A Regret Bound for Online Gradient Descent with Momentum

Laurent Orseau (but I didn't do much)

June 2025

Abstract

Finding a paper that provides a non-vacuous regret bound for online gradient descent (OGDM) with fixed momentum was surprisingly difficult. The proofs in the Adam paper and variants (like AMSGrad) make use of an exponentially decaying momentum *parameter*, which basically reverts to no momentum very quickly. Fortunately, [AMMC20] have already solved this problem — the core ingredient is their simple Lemma 1. In this document we use this lemma to provide a simple regret bound for OGDM, to help understand the simplest case. For a momentum parameter β, OGDM has a worst-case adversarial regret of $O(DG\sqrt{T/(1-\beta)})$. While this bound does not show acceleration, it shows at least that OGDM is consistent. It also follows that Adagrad can (likely) be made to use of momentum without losing consistency.

Define Online Gradient Descent with Momentum (OGDM), in a convex compact (closed, bounded) domain $\mathcal{D}$:

$$m_t = \beta m_{t-1} + (1-\beta) g_t \tag{1}$$ {eq:momentum}

$$x_{t+1} = \Pi_{\mathcal{D}}(x_t - \eta m_t)\,. \tag{2}$$ {eq:update}

where $\Pi_{\mathcal{D}}(x) = \operatorname{argmin}_{y\in\mathcal{D}} \|x-y\|_2^2$, and $m_0 = 0$.

Assumptions. We consider some horizon T. Let $G \geq \max_{t\leq T} \|g_t\|$, then we also have $\max_{t\leq T}\|m_t\| \leq G$. For all x and y of the domain $\mathcal{D}$, $\|x-y\| \leq D$.

Theorem 1. For a fixed horizon T, when optimizing η, the regret of OGDM compared to any point x^* of the domain after T steps is bounded by

$$R_T \;\leq\; DG\sqrt{\frac{1+\beta}{1-\beta}T} + \frac{\beta}{1-\beta}DG\,. \qquad \diamond$$

Proof. From Eq. (1),

$$g_t = \frac{m_t}{1-\beta} - \frac{\beta m_{t-1}}{1-\beta}\,.$$

Following [AMMC20, Lemma 1], the regret can be written

$$
\begin{aligned}
R_T &\leq \sum_{t\leq T} \langle x_t - x^*, g_t \rangle \\
&= \frac{1}{1-\beta} \sum_{t\leq T} \langle x_t - x^*, m_t \rangle - \frac{\beta}{1-\beta} \sum_{t\leq T} \langle x_t - x^*, m_{t-1} \rangle \\
&= \frac{1}{1-\beta} \sum_{t\leq T} \langle x_t - x^*, m_t \rangle - \frac{\beta}{1-\beta} \underbrace{\sum_{t\leq T} \langle x_{t-1} - x^*, m_{t-1} \rangle}_{(A)} + \frac{\beta}{1-\beta} \underbrace{\sum_{t\leq T} \langle x_{t-1} - x_t, m_{t-1} \rangle}_{(B)} .
\end{aligned}
$$

With $m_0 = 0$,

$$
(A) = \sum_{t\leq T} \langle x_t - x^*, m_t \rangle - \langle x_T - x^*, m_T \rangle ,
$$

and by Cauchy-Schwartz $\langle x_T - x^*, m_T \rangle \leq \|x_T - x^*\| \, \|m_T\| \leq DG$. Similarly, using that projection is non-expansive,

$$
(B) \leq \sum_{t\leq T} \|x_t - x_{t-1}\| \, \|m_{t-1}\| \leq \sum_{t\leq T} \|\eta m_{t-1}\| \, \|m_{t-1}\| \leq \eta T G^2 .
$$

Thus:

$$
R_T \leq \sum_{t\leq T} \langle x_t - x^*, m_t \rangle + \frac{\beta}{1-\beta} DG + \eta \frac{\beta}{1-\beta} TG^2 .
$$

Applying the standard regret bound for OGD where the gradients have been substituted with the momentum terms:

$$
R_T \leq \frac{1}{2\eta} D^2 + \frac{1}{2} \eta T G^2 + \frac{\beta}{1-\beta} DG + \eta \frac{\beta}{1-\beta} TG^2
$$

Finally, optimizing for η gives the result. □

References

alacaoglu20adam

[AMMC20] Ahmet Alacaoglu, Yura Malitsky, Panayotis Mertikopoulos, and Volkan Cevher. A new regret analysis for Adam-type algorithms. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 202–210. PMLR, 13–18 Jul 2020.