DECENTRALIZED ONLINE EXP-CONCAVE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Decentralized Online Convex Optimization (D-OCO) is a fundamental framework in which a set of local agents aims to minimize a sequence of global loss functions using only local computations and communications. While recent work has established near-optimal regret guarantees for convex and strongly convex losses, the case of exp-concave losses remains an open challenge: specifically, whether logarithmic regret is achievable. In this paper, we answer this question affirmatively by proposing Decentralized Online Newton Step (D-ONS), which adapts the idea of blocking and accelerated gossip to the Online Newton Step algorithm in a non-trivial manner. Concretely, our algorithm achieves a regret bound of $\mathcal{O}\left(Nd(1-\sigma_2)^{-1/2}\log^2(T)\right)$ where N is the number of agents, d is the dimension, T is the horizon, and $1-\sigma_2$ represents the spectral gap of the network. To our knowledge, this is the first logarithmic regret bound for decentralized online exp-concave optimization. We further establish a nearly matching lower bound of $\Omega\left(Nd(1-\sigma_2)^{-1/2}\log(T/N)\right)$ showing the near optimality of our upper bound. Finally, we extend our algorithm to decentralized online linear regression in which the feasible domain is unbounded, and achieve logarithmic regret.

1 INTRODUCTION

Decentralized online convex optimization (D-OCO) (Yan et al., 2013; Hosseini et al., 2013; Wan et al., 2024) provides a powerful framework for sequential decision making in distributed systems, where multiple agents *collaboratively* optimize a global objective while only observing *local* information and communicating through a network. In this setting, agents repeatedly choose decisions, incur global losses but only observe local feedback, and exchange information with neighboring agents in order to compete with the best fixed decision for the global losses in hindsight. This paradigm naturally captures a variety of applications including multi-robot systems (Shorinwa et al., 2024), sensor networks (Li et al., 2002), online packet routing (Awerbuch & Kleinberg, 2004), where centralized coordination is either infeasible due to communication constraints or undesirable due to privacy and robustness considerations.

For general convex losses, early studies by Yan et al. (2013); Hosseini et al. (2013) established decentralized online algorithms with regret bounds $\mathcal{O}(N^{5/4}(1-\sigma_2)^{-1/2}\sqrt{T})$ and $\mathcal{O}(N^{3/2}(1-\sigma_2)^{-1}\log(T))$ for convex and strongly convex losses, respectively. Here, N denotes the number of agents, T is the horizon, and $1-\sigma_2$ represents the spectral gap of the communication matrix. More recently, Wan et al. (2024; 2025) tightened these guarantees using accelerated gossip and block-based scheduling, achieving $\mathcal{O}(N(1-\sigma_2)^{-1/4}\sqrt{T\log N})$ regret for convex losses and $\mathcal{O}(N(1-\sigma_2)^{-1/2}\log(N)\log T)$ for strongly convex losses, alongside nearly matching lower bounds.

Despite these advances, the landscape for decentralized OCO with exp-concave losses remains largely unexplored. This class of functions captures a wide range of applications, including online portfolio selection, logistic regression, and online linear regression on bounded domains. Similar to the strongly convex case, exp-concavity allows for significantly faster convergence rates in centralized settings than general convexity. Specifically, the seminal Online Newton Step (ONS) algorithm (Hazan et al., 2007) achieves $\mathcal{O}(d\log T)$ regret for d-dimensional exp-concave losses. However, the possibility of achieving such fast rates with the decentralized communication remains an open challenge. This leads to the following central question of this work:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Can decentralized agents exploit exp-concavity to achieve logarithmic regret while communicating only through local gossip?

Our contributions. In this paper, we answer the above question affirmatively. More concretely:

- We propose Decentralized Online Newton Step (D-ONS), a decentralized variant of Online Newton Step for exp-concave losses, achieving a regret bound of $\mathcal{O}(Nd(1-\sigma_2)^{-1/2}\log^2(T))$. To the best of our knowledge, this is the first algorithm to achieve logarithmic regret for exp-concave losses under D-OCO, answering an open question proposed in Wan (2025). Technically, we combine block-based accelerated gossip with delayed ONS while carefully preserving the gradient-dependent curvature of exp-concave losses under decentralized communication.
- We further complement our upper bound with a nearly matching lower bound, demonstrating that the polynomial dependence on the number of agents, the dimension, and the network spectral gap is essentially unavoidable. In particular, we adapt the hard instance of online exp-concave optimization from Ordentlich & Cover (1998) to the decentralized setting. This establishes that no algorithm can achieve a regret better than $\Omega\left(Nd\log(T/N)(1-\sigma_2)^{-1/2}\right)$, showing that our upper bound is near-optimal up to logarithmic factors.
- Finally, we study decentralized online linear regression, where the decision domain for each agent is *unbounded* and thus not captured by standard exp-concave analysis. Specifically, we propose a decentralized online ridge regression algorithm with clipped predictions that achieves a regret bound of $\mathcal{O}(Nd(1-\sigma_2)^{-1/2}\log^2(T))$, demonstrating that logarithmic regret remains attainable even in the absence of a compact feasible set in online linear regression.

1.1 Related Works

Decentralized online convex optimization. Foundational work in decentralized optimization was developed primarily in offline settings, where a standard gossip step Xiao & Boyd (2004) was used to enable decentralized consensus and optimization (Nedic & Ozdaglar, 2009; Duchi et al., 2011). The online counterpart was first systematically studied by Yan et al. (2013), who proposed a Decentralized Online Gradient Descent and established sublinear regret guarantees. In particular, their algorithm achieves a regret bound of $\mathcal{O}(N^{5/4}(1-\sigma_2)^{-1/2}\sqrt{T})$. More recently, Wan et al. (2024; 2025) leverage an accelerated gossip strategy (Liu & Morse, 2011; Ye et al., 2023) together with a blocking update mechanism to develop new D-OCO algorithms that achieve regret bounds of $\mathcal{O}(N(1-\sigma_2)^{-1/4}\sqrt{T\log N})$ for convex functions and $\mathcal{O}(N\log N(1-\sigma_2)^{-1/2}\log T)$ for strongly convex functions. They further establish lower bounds of $\Omega(N(1-\sigma_2)^{-1/4}\sqrt{T})$ and $\Omega(N(1-\sigma_2)^{-1/2}\log T)$ for the convex and strongly convex settings, respectively, showing that their upper bounds are optimal up to logarithmic factors.

Online learning with curved losses. Curvature assumptions on the loss function are well known to yield faster regret rates in OCO. For strongly convex losses with G-bounded gradients, Online Gradient Descent achieves $\mathcal{O}(G^2\log T)$ regret, while for d-dimensional exp-concave losses, the Online Newton Step of Hazan et al. (2007) attains $\mathcal{O}(d\log T)$ regret. Closely related second-order methods for online linear regression include the Vovk–Azoury–Warmuth forecaster (Vovk, 2001) and online ridge regression (Hoerl & Kennard, 1970). Beyond these foundations, research has expanded into several advanced perspectives, including scale-free guarantees (Mayo et al., 2022), non-stationarity (Zhang et al., 2025), adaptive guarantees (Bartlett et al., 2007; Luo et al., 2022), and robustness to delayed feedback (Qiu et al., 2025). In contrast to these established paradigms, our work investigates exp-concave online learning in a decentralized setting, where multiple agents must collaboratively minimize the global objective over a communication network.

