Gradient Methods with Online Scaling

Part I. Theoretical Foundations

Wenzhi Gao*[1], Ya-Chi Chu†[2], Yinyu Ye‡[1,3], and Madeleine Udell§[1,3]

[1]ICME, Stanford University
[2]Department of Mathematics, Stanford University
[3]Department of Management Science and Engineering, Stanford University

September 8, 2025

Abstract

This paper establishes the theoretical foundations of the online scaled gradient methods (`OSGM`)[1], a framework that utilizes online learning to adapt stepsizes and provably accelerate first-order methods. `OSGM` quantifies the effectiveness of a stepsize by a feedback function motivated from a convergence measure and uses the feedback to adjust the stepsize through an online learning algorithm. Consequently, instantiations of `OSGM` achieve convergence rates that are asymptotically no worse than the optimal stepsize. `OSGM` yields desirable convergence guarantees on smooth convex problems, including 1) trajectory-dependent global convergence on smooth convex objectives; 2) an improved complexity result on smooth strongly convex problems, and 3) local superlinear convergence. Notably, `OSGM` constitutes a new family of first-order methods with non-asymptotic superlinear convergence, joining the celebrated quasi-Newton methods. Finally, `OSGM` explains the empirical success of the popular hypergradient-descent heuristic in optimization for machine learning.

1 Introduction

Consider the L-smooth and μ-strongly convex optimization problem $\min_{x\in\mathbb{R}^n} f(x)$. When f has a large condition number $\kappa \coloneqq L/\mu$, vanilla gradient descent with constant scalar stepsize $1/L$ converges slowly. Instead of using a constant scalar stepsize, preconditioned gradient descent chooses a preconditioner $P_k \in \mathbb{R}^{n\times n}$, a matrix stepsize, to scale the gradient and accelerate convergence at iteration k:

$$x^{k+1} = x^k - P_k \nabla f(x^k). \tag{1}$$

We can locally measure the quality of stepsize P_k by the contraction ratio of suboptimality at x^k:

$$r_{x^k}(P_k) := \tfrac{f(x^{k+1})-f^\star}{f(x^k)-f^\star} = \tfrac{f(x^k - P_k\nabla f(x^k))-f^\star}{f(x^k)-f^\star}.$$

The suboptimality after K iterations of preconditioned gradient descent (1) is the product of these ratios $\{r_{x^k}(P_k)\}$ and can be further bounded by their cumulative sum using the arithmetic-geometric mean inequality:

$$\tfrac{f(x^{K+1})-f^\star}{f(x^1)-f^\star} = \prod_{k=1}^K \tfrac{f(x^{k+1})-f^\star}{f(x^k)-f^\star} = \prod_{k=1}^K r_{x^k}(P_k) \le (\tfrac{1}{K}\sum_{k=1}^K r_{x^k}(P_k))^K. \tag{2}$$

*gwz@stanford.edu, equal contribution
†ycchu97@stanford.edu, equal contribution
‡yyye@stanford.edu
§udell@stanford.edu

[1]This paper extends two previous works [15] and [9].

Given this analysis, how should we choose the stepsize P_k? Fast convergence is achieved if we prespecify stepsizes $\{P_k\}$ to minimize $\sum_{k=1}^K r_{x^k}(P_k)$. However, solving this problem is hard for a given instance f, as the function $r_{x^k}(P)$ depends on the previous sequence of stepsizes and the optimization landscape. However, this problem is ideally suited to online learning, which optimizes a cumulative sum of functions with provable regret guarantees even when the functions are adversarially chosen. For example, updating $\{P_k\}$ with online gradient descent $P_{k+1} = P_k - \eta \nabla r_{x^k}(P_k)$ guarantees sublinear regret with respect to any fixed stepsize $\hat{P}$ [51]:

$$\tfrac{1}{K}\sum_{k=1}^K r_{x^k}(P_k) \le \tfrac{1}{K}\sum_{k=1}^K r_{x^k}(\hat{P}) + \mathcal{O}(\tfrac{1}{\sqrt{K}}). \tag{3}$$

Given the freedom to choose $\hat{P}$, we may take $\hat{P}$ equal to $P_r^\star$ that achieves the optimal condition number $\kappa^\star < \kappa$ and hence the ratio $r_{x^k}(P_r^\star) \le 1 - \frac{1}{\kappa^\star}$. The regret guarantee (3) together with (2) implies a convergence rate

$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](1 - \tfrac{1}{\kappa^\star} + \mathcal{O}(\tfrac{1}{\sqrt{K}}))^K, \tag{4}$$

as well as an asymptotic iteration complexity $\mathcal{O}(\kappa^\star \log(1/\varepsilon))$. Moreover, if f is (locally) quadratic with Hessian matrix $H \in \mathbb{S}^n_{++}$, then $\hat{P} = H^{-1}$ yields perfect conditioning $\kappa^\star = 1$ and hence superlinear convergence:

$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](\mathcal{O}(\tfrac{1}{\sqrt{K}})^K).$$

This concise proof establishes a new problem-dependent acceleration mechanism for gradient methods, along with a notable local superlinear convergence guarantee. More generally, our *online scaled gradient methods* (`OSGM`) is a family of first-order methods that updates the stepsize on the fly using online learning on a convergence measure.

1.1 Contributions

This paper establishes the theoretical foundations of `OSGM` and showcases several instantiations of `OSGM`. Notable features of `OSGM` include:

Non-asymptotic complexity improvement on gradient descent. One instance of `OSGM` (**Section 6.1**) is guaranteed to find an ε-optimal solution in

$$K_\varepsilon = \left\lceil \min\left\{\kappa^\star L^2 \|P_1 - P_r^\star\|_F^2 + \kappa^\star \log\left(\tfrac{f(x^1)-f^\star}{\varepsilon}\right),\ \kappa \log(\tfrac{f(x^1)-f^\star}{\varepsilon})\right\}\right\rceil \tag{5}$$

iterations, where $P_r^\star$ is the stepsize that achieves the optimal condition number $\kappa^\star$. The complexity in (5) is no worse than $\mathcal{O}(\kappa \log(1/\varepsilon))$, the complexity of vanilla gradient descent. When $\varepsilon \to 0$, (5) becomes competitive with $\mathcal{O}(\kappa^\star \log(1/\varepsilon))$, the complexity achieved by $P_r^\star$. In general, `OSGM` guarantees a problem-dependent complexity $\mathcal{O}(\kappa^\star \log(1/\varepsilon))$ on smooth strongly convex problems, improving on classical accelerated methods when $\kappa^\star < \sqrt{\kappa}$ and $\varepsilon \to 0$. `OSGM` can be viewed as an acceleration mechanism of gradient descent that uses preconditioning to achieve acceleration. It differs from the existing momentum-based schemes [47, 48, 46].

Non-asymptotic local superlinear convergence. To our knowledge, `OSGM` represents the second family of first-order methods that achieve non-asymptotic local superlinear convergence after the celebrated quasi-Newton methods [49, 14]. The non-asymptotic superlinear convergence rate of `OSGM` matches or surpasses that of quasi-Newton methods, which were analyzed recently [58, 59, 60, 31, 30, 26, 27]. Our superlinear convergence analysis is simple and can be of independent interest.

Analysis of hypergradient descent heuristic. An instantiation of `OSGM` (**Section 6.2**) simplifies to hypergradient descent [62, 7], a popular heuristic that has received limited analysis [33]. `OSGM` provides a rigorous analysis of hypergradient descent [62, 7] through the lens of both online learning (**Theorem 6.6**) and provides new analysis tool using potential function (**Theorem 6.2**, **Theorem 6.7**). Our analysis explains the empirical success of hypergradient descent and motivates its practical extensions.

OSGM also complements the fruitful line of research in adaptive gradient methods [32, 13, 22], parameter-free first-order methods [34, 35], and parameter-free online learning algorithms [51, 52]. In particular, OSGM offers a new perspective to model stepsize selection as an online learning problem in which the sequence of decisions lies in the *stepsize space*. This perspective contrasts with the existing literature on adaptive gradient methods: online learning applied to the *primal space* and the stepsizes are chosen by a fixed rule to minimize a regret upper bound [13, 44, 21]. This shift in the perspective delivers sharp guarantees for smooth convex objectives. To our knowledge, the only prior work in a similar spirit is [70], which adapts the stepsize to stochastic noise in nonconvex optimization.

Section 2 introduces the OSGM framework along with a roadmap for the rest of the paper. **Section 9** surveys the related works in detail.

Not in this paper. OSGM also demonstrates excellent practical convergence. However, discussion of the setup most relevant to practical problems does not fit into this paper. We refer the interested readers to a preliminary conference version of our paper [9] and Part II of this paper.

1.2 Notations

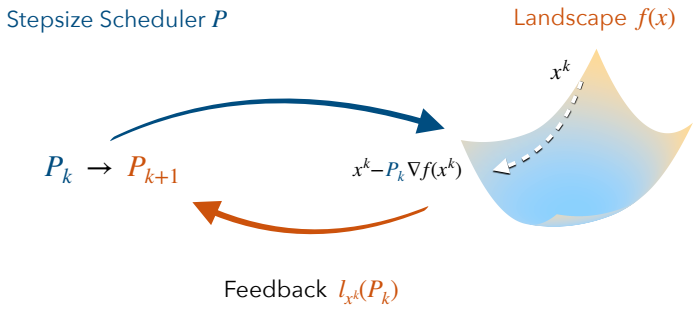
We use $\|\cdot\|$ to denote vector Euclidean norm or matrix operator norm and $\langle\cdot,\cdot\rangle$ to denote Euclidean or Frobenius inner product. Letters A, a denote matrices and scalars. $\|A\|_F \coloneqq \sqrt{\sum_{ij} a_{ij}^2}$ denotes the matrix Frobenius norm. Given a positive definite matrix $A \in \mathbb{S}^n_{++}$, we define $\|\cdot\|_A \coloneqq \sqrt{\langle x, Ax\rangle}$. Given two symmetric matrices A, B, $A \succeq B$ if $A - B$ is positive semidefinite. $\Pi_{\mathcal{P}}[\cdot]$ denotes the orthogonal projection onto a closed convex set $\mathcal{P}$. 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; $\mathrm{dist}(P, \mathcal{P}) \coloneqq \|P - \Pi_{\mathcal{P}}[P]\|_F$ denotes the distance between a point P and a closed convex set $\mathcal{P}$; $\mathrm{diam}(\mathcal{P}) = \max_{X,Y\in\mathcal{P}} \|X - Y\|_F$ denotes the diameter of set $\mathcal{P}$ in Frobenius norm. 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 P_k to index stepsize sequence.

2 Online scaled gradient methods

OSGM models stepsize selection as a sequential decision-making problem in a collaborative environment. There are two agents: a stepsize scheduling agent (Scheduler) and a landscape agent (Landscape). Scheduler chooses stepsize from a closed convex candidate set $\mathcal{P} \subseteq \mathbb{R}^{n\times n}$. An iteration of OSGM has three steps (**Algorithm 1**):

- Step 1. Scheduler makes decision $P_k \in \mathcal{P}$ and proposes an update $x^{k+1/2} = x^k - P_k \nabla f(x^k)$.
- Step 2. Landscape queries the objective function f and 1) chooses the next iterate $x^{k+1} = \mathcal{M}(x^k, x^{k+1/2})$ based on current x^k and the proposal $x^{k+1/2}$; 2) evaluates the quality of P_k and provides feedback $\ell_{x^k}(P_k)$ to the scheduler.
- Step 3. Scheduler updates the stepsize using an online learning algorithm $P_{k+1} = \mathcal{A}(P_k, \{\ell_{x^j}\}_{j\le k})$.

Feedback ℓ_x, landscape action $\mathcal{M}$, and online algorithm $\mathcal{A}$ combine to yield different variants of OSGM.



Algorithm 1: Online scaled gradient methods (`OSGM`)

input Initial point x^1, initial stepsize P_1, feedback ℓ_x, online algorithm $\mathcal{A}$, landscape action $\mathcal{M}$
for $k = 1, 2, \ldots$ **do**
  $x^{k+1/2} = x^k - P_k \nabla f(x^k)$ ▷ Scheduler suggests the next step based on P_k
  $x^{k+1} = \mathcal{M}(x^k, x^{k+1/2})$ ▷ Landscape evaluates P_k, chooses x^{k+1}, and gives feedback $\ell_{x^k}(P_k)$
  $P_{k+1} = \mathcal{A}(P_k, \{\ell_j\}_{j \le k})$ ▷ Scheduler learns a better stepsize from feedback
end

Comparison with online learning. Our framework builds on online learning but slightly differs in the spirit: the environment (Landscape) in online learning is often assumed to be adversarial, while the landscape in `OSGM` shares the same objective of minimizing f and collaborates with the scheduler in two ways: 1) Landscape provides faithful feedback to the scheduler to improve the stepsize. 2) Landscape can reject bad decisions P_k made by the scheduler and suppress regret in online learning through the landscape action.

Structure of the paper. The paper is organized as follows to introduce different components of `OSGM`.

- *Feedback* $\ell_x(P)$. A function that locally measures the quality of stepsize P at x. In our context, *smaller* feedback means *better* quality. **Section 3** introduces two important feedback functions.
- *Landscape action* $\mathcal{M}$. Landscape selectively accepts stepsizes P_k with small feedback and suppresses the regret of the learning algorithm $\mathcal{A}$. **Section 4** discusses different landscape actions and their consequences.
- *Learning algorithm* $\mathcal{A}$. An online learning algorithm that guarantees performance concerning feedback $\ell_x(P)$. In our context, the guarantee refers to the (sublinear) regret with respect to any stepsize $P \in \mathcal{P}$:

$$\sum_{k=1}^K \ell_{x^k}(P_k) \le \sum_{k=1}^K \ell_{x^k}(P) + o(K).$$

  Section 5 discusses the regret guarantees of online gradient descent on our feedback functions.

After establishing these components, we introduce several important instantiations of `OSGM` in **Section 6** and discuss the practical aspects of `OSGM` in **Section 7**.

3 Feedback design and analysis

We introduce two feedback functions $\ell_x(P)$: *ratio* feedback and *hypergradient* feedback[2]. The ratio feedback requires knowledge of the optimal value $f^\star$ and has strong theoretical guarantees. The hypergradient feedback does not require $f^\star$ and is more practical. We analyze the properties of each feedback function and associate each feedback function with a minimax optimal stepsize.

3.1 Ratio feedback

Given any $x \notin \mathcal{X}^\star$, the *ratio feedback* $r_x(P)$ measures the quality of stepsize P by the contraction ratio of the suboptimality as the iterate moves from x to $x - P\nabla f(x)$:

$$r_x(P) := \tfrac{f(x - P\nabla f(x)) - f^\star}{f(x) - f^\star}.$$

The ratio feedback inherits both convexity and smoothness from f. Since $r_x(P)$ simply translates and scales the function value at the proposed iterate $u_x(P) := f(x - P\nabla f(x))$, we begin by analyzing $u_x(P)$.

Proposition 3.1 (Properties of u_x)**.** *For any $x \in \mathbb{R}^n$, the function $u_x(P) = f(x - P\nabla f(x))$ has gradient*

$$\nabla u_x(P) = -\nabla f(x - P\nabla f(x)) \nabla f(x)^\top.$$

[2]We consider two feedback functions in this paper for simplicity. More feedback functions can be found in [15, 9]

Moreover, it satisfies the following properties:

1. *If f is convex and L-smooth, then $u_x(P)$ is convex and $L\|\nabla f(x)\|^2$-smooth in P.*
2. *If f is L-smooth and $\operatorname{diam}(\mathcal{P}) \le D$, then $u_x(P)$ is $(LD+1)\|\nabla f(x)\|^2$-Lipschitz in P.*

With the properties of u_x, the analytic properties of $r_x(P)$ follow immediately.

Lemma 3.1 (Properties of r_x)**.** *For any $x \notin \mathcal{X}^\star$, the ratio feedback $r_x(P)$ has gradient*

$$\nabla r_x(P) = -\tfrac{\nabla f(x - P\nabla f(x))\nabla f(x)^\top}{f(x) - f^\star}.$$

Moreover, it satisfies the following properties:

1. *If f is convex and L-smooth, then $r_x(P)$ is convex, non-negative, and $2L^2$-smooth in P.*
2. *If f is L-smooth and $\operatorname{diam}(\mathcal{P}) \le D$, then $r_x(P)$ is $2L(LD+1)$-Lipschitz in P.*

3.2 Hypergradient feedback

Given any $x \notin \mathcal{X}^\star$, the *hypergradient feedback* $h_x(P)$ measures the quality of P by the function value progress relative to the size of the gradient as the iterate moves from x to $x - P\nabla f(x)$:

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

The hypergradient feedback is motivated by the descent lemma for L-smooth functions:

$$f(x - \tfrac{1}{L}\nabla f(x)) - f(x) \le -\tfrac{1}{2L}\|\nabla f(x)\|^2,$$

which states that the improvement in the function value of a gradient step with stepsize $\frac{1}{L}I$ is proportional to the size of the gradient $\|\nabla f(x)\|^2$ with ratio $\frac{1}{2L}$. In a similar vein, hypergradient feedback $h_x(P)$ quantifies the progress in terms of this ratio when stepsize P is used. The properties of hypergradient feedback $h_x(P)$ also follow from **Proposition 3.1**.

Lemma 3.2 (Properties of h_x)**.** *For any $x \notin \mathcal{X}^\star$, the hypergradient feedback $h_x(P)$ has gradient*

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

Moreover, it satisfies the following properties:

1. *If f is convex and L-smooth, then $h_x(P)$ is convex and L-smooth in P.*
2. *If f is L-smooth and $\operatorname{diam}(\mathcal{P}) \le D$, then $h_x(P)$ is $(LD+1)$-Lipschitz in P.*

3.3 Minimax optimal stepsizes

Given a feedback function $\ell_x(P)$, we can define its corresponding minimax optimal stepsize $P_\ell^\star$ by

$$P_\ell^\star \in \operatorname*{arg\,min}_{P \in \mathcal{P}} \max_{x \notin \mathcal{X}^\star} \ell_x(P).$$

Minimax optimal stepsize $P_\ell^\star$ represents the best decision the scheduler can make when x is selected adversarially. We are interested in the feedback achieved by the minimax optimal stepsize $\ell_x(P_\ell^\star)$ since it represents the performance of the scheduler in the worst case. The minimax optimal feedback $\ell_x(P_\ell^\star)$ can be bounded by quantities studied in the literature. For example, when $\frac{1}{L}I \in \mathcal{P}$, the minimax optimal hypergradient feedback $h_x(P_h^\star)$ is no worse than $h_x(\frac{1}{L}I) \le -\frac{1}{2L}$ guaranteed by the descent lemma. For ratio feedback, the performance of $P_r^\star$ can be quantified using the globally optimal preconditioner defined below.

Definition 3.1 (Globally optimal preconditioner [33]). *Suppose f is L-smooth, μ-strongly convex, and twice-differentiable. Given a candidate set of stepsizes $\mathcal{P}$ such that $\frac{1}{L}I \in \mathcal{P}$, the globally optimal preconditioner $P_+^\star \in \mathbb{S}_+^n$ and the optimal condition number $\kappa^\star$ are defined as the optimal solution and the optimal value of*

$$\min_{\hat{\kappa}\geq 0, P\in\mathcal{P}\cap\mathbb{S}_+^n} \hat{\kappa} \quad \text{subject to} \quad \tfrac{1}{\hat{\kappa}}P^{-1} \preceq \nabla^2 f(x) \preceq P^{-1} \quad \text{for all } x.$$

The globally optimal condition number satisfies $\kappa^\star \leq \kappa = L/\mu$ since $\frac{1}{L}I \in \mathcal{P}$. In practice, recent works [16, 55, 33] observe that $\kappa^\star \ll \kappa$ is often the case. Gradient descent with stepsize $P_+^\star$ guarantees a contraction ratio of $1 - \frac{1}{\kappa^\star}$ (see Appendix **A.4** for the proof):

$$f(x - P_+^\star \nabla f(x)) - f^\star \leq (1 - \tfrac{1}{\kappa^\star})[f(x) - f^\star]. \tag{6}$$

Since $P_+^\star \in \mathcal{P}$, the ratio feedback of the globally optimal preconditioner $r_x(P_+^\star)$ upper bounds that of the minimax optimal stepsize $r_x(P_r^\star)$. The bound is tight on strongly convex quadratics [15].

Proposition 3.2 (Feedback of the minimax stepsize). *For any $x \notin \mathcal{X}^\star$, the minimax optimal stepsizes satisfy:*

- *If f is L-smooth and μ-strongly convex, then $r_x(P_r^\star) \leq 1 - \frac{1}{\kappa^\star}$.*
- *If f is L-smooth and $\frac{1}{L}I \in \mathcal{P}$, then $h_x(P_h^\star) \leq -\frac{1}{2L}$.*

In addition, if f is a strongly convex quadratic, then the minimax optimal stepsizes are $P_r^\star = P_+^\star$ and $P_h^\star = \frac{1}{L}I$.

Limitations of minimax optimal stepsize. Minimax stepsizes are useful for establishing global convergence results. However, they are not affine invariant and may not reflect the properties of the local landscape. For example, the globally optimal condition number $\kappa^\star$ depends on the choice of coordinate system when $\mathcal{P} \neq \mathbb{R}^{n\times n}$.

Example 3.1 (Limitations of globally optimal preconditioner). *Consider two strongly convex quadratics*

$$f_1(x) = \tfrac{1}{2}\langle x, \Lambda x\rangle, \quad f_2(x) = \tfrac{1}{2}\langle x, T_n x\rangle, \qquad T_n = \begin{pmatrix} 2 & -1 & & & \\ -1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix},$$

and Λ has eigenvalues of T_n on the diagonal $\lambda_k = 4\sin^2\left[\frac{\pi k}{2(n+1)}\right]$ for $k = 1, \ldots, n$. The functions f_1 and f_2 are related by a rotation: they share the same spectrum, and thus the same smoothness constant, strong convexity constant, and condition number $\kappa(T_n) = \Theta(n^2)$. However, their optimal condition numbers with respect to diagonal stepsizes are different:

$$\begin{aligned} \kappa_{f_1}^\star &= 1 \quad \textit{with optimal diagonal stepsize } P_1^\star = \Lambda^{-1}; \\ \kappa_{f_2}^\star &= \kappa(T_n) \quad \textit{with optimal diagonal stepsize } P_2^\star = I_n. \end{aligned}$$

The optimal diagonal preconditioner for f_1 gives perfect conditioning $\kappa_{f_1}^\star = 1$, yet no diagonal stepsize improves the conditioning of f_2. The orientation of eigenvectors affects the optimal condition number $\kappa^\star$ when $\mathcal{P} \neq \mathbb{R}^{n\times n}$.

