Convergence Acceleration of Hypergradient Descent under Self-Bounded Smoothness

Yifa Yu[*,1] **Zhen Bao**[†,1]

Mentor: Wenzhi Gao[§ 1], Madeleine Udell[‡ 1,2]
Project Idea: Wenzhi Gao[§ 1]

[1]Institute for Computational and Mathematical Engineering (ICME)
Stanford University

[2]Department of Management Science and Engineering (MS&E)
Stanford University

[*]yifacop0@stanford.edu
[†]zbao7@stanford.edu
[‡]udell@stanford.edu
[§]gwz@stanford.edu

Abstract

Recent results have shown that Monotone Online Scaled Gradient Methods with Hypergradients (OSGM-H) achieve trajectory-based convergence under general convexity conditions, although the source of their accelerated behavior remains not fully understood. In contrast, a vanilla hypergradient descent method (which unconditionally accepts each proposed update) can diverge on convex problems without global smoothness (for example, $f(x) = x^4$), since a global L-smoothness assumption becomes the only effective safeguard against the stepsize growing without bound. The monotone OSGM-H algorithm addresses this issue by augmenting hypergradient descent with a null-step safeguard: the current iterate x_k is held fixed while the stepsize α_k continues to be adjusted until a sufficiently small, safe value is identified. We formally analyze this monotone OSGM-H scheme and show that the null-step mechanism enforces monotonic descent and keeps all iterates within a bounded level set. This confinement prevents gradient explosions and induces an effective local L-smoothness along the trajectory, even when the objective function lacks a global Lipschitz constant. Moreover, under the self-boundedness condition (where the local curvature vanishes as the optimum is approached), the hypergradient adaptation causes the stepsize to grow unboundedly, allowing OSGM-H to exploit the increasingly flat geometry near the minimizer. We prove that this adaptive behavior yields an asymptotic acceleration beyond the classical $O(1/k)$ rate of first-order methods – in fact, monotone OSGM-H achieves a strictly faster $o(1/k)$ convergence rate under self-boundedness. Finally, we validate our theoretical findings with experiments on both synthetic and real-world tasks. In particular, on a logistic regression problem from the MNIST dataset (which does not satisfy global L-smoothness and exhibits vanishing curvature near optimality), OSGM-H outperforms standard methods (such as GD, Adam, and AdaGrad), confirming that the null-step hypergradient strategy robustly accelerates convergence on objectives with flat local geometry.

Contents

Abstract **1**

1 Introduction **3**
1.1 Problem Background 3
1.2 Previous work 3
1.3 Contributions 4

2 Convergence Behavior of Monotone OSGM-H under Self-Boundedness **4**
2.1 Notations 4
2.2 Monotone OSGM-H Algorithm 4
2.3 Mathematical Background 5
2.4 Convergence Analysis 6

3 Experiments & Numerical Results **8**

4 Future Work **10**

5 Conclusion **11**

Appendix **13**

A Proof of results in Section 2 **13**
A.1 Proof of Theorem 2 13
A.2 Proof of Lemma 1 14
A.3 Proof of Lemma 2 15
A.4 Proof of Lemma 3 16
A.5 Proof of Lemma 4 16

B Code for Numerical Results **18**

C AI Contribution **18**

1 Introduction

1.1 Problem Background

Gradient descent remains a fundamental tool in modern optimization, but its performance is highly sensitive to the choice of stepsize. Classical strategies, such as fixed stepsizes or global Lipschitz constants, often struggle on problems whose smoothness and curvature vary across the landscape, leading to slow convergence or overly conservative steps.

Online Scaled Gradient Methods with Hypergradients (OSGM–H) provide a principled framework for learning stepsizes using an online update rule. The method consists of two components: a scheduler that proposes a stepsize based on hypergradient feedback, and a landscape action that decides whether the proposed update is accepted.

In the monotone variant of OSGM–H, the landscape action is chosen to be a null–step rule that enforces monotonicity of the objective values and keeps all iterates inside a bounded level set. As a consequence, the algorithm enjoys an *effective* smoothness property during the run, even when f itself is not globally smooth.

In this work, we focus on objective functions that are globally L–smooth, but whose *local* smoothness improves as the iterates approach the minimizer. Intuitively, when the curvature decays near optimality—equivalently, when the local Lipschitz constant of the gradient becomes smaller—the hypergradient mechanism of OSGM–H is able to automatically increase its stepsize. Therefore, as the iterates move closer to $x^\star$, the algorithm takes increasingly aggressive steps, in contrast with classical gradient descent whose stepsizes remain limited by a global smoothness bound.

This motivates the central question of our analysis:

> *Can monotone OSGM–H exploit vanishing curvature to surpass the classical $\mathcal{O}(1/k)$ convergence under global smoothness assumptions?*

The monotone version of OSGM–H therefore offers a framework for studying hypergradient–based stepsize adaptation beyond standard smoothness assumptions and forms the basis of our analysis in the sequel.

1.2 Previous work

Recent results have established global convergence of monotone Online Scaled Gradient Methods with Hypergradients (OSGM–H) for convex objectives, relying on the null–step mechanism to maintain bounded level sets and ensure monotone descent along the iterates [1, 2, 3]. These works clarify the behavior of the hypergradient scheduler and confirm that stepsizes can be adapted efficiently from local information observed along the trajectory, and they establish the classical $\mathcal{O}(1/K)$ convergence rate.

However, existing analyses focus primarily on establishing convergence under classical smoothness assumptions, without addressing whether hypergradient-based methods can benefit from structural properties of the objective to achieve faster rates. In particular, no prior result shows that the decay of local curvature near the optimum can be leveraged to obtain rates beyond the standard $\mathcal{O}(1/K)$ behavior.

1.3 Contributions

This paper shows that monotone hypergradient descent achieves a strictly faster rate under the self–boundedness property, namely an $o(1/K)$ convergence of the objective values. To the best of our knowledge, this is the first result that goes beyond the $\mathcal{O}(1/K)$ rate under self–boundedness property.

2 Convergence Behavior of Monotone OSGM-H under Self-Boundedness

In this section we analyze the convergence behavior of the monotone Online Scaled Gradient Method with Hypergradients (OSGM–H) under the self–boundedness assumption. Building on the monotone mechanism discussed in Section 1.2, we move beyond global convergence and focus on the effect of vanishing local curvature near the optimal set.
Our goal is to show that, when the local Lipschitz constant of the gradient decreases along the trajectory, monotone OSGM–H automatically adapts its stepsizes and achieves an accelerated $o(1/K)$ convergence rate of the objective values.