2 Preliminaries

For any positive integer m, let $[m] \triangleq \{1, 2, \ldots, m\}$, and let $\mathbf{1}$ denote the all-ones vector of an appropriate dimension. For a matrix $M \in \mathbb{R}^{m_1\times m_2}$, we denote its (i,j)-th entry by M_{ij}. We study decentralized online convex optimization (D-OCO). The framework consists of a group of N agents interacting with a dynamic environment and communicating over a network modeled by a connected, undirected graph $\mathcal{G} \triangleq (V, E)$. The vertex set $V = [N]$ represents the agents, while

the edge set E specifies the available communication links. For each agent $i \in V$, we denote its neighborhood by $\mathcal{N}_i \triangleq \{j \in V : (i,j) \in E\} \cup \{i\}$. Throughout the paper, we use V and $[N]$ interchangeably. Each agent $i \in [N]$ encounters a sequence of local loss functions $\{f_{t,i}\}_{t=1}^T$, where each $f_{t,i} : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ is chosen arbitrarily by an adversary and is unknown to the agents. Below, we list several standard assumptions (Yan et al., 2013; Hosseini et al., 2013; Wan et al., 2024) and will explicitly specify which are required for each of our results.

Assumption 1. *For every $t \in [T]$ and $i \in [N]$, the local loss $f_{t,i} : \mathcal{X} \to \mathbb{R}$ is differentiable and convex, and satisfies $\|\nabla f_{t,i}(\mathbf{x})\|_2 \le G, \forall \mathbf{x} \in \mathcal{X}$.*

Assumption 2. *The feasible set $\mathcal{X} \subseteq \mathbb{R}^d$ is compact and convex, and contains the origin ($\mathbf{0} \in \mathcal{X}$). Furthermore, its diameter is bounded by D, i.e., $\|\mathbf{x} - \mathbf{y}\|_2 \le D, \forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X}$.*

In this paper, we focus on *exp-concave losses* Hazan et al. (2007), a broader class than strongly convex functions that also enjoys faster rates in OCOz.

Assumption 3. *For each round $t \in [T]$, the loss function $f_{t,i} : \mathcal{X} \mapsto \mathbb{R}$ of each agent $i \in [N]$ is α-exp-concave over $\mathcal{X}$. That is, the mapping $\mathbf{x} \mapsto \exp(-\alpha f_{t,i}(\mathbf{x}))$ is concave for all $\mathbf{x} \in \mathcal{X}$.*

The learning protocol of D-OCO involves an interaction between the agents and the environments over T rounds. Without loss of generality, we assume that $T \ge N$. At each round $t \in [T]$, each agent $i \in [N]$ chooses a decision $\mathbf{x}_i(t)$ from the feasible set $\mathcal{X}$. Simultaneously, the adversary selects the local loss functions $f_{t,i}$ for all $i \in [N]$. Each agent i then observes only the gradient of her *local loss* $\nabla f_{t,i}(\mathbf{x}_i(t))$. Subsequently, agents are permitted to exchange information with their neighbors in the graph $\mathcal{G}$ to update their internal states for the next round. The objective of each agent i is then to minimize the regret in terms of *the global function*, defined as:

$$\mathrm{Reg}_{T,i} \triangleq \max_{\mathbf{x} \in \mathcal{X}} \mathrm{Reg}_{T,i}(\mathbf{x}) \triangleq \max_{\mathbf{x} \in \mathcal{X}} \left(\sum_{t=1}^{T} \sum_{j=1}^{N} f_{t,j}(\mathbf{x}_i(t)) - \sum_{t=1}^{T} \sum_{j=1}^{N} f_{t,j}(\mathbf{x}) \right). \tag{1}$$

It remains to specify how agents communicate over the network $\mathcal{G}$. Following Hosseini et al. (2013); Yan et al. (2013); Wan et al. (2024), we model information exchange through a gossip protocol induced by a communication matrix $P \in \mathbb{R}^{N \times N}$. Intuitively, the entry P_{ij} specifies the weight that agent i assigns to information received from agent j, so repeated multiplication by P implements local averaging over the network. We formalize the required properties of P below.

Definition 1. *A communication matrix $P \in \mathbb{R}^{N \times N}$ associated with $\mathcal{G} = ([N], E)$ is symmetric, positive semidefinite, and doubly stochastic, with $P_{ij} = 0$ whenever $(i,j) \notin E$ and $i \neq j$. We assume its second-largest singular value satisfies $\sigma_2(P) < 1$, and write $\sigma_2 = \sigma_2(P)$ when clear.*

Given a collection of local vectors $\{\mathbf{u}_i\}_{i=1}^N$ (for example, local gradients or iterates), the goal of gossip is to approximately recover the network average $\overline{\mathbf{u}} = \frac{1}{N}\sum_{i=1}^N \mathbf{u}_i$ for every agent using only neighbor-to-neighbor communication. To this end, agents run L rounds of gossip. In each round k, agent i updates its local state according to

$$\mathbf{u}_i^{k+1} = (1+\theta) \sum_{j \in \mathcal{N}_i} P_{ij} \mathbf{u}_j^k - \theta \mathbf{u}_i^{k-1}, \tag{2}$$

for $k \ge 0$, where $\mathbf{u}_i^0 = \mathbf{u}_i^{-1} = \mathbf{u}_i$. When $\theta = 0$, this reduces to the standard gossip iteration (Xiao & Boyd, 2004); when $\theta > 0$, it becomes the accelerated gossip scheme of Liu & Morse (2011), which improves the convergence rate to $\overline{\mathbf{u}}$.

The following lemma extended from Ye et al. (2023) shows that accelerated gossip allows each local vector to approximate the global average with an error of $\mathcal{O}((TN)^{-2})$ after $L = \Theta(\log(NT)(1-\sigma_2)^{-1/2})$ by picking θ appropriately. For completeness, we provide the proof in Appendix A.

Lemma 1. *Let P be the communication matrix defined in Definition 1. Let*

$$\theta \triangleq \frac{1}{1 + \sqrt{1 - \sigma_2^2(P)}} \quad \text{and} \quad L \triangleq \left\lceil \frac{\sqrt{2}\log(T^2\sqrt{14}N^{5/2})}{(\sqrt{2}-1)\sqrt{1-\sigma_2(P)}} \right\rceil. \tag{3}$$

Consider arbitrary vectors $\mathbf{u}_1, \ldots, \mathbf{u}_N \in \mathbb{R}^d$ satisfying $\|\mathbf{u}_i\|_2 \le R, i \in [N]$. Let $U^0 = U^{-1} = [\mathbf{u}_1, \ldots, \mathbf{u}_N]^\top, \overline{\mathbf{u}} = \frac{1}{N}\sum_{i=1}^N \mathbf{u}_i^0, \overline{U} = \mathbf{1}\overline{\mathbf{u}}^\top$. Suppose accelerated gossip is run for L iterations with update $U^{k+1} = (1+\theta)PU^k - \theta U^{k-1}$ for $k \ge 0$. Then, for every agent $i \in [N]$, $\|\mathbf{u}_i^L - \overline{\mathbf{u}}\|_2 \le \frac{2R}{T^2N^2}$.