4 Progress reduction and landscape action

The example algorithm in **Section 1** is based on a simple observation that the suboptimality after K iterations can be bounded by the sum of contraction ratios at each step (see (2)). This observation, which we call *reduction*, reduces the optimization problem $\min_{x\in\mathbb{R}^n} f(x)$ to minimizing the sum of *per iteration progress*. The per iteration progress with respect to ratio feedback and hypergradient feedback is measured by

$$r_k \coloneqq \tfrac{f(x^{k+1}) - f^\star}{f(x^k) - f^\star}, \qquad h_k \coloneqq \tfrac{f(x^{k+1}) - f(x^k)}{\|\nabla f(x^k)\|^2}.$$

Two reductions below guarantee that smaller cumulative progress $\sum_{k=1}^K r_k$ or $\sum_{k=1}^K h_k$ yields faster convergence.

Theorem 4.1 (Reduction for r_k)**.** *Let $\{x^k\}$ be any sequence such that $x^k \notin \mathcal{X}^\star$. Then*

$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](\tfrac{1}{K}\textstyle\sum_{k=1}^K r_k)^K.$$

Theorem 4.2 (Reductions for h_k)**.** *Let $\{x^k\}$ be any sequence such that $x^k \notin \mathcal{X}^\star$ and $f(x^{k+1}) \le f(x^k)$.*

- *If f is convex, then*
$$f(x^{K+1}) - f^\star \le \min\{\tfrac{\Delta^2}{K}\tfrac{1}{\frac{1}{K}\sum_{k=1}^K -h_k}, f(x^1) - f^\star\},$$
where $\Delta \coloneqq \max_{x\in\{x: f(x)\le f(x^1)\}} \min_{x^\star\in\mathcal{X}^\star} \|x - x^\star\|$.
- *If f is μ-strongly convex, then*
$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](1 - \tfrac{2\mu}{K}\textstyle\sum_{k=1}^K -h_k)^K.$$

These two reductions are blackbox: they are independent of the mechanism generating the iterates $\{x^k\}$ and universally apply to (monotone) algorithms other than `OSGM`. **Theorem 4.1** even does not require f to be convex. **Theorem 4.2** comes with a restriction of non-increasing function values $f(x^k)$. In `OSGM`, when the landscape always accepts the scheduler's proposal by taking $x^{k+1} = x^{k+1/2}$, the algorithm may not be monotone, and the reductions for h_k in **Theorem 4.2** may not apply. However, we can design the landscape agent so that it takes action that filters out bad stepsizes and ensures monotonicity of the iterates.

4.1 Landscape actions $\mathcal{M}$ and progress

The landscape action in `OSGM` is denoted by $x^{k+1} = \mathcal{M}(x^{k+1/2}, x^k)$. We consider four landscape actions below.

- *Vanilla.* $x^{k+1} = x^{k+1/2}$

  Landscape accepts all decisions from the scheduler.

- *Monotone.* x^{k+1} satisfies $f(x^{k+1}) \le \min\{f(x^{k+1/2}), f(x^k)\}$

  Landscape moves to x^{k+1} that is no worse than the current iterate x^k and the suggested iterate $x^{k+1/2}$ in terms of objective value. This action can be implemented by line search or a null step: $x^{k+1} = \arg\min_{x\in\{x^{k+1/2}, x^k\}} f(x)$.

- *Lookahead.* $x^{k+1} = x^{k+1/2} - \frac{1}{L}\nabla f(x^{k+1/2})$

  Landscape takes an additional gradient descent step on top of the suggested iterate $x^{k+1/2}$ but does not enforce monotonicity.

- *Monotone Lookahead.* x^{k+1} satisfies $f(x^{k+1}) \le \min\{f(x^{k+1/2} - \frac{1}{L}\nabla f(x^{k+1/2})), f(x^k)\}$

  Landscape moves to x^{k+1} that is no worse than the current iterate x^k and the lookahead iterate $x^{k+1/2} - \frac{1}{L}\nabla f(x^{k+1/2})$ in terms of objective value.

All these actions, except vanilla, improve the iteration progress r_k and h_k upon the feedback $r_{x^k}(P_k)$ and $h_{x^k}(P_k)$ achieved by the scheduler, which we illustrate in **Lemma 4.1** below.

Lemma 4.1 (Feedback and progress)**.** *Let f be convex and L-smooth. Each of the above four landscape actions guarantees the following relation between the feedback and per iteration progress:*

- *Vanilla. $r_k = r_{x^k}(P_k)$ and $h_k = h_{x^k}(P_k)$.*
- *Monotone. $r_k \le \min\{r_{x^k}(P_k), 1\}$ and $h_k \le \min\{h_{x^k}(P_k), 0\}$.*
- *Lookahead. $r_k \le r_{x^k}(P_k) - \frac{1}{4L^2}\|\nabla r_{x^k}(P_k)\|_F^2$ and $h_k \le h_{x^k}(P_k) - \frac{1}{2L}\|\nabla h_{x^k}(P_k)\|_F^2$.*
- *Monotone Lookahead. $r_k \le \min\{r_{x^k}(P_k) - \frac{1}{4L^2}\|\nabla r_{x^k}(P_k)\|_F^2, 1\}$ and $h_k \le \min\{h_{x^k}(P_k) - \frac{1}{2L}\|\nabla h_{x^k}(P_k)\|_F^2, 0\}$.*

The per iteration progress r_k and h_k under these landscape actions are never worse than the feedback $r_{x^k}(P_k)$ and $h_{x^k}(P_k)$. To accelerate convergence, it suffices for the scheduler to minimize the cumulative feedback $\sum_{k=1}^K r_{x^k}(P_k)$ and $\sum_{k=1}^K h_{x^k}(P_k)$, a task well-suited for online learning algorithms.

5 Stepsize update and online learning

The scheduler aims to generate a sequence of stepsizes $\{P_k\}$ that reduces the cumulative feedback $\sum_{k=1}^K \ell_{x^k}(P_k)$ as much as possible. This can be done by the existing online learning algorithms with sublinear regret guarantees:

$$\frac{1}{K}\sum_{k=1}^K \ell_{x^k}(P_k) \le \frac{1}{K}\sum_{k=1}^K \ell_{x^k}(\hat{P}) + o(1) \quad \text{for any } \hat{P} \in \mathcal{P}. \tag{7}$$

This guarantee (7) says that the average feedback $\frac{1}{K}\sum_{k=1}^K \ell_{x^k}(P_k)$ achieved by the scheduler is asymptotically no worse than that achieved by any fixed stepsize $\hat{P} \in \mathcal{P}$. We consider the online learning algorithm $\mathcal{A}$ to be *online gradient descent* in this paper for concreteness, but our arguments for the convergence of `OSGM` apply to other online learning algorithms as well. We will discuss the practical choice of online algorithm in **Section 7**.

5.1 Regret guarantees of online gradient descent

Online gradient descent (`OGD`) updates the stepsize by

$$P_{k+1} = \Pi_{\mathcal{P}}[P_k - \eta_k \nabla \ell_{x^k}(P_k)], \tag{8}$$

where $\{\eta_k\}$ is a sequence of non-negative and non-increasing online learning stepsizes which will be specified later and $\Pi_{\mathcal{P}}$ is the projection operator onto the candidate set $\mathcal{P}$. `OGD` on convex feedback ℓ_{x^k} satisfies

$$\|P_{k+1} - \hat{P}\|_F^2 \le \|P_k - \hat{P}\|_F^2 - 2\eta_k[\ell_{x^k}(P_k) - \ell_{x^k}(\hat{P})] + \eta_k^2 \|\nabla \ell_{x^k}(P_k)\|_F^2 \quad \text{for any } \hat{P} \in \mathcal{P}. \tag{9}$$

According to (9), whenever the scheduler makes a stepsize decision P_k that underperforms $\hat{P}$ (i.e., $\ell_{x^k}(P_k) - \ell_{x^k}(\hat{P}) > 0$), the refined stepsize P_{k+1} approaches $\hat{P}$ up to some error $\eta_k^2\|\nabla \ell_{x^k}(P_k)\|_F^2$. Rearranging the inequality (9) and telescoping from $k = 1$ to K, we obtain the following regret guarantee:

Lemma 5.1 (Static regret)**.** *Suppose $\{\ell_{x^k}(P)\}$ are convex in P. Then `OGD` with constant stepsize $\eta_k \equiv \eta > 0$ generates a sequence $\{P_k\}$ such that*

$$\sum_{k=1}^K \ell_{x^k}(P_k) \le \sum_{k=1}^K \ell_{x^k}(\hat{P}) + \frac{1}{2\eta}\|P_1 - \hat{P}\|_F^2 + \frac{\eta}{2}\sum_{k=1}^K \|\nabla \ell_{x^k}(P_k)\|_F^2 \quad \textit{for any } \hat{P} \in \mathcal{P}. \tag{10}$$

Besides, if $\{\ell_{x^k}(P)\}$ are σ-Lipschitz w.r.t. $\|\cdot\|_F$ and $\operatorname{diam}(\mathcal{P}) \le D$, then `OGD` with $\eta_k \equiv \frac{c}{\sqrt{K}}$ or $\eta_k = \frac{c}{\sqrt{k}}$ satisfies

$$\sum_{k=1}^K \ell_{x^k}(P_k) \le \sum_{k=1}^K \ell_{x^k}(\hat{P}) + (\tfrac{D^2}{2c} + c\sigma^2)\sqrt{K} \quad \textit{for any } \hat{P} \in \mathcal{P}. \tag{11}$$

Lemma 5.1 competes the adaptive stepsizes $\{P_k\}$ against a static stepsize $\hat{P} \in \mathcal{P}$. A more refined analysis of online gradient descent, referred to as *dynamic regret* [19, 51], enables the scheduler to compete with a sequence of stepsizes $\{\hat{P}_k\}, \hat{P}_k \in \mathcal{P}$, at the cost of the *path length* defined by

$$\mathsf{PL}(\{\hat{P}_k\}) := \sum_{k=1}^{K-1} \|\hat{P}_k - \hat{P}_{k+1}\|_F. \tag{12}$$

Lemma 5.2 (Dynamic regret)**.** *Suppose $\{\ell_{x^k}(P)\}$ are convex in P. For any benchmark sequence of stepsizes $\{\hat{P}_k\}, P_k \in \mathcal{P}$, `OGD` with constant stepsize $\eta > 0$ generates $\{P_k\}$ such that*

$$\sum_{k=1}^K \ell_{x^k}(P_k) \le \sum_{k=1}^K \ell_{x^k}(\hat{P}_k) + \frac{\eta}{2}\sum_{k=1}^K \|\nabla \ell_{x^k}(P_k)\|_F^2 + \frac{\|\hat{P}_K - P_1\|_F^2}{2\eta} + \frac{\max_{k\le K}\|P_k - P_1\|_F}{\eta}\mathsf{PL}(\{\hat{P}_k\}). \tag{13}$$

Besides, if $\{\ell_{x^k}(P)\}$ are σ-Lipschitz w.r.t. $\|\cdot\|_F$ and $\operatorname{diam}(\mathcal{P}) \le D$, then `OGD` with $\eta_k \equiv \frac{c}{\sqrt{K}}$ satisfies

$$\sum_{k=1}^K \ell_{x^k}(P_k) \le \sum_{k=1}^K \ell_{x^k}(\hat{P}_k) + [\tfrac{c\sigma^2}{2} + \tfrac{D^2}{2c} + \tfrac{D}{c}\mathsf{PL}(\{\hat{P}_k\})]\sqrt{K} \quad \textit{for any } \hat{P}_k \in \mathcal{P}. \tag{14}$$

When $\hat{P}_k \equiv \hat{P}$, the path length vanishes $\mathsf{PL}(\{\hat{P}_k\}) = 0$ and dynamic regret (**Lemma 5.2**) reduce to static regret (**Lemma 5.1**). The relation (13) holds for *any* sequence $\{\hat{P}_k\}, \hat{P}_k \in \mathcal{P}$, indicating that `OGD` is asymptotically

competitive with the optimal sequence $\{\hat{P}_k\}$ that minimizes the right-hand side of the bounds. The optimal sequence balances the trade-off between the cumulative feedback $\sum_{k=1}^{K} \ell_{x^k}(\hat{P}_k)$ and the path length $\mathsf{PL}(\{\hat{P}_k\})$. **Example 5.1** illustrates the effect of dynamic regret guarantees.

Example 5.1. *Consider diagonal stepsizes* $\{P_k\}, P_k \in \mathcal{P} := \{P = \mathrm{Diag}(d) : d \in \mathbb{R}^n\}$ *for a 2-dimensional smooth convex objective defined as*

$$f(x_1, x_2) = \begin{cases} \frac{1}{4}x_1^2 + \frac{1}{2}x_2^2, & \text{if } (x_1, x_2) \in R_1 := \{(x_1, x_2) : x_1 \geq 0\}; \\ \frac{3}{4}x_1^2 + \frac{1}{2}x_2^2, & \text{if } (x_1, x_2) \in R_2 := \{(x_1, x_2) : x_1 < 0\}. \end{cases}$$

The function f is strongly convex and piecewise quadratic with different curvatures in the regions R_1 and R_2. For $x \in R_1$, the Hessian inverse $\mathrm{Diag}(0.5, 1)$ achieves zero ratio feedback $r_x(\mathrm{Diag}(0.5, 1)) = 0$ since it sends any $x \in R_1$ to the optimal solution $x^\star = (0, 0)$ in one preconditioned gradient step, resulting in $f(x - \mathrm{Diag}(0.5, 1)\nabla f(x)) - f^\star = 0$. Similarly, for $x \in R_2$, the Hessian inverse $\mathrm{Diag}(1.5, 1)$ also achieves zero ratio feedback. The trajectory $\{x^k\}$ of gradient descent (with appropriate stepsize) remains in the same region as the initial point x^1, ensuring that the common Hessian inverse on that region minimizes the cumulative feedback over the trajectory with an optimal value of zero: $\min_{P \in \mathcal{P}} \sum_{k=1}^{K} r_{x^k}(P) = 0$.

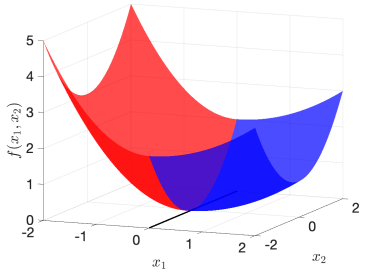


Figure 1: Illustration of $f(x_1, x_2)$.

However, no diagonal stepsize can achieve $r_x(P) = 0$ for all x simultaneously. Thus, the minimax optimal stepsize $P_r^\star$ must incur positive cumulative feedback $\sum_{k=1}^{K} r_{x^k}(P_r^\star) > 0$, which is strictly worse than the trajectory-based bound. Even if the trajectory $\{x^k\}$ does not always stay in the same region, the dynamic regret bound in (14) allows us to compare with a benchmark sequence $\{\hat{P}_k\}$. For example, let $\hat{P}_k = \mathrm{Diag}(0.5, 1)$ if $x^k \in R_1$; and $\hat{P}_k = \mathrm{Diag}(1.5, 1)$ if $x^k \in R_2$. Then $\sum_{k=1}^{K} r_{x^k}(\hat{P}_k) = 0$ and the path length $\mathsf{PL}(\{\hat{P}_k\}) = \sum_{k=1}^{K-1} \|\hat{P}_k - \hat{P}_{k+1}\|_F$ is proportional to the number of times the trajectory $\{x^k\}$ switches from one region to the other.

6 Algorithm design and analysis

This section demonstrates the convergence guarantees of `OSGM` through two variants, one with ratio feedback and the other with hypergradient feedback. We denote each variant of `OSGM` as `{Action} OSGM-{Feedback}`, where `Action` refers to the landscape action and `Feedback` $\in \{\mathtt{R}, \mathtt{H}\}$ represents the initial letter of the type of feedback. Without loss of generality, we assume $x^k \notin \mathcal{X}^\star$; otherwise the algorithm can be stopped.

6.1 `Lookahead OSGM-R`

In this section, we assume f is L-smooth and μ-strongly convex and instantiate `OSGM` with

$$\ell_x(P) := r_x(P) \qquad \text{Lookahead landscape: } x^{k+1} = x^{k+1/2} - \tfrac{1}{L}\nabla f(x^{k+1/2}) \qquad \mathcal{A} := \text{Online gradient descent.}$$

The algorithm is called `Lookahead OSGM-R` (**Algorithm 2**). The candidate stepsize set is $\mathcal{P} = \mathbb{R}^{n \times n}$.

Algorithm 2: `Lookahead OSGM-R`

input: Initial point x^1, initial stepsize $P_1 \in \mathcal{P} = \mathbb{R}^{n \times n}$, online gradient stepsize $\eta_k \equiv \eta > 0$
for $k = 1, 2, \ldots$ **do**
  $x^{k+1/2} = x^k - P_k \nabla f(x^k)$
  $x^{k+1} = x^{k+1/2} - \frac{1}{L}\nabla f(x^{k+1/2})$
  $P_{k+1} = P_k - \eta \nabla r_{x^k}(P_k)$
end

Recall that the gradient of ratio feedback in Line **5** of **Algorithm 2** takes the form

$$\nabla r_{x^k}(P_k) = -\frac{\nabla f(x^{k+1/2})\nabla f(x^k)^\top}{f(x^k)-f^\star}.$$

The rest of the section analyzes the convergence behavior of `Lookahead OSGM-R`, including global convergence and local superlinear convergence.

6.1.1 Global convergence

We analyze the convergence of `Lookahead OSGM-R` in three steps. First, according to **Theorem 4.1**, the convergence of `Lookahead OSGM-R` follows from bounding the suboptimality in terms of the cumulative progress

$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](\tfrac{1}{K}\textstyle\sum_{k=1}^K r_k)^K. \tag{15}$$

Second, the cumulative progress $\sum_{k=1}^K r_k$ under lookahead action is bounded by (**Lemma 4.1**):

$$\textstyle\sum_{k=1}^K r_k \le \sum_{k=1}^K r_{x^k}(P_k) - \frac{1}{4L^2}\sum_{k=1}^K \|\nabla r_{x^k}(P_k)\|_F^2. \tag{16}$$

Third, the regret guarantee of online gradient descent bounds the cumulative feedback $\sum_{k=1}^K r_{x^k}(P_k)$ by (**Lemma 5.1**):

$$\textstyle\sum_{k=1}^K r_{x^k}(P_k) \le \sum_{k=1}^K r_{x^k}(\hat{P}) + \frac{1}{2\eta}\|P_1 - \hat{P}\|_F^2 + \frac{\eta}{2}\sum_{k=1}^K \|\nabla r_{x^k}(P_k)\|_F^2. \tag{17}$$

Putting (16) and (17) together, the cumulative progress is bounded by

$$\textstyle\sum_{k=1}^K r_k \le \sum_{k=1}^K r_{x^k}(\hat{P}) + \frac{1}{2\eta}\|P_1 - \hat{P}\|_F^2 + (\frac{\eta}{2} - \frac{1}{4L^2})\sum_{k=1}^K \|\nabla r_{x^k}(P_k)\|_F^2. \tag{18}$$

When $\eta \le \frac{1}{2L^2}$, lookahead action helps the scheduler suppress the $\frac{\eta}{2}\|\nabla r_{x^k}(P_k)\|_F^2$ error term from online gradient descent. In particular, stepsize $\eta = \frac{1}{2L^2}$ simplifies (18) to

$$\textstyle\sum_{k=1}^K r_k \le \sum_{k=1}^K r_{x^k}(\hat{P}) + L^2\|P_1 - \hat{P}\|_F^2. \tag{19}$$

The global convergence guarantee follows immediately by plugging (19) into the reduction (15), summarized in **Theorem 6.1** below. The collaboration between landscape and scheduler allows `Lookahead OSGM-R` to converge as if the online algorithm incurred only constant regret.

Theorem 6.1 (Global convergence)**.** *Let f be L-smooth and μ-strongly convex. For any benchmark stepsize $\hat{P} \in \mathbb{R}^{n\times n}$, `Lookahead OSGM-R` (**Algorithm 2**) with $\eta = 1/(2L^2)$ satisfies*

$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](\tfrac{1}{K}\textstyle\sum_{k=1}^K r_{x^k}(\hat{P}) + \frac{L^2\|P_1-\hat{P}\|_F^2}{K})^K. \tag{20}$$

In particular, if `Lookahead OSGM-R` is initialized with $P_1 = \frac{1}{L}I$, then

$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star]\min\{(1-\tfrac{1}{\kappa})^K, (1 - \tfrac{1}{\kappa^\star} + \tfrac{L^2}{K}\|\tfrac{1}{L}I - P_r^\star\|_F^2)^K\}, \tag{21}$$

*where κ and $\kappa^\star$ are the condition number and optimal condition number (**Definition 3.1**) of f.*

The convergence guarantee in (20) is powerful since it holds for any benchmark stepsize $\hat{P}$. In particular, one can choose the benchmark stepsize $\hat{P}$ that achieves the best average feedback along the algorithm trajectory. Moreover, if `Lookahead OSGM-R` is initialized with $P_1 = \frac{1}{L}I$, according to (21), `Lookahead OSGM-R` can automatically adapt to the best convergence rate among $1 - \frac{1}{\kappa}$ and $1 - \frac{1}{\kappa^\star} + \frac{L^2}{K}\|\frac{1}{L}I - P_r^\star\|_F^2$. The former is the rate of vanilla gradient descent, and the latter is asymptotically the rate of gradient descent with the optimal stepsize. When K is small, the first rate $1 - \frac{1}{\kappa}$ is smaller; when K is large, the second is smaller. Relation (21) guarantees the best-of-both-worlds: `Lookahead OSGM-R` at least matches vanilla gradient descent, and asymptotically converges at least as fast as gradient descent with the best possible stepsize. **Figure 2** illustrates the

expected convergence behavior. An asymptotic $\mathcal{O}(\kappa^\star \log(1/\varepsilon))$ complexity can be obtained from **Theorem 6.1**.