2.1 Notations

We use $\|\cdot\|$ to denote vector Euclidean norm or matrix operator norm and $\langle\cdot,\cdot\rangle$ to denote Euclidean or Frobenius inner product. Given a vector $d \in \mathbb{R}^n$, $\mathrm{Diag}(d)$ denotes the diagonal matrix with elements of d on its diagonal. We use $\mathcal{X}^\star = \{x : f(x) = f^\star\}$ to denote the optimal set of f. A function f is L-smooth (has L-Lipschitz continuous gradient) if it satisfies $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$ for all $x, y \in \mathbb{R}^n$; a function f has H-Lipschitz Hessian if $\|\nabla^2 f(x) - \nabla^2 f(y)\| \le H\|x - y\|$ for all $x, y \in \mathbb{R}^n$. We use superscript x^k to index algorithm iterates and subscript α_k to index stepsize sequence.

2.2 Monotone OSGM-H Algorithm

In OSGM-H, the stepsize is treated as a learnable parameter and updated using a *hypergradient feedback* $h_{x_k}(\alpha_k)$ at each iteration [1, 2]. For a scalar stepsize α, the feedback is defined by the normalized function decrease

$$h_x(\alpha) := \frac{f(x - \alpha\nabla f(x)) - f(x)}{\|\nabla f(x)\|^2}, \qquad h'_x(\alpha) = -\frac{\big\langle \nabla f\big(x - \alpha\nabla f(x)\big), \nabla f(x)\big\rangle}{\|\nabla f(x)\|^2},$$

and the stepsize update is an online gradient step $\alpha_{k+1} = \alpha_k - \eta_k\, h'_{x_k}(\alpha_k)$.
At each iteration, a tentative point is proposed by $x_{k+\frac{1}{2}} = x_k - \alpha_k \nabla f(x_k)$. The *monotone* OSGM-H variant applies a null-step rule,

$$x_{k+1} = \begin{cases} x_{k+\frac{1}{2}}, & f(x_{k+\frac{1}{2}}) \le f(x_k), \\ x_k, & \text{otherwise}, \end{cases}$$

which guarantees $f(x_{k+1}) \le f(x_k)$ and therefore keeps all iterates inside the initial level set, preventing gradient explosion even when f is not globally L-smooth.

Algorithm 1 Monotone OSGM-H (with scalar stepsize α_k)

Require: Initial point x^1, initial stepsize $\alpha_1 \in \mathbb{R}$, online learning stepsize schedule $\eta > 0$

1: **for** $k = 1, 2, \ldots$ **do**
2: $\quad x^{k+\frac{1}{2}} = x^k - \alpha_k \nabla f(x^k)$
3: $\quad$ Choose x^{k+1} satisfying

$$x^{k+1} \in \arg \min_{z \in \{x^{k+\frac{1}{2}}, x^k\}} f(z)$$

4: $\quad \alpha_{k+1} = \alpha_k - \eta\, h'_{x^k}(\alpha_k)$
5: **end for**

2.3 Mathematical Background

We consider the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

where f is convex and twice differentiable with minimizer set $\mathcal{X}^\star := \{x : f(x) = f^\star\} \neq \emptyset$.

Definition 1 (Self–bounded loss) *A twice differentiable convex function f is said to be* self–bounded *if its local smoothness satisfies*

$$\|\nabla^2 f(x)\| \longrightarrow 0 \qquad \text{as } f(x) \to f(x^\star).$$

Throughout this section we impose the following assumptions:

- f is convex, twice differentiable and $\mathcal{X}^\star \neq \emptyset$;
- f is globally L–smooth;
- f is self–bounded in the sense of Definition 1.

Definition 2 (Local smoothness on level sets) *For $\theta > 0$, define the level set*

$$D_\theta := \{x \in \mathbb{R}^n : \ f(x) - f^\star \leq \theta\}.$$

We say that f is locally L_θ–smooth on D_θ if

$$\|\nabla f(x) - \nabla f(y)\| \ \leq \ L_\theta \, \|x - y\| \qquad \text{for all } x, y \in D_\theta.$$

Under the self–boundedness assumption, each L_θ exists and satisfies

$$L_\theta \ \longrightarrow \ 0 \qquad \text{as } \theta \downarrow 0.$$

We remark that global L–smoothness trivially implies this local level–set smoothness condition, and corresponds to the "glocal" framework of [8].
Under these conditions, every sublevel set D_θ admits a local smoothness constant L_θ which tends to zero as $\theta \downarrow 0$.

2.4 Convergence Analysis

We start from the following reduction theorem for the normalized decrease

$$h_k := \frac{f(x^{k+1}) - f(x^k)}{\|\nabla f(x^k)\|^2},$$

established in [2].

Theorem 1 (Reductions for h_k [2, Theorem 4.2]) *Let $\{x^k\}$ be any sequence such that $x^k \notin \mathcal{X}^\star$ and $f(x^{k+1}) \le f(x^k)$ for all k. If f is convex, then*

$$f(x^{K+1}) - f^\star \;\le\; \min\left\{ \frac{\Delta^2}{K}\,\frac{1}{\frac{1}{K}\sum_{k=1}^{K}(-h_k)},\;\; f(x^1) - f^\star \right\},$$

where

$$\Delta \;:=\; \max_{x:\; f(x)\le f(x^1)}\; \min_{x^\star\in\mathcal{X}^\star} \|x - x^\star\|.$$

In the monotone OSGM–H algorithm, the landscape action is a null–step rule, that is,

$$x^{k+1} \;\in\; \arg\min_{z\in\{x^k,\; x^k-\alpha_k\nabla f(x^k)\}} f(z),$$

and therefore

$$h_k \;=\; \frac{f(x^{k+1}) - f(x^k)}{\|\nabla f(x^k)\|^2} \;\le\; \min\{\, h_{x_k}(P_k),\, 0\,\}.$$

Thus the conditions of Theorem 1 are automatically satisfied by the iterates of monotone OSGM–H. To exploit this reduction, we next study how the scalar stepsizes α_k evolve under the hypergradient update.

Since f is globally L–smooth, the one–dimensional map $\alpha \mapsto h_{x_k}(\alpha)$ is also L–smooth on $\mathbb{R}$. Moreover, in monotone OSGM–H we always have

$$h_k \;\le\; \min\{h_{x_k}(\alpha_k),\, 0\},$$

because x^{k+1} is selected as the better point between x^k and $x^k - \alpha_k\nabla f(x_k)$. In particular, $h_k \le 0$ for all k, and hence the assumptions of Theorem 1 are automatically satisfied by the iterates of monotone OSGM–H.