Other notations. Define $\Delta_d \triangleq \{x \in \mathbb{R}^d : \sum_{j=1}^d x_j = 1, x_j \geq 0, \forall j \in [d]\}$ to be the $(d-1)$-dimensional simplex. I_d denotes the $d \times d$ identity matrix. Let $A_{\mathcal{G}}$ be the adjacency matrix of $\mathcal{G}$. Let $D_{\mathcal{G}} = \mathrm{diag}(d_1, \ldots, d_N)$ be the diagonal degree matrix, and let $d_{\max} = \max_{i\in[N]} d_i$ denote the maximum degree of $\mathcal{G}$. The graph Laplacian of $\mathcal{G}$ is defined as $L_{\mathcal{G}} \triangleq D_{\mathcal{G}} - A_{\mathcal{G}}$. For a positive semidefinite matrix $M \in \mathbb{R}^{d\times d}$ and $x \in \mathbb{R}^d$, we denote $\|x\|_M = \sqrt{x^\top M x}$ to be the Mahalanobis norm induced by M and, if M is positive definite, let $\|x\|_{M^{-1}} = \sqrt{x^\top M^{-1} x}$ be the dual norm.

3 Decentralized Online Exp-concave Optimization

In this section, we introduce our algorithm for decentralized online exp-concave optimization. We first provide an algorithm that achieves a logarithmic regret upper bound (Section 3.1), followed by a nearly-matching lower bound (Section 3.3) to show the optimality of our algorithm.

3.1 Upper Bound

In this subsection, we present our algorithm for decentralized online exp-concave optimization. Our algorithm builds upon the block-based communication schedule and accelerated gossip mechanism developed in recent decentralized OCO reductions for the convex and strongly convex settings (Wan et al., 2024; Wan, 2025). However, the exp-concave case cannot be handled by a direct black-box reduction, as unlike the strongly convex setting where the curvature is characterized by a fixed global parameter, the curvature information in exp-concave optimization is decision-dependent and may incur larger consensus errors during communication. To address this challenge, we propose Decentralized Online Newton Step (D-ONS), which gossips both block gradients and decisions before performing a delayed ONS update.

Algorithm 1: Decentralized Online Newton Step for each agent $i \in [N]$

Input: gossip parameter θ, epoch length L, regularization parameter β, learning rate $\eta > 0$, communication matrix P
Initialization: set $\mathbf{x}_i(1) = \hat{\mathbf{x}}_i(1) = \mathbf{0}$.
for $z = 1, \ldots, T/L$ **do**
  Compute $\hat{\mathbf{x}}_i(z+1) = \arg\min_{\mathbf{x}\in\mathcal{X}} \sum_{s=1}^{z-2} (\langle \mathbf{g}_i(s), \mathbf{x}\rangle + \frac{\beta}{2L} \langle \mathbf{g}_i(s), \mathbf{x}_i(s) - \mathbf{x}\rangle^2) + \frac{\eta}{2}\|\mathbf{x}\|_2^2$.
  Set $k_x = 0$ and $k_g = 0$ and $\mathbf{x}_i^0(z+1) = \mathbf{x}_i^{-1}(z+1) = \hat{\mathbf{x}}_i(z+1)$
  for $t = (z-1)L + 1, \ldots, zL$ **do**
    Play $\mathbf{x}_i(z)$ and query $\nabla f_{t,i}(\mathbf{x}_i(z))$
    if $z \geq 2$ **then** update $\mathbf{g}_i^{k_g+1}(z-1) = (1+\theta)\sum_{j\in\mathcal{N}_i} P_{ij}\mathbf{g}_j^{k_g}(z-1) - \theta \mathbf{g}_i^{k_g-1}(z-1)$;
    Update $\mathbf{x}_i^{k_x+1}(z+1) = (1+\theta)\sum_{j\in\mathcal{N}_i} P_{ij}\mathbf{x}_j^{k_x}(z+1) - \theta \mathbf{x}_i^{k_x-1}(z+1)$
    Set $k_x \leftarrow k_x + 1$ and $k_g \leftarrow k_g + 1$
  Update $\mathbf{x}_i(z+1) = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}_i^L(z+1) - \mathbf{x}\|_2$
  if $z \geq 2$ **then** set $\mathbf{g}_i(z-1) = \mathbf{g}_i^L(z-1)$;
  Set $\mathbf{g}_i^0(z) = \mathbf{g}_i^{-1}(z) = \sum_{\tau\in\mathcal{T}_z} \nabla f_{\tau,i}(\mathbf{x}_i(z))$

More specifically, the algorithm partitions the time horizon T into T/L blocks indexed by z, where $\mathcal{T}_z \triangleq \{(z-1)L + 1, \ldots, zL\}$ denotes the set of time steps in block z.[1] Before introducing the algorithm, for notational convenience, we define $\overline{\mathbf{g}}(z)$ and $\overline{\mathbf{x}}(z)$ as follows:

$$\overline{\mathbf{g}}(z) \triangleq \frac{1}{N}\sum_{t\in\mathcal{T}_z}\sum_{j=1}^{N} \nabla f_{t,j}\left(\mathbf{x}_j(z)\right) \quad \text{and} \quad \overline{\mathbf{x}}(z) \triangleq \frac{1}{N}\sum_{j=1}^{N} \hat{\mathbf{x}}_j(z). \tag{4}$$

Specifically, $\overline{\mathbf{g}}(z)$ denotes the network average of the cumulative gradients collected during block z, while $\overline{\mathbf{x}}(z)$ denotes the average of the preparatory decisions. Algorithm 1 proceeds over these blocks as follows.

[1] For notational convenience, we assume that T/L is an integer.

- **Step 1: Local Preparatory Update.** Before making a decision for the current block z, each agent i computes a local preparatory decision $\hat{\mathbf{x}}_i(z+1)$. This is obtained via a delayed ONS update (Line 4) using the gossiped aggregated gradient $\mathbf{g}_i(s)$ and decision information $\mathbf{x}_i(s)$ collected from blocks $s = 1$ to $z-2$.
- **Step 2: Decision Execution and Parallel Gossip.** Each agent i plays the fixed decision $\mathbf{x}_i(z) \in \mathcal{X}$ for all rounds $t \in \mathcal{T}_z$ and observes the local gradients $\nabla f_{t,i}(\mathbf{x}_i(z))$ (Line 7). Simultaneously, two accelerated gossip processes are executed in parallel:
  - *Gradient Gossip:* Agents run accelerated gossip on the gradients collected during the previous block $z-1$ (Line 8). According to Lemma 1, by the end of block z, the local state $\mathbf{g}_i^L(z-1)$ will approximate the network average of the cumulative gradients $\overline{\mathbf{g}}(z-1)$ within a poly$(1/T)$ error. This will serve as the feedback for future ONS updates.
  - *Decision Gossip:* Agents also run accelerated gossip on their local preparatory decisions $\hat{\mathbf{x}}_i(z+1)$. While one might intuitively use $\hat{\mathbf{x}}_i(z+1)$ directly as the decision, these local ONS outputs may *diverge* significantly across agents, complicating the regret analysis. To ensure consensus on the decision over all the agents, we also gossip these vectors so that $\mathbf{x}_i^L(z+1)$ converges toward the network average $\overline{\mathbf{x}}(z+1) = \frac{1}{N}\sum_j \hat{\mathbf{x}}_j(z+1)$ up to a poly$(1/T)$ consensus error.
- **Step 3: Projection and Finalization.** At the end of block z, the output of the decision-gossip process $\mathbf{x}_i^L(z+1)$ is projected back onto the feasible set $\mathcal{X}$. This projection is necessary because the linear combinations involved in accelerated gossip may produce iterates that lie outside the convex constraints.[2] The resulting projected vector is then used as the played decision $\mathbf{x}_i(z+1)$ for the next block. This ensures that the decisions made by each agent at each epoch are close to each other. For $z \geq 2$, the refined gradient estimate $\mathbf{g}_i(z-1)$ is fed into the local ONS subroutine to guide the next iteration.

