Hypergradient acceleration

Wenzhi Gao 
gwz@stanford.edu

Yifa Yu
yifacop0@stanford.edu

ICME, Stanford University

Abstract

This blog post[1] considers hypergradient descent (HDM), a stepsize adaptation heuristic for gradient-based methods [BCR+17, Sch99]. Recent advances on HDM use online learning to provide a theoretical justification of HDM on smooth convex functions, and the main finding is that, HDM can perform competitively with the "optimal stepsize $\alpha^\star$" for the iteration trajectory [GCYU25a, CGYU25b, GCYU25b, CGYU25a]. As a consequence, HDM is able to improve dependence on problem conditioning (e.g., smoothness) in the convergence rate. However, in the general smooth convex case, improving dependence on smoothness would not improve the $\mathcal{O}(\frac{1}{K})$ convergence rate. In this post, we consider a special case where HDM is able to produce $o(\frac{1}{K})$ rate of convergence. The idea is that, if the Hessian of a function vanishes around its optimum (i.e. $\|\nabla^2 f(x^\star)\| = 0$), then HDM will automatically increase the stepsize and obtain accelerated rate.

1 Introduction and background

HDM used to be a popular heuristic in optimization for deep learning. Given an unconstrained problem

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

HDM works by doing a *hypergradient descent* step (step (1)) before standard gradient descent (step (2)):

$$\alpha_{k+1} = \alpha_k + \eta \frac{\langle \nabla f(x^k - \alpha_k \nabla f(x^k)), \nabla f(x^k) \rangle}{\|\nabla f(x^k)\|^2} \tag{1}$$

$$x^{k+1} = x^k - \alpha_{k+1} \nabla f(x^k), \tag{2}$$

where $\{\alpha_k\}$ is known as a stepsize sequence and $\{x^k\}$ is the iterate sequence; $\eta > 0$ is an algorithm parameter known as hypergradient descent learning rate. Here's what (1) does

1. construct a test point $x^k - \alpha_k \nabla f(x^k)$ with the current stepsize;
2. increase/decrease the stepsize by the (normalized) inner product $\frac{\langle \nabla f(x^k - \alpha_k \nabla f(x^k)), \nabla f(x^k) \rangle}{\|\nabla f(x^k)\|^2}$

To understand the intuition of (1), consider a 1D quadratic function $f(x) = \frac{1}{2} x^2$. We have $\nabla f(x) = x$ and

$$x - \alpha \nabla f(x) = x - \alpha x = (1 - \alpha) x.$$

Here are some observations:

- From any point x, we know that stepsize $\alpha^\star = 1$ would bring us to $x^\star = 0$. So let's call $\alpha^\star$ optimal.
- If we use stepsize $\alpha > 1$, then a gradient descent step "overshoots" the solution and zig-zag between two branches of the graph (**Figure 1**, (left)), hence the (negative) gradient direction between two consecutive iterates will have different signs: $\langle \nabla f(x - \alpha \nabla f(x)), \nabla f(x) \rangle < 0$. Step (1) (with small $\eta > 0$) encourages α_{k+1} to get closer to $\alpha^\star$.

1. Blog generated by TeXmacs

- Similarly, if $\alpha < 1$, the iterate will stay on one branch of the graph (**Figure 1**, (right)), making $\langle \nabla f(x - \alpha \nabla f(x)), \nabla f(x)\rangle > 0$. Again, step (1) encourages α_{k+1} to get closer to $\alpha^\star$.

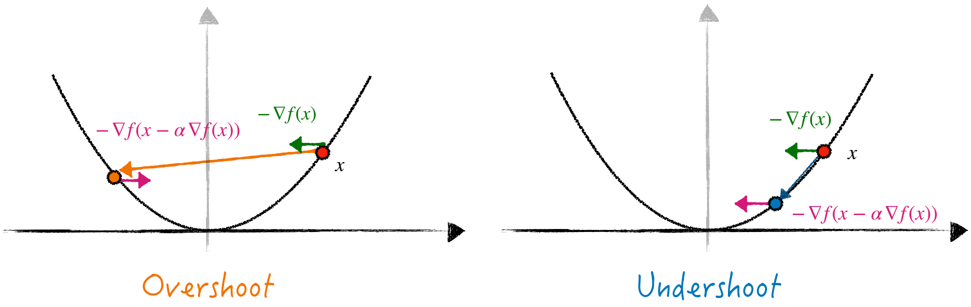

Figure 1. When the step over/undershoots the optimum, HDM will decrease/increase the stepsize accordingly.