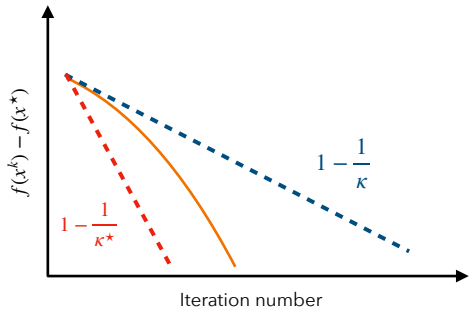


Figure 2: Theoretical performance of `Lookahead OSGM-R`. The linear convergence rate initially matches the $1 - \frac{1}{\kappa}$ rate of vanilla gradient descent and accelerates to (at least) $1 - \frac{1}{\kappa^\star}$.

Alternatively, we can obtain an explicit complexity bound through the lens of a potential function.

Potential reduction. Let $\hat{P}$ be a benchmark stepsize such that $r_x(\hat{P}) \leq 1 - \frac{1}{\kappa_{\hat{P}}} < 1$ for all x. For example, it suffices to take the optimal stepsize $\hat{P} = P_r^\star$ with its associated optimal condition number $\kappa_{\hat{P}} = \kappa^\star$. In each iteration of `Lookahead OSGM-R`, either

i) the feedback $r_{x^k}(P_k)$ is smaller (better) than $r_{x^k}(\hat{P}) < 1$ and the suboptimality contracts:

$$f(x^{k+1}) - f^\star = r_k[f(x^k) - f^\star] < f(x^k) - f^\star,$$

ii) or the feedback $r_{x^k}(P_k)$ is larger (worse) than $r_{x^k}(\hat{P})$ but $\|P_{k+1} - \hat{P}\|_F^2$ shrinks:

$$\|P_{k+1} - \hat{P}\|_F^2 \leq \|P_k - \hat{P}\|_F^2 - 2\eta[r_{x^k}(P_k) - r_{x^k}(\hat{P})] + \eta^2 \|\nabla r_{x^k}(P_k)\|_F^2.$$

This observation motivates the definition of a joint potential in x and P, parametrized by the benchmark $\hat{P}$:

$$\varphi(x, P) := \rho \log(f(x) - f^\star) + \|P - \hat{P}\|_F^2, \quad \rho > 0.$$

This potential combines the primal suboptimality $f(x) - f^\star$ and the distance to the benchmark stepsize $\|P - \hat{P}\|_F^2$. `Lookahead OSGM-R` decreases this potential by at least a constant at every iteration.

Theorem 6.2 (Potential reduction)**.** *Let f be L-smooth and μ-strongly convex. At every iteration, `Lookahead OSGM-R` with $\eta = 1/(2L^2)$ decreases the potential $\varphi(x, P)$ with $\rho = 1/L^2$ by*

$$\varphi(x^{k+1}, P_{k+1}) \leq \varphi(x^k, P_k) - \tfrac{1}{\kappa_{\hat{P}} L^2}.$$

In particular, `Lookahead OSGM-R` finds an ε-optimal solution within complexity

$$K_\varepsilon := \left\lceil \min_{\hat{P}} \left\{ \kappa_{\hat{P}} L^2 \|P_1 - \hat{P}\|_F^2 + \kappa_{\hat{P}} \log\left(\tfrac{f(x^1) - f^\star}{\varepsilon}\right) \right\} \right\rceil.$$

The simple algorithm `Lookahead OSGM-R` (**Algorithm 2**) delivers strong worst-case convergence guarantees without additional assumptions beyond smoothness and strong convexity. In the next subsection, we show that `Lookahead OSGM-R` also achieves strong problem-dependent convergence as well as local convergence guarantees.

6.1.2 Local convergence

Online gradient descent replaces stale information with new updates, which allows `OSGM` to adapt to the local landscape. This subsection highlights three consequences of this adaptivity: 1) trajectory-level adaptivity to

the local landscape 2) local superlinear convergence 3) asymptotic optimality compared to any fixed stepsize.

Trajectory-level adaptivity. We combine **Theorem 4.1** and the dynamic regret guarantee in **Lemma 5.2** to show local adaptivity to the trajectory.

Theorem 6.3 (Local adaptivity)**.** *Let f be L-smooth and μ-strongly convex. For any benchmark sequence of stepsizes $\{\hat{P}_k\}$, `Lookahead OSGM-R` with $\eta = 1/(2L^2)$ satisfies*

$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](\tfrac{1}{K}\textstyle\sum_{k=1}^K r_{x^k}(\hat{P}_k) + \tfrac{L^2}{K}[\|\hat{P}_K - P_1\|_F^2 + 2\max_{k\le K}\{\|P_k - P_1\|_F\}\mathsf{PL}(\{\hat{P}_k\})])^K,$$

where $\mathsf{PL}(\{\hat{P}_k\}) \coloneqq \sum_{k=1}^{K-1}\|\hat{P}_k - \hat{P}_{k+1}\|_F$ is the path length defined in (12)*.*

The convergence rate of `Lookahead OSGM-R` competes with the average contraction ratio $\frac{1}{K}\sum_{k=1}^K r_{x^k}(\hat{P}_k)$ achieved by any benchmark sequence of stepsizes $\{\hat{P}_k\}$, incurring an additional cost from the path length.

Local superlinear convergence. When f is twice continuously differentiable, f behaves like a quadratic function around $x^\star$: $f(x) = f^\star + \frac{1}{2}\langle x - x^\star, \nabla^2 f(x^\star)(x - x^\star)\rangle + \mathcal{O}(\|x - x^\star\|^3)$, for which the fixed stepsize $[\nabla^2 f(x^\star)]^{-1}$ enjoys (local) quadratic convergence. Since `Lookahead OSGM-R` can compete with any fixed stepsize, including $[\nabla^2 f(x^\star)]^{-1}$, `Lookahead OSGM-R` converges superlinearly. First, we formalize the intuition that $[\nabla^2 f(x^\star)]^{-1}$ is locally a good fixed stepsize by bounding its ratio feedback.

Lemma 6.1. *Suppose f is L-smooth μ-strongly convex and has H-Lipschitz Hessian. Then the ratio feedback of Hessian inverse at $x^\star$ is bounded by $r_x([\nabla^2 f(x^\star)]^{-1}) \le \frac{H^2\kappa}{4\mu^2}\|x - x^\star\|^2$ for all $x \notin \mathcal{X}^\star$.*

When f is a quadratic, $H = 0$ and the Hessian inverse drives any point x to the optimal solution $x^\star$ in one step and thus achieves zero ratio feedback. The superlinear convergence of `Lookahead OSGM-R` follows from **Lemma 6.1** and **Theorem 6.1**, along with the observation that $\|x^k - x^\star\| \to 0$ geometrically.

Theorem 6.4 (Superlinear convergence)**.** *Suppose f is L-smooth and μ-strongly convex and has H-Lipschitz Hessian. Then `Lookahead OSGM-R` (**Algorithm 2**) with $\eta = 1/(2L^2)$ and $P_1 = \frac{1}{L}I$ satisfies*

$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](\tfrac{C}{K})^K,$$

where $C \coloneqq \frac{H^2\kappa^2}{2\mu^3}[f(x^1) - f^\star] + L^2\|\frac{1}{L}I - [\nabla^2 f(x^\star)]^{-1}\|_F^2$.

This result shows superlinear convergence of `Lookahead OSGM-R`. The convergence appears linear (by **Theorem 6.1**) until $K \sim C$ and superlinear convergence becomes apparent when K is large. The convergence behavior of `Lookahead OSGM-R` in practice is the best of **Theorem 6.1** and **Theorem 6.4**.

Negative regret. Relation (19) suggests that the progress of `Lookahead OSGM-R` is no worse than any benchmark stepsize $\hat{P}$ up to a fixed additive constant. Actually, we can further show that progress of `Lookahead OSGM-R` is strictly better than any benchmark stepsize $\hat{P}$ unless there is a perfect stepsize that drives x to $x^\star$ in one step (e.g., $\hat{P} = [\nabla^2 f(x^\star)]^{-1}$).

Theorem 6.5 (Negative regret)**.** *Fix an arbitrary benchmark stepsize $\hat{P}$. For each $K \ge 1$, `Lookahead OSGM-R` with $\eta \in (0, \frac{1}{4L^2}]$ satisfies exactly one of the cases below:*

- *the average progress is smaller than that achieved by $\hat{P}$: $\frac{1}{K}\sum_{k=1}^K r_k \le \frac{1}{K}\sum_{k=1}^K r_{x^k}(\hat{P})$,*
- *the suboptimality satisfies a superlinear convergence bound: $f(x^{K+1}) - f(x^1) \le \frac{1}{2\mu}\|\nabla f(x^1)\|^2(\frac{\kappa^2\|P_1 - \hat{P}\|_F^2}{\eta K})^K$.*

In other words, `Lookahead OSGM-R` outperforms the linear convergence rate achievable by any fixed stepsize. The only exception is when a fixed stepsize achieves a convergence rate of 0, the case of superlinear convergence.

Knowledge of $f^\star$. One downside of `OSGM-R` is the requirement of optimal value $f^\star$. We can relax this requirement by running an outer loop to search for $f^\star$ and obtain an $\mathcal{O}(\kappa^\star \log^2(1/\varepsilon))$ complexity result. See the conference version [15] for more details.

6.2 Monotone Lookahead OSGM-H

In this section, we assume f is L-smooth, optionally μ-strongly convex and instantiate OSGM with

$$\ell_x(P) \coloneqq h_x(P), \text{ Monotone Lookahead landscape: } f(x^{k+1}) \le \min\{f(x^{k+1/2} - \tfrac{1}{L}\nabla f(x^{k+1/2})), f(x^k)\}, \ \mathcal{A} \coloneqq \texttt{OGD}.$$

The algorithm is called Monotone Lookahead OSGM-H (**Algorithm 3**). The candidate stepsize set is $\mathcal{P} = \mathbb{R}^{n\times n}$.

Algorithm 3: Monotone Lookahead OSGM-H

1 **input:** Initial point x^1, initial stepsize $P_1 \in \mathcal{P} = \mathbb{R}^{n\times n}$, online gradient stepsize $\eta_k \equiv \eta > 0$
2 **for** $k = 1, 2, \dots$ **do**
3 $\quad x^{k+1/2} = x^k - P_k \nabla f(x^k)$
4 $\quad$ Choose x^{k+1} that satisfies $f(x^{k+1}) \le \min\{f(x^{k+1/2} - \frac{1}{L}\nabla f(x^{k+1/2})), f(x^k)\}$
5 $\quad P_{k+1} = P_k - \eta \nabla h_{x^k}(P_k)$
6 **end**

Recall that the gradient of hypergradient feedback in Line **5** of **Algorithm 3** takes the form

$$\nabla h_{x^k}(P_k) = -\frac{\nabla f(x^{k+1/2})\nabla f(x^k)^\top}{\|\nabla f(x^k)\|^2}.$$

The rest of the section analyzes the convergence behavior of Monotone Lookahead OSGM-H, including global convergence and local superlinear convergence.

6.2.1 Global convergence.

Monotone Lookahead OSGM-H enjoys similar convergence guarantees to Lookahead OSGM-R.

Theorem 6.6 (Global convergence)**.** *Let f be L-smooth and (μ-strongly) convex. For any benchmark stepsize $\hat{P} \in \mathbb{R}^{n\times n}$, Monotone Lookahead OSGM-H with $\eta = 1/L$ satisfies*

$$f(x^{K+1}) - f^\star \le \min\{\frac{\Delta^2}{K\max\{\frac{1}{K}\sum_{k=1}^K -h_{x^k}(\hat{P}) - \frac{L}{2K}\|P_1 - \hat{P}\|_F^2, 0\}}, f(x^1) - f^\star\}, \qquad \text{(convex)}$$

$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](1 - 2\mu\max\{\tfrac{1}{K}\textstyle\sum_{k=1}^K -h_{x^k}(\hat{P}) - \frac{L}{2K}\|P_1 - \hat{P}\|_F^2, 0\})^K, \qquad (\mu\text{-strongly convex})$$

*where the constant Δ is defined in **Theorem 4.2**. In particular, if Monotone Lookahead OSGM-H is initialized with $P_1 = \frac{1}{L}I$, then*

$$f(x^{K+1}) - f^\star \le \min\{\frac{2L\Delta^2}{K}, f(x^1) - f^\star\}, \qquad \text{(convex)}$$

$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](1 - \tfrac{1}{\kappa})^K. \qquad (\mu\text{-strongly convex})$$

According to **Theorem 6.6**, Monotone Lookahead OSGM-H converges no slower than vanilla gradient descent. More importantly, whenever there exists a $\hat{P}$ such that the average hypergradient feedback along the trajectory satisfies $\frac{1}{K}\sum_{k=1}^K -h_{x^k}(\hat{P}) \gg \frac{1}{2L}$, Monotone Lookahead OSGM-H converges faster.

Potential reduction. Monotone Lookahead OSGM-H also admits a potential function analysis. Define

$$\omega(x, P) \coloneqq \tfrac{\rho_\omega}{f(x) - f^\star} + \|P - \tfrac{1}{L}I\|_F^2, \qquad \text{(convex)}$$

$$\varphi(x, P) \coloneqq \rho_\varphi \log(f(x) - f^\star) + \|P - \tfrac{1}{L}I\|_F^2. \qquad (\mu\text{-strongly convex})$$

Monotone Lookahead OSGM-H decreases these potentials by at least a constant at every iteration.

Theorem 6.7 (Potential reduction)**.** *Let f be L-smooth and (μ-strongly) convex. At every iteration, Monotone Lookahead OSGM-H with $\eta = 1/L$ decreases the potential $\varphi(x, P)$ with $\rho_\varphi = 1/(L\mu)$ and potential $\omega(x, P)$ with*

$\rho_\omega = 2\Delta^2/L$ *by*

$$\omega(x^{k+1}, P_{k+1}) - \omega(x^k, P_k) \le -\tfrac{1}{L^2}, \qquad \text{(convex)}$$
$$\varphi(x^{k+1}, P_{k+1}) - \varphi(x^k, P_k) \le -\tfrac{1}{L^2}. \qquad (\mu\text{-strongly convex})$$

In particular, ***Monotone Lookahead OSGM-H*** *with* $P_1 = \frac{1}{L} I$ *finds an* ε*-optimal solution within complexity*

$$K_\varepsilon \coloneqq \left\lceil \tfrac{2L\Delta^2}{\varepsilon} \right\rceil \quad \textit{for convex } f; \quad \textit{and} \quad K_\varepsilon \coloneqq \left\lceil \kappa \log(\tfrac{1}{\varepsilon}) \right\rceil \quad \textit{for } \mu\textit{-strongly convex } f.$$

6.2.2 Local convergence

The local convergence guarantees of `Monotone Lookahead OSGM-H` are similar to `Lookahead OSGM-R`.

Theorem 6.8 (Local adaptivity)**.** *Let* f *be* L*-smooth and* (μ*-strongly*) *convex. For any benchmark sequence of stepsizes* $\{\hat{P}_k\}$*,* ***Monotone Lookahead OSGM-H*** *with stepsize* $\eta = 1/L$ *satisfies*

$$f(x^{K+1}) - f^\star \le \min\{\tfrac{\Delta^2}{K \max\{\frac{1}{K}\sum_{k=1}^K -h_{x^k}(\hat{P}_k) - \rho_K(\{\hat{P}_k\}), 0\}}, f(x^1) - f^\star\}, \qquad \text{(convex)}$$
$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](1 - 2\mu \max\{\tfrac{1}{K}\textstyle\sum_{k=1}^K -h_{x^k}(\hat{P}_k) - \rho_K(\{\hat{P}_k\}), 0\})^K. \qquad (\mu\text{-strongly convex})$$

where $\rho_K(\{\hat{P}_k\}) \coloneqq \frac{L}{2K}[\|\hat{P}_K - P_1\|_F^2 + 2\max_{k\le K}\{\|P_k - P_1\|_F\}\mathsf{PL}(\{\hat{P}_k\})]$ *and* $\mathsf{PL}(\{\hat{P}_k\})$ *is defined in* (12)*.*

Theorem 6.9 (Superlinear convergence)**.** *Suppose* f *is* L*-smooth,* μ*-strongly convex and has* H*-Lipschitz Hessian. Then* ***Monotone Lookahead OSGM-H*** *with stepsize* $\eta = 1/L$ *and* $P_1 = \frac{1}{L} I$ *satisfies*

$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](\tfrac{C}{K})^K,$$

where $C \coloneqq \frac{H^2\kappa^3}{2\mu^3}[f(x^1) - f^\star] + L^2\|\frac{1}{L} I - [\nabla^2 f(x^\star)]^{-1}\|_F^2$*.*

Theorem 6.10 (Negative regret)**.** *Fix an arbitrary benchmark stepsize* $\hat{P}$*. For each* $K \ge 1$*,* ***Monotone Lookahead OSGM-H*** *with* $\eta \in (0, \frac{1}{2L}]$ *satisfies exactly one of the cases below:*

- *the average progress is smaller than achieved by* $\hat{P}$*:* $\frac{1}{K}\sum_{k=1}^K h_k \le \frac{1}{K}\sum_{k=1}^K h_{x^k}(\hat{P})$*,*
- *the suboptimality satisfies a superlinear convergence bound:* $f(x^{K+1}) - f(x^1) \le \frac{1}{2\mu}\|\nabla f(x^1)\|^2(\frac{2L\|P_1 - \hat{P}\|_F^2}{\eta K})^K$*.*

6.3 Other instances of `OSGM`

Variants of `OSGM` can be obtained by enumerating combinations of feedback and landscape actions, but noting that hypergradient feedback should be used with monotone (lookahead) action due to the reduction for h_k. `Vanilla OSGM-R` introduced in **Section 1** is the only variant that requires no environment action and is still guaranteed to converge on smooth strongly convex problems. `Monotone OSGM-H` often demonstrates the strongest empirical performance among all variants. The analyses for other variants resemble `Lookahead OSGM-R` and `Monotone Lookahead OSGM-H` presented in the previous two subsections, except for the choice of η_k in online gradient descent. We defer the highly repetitive details to Appendix **E** and summarize the convergence rates of each variant in **Table 1**. Notably, `OSGM` is the second family of first-order methods with nonasymptotic superlinear convergence guarantees on smooth convex optimization problems, following the renowned quasi-Newton methods. The online learning argument offers a simple superlinear convergence analysis of `OSGM`.

7 Practical algorithm design with `OSGM`

This section discusses the implementation of `OSGM` for practical performance.

Table 1: Global convergence rates of `OSGM`

Feedback	Convexity	Landscape action $\mathcal{M}$	Worst-case global convergence	Superlinear convergence
r_x	Strongly convex	Lookahead Monotone Lookahead	$\min\{(1-\frac{1}{\kappa^\star}+\mathcal{O}(\frac{1}{K}))^K, (1-\frac{1}{\kappa})^K\}$	$\mathcal{O}(e^{-K\log K})$
		Vanilla Monotone	$(1-\frac{1}{\kappa^\star}+\mathcal{O}(\frac{1}{\sqrt{K}}))^K$	$\mathcal{O}(e^{-\frac{1}{2}K\log K})$
h_x	Strongly convex	Monotone Lookahead	$(1-\frac{1}{\kappa})^K$	$\mathcal{O}(e^{-K\log K})$
		Monotone	$(1-\frac{1}{\kappa}+\mathcal{O}(\frac{1}{\sqrt{K}}))^K$	$\mathcal{O}(e^{-\frac{1}{2}K\log K})$
	Convex	Monotone Lookahead	$\frac{2L\Delta^2}{K}$	—
		Monotone	$\frac{2L\Delta^2}{K}(\frac{1}{1-\mathcal{O}(\frac{1}{\sqrt{K}})})$	

7.1 Choice of candidate stepsize

The candidate set $\mathcal{P}$ affects the convergence of `OSGM` through its expressive power and regret and determines the computational efficiency of `OSGM`.

Expressiveness of $\mathcal{P}$ and superlinear convergence. Richer candidate sets $\mathcal{P}$ leads to smaller minimum average feedback $\min_{\hat{P}\in\mathcal{P}} \frac{1}{K}\sum_{k=1}^{K} \ell_{x^k}(\hat{P})$, indicating faster convergence rate `OSGM` can achieve asymptotically. In particular, inverse Hessian at optimal solution $[\nabla^2 f(x^\star)]^{-1}$ guarantees local superlinear convergence. If $[\nabla^2 f(x^\star)]^{-1} \in \mathcal{P}$, then it is a legitimate benchmark $\hat{P}$ for `OSGM` to compete against, indicating local superlinear convergence of `OSGM` too.

Online regret. Richer candidate sets however may lead to larger dimension-dependent constant $\|P_1 - \hat{P}\|_F^2$ and online gradient norm $\|\nabla \ell_x(P)\|_F^2$ in the regret and increase the kick-in period. A scalar or diagonal stepsize often outperforms a full matrix if we expect to perform only a few iterations K.

Three common choices for $\mathcal{P}$ in practice are listed in **Table 2**.

Table 2: Examples of stepsize patterns and corresponding online gradients.

Pattern	Candidate set $\mathcal{P}$	Online gradient	
		Ratio r_x	Hypergradient h_x
Scalar	$\{\alpha I : \alpha \in \mathbb{R}\}$	$r_x'(\alpha) = -\frac{\langle \nabla f(x-\alpha\nabla f(x)), \nabla f(x)\rangle}{f(x)-f^\star}$	$h_x'(\alpha) = -\frac{\langle \nabla f(x-\alpha\nabla f(x)), \nabla f(x)\rangle}{\|\nabla f(x)\|^2}$
Diagonal	$\{\mathrm{Diag}(d) : d \in \mathbb{R}^n\}$	$\nabla r_x(d) = -\frac{\nabla f(x-d\circ\nabla f(x))\circ\nabla f(x)}{f(x)-f^\star}$	$\nabla h_x(d) = -\frac{\nabla f(x-d\circ\nabla f(x))\circ\nabla f(x)}{\|\nabla f(x)\|^2}$
Full matrix	$\mathbb{R}^{n\times n}$	$\nabla r_x(P) = -\frac{\nabla f(x-P\nabla f(x))\nabla f(x)^\top}{f(x)-f^\star}$	$\nabla h_x(P) = -\frac{\nabla f(x-P\nabla f(x))\nabla f(x)^\top}{\|\nabla f(x)\|^2}$

Additional constraints, such as positive-semidefiniteness and boundedness, can be enforced in `OGD` update $P_{k+1} = \Pi_{\mathcal{P}}[P_k - \eta\nabla\ell_{x^k}(P_k)]$ by an explicit projection but at different computational costs. For example, entrywise projection onto an interval simply truncates the entries to values in the interval, while projection on the semidefinite cone requires a full eigendecomposition. Our theory for `OSGM` applies even when `OGD` involves the projector $\Pi_{\mathcal{P}}[\cdot]$ due to the non-expansiveness of the projection operator:

$$\|P_{k+1} - \hat{P}\|_F^2 = \|\Pi_{\mathcal{P}}[P_k - \eta\nabla\ell_{x^k}(P_k)] - \hat{P}\|_F^2 \le \|P_k - \eta\nabla\ell_{x^k}(P_k) - \hat{P}\|_F^2, \quad \text{for all} \quad \hat{P}\in\mathcal{P}.$$

Memory and iteration cost. Storing and updating a full matrix costs $\mathcal{O}(n^2)$ memory and time, whereas a diagonal or scalar parameterization is essentially free, with memory and computational costs that match the cost of a gradient step. Projecting onto the set of positive semidefinite matrices $\mathcal{P} = \mathbb{S}_+^n$ requires a full eigendecomposition, at a cost of $\mathcal{O}(n^3)$, which is very rarely worthwhile. Indeed, one of the advantages of `OSGM` over traditional preconditioning is that we directly parametrize the scaling of the gradient, and so need not ensure the learned scaling is positive semidefinite.