Consequently, our algorithm can be viewed as operating with a constant *2-block delay*. Specifically, after agents collect their local cumulative gradients in block z, they perform a gossip process during block $z+1$ to approximate the network average of the cumulative gradients. Following the ONS update, a second accelerated gossip process is executed in block $z+2$ to synchronize the decisions for use in block $z+3$. Although this delayed information flow is adapted from the blocking framework of Wan (2025), the analysis must account for the gradient-dependent curvature of exp-concave losses and the distinct preparatory and executed decisions.

Formally, we state our regret guarantee for decentralized online exp-concave optimization:

Theorem 1. *Suppose that the sequence of loss functions $\{f_{t,i}\}_{t\in[T],i\in[N]}$ satisfy Assumption 1, 2, and 3. Let P be the communication matrix defined in Definition 1. If every agent $i \in [N]$ runs Algorithm 1 with $\beta = \frac{1}{2}\min\{\frac{1}{4GD}, \alpha\}$ and with parameters θ and L chosen in Eq. (3), then, for any agent $i \in [N]$, the regret is bounded by:*

$$\mathrm{Reg}_{T,i}(\mathbf{x}) = \mathcal{O}\left(\eta N\|\mathbf{x}\|_2^2 + \frac{Nd(1+\beta GD)L}{\beta}\log\left(1+\frac{\beta G^2 T}{\eta d}\right)\right). \tag{5}$$

By substituting the choice of L from Eq. (3) into Eq. (5) and picking η, we obtain a total regret of $\mathcal{O}\left(Nd\log^2 T(1-\sigma_2(P))^{-1/2}\right)$. To our knowledge, this is the first logarithmic regret guarantee for decentralized online exp-concave optimization. We also establish a lower bound of $\Omega\left(Nd\log(T/N)(1-\sigma_2(P))^{-1/2}\right)$ in Section 3.3, showing the near-optimality of our upper bound with respect to T, the number of agents N, the dimension d, and the network spectral gap $1-\sigma_2(P)$.

3.2 Analysis of Algorithm 1

In this subsection, we provide a proof sketch of Theorem 1. The complete proof is deferred to Appendix B. The analysis consists of two main steps. First, we provide a decentralized-to-block-level reduction, showing that the regret can be controlled by a sum of local ONS-type surrogate terms up to lower-order gossip errors. For any comparator $\mathbf{x} \in \mathcal{X}$, utilizing the Lipschitz property, the property of exp-concavity, and the consensus guarantee in Lemma 1, we obtain that

$$f_{t,j}\big(\mathbf{x}_i(z)\big) - f_{t,j}(\mathbf{x}) \leq \big\langle \nabla f_{t,j}\big(\mathbf{x}_j(z)\big), \overline{\mathbf{x}}(z) - \mathbf{x}\big\rangle - \frac{\beta}{2}\big\langle \nabla f_{t,j}\big(\mathbf{x}_j(z)\big), \mathbf{x}_j(z) - \mathbf{x}\big\rangle^2$$

[2] We point out that the algorithm in Wan (2025) in fact also requires this projection to make sure that $\mathbf{x}_i(z)$ lies in the feasible set.

$$+ \mathcal{O}\big(G\|\overline{\mathbf{x}}(z) - \mathbf{x}_i(z)\| + \|\overline{\mathbf{x}}(z) - \mathbf{x}_j(z)\|\big). \tag{6}$$

Specifically, the last term of Eq. (6) can be bounded by $\mathcal{O}(GD/(TN)^2)$ since $\mathbf{x}_i(z)$ and $\mathbf{x}_j(z)$ are both projected approximations of the network-average decision $\overline{\mathbf{x}}(z)$, thus Lemma 1 implies that their deviations from $\overline{\mathbf{x}}(z)$ are of order $\mathcal{O}(D/(T^2N^2))$. Therefore, the last term in Eq. (6) contributes only a lower-order term after summing over all rounds and agents. For the first two terms, after taking a summation over all $j \in [N]$ and $t \in \mathcal{T}_z$ and applying Jensen's inequality, with some calculation, we know that they are bounded by

$$N\,\langle \overline{\mathbf{g}}(z), \overline{\mathbf{x}}(z) - \mathbf{x}\rangle - \frac{\beta}{2NL}\Big(\sum_{t\in\mathcal{T}_z}\sum_{j=1}^{N}\langle \nabla f_{t,j}(\mathbf{x}_j(z)), \mathbf{x}_j(z) - \mathbf{x}\rangle\Big)^2. \tag{7}$$

By definition of $\overline{\mathbf{x}}(z)$, the first term of Eq. (7) can be written as $\sum_{j=1}^{N}\langle \overline{\mathbf{g}}(z), \widehat{\mathbf{x}}_j(z) - \mathbf{x}\rangle$, which can be further bounded by $\sum_{j=1}^{N}\langle \mathbf{g}_j(z), \widehat{\mathbf{x}}_j(z) - \mathbf{x}\rangle$ plus an $\mathcal{O}(1/T)$ approximation error due to Lemma 1. Via some calculation, the second term of Eq. (7) can be bounded by $-\frac{\beta}{2L}\sum_{j=1}^{N}\langle \mathbf{g}_j(z), \mathbf{x}_j(z) - \mathbf{x}\rangle^2 + \mathcal{O}(\sum_{j=1}^{N}(\|\mathbf{x}_j(z) - \overline{\mathbf{x}}(z)\|_2 + \|\mathbf{g}_j(z) - \overline{\mathbf{g}}(z)\|_2))$, where the second term is bounded by $\mathcal{O}(1/T)$ due to Lemma 1 again. Combining all of the above, we reduce bounding $\mathrm{Reg}_{T,i}(\mathbf{x})$ to bounding

$$\sum_{z=1}^{T/L}\sum_{j=1}^{N}\left(\langle \mathbf{g}_j(z), \widehat{\mathbf{x}}_j(z) - \mathbf{x}\rangle - \frac{\beta}{2L}\langle \mathbf{g}_j(z), \mathbf{x}_j(z) - \mathbf{x}\rangle^2\right) \tag{8}$$

plus $\mathcal{O}(1)$ term induced by the gossip process. The second step is to bound Eq. (8). A natural approach would be to interpret each summand as a surrogate loss and then apply a standard FTRL or ONS regret bound to the preparatory decision $\widehat{\mathbf{x}}_j(z)$. However, unlike the strongly convex setting considered in Wan (2025), such a direct argument does not apply here. The difficulty is that the linear term is evaluated at the preparatory decision $\widehat{\mathbf{x}}_j(z)$, whereas the negative quadratic term depends on the executed decision $\mathbf{x}_j(z)$. This mismatch prevents the summand in Eq. (8) from being treated as the value of a standard ONS surrogate loss at a single decision point.

To overcome this issue, we analyze the delayed FTRL dynamics for $\widehat{\mathbf{x}}_j(z)$ more carefully. The key observation is that although $\widehat{\mathbf{x}}_j(z)$ is computed using only feedback up to block $z-3$, it remains close, in the adaptive second-order geometry induced by the accumulated gradients, to the "current" regularized leader that has access to feedback up to block z. This closeness allows us to control the price of the two-block delay by a logarithmic term using ellipsoid potential arguments. The following lemma formalizes this delayed ONS bound.