In this simple scalar case, HDM can be interpreted as adaptively adjusting the stepsize based on *how much the gradients in consecutive iterations align with each other*. This particular idea first appeared in a 30-year old heuristic called Delta-Bar-Delta [Jac88], the origin of HDM-type methods.

Optimal stepsize and HDM. The intuition from the 1D toy example suggests that HDM should somewhat match the performance of a good stepsize.

To get a better sense of what "a good stepsize" means, suppose f is convex and twice-continuously differentiable. Then the largest eigenvalue of Hessian, $\|\nabla^2 f(x)\|$, provides a good surrogate of a "locally" good stepsize: $\alpha \sim \mathcal{O}\left(\frac{1}{\|\nabla^2 f(x)\|}\right)$. In particular, if f has L-Lipschitz continuous gradient (L-smooth), then $\|\nabla^2 f(x)\| \le L$, and a good stepsize becomes $\alpha \sim \mathcal{O}\left(\frac{1}{L}\right)$, a quantity that often appears in the convergence analysis of gradient descent.

Very informally, we can summarize the convergence guarantee of HDM as follows.

Theorem 1. *(Very informal) For L-smooth convex problems, HDM can achieve the performance of a good constant stepsize* $\alpha \sim \mathcal{O}\left(\frac{1}{L}\right)$.

Theorem 1 was first established in [GCYU25a, CGYU25b, GCYU25b, CGYU25a]. These works show that, as a consequence of achieving the performance of a good constant stepsize, HDM obtain improved dependence on conditioning-related constants. In other words, if gradient descent achieves $\mathcal{O}\left(\frac{C}{K}\right)$ rate with a good constant stepsize, then HDM can match this performance (asymptotically).

This result is fine, but only being able to improve constants looks unsatisfactory. So a question arises:

Are there cases where HDM can provably obtain convergence rate improvement?

The rest of this blog addresses this question.

1.1 Functions with vanishing curvature around the optimum

Recall that we argued that HDM will match the performance of stepsize $\mathcal{O}\left(\frac{1}{L}\right)$ when f is L-smooth. But L-smoothness is a global property, and we emphasized that the performance of HDM should be determined by a locally good stepsize: $\alpha \sim \mathcal{O}\left(\frac{1}{\|\nabla^2 f(x)\|}\right)$. In particular, we focus on the local behavior around $x^\star$, which is determined by $\|\nabla^2 f(x^\star)\|$.

An interesting question here: what if $\|\nabla^2 f(x^\star)\| = 0$? Is it possible? Consider $f(x) = \frac{1}{4}x^4$ with $x^\star = 0$. Then $\nabla^2 f(x) = 3x^2$ and $\|\nabla^2 f(x^\star)\| = 0$. Graphically, this function is flat around its optimum (**Figure 2**).

Other examples. There are many other functions that satisfy $\|\nabla^2 f(x^\star)\| = 0$. Typical examples are cross entropy for logistic regression, exponential loss, ℓ_p regression or powered hinge loss [Ora19]. See Section 2.1 of [GHU26] and [FMVS25] for more details.

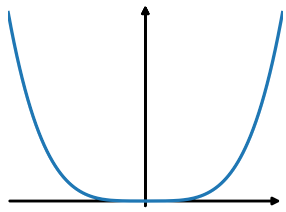

Figure 2. Function $f(x) = \frac{1}{4}x^4$ is flat around the optimum $x^\star = 0$.

What would happen if we run HDM on $\frac{1}{4}x^4$? Or, what does it mean by matching the $\mathcal{O}\big(\frac{1}{\|\nabla^2 f(x^\star)\|}\big)$ stepsize when $\|\nabla^2 f(x^\star)\| = 0$? Perhaps unsurprisingly, when we run HDM (**Figure 3** (right)), the stepsize consistently increases. In other words, $\alpha^\star = \infty$ and HDM will keep increasing the stepsize when x^k approaches $x^\star$.

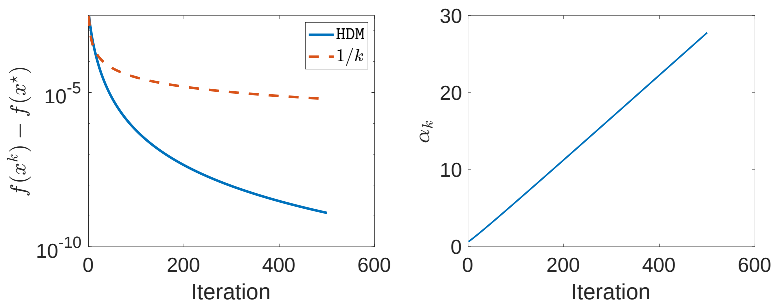

Figure 3. Convergence behavior of HDM on $f(x) = \frac{1}{4}x^4$. The stepsize $\{\alpha_k\}$ keeps increasing and the convergence rate is faster than $\mathcal{O}\big(\frac{1}{K}\big)$ predicted by theory.

