A Scale-Free MADGRAD Regret Bound

Shashank Manjunath
Boston University
manjuns@bu.edu

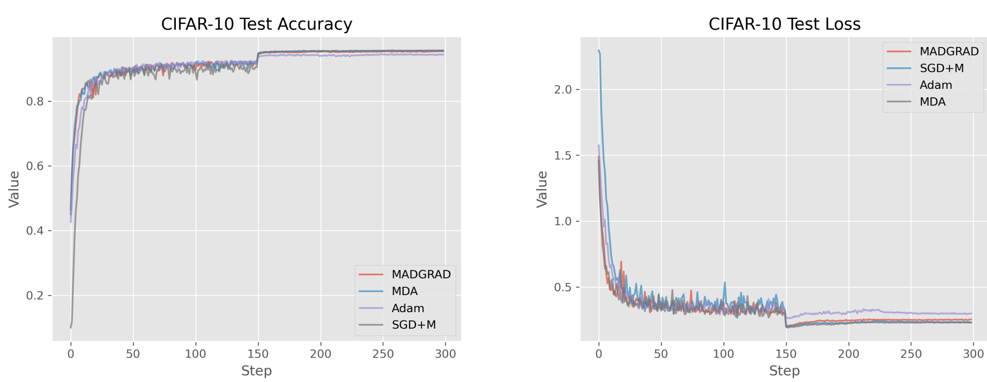
1 Introduction

This project is concerned with dual averaging algorithms applied to deep learning. So far, we have tested two dual averaging algorithms, Modernized Dual Averaging (MDA) (Jelassi and Defazio [2020]) and MADGRAD (Defazio and Jelassi), which use Follow the Regularized Leader (FTRL) style algorithms in order to optimize deep learning algorithms. For this project, we have focused on both implementing these algorithms in PyTorch (Paszke et al. [2019]) and testing them out on the CIFAR10 dataset (Krizhevsky). We then prove an alternate, scale-free regret bound for the MADGRAD algorithm.

2 Algorithm Details

3 Algorithm Implementation

So far, we have successfully replicated results on the CIFAR10 dataset for the MDA, MADGRAD, Adam, and Stochastic Gradient Descent with Momentum (SGD+M) algorithms. We show our test accuracy and test loss results in the plot below.



(a) Test Accuracy of Optimizers on CIFAR-10 (b) Test Loss of Optimizers on CIFAR-10

Figure 1: Comparison of optimizer performance on CIFAR-10 dataset

Our experiment setup and optimizer implementations for MDA and MADGRAD can be found at `https://github.com/shashankmanjunath/ftrl_deep_learning`.

4 Theory

While studying the convergence proof for the MADGRAD algorithm, we identified a potential improvement on the existing theory.

make above line sound more professional

When proving the convergence bound for MADGRAD, the authors require an alternative definition of MADGRAD than the one presented in the paper and implemented. In the original MADGRAD algorithm presented in the paper, the z_{k+1} is given by:

$$z_{k+1} = x_0 - \frac{1}{\sqrt[3]{v_{k+1}} + \epsilon} \circ s_{k+1}$$

where $\circ$ indicates the Hadamard product. ϵ is included for numerical stability in the early iterations of the algorithm, as the v_{k+1} parameter can be 0. However, in the convergence proof, the z_{k+1} parameter is given by:

$$z_{k+1} = x_0 - \frac{1}{\sqrt[3]{\lambda_{k+1} G^2 + v_{k+1}}} \circ s_{k+1}$$

Note the extra $\lambda_{k+1} G^2$ in the denominator, which is used to create the following upper bound leveraged in the overall convergence proof:

$$\sum_{t=0}^{k} \frac{\lambda_t^2 g_t^2}{\sqrt[3]{\lambda_t G^2 + \sum_{i=0}^{t-1} \lambda_i g_i^2}} \leq \frac{3}{2} \lambda_k \left(\sum_{i=1}^{k} \lambda_i g_i^2 \right)^{\frac{2}{3}}$$

This extra $\lambda_t G^2$ prevents the algorithm from being *scale-free*, or an algorithm that is invariant to the scaling of losses by a constant factor. Therefore, we aim to construct a convergence proof which maintains the scale-free nature of the algorithm.

