A Local Regret in Nonconvex Online Learning

Sergul Aydore
Stevens Institute of Technology, NJ, USA,
saydore@stevens.edu

Lee Dicker
Amazon, NY, USA
leehd@amazon.com

Dean Foster
Amazon, NY, USA
foster@amazon.com

Abstract

We consider an online learning process to forecast a sequence of outcomes for nonconvex models. A typical measure to evaluate online learning algorithms is regret but such standard definition of regret is intractable for nonconvex models even in offline settings. Hence, gradient based definition of regrets are common for both offline and online nonconvex problems. Recently, a notion of local gradient based regret was introduced. Inspired by the concept of calibration and a local gradient based regret, we introduce another definition of regret and we discuss why our definition is more interpretable for forecasting problems. We also provide bound analysis for our regret under certain assumptions.

1 Introduction

In typical forecasting problems, we make probabilistic estimates of future outcomes based on the previous observations. Recently, it has been shown that forecasting models can be complex nonconvex models [3, 7]. Frequent update of these models is desired as the relationship between the targets and outputs might change over time. However, re-training these models can be time consuming.

Online learning is a method of updating the model on each pattern as it is observed as opposed to batch learning where the training is performed over groups of pattern. It is a common technique to dynamically adapt to new patterns in the data or when training over the entire data set is infeasible. The literature in online learning is rich with interesting theoretical and practical applications but it is usually limited to the convex problems where global optimization is computationally tractable [8]. On the other hand, it is NP-hard to compute the global minimum of nonconvex functions over a convex domain [5, 6].

Due to the intractability of the nonconvex problems, various assumptions on the input have been used to design polynomial-time algorithms [2, 6]. However, these were too specific to the models and more generic approach was needed. One way to achieve this is by replacing the "global optimality" requirement with a more modest requirement of stationarity [1].

The idea of online learning was borrowed from game theory where an online player answers a sequence of questions. The true answers to the questions are unknown to the player at the time of each decision and the player suffers a loss after committing to a decision. These losses are unknown to the player and the performance of the sequence of decisions will be evaluated by the difference between this accumulated loss and the best fixed decision in hindsight. Most recently, Hazan et al. [5] proposed a notion of gradient based local regret for nonconvex games.

Inspired by Hazan's approach and incorporating the notion of calibration, we introduce a novel gradient based local regret for forecasting problems. Calibration is a well-studied concept in forecasting [4]. From game theoretic point of view, we call a forecasting procedure "calibrated" if the forecasts are consistent in hindsight. To the best of our knowledge, such definition of regret is new. We show that the proposed regret has logarithmic bound under certain circumstances and we provide insights to the proposed regret. We conjecture that more efficient algorithms can be developed that minimizes our regret.

2 Setting

In online forecasting, our goal is to update x_t at each t in order to incorporate the most recently available information. Assume that $t \in \mathcal{T} = \{1, \cdots, T\}$ represents a collection of T consecutive points where T is an integer and $t = 1$ represents an initial forecast point. $f_1, \cdots, f_T : \mathcal{K} \to \mathbb{R}$ are nonconvex loss functions on some convex subset $\mathcal{K} \subseteq \mathbb{R}^d$. To put in another way, x_t represents the parameters of a machine learning model at time t, $f_t(x_t)$ represents the loss function computed using the available data at time t given the model parameters x_t.

2.1 Regret Analysis

The performance of online learning algorithms is commonly evaluated by the regret, which is defined as the difference between the real cumulative loss and the minimum cumulative loss across $\mathcal{T}$:

$$R(T) \triangleq \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x). \tag{1}$$

If the regret grows linearly with T, it can be concluded that the player is not learning. If, on the other hand, the regret grows sub-linearly, the player is learning and its accuracy is improving. While such definition of regret makes sense for convex optimization problems, it is not appropriate for nonconvex problems, due to NP-hardness of nonconvex global optimization even in offline settings. Indeed, most research on nonconvex problems focuses on finding local optima. In literature on nonconvex optimization algorithms, it is common to use the magnitude of the gradient to analyze convergence. Hazan et al. [5] introduced a local regret measure - a new notion of regret that quantifies the objective of predicting points with small gradients on average. At each round of the game, the gradients of the loss functions from w where $1 \leq w \leq T$ most recent rounds of play are evaluated at the forecast, and these gradients are then averaged. Hazan et al. [5]'s local regret is defined to be the sum of the squared magnitude of the gradients averages.

