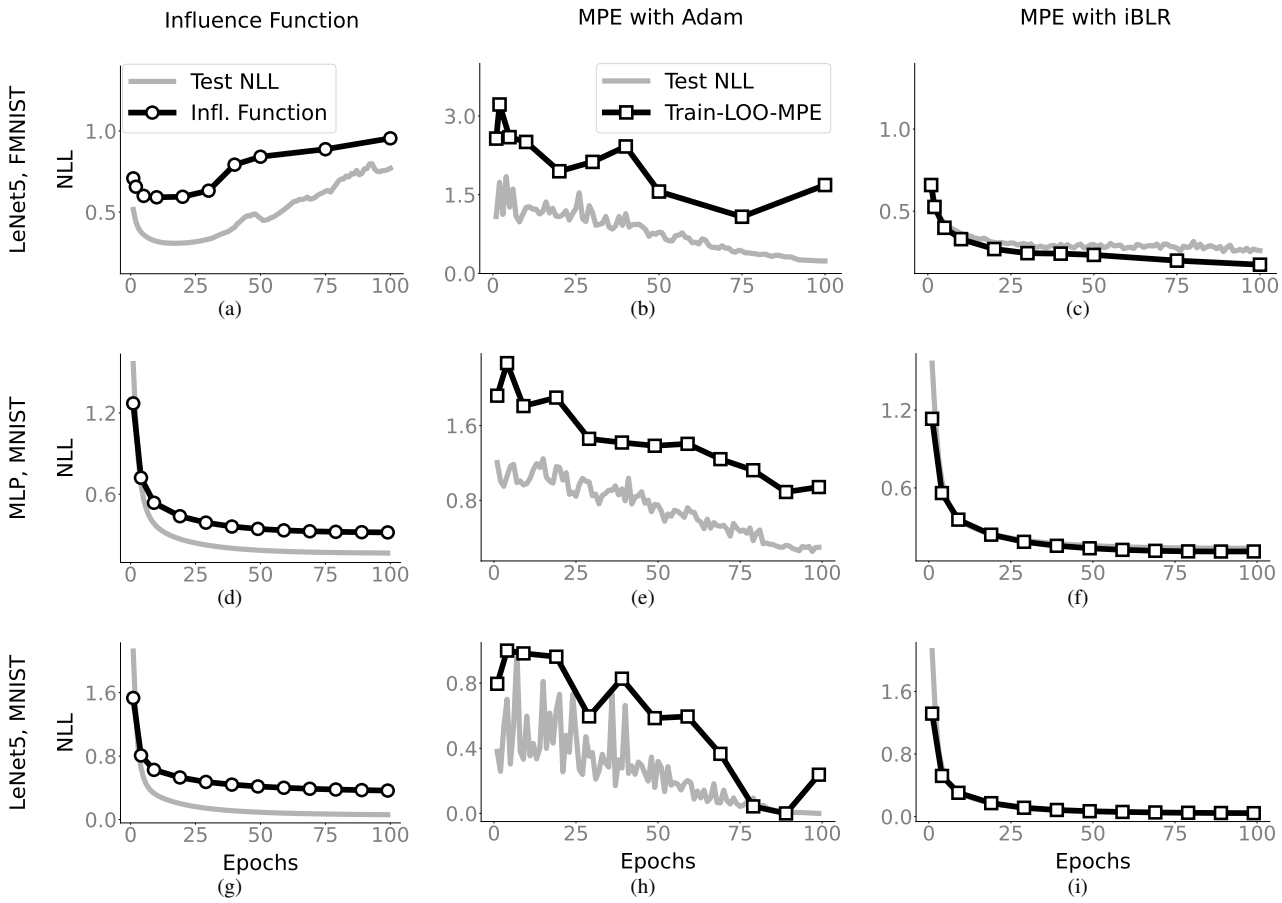
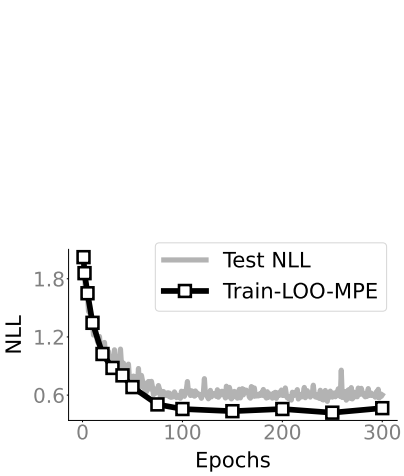
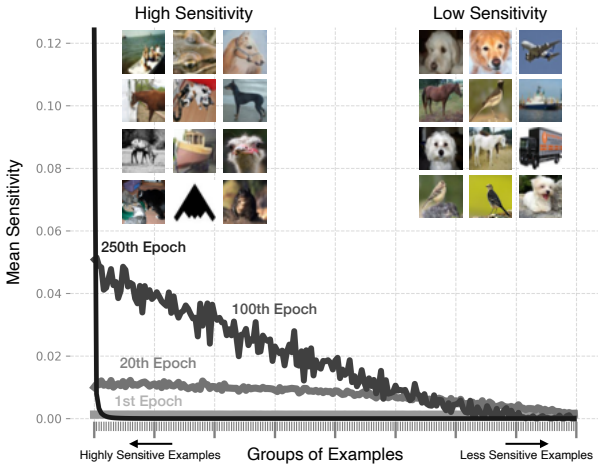


Figure I: Each row shows results on a different (model, data) pair to support our hypothesis: better uncertainty estimates result in better sensitivity, giving rise to a more faithful estimation of the test NLL by a train-LOO estimate (using Eq. 19). The third column shows the best results with closest match between true and estimated generalization. The first column uses the influence function, which is not valid during iterations. It is using the second measure in Eq. 16 with a diagonal Hessian $\nabla^2 \mathcal{L}(\boldsymbol{\theta}_t)$. In contrast, the next two columns use the new measure (the first measure in Eq. 16) with a diagonal covariance $\boldsymbol{\Sigma}_t$ estimated during training. $\boldsymbol{\Sigma}_t$ is a moving average of the past Hessians. Specifically, the third column uses a version of the BLR from [19]; the second column uses a (slightly worse) Adam-based method similar to [13]; finally, the first column uses the Hessian at an iterate of either AdamW (for Panel (a)) or SGD (for Panel (d) and (g)). As our theory predicts, the first two columns give worse estimates. In Panel (h), we normalize the y-axis, as the gap in absolute value of test NLL and LOO estimate is large.



(a) MPE with iBLR (ResNet–20, CIFAR10)



(b) MPE with iBLR (ResNet–20, CIFAR10)

Figure II: Panel (a) shows results on a large ResNet–20 (270K parameters) model on CIFAR10. We use the same setup as for the third column in Fig. I and see a similar trend for the ResNet architecture as well. In panel (b) we show the MPE estimate of the evolution of sensitivies during training with iBLR. We see that, with training, the model becomes more and more sensitive to a small fraction of data. We also show some examples of high and low sensitivies.