Another interesting observation is the convergence rate (**Figure 3** (left)). It seems HDM is converging faster than $\mathcal{O}\big(\frac{1}{K}\big)$. Why does this happen? Recall the convergence rate of gradient on L-smooth functions

$$f(x^{K+1}) - f(x^\star) \le \frac{LD^2}{K},$$

where D is some constant and L is the global smooth constant. This is a universal yet pessimistic estimate: when $x^k \to x^\star$, what really matters is the local smoothness around $x^\star$ determined by $\|\nabla^2 f(x^\star)\|$. When $\|\nabla^2 f(x^\star)\| = 0$, we are able to use a much smaller smoothness constant (say $\hat{L}$) than L. Since $\hat{L} \to 0$ as $x \to x^\star$, then we expect an improved convergence rate if $\hat{L}$ scales as $\mathrm{poly}(K^{-1})$. **Theorem 2** formalizes this idea.

Theorem 2. *If f is L-smooth and satisfies*

$$\|\nabla^2 f(x)\| \le c[f(x) - f(x^\star)]^\beta,$$

then HDM has convergence rate $\mathcal{O}\big(\frac{1}{K^{1+\beta}}\big)$.

2 Hypergradient acceleration

This section formalizes **Theorem 2** and starts with a short recap of the convergence analysis of hypergradient descent. The analysis is similar to [GCYU25b], but no background of online learning is required.

2.1 Online stepsize learning and hypergradient descent

Feel free to skip this section if you are familiar with online learning. Recall the previous intuition: `HDM` will adjust the stepsize to make it get closer to a good stepsize. How to mathematically quantify this intuition? Suppose there is some good stepsize $\alpha^\star$, then we can define a distance function $\text{dist}(\alpha, \alpha^\star)$ to quantify such "closeness". Let's choose the squared Euclidean distance

$$g(\alpha) := \text{dist}(\alpha, \alpha^\star) := \tfrac{1}{2}(\alpha - \alpha^\star)^2.$$

Then, getting closer to a good stepsize can be quantified by

$$\tfrac{1}{2}(\alpha_{k+1} - \alpha^\star)^2 \leq \tfrac{1}{2}(\alpha_k - \alpha^\star)^2 \tag{3}$$

Now another question arises: why/how can we ensure that (3) hold? Since we want to reduce the distance function $g(\alpha) = \text{dist}(\alpha, \alpha^\star)$, the most straightforward way is to find a descent direction of g.

Recall that a direction d is a descent direction of a function $g(\alpha)$ at α if

$$\langle [-\nabla g(\alpha)], d\rangle > 0.$$

In other words, d should form an acute angle with the negative gradient direction $-\nabla g(\alpha)$. To see why d is a descent direction, when g is L-smooth, we have

$$g(\alpha + \eta d) \leq g(\alpha) + \eta\langle \nabla g(\alpha), d\rangle + \tfrac{\eta^2}{2}\|d\|^2 = g(\alpha) - \eta\underbrace{\langle [-\nabla g(\alpha)], d\rangle}_{>0} + \tfrac{L\eta^2}{2}\|d\|^2$$

where $\langle -\nabla g(\alpha), d\rangle > 0$ is by definition of descent direction. Letting $\eta \to 0$, we always have $g(\alpha + \eta d) < g(\alpha)$. With $g(\alpha) = \frac{1}{2}(\alpha - \alpha^\star)^2$ and $\nabla g(\alpha) = \alpha - \alpha^\star$, we essentially need a direction d such that

$$0 < \langle -\nabla g(\alpha), d\rangle = -\langle \alpha - \alpha^\star, d\rangle.$$

Where can we find such a descent direction? The answer is convexity: a convex function satisfies

$$h(\alpha^\star) \geq h(\alpha) + \langle \nabla h(\alpha), \alpha^\star - \alpha\rangle \qquad \Leftrightarrow \quad \langle -\nabla g(\alpha), -\nabla h(\alpha)\rangle \geq h(\alpha) - h(\alpha^\star).$$