Definition 2.1. (Hazan's local regret) The w-local regret of an online algorithm is defined as:

$$HR_w(T) \triangleq \sum_{t=1}^{T} \|\nabla F_{t,w}(x_t)\|^2 \tag{2}$$

when $\mathcal{K} = \mathbb{R}^d$ and $F_{t,w}(x_t) \triangleq \frac{1}{w} \sum_{i=0}^{w-1} f_{t-i}(x_t)$. Hazan et al. [5] proposed various gradient descent algorithms where the regret HR is sublinear.

2.2 Proposed Local Regret

In order to introduce the concept of calibration [4], let's consider the first order Taylor series expansion of the cumulative loss:

$$\sum_{t=1}^{T} f_t(\text{proj}_{\mathcal{K}}(x_t + u)) = \sum_{t=1}^{T} f_t(x_t + D_u(x_t)) \approx \sum_{t=1}^{T} f_t(x_t) + \sum_{t=1}^{T} \langle D_u(x_t), \nabla f_t(x_t) \rangle \tag{3}$$

where $D_u(x_t) \triangleq \text{proj}_{\mathcal{K}}(x_t + u) - x_t$ for any $u \in \mathbb{R}^d$. If the forecasts $\{x_1, \cdots, x_T\}$ are well-calibrated, then perturbing x_t by any u cannot substantially reduce the cumulative loss. Hence, we can say that the sequence $\{x_1, \cdots, x_T\}$ is *asymptotically calibrated* with respect to $\{f_1, \cdots, f_T\}$, if:

$$\limsup_{T \to \infty} \sup_{u \in \mathbb{R}^d} -\frac{1}{T} \sum_{t=1}^{T} \langle D_u(x_t), \nabla f_t(x_t) \rangle \leq 0. \tag{4}$$

Definition 2.2. (Proposed Regret) We propose a w-local regret as:

$$PR_w(T) \triangleq \sum_{t=1}^{T} \left\| \frac{1}{w} \sum_{s=t-w+1}^{t} \langle D_u(x_s), \nabla f_s(x_s) \rangle \right\|^2 \tag{5}$$

where $f_t(x_t) = 0$ for $t \leq 0$. To motivate equation 5, we use the following equality:

$$\lim_{\delta \to 0} \frac{1}{\delta} \sup_{\|u\|=\delta} \left\| \frac{1}{w} \sum_{s=t-w+1}^{t} \langle D_u(x_s), \nabla f_s(x_s) \rangle \right\|^2 = \left\| \frac{1}{w} \sum_{s=t-w+1}^{t} \nabla f_s(x_s) \right\|^2 \tag{6}$$

which holds for the interior points. Using our definition of regret, we effectively evaluate an online learning algorithm by computing the average of losses at the corresponding forecast values over a sliding window. Hazan et al. [5]'s local regret, on the other hand, computes average of previous losses computed on the most recent forecast. We believe that our definition of regret is more applicable to forecasting problems as evaluating today's forecast on previous loss functions might be misleading.

3 Bound Analysis

We provide bound for different scenarios for the proposed regret in equation 5 for the interior points in the feasible set with the following assumptions: $\sup_{x,y \in \mathcal{K}} \|x - y\| = M$; $\sup_{x \in \mathcal{K}, t \in \mathcal{T}} \nabla f_t(x) = G$; parameter update at t is: $x_{t+1} = \text{proj}_{\mathcal{K}}(x_t - \eta_t \nabla f_t(x_t))$ where $\eta_t = \eta/\sqrt{t}$ is the learning rate for some small $\eta > 0$. We consider three scenarios: (i) $\eta_t = \eta$, w is constant and $\mathcal{K} = \mathbb{R}^d$, (ii) $\eta_t = \eta/\sqrt{t}$ and $w = t$, (iii) $\eta_t = \eta/\sqrt{t}$ and w is constant. We also note the following Theorem whose proof is provided in section 5.

