R1) Comments on the main proof strategy. We thank the reviewer for the insightful comments on the proof. We agree that the sketched argument based on the Bernstein-von Mises theorem is simpler, yet it relies - at least in its vanilla version- on the assumption that the target function is representable in a fixed parametric model. In order to avoid this assumption - which is not controllable a priori, we rely on limit results in the over-parameterised regime, in which the size of the model is larger than the input data. In this scenario, both the input data and the model size are taken to the limit, and so the limiting model is non-parameteric and guaranteed to have zero loss on the data manifold. It is not obvious to us if and under which conditions the BvM theorem generalises to this setting, but this is indeed a promising direction to investigate further. We still believe, though, that our proof provides some interesting geometric insights on adversarial attacks. We will clarify better in the main text notions like "overparamaterise" or "fully trained".

R2, R3, R4) Results reproducibility, convolutional layers, and number of different architectures. We remark that the hyper-parameters used for training are reported in the Supplementary (Section 3). The source code will be made available after the review phase. We remark that our experiments comprises both fully connected (up to 5 layers) and convolutional layers (see Table 2). Noted by R2, Table 4 in the supplementary was accidentally truncated by one value in several rows, we will update the parameters accordingly (in total we trained 1728 HMC BNNs).

R2) Using deterministic networks and deep ensembles as baseline models. We agree with the reviewer and in Table 1 we consider the same NN used to perform the experiments in Section 5.2 (hyper-parameters are reported in Table 3 in the Supplementary) and run a comparison with both deterministic NNs and deep ensembles (Lakshminarayanan, 2017). We further evaluate the robustness of deep ensembles on a subset of the NNs employed in Section 5.3. We find that deep ensemble NNs have a robustness similar to that of deterministic NNs suggesting that simply averaging predictions for different weight initialization and mini-batching is not enough to achieve a robust model. We will add these results in the main text.

R2) Priors. In our proof setting, an uniformative prior is one that gives equal density to all the possible weights realisations. This can be seen as the limit of a Gaussian with infinite variance. In practice, the relative importance of the prior w.r.t. the likelihood diminishes as more data are used for training, and the posterior distribution gets pulled apart from the prior. In the experiments reported in the paper, we have found an $\mathcal{N}(0, 1)$ prior to work well, as the posterior variance gets to around 0.05 after training. We performed evaluations with a higher prior variance (up to 10) and noticed a similar behaviour of the loss gradients.

Model	Test accuracy	FGSM accuracy	PGD accuracy
Deterministic NN	97.69	21.19	1.45
Ensemble NN	99.4	20.6	0.3
Bayesian NN	96.1	90.0	89.8

Table 1: FGSM and PGD attacks on the network employed in Section 5.2. We compare a deterministic NN, a deep ensamble NN (of size 100), and a BNN (trained with VI). Attacks are performed on 1k test points from the MNIST dataset. We observe that VI trained network achieve better robustness against PGD and FGSM.

R2) Correlation between accuracy and robustness. When a BNNs has a high number of neurons and high accuracy the conditions for Theorem 1 are approximately met. This guarantees that the network is protected against gradients attack, thanks to the cancelling effect of Theorem 1. For deterministic NNs Theorem 1 does not hold. The trade-off between robustness and accuracy in that case has been already observed and studied (Zhang et al.2019).

R3) Proof of Th 1: novelty and completeness. We would like to stress that Theorem 1 and its proof are novel. In fact, although we rely on known results for over-parametrised NNs, to the best of our knowledge, the application of these results in the context of robustness of Bayesian NNs and Lemma 2 are novel. Furthermore, we would like to clarify that for an input point and a fully trained model, Lemma 2 guarantees that there exists another model such that the NN has the same loss on the data manifold and opposite orthogonal gradients on that input point. Hence, by definition, the NN will have same likelihood on models. If they also have same prior (uninformative prior assumption) then they also have same posterior. This entails the cancellation of orthogonal gradient average.

R3) Are HMC and VI distributed according to the posterior? Unfortunately, for non-linear networks computation of the posterior is analytically intractable. Nevertheless, HMC converges to the true posterior in the limit of infinitely many Monte Carlo samples taken (we used 500 samples in our experiments). On the other side, VI is a more scalable approximate inference method, but has no convergence guarantees to the true posterior. This also explains why for MNIST, where both HMC and VI obtain $> 90\%$ accuracy, HMC tends to be more robust than VI.

R4) Why HMC performs better than VI in MNIST but not in FashionMNIST? HMC converges to the true posterior, but it is less scalable than VI. As a result, on MNIST where HMC is able to achieve $> 90\%$ accuracy, it tends to be more robust than VI. On the other hand, on FashionMNIST we were not able to train a BNN with HMC to have high accuracy, hence, we are far from the regime required by our Theorem 1 to achieve cancelling gradients.

R4) Add comparison with other training methods. Please, see R2) Using deterministic network and deep ensembles as baseline models.