Whenever we can find a function h such that

- h is convex and ∇h is computable
- h is (approximately) minimized by $\alpha^\star$ so that $h(\alpha) - h(\alpha^\star) \geq 0$,

then the gradient $-\nabla h(\alpha)$ is a descent direction of $g(\alpha) = \frac{1}{2}(\alpha - \alpha^\star)^2$. Let's instantiate the descent property. Say we already have such a convex function $h(\alpha)$ and do a gradient descent step on α:

$$\alpha_+ = \alpha + \eta d = \alpha - \eta \nabla h(\alpha).$$

Then combining the following facts

1. $d = -\nabla h(\alpha)$
2. $g(\alpha) = \frac{1}{2}(\alpha - \alpha^\star)^2$ is 1-smooth:

$$g(\alpha - \eta\nabla h(\alpha)) \leq g(\alpha) + \langle \nabla g(\alpha), \eta \nabla h(\alpha)\rangle + \tfrac{\eta^2}{2}\|d\|^2 = g(\alpha) - \eta\langle -\nabla g(\alpha), -\nabla h(\alpha)\rangle + \tfrac{\eta^2}{2}\|d\|^2$$

3. $\langle -\nabla g(\alpha), -\nabla h(\alpha)\rangle \geq h(\alpha) - h(\alpha^\star)$ by convexity

gives the descent inequality for $g(\alpha)$:

$$\begin{aligned}
\tfrac{1}{2}(\alpha_+ - \alpha^\star)^2 &= g(\alpha - \eta\nabla h(\alpha)) \\
\text{(By 1-smoothness)} &\leq g(\alpha) - \eta\langle -\nabla g(\alpha), -\nabla h(\alpha)\rangle + \tfrac{\eta^2}{2}\|\nabla h(\alpha)\|^2 \\
\text{(By convexity)} &\leq g(\alpha) - \eta[h(\alpha) - h(\alpha^\star)] + \tfrac{\eta^2}{2}\|\nabla h(\alpha)\|^2 \\
&= \tfrac{1}{2}(\alpha - \alpha^\star)^2 - \eta[h(\alpha) - h(\alpha^\star)] + \tfrac{\eta^2}{2}\|\nabla h(\alpha)\|^2.
\end{aligned}$$

The intuition from the above relation is clear: for any convex function h, if α underperforms $\alpha^\star$ (i.e., $h(\alpha) > h(\alpha^\star)$ since we are minimizing), then stepping in the direction $-\nabla h(\alpha)$ with a small stepsize $\eta > 0$ decreases the distance to $\alpha^\star$. In other words, we *learn* $\alpha^\star$ through loss function h.

The key intuition is that, the more we underperform, the more we learn:

$$\underbrace{\tfrac{1}{2}(\alpha_+ - \alpha^\star)^2 - \tfrac{1}{2}(\alpha - \alpha^\star)^2}_{\text{How much we learn}} \leq -\eta[\underbrace{h(\alpha) - h(\alpha^\star)}_{\text{How much we underperform}}] + \tfrac{\eta^2}{2}\underbrace{\|\nabla h(\alpha)\|^2}_{\mathcal{O}(\eta^2)\text{ error term}}.$$

Back to the context of hypergradient descent. We need some function $h(\alpha)$ (approximately) minimized by a good stepsize $\alpha^\star$. What function takes a small value when a good stepsize is used? A straightforward one is the function value $f(x - \alpha\nabla f(x))$, or equivalently, the relative progress associated with α:

$$h_x(\alpha) = \tfrac{f(x - \alpha\nabla f(x)) - f(x)}{\|\nabla f(x)\|^2}.$$

It is easy to verify that $\nabla h_x(\alpha) = \frac{-\langle \nabla f(x - \alpha\nabla f(x)), \nabla f(x)\rangle}{\|\nabla f(x)\|^2}$, and the hypergradient descent step (1) becomes

$$\alpha_{k+1} = \alpha_k + \eta\frac{\langle \nabla f(x^k - \alpha_k\nabla f(x^k)), \nabla f(x^k)\rangle}{\|\nabla f(x^k)\|^2} = \alpha_k - \eta\nabla h_x(\alpha_k),$$

which is known as online gradient descent. How to analyze `HDM`? Recall the relation

$$\tfrac{1}{2}(\alpha_{k+1} - \alpha^\star)^2 - \tfrac{1}{2}(\alpha_k - \alpha^\star)^2 \leq -\eta[h_{x^k}(\alpha_k) - h_{x^k}(\alpha^\star)] + \tfrac{\eta^2}{2}\|\nabla h_{x^k}(\alpha_k)\|^2.$$