Recommended choice. If a good guess of $[\nabla^2 f(x^\star)]^{-1}$ is available, `OSGM` locates $[\nabla^2 f(x^\star)]^{-1}$ efficiently in the first few iterations, and a full matrix stepsize is preferred. Otherwise, a diagonal stepsize is generally the most practical option. We observe that projection onto positive scalings is sometimes helpful for diagonal stepsizes.

7.2 Choice of feedback

Hypergradient feedback h_x generally has stronger empirical performance than ratio feedback r_x, even though the latter has desirable theoretical properties. There are two possible reasons:

- Hypergradient feedback does not require the knowledge of optimal value $f^\star$.
- Hypergradient feedback has a better dependence on the smoothness constant: h_x is L-smooth but r_x is $2L^2$-smooth. Consequently, r_x may incur larger regret and its theoretical advantage can only be observed for very large K.

An exception where ratio feedback is preferred occurs when 1) $\kappa^\star \leq 100$ and 2) a good initial stepsize P_1 is available. In this case, `OSGM-R` sometimes outperforms `OSGM-H`.

7.3 Choice of landscape action

Vanilla landscape action is less stable and not recommended. In most cases, a monotone landscape action suffices to yield satisfying performance. A null step $x^{k+1} = \arg\min_{x\in\{x^{k+1/2},x^k\}} f(x)$ is particularly recommended since the number of gradient oracle calls at each iteration equals that of vanilla gradient descent.

7.4 Choice of online learning algorithm

Other advanced online learning algorithms can be used in place of `OGD`, and our convergence analysis still applies. Advanced online learning algorithms often outperform `OGD`, and we recommend using `AdaGrad` [13] and other advanced online learning algorithms from [52, 25] in `OSGM`. Most convergence guarantees in this paper hold for `AdaGrad` variant of `OSGM`, though possibly in a weaker statement.

Online learning parameters. Online learning algorithms often come with their own hyperparameters, such as the stepsize η in `OGD`. Our analysis assumes the knowledge of L or at least an upper bound. One option to relax this assumption is to use parameter-free online algorithms such as [52], which typically require $\operatorname{diam}(\mathcal{P}) < \infty$. Another option is to apply back-tracking line-search to estimate L, relying on the fact that both f and feedback $r_x(P)/h_x(P)$ are smooth. For example, given an estimate L' of L, we can monitor two conditions below:

$$h_{x^k}(P_k - \tfrac{1}{L'}\nabla h_{x^k}(P_k)) - h_{x^k}(P_k) \leq -\tfrac{1}{2L'}\|\nabla h_{x^k}(P_k)\|_F^2,$$
$$f(x^{k+1/2} - \tfrac{1}{L'}\nabla f(x^{k+1/2})) - f(x^{k+1/2}) \leq -\tfrac{1}{2L'}\|\nabla f(x^{k+1/2})\|^2.$$

Whenever either condition fails, one can backtrack L' by a certain fraction to find new estimates.

7.5 Momentum-based optimization step

The most practical variant of `OSGM` adds the heavy-ball momentum to the step of gradient descent:

$$x^{k+1} = x^k - P_k\nabla f(x^k) + \beta_k(x^k - x^{k-1})$$

and jointly learns the stepsize P_k and momentum parameter β_k by online learning. The hypergradient feedback is modified based on the potential function for heavy-ball momentum [10]:

$$\varphi(x, x^-) = f(x) - f^\star + \tfrac{\omega}{2}\|x - x^-\|^2.$$

Heavy-ball variant of `OSGM-H` demonstrates promising performance in **Section 8** and [9].

8 Experiments

We have developed an efficient variant of `OSGM-H` for large-scale machine learning tasks. See Part II of this paper for more details.

9 Related work

`OSGM` is closely related to several lines of research, which we detail below.

Preconditioned gradient descent. The update $x^{k+1} = x^k - P_k \nabla f(x^k)$ that scales the gradient by P_K is known as preconditioned gradient descent in literature. Preconditioning has become a standard tool in both optimization and numerical linear algebra and is widely used in modern optimization algorithms [38, 66, 39, 40, 17, 37, 11, 50, 5, 23]. Some recent research has tried to empirically quantify the effect of preconditioning on linear systems [55, 16]. However, identifying a good preconditioner with theoretical guarantees is often challenging in practice. `OSGM` allows gradient-based methods to converge as if it is preconditioned by the best possible preconditioner, with the cost amortized along the iterations.

Adaptive first-order methods and stepsize scheduling. `OSGM` is closely related to the literature on stepsize scheduling and the widely used adaptive gradient methods. Examples in this family include stepsize schedulers such as [36, 65] and adaptive methods such as `AdaGrad` [13], `Adam` [68, 47], `RMSProp` [22] among many other popular variants [18, 57, 69, 1]. Many of the adaptive gradient methods originate from online learning and have provable regret guarantees, and they also relate to parameter-free online learning algorithms (see [51] for more details). Aside from stepsize selection strategies motivated by online learning, there are well-known stepsize selection strategies such as the Barzilai-Borwein (BB) step [6], Polyak stepsize [54, 20] and notable recently developed adaptive stepsizes [42, 43]. Another relevant line of research focuses on uniformly optimal first-order methods [35, 12], which choose stepsize to adapt to problem parameters, such as the smoothness constant. The idea of analyzing the multi-step convergence relates to recent work on stepsize hedging [3, 4].

Quasi-Newton methods. One notable feature of `OSGM` is its non-asymptotic superlinear convergence. See **Table 3** for a comparison with recent results showcasing superlinear convergence rates of quasi-Newton methods.

Table 3: Superlinear convergence rates of various methods

Algorithm	Reference	Superlinear convergence rate
Greedy quasi-Newton	[58]	$\mathcal{O}(e^{-\frac{1}{2}K^2})$
Broyden family	[60, 59]	$\mathcal{O}(e^{-\frac{1}{2}K\log K})$
Online-learning guided quasi-Newton	[26]	$\mathcal{O}(e^{-\frac{1}{2}K\log K})$
BFGS with line-search	[30, 29]	$\mathcal{O}(e^{-K\log K})$
Monotone Lookahead `OSGM`	This paper	$\mathcal{O}(e^{-K\log K})$
Vanilla/Monotone `OSGM`	This paper	$\mathcal{O}(e^{-\frac{1}{2}K\log K})$

Our results identify a similarity between `OSGM` and quasi-Newton methods. Both `OSGM` and `HDM` learn the inverse Hessian operator $g \mapsto [\nabla^2 f(x^\star)]^{-1} g$ as the algorithm progresses, but through different properties of the operator. The quasi-Newton methods use the secant equation $x - y \approx [\nabla^2 f(x^\star)]^{-1}(\nabla f(x) - \nabla f(y))$ for x, y close to $x^\star$ and enforce this equation, replacing the inverse Hessian by P_k, to guide learning [26, 27]. In contrast, `OSGM` learns an optimal stepsize for the function. Since the function is locally quadratic, this optimal stepsize is the inverse Hessian. `OSGM` uses the ratio/hypergradient feedback to measure the quality of the stepsize directly and can search for an optimal stepsize in a given closed convex set $\mathcal{P}$. Both approaches require a safeguard to prevent divergence in the warm-up phase, which is achieved by line-search in quasi-Newton and landscape actions in `OSGM`. In a word, both `OSGM` and quasi-Newton leverage complementary perspectives on $g \mapsto [\nabla^2 f(x^\star)]^{-1} g$, so it is natural that they achieve similar convergence guarantees.

Hypergradient descent heuristic. OSGM is closely relevant to the hypergradient descent heuristic. Hypergradient descent method (HDM) dates back to 1999 [2], which was first proposed as a heuristic to accelerate stochastic gradient descent. Similar updates were also explored in [63, 62, 24, 41], while those works employed slightly different algorithmic updates. Later, [7] rediscovered the HDM and named it "hypergradient descent"; [7] also extended HDM to other first-order methods with extensive experimental validation of its practical efficacy. Recent studies [28, 8, 53] further empirically enhanced HDM for broader applicability, reporting promising numerical results. Despite these empirical successes, a rigorous theoretical understanding of HDM has emerged only recently. [61] showed that a variant of HDM converges on convex quadratic functions and established several analytic properties. [67] showed the convergence of hypergradient descent with a "powerball" technique and variance-reduction on stochastic finite-sum optimization. However, these results do not explain the empirical performance of HDM. Subsequently, [33] demonstrated that when using a diagonal preconditioner, hypergradient can be employed to generate cutting planes in the preconditioner space, achieving an $\mathcal{O}(\sqrt{n}\kappa^\star \log(1/\varepsilon))$ complexity result on smooth strongly convex functions. Prior to [33], a similar idea was exploited in [45], where the authors used the ellipsoid method to update the preconditioner in the conjugate gradient method for solving ill-conditioned linear systems. The preliminary versions of this work [15, 9] showed that HDM can be viewed as online gradient descent applied to some surrogate loss function and that HDM has strong trajectory-based convergence guarantees.

Connection between OSGM and HDM in literature. The most typical version of hypergradient descent [7, 61] swaps the order of primal update and stepsize update in OSGM-H:

HDM	Lookahead OSGM-H
$P_{k+1} = P_k - \eta \nabla h_{x^k}(P_k)$	$x^{k+1/2} = x^k - P_k \nabla f(x^k)$
$x^{k+1}_{\texttt{HDM}} = x^k - P_{k+1}\nabla f(x^k)$	$x^{k+1}_{\texttt{OSGM}} = x^{k+1/2} - \frac{1}{L}\nabla f(x^{k+1/2})$
	$P_{k+1} = P_k - \eta \nabla h_{x^k}(P_k)$

HDM first updates the matrix stepsize P_k using the feedback $h_{x^k}(P)$ and then makes a gradient step on x^k using P_{k+1} to arrive at the next iterate x^{k+1}. The per-iteration progress of HDM is therefore

$$h_k = \frac{f(x^{k+1}_{\texttt{HDM}}) - f(x^k)}{\|\nabla f(x^k)\|^2} = \frac{f(x^k - P_{k+1}\nabla f(x^k)) - f(x^k)}{\|\nabla f(x^k)\|^2} = h_{x^k}(P_{k+1}),$$

which is the hypergradient feedback evaluated at P_{k+1} whose update already uses the feedback function $h_{x^k}(P)$. This setting of HDM can be modeled as *prescient* online learning [64], in which $h_{x^k}(P)$ is revealed to the scheduler before making a decision. The received feedback $h_{x^k}(P_{k+1})$ is evaluated at the newly chosen stepsize. The knowledge of the future allows the stepsize scheduler in HDM to achieve a constant regret [64, 51, 56], which is similar to the guarantee achieved by Lookahead OSGM. This is not a coincidence: when $\mathcal{P} = \mathbb{R}^{n\times n}$, HDM reduces to Lookahead OSGM-H except the stepsize of additional gradient step is not $\frac{1}{L}$ but η:

$$\begin{aligned}
x^{k+1}_{\texttt{HDM}} &= x^k - P_{k+1}\nabla f(x^k) \\
&= x^k - (P_k - \eta \nabla h_{x^k}(P_k))\nabla f(x^k) && \text{(By update of } P_{k+1}) \\
&= x^k - (P_k + \eta \tfrac{\nabla f(x^k - P_k \nabla f(x^k))\nabla f(x^k)^\top}{\|\nabla f(x^k)\|^2})\nabla f(x^k) && \text{(By definition of } \nabla h_x(P)) \\
&= x^k - P_k \nabla f(x^k) - \eta \nabla f(x^k - P_k \nabla f(x^k)) \\
&= x^{k+1/2} - \eta \nabla f(x^{k+1/2}). && (x^{k+1/2} = x^k - P_k \nabla f(x^k))
\end{aligned}$$

In other words, HDM uses the same stepsize η for both additional gradient step and online gradient descent, while Lookahead OSGM-H decouples the two stepsizes. HDM and Lookahead OSGM are equivalent if the stepsize of online gradient descent is set to $\eta = \frac{1}{L}$ and $\mathcal{P} = \mathbb{R}^{n\times n}$ (no projection in online gradient descent). When $\eta < 1/L$, the lookahead landscape action makes more progress than the swapped update $x^{k+1}_{\texttt{HDM}} = x^k - P_{k+1}\nabla f(x^k)$.

Reconcile with lower bound. `OSGM` provides an problem-dependent $\mathcal{O}(\kappa^\star \log(1/\varepsilon))$ acceleration result that can outperform the accelerated rate $\mathcal{O}(\sqrt{\kappa}\log(1/\varepsilon))$ when $\kappa^\star < \sqrt{\kappa}$ and $\varepsilon \to 0$. However, our result does not conflict with the known lower bound for smooth strongly convex optimization [48]. Instead, the acceleration effect should be considered as the effect of implicit preconditioning.

10 Conclusions

Online scaled gradient methods (`OSGM`) offer a new mechanism for accelerating gradient-based methods and constitute a new family of first-order methods with superlinear convergence. `OSGM` has global, local, and trajectory-dependent convergence guarantees, laying the theoretical foundation for hypergradient-descent-type first-order methods. We hope `OSGM` paves the way for new directions in the design of adaptive optimization methods.

References

[1] Naman Agarwal, Rohan Anil, Elad Hazan, Tomer Koren, and Cyril Zhang. Disentangling adaptive gradient methods from learning rates. *CoRR*, abs/2002.11803, 2020. (cited on 17)

[2] Luís B Almeida, Thibault Langlois, José D Amaral, and Alexander Plakhov. Parameter adaptation in stochastic optimization. In *On-line learning in neural networks*, pages 111–134. 1999. (cited on 18)

[3] Jason M Altschuler and Pablo A Parrilo. Acceleration by stepsize hedging: Silver stepsize schedule for smooth convex optimization. *Mathematical Programming*, pages 1–14, 2024. (cited on 17)

[4] Jason M Altschuler and Pablo A Parrilo. Acceleration by stepsize hedging: Multi-step descent and the silver stepsize schedule. *Journal of the ACM*, 72(2):1–38, 2025. (cited on 17)

[5] David Applegate, Mateo Diaz, Oliver Hinder, Haihao Lu, Miles Lubin, Brendan O'Donoghue, and Warren Schudy. Practical large-scale linear programming using primal-dual hybrid gradient. *Advances in Neural Information Processing Systems*, 34:20243–20257, 2021. (cited on 17)

[6] Jonathan Barzilai and Jonathan M Borwein. Two-point step size gradient methods. *IMA journal of numerical analysis*, 8(1):141–148, 1988. (cited on 17)

[7] Atilim Gunes Baydin, Robert Cornish, David Martinez Rubio, Mark Schmidt, and Frank Wood. Online learning rate adaptation with hypergradient descent. In *International Conference on Learning Representations*, 2018. (cited on 2, 18)

[8] Kartik Chandra, Audrey Xie, Jonathan Ragan-Kelley, and Erik Meijer. Gradient descent: The ultimate optimizer. *Advances in Neural Information Processing Systems*, 35:8214–8225, 2022. (cited on 18)

[9] 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. (cited on 1, 3, 4, 16, 18)

[10] Marina Danilova, Anastasiia Kulakova, and Boris Polyak. Non-monotone behavior of the heavy ball method. In *Difference Equations and Discrete Dynamical Systems with Applications: 24th ICDEA, Dresden, Germany, May 21–25, 2018 24*, pages 213–230. Springer, 2020. (cited on 16)

[11] Qi Deng, Qing Feng, Wenzhi Gao, Dongdong Ge, Bo Jiang, Yuntian Jiang, Jingsong Liu, Tianhao Liu, Chenyu Xue, Yinyu Ye, et al. An enhanced alternating direction method of multipliers-based interior point method for linear and conic optimization. *INFORMS Journal on Computing*, 2024. (cited on 17)

[12] Qi Deng, Guanghui Lan, and Zhenwei Lin. Uniformly optimal and parameter-free first-order methods for convex and function-constrained optimization. *arXiv preprint arXiv:2412.06319*, 2024. (cited on 17)

[13] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011. (cited on 3, 16, 17)

[14] Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2000. (cited on 2)

[15] Wenzhi Gao, Ya-Chi Chu, Yinyu Ye, and Madeleine Udell. Gradient methods with online scaling. *arXiv preprint arXiv:2411.01803*, 2024. (cited on 1, 4, 6, 12, 18)

[16] Wenzhi Gao, Zhaonan Qu, Madeleine Udell, and Yinyu Ye. Scalable approximate optimal diagonal preconditioning. *arXiv preprint arXiv:2312.15594*, 2023. (cited on 6, 17)

[17] Paul J Goulart and Yuwen Chen. Clarabel: An interior-point solver for conic programs with quadratic objectives. *arXiv preprint arXiv:2405.12762*, 2024. (cited on 17)

[18] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018. (cited on 17)

[19] Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016. (cited on 8)

[20] Elad Hazan and Sham Kakade. Revisiting the polyak step size. *arXiv preprint arXiv:1905.00313*, 2019. (cited on 17)

[21] Elad Hazan, Alexander Rakhlin, and Peter Bartlett. Adaptive online gradient descent. *Advances in neural information processing systems*, 20, 2007. (cited on 3)

[22] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012. (cited on 3, 17)

[23] Yicheng Huang, Wanyu Zhang, Hongpei Li, Dongdong Ge, Huikang Liu, and Yinyu Ye. Restarted primal-dual hybrid conjugate gradient method for large-scale quadratic programming. *arXiv preprint arXiv:2405.16160*, 2024. (cited on 17)

[24] Robert A Jacobs. Increased rates of convergence through learning rate adaptation. *Neural networks*, 1(4):295–307, 1988. (cited on 18)

[25] Andrew Jacobsen and Ashok Cutkosky. Unconstrained online learning with unbounded losses. In *International Conference on Machine Learning*, pages 14590–14630. PMLR, 2023. (cited on 16)

[26] Ruichen Jiang, Qiujiang Jin, and Aryan Mokhtari. Online learning guided curvature approximation: A quasi-newton method with global non-asymptotic superlinear convergence. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1962–1992. PMLR, 2023. (cited on 2, 17)

[27] Ruichen Jiang and Aryan Mokhtari. Online learning guided quasi-newton methods with global non-asymptotic convergence. *arXiv preprint arXiv:2410.02626*, 2024. (cited on 2, 17)

[28] Renlong Jie, Junbin Gao, Andrey Vasnev, and Minh-Ngoc Tran. Adaptive hierarchical hyper-gradient descent. *International Journal of Machine Learning and Cybernetics*, 13(12):3785–3805, 2022. (cited on 18)

[29] Qiujiang Jin, Ruichen Jiang, and Aryan Mokhtari. Non-asymptotic global convergence analysis of bfgs with the armijo-wolfe line search. *arXiv preprint arXiv:2404.16731*, 2024. (cited on 17)

[30] Qiujiang Jin, Ruichen Jiang, and Aryan Mokhtari. Non-asymptotic global convergence rates of bfgs with exact line search. *arXiv preprint arXiv:2404.01267*, 2024. (cited on 2, 17)

[31] Qiujiang Jin and Aryan Mokhtari. Non-asymptotic superlinear convergence of standard quasi-newton methods. *Mathematical Programming*, 200(1):425–473, 2023. (cited on 2)

[32] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. (cited on 3)

[33] Frederik Kunstner, Victor Sanches Portella, Mark Schmidt, and Nicholas Harvey. Searching for optimal per-coordinate step-sizes with multidimensional backtracking. *Advances in Neural Information Processing Systems*, 36, 2024. (cited on 2, 6, 18)

[34] Guanghui Lan, Yuyuan Ouyang, and Zhe Zhang. Optimal and parameter-free gradient minimization methods for smooth optimization. *arXiv preprint arXiv:2310.12139*, 2023. (cited on 3)

[35] Tianjiao Li and Guanghui Lan. A simple uniformly optimal method without line search for convex optimization. *arXiv preprint arXiv:2310.10082*, 2023. (cited on 3, 17)

[36] Xiaoyu Li, Zhenxun Zhuang, and Francesco Orabona. A second look at exponential and cosine step sizes: Simplicity, adaptivity, and performance. In *International Conference on Machine Learning*, pages 6553–6564. PMLR, 2021. (cited on 17)

[37] Tianyi Lin, Shiqian Ma, Yinyu Ye, and Shuzhong Zhang. An admm-based interior-point method for large-scale linear programming. *Optimization Methods and Software*, 36(2-3):389–424, 2021. (cited on 17)

[38] Zhenwei Lin, Zikai Xiong, Dongdong Ge, and Yinyu Ye. Pdcs: A primal-dual large-scale conic programming solver with gpu enhancements. *arXiv preprint arXiv:2505.00311*, 2025. (cited on 17)

[39] Haihao Lu and Jinwen Yang. cupdlp. jl: A gpu implementation of restarted primal-dual hybrid gradient for linear programming in julia. *arXiv preprint arXiv:2311.12180*, 2023. (cited on 17)

[40] Haihao Lu, Jinwen Yang, Haodong Hu, Qi Huangfu, Jinsong Liu, Tianhao Liu, Yinyu Ye, Chuwen Zhang, and Dongdong Ge. cupdlp-c: A strengthened implementation of cupdlp for linear programming by c language. *arXiv preprint arXiv:2312.14832*, 2023. (cited on 17)

[41] Ashique Rupam Mahmood, Richard S Sutton, Thomas Degris, and Patrick M Pilarski. Tuning-free step-size adaptation. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2121–2124. IEEE, 2012. (cited on 18)