Lemma 2. *Under the same conditions as Algorithm 1. For any $j \in [N]$ and every comparator $\mathbf{x} \in \mathcal{X}$,*

$$\sum_{z=1}^{T/L}\big(\langle \mathbf{g}_j(z), \hat{\mathbf{x}}_j(z) - \mathbf{x}\rangle - \frac{\beta}{2L}\langle \mathbf{g}_j(z), \mathbf{x}_j(z) - \mathbf{x}\rangle^2\big) \le \mathcal{O}\Big(\eta\|\mathbf{x}\|_2^2 + \frac{d(1+\beta GD)L\log\big(1+\frac{\beta G^2 T}{\eta d}\big)}{\beta}\Big).$$

Its proof uses the same local-norm stability and elliptical-potential tools as Qiu et al. (2025, Theorem 4.1); in particular, their Lemma A.2 gives the intermediate stability comparison. Here, $\mathbf{g}_j(z)$ aggregates gradients evaluated at different agents' executed decisions, while the linear and quadratic terms involve $\widehat{\mathbf{x}}_j(z)$ and $\mathbf{x}_j(z)$, respectively. We use the lemma to examine this two-decision structure, rather than highlight it as a separate novel contribution. The self-contained derivation is retained in Appendix B. Applying Lemma 2 to each $j \in [N]$ and combining it with the regret decomposition above yields the claimed bound in Theorem 1. This completes the proof sketch.

3.3 Lower Bound

In this section, we provide the lower bound of decentralized online exp-concave optimization. Specifically, we show that, even when $\alpha GD = \Theta(1)$, no algorithm can achieve a regret better than $\Omega\left(Nd\alpha^{-1}(1-\sigma_2)^{-1/2}\log(T/(Nd))\right)$ in decentralized exp-concave online optimization. The regime $\alpha GD = \Theta(1)$ is the one in which Theorem 1 reads $\mathcal{O}(Nd\alpha^{-1}L\log T)$, so the two bounds can be compared directly.

Theorem 2. *Let $N = 4K + 2$ for some positive integer K, let $d \geq 1$, let $\alpha > 0$, fix any $B > 0$, and suppose that $T \geq (\pi e)^2 N d^3$. Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq B\}$, a compact convex set containing the origin with diameter $D = 2B$, and let $\mathcal{G} = ([N], E)$ be the cycle graph with $P = I - \frac{1}{2(d_{\max}+1)} L_{\mathcal{G}}$. Then, for any (possibly randomized) decentralized online algorithm, there exists a sequence of local losses $\{f_{t,i} : \mathcal{X} \to \mathbb{R}\}_{i \in [N], t \in [T]}$ satisfying Assumption 1–Assumption 3 with $G = \frac{1}{\alpha B}$, for which*

$$\mathbb{E}[\mathrm{Reg}_{T,1}] \geq \frac{Nd}{20\,\alpha\sqrt{1-\sigma_2(P)}} \log \frac{T}{Nd}.$$

On this instance, $\beta = \frac{\alpha}{16}$, so Theorem 1 with $\eta = \frac{1}{\beta D^2}$ gives $\mathrm{Reg}_{T,i} = \mathcal{O}\big(Nd\alpha^{-1}(1 - \sigma_2(P))^{-1/2} \log^2 T\big)$. Hence our upper bound shown in Theorem 1 is tight in N, d, α, and $1-\sigma_2(P)$, up to a logarithmic factor in T. The proof combines a hard instance for sequential prediction under the logarithmic loss with the cycle-based construction of Wan et al. (2024; 2025), in which information about the current loss cannot reach agent 1 in time. The complete proof is deferred to Appendix B.1.

4 Decentralized Online Linear Regression

In this section, we investigate decentralized online linear regression (D-OLR) over an unbounded domain. The learning protocol proceeds as follows. At each round $t \in [T]$, before making its prediction, agent i observes only its local feature vector $\mathbf{h}_{t,i}$, with $\|\mathbf{h}_{t,i}\|_2 \leq H$ for some $H \geq 0$ and all $t \in [T]$, $i \in [N]$. Each agent i decides a model $\mathbf{x}_i(t) \in \mathbb{R}^d$ to generate its prediction. Crucially, after making this decision, agent i only observes its local true label $y_{t,i} \in [-Y, Y]$ corresponding to the feature vector $\mathbf{h}_{t,i}$, while the features and labels of other agents are not directly available to agent i. The agents communicate the sufficient statistics needed for their model updates through the network. Because the true labels are bounded such that $y_{t,i} \in [-Y, Y]$ for all $t \in [T]$ and $i \in [N]$, we define the prediction for a given feature $\mathbf{h}$ and model $\mathbf{x}$ as the clipped inner product $\Pi_{[-Y,Y]}(\langle \mathbf{h}, \mathbf{x} \rangle)$, where $\Pi_{[-Y,Y]}(a) \triangleq \max\{-Y, \min\{a, Y\}\}$. Furthermore, let $\widehat{y}^i_{t,j} \triangleq \Pi_{[-Y,Y]}\left(\langle \mathbf{h}_{t,j}, \mathbf{x}_i(t) \rangle\right)$ denote the prediction of agent i's model evaluated on agent j's feature at round t. For $j \neq i$, this is a virtual prediction used only to define regret; its evaluation does not require agent i to observe $\mathbf{h}_{t,j}$ online.

We then measure the performance of each agent i using the regret evaluated on the squared loss of their predictions across all global features. Specifically, the regret of agent i with respect to a fixed comparator $\mathbf{x}$ is defined as follows:

$$\mathrm{Reg}^{\mathrm{OLR}}_{T,i}(\mathbf{x}) \triangleq \sum_{t=1}^{T} \sum_{j=1}^{N} \frac{1}{2} \left(\Pi_{[-Y,Y]}\left(\langle \mathbf{h}_{t,j}, \mathbf{x}_i(t) \rangle\right) - y_{t,j}\right)^2 - \sum_{t=1}^{T} \sum_{j=1}^{N} \frac{1}{2} \left(\langle \mathbf{h}_{t,j}, \mathbf{x} \rangle - y_{t,j}\right)^2 .$$

Notably, the squared loss is not globally exp-concave over the unbounded decision domain $\mathbb{R}^d$, and thus the unbounded nature of the decision domain in this setting prevents a direct application of Algorithm 1 discussed in the previous section.

For each block z, agent i forms the covariance and label-feature statistics $C^0_i(z) = \sum_{\tau \in \mathcal{T}_z} \mathbf{h}_{\tau,i} \mathbf{h}^\top_{\tau,i}$ and $Y^0_i(z) = \sum_{\tau \in \mathcal{T}_z} y_{\tau,i} \mathbf{h}_{\tau,i}$. Both statistics are gossiped during block $z+1$ and are available for use in block $z+2$.