When we measure the cumulative amount we have learned over a time horizon K, we have, for $\eta \leq \frac{1}{L}$, that

$$\begin{aligned}
\underbrace{\tfrac{1}{2}(\alpha_{K+1} - \alpha^\star)^2 - \tfrac{1}{2}(\alpha_1 - \alpha^\star)^2}_{\text{How much we learn over } K \text{ steps}} &= \sum_{k=1}^{K} \underbrace{\tfrac{1}{2}(\alpha_{k+1} - \alpha^\star)^2 - \tfrac{1}{2}(\alpha_k - \alpha^\star)^2}_{\text{How much we learn in step } k} \\
&\leq -\eta\sum_{k=1}^{K}[\underbrace{h_{x^k}(\alpha_k) - h_{x^k}(\alpha^\star)}_{\text{How much } \{\alpha_k\} \text{ underperform in step } k}] + \tfrac{\eta^2}{2}\sum_{k=1}^{K}\|\nabla h_{x^k}(\alpha_k)\|^2 \\
&= -\eta \underbrace{\sum_{k=1}^{K}[h_{x^k}(\alpha_k) - h_{x^k}(\alpha^\star)]}_{\text{How much } \{\alpha_k\} \text{ underperform over } K \text{ steps}} + \tfrac{\eta^2}{2}\sum_{k=1}^{K}\|\nabla h_{x^k}(\alpha_k)\|^2 \\
&\leq -\eta \underbrace{\sum_{k=1}^{K}[h_{x^k}(\alpha_{k+1}) - h_{x^k}(\alpha^\star)]}_{\text{How much } \{\alpha_{k+1}\} \text{ underperform over } K \text{ steps}},
\end{aligned}$$

where the last step uses the fact that h_x is L-smooth and that $h_{x^k}(\alpha_{k+1}) \leq h_{x^k}(\alpha_k) - \left(\eta - \frac{L\eta^2}{2}\right)\|\nabla h_{x^k}(\alpha_k)\|^2$ to cancel the error term. Over the K steps, the more $\{\alpha_{k+1}\}$ underperform, the more we learn. Let's rephrase this intuition:

- suppose $\alpha^\star$ is finite

⇒ there is finite that we can learn

⇒ there is finite that we can underperform

Mathematically, we have

$$\underbrace{\sum_{k=1}^{K}[h_{x^k}(\alpha_{k+1})-h_{x^k}(\alpha^\star)]}_{\text{How much } \{\alpha_{k+1}\} \text{ underperform}} \leq \underbrace{\frac{1}{2\eta}(\alpha_1-\alpha^\star)^2-\frac{1}{2\eta}(\alpha_{K+1}-\alpha^\star)^2 \leq \frac{1}{2\eta}(\alpha_1-\alpha^\star)^2}_{\text{Finite to learn}}. \tag{4}$$

Finally, given that $\{\alpha_{k+1}\}$ has finite underperformance over K steps, it remains to connect it with the convergence rate, which is given by the following reduction lemma.

Lemma 1. (Reduction lemma [GCYU25a]) *Suppose f is convex, then* `HDM` *generates* $\{x^k\}$ *such that*

$$f(x^{K+1})-f(x^\star) \leq \frac{\Delta^2}{K}\frac{1}{\max\left\{-\frac{1}{K}\sum_{k=1}^{K}h_{x^k}(\alpha_{k+1}),0\right\}} \leq \frac{\Delta^2}{K}\frac{1}{\max\left\{-\frac{1}{K}\sum_{k=1}^{K}h_{x^k}(\alpha^\star)-\frac{1}{2\eta}(\alpha_1-\alpha^\star)^2,0\right\}}$$

for any $\alpha^\star \in \mathbb{R}$, *where* $\Delta := \max_{x\in\{x:x\leq f(x^1)\}}\min_{x^\star\in\{x:f(x)=f(x^\star)\}}\|x-x^\star\|$.

The power of **Lemma 1** lies in the freedom for choosing $\alpha^\star$. First, for an L-smooth function, we always $f\big(x-\frac{1}{L}\nabla f(x)\big)-f(x^\star)\leq\frac{1}{2L}\|\nabla f(x)\|^2$, which implies $h_{x^k}\big(\frac{1}{L}\big)\leq-\frac{1}{2L}$: plugging in $\alpha^\star=\frac{1}{L}$ gives the following baseline convergence result:

Lemma 2. *Suppose f is L-smooth, convex. Then* `HDM` *with* $\eta=\frac{1}{L}$ *generates* $\{x^k\}$ *such that*

$$f(x^{K+1})-f(x^\star)\leq\frac{2L\Delta^2}{K}\frac{2K}{2K-L^2\big(\alpha_1-\frac{1}{L}\big)^2},$$

where Δ *is defined in* ***Lemma 1****. In particular, if* $\alpha_1=\frac{1}{L}$, *we have* $f(x^{K+1})-f(x^\star)\leq\frac{2L\Delta^2}{K}$.