To exploit this reduction, we next examine how the scalar stepsizes α_k evolve under the hypergradient update.

Theorem 2 (One–step descent inequality for α_k) *Fix $k \ge 0$ and let*

$$\alpha_{k+1} = \alpha_k - \eta\, h'_{x_k}(\alpha_k), \qquad 0 < \eta < 1/L.$$

Then for every fixed $\hat\alpha \ge 0$,

$$(\alpha_{k+1} - \hat\alpha)^2 \;\le\; (\alpha_k - \hat\alpha)^2 - 2\eta\, h_{x_k}(\alpha_k) + 2\eta\, h_{x_k}(\hat\alpha) + \frac{\eta^2}{1-\eta L}\,\|h'_{x_k}(\hat\alpha)\|^2. \tag{1}$$

Equivalently,

$$h_{x_k}(\alpha_k) \;\le\; \frac{(\alpha_k - \hat\alpha)^2 - (\alpha_{k+1} - \hat\alpha)^2}{2\,\eta} \;+\; h_{x_k}(\hat\alpha) \;+\; \frac{\eta}{2(1-\eta L)}\,\big\|h'_{x_k}(\hat\alpha)\big\|^2. \tag{2}$$

Moreover, as the iterates x_k approach the optimal set $\mathcal{X}^\star$, the self–boundedness of f allows us to replace the global smoothness constant L by a smaller *local* smoothness parameter. Indeed, on any sublevel set the local Lipschitz constant of the gradient becomes strictly smaller and tends to zero as $x_k \to \mathcal{X}^\star$. Consequently, the one–step relation above becomes progressively tighter along the trajectory, indicating that the hypergradient update may take more aggressive steps when the curvature is small.

An immediate question is whether the sequence of learned stepsizes (α_k) remains bounded. In monotone OSGM–H we have $h_{x_k}(0) = 0$ and $h'_{x_k}(0) = -1$ by construction, so the hypergradient initially points in an increasing direction. Combined with the vanishing local curvature implied by self–boundedness, this prevents the stepsizes from stabilizing at any finite upper bound.
To bound the right–hand side of Theorem 2, we first show that any fixed reference stepsize yields a strict local descent whenever x_k is sufficiently close to the minimizer set.

Lemma 1 (Local descent near a minimizer) *Let f be convex, twice differentiable, and self–bounded. Fix any $p > 0$. Then there exists $\theta > 0$ such that for all x with* $\operatorname{dist}(x, \mathcal{X}^\star) \le \theta$,

$$f(x - p\nabla f(x)) \;\le\; f(x) - \tfrac{p}{2}\|\nabla f(x)\|^2 \;\le\; f(x),$$

and in particular $f(x - p\nabla f(x)) < f(x)$ whenever $\nabla f(x) \neq 0$.

Having established this, we next show that the learned stepsizes cannot remain bounded from above. The following lemma formalizes this fact.

Lemma 2 (Stepsizes cannot remain bounded) *Let (x_k, α_k) be generated by monotone OSGM–H with f convex, and self–bounded. Then the sequence (α_k) cannot remain bounded from above.*

This observation provides a key mechanism behind the acceleration phenomenon: as the local curvature vanishes near $\mathcal{X}^\star$, the hypergradient rule automatically drives the stepsizes upward, enabling progressively more aggressive descent steps.
To prepare for the asymptotic argument, it is convenient to record the finite–horizon summation of (2).

Lemma 3 (Summed bound for the learned stepsizes) *Under the assumptions of Theorem 2, fix any $\hat{\alpha} \ge 0$ and let $\eta \in (0, 1/L)$. Then for every integer $K \ge 1$,*

$$\sum_{k=1}^{K} h_{x_k}(\alpha_k) \;\le\; \frac{(\alpha_1 - \hat{\alpha})^2}{2\eta} + \sum_{k=1}^{K} h_{x_k}(\hat{\alpha}) + \frac{\eta}{2(1-\eta L)} \sum_{k=1}^{K} \big\|h'_{x_k}(\hat{\alpha})\big\|^2. \tag{3}$$

Having established that any fixed reference stepsize induces a strict local descent whenever x_k is sufficiently close to the minimizer set, we next use this fact to control the average of the normalized decreases $h_{x_k}(\alpha_k)$. Recall that Theorem 2 gives a one–step relation in which the leading term telescopes under summation and the remaining terms are evaluated at a fixed reference stepsize. Since $x_k \to \mathcal{X}^\star$ while $x_k \notin \mathcal{X}^\star$ for all k, Lemma 1 guarantees a uniform amount of decrease whenever α_k is compared against any fixed $\hat{\alpha} > 0$. This leads to the following quantitative bound.

Lemma 4 (Asymptotic average decrease) *Let $\{(x_k, \alpha_k)\}$ be generated by monotone OSGM–H with f globally L–smooth and self–bounded, and suppose that $x_k \to \mathcal{X}^\star$ while $x_k \notin \mathcal{X}^\star$ for all k. Then for any fixed $\hat{\alpha} > 0$,*

$$\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \left\{ h_{x_k}(\hat{\alpha}) \; + \; \frac{\eta}{2(1-\eta L)} \left\| h'_{x_k}(\hat{\alpha}) \right\|^2 \right\} \; \leq \; -\frac{\hat{\alpha}}{2}.$$

Convergence rate. We now combine the previous ingredients to obtain the final rate. Recall from Theorem 1 that the optimality gap is controlled by the reciprocal of the averaged normalized decrease $(-h_k)$. In particular,

$$f(x^{K+1}) - f^\star \; \leq \; \frac{\Delta^2}{K} \cdot \frac{1}{\frac{1}{-K}\sum_{k=1}^{K}\left(\frac{(\alpha_1-\hat{\alpha})^2}{2\eta} + \sum_{k=1}^{K} h_{x_k}(\hat{\alpha}) + \frac{\eta}{2(1-\eta L)} \sum_{k=1}^{K} \left\| h'_{x_k}(\hat{\alpha}) \right\|^2\right)}. \tag{4}$$

On the other hand, Lemma 3 shows that the finite–horizon sum $\sum_{k=1}^{K} h_{x_k}(\alpha_k)$ can be bounded above by quantities evaluated at any fixed reference stepsize $\hat{\alpha} > 0$. Substituting this into (4) yields a uniform bound on $f(x^{K+1}) - f^\star$ for every $\hat{\alpha} > 0$.

