Not All Samples Are Created Equal
Supplementary material

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

1. Differences of variances

In the following equations we quantify the variance reduction achieved with importance sampling using the gradient norm. Let $g_i \propto \left\|\nabla_{\theta_t} \mathcal{L}(\Psi(x_i;\theta_t), y_i)\right\|_2 = \left\|G_i\right\|_2$ and $u = \frac{1}{B}$ the uniform probability.

We want to compute

$$\begin{aligned} &\text{Tr}\left(\mathbb{V}_u[G_i]\right) - \text{Tr}\left(\mathbb{V}_g[w_i G_i]\right) \\ &\qquad = \mathbb{E}_u\left[\left\|G_i\right\|_2^2\right] - \mathbb{E}_g\left[w_i^2 \left\|G_i\right\|_2^2\right]. \quad (1) \end{aligned}$$

Using the fact that $w_i = \frac{1}{B g_i}$ we have

$$\mathbb{E}_g\left[w_i^2 \left\|G_i\right\|_2^2\right] = \left(\frac{1}{B}\sum_{i=1}^{B} \left\|G_i\right\|_2\right)^2, \quad (2)$$

thus

$$\begin{aligned} &\text{Tr}\left(\mathbb{V}_u[G_i]\right) - \text{Tr}\left(\mathbb{V}_g[w_i G_i]\right) && (3) \\ &= \frac{1}{B}\sum_{i=1}^{B}\left\|G_i\right\|_2^2 - \left(\frac{1}{B}\sum_{i=1}^{B}\left\|G_i\right\|_2\right)^2 && (4) \\ &= \frac{\left(\sum_{i=1}^{B}\left\|G_i\right\|_2\right)^2}{B^3}\sum_{i=1}^{B}\left(B^2\frac{\left\|G_i\right\|_2^2}{\left(\sum_{i=1}^{B}\left\|G_i\right\|_2\right)^2} - 1\right) && (5) \\ &= \frac{\left(\sum_{i=1}^{B}\left\|G_i\right\|_2\right)^2}{B}\sum_{i=1}^{B}\left(g_i^2 - u^2\right). && (6) \end{aligned}$$

Completing the squares at equation 6 and using the fact that $\sum_{i=1}^{B} u = 1$ we complete the derivation.

$$\begin{aligned} &\text{Tr}\left(\mathbb{V}_u[G_i]\right) - \text{Tr}\left(\mathbb{V}_g[w_i G_i]\right) && (7) \\ &= \frac{\left(\sum_{i=1}^{B}\left\|G_i\right\|_2\right)^2}{B}\sum_{i=1}^{B}\left(g_i - u\right)^2 && (8) \\ &= \left(\frac{1}{B}\sum_{i=1}^{B}\left\|G_i\right\|_2\right)^2 B\left\|g-u\right\|_2^2. && (9) \end{aligned}$$

2. An upper bound to the gradient norm

In this section, we reiterate the analysis from the main paper (§ 3.2) with more details.

Let $\theta^{(l)} \in \mathbb{R}^{M_l \times M_{l-1}}$ be the weight matrix for layer l and $\sigma^{(l)}(\cdot)$ be a Lipschitz continuous activation function. Then, let

$$\begin{aligned} x^{(0)} &= x && (10) \\ z^{(l)} &= \theta^{(l)}\, x^{(l-1)} && (11) \\ x^{(l)} &= \sigma^{(l)}(z^{(l)}) && (12) \\ \Psi(x;\Theta) &= x^{(L)}. && (13) \end{aligned}$$

Equations 10-13 define a simple fully connected neural network without bias to simplify the closed form definition of the gradient with respect to the parameters Θ.

