Motivation: SGD with uniform sampling wastes time on examples with negligible gradients

To evaluate the performance of our upper bound we compute the variance reduction at different stages of training. Our importance distribution performs on par with the optimal even in cases where the loss completely fails to reduce the variance.

To avoid computing $\|G\|_2$, we use an upper bound $\hat{G}_i$:

$\|G\|_2 \leq \hat{G}_i \iff \min_p \mathbb{E}_P \left[ \frac{1}{p} \hat{G}_i^2 \right] \leq \mathbb{E}_P \left[ w_i^2 \|G_i\|_2^2 \right].$

Observation: Due to normalization and initialization the gradient norm variation is captured by the last layers.

Result: We upper bound the gradient norm with the norm of the gradient of the last layer output (pre-activation).

Only one forward pass is required to compute this upper bound for all samples in a batch.

**Sequence classification:** LSTM on pixel-by-pixel permuted MNIST

**Finetuning:** Pretrained ResNet-50 finetuned on the MIT-67 dataset