As $K \to \infty$, Lemma 4 implies that the averaged normalized decrease satisfies

$$\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} (-h_k) \; \geq \; \frac{\hat{\alpha}}{2},$$

again for every fixed $\hat{\alpha} > 0$. Hence

$$\lim_{K\to\infty} K\big(f(x^{K+1}) - f^\star\big) \; \leq \; \frac{2\Delta^2}{\hat{\alpha}}.$$

Finally, letting $\hat{\alpha} \to +\infty$ shows that the right–hand side vanishes, which proves that the optimality gap decays strictly faster than $1/K$, namely

$$f(x^K) - f^\star \; = \; o\left(\frac{1}{K}\right).$$

This completes the convergence analysis of monotone OSGM–H. All detailed proofs are provided in Appendix 5.

3 Experiments & Numerical Results

We compare GD, Adam, AdaGrad, and the scalar OSGM–H on a family of least–squares–type problems and on logistic regression. All runs start from the zero vector and are run for $2 \cdot 10^3$ iterations. For the synthetic regression tasks we generate a matrix $A \in \mathbb{R}^{50\times 20}$ with condition number $\kappa(A) \approx 10^2$ and set $b = Ax_\star + 0.01\,\xi$ with $x_\star \sim \mathcal{N}(0, I)$ and Gaussian noise ξ of small variance. We then consider

$$f_p(x) = \|Ax - b\|_2^p, \qquad p \in \{2, 4, 6, 8\},$$

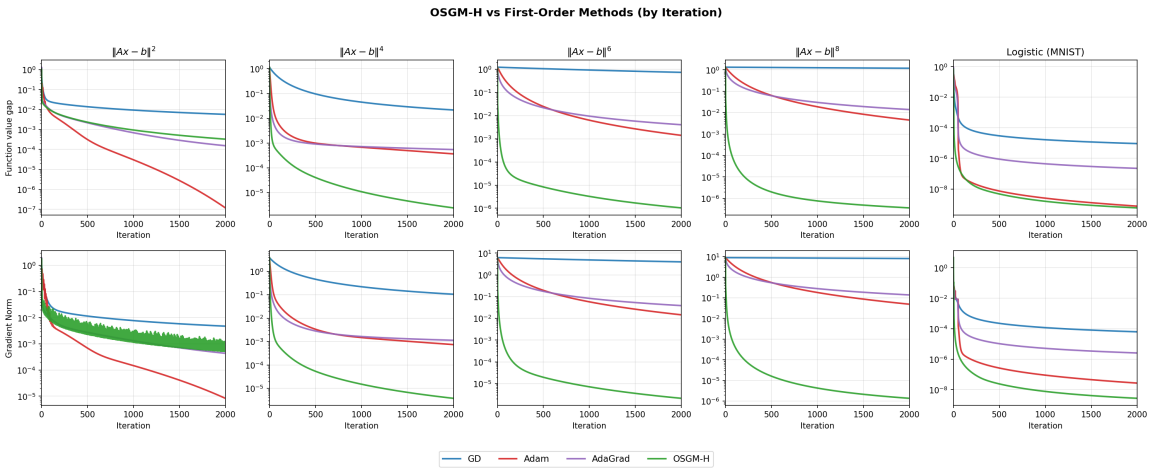

Figure 1: Convergence of GD, Adam, AdaGrad, and OSGM-H on the synthetic regression tasks and MNIST logistic regression. Top row: function value gap $f(x_k) - f^\star$ vs. iteration (log scale). Bottom row: gradient norm $\|\nabla f(x_k)\|$ vs. iteration (log scale).

so that $p = 2$ corresponds to the classical quadratic loss while $p > 2$ yields convex self–bounded objectives whose local smoothness vanishes at the minimizer. For the classification task we train an unregularized logistic regression model on the 0-vs-1 subset of MNIST, using 2 000 examples with standardized features. For each method and problem we record the function value gap $f(x_k) - f^\star$ (where $f^\star$ is the loss at the least–squares solution for the regression tasks and 0 for logistic regression) and the gradient norm $\|\nabla f(x_k)\|$. Figure 1 summarizes convergence as a function of iteration.

On the synthetic regression problems we minimize $f_p(x) = \|Ax-b\|_2^p$ for $p \in \{2, 4, 6, 8\}$ (Figure 1). For the quadratic case $p = 2$, Adam is the strongest baseline and OSGM-H is competitive but not clearly better, while GD and AdaGrad converge more slowly and plateau at larger function gaps and gradient norms. As p increases and the objectives become more self-bounded, OSGM-H progressively pulls ahead: for $p = 4$ it attains function gaps about one–two orders of magnitude smaller than Adam and AdaGrad after 2 000 iterations, and for $p = 6$ and $p = 8$ GD essentially stalls and Adam/AdaGrad level off around 10^{-2}–10^{-3}, whereas OSGM-H continues to decrease almost linearly on the log–scale, reaching gaps on the order of 10^{-6} and driving the gradient norm down by five–six orders of magnitude. This pattern is consistent with the theoretical prediction that HDM benefits most when the curvature vanishes near the optimum and no single global stepsize is adequate.

For logistic regression on MNIST (0 vs. 1), all methods eventually drive the training loss close to zero and achieve near-perfect training accuracy, but their convergence speeds differ significantly. Gradient Descent (GD) is consistently the slowest and remains two to three orders of magnitude above the best methods even after 2,000 iterations. AdaGrad performs better, yet still lags behind Adam and OSGM–H.
Adam and OSGM–H exhibit a very similar and steep initial decay, quickly reducing the loss below 10^{-6}. After this transient phase, OSGM–H continues to decrease slightly faster and eventually attains the lowest final loss among all tested methods.

The gradient norms display a consistent picture: OSGM–H achieves the sharpest decline, driving $\|\nabla f(x_k)\|$ down to about 10^{-8} by the end of the run, followed by Adam, AdaGrad, and GD. This indicates that the additional hypergradient update does not impair performance on a realistic, high-dimensional classification problem and can even yield a mild advantage over well–tuned adaptive baselines.
We remark that logistic regression does not satisfy our self-boundedness assumption in the strict sense; nonetheless, its curvature decreases as the loss approaches zero, and the observed behavior of OSGM–H remains consistent with our theoretical predictions.

These results empirically support our theoretical intuition that hypergradient methods benefit from improving local smoothness along the trajectory rather than from a single global stepsize. Although our logistic regression and $f_p(x) = \|Ax - b\|_2^p$ experiments are not globally L–smooth, the monotone OSGM–H null–step rule keeps iterates in bounded level sets, effectively inducing a local L_k–smooth region where the theory applies. We also note that vanilla hypergradient descent is known to diverge even on the simple function $f(x) = x^4$ [7], whereas the null–step mechanism in monotone OSGM–H prevents such instability in principle.