We now introduce Algorithm 2 for the D-OLR problem. Our approach builds upon the framework of block-based accelerated gossip, integrating it with the classic online ridge regression algorithm (Hoerl & Kennard, 1970; Vovk, 2001). In contrast to the general exp-concave setting over bounded domains which necessitates the communication of both cumulative gradients and decision vectors, our algorithm exploits a fundamental structural property of D-OLR: the model update depends only on covariance and label-feature statistics, both of which can be aggregated locally and communicated through gossip. Specifically, Algorithm 2 proceeds over these blocks as follows:

- **Step 1: Local Ridge Prediction.** Before making predictions for the current block z, agent i computes a ridge-regression estimate $\mathbf{x}_i(z)$ (Line 3) using the gossiped covariance and label-feature statistics collected from blocks $s = 1$ to $z - 2$. The model is fixed during the block. Here $\Pi_{\mathbb{S}^d_+}$ denotes projection onto the positive semidefinite cone in Frobenius norm. This projection

Algorithm 2: Decentralized online ridge regression for agent $i \in [N]$

Input: gossip parameter θ, epoch length L, regularization parameter $\lambda > 0$, communication matrix P, clipping level Y;
for $z = 1, \dots, T/L$ **do**
  Compute $\mathbf{x}_i(z) = \left(\lambda I_d + N \sum_{s=1}^{z-2} \Pi_{\mathbb{S}_+^d}(C_i(s))\right)^{-1} N \sum_{s=1}^{z-2} Y_i(s)$.
  Set $k_Y = 0$
  for $t = (z-1)L + 1, \dots, zL$ **do**
    Observe the local feature $\mathbf{h}_{t,i}$
    Predict $\Pi_{[-Y,Y]}(\langle \mathbf{h}_{t,i}, \mathbf{x}_i(z) \rangle)$ and receive only the local label $y_{t,i}$
    if $z \geq 2$ **then**
      Update $C_i^{k_Y+1}(z-1) = (1+\theta) \sum_{j \in \mathcal{N}_i} P_{ij} C_j^{k_Y}(z-1) - \theta C_i^{k_Y-1}(z-1)$
      Update $Y_i^{k_Y+1}(z-1) = (1+\theta) \sum_{j \in \mathcal{N}_i} P_{ij} Y_j^{k_Y}(z-1) - \theta Y_i^{k_Y-1}(z-1)$
    Set $k_Y \leftarrow k_Y + 1$
  if $z \geq 2$ *and* $k_Y = L$ **then**
    Set $C_i(z-1) = C_i^L(z-1)$ and $Y_i(z-1) = Y_i^L(z-1)$;
  Set $C_i^0(z) = C_i^{-1}(z) = \sum_{\tau \in \mathcal{T}_z} \mathbf{h}_{\tau,i} \mathbf{h}_{\tau,i}^\top$ and Set $Y_i^0(z) = Y_i^{-1}(z) = \sum_{\tau \in \mathcal{T}_z} y_{\tau,i} \mathbf{h}_{\tau,i}$

is needed because accelerated gossip uses signed combinations and need not preserve positive semidefiniteness and it ensures that the regularized matrix is bounded below by λI_d.

- **Step 2: Prediction and Local Feedback.** Each agent i plays the fixed model $\mathbf{x}_i(z)$ for all rounds $t \in \mathcal{T}_z$ and observes only its local ground-truth label $y_{t,i}$. With a slight abuse of notation, we define $\widehat{y}_{t,j}^i(z) \triangleq \Pi_{[-Y,Y]}(\langle \mathbf{h}_{t,j}, \mathbf{x}_i(z) \rangle)$ as the virtual evaluation of agent i's model on agent j's feature using Algorithm 2.
- **Step 3: Gossip of Covariance and Label-Feature Statistics.** While collecting new local labels during block z, the agents simultaneously perform accelerated gossip on both statistics collected during the previous block $z-1$, applying the covariance update entrywise. After L gossip iterations, $Y_i^L(z-1)$ approximates $\frac{1}{N} \sum_{\tau \in \mathcal{T}_{z-1}} \sum_{j=1}^N y_{\tau,j} \mathbf{h}_{\tau,j}$, while $C_i^L(z-1)$ approximates $\frac{1}{N} \sum_{\tau \in \mathcal{T}_{z-1}} \sum_{j=1}^N \mathbf{h}_{\tau,j} \mathbf{h}_{\tau,j}^\top$. Both updates use the same neighbor-communication step, with $O(d^2)$ real numbers per message.

We remark again that in contrast to Algorithm 1 for the general exp-concave setting which requires gossiping both the cumulative gradients and the decision vectors, Algorithm 2 requires gossiping covariance and label-feature statistics, but not the decision vectors. The absence of a decision-gossip stage allows the procedure to operate with only a 1-block communication delay, rather than the 2-block delay inherent to Algorithm 1 as discussed before. Specifically, after agents collect their local covariance and label-feature statistics in block z, they execute the two gossip recursions simultaneously during block $z+1$ to approximate the network averages. These statistics are then incorporated into the ridge-regression model for use in block $z+2$ and all subsequent blocks.

The following theorem demonstrates that Algorithm 2 guarantees a logarithmic regret for D-OLR.

Theorem 3. *Let P be the communication matrix defined in Definition 1. In the D-OLR problem, if every agent $i \in [N]$ executes Algorithm 2 with $\lambda > 0$, θ chosen according to Eq. (3), and $L = \left\lceil \frac{\sqrt{2}}{(\sqrt{2}-1)\sqrt{1-\sigma_2(P)}} \log\left(\sqrt{14} T^3 N^{7/2} (1 + H^2/\lambda)\right) \right\rceil$, then for any agent $i \in [N]$ and comparator $\mathbf{x} \in \mathbb{R}^d$, the regret satisfies $\mathrm{Reg}_{T,i}^{\mathrm{OLR}}(\mathbf{x}) = \mathcal{O}\big(\lambda \|\mathbf{x}\|_2^2 + Y^2 H^2/\lambda + Y^2 L N d \log(1 + NTH^2/(\lambda d))\big)$.*

By substituting the choice of L into the bound of Theorem 3, we obtain a total regret of $\mathcal{O}\left(Nd(1-\sigma_2(P))^{-1/2} \log^2(T)\right)$. To our knowledge, Theorem 3 establishes the first logarithmic regret guarantee for D-OLR with locally observed features and labels.

4.1 Analysis of Algorithm 2

We briefly explain the main ideas behind the proof of Theorem 3 and the complete proof is deferred to Appendix C. The analysis compares Algorithm 2 with the auxiliary exact delayed ridge predictor

$$\mathbf{x}'(z) \triangleq \left(\lambda I_d + \sum_{s=1}^{z-2}\sum_{\tau\in\mathcal{T}_s}\sum_{j=1}^{N}\mathbf{h}_{\tau,j}\mathbf{h}_{\tau,j}^\top\right)^{-1}\sum_{s=1}^{z-2}\sum_{\tau\in\mathcal{T}_s}\sum_{j=1}^{N}y_{\tau,j}\mathbf{h}_{\tau,j},$$

which uses exactly the blocks available to every agent at block z, but with exact network averaging in place of gossip. Let $f_{t,j}(\mathbf{x}) = \frac{1}{2}(\langle\mathbf{h}_{t,j},\mathbf{x}\rangle - y_{t,j})^2$ and, for $|y| \le Y$, let $\phi_y(u) \triangleq \frac{1}{2}(\Pi_{[-Y,Y]}(u) - y)^2$ be the clipped squared loss. Since agent i plays $\mathbf{x}_i(z)$ throughout block z, the regret decomposes as

$$\begin{aligned}\mathrm{Reg}_{T,i}^{\mathrm{OLR}}(\mathbf{x}) = &\sum_{z=1}^{T/L}\sum_{t\in\mathcal{T}_z}\sum_{j=1}^{N}\big(\phi_{y_{t,j}}(\langle\mathbf{h}_{t,j},\mathbf{x}'(z)\rangle) - f_{t,j}(\mathbf{x})\big)\\ &+\sum_{z=1}^{T/L}\sum_{t\in\mathcal{T}_z}\sum_{j=1}^{N}\big(\phi_{y_{t,j}}(\langle\mathbf{h}_{t,j},\mathbf{x}_i(z)\rangle) - \phi_{y_{t,j}}(\langle\mathbf{h}_{t,j},\mathbf{x}'(z)\rangle)\big).\end{aligned} \tag{9}$$

The first term is a *delay error*, due to the absence of the previous block and the current block's statistics, and the second term is a *gossip error*, due to imperfect averaging of completed blocks. Both are controlled through three elementary properties of the clipped loss: $\phi_y(u) \le \frac{1}{2}(u-y)^2$, $\phi_y(u) \le 2Y^2$, and ϕ_y is $2Y$-Lipschitz. This is why clipping is useful here: the models themselves can be unbounded, whereas the error estimates depend only on the bounded label range.