[42] Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. In *International Conference on Machine Learning*, pages 6702–6712. PMLR, 2020. (cited on 17)

[43] Yura Malitsky and Konstantin Mishchenko. Adaptive proximal gradient method for convex optimization. *Advances in Neural Information Processing Systems*, 37:100670–100697, 2024. (cited on 17)

[44] H Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. *arXiv preprint arXiv:1002.4908*, 2010. (cited on 3)

[45] Renato D. C. Monteiro, Jerome W. O'Neal, and Arkadi Nemirovski. A new conjugate gradient algorithm incorporating adaptive ellipsoid preconditioning. Technical Report Optimization Online e-print 2004-10-973, School of Industrial and Systems Engineering, Georgia Institute of Technology, October 2004. Submitted October 6, 2004. (cited on 18)

[46] Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175:69–107, 2019. (cited on 2)

[47] Yurii Nesterov. A method for solving the convex programming problem with convergence rate o (1/k2). In *Dokl akad nauk Sssr*, volume 269, page 543, 1983. (cited on 2, 17)

[48] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013. (cited on 2, 19)

[49] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999. (cited on 2)

[50] Brendan O'donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169:1042–1068, 2016. (cited on 17)

[51] Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019. (cited on 2, 3, 8, 17, 18)

[52] Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 29, 2016. (cited on 3, 16)

[53] Kaan Ozkara, Can Karakus, Parameswaran Raman, Mingyi Hong, Shoham Sabach, Branislav Kveton, and Volkan Cevher. MADA: Meta-adaptive optimizers through hyper-gradient descent. In *Forty-first International Conference on Machine Learning*, 2024. (cited on 18)

[54] Boris T Polyak. Introduction to optimization. 1987. (cited on 17)

[55] Zhaonan Qu, Wenzhi Gao, Oliver Hinder, Yinyu Ye, and Zhengyuan Zhou. Optimal diagonal preconditioning. *Operations Research*, 2024. (cited on 6, 17)

[56] Alexander Rakhlin, J Abernethy, A Agarwal, P Bartlett, E Hazan, and A Tewari. Lecture notes on online learning draft, 2009. (cited on 18)

[57] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019. (cited on 17)

[58] Anton Rodomanov and Yurii Nesterov. Greedy quasi-newton methods with explicit superlinear convergence. *SIAM Journal on Optimization*, 31(1):785–811, 2021. (cited on 2, 17)

[59] Anton Rodomanov and Yurii Nesterov. New results on superlinear convergence of classical quasi-newton methods. *Journal of optimization theory and applications*, 188:744–769, 2021. (cited on 2, 17)

[60] Anton Rodomanov and Yurii Nesterov. Rates of superlinear convergence for classical quasi-newton methods. *Mathematical Programming*, pages 1–32, 2022. (cited on 2, 17)

[61] David Martinez Rubio. Convergence analysis of an adaptive method of gradient descent. *University of Oxford, Oxford, M. Sc. thesis*, 2017. (cited on 18)

[62] Nicol N Schraudolph. Local gain adaptation in stochastic gradient descent. 1999. (cited on 2, 18)

[63] Richard S Sutton. Adapting bias by gradient descent: An incremental version of delta-bar-delta. In *AAAI*, volume 92, pages 171–176. Citeseer, 1992. (cited on 18)

[64] Jun-Kun Wang, Jacob Abernethy, and Kfir Y Levy. No-regret dynamics in the fenchel game: A unified framework for algorithmic convex optimization. *Mathematical Programming*, 205(1):203–268, 2024. (cited on 18)

[65] Xiaoyu Wang and Ya-xiang Yuan. On the convergence of stochastic gradient descent with bandwidth-based step size. *Journal of Machine Learning Research*, 24(48):1–49, 2023. (cited on 17)

[66] Zikai Xiong and Robert M Freund. The role of level-set geometry on the performance of pdhg for conic linear optimization. *arXiv preprint arXiv:2406.01942*, 2024. (cited on 17)

[67] Zhuang Yang. Adaptive powerball stochastic conjugate gradient for large-scale learning. *IEEE Transactions on Big Data*, 9(6):1598–1606, 2023. (cited on 18)

[68] Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. *arXiv preprint arXiv:2406.16793*, 2024. (cited on 17)

[69] Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar C Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James Duncan. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. *Advances in neural information processing systems*, 33:18795–18806, 2020. (cited on 17)

[70] Zhenxun Zhuang, Ashok Cutkosky, and Francesco Orabona. Surrogate losses for online learning of stepsizes in stochastic non-convex optimization. In *International Conference on Machine Learning*, pages 7664–7672. PMLR, 2019. (cited on 3)

Appendix

Table of Contents

A Proof of results in Section 3

A.1 Proof of Proposition 3.1

By the chain rule, the gradient of $u_x(P)$ with respect to P takes the form

$$\nabla u_x(P) = -\nabla f(x - P\nabla f(x))\nabla f(x)^\top. \tag{22}$$

Since $u_x(P) = f(x - P\nabla f(x))$ is the composition between affine function $x - P\nabla f(x)$ and convex function f, it is convex by the composition rule. Next, we show the smoothness of $u_x(P)$. For any P_1, P_2, we deduce that

$$\begin{aligned}
\|\nabla u_x(P_1) - \nabla u_x(P_2)\|_F &= \|[\nabla f(x - P_1\nabla f(x)) - \nabla f(x - P_2\nabla f(x))]\nabla f(x)^\top\|_F && \text{(by definition (22))} \\
&= \|\nabla f(x - P_1\nabla f(x)) - \nabla f(x - P_2\nabla f(x))\|\|\nabla f(x)\| && \text{(by } \|ab^\top\|_F = \|a\|\|b\|\text{)} \\
&\le L\|(P_1 - P_2)\nabla f(x)\|\|\nabla f(x)\| && \text{(by } L\text{-smoothness)} \\
&\le L\|\nabla f(x)\|^2\|P_1 - P_2\| && \text{(by submultiplicativity of } \|\cdot\|\text{)} \\
&\le L\|\nabla f(x)\|^2\|P_1 - P_2\|_F. && \text{(by } \|\cdot\| \le \|\cdot\|_F\text{)}
\end{aligned}$$

Now, suppose $\operatorname{diam}(\mathcal{P}) \le D$. We show the Lipschitz continuity of $u_x(P)$. For any $P \in \mathcal{P}$, we deduce that

$$\begin{aligned}
\|\nabla u_x(P)\|_F &= \|\nabla f(x - P\nabla f(x))\nabla f(x)^\top\|_F && \text{(by definition (22))} \\
&= \|\nabla f(x - P\nabla f(x))\|\|\nabla f(x)\| && \text{(by } \|ab^\top\|_F = \|a\|\|b\|\text{)} \\
&\le (\|\nabla f(x - P\nabla f(x)) - \nabla f(x)\| + \|\nabla f(x)\|)\|\nabla f(x)\| && \text{(by } \|a\| \le \|a-b\| + \|b\|\text{)} \\
&\le (L\|P\nabla f(x)\| + \|\nabla f(x)\|)\|\nabla f(x)\| && \text{(by } L\text{-smoothness)} \\
&\le (L\|P\| + 1)\|\nabla f(x)\|^2 && \text{(by submultiplicativity of } \|\cdot\|\text{)} \\
&\le (LD + 1)\|\nabla f(x)\|^2. && \text{(by } \|P\| \le \operatorname{diam}(\mathcal{P}) \le D\text{)}
\end{aligned}$$

A.2 Proof of Lemma 3.1

When $x \notin \mathcal{X}^\star$, the denominator $f(x) - f^\star > 0$ and the ratio feedback is well-defined. Note that $r_x(P)$ simply translates and scales $u_x(P)$ by a positive factor $f(x) - f^\star > 0$ (since $x \notin \mathcal{X}^\star$):

$$r_x(P) = \tfrac{u_x(P) - f^\star}{f(x) - f^\star}.$$

Hence the gradient of $r_x(P)$ with respect to P is $\nabla r_x(P) = \frac{\nabla u_x(P)}{f(x)-f^\star} = -\frac{\nabla f(x - P\nabla f(x))\nabla f(x)^\top}{f(x)-f^\star}$. Non-negativity of $r_x(P)$ follows from the fact that $f(x) > f^\star$ for all $x \notin \mathcal{X}^\star$ and the numerator is non-negative. The convexity of $r_x(P)$ follows from the convexity of $u_x(P)$ in **Proposition 3.1**. Since $u_x(P)$ is $L\|\nabla f(x)\|^2$-smooth, the ratio feedback $r_x(P)$ is also smooth with smoothness constant $\frac{L\|\nabla f(x)\|^2}{f(x)-f^\star} \le 2L^2$. Suppose further $\operatorname{diam}(\mathcal{P}) \le D$. Since $u_x(P)$ is $(LD+1)\|\nabla f(x)\|^2$-Lipschitz, the ratio feedback has Lipschitz constant $\frac{(LD+1)\|\nabla f(x)\|^2}{f(x)-f^\star} \le 2L(LD+1)$.

A.3 Proof of Lemma 3.2

Hypergradient feedback $h_x(P) = \frac{u_x(P) - f(x)}{\|\nabla f(x)\|^2}$ simply translates $u_x(P)$ and divides by a factor of $\|\nabla f(x)\|^2$. The convexity, smoothness, and Lipschitz continuity of $h_x(P)$ follow immediately from the same properties of $u_x(P)$ in **Proposition 3.1**. The expression of gradient $\nabla h_x(P)$, smoothness constant, and Lipschitz constant are obtained by dividing those of $u_x(P)$ by the scaling factor $\|\nabla f(x)\|^2$.

A.4 Proof of Equation 6

For any $x, y \in \mathbb{R}^n$ and $t \in [0,1]$, define $x_t \coloneqq x + t(y-x)$. Since $(P_+^\star)^{1/2}\nabla^2 f(z)(P_+^\star)^{1/2} \succeq \frac{1}{\kappa^\star} I$ for any $z \in \mathbb{R}^n$,

$$\begin{aligned}
f(y) &= f(x) + \langle \nabla f(x), y-x\rangle + \textstyle\int_0^1 (1-t)\langle y-x, \nabla^2 f(x_t)(y-x)\rangle \mathrm{d}t \\
&\geq f(x) + \langle \nabla f(x), y-x\rangle + \tfrac{1}{\kappa^\star}\|x-y\|^2_{(P_+^\star)^{-1}} \textstyle\int_0^1 (1-t)\mathrm{d}t \qquad \text{(by } (P_+^\star)^{1/2}\nabla^2 f(z)(P_+^\star)^{1/2} \succeq \tfrac{1}{\kappa^\star} I\text{)} \\
&= f(x) + \langle \nabla f(x), y-x\rangle + \tfrac{1}{2\kappa^\star}\|x-y\|^2_{(P_+^\star)^{-1}}.
\end{aligned}$$

Minimize both sides over $y \in \mathbb{R}^n$ to obtain

$$f^\star \geq f(x) + \min_{d\in\mathbb{R}^n} \langle \nabla f(x), d\rangle + \tfrac{1}{2\kappa^\star}\|d\|^2_{(P_+^\star)^{-1}} = f(x) - \tfrac{\kappa^\star}{2}\|\nabla f(x)\|^2_{P_+^\star},$$

which rearranges to

$$f(x) - f^\star \leq \tfrac{\kappa^\star}{2}\|\nabla f(x)\|^2_{P_+^\star}. \tag{23}$$

Using $(P_+^\star)^{1/2}\nabla^2 f(z)(P_+^\star)^{1/2} \preceq I$, we have

$$\begin{aligned}
f(x - P_+^\star \nabla f(x)) &= f(x) - \langle \nabla f(x), P_+^\star \nabla f(x)\rangle + \textstyle\int_0^1 (1-t)\langle \nabla f(x), P_+^\star \nabla^2 f(x_t) P_+^\star \nabla f(x)\rangle \mathrm{d}t \\
&\leq f(x) - \langle \nabla f(x), P_+^\star \nabla f(x)\rangle + \textstyle\int_0^1 (1-t)\|\nabla f(x)\|^2_{P_+^\star}\mathrm{d}t \qquad \text{(by } (P_+^\star)^{1/2}\nabla^2 f(z)(P_+^\star)^{1/2} \preceq I\text{)} \\
&= f(x) - \|\nabla f(x)\|^2_{P_+^\star} + \tfrac{1}{2}\|\nabla f(x)\|^2_{P_+^\star} \\
&= f(x) - \tfrac{1}{2}\|\nabla f(x)\|^2_{P_+^\star} \\
&\leq f(x) - \tfrac{1}{\kappa^\star}[f(x) - f^\star]. \qquad \text{(by (23))}
\end{aligned}$$

Rearrange to conclude $f(x - P_+^\star \nabla f(x)) - f^\star \leq (1 - \frac{1}{\kappa^\star})[f(x) - f^\star]$.

A.5 Proof of Proposition 3.2

- Ratio feedback: Definition of $P_r^\star$ and $P_+^\star \in \mathcal{P}$ together imply $r_x(P_r^\star) \leq \max_{x\notin\mathcal{X}^\star} r_x(P_+^\star) \overset{(6)}{\leq} 1 - \frac{1}{\kappa^\star}$. If f is a strongly convex quadratic with Hessian $A \succ 0$, then both global optimal preconditioner $P_+^\star$ and the minimax stepsize $P_r^\star$ are Hessian inverse with optimal condition number $\kappa^\star = 1$ and minimax feedback $r_x(A^{-1}) = 0$.
- Hypergradient feedback: Since $\frac{1}{L} I \in \mathcal{P}$, the definition of $P_h^\star$ and descent lemma together imply

  $$h_x(P_h^\star) \leq \max_{x\notin\mathcal{X}^\star} h_x(\tfrac{1}{L} I) \leq -\tfrac{1}{2L}.$$

  Now we show that $P_h^\star = \frac{1}{L} I$ if f is a strongly convex quadratic. Without loss of generality, assume $f(x) = \frac{1}{2}\langle x, Ax\rangle$ is homogeneous. Then $x^\star = 0$ is the unique optimal solution and

  $$\begin{aligned}
  \min_{P\in\mathcal{P}} \max_{x\neq 0}\ h_x(P) &= \min_{P\in\mathcal{P}} \max_{x\neq 0} \tfrac{f(x - P\nabla f(x)) - f(x)}{\|\nabla f(x)\|^2} \\
  &\geq \min_{P\in\mathbb{R}^{n\times n}} \max_{x\neq 0} \tfrac{f(x - P\nabla f(x)) - f(x)}{\|\nabla f(x)\|^2} \qquad \text{(by } \mathcal{P} \subseteq \mathbb{R}^{n\times n}\text{)} \\
  &\geq \max_{x\neq 0} \min_{P\in\mathbb{R}^{n\times n}} \tfrac{f(x - P\nabla f(x)) - f(x)}{\|\nabla f(x)\|^2} \qquad \text{(by weak duality)} \\
  &= \max_{x\neq 0} \tfrac{f^\star - f(x)}{\|\nabla f(x)\|^2} \qquad (24) \\
  &= \max_{x\neq 0} -\tfrac{\langle x, Ax\rangle}{2\|Ax\|^2} \qquad \text{(by } f^\star = 0 \text{ and definition of } f\text{)} \\
  &= \max_{x\neq 0} \tfrac{\langle Ax, (-A^{-1})Ax\rangle}{2\|Ax\|^2} = \tfrac{1}{2}\lambda_{\max}(-A^{-1}) = -\tfrac{1}{2L},
  \end{aligned}$$

where (24) plugs in the minimizer $P = A^{-1}$ that drives x to optimal solution $x^\star$ in one step. On the other hand, descent lemma guarantees $\max_{x\neq 0} h_x(\frac{1}{L}I) \leq -\frac{1}{2L}$. Hence, $\frac{1}{L}I$ achieves the minimax feedback and is the minimax optimal stepsize with respect to hypergradient feedback.

B Proof of results in Section 4

B.1 Proof of Theorem 4.1

The result follows immediately from the definition of r_k and the AM-GM inequality:

$$\frac{f(x^{K+1})-f^\star}{f(x^1)-f^\star} = \prod_{k=1}^K \frac{f(x^{k+1})-f^\star}{f(x^k)-f^\star} = \prod_{k=1}^K r_k \leq (\frac{1}{K}\sum_{k=1}^K r_k)^K.$$

B.2 Proof of Theorem 4.2

Convex f. Observe that

$$\begin{aligned}\frac{1}{f(x^{K+1})-f^\star} &= \sum_{k=1}^K \left[\frac{1}{f(x^{k+1})-f^\star} - \frac{1}{f(x^k)-f^\star}\right] + \frac{1}{f(x^1)-f^\star} \\ &= \sum_{k=1}^K \left[\frac{f(x^k)-f(x^{k+1})}{[f(x^{k+1})-f^\star][f(x^k)-f^\star]}\right] + \frac{1}{f(x^1)-f^\star} \\ &= \sum_{k=1}^K \left[\frac{-h_k\|\nabla f(x^k)\|^2}{[f(x^{k+1})-f^\star][f(x^k)-f^\star]}\right] + \frac{1}{f(x^1)-f^\star}. \end{aligned} \tag{25}$$

Since $h_k = \frac{f(x^{k+1})-f(x^k)}{\|\nabla f(x^k)\|^2} \leq 0$ by monotonicity $f(x^{k+1}) \leq f(x^k)$ and $f(x)-f^\star \leq \|\nabla f(x)\|\cdot\|x-x^\star\|$ by convexity of f, we have

$$\frac{-h_k\|\nabla f(x^k)\|^2}{[f(x^{k+1})-f^\star][f(x^k)-f^\star]} \geq \frac{-h_k\|\nabla f(x^k)\|^2}{[f(x^k)-f^\star]^2} \geq \frac{-h_k}{\mathrm{dist}(x^k,\mathcal{X}^\star)^2} \geq \frac{-h_k}{\Delta^2}.$$

Hence, (25) can be lower bounded by

$$\frac{1}{f(x^{K+1})-f^\star} \geq -\frac{1}{\Delta^2}\sum_{k=1}^K h_k + \frac{1}{f(x^1)-f^\star} \geq \max\{-\frac{1}{\Delta^2}\sum_{k=1}^K h_k, \frac{1}{f(x^1)-f^\star}\}.$$

The desired result follows by taking the reciprocal on both sides of the inequality.

μ-strongly convex f. The desired inequality follows from the following steps:

$$\begin{aligned}\frac{f(x^{K+1})-f^\star}{f(x^1)-f^\star} &\leq (\frac{1}{K}\sum_{k=1}^K \frac{f(x^{k+1})-f^\star}{f(x^k)-f^\star})^K && \text{(by AM-GM inequality)} \\ &= (1+\frac{1}{K}\sum_{k=1}^K \frac{f(x^{k+1})-f(x^k)}{f(x^k)-f^\star})^K \\ &= (1+\frac{1}{K}\sum_{k=1}^K h_k\frac{\|\nabla f(x^k)\|^2}{f(x^k)-f^\star})^K && \text{(by definition of } h_k) \\ &\leq (1-\frac{2\mu}{K}\sum_{k=1}^K -h_k)^K. && \text{(by } h_k\leq 0 \text{ and } \frac{\|\nabla f(x^k)\|^2}{f(x^k)-f^\star}\geq 2\mu)\end{aligned}$$

B.3 Proof of Lemma 4.1

Recall that $x^{k+1} = \mathcal{M}(x^{k+1/2}, x^k)$ and $x^{k+1/2} = x^k - P_k\nabla f(x^k)$. The per iteration progress and feedback are

$$r_k = \frac{f(x^{k+1})-f^\star}{f(x^k)-f^\star}, \quad r_{x^k}(P_k) = \frac{f(x^{k+1/2})-f^\star}{f(x^k)-f^\star}, \quad h_k = \frac{f(x^{k+1})-f(x^k)}{\|\nabla f(x^k)\|^2}, \quad h_{x^k}(P_k) = \frac{f(x^{k+1/2})-f(x^k)}{\|\nabla f(x^k)\|^2}.$$

- *Vanilla* landscape satisfies $x^{k+1} = x^{k+1/2}$. Then $r_k = r_{x^k}(P_k)$ and $h_k = h_{x^k}(P_k)$ follow immediately.

- *Monotone* landscape satisfies $f(x^{k+1}) \le \min\{f(x^{k+1/2}), f(x^k)\}$ and thus

$$r_k = \tfrac{f(x^{k+1})-f^\star}{f(x^k)-f^\star} \le \tfrac{\min\{f(x^{k+1/2}),f(x^k)\}-f^\star}{f(x^k)-f^\star} = \min\{r_{x^k}(P_k), 1\};$$
$$h_k = \tfrac{f(x^{k+1})-f(x^k)}{\|\nabla f(x^k)\|^2} \le \tfrac{\min\{f(x^{k+1/2}),f(x^k)\}-f(x^k)}{\|\nabla f(x^k)\|^2} = \min\{h_{x^k}(P_k), 0\}.$$

- *Lookahead* landscape satisfies $x^{k+1} = x^{k+1/2} - \frac{1}{L}\nabla f(x^{k+1/2})$ and descent lemma implies

$$f(x^{k+1}) = f(x^{k+1/2} - \tfrac{1}{L}\nabla f(x^{k+1/2})) \le f(x^{k+1/2}) - \tfrac{1}{2L}\|\nabla f(x^{k+1/2})\|^2.$$

  Then it follows that

$$\begin{aligned} r_k = \tfrac{f(x^{k+1})-f^\star}{f(x^k)-f^\star} &\le \tfrac{f(x^{k+1/2})-f^\star}{f(x^k)-f^\star} - \tfrac{1}{2L}\tfrac{\|\nabla f(x^{k+1/2})\|^2}{f(x^k)-f^\star} \\ &= \tfrac{f(x^{k+1/2})-f^\star}{f(x^k)-f^\star} - \tfrac{1}{2L}\tfrac{\|\nabla f(x^{k+1/2})\|^2\|\nabla f(x^k)\|^2}{[f(x^k)-f^\star]^2}\tfrac{f(x^k)-f^\star}{\|\nabla f(x^k)\|^2} \\ &= r_{x^k}(P_k) - \tfrac{1}{2L}\|\nabla r_{x^k}(P_k)\|_F^2 \tfrac{f(x^k)-f^\star}{\|\nabla f(x^k)\|^2} \qquad (26) \\ &\le r_{x^k}(P_k) - \tfrac{1}{4L^2}\|\nabla r_{x^k}(P_k)\|_F^2, \qquad (27) \end{aligned}$$

  where (26) uses the relation

$$\|\nabla r_{x^k}(P_k)\|_F^2 = \left\|\tfrac{\nabla f(x^{k+1/2})\nabla f(x^k)^\top}{f(x^k)-f^\star}\right\|_F^2 = \tfrac{\|\nabla f(x^{k+1/2})\|^2\|\nabla f(x^k)\|^2}{[f(x^k)-f^\star]^2}$$

  and the inequality in (27) uses the fact $f(x) - f^\star \ge \frac{1}{2L}\|\nabla f(x)\|^2$ for L-smooth f. Similarly,

$$\begin{aligned} h_k = \tfrac{f(x^{k+1})-f(x^k)}{\|\nabla f(x^k)\|^2} &\le \tfrac{f(x^{k+1/2})-f^\star}{\|\nabla f(x^k)\|^2} - \tfrac{1}{2L}\tfrac{\|\nabla f(x^{k+1/2})\|^2}{\|\nabla f(x^k)\|^2} \\ &= h_{x^k}(P_k) - \tfrac{1}{2L}\|\nabla h_{x^k}(P_k)\|_F^2, \qquad (28) \end{aligned}$$

  where (28) uses the relation

$$\|\nabla h_{x^k}(P_k)\|_F^2 = \left\|\tfrac{\nabla f(x^{k+1/2})\nabla f(x^k)^\top}{\|\nabla f(x^k)\|^2}\right\|_F^2 = \tfrac{\|\nabla f(x^{k+1/2})\|^2}{\|\nabla f(x^k)\|^2}.$$

- *Monotone Lookahead* landscape satisfies

$$f(x^{k+1}) \le \min\{f(x^{k+1/2} - \tfrac{1}{L}\nabla f(x^{k+1/2})), f(x^k)\}.$$

  Combining the analysis of lookahead and monotone landscape completes the proof.

