A Experiment Details

All experiments unless otherwise stated present the average/standard deviations of $n = 3$ runs. Each run consists of a random subset of MNIST 0/1 digits for MNIST binary classification, or random positions of sampled datapoints for synthetic data, and different samples from the GP for the Gaussian Random Field experiment. Distilled datasets are initialized as subsets of the original training data. We distill for 20000 iterations with Adam optimizer with a learning rate of 0.002 optimizing both images/data positions and labels. We use full batch gradient descent for the synthetic datasets and a maximum batch size of 2000 for the MNIST experiment. For the MNIST experiment we found that particularly for larger values of n, with minibatch training, we could obtain lower distillation losses by optimizing for longer, so the closing of the gap between the upper bound and experiment values in fig. 4 may be misleading: longer optimization could bring the actual distillation loss lower.

To ensure that assumption (II) is fulfilled, we scale the labels such that $\left\| f^{\lambda}_{[\mathbf{X},y,k]} \right\|_{\mathcal{H}} = 1$. For example, if we are working with MNIST binary classification, with labels $\{+1, -1\}$, we first compute $\left\| f^{\lambda}_{[\mathbf{X},y,k]} \right\|_{\mathcal{H}} = r$ using $\{+1, -1\}$ labels, then rescale the labels by $1/r$ so that the labels are $\{+\frac{1}{r}, -\frac{1}{r}\}$. Suppose this results in some upper bound $\mathcal{L}_U$ and some real distillation loss $\mathcal{L}_R$. For the corresponding plots in figs. 2 to 4, we plot $r^2\mathcal{L}_U$ and $r^2\mathcal{L}_R$. We do this because the r values for different parameters (such as n or σ_x) could be different, and scaling for the plots allows the values to be comparable.

In the figures for the upper bounds on the distillation loss we plot the smallest value of the upper bounds in remark 6.

A.1 Code

Code is available in the supplementary material.