4.1 Proof of Scale-Free Regret Bound for MADGRAD

Consider the MADGRAD algorithm. This algorithm implements FTRL with the regularizer:

$$\psi_t(\mathbf{x}) = \frac{1}{2} \|\mathbf{x} - \mathbf{x}_0\|_{A_t}$$

where $A_t = \text{diag}(\alpha_t)$, and $\alpha_t = \sqrt[3]{\sum_{i=1}^{t-1} \lambda_i g_i^2}$. Note that $\psi_t(\mathbf{x})$ is strongly-convex with respect to the norm $\|\cdot\|_{A_t}$. Let us denote the Bregman divergence of a function f over two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$ by $B_f(\mathbf{x}; \mathbf{y})$ and let $f^\star$ indicate the Fenchel conjugate of f. In order to prove the scale-free bound, we make the reduction that $c_t = 1$ for all rounds. This effectively removes the momentum operation, and leaves us with only FTRL iterates. Let us first define some useful lemmas.

Lemma 4.1 (Lemma 1 in Orabona et al. [2014]). *Let $\{\psi_t\}_{t=1}^{\infty}$ be a sequence of functions defined on a common convex domain $S \subseteq \mathbb{R}^n$ and such that each ψ_t is μ_t-strongly convex with respect to the norm $\|\cdot\|_t$. Let $\|\cdot\|_{t,\star}$ be the dual norm of $\|\cdot\|_t$, for $t = 1, 2, \cdots, T$. Then, for any $\mathbf{u} \in S$,*

$$\textit{Regret}_T(\mathbf{u}) \leq \sum_{t=1}^{T} \langle g_t, \mathbf{u} - \mathbf{x}_t \rangle \leq \psi_T(\mathbf{u}) + \psi_1^\star(0) + \sum_{t=1}^{T} B_{\psi_t^\star}(-\theta_t, -\theta_{t-1}) - \psi_t^\star(-\theta_t) + \psi_{t+1}^\star(-\theta_t)$$

Proof. Given in (Orabona et al. [2014]) □

Lemma 4.2. *Let $a_1, a_2, \cdots, a_t$ be non-negative real numbers. If $a_1 > 0$, then*

$$\sum_{t=1}^{T} \frac{a_t}{\sqrt[3]{\sum\limits_{s=1}^{t} a_s}} \leq \frac{3}{2} \left(\sum_{t=1}^{T} a_t \right)^{\frac{2}{3}}$$

Proof. Note that if $0 \leq x \leq 1$,

$$\frac{2}{3} x \leq 1 - (1-x)^{\frac{2}{3}}$$

Let $L_t = \sum\limits_{i=1}^{t} \ell_i$, and let $x = \frac{\ell_t}{L_t}$. Let $\ell_0 = 0$.

$$\begin{aligned}
\frac{2}{3} \frac{\ell_t}{L_t} &\leq 1 - (1 - \frac{\ell_t}{L_t})^{\frac{2}{3}} = 1 - (\frac{L_{t-1}}{L_t})^{\frac{2}{3}} \\
\frac{2}{3} \frac{\ell_t}{L_t} L_t^{\frac{2}{3}} &\leq L_t^{\frac{2}{3}} - L_{t-1}^{\frac{2}{3}} \\
\frac{2}{3} \frac{\ell_t}{\sqrt[3]{L_t}} &\leq L_t^{\frac{2}{3}} - L_{t-1}^{\frac{2}{3}} \\
\therefore \frac{2}{3} \sum_{t=1}^{T} \frac{\ell_t}{\sqrt[3]{L_t}} &\leq \sum_{t=1}^{T} L_t^{\frac{2}{3}} - L_{t-1}^{\frac{2}{3}} \\
\sum_{t=1}^{T} \frac{\ell_t}{\sqrt[3]{L_t}} &\leq \frac{3}{2} L_T^{\frac{3}{2}} \\
\sum_{t=1}^{T} \frac{\ell_t}{\sqrt[3]{L_t}} &\leq \frac{3}{2} \left(\sum_{t=1}^{T} \ell_t \right)^{\frac{2}{3}}
\end{aligned}$$

Letting $\ell_i = a_i \forall i$ yields the lemma.

$\square$

Lemma 4.3. *Let $C, a_1, a_2, \cdots, a_T \geq 0$, and $\alpha \geq 1$. Then,*

$$\sum_{t=1}^{T} \min \left\{ \frac{a_t^2}{\sqrt[3]{\sum\limits_{s=1}^{t-1} a_s^2}}, C a_t \right\} \leq \frac{C\alpha}{\alpha - \left(\sum\limits_{s=1}^{t-1} a_s^2 \right)^{\frac{2}{3}}} \max_{t=1,2,\cdots,T} a_t + 2\sqrt[3]{1+\alpha^2} \sqrt{\sum_{s=1}^{T} a_s^3}$$

Proof. We will prove this bound by proving each individual case in the minimum, then summing them.