In addition we define the gradient of the loss with respect to the output of the network as

$$\nabla_{x_i^{(L)}}\mathcal{L} = \nabla_{x_i^{(L)}}\mathcal{L}(\Psi(x_i;\Theta), y_i) \quad (14)$$

and the gradient of the loss with respect to the output of layer l as

$$\nabla_{x_i^{(l)}}\mathcal{L} = \Delta_i^{(l)}\Sigma_L'(z_i^{(L)})\nabla_{x_i^{(L)}}\mathcal{L} \quad (15)$$

where

$$\Delta_i^{(l)} = \Sigma_l'(z_i^{(l)})\theta_{l+1}^T \dots \Sigma_{L-1}'(z_i^{(L-1)})\theta_L^T \quad (16)$$

propagates the gradient from the last layer (pre-activation) to layer l and

$$\Sigma_l'(z) = diag\left(\sigma'^{(l)}(z_1), \dots, \sigma'^{(l)}(z_{M_l})\right) \quad (17)$$

[1] Idiap Research Institute, Martigny, Switzerland [2] EPFL, Lausanne, Switzerland. Correspondence to: Angelos Katharopoulos <firstname.lastname@idiap.ch>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

defines the gradient of the activation function of layer l.

Finally, the gradient with respect to the parameters of the l-th layer can be written

$$
\begin{aligned}
&\left\|\nabla_{\theta_l}\mathcal{L}(\Psi(x_i;\Theta),y_i)\right\|_2 && (18)\\
&= \left\|\left(\Delta_i^{(l)}\Sigma'_L(z_i^{(L)})\nabla_{x_i^{(L)}}\mathcal{L}\right)\left(x_i^{(l-1)}\right)^T\right\|_2 && (19)\\
&\leq \left\|x_i^{(l-1)}\right\|_2\left\|\Delta_i^{(l)}\right\|_2\left\|\Sigma'_L(z_i^{(L)})\nabla_{x_i^{(L)}}\mathcal{L}\right\|_2 . && (20)
\end{aligned}
$$

We observe that $x_i^{(l)}$ and $\Delta_i^{(l)}$ depend only on z_i and Θ. However, we theorize that due to various weight initialization and activation normalization techniques those quantities do not capture the important per sample variations of the gradient norm. Using the above, which is also shown experimentally to be true in § 4.1, we deduce the following upper bound per layer

$$
\begin{aligned}
&\left\|\nabla_{\theta_l}\mathcal{L}(\Psi(x_i;\Theta),y_i)\right\|_2 && (21)\\
&\leq \max_{l,i}\left(\left\|x_i^{(l-1)}\right\|_2\left\|\Delta_i^{(l)}\right\|_2\right)\left\|\Sigma'_L(z_i^{(L)})\nabla_{x_i^{(L)}}\mathcal{L}\right\|_2 && (22)\\
&= \rho\left\|\Sigma'_L(z_i^{(L)})\nabla_{x_i^{(L)}}\mathcal{L}\right\|_2 , && (23)
\end{aligned}
$$

which can then be used to derive our final upper bound

$$
\left\|\nabla_{\Theta}\mathcal{L}(\Psi(x_i;\Theta),y_i)\right\|_2 \leq \underbrace{L\rho\left\|\Sigma'_L(z_i^{(L)})\nabla_{x_i^{(L)}}\mathcal{L}\right\|_2}_{\hat{G}_i} . \tag{24}
$$

Intuitively, equation 24 means that the variations of the gradient norm are mostly captured by the final classification layer. Consequently, we can use the gradient of the loss with respect to the pre-activation outputs of our neural network as an upper bound to the per-sample gradient norm.

3. Comparison with SVRG methods

For completeness, we also compare our proposed method with Stochastic Variance Reduced Gradient methods and present the results in this section. We follow the experimental setup of § 4.2 and evaluate on the augmented CIFAR10 and CIFAR100 datasets. The algorithms we considered were SVRG (Johnson & Zhang, 2013), accelerated SVRG with Katyusha momentum (Allen-Zhu, 2017) and, the most suitable for Deep Learning, SCSG (Lei et al., 2017) which in practice is a mini-batch version of SVRG. SAGA (Defazio et al., 2014) was not considered due to the prohibitive memory requirements for storing the per sample gradients.

For all methods, we tune the learning rate and the epochs per batch gradient computation (m in SVRG literature). For SCSG, we also tune the large batch (denoted as B_j in Lei et al. (2017)) and its growth rate. The results are depicted in figure 1. We observe that SGD with momentum performs significantly better than all SVRG methods. Full batch SVRG and Katyusha perform a small number of parameter updates thus failing to optimize the networks. In all cases, the best variance reduced method achieves more than an order of magnitude higher training loss than our proposed importance sampling scheme.