Proof. The proof follows by plugging in $\alpha^\star=\frac{1}{L}$ and applying $h_{x^k}\big(\frac{1}{L}\big)\leq-\frac{1}{2L}$. □

2.2 Convergence rate acceleration

Lemma 2 provides a baseline $\mathcal{O}\big(\frac{1}{K}\big)$ convergence rate. But as we previously mentioned, we should expect a faster rate than $\mathcal{O}\big(\frac{1}{K}\big)$ when the function becomes flat when the iterates approach optimality:

A1. We have $\|\nabla f(x)-\nabla f(y)\|\leq c_\delta\|x-y\|$ for all $x,y\in\mathcal{L}_\delta:=\{x:f(x)-f(x^\star)\leq\delta\}$, where $c_\delta\leq L$.

Under **A1**, it is feasible to plug in $\alpha^\star\gg\frac{1}{L}$ to reduce $h_{x^k}(\alpha^\star)$, as shown by **Lemma 3**.

Lemma 3. *Under* ***A1****, we have* $h_x\big(\frac{1}{c_\delta}\big)\leq-\frac{1}{2c_\delta}$ *for all* $x\in\mathcal{L}_\delta$.

Now we are ready to prove the main result.

Theorem 3. *Suppose f is L-smooth, convex and satisfies* ***A1****, then for all* $K\geq\frac{2L\Delta^2}{\delta}$, `HDM` *with* $\eta=\frac{1}{L}$ *satisfies*

$$f(x^{2K+1})-f(x^\star)\leq\frac{4\Delta^2}{K}\frac{c_{2L\Delta^2K^{-1}}}{\max\left\{1-\frac{L}{Kc_{2L\Delta^2K^{-1}}},0\right\}}.$$

Proof. Our analysis will follow a two-phase argument. Suppose HDM is run for $2K$ iterations. Then

$$f(x^{2K+1})-f(x^\star)\le \frac{c_{2L\Delta^2K^{-1}}}{K}\frac{4\Delta^2}{\max\left\{1-\frac{L}{K\eta}(\alpha_2^\star)^2,0\right\}}$$

by **Lemma 1**. According to (4), we can decompose $\Sigma := \sum_{k=1}^{2K} h_{x^k}(\alpha_{k+1})$ into

$$\begin{aligned}\Sigma_1 := \textstyle\sum_{k=1}^{K} h_{x^k}(\alpha_{k+1}) &\le \textstyle\sum_{k=1}^{K} h_{x^k}(\alpha_1^\star)+\frac{1}{2\eta}(\alpha_1-\alpha_1^\star)^2-\frac{1}{2\eta}(\alpha_{K+1}-\alpha_1^\star)^2\\ \Sigma_2 := \textstyle\sum_{k=K+1}^{2K} h_{x^k}(\alpha_{k+1}) &\le \textstyle\sum_{k=1}^{K} h_{x^k}(\alpha_2^\star)+\frac{1}{2\eta}(\alpha_{K+1}-\alpha_2^\star)^2-\frac{1}{2\eta}(\alpha_{2K+1}-\alpha_2^\star)^2.\end{aligned}$$

To bound Σ_1, we take $\alpha_1^\star=\alpha_1=\frac{1}{L}$ and $\Sigma_1\le -\frac{K}{2L}+\frac{1}{2\eta}\underbrace{(\alpha_1-\alpha_1^\star)^2}_{=0}-\frac{1}{2\eta}(\alpha_{K+1}-\alpha_1^\star)^2$.

To bound Σ_2, we notice that after the first K iterations, $f(x^{K+k})-f(x^\star)\le\frac{2L\Delta^2}{K}$ for all $k\in[K]$ by monotonicity of the algorithm and **Lemma 2**, which implies $x^{K+k}\in\mathcal{L}_\delta$ since $K\ge\frac{2L\Delta^2}{\delta}$. Taking $\alpha_2^\star=\frac{1}{c_{2L\Delta^2K^{-1}}}$, we have $h_{x^k}(\alpha_2^\star)\le-\frac{1}{2c_{2L\Delta^2K^{-1}}}$ and that

$$\textstyle\sum_{k=K+1}^{2K} h_{x^k}(\alpha_{k+1})\le -\frac{K}{2c_{2L\Delta^2K^{-1}}}+\frac{1}{2\eta}(\alpha_{K+1}-\alpha_2^\star)^2-\frac{1}{2\eta}(\alpha_{2K+1}-\alpha_2^\star)^2.$$

