Not All Samples Are Created Equal: Deep Learning with Importance Sampling

Angelos Katharopoulos[1,2] François Fleuret[1,2]

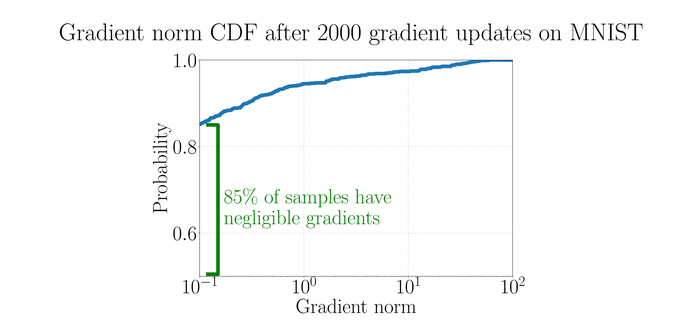
[1]Idiap Research Institute [2]École Polytechnique Fédérale de Lausanne

Funded by FNSNF

http://importance-sampling.com/

Idea

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



Use importance sampling to select informative examples!

Derivation of importance distribution

Similar to Zhao and Zhang (2015) we search for the **importance distribution** P^*, such that

$$P^* = \arg\min_P \operatorname{Tr}\left(\mathbb{V}_P[w_i G_i]\right) \iff p_i \propto \|G_i\|_2$$

where G_i is the *per sample gradient* and w_i are *sampling weights* that ensure that we have an unbiased estimator of the gradient.

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

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

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

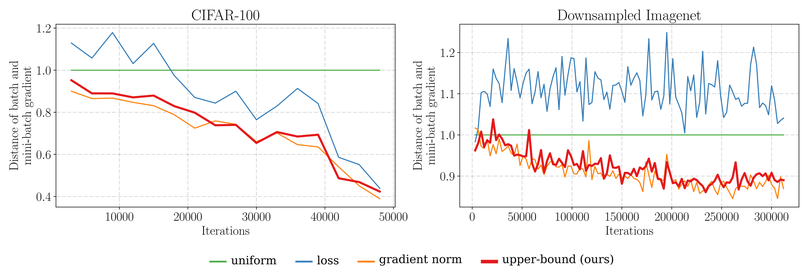
Result: We upper bound the gradient norm with the norm of the gradient of the last layer output (pre-activation).

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

Evaluation of importance distribution

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.



When to start importance sampling?

Given a large batch consisting of B samples we derive the variance reduction in terms of equivalent batch increment τ in closed form.

$$\tau \geq \left(1 - \frac{\sum_{i=1}^B (p_i - u)^2}{\sum_{i=1}^B p_i^2}\right)^{-1}$$

Assuming that the backward pass requires twice the amount of time of the forward, we achieve speedup by

1. resampling $b < B$ samples with importance
2. starting importance sampling iff $B + 3b < 3\tau b$

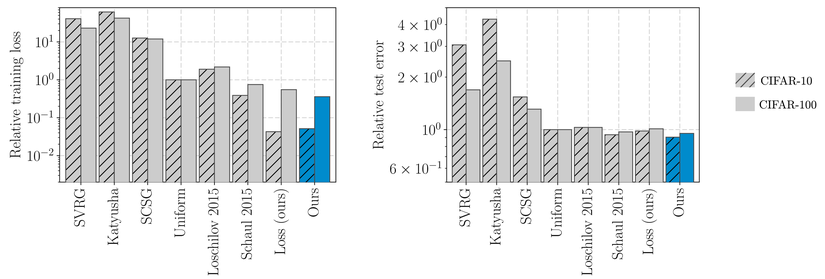
One SGD iteration with importance sampling

1: **if** $\tau > \tau_{th}$ **then**
2: Sample B datapoints uniformly and compute the forward pass
3: Resample b with importance and compute the forward-backward pass
4: **else**
5: Sample b uniformly and compute the forward-backward pass
6: **end if**
7: Update τ with the uniformly sampled scores

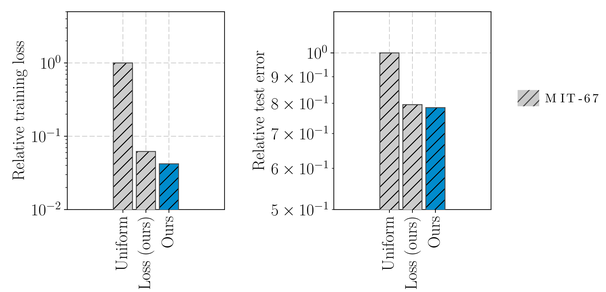
Experiments

- We allocate to all methods **the same time budget**
- Our importance sampling achieves **improved training loss and test error across 3 tasks and 4 datasets**

Image classification: WideResnet-28-2 on the CIFAR datasets



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



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