We thank the reviewers for their insights, which we have incorporated into our work. We were pleased that reviewers highlighted the novelty of our theory (**R1**,R2,R3,R4), its clarity (**R1**,R4), its correctness (R2,R3,R4), and its relevance and potential impact for the research community (**R1**,R2,R3,R4).

To bolster our claims, we have run additional experiments that support our approximations, which ignored $O(\kappa)$ in Theorem 1 for 0.1 variance GNIs. In Fig 1 we show the estimation error of the true noisy loss this approximation induces for 12-layer MLP sigmoid networks for which the interaction effects described by **R1** would be most prominent. Clearly the error is small for the variance we used, which is highlighted in red. This experiment strengthens our arguments and we thank **R1** and R4 for suggesting it.

To further support this result, we ran experiments which demonstrate that the *Exp Reg* terms (first term of Theorem 1 used in our approximations) dominates $O(\boldsymbol{\kappa})$ in value for a range of small and large σ^2 (see Fig 2). The findings of Figs 1, 2 also hold for ReLU and ELU activations and we will include these results in the update. Combined, these new results show that our approximations hold even for large variance GNIs, making our work even more impactful.

Reviewer 1: We thank the reviewer for their insights and as seen in the results above, they have already stimulated new experiments which strengthen our arguments. To further alleviate your concerns, we have run experiments which demonstrate that the regularisation induced by *Exp Reg* (first term of Theorem 1) matches that of GNIs and is not just a byproduct of any generic regularisation method. See Fig 3 for a demonstration of this. We also thank you for pointing out the typo in Proposition 1.

Reviewer 2: The restriction of Plancherel's Theorem to a compact subset is straightforward, and we will improve the clarity surrounding this, as suggested. We will also make sure to include the justification for this in the Appendix. We will also clarify the connection between Sec.4, 5. **Note**: 'Elements of $\|f\|^2$' below Eq (7) should read as the 'constituent terms of $\|f\|^2$'. Take the Sobolev norm for instance, we could penalise the L_p norm of the function and the L_p norm of its derivative with different weightings. $\mathbf{x}$ in Proposition 1 is elements of the minibatch; we'll clarify this.

Reviewer 3: Our theory encompasses the case of injecting noise solely on data, as this corresponds to layer 0 in our formulation. This translates into a lessened penalisation of the network Jacobians. In particular, we ran experiments following this review which showed that the dampening of higher-order frequencies in the Fourier domain is less when injecting noise only on data. We will include these results in our update. See also **R1** and Fig 3 for a comparison of GNIs to L2 regularisation. **Note**: We thank you for your detailed review of Figs 3 and 4. We will improve the figures given your feedback.

Reviewer 4: We thank the reviewer for providing such detailed comments. As seen above, we have run experiments testing the limits of our assumptions (see also **R1**). As our theory is in early stages we wanted to test its efficacy primarily on simpler models. We are working on scaling it to larger models, eg VGG13: currently the calculation of each layer's Jacobian is memory hungry, as is $\mathrm{Tr}(\mathbf{H})$ in Fig 1 [paper]. We thank you for pointing us to works on the brittleness of neural networks and on the tradeoffs between robustness to noise vs. other types of data corruption. We will include these in our update along with with more detailed commands in the README to replicate experiments. We will also clarify the connection between Secs 4 & 5. **Note**: A "PSD" scalar is incorrect as you have pointed out, it should read that the "constituent terms" of the scalar are PSD.

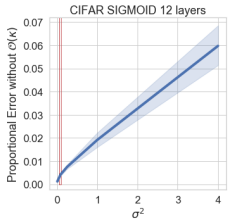

Figure 1: Proportional estimation error (maximum of 1.0) of $\mathbb{E}[\tilde{\mathcal{L}}_{\mathrm{SGD}}(\mathcal{B};\boldsymbol{\theta},\boldsymbol{\epsilon})]$ when ignoring $O(\boldsymbol{\kappa})$ from Theorem 1. We plot this for 12-layer MLPs trained on CIFAR10 with sigmoid activations, non-linearities and network depth which heavily test our approximations. GNIs are applied to each layer bar the final layer. Even under these conditions, 0.1 variance GNIs (highlighted in red) are well approximated by our estimations. Shading is the standard deviation over 250 points.

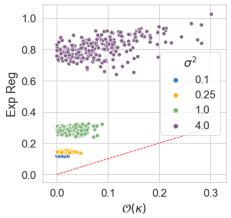

Figure 2: Here we plot *Exp Reg*, the first term of Theorem 1, against $O(\boldsymbol{\kappa})$ for $\sigma^2 \in [0.1, 0.25, 1.0, 4.0]$. Networks are the same as in Fig 1. For reference we plot $y = x$ in red. For all values of σ^2, *Exp Reg* lies above this line and we can claim that empirically *Exp Reg* > $O(\boldsymbol{\kappa})$.

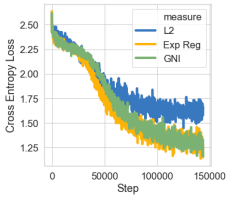

Figure 3: Training set loss for 4-layer ELU MLPs trained on SVHN. Three regularisation methods are compared: L_2 regularisation with λ =0.01 penalisation, 0.1 variance GNIs, and *Exp Reg*, the first term of Theorem 1. λ was picked such that the magnitude of the regularisation roughly matched that of GNIs/Exp Reg at the start of training. Clearly GNIs/Exp Reg have distinct training curves to L2.

You make an interesting point about Fig 1: All models were trained with a relatively low learning rate (lr) of 0.001, supporting your claim of low lr Hessian 'blowup'. In light of this we have run the baseline with lr=0.1 and found that $\mathrm{Tr}(\mathbf{H})$ *decreases* with training instead. As you suggest, there could be a lr for which we recover some of the benefits of GNIs. Exploring this connection further would be a very interesting stream of research.