Summing up Σ_1 and Σ_2, we have

$$\begin{aligned}\Sigma &= \Sigma_1+\Sigma_2\\ &\le -\frac{K}{2L}-\frac{1}{2\eta}(\alpha_{K+1}-\alpha_1^\star)^2\\ &\quad -\frac{K}{2c_\delta}+\frac{1}{2\eta}(\alpha_{K+1}-\alpha_2^\star)^2-\frac{1}{2\eta}(\alpha_{2K+1}-\alpha_2^\star)^2\\ &\le -\frac{K}{2c_\delta}-\frac{1}{2\eta}(\alpha_{K+1}-\alpha_1^\star)^2+\frac{1}{2\eta}(\alpha_{K+1}-\alpha_2^\star)^2\\ &= -\frac{K}{2c_\delta}-\frac{1}{2\eta}(2\alpha_{K+1}-\alpha_1^\star-\alpha_2^\star)(\alpha_2^\star-\alpha_1^\star).\end{aligned}$$

Since $\alpha_2^\star=\frac{1}{c_{2L\Delta^2K^{-1}}}\ge\alpha_1^\star$, we have $(2\alpha_{K+1}-\alpha_1^\star-\alpha_2^\star)(\alpha_2^\star-\alpha_1^\star)\ge(\alpha_1^\star)^2-(\alpha_2^\star)^2$ and

$$\Sigma\le-\frac{K}{2c_{2L\Delta^2K^{-1}}}+\frac{1}{2\eta}(\alpha_2^\star)^2=-\frac{K}{2c_{2L\Delta^2K^{-1}}}+\frac{1}{2\eta}\Big(\frac{1}{c_{2L\Delta^2K^{-1}}}\Big)^2$$

Plugging Σ back gives

$$\begin{aligned}f(x^{2K+1})-f(x^\star) &\le \frac{\Delta^2}{K}\frac{1}{\max\left\{-\frac{1}{2K}\sum_{k=1}^{2K}h_{x^k}(\alpha_{k+1}),0\right\}}\\ &\le \frac{4\Delta^2}{K}\frac{c_{2L\Delta^2K^{-1}}}{\max\left\{1-\frac{L}{Kc_{2L\Delta^2K^{-1}}},0\right\}}\end{aligned}$$

and completes the proof. □

By specifying concrete expressions of c_δ, we can get convergence rate improvement.

Corollary 1. *Under the same settings as* ***Theorem 3****, suppose* $c_\delta\le L\delta^\beta,\beta\in[0,1)$*, then*

$$f(x^{2K+1})-f(x^\star)=\mathcal{O}\Big(\frac{1}{K^{1+\beta}}\Big).$$

As a special case, if f is twice continuously differentiable with $\|\nabla^2 f(x)\|\le c[f(x)-f(x^\star)]^\beta$*, then it satisfies* ***A1*** *with the same value of* β.

Proof. Since $\beta\in[0,1), c_{2L\Delta^2K^{-1}}=\mathcal{O}(K^{-\beta})$ and $\frac{1}{Kc_{2L\Delta^2K^{-1}}}=o(1)$ and

$$\frac{4\Delta^2}{K}\frac{c_{2L\Delta^2K^{-1}}}{\max\left\{1-\frac{L}{Kc_{2L\Delta^2K^{-1}}},0\right\}}=\mathcal{O}\Big(\frac{1}{K^{1+\beta}}\Big).$$ □

References

[BCR+17] Atilim Gunes Baydin, Robert Cornish, David Martinez Rubio, Mark Schmidt, and Frank Wood. Online learning rate adaptation with hypergradient descent. *ArXiv preprint arXiv:1703.04782*, 2017.

[CGYU25a] Ya-Chi Chu, Wenzhi Gao, Yinyu Ye, and Madeleine Udell. Gradient methods with online scaling part ii. practical aspects. *ArXiv preprint arXiv:2509.11007*, 2025.

[CGYU25b] Ya-Chi Chu, Wenzhi Gao, Yinyu Ye, and Madeleine Udell. Provable and practical online learning rate adaptation with hypergradient descent. *ArXiv preprint arXiv:2502.11229*, 2025.

[FMVS25] Curtis Fox, Aaron Mishkin, Sharan Vaswani, and Mark Schmidt. Glocal smoothness: line search can really help! *ArXiv preprint arXiv:2506.12648*, 2025.

[GCYU25a] Wenzhi Gao, Ya-Chi Chu, Yinyu Ye, and Madeleine Udell. Gradient methods with online scaling. In *The Thirty Eighth Annual Conference on Learning Theory*, pages 2192–2226. PMLR, 2025.

