We thank the reviewers for their thoughtful and constructive reviews. We appreciate that they found the paper to be 'theoretically and experimentally grounded,' and 'extremely well-written,' that it'could easily be used in practice,' and has 'practical impact in a number of real-world applications,' specifically 'where security and privacy are important.' Below we respond to the major comments; we will fix the minor ones in the final version.

Reviewer #1 • *results on more benchmarks [...] would further improve the confidence.* Agreed. As suggested, we ran experiments on the Fashion MNIST benchmark; The results (Table I) are substantially in line with those in the paper. Experiments for CIFAR-100 are in progress and we will add those and the results of Table I to the final paper. • *The MLP and CNN are a bit old models [...]* We used MLP and CNNs since they were used in studies that we compared to, e.g. Deep Ensembles (DE) (arXiv:1612.01474). Furthermore, in the submitted paper, we showed the effectiveness of our proposed methods on larger and deeper pretrained networks. In the new sets of experiments, we used Wide-ResNet-28$\times$10 (arXiv:1605.07146v4 2016) for CIFAR-10 out-of-distribution (OOD) detection experiments. • *Ref. [39] reported improved results [...] using 'mixup'* Agreed. Due to the limited time of the response period, that experiment cannot be completed now. We will add the 'mixeup training' results into the revised paper. • *uniform distribution prior or some other distribution [instead of normal for perturbation] ....?* We tested a uniform distribution with MNIST (MLP) and observed similar performance to a normal distribution on this small problem. We will run experiments on other applications, including larger ImageNet networks, and add the results to the final version of the paper. • *the performance of PEP on out-of-distribution images.* We performed experiments similar to arXiv:1902.02476 for OOD detection. We trained a WideResNet-28x10 on data from five classes of the CIFAR-10 dataset and then evaluated on the whole test set. We measured the symmetrized KL divergence (KLD) between the in-distribution and out-of-distributions samples. The results show that KLD increased from 0.47 (baseline) to 0.72 by using PEP. Temp. scaling also increased KLD to 0.71. We will add these results to the paper.

Table 1: Aditional experiments on fashion MNIST (For all metrics smaller is better).

Metric	Baseline	PEP	Temp. Scaling	MCD	Deep Ensembles
NLL	0.360 ± 0.01	0.275 ± 0.01	0.271 ± 0.01	0.218 ± 0.01	0.198 ± 0.00
Brier	0.137 ± 0.01	0.127 ± 0.01	0.126 ± 0.00	0.111 ± 0.00	0.096 ± 0.00
ECE %	5.269 ± 0.22	1.784 ± 0.54	1.098 ± 0.18	1.466 ± 0.30	0.942 ± 0.13
Classification Error	8.420 ± 0.32	8.522 ± 0.34	8.420 ± 0.32	7.692 ± 0.34	6.508 ± 0.10

Reviewer #2 • *there are many missing references to very relevant and related pieces of work.* We thank R2 for pointing out the related work of (Ritter et al. 2010), (Izmailov et al. 2018) and (Maddox et al. 2019), especially about Laplace approximations. From that point of view, PEP is perhaps the simplest possible Laplace approximation - an isotropic Gaussian with one variance parameter, though we set the parameter with simple ML/cross-validation rather than calculating curvature. The trade-off is that while performance is expected to be better with richer covariance models, there is some overhead in calculating them, and they are not practical for use with pre-trained models. We will revise the paper accordingly. • *In the discussion of PEP effect vs. overfitting [... (Goodfellow ICLR 2015) ...] may be a good paper to include in the related work.* Agreed. We will include it in the final version. • *SWA methods of Izmailov et al. + [SWAG method] Maddox et al. should be included as baselines.* We agree that addition of SWA/G results will strengthen the conclusions. We are addressing the implementation logistics between us and SWA/G and (SWAG, more recent, is in PyTorch, SWA TensorFlow implementation has bugs, we are in TensorFlow) which need additional time. We will add results of the experiments in the revision. • *No broader impact section* We will add a 'broader impact' section that will discuss the importance of reducing carbon footprint via reduced compute resources, and the importance of improved calibration, security and privacy in medical applications. • *While the theoretical analysis is all correct, much of it is also well established (... Taylor expansions as Laplace approximations ... odd moments of ... Gaussians are zero, etc).* We agree that we are using well-established methods; we think of this as an advantage. Our work shows how a simple formalism yields improvement in the calibration of pre-trained networks. It also enables us to provide an in-depth analysis of why our method can improve NLL, and under what conditions. • *[what are] "different empirical FI?" ... "first term" ... "second term"* FI is Fisher Information, 'First term' and 'Second term' refer to the preceding equation. We will clarify in the revision accordingly.

Reviewer #3 Thank you for appreciating the novelty of our approach. • *The performance [...] is still much worse than for two competing approaches.* True, but DE has additional training cost, and MCD requires model modification.

Reviewer #4 • *the evaluation measures that should matter are ECE and the classification error. However, PEP does not seem to necessarily improve these measures.* PEP is mainly aimed at improving calibration, though it can provide classification improvement for overfitted models. NLL, Brier score, and ECE are all commonly used metrics to assess calibration. PEP improved ECE of the baselines in 8 out of 9 experiments. It also consistently improved NLL and Brier score of the baselines. • *The paper doesn't mention the reason this specific approach was chosen; i.e., please given an intuitive explanation [...]* There are general arguments about why ensembles can work [ref 5]. Also, Jensen's inequality suggested to us that simple probabilistic perturbations about θ^* might be effective, depending on the curvature of the validation log likelihood function at θ^* (from training), (which might depend on overfitting). • *notation, while clear, is not always defined. please state clearly what i, j, and m are.* They are index of data item, index of Gaussian sample, number of Gaussian samples, respectively. We will revise the paper accordingly.