C Proof of results in Section 5

C.1 Proof of Lemma 5.1

For any $\hat{P} \in \mathcal{P}$, online gradient descent update $P_{k+1} = \Pi_{\mathcal{P}}[P_k - \eta_k \nabla \ell_{x^k}(P_k)]$ satisfies

$$\begin{aligned} \|P_{k+1} - \hat{P}\|_F^2 &= \|\Pi_{\mathcal{P}}[P_k - \eta_k \nabla \ell_{x^k}(P_k)] - \hat{P}\|_F^2 \\ &\le \|P_k - \eta_k \nabla \ell_{x^k}(P_k) - \hat{P}\|_F^2 \qquad \text{(by non-expansiveness of projection)} \\ &= \|P_k - \hat{P}\|_F^2 - 2\eta_k \langle \nabla \ell_{x^k}(P_k), P_k - \hat{P}\rangle + \eta_k^2 \|\nabla \ell_{x^k}(P_k)\|_F^2 \\ &\le \|P_k - \hat{P}\|_F^2 - 2\eta_k[\ell_{x^k}(P_k) - \ell_{x^k}(\hat{P})] + \eta_k^2 \|\nabla \ell_{x^k}(P_k)\|_F^2, \qquad (29) \end{aligned}$$

where (29) uses convexity $\ell_{x^k}(\hat{P}) \geq \ell_{x^k}(P_k) + \langle \nabla \ell_{x^k}(P_k), \hat{P} - P_k \rangle$. Dividing both sides by $2\eta_k$, we get

$$\ell_{x^k}(P_k) \leq \ell_{x^k}(\hat{P}) + \tfrac{1}{2\eta_k}[\|P_k - \hat{P}\|_F^2 - \|P_{k+1} - \hat{P}\|_F^2] + \tfrac{\eta_k}{2}\|\nabla \ell_{x^k}(P_k)\|_F^2.$$

Telescoping the relation from $k = 1$ to K and dropping the negative term $-\frac{1}{2\eta_K}\|P_{K+1} - \hat{P}\|_F^2$, we have

$$\begin{aligned}\textstyle\sum_{k=1}^K \ell_{x^k}(P_k) &\leq \textstyle\sum_{k=1}^K \ell_{x^k}(\hat{P}) + \frac{1}{2\eta_1}\|P_1 - \hat{P}\|_F^2 + \sum_{k=1}^{K-1}(\frac{1}{2\eta_{k+1}} - \frac{1}{2\eta_k})\|P_{k+1} - \hat{P}\|_F^2 + \sum_{k=1}^K \frac{\eta_k}{2}\|\nabla \ell_k(P_k)\|_F^2 \\ &\leq \textstyle\sum_{k=1}^K \ell_{x^k}(\hat{P}) + \frac{1}{2\eta_1}\|P_1 - \hat{P}\|_F^2 + \sum_{k=1}^K \frac{\eta_k}{2}\|\nabla \ell_k(P_k)\|_F^2, \end{aligned} \tag{30}$$

where (30) uses the fact $\eta_{k+1} \leq \eta_k$. Equation (30) reduces to (10) when $\eta_k \equiv \eta$. Now, suppose that $\{\ell_{x^k}(P)\}$ are σ-Lipschitz and $\mathrm{diam}(\mathcal{P}) \leq D$. Then $\|\nabla \ell_k(P_k)\|_F^2 \leq \sigma^2$ and $\|P_1 - \hat{P}\|_F^2 \leq D^2$. And (30) implies

$$\textstyle\sum_{k=1}^K \ell_{x^k}(P_k) \leq \sum_{k=1}^K \ell_{x^k}(\hat{P}) + \frac{D^2}{2\eta_1} + \frac{\sigma^2}{2}\sum_{k=1}^K \eta_k.$$

- The choice of the stepsize $\eta_k \equiv \frac{c}{\sqrt{K}}$ gives

$$\textstyle\sum_{k=1}^K \ell_{x^k}(P_k) \leq \sum_{k=1}^K \ell_{x^k}(\hat{P}) + (\frac{D^2}{2c} + \frac{c\sigma^2}{2})\sqrt{K}.$$

- The choice of the stepsize $\eta_k = \frac{c}{\sqrt{k}}$ and the fact $\sum_{k=1}^K \frac{1}{\sqrt{k}} \leq 2\sqrt{K}$ together give

$$\textstyle\sum_{k=1}^K \ell_{x^k}(P_k) \leq \sum_{k=1}^K \ell_{x^k}(\hat{P}) + \frac{D^2}{2c} + \frac{c\sigma^2}{2}\sum_{k=1}^K \frac{1}{\sqrt{k}} \leq \sum_{k=1}^K \ell_{x^k}(\hat{P}) + \frac{D^2}{2c} + c\sigma^2\sqrt{K}.$$

Either case can be further bounded by the right-hand side of (11).

C.2 Proof of Lemma 5.2

Recall from (29) that for any $\hat{P} \in \mathcal{P}$ we have

$$\|P_{k+1} - \hat{P}\|_F^2 \leq \|P_k - \hat{P}\|_F^2 - 2\eta[\ell_{x^k}(P_k) - \ell_{x^k}(\hat{P})] + \eta^2\|\nabla \ell_{x^k}(P_k)\|_F^2.$$

Plug in $\hat{P} = \hat{P}_k$ to obtain $\|P_{k+1} - \hat{P}_k\|_F^2 \leq \|P_k - \hat{P}_k\|_F^2 - 2\eta[\ell_{x^k}(P_k) - \ell_{x^k}(\hat{P}_k)] + \eta^2\|\nabla \ell_{x^k}(P_k)\|_F^2$ and telescoping from $k = 1$ to K yields

$$\begin{aligned}&\textstyle\sum_{k=1}^K \ell_{x^k}(P_k) - \sum_{k=1}^K \ell_{x^k}(\hat{P}_k) - \frac{\eta}{2}\sum_{k=1}^K \|\nabla \ell_k(P_k)\|_F^2 \\ \leq\ & \tfrac{\|\hat{P}_1 - P_1\|_F^2 - \|\hat{P}_K - P_{K+1}\|_F^2}{2\eta} + \tfrac{1}{2\eta}\textstyle\sum_{k=1}^{K-1}[\|\hat{P}_{k+1} - P_{k+1}\|_F^2 - \|\hat{P}_k - P_{k+1}\|_F^2]. \end{aligned} \tag{31}$$

Observe that the sum in the last term can be simplified to

$$\begin{aligned}&\textstyle\sum_{k=1}^{K-1}[\|\hat{P}_{k+1} - P_{k+1}\|_F^2 - \|\hat{P}_k - P_{k+1}\|_F^2] \\ =\ & \textstyle\sum_{k=1}^{K-1}[\|\hat{P}_{k+1}\|_F^2 - \|\hat{P}_k\|_F^2 + 2\langle P_{k+1}, \hat{P}_k - \hat{P}_{k+1}\rangle] \\ =\ & \textstyle\sum_{k=1}^{K-1}[\|\hat{P}_{k+1}\|_F^2 - \|\hat{P}_k\|_F^2 + 2\langle P_{k+1} - P_1, \hat{P}_k - \hat{P}_{k+1}\rangle + 2\langle P_1, \hat{P}_k - \hat{P}_{k+1}\rangle] \\ =\ & \|\hat{P}_K\|_F^2 - \|\hat{P}_1\|_F^2 + 2\textstyle\sum_{k=1}^{K-1}\langle P_{k+1} - P_1, \hat{P}_k - \hat{P}_{k+1}\rangle + 2\langle P_1, \hat{P}_1 - \hat{P}_K\rangle \\ \leq\ & \|\hat{P}_K\|_F^2 - \|\hat{P}_1\|_F^2 + 2\max_{k\leq K}\|P_k - P_1\|_F \textstyle\sum_{k=1}^{K-1}\|\hat{P}_k - \hat{P}_{k+1}\|_F + 2\langle P_1, \hat{P}_1 - \hat{P}_K\rangle \\ =\ & \|\hat{P}_K - P_1\|_F^2 - \|\hat{P}_1 - P_1\|_F^2 + 2\max_{k\leq K}\|P_k - P_1\|_F \textstyle\sum_{k=1}^{K-1}\|\hat{P}_k - \hat{P}_{k+1}\|_F. \end{aligned} \tag{32}$$

Combining (31) and (32), we get

$$\begin{aligned}&\textstyle\sum_{k=1}^K \ell_{x^k}(P_k) - \sum_{k=1}^K \ell_{x^k}(\hat{P}_k) - \frac{\eta}{2}\sum_{k=1}^K \|\nabla \ell_k(P_k)\|_F^2 \\ \le\ & \tfrac{\|\hat{P}_K - P_1\|_F^2 - \|\hat{P}_K - P_{K+1}\|_F^2}{2\eta} + \tfrac{\max_{k\le K}\|P_k - P_1\|_F}{\eta}\textstyle\sum_{k=1}^{K-1}\|\hat{P}_k - \hat{P}_{k+1}\|_F.\end{aligned}$$

Moving $\frac{\eta}{2}\sum_{k=1}^K \|\nabla \ell_{x^k}(P_k)\|_F^2$ to the right and dropping the negative term $-\frac{\|\hat{P}_K - P_{K+1}\|_F^2}{2\eta}$ prove (13). When $\{\ell_{x^k}(P_k)\}$ are σ-Lipschitz, we can bound $\|\nabla \ell_{x^k}(P_k)\|_F^2 \le \sigma^2$ and plug in $\eta = \frac{c}{\sqrt{K}}$ to arrive at (14).

D Proof of results in Section 6

D.1 Proof of Theorem 6.1

Equation (20) follows by plugging the bound (19) into **Theorem 4.1**. Now, suppose `Lookahead OSGM-R` is initialized with $P_1 = \frac{1}{L}I$. To show (21), we plug in $\hat{P} = \frac{1}{L}I$ into (20) and use $r_x(\frac{1}{L}I) \le 1 - \frac{1}{\kappa}$ to obtain

$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](\tfrac{1}{K}\textstyle\sum_{k=1}^K r_{x^k}(\frac{1}{L}I) + \frac{L^2}{K}\|\frac{1}{L}I - \frac{1}{L}I\|_F^2)^K \le [f(x^1) - f^\star](1 - \frac{1}{\kappa})^K. \tag{33}$$

On the other hand, we can plug in $\hat{P} = P_r^\star$ into (20) and use $r_x(P_r^\star) \le 1 - \frac{1}{\kappa^\star}$ to obtain

$$\begin{aligned}f(x^{K+1}) - f^\star &\le [f(x^1) - f^\star](\tfrac{1}{K}\textstyle\sum_{k=1}^K r_{x^k}(P_r^\star) + \frac{L^2}{K}\|\frac{1}{L}I - P_r^\star\|_F^2)^K \\ &\le [f(x^1) - f^\star](1 - \tfrac{1}{\kappa^\star} + \tfrac{L^2}{K}\|\tfrac{1}{L}I - P_r^\star\|_F^2)^K.\end{aligned} \tag{34}$$

Take the minimum of two bounds in (33) and (34) to conclude (21).

D.2 Proof of Theorem 6.2

At every iteration, the function value part of the potential changes by

$$\rho\log(f(x^{k+1}) - f^\star) - \rho\log(f(x^k) - f^\star) = \rho\log\tfrac{f(x^{k+1}) - f^\star}{f(x^k) - f^\star} = \rho\log r_k. \tag{35}$$

On the other hand, the distance part of the potential changes by

$$\begin{aligned}\|P_{k+1} - \hat{P}\|_F^2 &\overset{(9)}{\le} \|P_k - \hat{P}\|_F^2 - 2\eta[r_{x^k}(P_k) - r_{x^k}(\hat{P})] + \eta^2\|\nabla r_{x^k}(P_k)\|_F^2 \\ &= \|P_k - \hat{P}\|_F^2 - 2\eta[r_{x^k}(P_k) - \tfrac{\eta}{2}\|\nabla r_{x^k}(P_k)\|_F^2 - r_{x^k}(\hat{P})] \\ &= \|P_k - \hat{P}\|_F^2 - 2\eta[r_{x^k}(P_k) - \tfrac{1}{4L^2}\|\nabla r_{x^k}(P_k)\|_F^2 - r_{x^k}(\hat{P})] && (\text{by } \eta = \tfrac{1}{2L^2}) \\ &\le \|P_k - \hat{P}\|_F^2 - 2\eta[r_k - r_{x^k}(\hat{P})], && (36)\end{aligned}$$

where (36) uses $r_k \le r_{x^k}(P_k) - \frac{1}{4L^2}\|\nabla r_{x^k}(P_k)\|_F^2$ by **Lemma 4.1** with lookahead landscape action. Combining (35) and (36) gives

$$\begin{aligned}&\varphi(x^{k+1}, P_{k+1}) - \varphi(x^k, P_k) \\ =\ & \rho\log(f(x^{k+1}) - f^\star) - \rho\log(f(x^k) - f^\star) + \|P_{k+1} - \hat{P}\|_F^2 - \|P_k - \hat{P}\|_F^2 \\ \le\ & \rho\log r_k - 2\eta[r_k - r_{x^k}(\hat{P})] && (\text{by (36)}) \\ =\ & \rho\log r_k - 2\eta r_k + 2\eta r_{x^k}(\hat{P}) \\ \le\ & \rho\log r_k - 2\eta r_k + 2\eta(1 - \tfrac{1}{\kappa_{\hat{P}}}). && (\text{by } r_{x^k}(\hat{P}) \le 1 - \tfrac{1}{\kappa_{\hat{P}}}) \\ =\ & \alpha(r_k) + 2\eta(1 - \tfrac{1}{\kappa_{\hat{P}}}),\end{aligned}$$

where the function $\alpha(x) \coloneqq \rho \log x - 2\eta x$ is maximized at $x = \frac{\rho}{2\eta}$ and $\rho \log r_k - 2\eta r_k \le \rho \log \frac{\rho}{2\eta} - \rho$, which implies

$$\varphi(x^{k+1}, P_{k+1}) - \varphi(x^k, P_k) \le \rho \log \tfrac{\rho}{2\eta} - \rho + 2\eta(1 - \tfrac{1}{\kappa_{\hat{P}}})$$

Taking $\rho = 1/L^2 = 2\eta$, we get the desired reduction in the potential function:

$$\varphi(x^{k+1}, P_{k+1}) - \varphi(x^k, P_k) \le -\tfrac{1}{\kappa_{\hat{P}} L^2}.$$

To obtain the iteration complexity, note that

$$\begin{aligned}
\tfrac{1}{L^2} \log(f(x^{K+1}) - f^\star) &\le \varphi(x^{K+1}, P_{K+1}) \\
&\le \varphi(x^1, P_1) - \tfrac{1}{\kappa_{\hat{P}} L^2} K \\
&= \tfrac{1}{L^2} \log[f(x^1) - f^\star] + \|P_1 - P^\star\|_F^2 - \tfrac{1}{\kappa_{\hat{P}} L^2} K.
\end{aligned}$$

Hence, `Lookahead OSGM-R` achieves $f(x^{K+1}) - f^\star \le \varepsilon$ for $K \ge \kappa_{\hat{P}} L^2 \|P_1 - P^\star\|_F^2 + \kappa_{\hat{P}} \log(\frac{f(x^1) - f^\star}{\varepsilon})$. This holds for arbitrary benchmark stepsize $\hat{P}$. Taking the minimum (infimum) over $\hat{P}$ completes the proof.

D.3 Proof of Theorem 6.3

The result follows from the following chain of inequalities:

$$\begin{aligned}
& f(x^{K+1}) - f^\star \\
\le\ & [f(x^1) - f^\star](\tfrac{1}{K} \textstyle\sum_{k=1}^K r_k)^K && \text{(by } \textbf{Theorem 4.1}\text{)} \\
\le\ & [f(x^1) - f^\star](\tfrac{1}{K} \textstyle\sum_{k=1}^K r_{x^k}(P_k) - \tfrac{1}{4L^2} \|\nabla r_{x^k}(P_k)\|_F^2)^K && \text{(by } \textbf{Lemma 4.1} \text{ + lookahead action)} \\
=\ & [f(x^1) - f^\star](\tfrac{1}{K} \textstyle\sum_{k=1}^K r_{x^k}(P_k) - \tfrac{\eta}{2} \|\nabla r_{x^k}(P_k)\|_F^2)^K && \text{(by } \eta = \tfrac{1}{2L^2}\text{)} \\
\le\ & [f(x^1) - f^\star](\tfrac{1}{K} \textstyle\sum_{k=1}^K r_{x^k}(\hat{P}_k) + \tfrac{L^2}{K} \rho_K(\{\hat{P}_k\}))^K, && \text{(by } \textbf{Lemma 5.2}\text{)}
\end{aligned}$$

where $\rho_K(\{\hat{P}_k\}) \coloneqq \|\hat{P}_{K+1} - P_1\|_F^2 + 2 \max_{k \le K}\{\|P_k - P_1\|_F\} \mathsf{PL}(\{\hat{P}_k\})$.

D.4 Proof of Lemma 6.1

We start by bounding the numerator $f(x - [\nabla^2 f(x^\star)]^{-1} \nabla f(x)) - f^\star$ of ratio feedback $r_x([\nabla^2 f(x^\star)]^{-1})$. By $f(x) - f^\star \le \frac{L}{2} \|x - x^\star\|^2$ from L-smoothness and $[\nabla^2 f(x^\star)]^{-1} \preceq \frac{1}{\mu} I$ from μ-strong convexity, we upper bound the numerator as follows:

$$\begin{aligned}
f(x - [\nabla^2 f(x^\star)]^{-1} \nabla f(x)) - f^\star &\le \tfrac{L}{2} \|x - [\nabla^2 f(x^\star)]^{-1} \nabla f(x) - x^\star\|^2 \\
&= \tfrac{L}{2} \|[\nabla^2 f(x^\star)]^{-1} [\nabla^2 f(x^\star)(x - x^\star) - (\nabla f(x) - \nabla f(x^\star))]\|^2 \\
&\le \tfrac{L}{2\mu^2} \|\nabla^2 f(x^\star)(x - x^\star) - (\nabla f(x) - \nabla f(x^\star))\|^2. \qquad (37)
\end{aligned}$$

Using $\nabla f(x) - \nabla f(x^\star) = \int_0^1 \nabla^2 f(x^\star + t(x - x^\star))(x - x^\star) \mathrm{d}t$, the norm in (37) can be further bounded by

$$\begin{aligned}
& \|\nabla^2 f(x^\star)(x - x^\star) - (\nabla f(x) - \nabla f(x^\star))\| \\
=\ & \|\nabla^2 f(x^\star)(x - x^\star) - \textstyle\int_0^1 \nabla^2 f(x^\star + t(x - x^\star))(x - x^\star) \mathrm{d}t\| \\
=\ & \|\textstyle\int_0^1 [\nabla^2 f(x^\star) - \nabla^2 f(x^\star + t(x - x^\star))](x - x^\star) \mathrm{d}t\| \\
\le\ & \textstyle\int_0^1 tH \|x - x^\star\|^2 \mathrm{d}t = \frac{H}{2} \|x - x^\star\|^2, && \text{(by } H\text{-Lipschitz Hessian)}
\end{aligned}$$

and thus (37) becomes

$$f(x-[\nabla^2 f(x^\star)]^{-1}\nabla f(x))-f^\star \le \tfrac{L}{2\mu^2}(\tfrac{H}{2}\|x-x^\star\|^2)^2 = \tfrac{LH^2}{8\mu^2}\|x-x^\star\|^4. \tag{38}$$

Dividing both sides by $f(x)-f^\star$ and using $f(x)-f^\star \ge \frac{\mu}{2}\|x-x^\star\|^2$, we conclude that

$$\begin{aligned} r_x([\nabla^2 f(x^\star)]^{-1}) &= \tfrac{f(x-[\nabla^2 f(x^\star)]^{-1}\nabla f(x))-f^\star}{f(x)-f^\star} \\ &\le \tfrac{LH^2}{8\mu^2}\tfrac{\|x-x^\star\|^4}{f(x)-f^\star} \le \tfrac{LH^2}{4\mu^3}\|x-x^\star\|^2 = \tfrac{H^2\kappa}{4\mu^2}\|x-x^\star\|^2. \end{aligned}$$

D.5 Proof of Theorem 6.4

Plugging $r_x([\nabla^2 f(x^\star)]^{-1}) \le \frac{H^2\kappa}{4\mu^2}\|x-x^\star\|^2$ from **Lemma 6.1** into **Theorem 6.1**, we get

$$f(x^{K+1})-f^\star \le [f(x^1)-f^\star](\tfrac{H^2\kappa}{4\mu^2}\tfrac{1}{K}\textstyle\sum_{k=1}^K \|x^k-x^\star\|^2 + \tfrac{L^2}{K}\|P_1-[\nabla^2 f(x^\star)]^{-1}\|_F^2)^K. \tag{39}$$

Recall from **Theorem 6.1** that `Lookahead OSGM-R` satisfies $f(x^k)-f^\star \le (1-\frac{1}{\kappa})^{k-1}[f(x^1)-f^\star]$ and thus together with $\frac{\mu}{2}\|x^k-x^\star\|^2 \le f(x^k)-f^\star$, we have

$$\begin{aligned} \textstyle\sum_{k=1}^K \|x^k-x^\star\|^2 &\le \textstyle\sum_{k=1}^K \frac{2}{\mu}[f(x^k)-f^\star] \\ &\le \tfrac{2}{\mu}[f(x^1)-f^\star]\textstyle\sum_{k=1}^K(1-\frac{1}{\kappa})^{k-1} \\ &\le \tfrac{2}{\mu}[f(x^1)-f^\star]\kappa, \end{aligned} \tag{40}$$

where the last inequality uses the relation $\sum_{k=1}^K \gamma^{k-1} \le \frac{1}{1-\gamma}$ for $\gamma \in [0,1)$. Plugging (40) back into (39) gives

$$\begin{aligned} f(x^{K+1})-f^\star &\le [f(x^1)-f^\star](\tfrac{H^2\kappa}{4\mu^2}\tfrac{1}{K}\textstyle\sum_{k=1}^K \|x^k-x^\star\|^2 + \tfrac{L^2}{K}\|P_1-[\nabla^2 f(x^\star)]^{-1}\|_F^2)^K \\ &\le [f(x^1)-f^\star](\tfrac{H^2\kappa^2}{2\mu^3}\tfrac{f(x^1)-f^\star}{K} + \tfrac{L^2}{K}\|P_1-[\nabla^2 f(x^\star)]^{-1}\|_F^2)^K \\ &= [f(x^1)-f^\star](\tfrac{C}{K})^K, \end{aligned} \tag{41}$$

where $C = \frac{H^2\kappa^2}{2\mu^3}[f(x^1)-f^\star] + L^2\|\frac{1}{L}I-[\nabla^2 f(x^\star)]^{-1}\|_F^2$.