4. Ablation study on B

The only hyperparameter that is somewhat hard to define in our algorithm is the pre-sampling size B. As mentioned in the main paper, it controls the maximum possible variance reduction and also how much wall-clock time one iteration with importance sampling will require.

In figure 2 we depict the results of training with importance sampling and different pre-sampling sizes on CIFAR10. We follow the same experimental setup as in the paper.

We observe that larger presampling size results in lower training loss, which follows from our theory since the maximum variance reduction is smaller with small B. In this experiment we use the same τ_{th} for all the methods and we observe that $B = 384$ reaches first to 0.6 training loss. This is justified because computing the importance for $1,024$ samples in the beginning of training is wasteful according to our analysis.

According to this preliminary ablation study for B, we conclude that choosing $B = kb$ with $2 < k < 6$ is a good strategy for achieving a speedup. However, regardless of the choice of B, pairing it with a threshold τ_{th} designated by the analysis in the paper guarantees that the algorithm will be spending time on importance sampling only when the variance can be greatly reduced.

5. Importance Sampling with the Loss

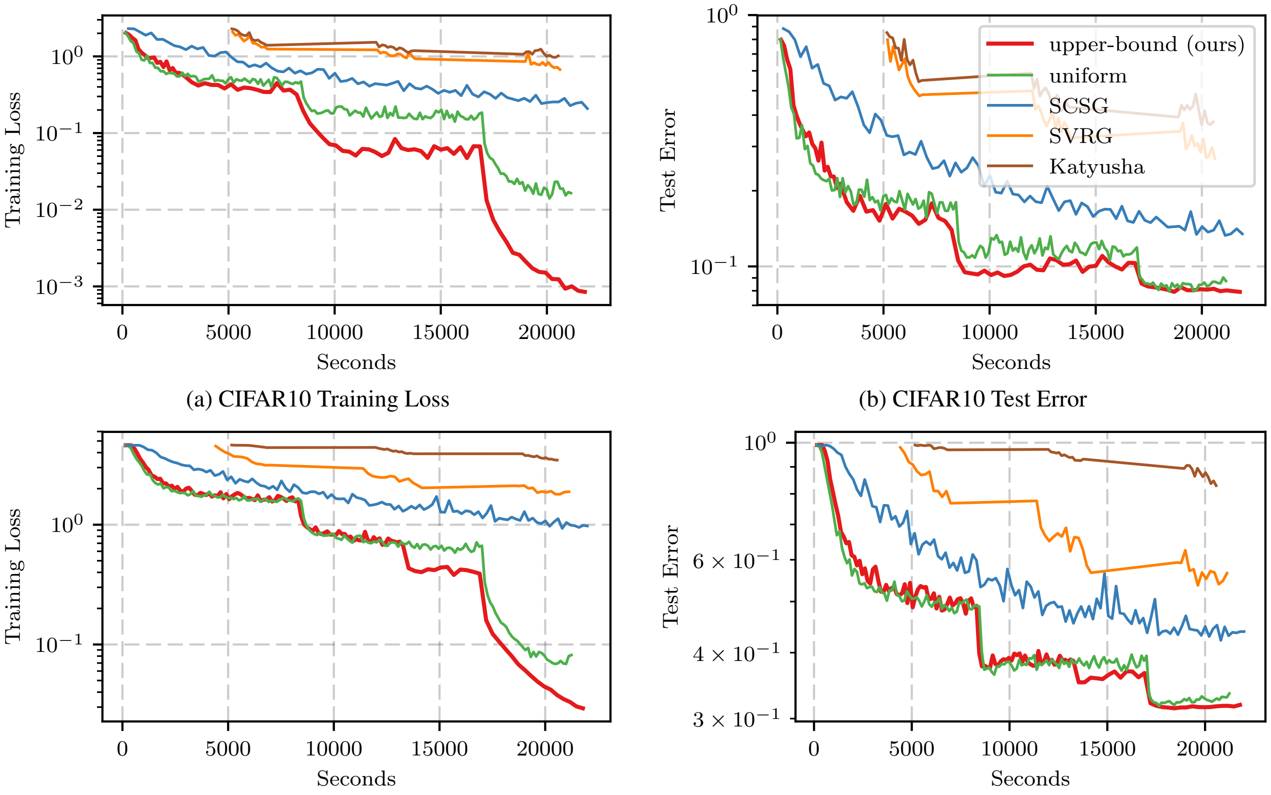
In this section we will present a small analysis that provides intuition regarding using the loss as an approximation or an upper bound to the per sample gradient norm.