Theorem 3.1. $\sum_{s=t-w+1}^{t} \langle D_u(x_s), \nabla f_s(x_s) \rangle \geq 2\eta G^2 \sqrt{t-w+1} - \left(\frac{3M^2}{2\eta} + 2\eta G^2 \right) \sqrt{t}$ where $\eta_s = \eta/\sqrt{s}$.

3.1 Scenario 1: $\eta_t = \eta$, w is constant and $\mathcal{K} = \mathbb{R}^d$

Since $\mathcal{K} = \mathbb{R}^d$, the update rule becomes $x_{t+1} = x_t - \eta \nabla f_t(x_t)$; in other words, no projection operator is necessary. Hence we can write:

$$
\begin{aligned}
\sum_{s=t-w+1}^{t} \langle D_u(x_s), \nabla f_s(x_s) \rangle &= \sum_{s=t-w+1}^{t} \langle u, \nabla f_s(x_s) \rangle = \left\langle u, \sum_{s=t-w+1}^{t} \nabla f_s(x_s) \right\rangle \tag{7} \\
&= \left\langle u, \frac{1}{\eta} \sum_{s=t-w+1}^{t} (x_s - x_{s+1}) \right\rangle = \frac{1}{\eta} \langle u, (x_{t-w+1} - x_{t+1}) \rangle \\
&\leq \frac{1}{\eta} \|u\| \|x_{t-w+1} - x_{t+1}\| \leq \frac{M\|u\|}{\eta}
\end{aligned}
$$

Taking u as a unit vector such that $u = \frac{\sum_{s=t-w+1}^{t} \nabla f_s(x_s)}{\|\sum_{s=t-w+1}^{t} \nabla f_s(x_s)\|}$, we can write $\|\sum_{s=t-w+1}^{t} \nabla f_s(x_s)\|^2 \leq M^2/\eta^2$. Hence; the bound for the proposed regret becomes:

$$PR_w(T) = \sum_{t=1}^{T} \left\| \frac{1}{w} \sum_{s=t-w+1}^{t} \nabla f_s(x_s) \right\|^2 \leq \frac{M^2 T}{w^2 \eta^2} \tag{8}$$

which can be made sublinear in T if w is selected large enough.

3.2 Scenario 2: $\eta_t = \eta/\sqrt{t}$ and $w = t$

Assuming $x_s + u$ is interior of the feasible set for all u and s and setting $w = t$, we can write the result in theorem 3.1 as:

$$\sum_{s=1}^{w} \langle u, \nabla f_s(x_s) \rangle = \left\langle u, \sum_{s=1}^{t} \nabla f_s(x_s) \right\rangle \quad = \quad -\frac{\|\sum_{s=1}^{t} \nabla f_s(x_s)\|^2}{\|\sum_{s=1}^{t} \nabla f_s(x_s)\|} \tag{9}$$

$$= \quad -\|\sum_{s=1}^{t} \nabla f_s(x_s)\| \tag{10}$$

$$\geq \quad -\left(\frac{3M^2}{2\eta} + 2\eta G^2\right)\sqrt{t} \tag{11}$$

where u is set to $-\frac{\sum_{s=1}^{t} \nabla f_s(x_s)}{\|\sum_{s=1}^{t} \nabla f_s(x_s)\|}$. Hence, we get:

$$\left\| \frac{1}{t} \sum_{s=1}^{t} \nabla f_s(x_s) \right\|^2 \leq \left(\frac{3M^2}{2\eta} + 2\eta G^2\right)^2 \frac{1}{t} \tag{12}$$

Summing this over t yields:

$$PR_w(T) = \sum_{t=1}^{T} \left\| \frac{1}{t} \sum_{s=1}^{t} \nabla f_s(x_s) \right\|^2 \leq \sum_{t=1}^{T} \left(\frac{3M^2}{2\eta} + 2\eta G^2\right)^2 \frac{1}{t} \leq \left(\frac{3M^2}{2\eta} + 2\eta G^2\right)^2 \log(T) \tag{13}$$

which concludes the logarithmic bound for the proposed regret for interior points when $\eta_t = \eta/\sqrt{t}$ and $w = t$.