[GCYU25b] Wenzhi Gao, Ya-Chi Chu, Yinyu Ye, and Madeleine Udell. Gradient methods with online scaling part i. theoretical foundations. *ArXiv preprint arXiv:2505.23081*, 2025.

[GHU26] Wenzhi Gao, Chang He, and Madeleine Udell. Small gradient norm regret for online convex optimization. *ArXiv preprint arXiv:2601.13519*, 2026.

[Jac88] Robert A Jacobs. Increased rates of convergence through learning rate adaptation. *Neural networks*, 1(4):295–307, 1988.

[Ora19] Francesco Orabona. Online linear classification. oct 2019. Blog post on *Parameter-free Learning and Optimization Algorithms*. Updated Nov 26, 2019. Accessed 2026-03-03.

[Sch99] Nicol N Schraudolph. Local gain adaptation in stochastic gradient descent. In *1999 Ninth international conference on artificial neural networks ICANN 99.(Conf. Publ. No. 470)*, volume 2, pages 569–574. IET, 1999.

A Proof of main results

A.1 Proof of Lemma 3

Proof. Suppose $x \in \mathcal{L}_\delta$ and let $\alpha_x^\star = \arg\min_\alpha h_x(\alpha)$ be the steepest descent stepsize. We start by showing $\alpha_x^\star \geq \frac{1}{c_\delta}$. Without loss of generality, we assume $\alpha_x^\star$ is finite. Then $h_x'(\alpha_x^\star) = 0$, $h_x'(0) = -\frac{\langle \nabla f(x), \nabla f(x) \rangle}{\|\nabla f(x)\|^2} = -1$, and

$$\begin{aligned} 1 = |h_x'(\alpha_x^\star) - h_x'(0)| &= \left| \tfrac{\langle \nabla f(x - \alpha_x^\star \nabla f(x)) - \nabla f(x), \nabla f(x) \rangle}{\|\nabla f(x)\|^2} \right| \\ &\leq \tfrac{\|\nabla f(x - \alpha_x^\star \nabla f(x)) - \nabla f(x)\|}{\|\nabla f(x)\|} \\ &\leq \tfrac{c_\delta \alpha_x^\star \|\nabla f(x)\|}{\|\nabla f(x)\|} = c_\delta \alpha_x^\star, \end{aligned}$$

giving $\alpha_x^\star \geq \frac{1}{c_\delta}$. Next we deduce that

$$\begin{aligned} & f(x - \alpha \nabla f(x)) - f(x) \\ = & -\alpha \int_0^1 \langle \nabla f(x - \alpha t \nabla f(x)), \nabla f(x) \rangle \mathrm{d}t \\ = & -\alpha \int_0^1 \langle \nabla f(x - \alpha t \nabla f(x)) - \nabla f(x), \nabla f(x) \rangle \mathrm{d}t - \alpha \int_0^1 \|\nabla f(x)\|^2 \mathrm{d}t \\ \leq & -\alpha \|\nabla f(x)\|^2 + \alpha \int_0^1 \|\nabla f(x - \alpha t \nabla f(x)) - \nabla f(x)\| \cdot \|\nabla f(x)\| \mathrm{d}t \end{aligned}$$

and for any $\alpha \leq \alpha_x^\star$, we have $f(x - \alpha \nabla f(x)) \leq f(x)$ and that $x - \alpha \nabla f(x) \in \mathcal{L}_\delta$. Hence $\|\nabla f(x - \alpha t \nabla f(x)) - \nabla f(x)\| \leq \alpha t c_\delta \|\nabla f(x)\|$, and we get

$$\begin{aligned} & f(x - \alpha \nabla f(x)) - f(x) \\ \leq & -\alpha \|\nabla f(x)\|^2 + \alpha \int_0^1 \alpha t c_\delta \|\nabla f(x)\|^2 \mathrm{d}t \\ = & -\alpha \|\nabla f(x)\|^2 + \tfrac{\alpha^2 c_\delta}{2} \|\nabla f(x)\|^2 = \left[\tfrac{\alpha^2 c_\delta}{2} - \alpha \right] \|\nabla f(x)\|^2. \end{aligned}$$

Taking $\alpha = \frac{1}{c_\delta} \leq \alpha_x^\star$, we get

$$h_x(\alpha_x^\star) \leq h_x(\alpha) = \tfrac{f(x - \alpha \nabla f(x)) - f(x)}{\|\nabla f(x)\|^2} \leq -\tfrac{1}{2c_\delta}$$

and this completes the proof. □