Let $\mathcal{L}(\psi, y) : D \to \mathbb{R}$ be either the negative log likelihood through a sigmoid or the squared error loss function defined respectively as

$$
\begin{aligned}
\mathcal{L}_1(\psi,y) &= -\log\left(\frac{\exp(y\psi)}{1+\exp(y\psi)}\right) \quad && y\in\{-1,1\} \quad \psi\in\mathbb{R}\\
\mathcal{L}_2(\psi,y) &= \|y-\psi\|_2^2 && y\in\mathbb{R}^d \quad \psi\in\mathbb{R}^d
\end{aligned}
\tag{25}
$$

Given our upper bound to the gradient norm, we can write

$$
\left\|\nabla_{\theta_t}\mathcal{L}(\Psi(x_i;\theta_t),y_i)\right\|_2 \leq L\rho\left\|\nabla_{\psi}\mathcal{L}(\Psi(x_i;\theta_t),y_i)\right\|_2 . \tag{26}
$$



(a) CIFAR10 Training Loss

(b) CIFAR10 Test Error

(c) CIFAR100 Training Loss

(d) CIFAR100 Test Error

Figure 1: Comparison of our proposed importance sampling scheme (*upper-bound*) to SGD with uniform sampling and variance reduced methods. Only SCSG can actually perform enough iterations to optimize the network. However, SGD with uniform sampling and our *upper-bound* greatly outperform SCSG.

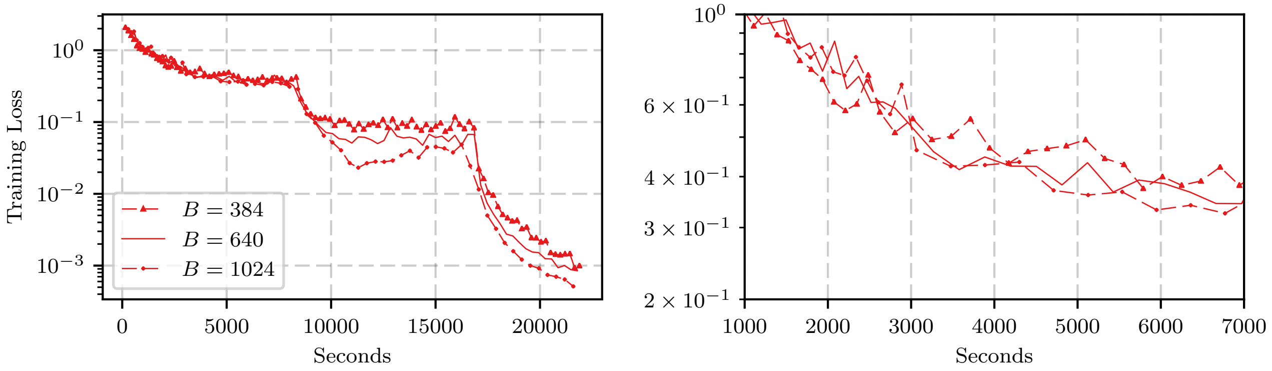


Figure 2: Results on training with different B on CIFAR10. See the paper for the experimental setup.

Moreover, for the losses that we are considering, when $\mathcal{L}(\psi, y) \to 0$ then $\|\nabla_\psi \mathcal{L}(\Psi(x_i; \theta_t), y_i)\|_2 \to 0$. Using this fact in combination to equation 26, we claim that so does the per sample gradient norm thus small loss values imply small gradients. However, large loss values are not well correlated with the gradient norm which can also be observed in § 4.1 in the paper.

To summarize, we conjecture that due to the above facts, sampling proportionally to the loss reduces the variance only when the majority of the samples have losses close to 0. Our assumption is validated from our experiments, where the *loss* struggles to achieve a speedup in the early stages of training where most samples still have relatively large loss values.

References

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1200–1205. ACM, 2017.

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pp. 1646–1654, 2014.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pp. 315–323, 2013.

Lei, L., Ju, C., Chen, J., and Jordan, M. I. Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems*, pp. 2345–2355, 2017.