We thank the reviewers for their useful feedback. We will carefully address all raised questions in the camera-ready version. Additionally, we will incorporate detailed new material based on the clarifications outlined below. This will go along with a treatment of the reviewers' smaller improvement suggestions on notation and figure design. Further, we intend to increase the number of facilitating pointers to the appendix, where detailed theorem formulations and supplementary information on numerical quantities such as time interval parameters can be found.

(Non-) Necessity of Affine-Linear Coefficients (reviewer #3): In our numerical examples, we focus on affine-linear coefficient functions because they are important in practical applications, computationally fast to evaluate, and easy to parametrize. As rightfully pointed out by one reviewer, however, the presented method is indeed not restricted to the case of affine-linear coefficients and can as well be used in a substantially more general setting. In particular, note that affine-linear coefficients are not assumed when the rigorous validity of the core learning problem is established in Theorem 1.

Robustness w.r.t. Approximate Data Generation via Euler-Maruyama (reviewers #2, #5) : Let S_Λ^N be the Euler-Maruyama approximation of the solution to the parametric SDE $S_\Lambda = S_{\Gamma,X,\mathcal{T}}$ (as defined in (10) in the paper) with $N \in \mathbb{N}$ equidistant steps, given by

$$S_\Lambda^0 = X \quad \text{and} \quad S_\Lambda^{n+1} = S_\Lambda^n + \mu_\Gamma(S_\Lambda^n)\frac{\mathcal{T}}{N} + \sigma_\Gamma(S_\Lambda^n)\big(B_{\frac{(n+1)\mathcal{T}}{N}} - B_{\frac{n\mathcal{T}}{N}}\big), \quad n = 0, \dots, N-1.$$

We managed to prove a theorem which shows that using our method with data obtained via the Euler-Maruyama scheme must result in the expected approximation of the parametric PDE solution map $\bar{u}$.

Theorem. *Assume Assumptions* 1 *and* 2 *from the appendix and further assume that φ_γ has an at most polynomially growing derivative. Let $\bar{u}$ be the parametric solution map of the Kolmogorov PDE. Then there exists a constant C depending only on v, w, T, and the growth rates and (local) Lipschitz constants of σ_γ, μ_γ, and φ_γ such that the solution to the approximated learning problem $\bar{u}^N = \operatorname{argmin}_f \mathbb{E}\Big[\big(f(\Lambda) - \varphi_\Gamma(S_\Lambda^N)\big)^2\Big]$ satisfies that*

$$\max_{(\gamma,x,t)\in D\times[v,w]^d\times[0,T]} |\bar{u}^N - \bar{u}| \leq \frac{C}{\sqrt{N}}.$$

Proof (Sketch). Extending results on the Euler-Maruyama scheme (see, e.g., [Kloeden and Platen, 1992, Theorem 10.2.2]) one can prove that also in the parametric SDE case for $p \geq 2$ the p-th moments of S_Λ and S_Λ^N are bounded and that it holds that $\big(\mathbb{E}\big[\|S_\Lambda - S_\Lambda^N\|_{\mathbb{R}^d}^p\big]\big)^{1/p} \leq \frac{C}{\sqrt{N}}$. The local Lipschitz property of φ_γ then proves the claim. □

This result provides a theoretical guarantee for the robustness of our machine learning method; it can easily be used to prove that our generalization results are not compromised by using data simulated by the Euler-Maruyama scheme.

The factor $1/V$ (reviewer #2): The factor $\frac{1}{V} = \frac{1}{\operatorname{vol}(D\times[v,w]^d\times[0,T])}$ naturally appears when transforming L^∞- to L^p-results and can be omitted by viewing the error in the space $L^p(\mathbb{P}_\Lambda)$ (where $\mathbb{P}_\Lambda$ is the uniform probability measure on V) via

$$\|\cdot\|_{L^p(\mathbb{P}_\Lambda)}^p = \frac{1}{V}\|\cdot\|_{L^p(D\times[v,w]^d\times[0,T])}^p \leq \|\cdot\|_{L^\infty(D\times[v,w]^d\times[0,T])}^p.$$

All results within the established standard setting of statistical learning theory (including our generalization bound) give rise to L^p-bounds w.r.t. a given probability measure on the input domain. In fact, note that our setting easily allows us to choose arbitrary probability measures $\mathbb{P}$ on $D \times [v,w]^d \times [0,T]$ and prove analogous results w.r.t. the $L^2(\mathbb{P})$ norm. Thus, following the conventional terminology used in statistical learning, we can indeed claim that the presented bound overcomes the curse of dimensionality. To underline this we mention Barron [1993] as one of many examples of a classical and well-known approximation result where the terminology "avoiding/overcoming the curse of dimensionality" is used in strictly the same context as in our paper. We aim to further clarify this in the camera-ready version.

Overcoming the Curse of Dimensionality (reviewer #5): Based on the feedback of reviewer #5 we will further clarify in the camera-ready version that the curse of dimensionality is overcome with respect to the neural network size as well as the sample size. We emphasize that our empirical results strongly suggest that also the ERM algorithm does not suffer from the curse of dimensionality but proving this is out of scope of this paper.

References

A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*, volume 23 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1992. ISBN 3-540-54062-8.