3.3 Scenario 3: $\eta_t = \eta/\sqrt{t}$ and w is constant

Similar to 3.2, we can write:

$$\left\| \frac{1}{w} \sum_{s=1}^{t} \nabla f_s(x_s) \right\|^2 \leq \left(\frac{3M^2}{2\eta} + 2\eta G^2\right)^2 \frac{t}{w^2}. \tag{14}$$

Summing this result across t yields:

$$PR_w(T) = \sum_{t=1}^{T} \left\| \frac{1}{w} \sum_{s=1}^{t} \nabla f_s(x_s) \right\|^2 \quad \leq \quad \left(\frac{3M^2}{2\eta} + 2\eta G^2\right)^2 \frac{1}{w^2} \sum_{t=1}^{T} t \tag{15}$$

$$= \quad \left(\frac{3M^2}{2\eta} + 2\eta G^2\right)^2 \frac{T(T+1)}{2w^2} \tag{16}$$

which is quadratic in T but w can be selected accordingly to make the upper bound sub-linear.

4 Conclusion

We introduced a new definition of a local regret to study nonconvex problems in forecasting. We used the concept of a calibration and showed that our regret can be written as a local regret for the interior points in the feasible set. Our regret differs from Hazan's regret in the sense that it emphasizes today's reward as opposed to past reward. We also showed that our definition of regret has a logarithmic bound under some constraints. As a future direction, we plan to study the insights of our regret for the boundary points in the feasible set and propose efficient machine learning algorithms for nonconvex online learning that are optimal in terms of our definition of regret.

References

[1] Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pages 699–707, 2016.

[2] Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Conference on Learning Theory*, pages 779–806, 2014.

[3] Valentin Flunkert, David Salinas, and Jan Gasthaus. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *arXiv preprint arXiv:1704.04110*, 2017.

[4] Dean P Foster and Rakesh V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

[5] Elad Hazan, Karan Singh, and Cyril Zhang. Efficient regret minimization in non-convex games. *arXiv preprint arXiv:1708.00075*, 2017.

[6] Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

[7] Ruofeng Wen, Kari Torkkola, and Balakrishnan Narayanaswamy. A multi-horizon quantile recurrent forecaster. *arXiv preprint arXiv:1711.11053*, 2017.

[8] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 928–936, 2003.

5 Appendix

Lemma 5.1. $\eta_t \langle D_u(x_t), \nabla f_t(x_t) \rangle \geq \langle u_t - u_{t+1}, u \rangle + \frac{1}{2} \left(\|u_{t+1} - x_{t+1}\|^2 - \|u_t - x_t\|^2 \right) - \eta_t^2 G^2$ where $u_t \triangleq \text{proj}_{\mathcal{K}}(x_t + u)$, $\sup_{x \in \mathcal{K}, t \in \mathcal{T}} \nabla f_t(x) = G$ for any $u \in \mathcal{K}$ such that $x_t + u \in \mathcal{K}$.

Proof. Let $y_{t+1} \triangleq x_t - \eta_t \nabla f_t(x_t)$ and recall that $x_{t+1} = \text{proj}_{\mathcal{K}}(y_{t+1})$. Then we have:

$$
\begin{aligned}
\eta_t \langle D_u(x_t), \nabla f_t(x_t) \rangle &= \langle u_t - x_t, x_t - y_{t+1} \rangle = \langle u_t - x_t, x_t - x_{t+1} \rangle + \langle u_t - x_t, x_{t+1} - y_{t+1} \rangle \\
&= \langle u_t - x_t, x_t - x_{t+1} \rangle + \langle u_t - x_{t+1}, x_{t+1} - y_{t+1} \rangle + \langle x_{t+1} - x_t, x_{t+1} - y_{t+1} \rangle \\
&\geq \langle u_t - x_t, x_t - x_{t+1} \rangle + \langle x_{t+1} - x_t, x_{t+1} - y_{t+1} \rangle
\end{aligned}
\tag{17}
$$