D.6 Proof of Theorem 6.5

Fix a benchmark stepsize $\hat{P}$ and $\eta \le \frac{1}{4L^2}$. Recall from (18) that

$$\begin{aligned} \textstyle\sum_{k=1}^K r_k &\le \textstyle\sum_{k=1}^K r_{x^k}(\hat{P}) + \frac{1}{2\eta}\|P_1-\hat{P}\|_F^2 + \frac{1}{2}(\eta-\frac{1}{2L^2})\sum_{k=1}^K \|\nabla r_{x^k}(P_k)\|_F^2 && \text{(by (18))} \\ &\le \textstyle\sum_{k=1}^K r_{x^k}(\hat{P}) + \frac{1}{2\eta}\|P_1-\hat{P}\|_F^2 - \frac{1}{8L^2}\sum_{k=1}^K \|\nabla r_{x^k}(P_k)\|_F^2. && \text{(by } \eta \le \tfrac{1}{4L^2}) \end{aligned}$$

To lower bound the gradient norm of ratio feedback, observe that

$$\|\nabla r_{x^k}(P_k)\|_F^2 = \tfrac{\|\nabla f(x^k)\|^2\|\nabla f(x^{k+1/2})\|^2}{[f(x^k)-f^\star]^2} = \tfrac{\|\nabla f(x^{k+1})\|^2}{\|\nabla f(x^k)\|^2}\tfrac{\|\nabla f(x^{k+1/2})\|^2}{\|\nabla f(x^{k+1})\|^2}\tfrac{\|\nabla f(x^k)\|^4}{[f(x^k)-f^\star]^2}.$$

The middle fraction can be bounded using $\|\nabla f(x-\frac{1}{L}\nabla f(x))\| \le \|\nabla f(x)\|$; the last fraction can be lower bounded using the relation $\frac{1}{2\mu}\|\nabla f(x^k)\|^2 \ge f(x^k)-f^\star$.

Hence, we have $\|\nabla r_{x^k}(P_k)\|_F^2 \geq 4\mu^2 \frac{\|\nabla f(x^{k+1})\|^2}{\|\nabla f(x^k)\|^2}$ and the bound on the cumulative progress becomes

$$\begin{aligned}\textstyle\sum_{k=1}^K r_k &\leq \textstyle\sum_{k=1}^K r_{x^k}(\hat{P}) + \frac{1}{2\eta}\|P_1 - \hat{P}\|_F^2 - \frac{1}{8L^2}\sum_{k=1}^K \|\nabla r_{x^k}(P_k)\|_F^2 \\ &\leq \textstyle\sum_{k=1}^K r_{x^k}(\hat{P}) + \frac{1}{2\eta}\|P_1 - \hat{P}\|_F^2 - \frac{1}{2\kappa^2}\sum_{k=1}^K \frac{\|\nabla f(x^{k+1})\|^2}{\|\nabla f(x^k)\|^2}.\end{aligned}$$

A rearrangement gives

$$\textstyle\sum_{k=1}^K \frac{\|\nabla f(x^{k+1})\|^2}{\|\nabla f(x^k)\|^2} \leq 2\kappa^2[\sum_{k=1}^K r_{x^k}(\hat{P}) - \sum_{k=1}^K r_k] + \frac{\kappa^2}{\eta}\|P_1 - \hat{P}\|_F^2. \tag{42}$$

Then we can bound the suboptimality by

$$\begin{aligned}f(x^{K+1}) - f^\star &\leq \tfrac{1}{2\mu}\|\nabla f(x^{K+1})\|^2 \\ &\leq \tfrac{1}{2\mu}\|\nabla f(x^1)\|^2(\tfrac{1}{K}\textstyle\sum_{k=1}^K \frac{\|\nabla f(x^{k+1})\|^2}{\|\nabla f(x^k)\|^2})^K && \text{(by AM-GM inequality)} \\ &\overset{(42)}{\leq} \tfrac{1}{2\mu}\|\nabla f(x^1)\|^2(\tfrac{2\kappa^2}{K}[\textstyle\sum_{k=1}^K r_{x^k}(\hat{P}) - \sum_{k=1}^K r_k] + \frac{\kappa^2}{\eta K}\|P_1 - \hat{P}\|_F^2)^K. && (43)\end{aligned}$$

Finally, we do a case analysis on the sign of $\sum_{k=1}^K r_{x^k}(\hat{P}) - \sum_{k=1}^K r_k$.

Case 1. $\sum_{k=1}^K r_k \leq \sum_{k=1}^K r_{x^k}(\hat{P})$: The progress of `Lookahead OSGM-R` is better than stepsize $\hat{P}$.

Case 2. $\sum_{k=1}^K r_k \geq \sum_{k=1}^K r_{x^k}(\hat{P})$: relation (43) can be further bounded by

$$f(x^{K+1}) - f(x^1) \leq \tfrac{1}{2\mu}\|\nabla f(x^1)\|^2(\tfrac{\kappa^2\|P_1 - \hat{P}\|_F^2}{\eta K})^K.$$

D.7 Proof of Theorem 6.6

Using **Lemma 4.1** with lookahead action, we bound the cumulative progress by

$$\textstyle\sum_{k=1}^K h_k \leq \sum_{k=1}^K \min\{h_{x^k}(P_k) - \frac{1}{2L}\|\nabla h_{x^k}(P_k)\|_F^2, 0\} \leq \sum_{k=1}^K h_{x^k}(P_k) - \frac{1}{2L}\sum_{k=1}^K \|\nabla h_{x^k}(P_k)\|_F^2. \tag{44}$$

By **Lemma 5.1** and the choice of stepsize $\eta = \frac{1}{L}$, we further bound the right-hand side of (44):

$$\begin{aligned}\textstyle\sum_{k=1}^K h_k &\leq \textstyle\sum_{k=1}^K h_{x^k}(P_k) - \frac{1}{2L}\sum_{k=1}^K \|\nabla h_{x^k}(P_k)\|_F^2 \\ &\leq \textstyle\sum_{k=1}^K h_{x^k}(\hat{P}) + \frac{1}{2\eta}\|P_1 - \hat{P}\|_F^2 + (\frac{\eta}{2} - \frac{1}{2L})\sum_{k=1}^K \|\nabla h_{x^k}(P_k)\|_F^2 && (45) \\ &\leq \textstyle\sum_{k=1}^K h_{x^k}(\hat{P}) + \frac{L}{2}\|P_1 - \hat{P}\|_F^2. && (46)\end{aligned}$$

Together with the hypergradient reduction from **Theorem 4.2** and $h_k \leq 0$, we conclude

Convex f.

$$\begin{aligned}f(x^{K+1}) - f^\star &\leq \min\{\tfrac{\Delta^2}{\sum_{k=1}^K -h_k}, f(x^1) - f^\star\} \\ &\leq \min\{\tfrac{\Delta^2}{\max\{\sum_{k=1}^K -h_{x^k}(\hat{P}) - \frac{L}{2}\|P_1 - \hat{P}\|_F^2, 0\}}, f(x^1) - f^\star\};\end{aligned}$$

μ-strongly convex f.

$$\begin{aligned}f(x^{K+1}) - f^\star &\leq [f(x^1) - f^\star](1 + \tfrac{2\mu}{K}\textstyle\sum_{k=1}^K h_k)^K \\ &\leq [f(x^1) - f^\star](1 - 2\mu\max\{\tfrac{1}{K}\textstyle\sum_{k=1}^K -h_{x^k}(\hat{P}) - \frac{L}{2K}\|P_1 - \hat{P}\|_F^2, 0\})^K.\end{aligned}$$

D.8 Proof of Theorem 6.7

For brevity, we drop the subscript φ and ω for parameters ρ_φ and ρ_ω in both potential functions.

Convex f. We analyze the potential function

$$\omega(x, P) = -\tfrac{\rho}{f(x)-f^\star} + \|P - \tfrac{1}{L}I\|_F^2.$$

At every iteration, the function value part of the potential changes by

$$\begin{aligned}
\tfrac{1}{f(x^{k+1})-f^\star} &= \tfrac{1}{f(x^k)-f^\star} - \tfrac{\|\nabla f(x^k)\|^2 h_k}{[f(x^{k+1})-f^\star][f(x^k)-f^\star]} \\
&\geq \tfrac{1}{f(x^k)-f^\star} - \tfrac{\|\nabla f(x^k)\|^2 h_k}{[f(x^k)-f^\star]^2} && \text{(by } f(x^{k+1}) \leq f(x^k) \text{ and } h_k \leq 0) \\
&\geq \tfrac{1}{f(x^k)-f^\star} - \tfrac{h_k}{\Delta^2}. && \text{(by } \tfrac{f(x^k)-f^\star}{\|\nabla f(x^k)\|^2} \geq \Delta^2)
\end{aligned}$$

On the other hand, the distance part changes by

$$\begin{aligned}
\|P_{k+1} - \tfrac{1}{L}I\|_F^2 &\leq \|P_k - \tfrac{1}{L}I\|_F^2 - 2\eta[h_{x^k}(P_k) - h_{x^k}(\tfrac{1}{L}I)] + \eta^2\|\nabla h_{x^k}(P_k)\|_F^2 \\
&\leq \|P_k - \tfrac{1}{L}I\|_F^2 - 2\eta[h_{x^k}(P_k) - \tfrac{\eta}{2}\|\nabla h_{x^k}(P_k)\|_F^2] - \tfrac{\eta}{L} && \text{(by descent lemma } h_{x^k}(\tfrac{1}{L}I) \leq -\tfrac{1}{2L}) \\
&\leq \|P_k - \tfrac{1}{L}I\|_F^2 - 2\eta[h_{x^k}(P_k) - \tfrac{1}{2L}\|\nabla h_{x^k}(P_k)\|_F^2] - \tfrac{\eta}{L} && \text{(by } \eta \leq \tfrac{1}{L}) \\
&\leq \|P_k - \tfrac{1}{L}I\|_F^2 - 2\eta h_k - \tfrac{\eta}{L}. && \text{(by } \textbf{Lemma 4.1} \text{ + lookahead action)}
\end{aligned}$$

Combining both parts to obtain the change of potential:

$$\omega(x^{k+1}, P_{k+1}) - \omega(x^k, P_k) \leq \tfrac{\rho h_k}{\Delta^2} - 2\eta h_k - \tfrac{\eta}{L} = (\tfrac{\rho}{\Delta^2} - 2\eta)h_k - \tfrac{\eta}{L}.$$

For $\rho \geq 2\eta\Delta^2$, the potential function will strictly decrease. Taking $\rho = 2\eta\Delta^2$ and $\eta = \frac{1}{L}$, we have

$$\omega(x^{k+1}, P_{k+1}) - \omega(x^k, P_k) \leq -\tfrac{1}{L^2}.$$

To obtain the iteration complexity, note that

$$-\tfrac{1}{L}\tfrac{2\Delta^2}{f(x^{K+1})-f^\star} + \|P_{K+1} - \tfrac{1}{L}I\|_F^2 \leq -\tfrac{1}{L}\tfrac{2\Delta^2}{f(x^1)-f^\star} + \|P_1 - \tfrac{1}{L}I\|_F^2 - \tfrac{1}{L^2}K.$$

Hence, `Monotone Lookahead OSGM-H` with $P_1 = \frac{1}{L}I$ achieves $f(x^{K+1}) - f^\star \leq \varepsilon$ for

$$K \geq L^2\|P_1 - \tfrac{1}{L}I\|_F^2 + \tfrac{2L\Delta^2}{\varepsilon} = \tfrac{2L\Delta^2}{\varepsilon}.$$

μ-strongly convex f. We analyze the potential function

$$\varphi(x, P) = \rho\log(f(x) - f^\star) + \|P - \tfrac{1}{L}I\|_F^2.$$

At every iteration, the function value part of the potential changes by

$$\begin{aligned}
f(x^{k+1}) - f^\star &= f(x^{k+1}) - f(x^k) + f(x^k) - f^\star \\
&= h_k\|\nabla f(x^k)\|^2 + f(x^k) - f^\star \\
&= (1 + h_k\tfrac{\|\nabla f(x^k)\|^2}{f(x^k)-f^\star})(f(x^k) - f^\star) \\
&\leq (1 + 2\mu h_k)(f(x^k) - f^\star),
\end{aligned}$$

where the last inequality uses $\frac{\|\nabla f(x^k)\|^2}{f(x^k)-f^\star} \geq 2\mu$. The change of distance part $\|P - \frac{1}{L}I\|_F^2$ is the same as convex case. Combine both parts to obtain the change of potential:

$$\begin{aligned}
&\varphi(x^{k+1}, P_{k+1}) - \varphi(x^k, P_k) \\
&= \rho \log(f(x^{k+1}) - f^\star) - \rho \log(f(x^k) - f^\star) + \|P_{k+1} - \tfrac{1}{L}I\|_F^2 - \|P_k - \tfrac{1}{L}I\|_F^2 \\
&\leq \rho \log(1 + 2\mu h_k) - 2\eta h_k - \tfrac{\eta}{L} = \alpha(h_x) - \tfrac{\eta}{L},
\end{aligned}$$

where the function $\alpha(x) \coloneqq \rho \log(1 + 2\mu x) - 2\eta x$ is maximized at $x = \frac{\mu\rho - \eta}{2\eta\mu}$. Taking $\rho = \eta/\mu$, we get

$$\varphi(x^{k+1}, P_{k+1}) - \varphi(x^k, P_k) \leq \rho \log(1 + \tfrac{\mu\rho-\eta}{\eta}) - \tfrac{\mu\rho-\eta}{\mu} - \tfrac{\eta}{L} = -\tfrac{\eta}{L}.$$

Take $\eta = 1/L$ to get

$$\varphi(x^{k+1}, P_{k+1}) - \varphi(x^k, P_k) \leq -\tfrac{1}{L^2}.$$

To obtain the iteration complexity, note that

$$\tfrac{1}{\mu L} \log(f(x^{K+1}) - f^\star) + \|P_{K+1} - \tfrac{1}{L}I\|_F^2 \leq \tfrac{1}{\mu L} \log[f(x^1) - f^\star] + \|P_1 - \tfrac{1}{L}I\|_F^2 - \tfrac{1}{L^2}K.$$

Hence, `Monotone Lookahead OSGM-H` with $P_1 = \frac{1}{L}I$ achieves $f(x^{K+1}) - f^\star \leq \varepsilon$ for

$$K \geq L^2\|P_1 - \tfrac{1}{L}I\|_F^2 + \kappa \log(\tfrac{f(x^1)-f(x)}{\varepsilon}) = \kappa \log(\tfrac{f(x^1)-f(x)}{\varepsilon}).$$

D.9 Proof of Theorem 6.8

By (44) and **Lemma 5.2**, the cumulative progress of `Monotone Lookahead OSGM-H` is bounded by

$$\begin{aligned}
\textstyle\sum_{k=1}^K h_k &\leq \textstyle\sum_{k=1}^K h_{x^k}(P_k) - \frac{1}{2L}\sum_{k=1}^K \|\nabla h_{x^k}(P_k)\|_F^2 && \text{(by (44))} \\
&\leq \textstyle\sum_{k=1}^K h_{x^k}(\hat{P}_k) + \frac{\|\hat{P}_K - P_1\|_F^2}{2\eta} + \frac{\max_{k\leq K}\|P_k - P_1\|_F}{\eta}\mathsf{PL}(\{\hat{P}_k\}) && \text{(by \textbf{Lemma 5.2})} \\
&\leq \textstyle\sum_{k=1}^K h_{x^k}(\hat{P}_k) + \frac{L}{2}[\|\hat{P}_K - P_1\|_F^2 + 2\max_{k\leq K}\{\|P_k - P_1\|_F\}\mathsf{PL}(\{\hat{P}_k\})]. && \text{(by } \eta = \tfrac{1}{L})
\end{aligned}$$

Plugging the minimum of this bound and the bound $h_k \leq 0$ into hypergradient reduction from **Theorem 4.2** completes the proof.

D.10 Proof of Theorem 6.9

First, we establish a bound on $h_x([\nabla^2 f(x^\star)]^{-1})$ below.

Lemma D.1. *Suppose f is L-smooth μ-strongly convex has H-Lipschitz Hessian. Then the hypergradient feedback of Hessian inverse at $x^\star$ satisfies $h_x([\nabla^2 f(x^\star)]^{-1}) + \frac{f(x)-f^\star}{\|\nabla f(x)\|^2} \leq \frac{H^2\kappa}{8\mu^3}\|x - x^\star\|^2$ for all $x \notin \mathcal{X}^\star$.*

Proof. Notice that

$$h_x([\nabla^2 f(x^\star)]^{-1}) + \tfrac{f(x)-f^\star}{\|\nabla f(x)\|^2} = \tfrac{f(x-[\nabla^2 f(x^\star)]^{-1}\nabla f(x))-f^\star}{\|\nabla f(x)\|^2}.$$

Using (38) and $\frac{\|x-x^\star\|^2}{\|\nabla f(x)\|^2} \leq \frac{1}{\mu^2}$, the right-hand side can be further bounded by

$$\tfrac{f(x-[\nabla^2 f(x^\star)]^{-1}\nabla f(x))-f(x^\star)}{\|\nabla f(x)\|^2} \overset{(38)}{\leq} \tfrac{LH^2}{8\mu^2}\tfrac{\|x-x^\star\|^4}{\|\nabla f(x)\|^2} \leq \tfrac{LH^2}{8\mu^4}\|x - x^\star\|^2 = \tfrac{H^2\kappa}{8\mu^3}\|x - x^\star\|^2.$$

□

Now, we are ready to prove **Theorem 6.9**. Note that

$$\begin{aligned} f(x^{K+1}) - f(x^k) &\leq [f(x^1) - f^\star](\tfrac{1}{K}\textstyle\sum_{k=1}^K \frac{f(x^{k+1})-f^\star}{f(x^k)-f(x^\star)}) && \text{(by AM-GM inequality)} \\ &= [f(x^1) - f^\star](\tfrac{1}{K}\textstyle\sum_{k=1}^K \frac{f(x^{k+1})-f^\star}{\|\nabla f(x^k)\|^2}\frac{\|\nabla f(x^k)\|^2}{f(x^k)-f^\star})^K \\ &\leq [f(x^1) - f^\star](\tfrac{2L}{K}\textstyle\sum_{k=1}^K \frac{f(x^{k+1})-f^\star}{\|\nabla f(x^k)\|^2})^K && \text{(by } L\text{-smoothness of } f) \\ &= [f(x^1) - f^\star](\tfrac{2L}{K}\textstyle\sum_{k=1}^K [h_k + \frac{f(x^k)-f^\star}{\|\nabla f(x^k)\|^2}])^K. && (47) \end{aligned}$$

We now further bound the sum on the right by

$$\begin{aligned} \textstyle\sum_{k=1}^K [h_k + \frac{f(x^k)-f^\star}{\|\nabla f(x^k)\|^2}] &\leq \textstyle\sum_{k=1}^K [h_{x^k}([\nabla^2 f(x^\star)]^{-1}) + \frac{f(x^k)-f^\star}{\|\nabla f(x^k)\|^2}] + \frac{L}{2}\|P_1 - [\nabla^2 f(x^\star)]^{-1}\|_F^2 && \text{(by (46))} \\ &\leq \tfrac{H^2\kappa}{8\mu^3}\textstyle\sum_{k=1}^K \|x^k - x^\star\|^2 + \frac{L}{2}\|P_1 - [\nabla^2 f(x^\star)]^{-1}\|_F^2. && \text{(by } \textbf{Lemma D.1}) \end{aligned}$$

The linear convergence rate of `Monotone Lookahead OSGM-H` in **Theorem 6.6** implies

$$\begin{aligned} \textstyle\sum_{k=1}^K \|x^k - x^\star\|^2 &\leq \tfrac{2}{\mu}\textstyle\sum_{k=1}^K [f(x^k) - f^\star] && \text{(by } \mu\text{-strong convexity of } f) \\ &\leq \tfrac{2}{\mu}[f(x^1) - f^\star]\textstyle\sum_{k=1}^K (1 - \frac{1}{\kappa})^k \leq \frac{2\kappa}{\mu}[f(x^1) - f^\star]. && \text{(by } \textbf{Theorem 6.6}) \end{aligned}$$

The above two inequalities, together with the choice $P_1 = \frac{1}{L}I$, imply

$$\textstyle\sum_{k=1}^K [h_k + \frac{f(x^k)-f^\star}{\|\nabla f(x^k)\|^2}] \leq \frac{H^2\kappa^2}{4\mu^4}[f(x^1) - f^\star] + \frac{L}{2}\|\frac{1}{L}I - [\nabla^2 f(x^\star)]^{-1}\|_F^2. \tag{48}$$

Plugging (48) back into (47) completes the proof:

$$\begin{aligned} f(x^{K+1}) - f(x^k) &\leq [f(x^1) - f^\star](\tfrac{2L}{K}[\tfrac{H^2\kappa^2}{4\mu^4}[f(x^1) - f^\star] + \tfrac{L}{2}\|\tfrac{1}{L}I - [\nabla^2 f(x^\star)]^{-1}\|_F^2])^K \\ &= [f(x^1) - f^\star](\tfrac{1}{K}[\tfrac{H^2\kappa^3}{2\mu^3}[f(x^1) - f^\star] + L^2\|\tfrac{1}{L}I - [\nabla^2 f(x^\star)]^{-1}\|_F^2])^K \\ &= [f(x^1) - f^\star](\tfrac{C}{K})^K, \end{aligned}$$

where $C \coloneqq \frac{H^2\kappa^3}{2\mu^3}[f(x^1) - f^\star] + L^2\|\frac{1}{L}I - [\nabla^2 f(x^\star)]^{-1}\|_F^2]$.