For the delay error, let M_k denote the minimum of the centralized ridge objective $\frac{\lambda}{2}\|\mathbf{x}\|_2^2 + \sum_{s=1}^{k}\sum_{\tau\in\mathcal{T}_s}\sum_{j=1}^{N} f_{\tau,j}(\mathbf{x})$ over the first k blocks. The increments $M_z - M_{z-1}$ are nonnegative and telescope to at most $\frac{\lambda}{2}\|\mathbf{x}\|_2^2 + \sum_{t=1}^{T}\sum_{j=1}^{N} f_{t,j}(\mathbf{x})$, which is the block form of the Be-The-Regularized-Leader comparison. It therefore suffices to charge the clipped loss of $\mathbf{x}'(z)$ on block z to $M_z - M_{z-1}$. We retain the quadratic structure of the ridge objective to do so. First, completing the square and a matrix-inversion identity show that the ridge solution after block $z-1$, which is fixed throughout block z, incurs a clipped block loss exceeding $M_z - M_{z-1}$ by at most $2Y^2LN$ times the log-determinant increase of block z. Second, $\mathbf{x}'(z)$ also misses block $z-1$. The same identity bounds the squared local-norm movement of the ridge solution caused by block $z-1$ by the trace potential of that block times $M_{z-1} - M_{z-2}$. By the Lipschitz property, this movement costs $\mathcal{O}(Y^2LN)$ times the log-determinant increases of blocks $z-1$ and z whenever $M_{z-1} - M_{z-2} = \mathcal{O}(Y^2LN)$. When $M_{z-1} - M_{z-2}$ is larger, the clipped loss of block $z-1$, which is at most $2Y^2LN$, is well below this increment, and the surplus pays for block z. Summing the log-determinant increases gives $\mathcal{O}\big(Y^2LNd\log(1 + NTH^2/(\lambda d))\big)$.

For the gossip error, the covariance estimates are projected onto the positive semidefinite cone, so both the exact and approximate regularized matrices dominate λI_d. Non-expansiveness of this projection and the matrix-inverse identity $\widehat{A}^{-1} - A^{-1} = \widehat{A}^{-1}(A - \widehat{A})A^{-1}$ reduce the model difference to the covariance and label-feature gossip errors. Using the same accelerated-gossip estimate as in Lemma 1, with the block length L defined in Theorem 3, their cumulative contribution to regret is at most $4Y^2H^2/\lambda$. Summing the delay and gossip terms yields the claimed regret bound.

5 Conclusion and Future Work

We studied decentralized online optimization with exp-concave losses and proposed D-ONS, which achieves $O\big(Nd(1-\sigma_2)^{-1/2}\log^2 T\big)$ regret. We further established a nearly matching lower bound and extended the framework to decentralized online linear regression over an unbounded decision domain. An important direction for future work is to develop adaptive algorithms that do not require prior knowledge of the exp-concavity parameter. It would also be interesting to close the remaining gap in the logarithmic dependence on T between the upper and lower bounds.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing of author-written text, including the presentation of mathematical proofs, and to review author-written proofs for potential logical gaps, algebraic errors, or unclear steps. The tools were not used to generate the main proof ideas, proof strategies, mathematical claims, algorithms, or theoretical results. All proofs and mathematical arguments were developed by the authors, and any issues or suggestions identified by the AI tools were independently checked and resolved by the authors. In particular, the authors manually verified the correctness of the final proofs and derivations rather than relying on the outputs of generative AI tools. We also reviewed all AI-assisted language edits to ensure that they preserved the intended mathematical meaning. We take responsibility for the final content of this work, including text and mathematical arguments produced or revised with the aid of generative AI.

ETHICS STATEMENT

This work is theoretical and focuses on decentralized online optimization and online learning. It does not involve human subjects, the collection or use of personal or sensitive data, or experiments requiring institutional ethical approval. We do not release any new dataset containing personal information. The methods developed in this work are general-purpose optimization methods, and we are not aware of additional ethical concerns specific to the theoretical results presented in this paper.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility. The problem setting, assumptions, and communication model are specified in Section 2. The proposed algorithms, including all update rules and parameter choices, are provided in Section 3.1and Section 4. Proof sketches and the main arguments for the theoretical results are included in the main text, while complete proofs and auxiliary lemmas are provided in Appendices. In particular, the appendix contains the full derivations of the regret upper and lower bounds and the analysis of the decentralized online linear regression algorithm. Since the results of this work are theoretical, reproducing the main claims does not require external datasets or experimental infrastructure.

REFERENCES

Baruch Awerbuch and Robert D Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pp. 45–53, 2004.

Peter Bartlett, Elad Hazan, and Alexander Rakhlin. Adaptive online gradient descent. *Advances in neural information processing systems*, 20, 2007.

John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3): 592–606, 2011.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Saghar Hosseini, Airlie Chapman, and Mehran Mesbahi. Online distributed optimization via dual averaging. In *52nd IEEE Conference on Decision and Control*, pp. 1484–1489. IEEE, 2013.

Pooria Joulani, András György, and Csaba Szepesvári. A modular analysis of adaptive (non-) convex optimization: Optimism, composite objectives, and variational bounds. In *International Conference on Algorithmic Learning Theory*, pp. 681–720. PMLR, 2017.

Dan Li, Kerry D Wong, Yu Hen Hu, and Akbar M Sayeed. Detection, classification, and tracking of targets. *IEEE signal processing magazine*, 19(2):17–29, 2002.

Ji Liu and A Stephen Morse. Accelerated linear iterations for distributed averaging. *Annual Reviews in Control*, 35(2):160–165, 2011.

Haipeng Luo, Mengxiao Zhang, and Peng Zhao. Adaptive bandit convex optimization with heterogeneous curvature. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 1576–1612. PMLR, 02–05 Jul 2022.

Jack J. Mayo, Hedi Hadiji, and Tim van Erven. Scale-free unconstrained online learning for curved losses. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4464–4497. PMLR, 02–05 Jul 2022.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on automatic control*, 54(1):48–61, 2009.

Francesco Orabona. A modern introduction to online learning, 2026. URL `https://arxiv.org/abs/1912.13213`.

Erik Ordentlich and Thomas M Cover. The cost of achieving the best portfolio in hindsight. *Mathematics of Operations Research*, 23(4):960–982, 1998.

Hao Qiu, Emmanuel Esposito, and Mengxiao Zhang. Exploiting curvature in online convex optimization with delayed feedback. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 50448–50479. PMLR, 2025.

Ola Shorinwa, Trevor Halsted, Javier Yu, and Mac Schwager. Distributed optimization methods for multi-robot systems: Part 1—a tutorial [tutorial]. *IEEE Robotics & Automation Magazine*, 31(3): 121–138, 2024.

Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Yuanyu Wan. Black-box reductions for decentralized online convex optimization in changing environments. In *The Thirty Eighth Annual Conference on Learning Theory*, pp. 5605–5631. PMLR, 2025.