4 Future Work

Our analysis in Section 2.4 is carried out under the standing assumption that f is globally L–smooth. In particular, the one–step inequality of Theorem 2 is derived by applying the global smoothness of the one–dimensional map $\alpha \mapsto h_{x_k}(\alpha)$, which in turn relies on global L–smoothness of f. However, all of our experiments in Section 3 are performed on objectives that are *not* globally L–smooth, yet the empirical behavior of monotone OSGM–H is fully consistent with the $\mathcal{O}(1/K)$ convergence bound, and often exhibits the $o(1/K)$ acceleration predicted by our theory.
The real obstacle to dropping global smoothness is not the behavior of the normalized decreases $h_{x_k}(\alpha_k)$ themselves, but the lack of a priori control on the tentative point

$$x_{k+\frac{1}{2}} \;=\; x_k - \alpha_k \nabla f(x_k).$$

Without a global Lipschitz constant we cannot exclude the possibility that $x_{k+\frac{1}{2}}$ escapes to regions of very large curvature before the landscape rule has a chance to reject the step. In practice, however, we observe that after a short warm–up phase the monotone mechanism almost always produces proposals satisfying

$$f\big(x_k - \alpha_k \nabla f(x_k)\big) \;\le\; f(x_k),$$

so that $x_{k+\frac{1}{2}}$ already lies in a low–curvature region where local smoothness constants L_k are small. This suggests that the global L–smoothness assumption may be largely technical.
A promising direction is therefore to introduce an explicit *local clipping* rule for the stepsizes, for example

$$\alpha_k^{\text{clip}} \;:=\; \min\Big\{\alpha_k,\; \frac{C}{\|\nabla f(x_k)\|}\Big\},$$

for some large constant $C > 0$. Such a clipping does not prevent the learned stepsizes from diverging to $+\infty$, but it does guarantee that $\|x_{k+\frac{1}{2}} - x_k\| \leq C$ in the worst case. When the minimizer set $\mathcal{X}^\star$ is nonempty, this would ensure that all tentative points remain in a bounded neighborhood of $\mathcal{X}^\star$, so that the local smoothness constants L_k on the relevant level sets can be used in place of a global L in Theorem 2. On the other hand, if $\mathcal{X}^\star$ lies at infinity (for example in the logistic regression problems), the clipping forces the growth of $\|x_k\|$ to be at most linear in k, which may significantly slow down convergence along the "optimal" direction. A first open problem is to rigorously analyze the behavior of the tentative iterates $x_{k+\frac{1}{2}}$: we conjecture that, under self–boundedness and convexity, $x_{k+\frac{1}{2}}$ remains bounded, and possibly even converges to $\mathcal{X}^\star$. Establishing such a result would allow us to replace the global L–smoothness assumption by a purely local one and to use the vanishing level–set smoothness $L_k \to 0$ directly in Theorem 2.
A second direction is to go beyond the qualitative $o(1/K)$ statement and quantify how the convergence rate depends on the *degree* of self–boundedness. In this paper we only assume that the Hessian norm satisfies $\|\nabla^2 f(x)\| \to 0$ as $f(x) \downarrow f^\star$, which is enough to prove that the normalized decreases $(h_{x_k}(\alpha_k))$ have strictly negative Cesàro limit. For more structured classes of objectives, such as

$$\|\nabla^2 f(x)\|_2 \;\leq\; C\big(f(x) - f^\star\big)^p \qquad \text{for some } p > 0,$$

it is natural to ask whether the accelerated rate can be sharpened to an explicit $\mathcal{O}(1/K^q)$ bound with $q > 1$ depending on p. Characterizing this mapping $p \mapsto q$ would connect the qualitative notion of self–boundedness to quantitative convergence guarantees, and could help explain the dramatic empirical gains observed on highly flat objectives such as high–order polynomial losses and logistic regression.

5 Conclusion

This work presents a convergence rate analysis for monotone OSGM–H under convex objectives that are either globally L–smooth and self–bounded, or only locally smooth near the minimizer. We proved that under the global smoothness and the self-boundedness condition – where the objective's curvature gradually vanishes near the optimum – the hypergradient-driven update schedule causes the stepsize to increase without bound as the algorithm approaches the minimizer. This adaptive stepsize growth enables the algorithm to exploit the favorable local geometry, yielding an accelerated asymptotic convergence rate of $\mathrm{o}(1/K)$ that is strictly faster than the standard $O(1/K)$ rate of classical first-order methods. Our numerical experiments on both synthetic power-law regression problems ($f_p(x) = |Ax - b|_2^p$ for various p) and an unregularized MNIST logistic regression task empirically validate these theoretical results. In particular, OSGM-H consistently outperformed baseline optimizers (Gradient Descent, Adam, and AdaGrad) on objectives with nearly flat curvature near the optimum (as in the logistic regression example), achieving final objective values and gradient norms several orders of magnitude smaller than those attained by the alternatives. Collectively, our results demonstrate that hypergradient-based step

size adaptation is a principled and effective mechanism for leveraging local smoothness properties to accelerate convergence, all without relying on any global Lipschitz smoothness assumption.

References

[1] Chu, Y.-C., Gao, W., Ye, Y., & Udell, M. (2025). Provable and Practical Online Learning Rate Adaptation with Hypergradient Descent. *arXiv preprint arXiv:2405.15682*.

[2] Gao, W., Chu, Y.-C., Ye, Y., & Udell, M. (2025). Gradient Methods with Online Scaling, Part I: Theoretical Foundations. *arXiv preprint arXiv:2505.23081*.

[3] Chu, Y.-C., Gao, W., Ye, Y., & Udell, M. (2025). Gradient Methods with Online Scaling, Part II: Practical Aspects. *arXiv preprint arXiv:2509.11007*.

[4] Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12: 2121–2159.

[5] Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*.

[6] Srebro, N., Sridharan, K., & Tewari, A. (2010). Smoothness, Low Noise and Fast Rates. *Advances in Neural Information Processing Systems (NeurIPS)*.

[7] Martínez Rubio, D. (2017). Convergence Analysis of an Adaptive Method of Gradient Descent. MSc Dissertation, University of Oxford. `https://github.com/damaru2/convergence_analysis_hypergradient_descent/blob/master/dissertation_hypergradients.pdf`

[8] Fox, C., Mishkin, A., Vaswani, S., & Schmidt, M. (2025). Glocal Smoothness: Line Search can really help! *arXiv preprint arXiv:2506.12648*.