The inequality in 17 can be justified by geometrical interpretation of projections as shown in Figure 1(a).
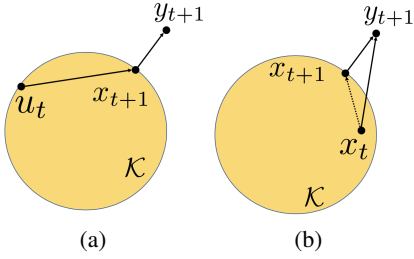

Figure 1: (a) Geometrical justification for inequality 17. The angle between $u_t - x_{t+1}$ and $x_{t+1} - y_{t+1}$ is always less than or equal to $\pi/2$; hence $\langle u_t - x_{t+1}, x_{t+1} - y_{t+1} \rangle \geq 0$ for all $u \in \mathbb{R}^d$. (b) Due to the triangle inequality, $\|x_t - x_{t+1}\| \leq \|x_t - y_{t+1}\| = \eta_t \nabla f_t(x_t)$. Hence $\langle x_t - x_{t+1}, \nabla f_t(x_t) \rangle \leq \eta_t \|\nabla f_t(x_t)\|^2 \leq \eta_t G^2$.

Plugging $y_{t+1} = x_{t+1} - \eta_t \nabla f_t(x_t)$, we have:

$$
\begin{aligned}
\eta_t \langle D_u(x_t), \nabla f_t(x_t) \rangle &= \langle u_t - x_t, x_t - x_{t+1} \rangle + \langle x_t - x_{t+1}, x_t - \eta_t \nabla f_t(x_t) - x_{t+1} \rangle \\
&= \langle u_t - x_t, x_t - x_{t+1} \rangle + \|x_t - x_{t+1}\|^2 - \eta_t \langle \nabla f_t(x_t), x_t - x_{t+1} \rangle \\
&\geq \langle u_t - x_t, x_t - x_{t+1} \rangle + \|x_t - x_{t+1}\|^2 - \eta_t^2 G^2
\end{aligned}
\tag{18}
$$

Inequality 18 is a result of triangle inequality as drawn in Figure 1(b). Using the fact that $\|u_t - x_t + x_t - x_{t+1}\|^2 = \|u_t - x_t\|^2 + 2 \langle u_t - x_t, x_t - x_{t+1} \rangle + \|x_t - x_{t+1}\|^2$ in equation 18 , we can write:

$$
\begin{aligned}
\eta_t \langle D_u(x_t), \nabla f_t(x_t) \rangle &\geq \frac{1}{2} \left(\|u_t - x_{t+1}\|^2 - \|u_t - x_t\|^2 + \|x_{t+1} - x_t\|^2 \right) - \eta_t^2 G^2 \tag{19} \\
&= \langle u_t - u_{t+1}, u_{t+1} - x_{t+1} \rangle + \frac{1}{2} \left(\|u_{t+1} - x_{t+1}\|^2 - \|u_t - x_t\|^2 \right) \\
&\quad + \frac{1}{2} \left(\|u_{t+1} - u_t\|^2 + \|x_{t+1} - x_t\|^2 \right) - \eta_t^2 G^2 \tag{20}
\end{aligned}
$$

where equation 20 is a result of $\|u_t - u_{t+1} + u_{t+1} - x_{t+1}\|^2 = \|u_t - u_{t+1}\|^2 + 2 \langle u_t - u_{t+1}, u_{t+1} - x_{t+1} \rangle + \|u_{t+1} - x_{t+1}\|^2$. By rewriting $\langle u_t - u_{t+1}, u_{t+1} - x_{t+1} \rangle$ as $\langle u_t - u_{t+1}, u_{t+1} - x_{t+1} - u + u \rangle$, we get:

$$
\begin{aligned}
\eta_t \langle D_u(x_t), \nabla f_t(x_t) \rangle &\geq \langle u_t - u_{t+1}, u \rangle + \langle u_t - u_{t+1}, u_{t+1} - (x_{t+1} + u) \rangle \\
&\quad + \frac{1}{2} \left(\|u_{t+1} - x_{t+1}\|^2 - \|u_t - x_t\|^2 \right) \\
&\quad + \frac{1}{2} \left(\|u_{t+1} - u_t\|^2 + \|x_{t+1} - x_t\|^2 \right) - \eta_t^2 G^2 \tag{21}
\end{aligned}
$$