D.11 Proof of Theorem 6.10

Fix a benchmark stepsize $\hat{P}$ and $\eta \leq \frac{1}{2L}$. Recall from (45) that

$$\begin{aligned} \textstyle\sum_{k=1}^K h_k &\leq \textstyle\sum_{k=1}^K h_{x^k}(\hat{P}) + \frac{1}{2\eta}\|P_1 - \hat{P}\|_F^2 + \frac{1}{2}(\eta - \frac{1}{L})\sum_{k=1}^K \|\nabla h_{x^k}(P_k)\|_F^2 && \text{(by (45))} \\ &\leq \textstyle\sum_{k=1}^K h_{x^k}(\hat{P}) + \frac{1}{2\eta}\|P_1 - \hat{P}\|_F^2 - \frac{1}{4L}\sum_{k=1}^K \|\nabla h_{x^k}(P_k)\|_F^2. && \text{(by } \eta \leq \tfrac{1}{2L}) \\ &\leq \textstyle\sum_{k=1}^K h_{x^k}(\hat{P}) + \frac{1}{2\eta}\|P_1 - \hat{P}\|_F^2 - \frac{1}{4L}\sum_{k=1}^K \frac{\|\nabla f(x^{k+1})\|^2}{\|\nabla f(x^k)\|^2}, && (49) \end{aligned}$$

where (49) applies the definition of h_x and that $\|\nabla f(x^{k+1})\| = \|\nabla f(x^{k+1/2} - \frac{1}{L}\nabla f(x^{k+1/2}))\| \leq \|\nabla f(x^{k+1/2})\|$:

$$\|\nabla h_{x^k}(P_k)\|_F^2 = \frac{\|\nabla f(x^k)\|^2\|\nabla f(x^{k+1/2})\|^2}{\|\nabla f(x^k)\|^4} = \frac{\|\nabla f(x^{k+1/2})\|^2}{\|\nabla f(x^k)\|^2} \geq \frac{\|\nabla f(x^{k+1})\|^2}{\|\nabla f(x^k)\|^2}.$$

Rearranging (49), we get

$$\textstyle\sum_{k=1}^K \frac{\|\nabla f(x^{k+1})\|^2}{\|\nabla f(x^k)\|^2} \leq 4L[\sum_{k=1}^K h_{x^k}(\hat{P}) - \sum_{k=1}^K h_k] + \frac{2L}{\eta}\|P_1 - \hat{P}\|_F^2. \tag{50}$$

Then we can bound the suboptimality by

$$\begin{aligned} f(x^{K+1}) - f^\star &\le \tfrac{1}{2\mu}\|\nabla f(x^{K+1})\|^2 \\ &\le \tfrac{1}{2\mu}\|\nabla f(x^1)\|^2 (\tfrac{1}{K}\textstyle\sum_{k=1}^K \frac{\|\nabla f(x^{k+1})\|^2}{\|\nabla f(x^k)\|^2})^K \qquad \text{(by AM-GM inequality)} \\ &\overset{(50)}{\le} \tfrac{1}{2\mu}\|\nabla f(x^1)\|^2 (\tfrac{4L}{K}[\textstyle\sum_{k=1}^K h_{x^k}(\hat{P}) - \sum_{k=1}^K h_k] + \tfrac{2L}{\eta K}\|P_1 - \hat{P}\|_F^2)^K. \end{aligned} \tag{51}$$

Finally, we do a case analysis on the sign of $\sum_{k=1}^K h_{x^k}(\hat{P}) - \sum_{k=1}^K h_k$.

Case 1. $\sum_{k=1}^K h_k \le \sum_{k=1}^K h_{x^k}(\hat{P})$: The progress of `Monotone Lookahead OSGM-H` is better than stepsize $\hat{P}$.

Case 2. $\sum_{k=1}^K h_k \ge \sum_{k=1}^K h_{x^k}(\hat{P})$: relation (51) can be further bounded by

$$f(x^{K+1}) - f(x^1) \le \tfrac{1}{2\mu}\|\nabla f(x^1)\|^2 (\tfrac{2L\|P_1 - \hat{P}\|_F^2}{\eta K})^K.$$

E Other instances of `OSGM`

This section presents additional instantiations of `OSGM`. We will invoke the following assumptions.

A1: f is L-smooth and convex.

A2: f is μ-strongly convex.

A3: $\mathcal{P}$ satisfies $0 \in \mathcal{P}$, $\frac{1}{L}I \in \mathcal{P}$ and $\operatorname{diam}(\mathcal{P}) \le D < \infty$.

A4: f has H-Lipschitz Hessian.

E.1 `Vanilla OSGM-R`

In this section, we assume f is L-smooth and μ-strongly convex (**A1**, **A2**) and instantiate `OSGM` with

$$\ell_x(P) \coloneqq r_x(P), \qquad \text{Vanilla landscape: } x^{k+1} = x^{k+1/2}, \qquad \mathcal{A} \coloneqq \texttt{OGD}.$$

The algorithm is called `Vanilla OSGM-R` (**Algorithm 4**).

Algorithm 4: `Vanilla OSGM-R`

input: $x^1, P_1 \in \mathcal{P}, \eta_k = \frac{c}{\sqrt{k}}$ or $\eta_k \equiv \frac{c}{\sqrt{K}}, c > 0$
for $k = 1, 2, \dots$ **do**
  $x^{k+1} = x^k - P_k \nabla f(x^k)$
  $P_{k+1} = \Pi_{\mathcal{P}}[P_k - \eta_k \nabla r_{x^k}(P_k)]$
end

The convergence analysis of `Vanilla OSGM-R` is similar to that for `Lookahead OSGM-R`. However, the convergence guarantees of `Vanilla OSGM-R` are weaker due to the vanilla landscape action.

Theorem E.1 (Global convergence)**.** *Under* **A1** *to* **A3**, *for any benchmark stepsize* $\hat{P} \in \mathcal{P}$, ***`Vanilla OSGM-R`*** *(**Algorithm 4**) with* $\eta_k \equiv \frac{D}{2L(LD+1)\sqrt{K}}$ *satisfies*

$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](\tfrac{1}{K}\textstyle\sum_{k=1}^K r_{x^k}(\hat{P}) + \frac{3LD(LD+1)}{\sqrt{K}})^K. \tag{52}$$

Moreover, ***`Vanilla OSGM-R`*** *with* $\eta_k = \frac{D}{2L(LD+1)\sqrt{k}}$ *satisfies* (52) *for all* $K \ge 1$.

Theorem E.1 suggests a divergence behavior of `Vanilla OSGM-R` in the earlier iterations. Indeed, when the landscape takes no action to filter out bad stepsizes, the algorithm will remain unstable until the scheduler learns a good stepsize.

Theorem E.2 (Local adaptivity)**.** *Under the same assumptions as* ***Theorem E.1****, for any benchmark sequence of stepsizes* $\{\hat{P}_k\}$*,* ***`Vanilla OSGM-R`*** *with* $\eta_k \equiv \frac{D}{2L(LD+1)\sqrt{K}}$ *satisfies*

$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](\tfrac{1}{K}\textstyle\sum_{k=1}^K r_{x^k}(\hat{P}_k) + \frac{3L(LD+1)(2D+\mathsf{PL}(\{\hat{P}_k\}))}{\sqrt{K}})^K \quad \text{for any } \hat{P}_k \in \mathcal{P}.$$

Theorem E.3 (Superlinear convergence)**.** *Instate* **A1** *to* **A4** *and suppose* $[\nabla^2 f(x^\star)]^{-1} \in \mathcal{P}$*,* ***`Vanilla OSGM-R`*** *with* $\eta_k = \frac{D}{2L(LD+1)\sqrt{k}}$ *satisfies*

$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](\tfrac{C_1}{K} + \tfrac{C_2}{\sqrt{K}})^K,$$

where $C_1 = \frac{H^2\kappa}{4\mu^2}\big[K_0(1 - \frac{1}{\kappa} + 3LD(LD+1))^{K_0} + 2\kappa\big], K_0 = \lceil 36\kappa^2[LD(LD+1)]^2\rceil$ *and* $C_2 = 3LD(LD+1)$.

E.2 `Monotone OSGM-H`

In this section, we assume f is L-smooth, optionally μ-strongly convex (**A1**, optionally **A2**), and instantiate `OSGM` with

$$\ell_x(P) \coloneqq h_x(P), \quad \text{Monotone landscape: } f(x^{k+1}) \le \min\{f(x^{k+1/2}), f(x^k)\}, \quad \mathcal{A} \coloneqq \texttt{OGD}.$$

The algorithm is called `Monotone OSGM-H` (**Algorithm 5**).

Algorithm 5: `Monotone OSGM-H`

input: $x^1, P_1 \in \mathcal{P}, \eta_k = \frac{c}{\sqrt{k}}$ or $\eta_k \equiv \frac{c}{\sqrt{K}}, c > 0$
for $k = 1, 2, \dots$ **do**
  $x^{k+1/2} = x^k - P_k \nabla f(x^k)$
  Choose x^{k+1} that satisfies $f(x^{k+1}) \le \min\{f(x^{k+1/2}), f(x^k)\}$
  $P_{k+1} = \Pi_{\mathcal{P}}[P_k - \eta_k \nabla r_{x^k}(P_k)]$
end

The convergence analysis of `Monotone OSGM-H` is similar to `Monotone Lookahead OSGM-H`.

Theorem E.4 (Global convergence)**.** *Under* **A1** *to* **A3***, for any benchmark stepsize* $\hat{P} \in \mathcal{P}$*,* ***`Monotone OSGM-H`*** *(**Algorithm 5**) with* $\eta_k \equiv \frac{D}{(LD+1)\sqrt{K}}$ *satisfies*

$$f(x^{K+1}) - f^\star \le \min\big\{\tfrac{\Delta^2}{K\max\{\frac{1}{K}\sum_{k=1}^K -h_{x^k}(\hat{P}) - \frac{3D(LD+1)}{\sqrt{K}}, 0\}}, f(x^1) - f^\star\big\} \qquad \text{(convex)}$$

$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](1 - 2\mu\max\{\tfrac{1}{K}\textstyle\sum_{k=1}^K -h_{x^k}(\hat{P}) - \frac{3D(LD+1)}{\sqrt{K}}, 0\}). \qquad (\mu\text{-strongly convex})$$

Moreover, ***`Monotone OSGM-H`*** *with* $\eta_k = \frac{D}{(LD+1)\sqrt{k}}$ *satisfies the same bounds for all* $K \ge 1$.

Theorem E.5 (Local adaptivity)**.** *Under the same assumptions as* ***Theorem E.4****, for any benchmark sequence of stepsizes* $\{\hat{P}_k\}$*,* ***`Monotone OSGM-H`*** *with* $\eta_k \equiv \frac{D}{(LD+1)\sqrt{K}}$ *satisfies*

$$f(x^{K+1}) - f^\star \le \min\big\{\tfrac{\Delta^2}{K\max\{\frac{1}{K}\sum_{k=1}^K -h_{x^k}(\hat{P}_k) - \rho_K(\{\hat{P}_k\}), 0\}}, f(x^1) - f^\star\big\}, \qquad \text{(convex)}$$

$$f(x^{K+1}) - f^\star \le [f(x^1) - f^\star](1 - 2\mu\max\{\tfrac{1}{K}\textstyle\sum_{k=1}^K -h_{x^k}(\hat{P}_k) - \rho_K(\{\hat{P}_k\}), 0\}), \qquad (\mu\text{-strongly convex})$$

where $\rho_K(\{\hat{P}_k\}) \coloneqq \frac{3D(LD+1)(2D+\mathsf{PL}(\{\hat{P}_k\}))}{\sqrt{K}}$.

Theorem E.6 (Superlinear convergence)**.** *Instate the same assumptions as* ***Theorem E.1*** *and suppose* $[\nabla^2 f(x^\star)]^{-1} \in \mathcal{P}$ *and* $\eta_k = \frac{D}{(LD+1)\sqrt{k}}$*.* `Monotone OSGM-H` *satisfies*

$$f(x^{K+1}) - f(x^k) \leq [f(x^1) - f(x^k)](\tfrac{C_1}{K} + \tfrac{C_2}{\sqrt{K}})^K,$$

where $C_1 = \frac{H^2\kappa^2}{4\mu^2}[K_0(1 - \frac{1}{\kappa} + 3D(LD+1))^{K_0} + 2\kappa], K_0 = \lceil 36\kappa^2[D(LD+1)]^2 \rceil$ *and* $C_2 = 3D(LD+1)$*.*

E.3 `Monotone OSGM-R` and `Monotone Lookahead OSGM-R`

`OSGM-R` does not require monotone landscape action to guarantee convergence, such as `Vanilla OSGM-R` and `Lookahead OSGM-R`, but it does not hurt to equip both variants with monotone landscape action. By **Lemma 4.1**, the per iteration progress of monotone variants of `OSGM-R` are bounded by

$$\begin{aligned}
\texttt{Monotone}: \ & r_k \leq \min\{r_{x^k}(P_k), 1\} \leq r_{x^k}(P_k); \\
\texttt{Monotone Lookahead}: \ & r_k \leq \min\{r_{x^k}(P_k) - \tfrac{1}{4L^2}\|\nabla r_{x^k}(P_k)\|_F^2, 1\} \leq r_{x^k}(P_k) - \tfrac{1}{4L^2}\|\nabla r_{x^k}(P_k)\|_F^2,
\end{aligned}$$

and the bounds on the right-hand side can be further bounded in the same way for `Vanilla OSGM-R` and `Lookahead OSGM-R`. In other words, the convergence analysis of monotone variants reduces to that of non-monotone variants, `Vanilla OSGM-R` and `Lookahead OSGM-R`, respectively.

E.4 Proof of results in Appendix E.1

E.4.1 Proof of Theorem E.1

We successively deduce that

$$\begin{aligned}
f(x^{K+1}) - f^\star &\leq (\tfrac{1}{K}\textstyle\sum_{k=1}^K r_k)^K && \text{(by } \textbf{Theorem 4.1}\text{)} \\
&= (\tfrac{1}{K}\textstyle\sum_{k=1}^K r_{x^k}(P_k))^K && \text{(by } \textbf{Lemma 4.1} \text{ + vanilla action)} \\
&\leq (\tfrac{1}{K}\textstyle\sum_{k=1}^K r_{x^k}(\hat{P}) + \tfrac{3LD(LD+1)}{\sqrt{K}})^K. && \text{(by } \textbf{Lemma 5.1}\text{)}
\end{aligned}$$

For stepsize $\eta_k = \mathcal{O}(\frac{1}{\sqrt{k}})$, the regret guarantee applies to any $K \geq 1$ and provides anytime convergence.

E.4.2 Proof of Theorem E.2

Combining $f(x^{K+1}) - f^\star \leq (\frac{1}{K}\sum_{k=1}^K r_{x^k}(P_k))^K$ with **Lemma 5.2** completes the proof.

E.4.3 Proof of Theorem E.3

Plugging $\hat{P} = \frac{1}{L}I$ into **Theorem E.1**, we have, for each $k = 1, \ldots, K$, that

$$f(x^{k+1}) - f^\star \leq [f(x^1) - f^\star](1 - \tfrac{1}{\kappa} + \tfrac{3LD(LD+1)}{\sqrt{k}})^k$$

and using strong convexity,

$$\|x^k - x^\star\|^2 \leq \tfrac{2}{\mu}[f(x^{k+1}) - f^\star] \leq \tfrac{2}{\mu}[f(x^1) - f^\star](1 - \tfrac{1}{\kappa} + \tfrac{3LD(LD+1)}{\sqrt{k}})^k.$$

and we bound $\sum_{k=1}^K \|x^k - x^\star\|^2$ using

$$\begin{aligned}\sum_{k=1}^K \|x^k - x^\star\|^2 &\leq \tfrac{2}{\mu}[f(x^1) - f^\star] \sum_{k=1}^K (1 - \tfrac{1}{\kappa} + \tfrac{3LD(LD+1)}{\sqrt{k}})^k \\ &\leq \tfrac{2}{\mu}[f(x^1) - f^\star] \sum_{k=1}^\infty (1 - \tfrac{1}{\kappa} + \tfrac{3LD(LD+1)}{\sqrt{k}})^k \\ &=: \tfrac{2}{\mu}[f(x^1) - f^\star] \sum_{k=1}^\infty e_k\end{aligned}$$

Let $K_0 = \lceil 36\kappa^2[LD(LD+1)]^2 \rceil$. We have $e_k \leq (1 - \frac{1}{2\kappa})^k$ for all $k \geq K_0$ and

$$\sum_{k=1}^\infty e_k = \sum_{k=1}^{K_0} e_k + \sum_{k=K_0+1}^\infty e_k \leq K_0(1 - \tfrac{1}{\kappa} + 3LD(LD+1))^{K_0} + 2\kappa$$

Using **Theorem E.1** and **Lemma 6.1**, we deduce that

$$\begin{aligned}f(x^{K+1}) - f^\star &\leq [f(x^1) - f^\star](\tfrac{1}{K}\sum_{k=1}^K r_{x^k}(\hat{P}) + \tfrac{3LD(LD+1)}{\sqrt{K}})^K \\ &\leq [f(x^1) - f^\star](\tfrac{H^2\kappa}{4\mu^2}\tfrac{1}{K}\sum_{k=1}^K \|x^k - x^\star\|^2 + \tfrac{3LD(LD+1)}{\sqrt{K}})^K \\ &\leq [f(x^1) - f^\star](\tfrac{H^2\kappa}{4\mu^2 K}[K_0(1 - \tfrac{1}{\kappa} + 3LD(LD+1))^{K_0} + 2\kappa] + \tfrac{3LD(LD+1)}{\sqrt{K}})^K \\ &= [f(x^1) - f^\star](\tfrac{C_1}{K} + \tfrac{C_2}{\sqrt{K}})^K\end{aligned}$$

and this completes the proof.

E.5 Proof of results in Appendix E.2

E.5.1 Proof of Theorem E.4

Similar to **Theorem E.1**, chaining **Theorem 4.2**, **Lemma 4.1** and **Lemma 5.1** completes the proof.

E.5.2 Proof of Theorem E.5

Similar to **Theorem E.2**, chaining **Theorem 4.2**, **Lemma 4.1** and **Lemma 5.2** completes the proof.

E.5.3 Proof of Theorem E.6

Plugging $\hat{P} = \frac{1}{L} I$ into **Theorem E.4**, we have, for each $k = 1, \ldots, K$, that

$$f(x^{k+1}) - f^\star \leq [f(x^1) - f^\star](1 - \tfrac{1}{\kappa} + \tfrac{3D(LD+1)}{\sqrt{k}})^k.$$

Using the same argument as **Theorem E.3**, we have

$$\|x^k - x^\star\|^2 \leq \tfrac{2}{\mu}[f(x^{k+1}) - f^\star] \leq \tfrac{2}{\mu}[f(x^1) - f^\star](1 - \tfrac{1}{\kappa} + \tfrac{3D(LD+1)}{\sqrt{k}})^k.$$

and we bound $\sum_{k=1}^K \|x^k - x^\star\|^2$ using

$$\begin{aligned}\sum_{k=1}^K \|x^k - x^\star\|^2 &\leq \tfrac{2}{\mu}[f(x^1) - f^\star] \sum_{k=1}^K (1 - \tfrac{1}{\kappa} + \tfrac{3D(LD+1)}{\sqrt{k}})^k \\ &\leq \tfrac{2}{\mu}[f(x^1) - f^\star] \sum_{k=1}^\infty (1 - \tfrac{1}{\kappa} + \tfrac{3D(LD+1)}{\sqrt{k}})^k \\ &=: \tfrac{2}{\mu}[f(x^1) - f^\star] \sum_{k=1}^\infty e_k\end{aligned}$$

Let $K_0 = \lceil 36\kappa^2[D(LD+1)]^2 \rceil$. We have $e_k \leq (1 - \frac{1}{2\kappa})^k$ for all $k \geq K_0$ and

$$\sum_{k=1}^\infty e_k = \sum_{k=1}^{K_0} e_k + \sum_{k=K_0+1}^\infty e_k \leq K_0(1 - \tfrac{1}{\kappa} + 3D(LD+1))^{K_0} + 2\kappa$$

Similar to the analysis of **Theorem 6.9**, using $\sum_{k=1}^{K} h_k \leq \sum_{k=1}^{K} h_{x^k}([\nabla^2 f(x^\star)]^{-1}) + \frac{3D(LD+1)}{\sqrt{K}}$, we further deduce that

$$f(x^{K+1}) - f(x^k) \leq [f(x^1) - f^\star](\tfrac{2L}{K} \sum_{k=1}^{K} [h_k + \tfrac{f(x^k)-f^\star}{\|\nabla f(x^k)\|^2}])^K$$

and that

$$\begin{aligned} \sum_{k=1}^{K} [h_k + \tfrac{f(x^k)-f^\star}{\|\nabla f(x^k)\|^2}] &\leq \sum_{k=1}^{K} [h_{x^k}([\nabla^2 f(x^\star)]^{-1}) + \tfrac{f(x^k)-f^\star}{\|\nabla f(x^k)\|^2}] + 3D(LD+1)\sqrt{K} \\ &\leq \tfrac{H^2\kappa}{8\mu^3} \sum_{k=1}^{K} \|x^k - x^\star\|^2 + 3D(LD+1)\sqrt{K}. \end{aligned}$$

Plugging the bounds back completes the proof.