We are very grateful that all four reviewers recognise the novelty and originality of the paper. Below, we first clarify the main contribution (in particular regarding Lemma 4.3), and then address the detailed points raised by the reviewers.

Main Contribution is the novel concept weight correlation (WC), which we believe is a key factor affecting the generalisation ability. To consolidate this, we inject WC into the well-received PAC-Bayesian framework to derive a *closed-form expression* of generalization gap bound with *mild assumption* on weight distribution, and then employ WC as an explicit reguraliser to enhance generalisation performance within training. More importantly, the regulariser is effective and computationally efficient in enhancing generalisation performance in practice.

Lemma 4.3 shows the positive correlation between WC and the bound (but not the generalisation itself). The effectiveness of WC in either predicting the generalisation (Section 6.1) or training (Section 6.2) is shown with experiments.

The central question raised in the reviews is the justification of the assumption that we have a **Gaussian posterior** distribution. We believe that this is a very mild assumption, which is partially justified by distributions tending to converge against Gaussian distributions. More importantly, our assumption significantly relaxes the assumptions used in prior works: an i.i.d. assumption made in [Dziugaite and Roy, 2017, Neyshabur et al. (2017); Jiang et al. (2020)]. Different to assuming a general Gaussian distribution, i.i.d. is unrealistic, such that we can lift an unrealistic assumption.

This also addresses the **R3: novelty of theoretical claim on WC correlation** concern: we have lifted the—unrealistic—i.i.d. assumption from [Dziugaite and Roy, 2017], landing practical relevance to their findings.

This puts us in a sweet spot between the techniques that make unrealistic assumptions about the posterior distribution (usually i.i.d.), and approaches that make no assumptions, but only allow for an a posteriori estimation. (Moreover, such estimations are hard to compute and inaccurate for high dimensional data.) As a result, we have gained the capacity to develop a regulariser, which is both meaningful and easy to compute.

R1, R5: comparison with the state-of-the-art While mainly focused on [Chatterji et al., ICLR'20], our results also shed light on the fantastic measures paper [Jiang et al, ICLR'20]. In particular, in [Jiang, et al, ICLR'20], it is concluded that "Sharpness-based measures such as sharpness PAC-Bayesian bounds ... perform better overall and seem to be promising candidates for further research". Our results advance this and show that the PAC-Bayesian bounds can be further improved with the weight correlation. Therefore, we believe our paper has advanced the state-of-the-art.

R1, R2, R5: Figure 1 Figure 1 (in Introduction) is an illustrative example to show the positive correlation between WC and generalisability. It is *not* to suggest a general trend for WC (the evolution of WC can be fluctuating) or a conclusive result for the positive correlation. The latter is obtained by the theoretical and empirical results in the following sections.

R2, R5: theoretical support of Lemma 4.3 to regulariser We agree PAC-Bayes can only provide partial theoretical support to a regulariser. Lemma 4.3 shows that WC is positively correlated with the generalisation bound. This result suggests that it *may* be effective to consider WC as a regulariser. Then, extensive experiments are conducted to validate.

R3: discussions/justification on the particular form of the posterior distributions We guess you are referring to Def. 4.2. Given the weight matrix $w_\ell \in \mathbb{R}^{N_{\ell-1} \times N_\ell}$ with i-th column $w_{\ell i}$ as a random vector, the posterior covariance matrix $\Sigma_{Q_{w_\ell}}$ is defined in a standard way as $\Sigma_{Q_{w_\ell}} = \mathbb{E}[\mathrm{vec}(w_\ell)\mathrm{vec}(w_\ell)^T] \in \mathbb{R}^{N_\ell N_{\ell-1} \times N_\ell N_{\ell-1}}$, where $\mathrm{vec}(\cdot)$ is the vectorisation of a matrix. The (i,j)-th block is $\left[\Sigma_{Q_{w_\ell}}\right]_{i,j} = \mathbb{E}[w_{\ell i} w_{\ell j}^T] \in \mathbb{R}^{N_{\ell-1} \times N_{\ell-1}}$. For computational simplicity, we use the arithmetic mean instead of the expected value, so that the weight correlation $\rho(w_\ell)$ can be used to represent $[\Sigma_{Q_{w_\ell}}]_{i,j} = \rho(w_\ell)\sigma_\ell^2 I_{N_{\ell-1}}$. Therefore we have $\Sigma_{Q_{w_\ell}} = \Sigma_{\rho(w_\ell)} \otimes \sigma_\ell^2 I_{N_{\ell-1}}$, where $\otimes$ is the Kronecker product.

R3: modest improvement in Section 6.2 While the improvement might look modest in individual cases, it is persistent across the experiments.

R3, R5: more experiments We conducted additional experiments on Caltech-256 dataset for this rebuttal. It is unrealistic to consider ImageNet, since a single training may take days (or months). The results on the Caltech-256 dataset are also promising, and similar to MNIST and CIFAR10. For complexity measure (like Table 2), we have Kendall's τ at 0.33, as opposed to others at 0.28, 0.28, 0.17, 0.22. For the comparison of models with and without WCD (like Table 3), those models with WCD achieve about 1% improvement on top-5 error over models without WCD.

R5: WCD and posterior We can confirm that our WCD is obtained from the expression of the posterior of PAC Bayes (Lemma 4.3), and therefore there is no discrepancy between posterior of Lemma 4.3 and WCD-based posterior.

R5: technical details of Lemma 4.3 There may be some misunderstanding to our technical details. For example, the prior is a diagonal matrix $\mathrm{diag}(\sigma_\ell^2)$ and the PCA Bayesian is conducted by summarising the results across layers.

R5: termination condition of training We use the same termination condition across all our experiments, so it is fair to all methods. The reason for us to consider the best loss across the last few training epoch is to make sure that our results are not affected by random factor – we find that there may still be uncertainty in the last few epochs.