Note that by replacing x_{t+1} with u_{t+1} and y_{t+1} with $x_{t+1} + u$ in Figure 1(a), we can see that $\langle u_t - u_{t+1}, u_{t+1} - (x_{t+1} + u) \rangle \geq 0$. Since $\frac{1}{2} \left(\|u_{t+1} - u_t\|^2 + \|x_{t+1} - x_t\|^2 \right) \geq 0$, we get:

$$
\eta_t \langle D_u(x_t), \nabla f_t(x_t) \rangle \geq \langle u_t - u_{t+1}, u \rangle + \frac{1}{2} \left(\|u_{t+1} - x_{t+1}\|^2 - \|u_t - x_t\|^2 \right) - \eta_t^2 G^2. \tag{22}
$$

$\square$

Proof of Theorem 3.1 :
As a result of lemma 5.1, we can write the following inequality:

$$\sum_{s=t-w+1}^{t} \langle D_u(x_s), \nabla f_s(x_s) \rangle \geq \sum_{s=t-w+1}^{t} \frac{1}{\eta_s} \langle u_s - u_{s+1}, u \rangle - \sum_{s=t-w+1}^{t} \eta_s G^2$$

$$+ \sum_{s=t-w+1}^{t} \frac{1}{2\eta_s} \left(\|u_{s+1} - x_{s+1}\|^2 - \|u_s - x_s\|^2 \right)$$

The first term can be rewritten as

$$\sum_{s=t-w+1}^{t} \frac{1}{\eta_s} \langle u_s - u_{s+1}, u \rangle = \sum_{s=t-w+1}^{t} \frac{\sqrt{s}}{\eta} \langle u_s - u_{s+1} - x + x, u \rangle \tag{23}$$

$$= \sum_{s=t-w+1}^{t} \frac{\sqrt{s}}{\eta} \langle u_s - x, u \rangle - \sum_{s=t-w+1}^{t} \frac{\sqrt{s}}{\eta} \langle u_{s+1} - x, u \rangle$$

$$= \frac{t-w+1}{\eta} \langle u_{t-w+1} - x, u \rangle - \frac{\sqrt{t}}{\eta} \langle u_{t+1} - x, u \rangle$$

$$+ \frac{1}{\eta} \sum_{s=t-w+2}^{t} \left(\sqrt{s} - \sqrt{s-1} \right) \langle u_s - x, u \rangle \tag{24}$$

The bound for the second term can be written as:

$$-\eta G^2 \sum_{s=t-w+1}^{t} \frac{1}{\sqrt{s}} \geq \eta G^2 \left(2\sqrt{t-w+1} - 2\sqrt{t} \right) \tag{25}$$

7

as a result of $\sum_{s=t-w+1}^{t} \frac{1}{\sqrt{s}} \leq \int_{t-w+1}^{t} \frac{1}{\sqrt{s}} ds = 2\sqrt{t} - 2\sqrt{t-w+1}$. The bound for the third term can be rewritten as:

