

Signature-Based Backdoors in Variational Quantum Models

Eleanor Kedem, Ryan Sweke, Francesco Petruccione

Variational Quantum Models (VQMs) represent a promising tool in the emerging field of quantum machine learning [CAB+21], and can potentially be used for a large variety of sectors and purposes. However, for the foreseeable future, only a select number of entities and organizations have access to machines that can run them. Therefore, training and inference of such models need to be delegated, outsourced, or externally influenced by those organizations, which may have ulterior motives than those of the original model owner. As such, understanding this vulnerability, particularly in adversarial settings, is of utmost importance [LDD20; SAP+21]. This raises a natural cryptographic question: can hidden malicious functionality be embedded into such models in a way that remains attacker-controlled yet difficult to recognize from the expected circuit structure?

In classical machine learning, recent work has shown the existence of undetectable backdoors using cryptographically structured triggers [GKV+22]. We investigate the corresponding question for parameterized quantum classifiers. Our approach is based on digital-signature verification [GPV08; CHK+12]. The core idea is to embed a verifier-like mechanism into a parameterized quantum classifier so that a signed input activates a hidden malicious behavior, while unsigned inputs continue through the ordinary classification pathway. Signature-based triggers are appealing because they are rare, secret-controlled, and non-replicable, making them structurally different from standard backdoor triggers. This digital-signature-based mechanism realizes a black-box, undetectable backdoor. We then study the setting of gray-box undetectability, namely how the verifier-like mechanism can be disguised within a given model family to avoid detection under partial circuit inspection.

We consider three construction families. The first is a *hard* verifier, which most closely follows exact GPV-style verification logic and retains explicit arithmetic structure [GPV08]. The second is a *soft* verifier, which replaces exact acceptance by a smoother approximate mechanism. The third is a *trainable* verifier, which learns trigger behavior without enforcing the full arithmetic structure of exact verification. These three constructions span a spectrum from cryptographic faithfulness to implementation flexibility. Our main message is that these constructions expose a central trade-off. The more faithfully one implements exact cryptographic verification inside a quantum classifier, the larger and more structurally visible the resulting circuit becomes. This effect is already apparent at the level of raw verifier overhead, such as excessive gate and ancilla requirements, even before additional camouflage is introduced to blend the verifier into an expected ansatz. Softer and trainable constructions offer greater flexibility and may integrate more naturally into variational circuit structure, but they move away from exact cryptographic semantics.

Our contribution is therefore to identify a gray-box threat setting relevant to VQMs, present three concrete verifier-based construction paradigms, and compare them through the combined lens of trigger functionality, concealment strategy, and circuit resource overhead. The results suggest that signature-based backdoors are a meaningful and underexplored cryptographic threat model for quantum learning systems, while also highlighting the substantial gap between cryptographic elegance and practical embedding cost.

References

- [CHK+12] David Cash, Dennis Hofheinz, Eike Kiltz, and Chris Peikert. “Bonsai trees, or how to delegate a lattice basis”. In: *Journal of cryptology* 25 (2012), pp. 601–639 (page 1).
- [CAB+21] Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. “Variational quantum algorithms”. In: *Nature Reviews Physics* 3.9 (2021), pp. 625–644 (page 1).
- [GPV08] Craig Gentry, Chris Peikert, and Vinod Vaikuntanathan. “Trapdoors for hard lattices and new cryptographic constructions”. In: *Proceedings of the fortieth annual ACM symposium on Theory of computing*. 2008, pp. 197–206 (page 1).
- [GKV+22] Shafi Goldwasser, Michael P Kim, Vinod Vaikuntanathan, and Or Zamir. “Planting undetectable backdoors in machine learning models”. In: *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2022, pp. 931–942 (page 1).
- [LDD20] Sirui Lu, Lu-Ming Duan, and Dong-Ling Deng. “Quantum adversarial machine learning”. In: *Physical Review Research* 2.3 (2020), p. 033212 (page 1).
- [SAP+21] Abdullah Ash Saki, Mahabubul Alam, Koustubh Phalak, Aakarshitha Suresh, Rasit Onur Topaloglu, and Swaroop Ghosh. “A survey and tutorial on security and resilience of quantum computing”. In: *2021 IEEE European Test Symposium (ETS)*. IEEE. 2021, pp. 1–10 (page 1).