y. (2017). SecureML: A system for scalable privacy-preserving machine learning. In 2017 IEEE Symposium on Security and Privacy (pp. 19–38). IEEE.
5. Riazi, M. S., Weinert, C., Tkachenko, O., Songhori, E. M., Schneider, T., & Koushanfar, F. (2018). Chameleon: A hybrid secure computation framework for machine learning applications. In Proceedings of the 2018 on Asia Conference on Computer and Communications Security (pp. 707–721). Association for Computing Machinery.
6. Juvekar, C., Vaikuntanathan, V., & Chandrakasan, A. (2018). GAZELLE: A low latency framework for secure neural network inference. In 27th USENIX Security Symposium (pp. 1651–1669). USENIX Association.
7. Liu, J., Juuti, M., Lu, Y., & Asokan, N. (2017). Oblivious neural network predictions via MiniONN transformations. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 619–631). Association for Computing Machinery.
8. Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., & Wernsing, J. (2016). CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy. In Proceedings of the 33rd International Conference on Machine Learning (pp. 201–210). PMLR.
9. Bourse, F., Minelli, M., Minihold, M., & Paillier, P. (2018). Fast homomorphic evaluation of deep discretized neural networks. In Advances in Cryptology – CRYPTO 2018 (pp. 483–512). Springer.
10. Dathathri, R., Saarikivi, O., Chen, H., Laine, K., Lauter, K., Maleki, S., Milder, P., & Pouchet, L. N. (2019). CHET: An optimizing compiler for fully homomorphic neural-

- network inference. In Proceedings of the 2019 USENIX Annual Technical Conference (pp. 757–772). USENIX Association.
11. Lou, Q., & Jiang, L. (2019). SHE: A fast and accurate deep neural network for encrypted data. In Advances in Neural Information Processing Systems 32 (pp. 10035–10043). Curran Associates.
 12. Mishra, P., Lehmkuhl, R., Srinivasan, A., Zheng, W., & Popa, R. A. (2020). Delphi: A cryptographic inference service for neural networks. In 29th USENIX Security Symposium (pp. 2505–2522). USENIX Association.
 13. Knott, B., Venkataraman, S., Hannun, A., Sengupta, S., & Ma, J. (2021). CrypTen: Secure multi-party computation meets machine learning. In Advances in Neural Information Processing Systems 34 (pp. 4961–4973). Curran Associates.
 14. Rathee, D., Rathee, M., Kumar, N., Chandran, N., Gupta, D., Rastogi, A., & Sharma, R. (2020). CrypTFlow2: Practical 2-party secure inference. In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (pp. 325–342). Association for Computing Machinery.
 15. Watson, J. L., Wagh, S., & Popa, R. A. (2022). Piranha: A GPU platform for secure computation. In 31st USENIX Security Symposium (pp. 827–844). USENIX Association.
 16. Peng, H., Zhou, S., Luo, Y., Xu, N., Duan, S., Ran, R., ... & Ding, C. (2023). Rrnet: Towards relu-reduced neural network for two-party computation based private inference. arXiv preprint arXiv:2302.02292.
 17. Wagh, S., Gupta, D., & Chandran, N. (2019). SecureNN: 3-party secure computation for neural network training. Proceedings on Privacy Enhancing Technologies, 2019(3), 26–49. <https://doi.org/10.2478/popets-2019-0035>
 18. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407. <https://doi.org/10.1561/04000000042>
 19. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210. <https://doi.org/10.1561/22000000083>
 20. Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). SoK: Security and privacy in machine learning. In 2018 IEEE European Symposium on Security and Privacy (pp. 399–414). IEEE.