Appendix

A Proof of results in Section 2

A.1 Proof of Theorem 2

Proof 1 ***Step 1: Expand*** $(\alpha_{k+1}-\hat{\alpha})^2$***.*** *By the update rule* $\alpha_{k+1}=\alpha_k-\eta h'_{x_k}(\alpha_k)$*,*

$$\begin{aligned}(\alpha_{k+1}-\hat{\alpha})^2 &= \big(\alpha_k-\eta h'_{x_k}(\alpha_k)-\hat{\alpha}\big)^2\\ &= (\alpha_k-\hat{\alpha})^2-2\eta\, h'_{x_k}(\alpha_k)(\alpha_k-\hat{\alpha})+\eta^2\big(h'_{x_k}(\alpha_k)\big)^2. \end{aligned}\tag{5}$$

Step 2: Use L***–smoothness of*** h_{x_k} ***to control the cross term.*** *The* L*–smoothness of* h_{x_k} *implies the inequality:*

$$h_{x_k}(u)\ \geq\ h_{x_k}(v)+h'_{x_k}(v)(u-v)+\frac{1}{2L}\big\|h'_{x_k}(u)-h'_{x_k}(v)\big\|^2,\qquad \forall u,v.$$

Apply this with $u=\hat{\alpha}$ *and* $v=\alpha_k$*:*

$$h_{x_k}(\hat{\alpha})\ \geq\ h_{x_k}(\alpha_k)+h'_{x_k}(\alpha_k)(\hat{\alpha}-\alpha_k)+\frac{1}{2L}\big\|h'_{x_k}(\hat{\alpha})-h'_{x_k}(\alpha_k)\big\|^2.$$

Rearranging,

$$h'_{x_k}(\alpha_k)(\alpha_k-\hat{\alpha})\ \geq\ h_{x_k}(\alpha_k)-h_{x_k}(\hat{\alpha})+\frac{1}{2L}\big\|h'_{x_k}(\hat{\alpha})-h'_{x_k}(\alpha_k)\big\|^2.$$

Multiplying both sides by -2η *yields*

$$-2\eta\, h'_{x_k}(\alpha_k)(\alpha_k-\hat{\alpha})\ \leq\ -2\eta\big(h_{x_k}(\alpha_k)-h_{x_k}(\hat{\alpha})\big)-\frac{\eta}{L}\big\|h'_{x_k}(\hat{\alpha})-h'_{x_k}(\alpha_k)\big\|^2.\tag{6}$$

Step 3: Combine (5) ***and*** (6)***.*** *Substituting* (6) *into* (5) *gives*

$$\begin{aligned}(\alpha_{k+1}-\hat{\alpha})^2 &\leq (\alpha_k-\hat{\alpha})^2-2\eta\big(h_{x_k}(\alpha_k)-h_{x_k}(\hat{\alpha})\big)-\frac{\eta}{L}\big\|h'_{x_k}(\hat{\alpha})-h'_{x_k}(\alpha_k)\big\|^2+\eta^2\big\|h'_{x_k}(\alpha_k)\big\|^2\\ &= (\alpha_k-\hat{\alpha})^2-2\eta h_{x_k}(\alpha_k)+2\eta h_{x_k}(\hat{\alpha})+\eta^2\big\|h'_{x_k}(\alpha_k)\big\|^2-\frac{\eta}{L}\big\|h'_{x_k}(\hat{\alpha})-h'_{x_k}(\alpha_k)\big\|^2.\end{aligned}\tag{7}$$

Step 4: Eliminate $h'_{x_k}(\alpha_k)$ ***using Young's inequality.*** *Let*

$$a:=h'_{x_k}(\hat{\alpha}),\qquad b:=h'_{x_k}(\alpha_k).$$

Consider the term

$$\eta^2\|b\|^2-\frac{\eta}{L}\|a-b\|^2.$$

We expand

$$\begin{aligned}\eta^2\|b\|^2-\frac{\eta}{L}\|a-b\|^2 &= \eta^2\|b\|^2-\frac{\eta}{L}\big(\|a\|^2+\|b\|^2-2\langle a,b\rangle\big)\\ &= -\frac{\eta}{L}\|a\|^2+\Big(\eta^2-\frac{\eta}{L}\Big)\|b\|^2+\frac{2\eta}{L}\langle a,b\rangle.\end{aligned}$$

Apply Young's inequality $\frac{2\eta}{L}\langle a, b\rangle \le \frac{\eta}{L\gamma}\|a\|^2 + \frac{\eta\gamma}{L}\|b\|^2$ *with* $\gamma := 1 - \eta L \in (0, 1)$ *(since* $\eta < 1/L$*). Then*

$$\eta^2\|b\|^2 - \frac{\eta}{L}\|a - b\|^2 \le \Big(-\frac{\eta}{L} + \frac{\eta}{L\gamma}\Big)\|a\|^2 + \Big(\eta^2 - \frac{\eta}{L} + \frac{\eta\gamma}{L}\Big)\|b\|^2.$$

By the choice of γ *we have*

$$\eta^2 - \frac{\eta}{L} + \frac{\eta\gamma}{L} = \eta^2 - \frac{\eta}{L} + \frac{\eta(1 - \eta L)}{L} = 0,$$

so the $\|b\|^2$*–term vanishes, and*

$$\eta^2\|b\|^2 - \frac{\eta}{L}\|a - b\|^2 \le \frac{\eta^2}{1 - \eta L}\,\|a\|^2.$$

Returning to $a = h'_{x_k}(\hat{\alpha})$*, we obtain*

$$\eta^2\big\|h'_{x_k}(\alpha_k)\big\|^2 - \frac{\eta}{L}\big\|h'_{x_k}(\hat{\alpha}) - h'_{x_k}(\alpha_k)\big\|^2 \le \frac{\eta^2}{1 - \eta L}\,\big\|h'_{x_k}(\hat{\alpha})\big\|^2. \tag{8}$$

Step 5: Plug (8) ***into*** (7)***.*** *Substituting* (8) *into* (7) *yields*

$$(\alpha_{k+1} - \hat{\alpha})^2 \le (\alpha_k - \hat{\alpha})^2 - 2\eta h_{x_k}(\alpha_k) + 2\eta h_{x_k}(\hat{\alpha}) + \frac{\eta^2}{1 - \eta L}\,\big\|h'_{x_k}(\hat{\alpha})\big\|^2,$$

which is exactly (1)*. Rearranging* (1) *gives* (2)*, completing the proof.*

A.2 Proof of Lemma 1

Proof 2 *Fix any* $p > 0$*. Since* f *is convex and differentiable, every minimizer* $x^\star \in \mathcal{X}^\star$ *satisfies* $\nabla f(x^\star) = 0$*.*