Case 1. Consider $a_t \leq \alpha^3 \left(\sum\limits_{s=1}^{t-1} a_s^2 \right)^{\frac{2}{3}}$.

$$\min\left\{\frac{a_t^2}{\sqrt[3]{\sum_{s=1}^{t-1} a_s^2}}, Ca_t\right\} \leq \frac{\alpha_t^2}{\sqrt[3]{\sum_{s=1}^{t-1} a_s^2}} = \frac{a_t^2}{\sqrt[3]{\frac{1}{1+\alpha^2}\left(\alpha^2 \sum_{s=1}^{t-1} a_s^2 + \sum_{s=1}^{t-1} a_s^2\right)}}$$

$$\leq \frac{a_t^2 \sqrt[3]{1+\alpha^2}}{\sqrt[3]{a_t^2 + \sum_{s=1}^{t-1} a_s^2}} = \frac{a_t^2 \sqrt[3]{1+\alpha^2}}{\sqrt[3]{\sum_{s=1}^{t} a_s^2}}$$

Note that $\frac{x^2}{\sqrt[3]{x^2+y^2}} \leq 2(\sqrt{x^3+y^3} - \sqrt{y^3})$. Using this inequality,

$$\sqrt[3]{1+\alpha^2}\frac{a_t^2}{\sqrt[3]{\sum_{s=1}^{t} a_s^2}} \leq 2\sqrt[3]{1+\alpha^2}\left(\sqrt{\sum_{s=1}^{t} a_s^3} - \sqrt{\sum_{s=1}^{t-1} a_s^3}\right)$$

Case 2 Consider $a_t^2 \geq \alpha^3 \left(\sum_{s=1}^{t-1} a_s^2\right)^{\frac{2}{3}}$. Note that this implies that $a_t \geq \alpha^{\frac{3}{2}} \sqrt[3]{\sum_{s=1}^{t-1} a_s^2}$. Additionally, let $A = \left(\sum_{s=1}^{t-1} a_s^2\right)^{\frac{2}{3}}$.

$$\min\left\{\frac{a_t^2}{\sqrt[3]{\sum_{s=1}^{t-1} a_s^2}}, Ca_t\right\} \leq Ca_t = Ca_t\left(\frac{\alpha - A}{\alpha - A}\right)$$

$$\leq \frac{C}{\alpha - A}\left(\alpha a_t - Aa_t\right) = \frac{C\alpha}{\alpha - A}\left(a_t - \alpha^{\frac{1}{2}} A \sqrt[3]{\sum_{s=1}^{t-1} a_s^2}\right)$$

$$\leq \frac{C\alpha}{\alpha - A}\left(a_t - \left(\sum_{s=1} a_s^2\right)^{\frac{2}{3}}\left(\sum_{s=1} a_s^2\right)^{\frac{1}{3}}\right)$$

$$\leq \frac{C\alpha}{\alpha - A}\left(a_t - \sqrt{\sum_{s=1}^{t-1} a_s^2}\right)$$

Let $M_t = \max\{a_t, \cdots, a_t\}$. Note that in this case, $a_t = M_t$, and $\sqrt{\sum_{s=1}^{t-1} a_s^2} \geq M_{t-1}$. Therefore,

$$\min\left\{\frac{a_t^2}{\sqrt[3]{\sum_{s=1}^{t-1} a_s^2}}, Ca_t\right\} \leq \frac{C\alpha}{\alpha - A}\left(M_t - M_{t-1}\right)$$

Therefore, combining the two cases, we have:

$$\min\left\{\frac{a_t^2}{\sqrt[3]{\sum\limits_{s=1}^{t-1} a_s^2}}, Ca_t\right\} \leq \frac{C\alpha}{\alpha - A}\left(M_t - M_{t-1}\right) + 2\sqrt[3]{1+\alpha^2}\left(\sqrt{\sum_{s=1}^{t} a_s^3} - \sqrt{\sum_{s=1}^{t-1} a_s^3}\right)$$

Therefore, summing from $t = 1$ to T,

$$\sum_{t=1}^{T}\min\left\{\frac{a_t^2}{\sqrt[3]{\sum\limits_{s=1}^{t-1} a_s^2}}, Ca_t\right\} \leq \frac{C\alpha}{\alpha - A}\left(\max_{t=1,\cdots,T} a_t\right) + 2\sqrt[3]{1+\alpha^2}\left(\sqrt{\sum_{t=1}^{T} a_t^3}\right)$$

□

Let us now state the overall convergence bound for our version of MADGRAD.