$$
\begin{aligned}
\sum_{s=t-w+1}^{t} \frac{1}{2\eta_s} \left(\|u_{s+1} - x_{s+1}\|^2 - \|u_s - x_s\|^2 \right) &= \sum_{s=t-w+1}^{t} \frac{\sqrt{s}}{2\eta} \left(\|u_{s+1} - x_{s+1}\|^2 - \|u_s - x_s\|^2 \right) \\
&= \frac{\sqrt{t}}{2\eta} \|u_{t+1} - x_{t+1}\|^2 - \frac{\sqrt{t-w+1}}{2\eta} \|u_{t-w+1} - x_{t-w+1}\|^2 \\
&\quad - \frac{1}{2\eta} \sum_{s=t-w+2}^{t} \left(\sqrt{s} - \sqrt{s-1} \right) \|u_s - x_s\|^2 \quad (26) \\
&\geq \frac{\sqrt{t}}{2\eta} \|u_{t+1} - x_{t+1}\|^2 - \frac{\sqrt{t-w+1}}{2\eta} \|u_{t-w+1} - x_{t-w+1}\|^2 \\
&\quad - \frac{1}{2\eta} \sum_{s=t-w+2}^{t} (\sqrt{s} - \sqrt{s-1}) \|u_s - x_s\|^2 \quad (27) \\
&\geq \frac{\sqrt{t}}{2\eta} \|u_{t+1} - x_{t+1}\|^2 - \frac{\sqrt{t-w+1}}{2\eta} \|u_{t-w+1} - x_{t-w+1}\|^2 \\
&\quad - \frac{M^2}{2\eta} \underbrace{\sum_{s=t-w+2}^{t} (\sqrt{s} - \sqrt{s-1})}_{\sqrt{t} - \sqrt{t-w+1}} \quad (28) \\
&= \frac{\sqrt{t}}{2\eta} \|u_{t+1} - x_{t+1}\|^2 - \frac{\sqrt{t-w+1}}{2\eta} \|u_{t-w+1} - x_{t-w+1}\|^2 \\
&\quad - \frac{M^2\sqrt{t}}{2\eta} + \frac{M^2\sqrt{t-w+1}}{2\eta} \quad (29) \\
&\geq \frac{\sqrt{t}}{2\eta} \|u_{t+1} - x_{t+1}\|^2 - \frac{\sqrt{t-w+1}}{2\eta} M^2 \\
&\quad - \frac{M^2\sqrt{t}}{2\eta} + \frac{M^2\sqrt{t-w+1}}{2\eta} \quad (30) \\
&= \frac{\sqrt{t}}{2\eta} \|u_{t+1} - x_{t+1}\|^2 - \frac{M^2\sqrt{t}}{2\eta} \geq -\frac{M^2\sqrt{t}}{2\eta} \quad (31)
\end{aligned}
$$

where equation 27 is a result of $\sup_{x,y\in\mathcal{K}} \|x-y\| = M$. Hence, we have:

$$
\begin{aligned}
\sum_{s=t-w+1}^{t} \langle D_u(x_s), \nabla f_s(x_s) \rangle &\geq \frac{t-w+1}{\eta} \langle u_{t-w+1} - x, u \rangle - \frac{\sqrt{t}}{\eta} \langle u_{t+1} - x, u \rangle \\
&\quad + \frac{1}{\eta} \sum_{s=t-w+2}^{t} (\sqrt{s} - \sqrt{s-1}) \langle u_s - x, u \rangle \\
&\quad - \frac{M^2\sqrt{t}}{2\eta} + \eta G^2 \left(2\sqrt{t-w+1} - 2\sqrt{t} \right) \quad (32) \\
& \quad (33)
\end{aligned}
$$

now, let's explore the bound for $\langle u_t - x, u \rangle$ for any $x \in \mathcal{K}$. By definition of u_t, we can write:

$$
\begin{aligned}
\|x_t + u - x\|^2 &\geq \|x_t + u - u_t\|^2 \quad (34) \\
&= \|x_t + u - x\|^2 + \|x - u_t\|^2 + 2\langle x_t + u - x, x - u_t \rangle \quad (35) \\
&= \|x_t + u - x\|^2 + \|x - u_t\|^2 + 2\langle x_t - x, x - u_t \rangle + 2\langle u, x - u_t \rangle \quad (36) \\
&\geq \|x_t + u - x\|^2 - 2M^2 + 2\langle u, x - u_t \rangle. \quad (37)
\end{aligned}
$$

Hence, $\langle u, u_t - x \rangle \geq -M^2$. Taking $x = u_{t+1}$ and combining 32 and 37, we get:

$$\sum_{s=t-w+1}^{t} \langle D_u(x_s), \nabla f_s(x_s) \rangle \geq -\frac{\sqrt{t-w+1}}{\eta} M^2 - \left(\frac{\sqrt{t} - \sqrt{t-w+1}}{\eta} \right) M^2 - \left(\frac{M^2}{2\eta} + 2\eta G^2 \right) \sqrt{t}$$

$$= 2\eta G^2 \sqrt{t-w+1} - \left(\frac{3M^2}{2\eta} + 2\eta G^2 \right) \sqrt{t} \qquad (38)$$