Local smoothness near the minimizer set. *By self–boundedness,*

$$\operatorname{dist}(x, \mathcal{X}^\star) \to 0 \quad \Longrightarrow \quad \lambda_{\max}(\nabla^2 f(x)) \to 0.$$

Taking $\varepsilon := 1/p$*, there exists* $\theta_1 > 0$ *such that*

$$\operatorname{dist}(x, \mathcal{X}^\star) \le \theta_1 \implies \lambda_{\max}(\nabla^2 f(x)) \le \frac{1}{p},$$

hence f *is* L*–smooth with* $L = 1/p$ *on the neighborhood*

$$\mathcal{N} := \{x : \operatorname{dist}(x, \mathcal{X}^\star) \le \theta_1\}.$$

The descent segment stays inside the smooth region. *Since* $\nabla f(x^\star) = 0$ *and* ∇f *is continuous, we also have* $\|\nabla f(x)\| \to 0$ *as* $\operatorname{dist}(x, \mathcal{X}^\star) \to 0$*. Thus there exists* $\theta_2 > 0$ *such that*

$$\operatorname{dist}(x, \mathcal{X}^\star) \le \theta_2 \implies \|p\nabla f(x)\| \le \frac{\theta_1}{2}.$$

Let $\theta := \min\{\theta_1/2, \theta_2\}$. Then, if $\operatorname{dist}(x, \mathcal{X}^\star) \le \theta$, for any $t \in [0,1]$ we have

$$\operatorname{dist}(x - tp\nabla f(x), \mathcal{X}^\star) \le \operatorname{dist}(x, \mathcal{X}^\star) + t\|p\nabla f(x)\| \le \frac{\theta_1}{2} + \frac{\theta_1}{2} = \theta_1,$$

so the entire segment $x - tp\nabla f(x)$ lies in $\mathcal{N}$, where f is $1/p$–smooth.

Apply the descent lemma. *Because f is L–smooth with $L = 1/p$ on this segment,*

$$\begin{aligned} f(x - p\nabla f(x)) &\le f(x) - p\|\nabla f(x)\|^2 + \frac{L}{2}p^2\|\nabla f(x)\|^2 \\ &= f(x) - \frac{p}{2}\|\nabla f(x)\|^2. \end{aligned}$$

If $\nabla f(x) \neq 0$, the inequality is strict.

A.3 Proof of Lemma 2

Proof 3 *We argue by contradiction and assume that (α_k) is bounded from above: there exists $M > 0$ such that $\alpha_k \le M$ for all k.*
Recall that:

$$h_{x_k}(p) = \frac{f(x_k - p\nabla f(x_k)) - f(x_k)}{\|\nabla f(x_k)\|^2}, \qquad p \ge 0.$$

Let $g_k := \nabla f(x_k)$. By construction we have

$$h_{x_k}(0) = 0, \qquad h'_{x_k}(0) = -1.$$

By Lemma 1, the segment $\{x_k - tpg_k : t \in [0,1]\}$ lies in a region where f is locally smooth with parameter $1/p$ for all sufficiently large k.
Along this segment we have

$$0 \le h''_{x_k}(t) = \frac{g_k^\top \nabla^2 f(x_k - tg_k) g_k}{\|g_k\|^2} \le \frac{1}{p}, \qquad t \in [0, p].$$

Now fix an arbitrary $M > 0$ and set

$$p := 2M.$$

For all sufficiently large k (say $k \ge K := K(p)$) we have $\alpha_k \le M = p/2$, so

$$h'_{x_k}(\alpha_k) = h'_{x_k}(0) + \int_0^{\alpha_k} h''_{x_k}(t)\, dt \le -1 + \int_0^{\alpha_k} \frac{1}{p}\, dt = -1 + \frac{\alpha_k}{p} \le -1 + \frac{1}{2} = -\frac{1}{2}.$$

Hence for all $k \ge K$,

$$\alpha_{k+1} = \alpha_k - \eta h'_{x_k}(\alpha_k) \;\ge\; \alpha_k + \frac{\eta}{2}.$$

By induction, for every $n \ge 0$,

$$\alpha_{K+n} \;\ge\; \alpha_K + n\,\frac{\eta}{2},$$

which diverges to $+\infty$ as $n \to \infty$. In particular, once $k \ge K$ the sequence (α_k) is strictly increasing and cannot stay below the fixed bound M. Since $M > 0$ was arbitrary, this contradicts the assumption that (α_k) is bounded from above. Therefore the stepsizes cannot remain bounded from above.

A.4 Proof of Lemma 3

Proof 4 *Starting from the one–step inequality* (2) *in Theorem 2, we have for every* $k \geq 0$ *and every fixed* $\hat{\alpha} \geq 0$,

$$h_{x_k}(\alpha_k) \;\leq\; \frac{(\alpha_k - \hat{\alpha})^2 - (\alpha_{k+1} - \hat{\alpha})^2}{2\eta} + h_{x_k}(\hat{\alpha}) + \frac{\eta}{2(1-\eta L)} \big\| h'_{x_k}(\hat{\alpha}) \big\|^2.$$

Summing this inequality from $k = 1$ *to* K *gives*

$$\sum_{k=1}^{K} h_{x_k}(\alpha_k) \;\leq\; \frac{1}{2\eta} \sum_{k=1}^{K} \big[(\alpha_k - \hat{\alpha})^2 - (\alpha_{k+1} - \hat{\alpha})^2 \big] + \sum_{k=1}^{K} h_{x_k}(\hat{\alpha}) + \frac{\eta}{2(1-\eta L)} \sum_{k=1}^{K} \big\| h'_{x_k}(\hat{\alpha}) \big\|^2.$$

The first sum on the right telescopes:

$$\sum_{k=1}^{K} \big[(\alpha_k - \hat{\alpha})^2 - (\alpha_{k+1} - \hat{\alpha})^2 \big] = (\alpha_1 - \hat{\alpha})^2 - (\alpha_{K+1} - \hat{\alpha})^2 \;\leq\; (\alpha_1 - \hat{\alpha})^2.$$

Substituting this into the previous display yields

$$\sum_{k=1}^{K} h_{x_k}(\alpha_k) \;\leq\; \frac{(\alpha_1 - \hat{\alpha})^2}{2\eta} + \sum_{k=1}^{K} h_{x_k}(\hat{\alpha}) + \frac{\eta}{2(1-\eta L)} \sum_{k=1}^{K} \big\| h'_{x_k}(\hat{\alpha}) \big\|^2,$$

which is exactly (3).

