We thank all the reviewers for their time and effort to evaluate our paper.

R1, R3, R4: **Assumptions:** We will discuss the assumptions in the context of the SDE in Eq. 4, which is our main focus. **H1:** The boundedness assumption requires the solutions to the SDE to be 'non-explosive' in the sense that for every S, $\|\mathrm{W}_t^{(S)}\| < \infty$ almost surely for all $t \in [0, T]$. As we shortly mentioned in L.269, H1 directly holds under compact $\mathcal{Z}$ and the conditions of [XZ20, Lem7.1], which requires standard regularity conditions on f, Σ_1, Σ_2, $\boldsymbol{\alpha}$. **H2.** only concerns ℓ. **H3.** follows from Prop.1. and it is not required in Thm2. **H4.** It is easier to analyze the Hausdorff dimension (HD) of the range of a stochastic process, it has henceforth become the most frequently studied notion of fractal dimension for Lévy-type processes. Yet, the Minkowski dimension (MD) is more relevant to the generalization error, hence we used MD as a tool in our proofs and used H4 to ensure it is equal to HD. However, we underline that for many fractal-like sets, HD is **already** equal to MD (see the discussions in [Mat99, Page 80-81]), which include several Feller processes like Brownian motion or α-stable processes (see [Fal04,Chapter16]). Hence, H4 is not an unrealistic assumption for our purposes. On the other hand, in Thm.S3, we already proved a bound which does not require H4, but requires $\mathcal{Z}$ to be finite. **H5:** This assumption is common in statistics (see [Bra83]) but hard to verify in practice. It essentially quantifies the dependence between S and the trajectory $\{\mathrm{W}_t^{(S)}\}_{t\in[0,1]}$, through the constant $M > 0$: small M indicates that the training error $\hat{\mathcal{R}}$ would not differ much if the training data S is slightly altered. This is similar to the mutual information used recently in [XR17,arXiv:1806.03803] and to the concept of stability. We agree that H4-5 seem technical, and we will add a more detailed discussion.

R1: **"Prop. 1, ...not...interpretable"** We agree that Prop.1 can be difficult to grasp at a first sight; hence, we provided a paragraph (L.245) to discuss its semantics in our context, and also we believe that it should be considered as a contribution as we make this surprising connection with probability theory literature.

R2: **Bounded loss assumption:** In the proof, we require the concentration of empirical risk $\frac{1}{n}\sum_i \ell(w, z_i)$ which, for a fixed w, is a sum of iid r.v.'s $\ell(w, z_i)$. In the current version, we assumed $\ell(w, z_i) \leq B$, and used Hoeffding; however, the same can be achieved by simply assuming $\exists K$, $\forall p$, $\mathbb{E}[\ell(w, z)^p]^{1/p} \leq K\sqrt{p}$, and using sub-Gaussian concentration: Thms.1-2 still hold with K in place of B. **Learning-rate schedule:** We are interested in the relationship between intrinsic dimensionality and generalization of the deep networks. Relationship between dimensionality and learning rate is an interesting topic which is out-of-scope for our paper, but a great candidate for future work. **Illustration:** We will add new figures illustrating the Hausdorff dimension. **Underlying mechanism:** At this stage, we can only speculate that if the process exhibit heavy-tails near a local minimum, it stays near that minimum even when large jumps occur due to the heavy-tails (otherwise it would go near a different minimum). In a non-convex setting, this would happen when the local minimum has a low curvature, connecting our framework to the notion of flatness.

R3: Understanding the mechanism of BN is an active research topic, and the relationship between generalization and BN is not fully understood. Our main focus is the relationship between generalization and intrinsic dimensionality. Relationship between BN and dimensionality is an interesting future direction which would complement our work.

R4: We believe we addressed all the raised issues in what follows. We hope the reviewer could reconsider their overall score. **Comparison to std. Gaussian SDEs:** **(1)** The current bounds for Brownian-SDEs (based on isotropic Gaussian noise) often contain an implicit term, e.g. KL div. (arXiv:1911.02151), which are hard to control and grow with d (see [NBMS17]), or the sum of step-sizes [MWZZ17] which may result in vacuous bounds (see Fig. 1 in arXiv:1911.02151). In this sense, our bounds also improve upon existing Brownian-SDE bounds: our result is the first generalization bound which is independent of d for **both** Brownian and heavy-tailed SDEs. Also our bounds are uniform in the **full path** of the algorithm; whereas [MWZZ17], arXiv:1911.02151 only control the error at the endpoint, which has to be determined in advance. **(2)** When the process exhibits heavy-tails, our bounds show that this intrinsic dimension gets even smaller, can become **arbitrarily** close to zero even when d is large, see Fig1.a. In this sense, we disagree with the reviewer on the comment "improvement of a constant factor": in our bounds d_{H} gets very small for networks with large d, hence the improvement is clearly not a constant factor. **Experiments:** We already conducted many experiments on several neural architectures of different sizes (fully connected, Alexnet, VGG), which all conformed with our theory. However, we decided to report our results only on VGG for the sake of conciseness, as they contain millions of parameters, which cannot be explained by existing generalization bounds. Experiments on other architectures produced similar figures. We will include others in the revised version. On the other hand, The fact that our theory is valid **even** on VGG networks should be considered as one of our strengths. Moreover, we thank the reviewer for suggesting to experiment the extreme cases. We agree it would be an interesting addition, we will run the experiments on extreme cases of pure generalization and pure memorization and add in the final version. **Countability:** Countability of $\mathcal{Z}$ is only assumed in Thm1 (Thm2 does not require it), and it has been considered in various studies, e.g. [BE02]. **Clarity:** We will describe in more detail why local decomposability reflects in the global behavior. **Add. Comments:** **(1)** Our theory is agnostic to the way the minibatch is drawn. **(2)** Contrary to works that exploit the implicit regularization of zero initialization, our bounds are valid for any fixed initial point, see [arXiv:1707.06618] for similar arguments. **(3)** We will revise and state they are close to SotA. **(4)** We ran SGD for 100 epochs (see L.340). Note that our bounds do not depend on T. **(5)** Thms 1-2 apply to gradient descent as well; however, it is not clear how $\dim_{\mathrm{H}} \mathcal{W}_S$ can be measured in that case.