Theorem 4.4. *Suppose $K \subseteq \mathbb{R}^D$ is a non-empty closed convex subset. Suppose that a regularizer $\psi_t : K \to R$ is a non-negative lower semi-continuous function that is μ_t strongly convex with respect to a norm $\|\cdot\|_t$. The regret of non-momentumized MADGRAD satisfies:*

$$\textit{Regret}_T(u) \leq \sum_{d=1}^{D}\left(\frac{\psi_d(u_d)}{\sqrt[3]{\sum\limits_{i=1}^{T}\lambda_i g_{id}^2}}\right) + \frac{3}{2}\left(\sum_{t=1}^{T}\lambda_t g_{td}^2\right)^{\frac{2}{3}} + \frac{2\sqrt{T-1}\left(\sum_{i=1}^{t-1}\lambda_i g_{id}^2\right)}{1-\left(\sum_{i=1}^{t-1} a_i^2\right)^{\frac{2}{3}}} \max_{t\leq T}\sqrt{\lambda_t}|g_{td}|$$
$$+ 2\sqrt[3]{2}\sqrt{\sum_{t-1}^{T}\lambda_t^{\frac{3}{2}}|g_{td}|^3}$$

Proof. Note that $\psi_t(x) = \frac{1}{2}\|\mathbf{x}-\mathbf{x}_0\|_{A_t}^2$, where $A_t = \text{diag}(\alpha_t)$. For this regularizer, $\alpha_t = \sqrt[3]{\sum\limits_{i=1}^{t-1}\lambda_i g_i^2} \in \mathbb{R}^D$. Let $L_t = \sum\limits_{i=1}^{t}\lambda_i g_i$.Let us perform this analysis per-coordinate.

$$\psi_{t,d}(\mathbf{x}) = \frac{1}{2}(\mathbf{x}_d - \mathbf{x}_{0,d})\sqrt[3]{\sum_{i=1}^{t-1}\lambda_i g_{id}^2}(\mathbf{x}_d - \mathbf{x}_{0,d}) = \frac{1}{2}\sqrt[3]{\sum_{i=1}^{t-1}\lambda_i g_{id}^2}(\mathbf{x}_d - \mathbf{x}_{0,d})^2.$$

Let $\eta_{t,d} = \frac{1}{\sqrt[3]{\sum\limits_{i=1}^{t-1}\lambda_i g_{id}^2}}$. Therefore, we have:

$$\psi_{t,d}(\mathbf{x}) = \frac{1}{\eta_{t,d}}\psi_d(\mathbf{x})$$
$$\psi_d(\mathbf{x}) = \frac{1}{2}(\mathbf{x}_d - \mathbf{x}_{0,d})$$

Since A_t is a diagonal matrix, $\psi_t(\mathbf{x})$ is $\min_D \sqrt[3]{\sum\limits_{i=1}^{t-1}\lambda_i g_i^2}$-strongly convex. We can tighten this bound by analyzing $\psi_{t,d}(\mathbf{x})$ and establishing a per-coordinate strong convexity bound. Recall that, since $\psi_d(\mathbf{x}) : \mathbb{R} \to \mathbb{R}$, we can find the strong convexity by finding the lower bound of the second derivative of $\psi_d(\mathbf{x})$.

References

Aaron Defazio and Samy Jelassi. Adaptivity without Compromise: A Momentumized, Adaptive, Dual Averaged Gradient Method for Stochastic Optimization. page 33.

Samy Jelassi and Aaron Defazio. Dual Averaging is Surprisingly Effective for Deep Learning Optimization. *arXiv:2010.10502 [cs, math, stat]*, October 2020. URL http://arxiv.org/abs/2010.10502. arXiv: 2010.10502.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. page 60.

Francesco Orabona, Koby Crammer, and Nicolò Cesa-Bianchi. A Generalized Online Mirror Descent with Applications to Classification and Regression. *arXiv:1304.2994 [cs]*, July 2014. URL http://arxiv.org/abs/1304.2994. arXiv: 1304.2994.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.