A.5 Proof of Lemma 4

Proof 5 *Fix a reference stepsize* $\hat{\alpha} > 0$ *and write* $g_k := \nabla f(x_k)$. *By assumption,* $x_k \to \mathcal{X}^\star$ *while* $x_k \notin \mathcal{X}^\star$ *for all* k, *so* $\|g_k\| > 0$ *for every* k *and* $\operatorname{dist}(x_k, \mathcal{X}^\star) \to 0$.

Step 1: A uniform negative bound for $h_{x_k}(\hat{\alpha})$**.** *By Lemma 1, there exists* $\theta > 0$ *(depending on* $\hat{\alpha}$*) such that whenever* $\operatorname{dist}(x, \mathcal{X}^\star) \leq \theta$,

$$f\big(x - \hat{\alpha} \nabla f(x)\big) \;\leq\; f(x) - \frac{\hat{\alpha}}{2} \|\nabla f(x)\|^2 \;\leq\; f(x).$$

Since $\operatorname{dist}(x_k, \mathcal{X}^\star) \to 0$, *there exists an index* $N_1 = N_1(\hat{\alpha})$ *such that* $\operatorname{dist}(x_k, \mathcal{X}^\star) \leq \theta$ *for all* $k \geq N_1$. *For such* k *we obtain*

$$h_{x_k}(\hat{\alpha}) \;=\; \frac{f(x_k - \hat{\alpha} g_k) - f(x_k)}{\|g_k\|^2} \;\leq\; -\frac{\hat{\alpha}}{2}. \tag{9}$$

Step 2: Controlling the derivative term via local smoothness. *For each* k, *consider the one–dimensional function* $t \mapsto h_{x_k}(t)$ *along the ray* $x_k - t g_k$. *By self–boundedness and the definition of local smoothness on level sets, there exists a sequence of local smoothness constants* (L_k) *such that* $L_k \to 0$ *and, for all sufficiently large* k, f *is* L_k*–smooth on the segment* $\{x_k - t g_k : t \in [0, \hat{\alpha}]\}$. *In particular,* h_{x_k} *is twice differentiable on* $[0, \hat{\alpha}]$ *with*

$$0 \;\leq\; h''_{x_k}(t) \;\leq\; L_k, \qquad t \in [0, \hat{\alpha}],$$

and $h'_{x_k}(0) = -1$ (by construction of the normalized decrease).
Integrating h''_{x_k} from 0 to $\hat{\alpha}$ gives

$$h'_{x_k}(\hat{\alpha}) = h'_{x_k}(0) + \int_0^{\hat{\alpha}} h''_{x_k}(t)\,dt = -1 + \varepsilon_k(\hat{\alpha}), \qquad |\varepsilon_k(\hat{\alpha})| \le L_k\hat{\alpha}.$$

Hence

$$\left\|h'_{x_k}(\hat{\alpha})\right\|^2 = \left(1 - \varepsilon_k(\hat{\alpha})\right)^2 \le 1 + 2|\varepsilon_k(\hat{\alpha})| + \varepsilon_k(\hat{\alpha})^2 \le 1 + C_1 L_k\hat{\alpha}$$

for some constant $C_1 = C_1(\hat{\alpha}) > 0$ and all sufficiently large k. Since $L_k \to 0$, we may choose $N_2 = N_2(\hat{\alpha})$ such that $k \ge N_2$ implies

$$\frac{\eta}{2(1-\eta L)}\left\|h'_{x_k}(\hat{\alpha})\right\|^2 \;\le\; \varepsilon_k(\hat{\alpha}), \qquad \varepsilon_k(\hat{\alpha}) \xrightarrow[k\to\infty]{} 0. \tag{10}$$

Step 3: Tail bound for the averaged quantity. *Let $N := \max\{N_1, N_2\}$. Combining (9) and (10), we obtain for all $k \ge N$,*

$$h_{x_k}(\hat{\alpha}) + \frac{\eta}{2(1-\eta L)}\left\|h'_{x_k}(\hat{\alpha})\right\|^2 \;\le\; -\frac{\hat{\alpha}}{2} + \varepsilon_k(\hat{\alpha}), \qquad \varepsilon_k(\hat{\alpha}) \to 0. \tag{11}$$

Now split the Cesàro average into a finite prefix and a tail:

$$\frac{1}{K}\sum_{k=1}^{K}\Big(h_{x_k}(\hat{\alpha}) + \frac{\eta}{2(1-\eta L)}\left\|h'_{x_k}(\hat{\alpha})\right\|^2\Big) = \frac{1}{K}\sum_{k=1}^{N-1}(\cdots) + \frac{1}{K}\sum_{k=N}^{K}(\cdots).$$

The first term involves only finitely many summands, hence

$$\lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{N-1}(\cdots) = 0.$$

For the tail, use (11):

$$\frac{1}{K}\sum_{k=N}^{K}(\cdots) \;\le\; \frac{K-N+1}{K}\Big(-\frac{\hat{\alpha}}{2}\Big) + \frac{1}{K}\sum_{k=N}^{K}\varepsilon_k(\hat{\alpha}).$$

Taking $\limsup$ *as $K \to \infty$ yields*

$$\limsup_{K\to\infty}\frac{1}{K}\sum_{k=N}^{K}(\cdots) \;\le\; -\frac{\hat{\alpha}}{2} + \limsup_{K\to\infty}\frac{1}{K}\sum_{k=N}^{K}\varepsilon_k(\hat{\alpha}).$$

Since $\varepsilon_k(\hat{\alpha}) \to 0$, the Cesàro average on the right-hand side vanishes, so

$$\limsup_{K\to\infty}\frac{1}{K}\sum_{k=N}^{K}(\cdots) \;\le\; -\frac{\hat{\alpha}}{2}.$$

Combining the prefix and tail contributions gives

$$\limsup_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K}\Big(h_{x_k}(\hat{\alpha}) + \frac{\eta}{2(1-\eta L)}\left\|h'_{x_k}(\hat{\alpha})\right\|^2\Big) \;\le\; -\frac{\hat{\alpha}}{2}.$$

Since the sequence of averages is bounded from above and below, the $\limsup$ *is in fact a limit, which proves the claim.*

B Code for Numerical Results

All code used to produce the numerical experiments and figures in this paper is publicly available at:

`https://github.com/Yifa-Yu/CME291-307-Project`

The repository contains the implementation of OSGM-H and the baseline first-order methods (GD, Adam, AdaGrad), along with scripts to reproduce all experiments and plots reported in Section 3.

C AI Contribution

We used AI to help with coding the experiments, generating data-visualization scripts, and polishing the writing (grammar and phrasing). All experimental designs, modeling choices, and final edits were made and verified by the authors.