Yuanyu Wan, Tong Wei, Mingli Song, and Lijun Zhang. Nearly optimal regret for decentralized online convex optimization. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 4862–4888. PMLR, 2024.

Yuanyu Wan, Tong Wei, Bo Xue, Mingli Song, and Lijun Zhang. Optimal and efficient algorithms for decentralized online convex optimization. *Journal of Machine Learning Research*, 26(135): 1–43, 2025.

Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.

Feng Yan, Shreyas Sundaram, SVN Vishwanathan, and Yuan Qi. Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2483–2493, 2013.

Haishan Ye, Luo Luo, Ziang Zhou, and Tong Zhang. Multi-consensus decentralized accelerated gradient descent. *Journal of machine learning research*, 24(306):1–50, 2023.

Yu-Jie Zhang, Peng Zhao, and Masashi Sugiyama. Non-stationary online learning for curved losses: Improved dynamic regret via mixability. *arXiv preprint arXiv:2506.10616*, 2025.

A AUXILIARY LEMMAS

In this section, we provide some auxiliary technical results that will be helpful throughout the paper.

The following lemma is a standard result standard be-the-leader lemma, see also (Joulani et al., 2017, Theorem 3).

Lemma 3 (Be-The-Regularized-Leader)**.** *Let $\ell_1, \dots, \ell_T : \mathcal{X} \to \mathbb{R}$ be arbitrary functions, and let $\psi_0, \psi_1, \dots, \psi_T : \mathcal{X} \to \mathbb{R}$ satisfy $\psi_0 \equiv 0$. Define $F_t(\mathbf{x}) \triangleq \sum_{s=1}^{t} \ell_s(\mathbf{x}) + \psi_t(\mathbf{x})$ and $\mathbf{x}^\star(t) \in \arg\min_{\mathbf{x}\in\mathcal{X}} F_t(\mathbf{x}), \quad t \in [T]$. Then, for any $\mathbf{x} \in \mathcal{X}$,*

$$\sum_{t=1}^{T} \ell_t\big(\mathbf{x}^\star(t)\big) \le \sum_{t=1}^{T} \ell_t(\mathbf{x}) + \psi_T(\mathbf{x}) - \sum_{t=1}^{T}\Big(\psi_t\big(\mathbf{x}^\star(t)\big) - \psi_{t-1}\big(\mathbf{x}^\star(t)\big)\Big). \tag{10}$$

In particular, if

$$\psi_t(\mathbf{x}) - \psi_{t-1}(\mathbf{x}) \ge 0, \qquad \forall t \in [T],\ \forall \mathbf{x} \in \mathcal{X},$$

then

$$\sum_{t=1}^{T} \ell_t\big(\mathbf{x}^\star(t)\big) \le \sum_{t=1}^{T} \ell_t(\mathbf{x}) + \psi_T(\mathbf{x}).$$

Proof. Define $\tilde{\ell}_t(\mathbf{x}) \triangleq \ell_t(\mathbf{x}) + \psi_t(\mathbf{x}) - \psi_{t-1}(\mathbf{x}), t \in [T]$. Then, for every $t \in [T]$,

$$\sum_{s=1}^{t} \tilde{\ell}_s(\mathbf{x}) = \sum_{s=1}^{t} \ell_s(\mathbf{x}) + \sum_{s=1}^{t}\big(\psi_s(\mathbf{x}) - \psi_{s-1}(\mathbf{x})\big) = \sum_{s=1}^{t} \ell_s(\mathbf{x}) + \psi_t(\mathbf{x}) = F_t(\mathbf{x}).$$

Hence, $\mathbf{x}^\star(t) \in \arg\min_{\mathbf{x}\in\mathcal{X}} \sum_{s=1}^{t} \tilde{\ell}_s(\mathbf{x})$.

Applying the Be-The-Leader lemma to the sequence $\{\tilde{\ell}_t\}_{t=1}^{T}$, we obtain that for any $\mathbf{x} \in \mathcal{X}$, $\sum_{t=1}^{T} \tilde{\ell}_t\big(\mathbf{x}^\star(t)\big) \le \sum_{t=1}^{T} \tilde{\ell}_t(\mathbf{x})$. Expanding both sides yields

$$\sum_{t=1}^{T}\Big(\ell_t\big(\mathbf{x}^\star(t)\big) + \psi_t\big(\mathbf{x}^\star(t)\big) - \psi_{t-1}\big(\mathbf{x}^\star(t)\big)\Big) \le \sum_{t=1}^{T}\Big(\ell_t(\mathbf{x}) + \psi_t(\mathbf{x}) - \psi_{t-1}(\mathbf{x})\Big).$$

Since the right-hand side telescopes and $\psi_0 \equiv 0$, we have $\sum_{t=1}^{T}\big(\psi_t(\mathbf{x}) - \psi_{t-1}(\mathbf{x})\big) = \psi_T(\mathbf{x})$. Therefore,

$$\sum_{t=1}^{T} \ell_t\big(\mathbf{x}^\star(t)\big) + \sum_{t=1}^{T}\Big(\psi_t\big(\mathbf{x}^\star(t)\big) - \psi_{t-1}\big(\mathbf{x}^\star(t)\big)\Big) \le \sum_{t=1}^{T} \ell_t(\mathbf{x}) + \psi_T(\mathbf{x}).$$

Rearranging proves Eq. (10). If, in addition, $\psi_t(\mathbf{x}) - \psi_{t-1}(\mathbf{x}) \ge 0, \forall t \in [T],\ \forall \mathbf{x} \in \mathcal{X}$, then

$$\sum_{t=1}^{T}\Big(\psi_t\big(\mathbf{x}^\star(t)\big) - \psi_{t-1}\big(\mathbf{x}^\star(t)\big)\Big) \ge 0.$$

Dropping this nonnegative term gives

$$\sum_{t=1}^{T} \ell_t\big(\mathbf{x}^\star(t)\big) \le \sum_{t=1}^{T} \ell_t(\mathbf{x}) + \psi_T(\mathbf{x}),$$

which completes the proof. □

The following lemma is the quadratic bound of α-exp-concave functions.

Lemma 4 ((Hazan et al., 2007, Lemma 3))**.** *At each round $t \in [T]$, the loss function $f_{t,i}(\mathbf{x})$ of each agent $i \in [N]$ is α-exp-concave over $\mathcal{X}$, i.e., $f_{t,i}(\mathbf{y}) \ge f_{t,i}(\mathbf{x}) + \langle \nabla f_{t,i}(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\beta}{2}\big(\langle \nabla f_{t,i}(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle\big)^2$ for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, where $\beta = \frac{1}{2}\min\big\{\frac{1}{4GD}, \alpha\big\}$.*

Here we provide a proof of Lemma 1, which shows that after $\Theta(\log(TN)/(1-\sigma_2)^{-1/2})$ rounds of accelerated gossiping, the approximation error to the global average is $\mathcal{O}(1/(T^2N^2))$.

Lemma 1. *Let P be the communication matrix defined in Definition 1. Let*

$$\theta \triangleq \frac{1}{1+\sqrt{1-\sigma_2^2(P)}} \text{ and } L \triangleq \left\lceil \frac{\sqrt{2}\log(T^2\sqrt{14}N^{5/2})}{(\sqrt{2}-1)\sqrt{1-\sigma_2(P)}} \right\rceil. \tag{3}$$

Consider arbitrary vectors $\mathbf{u}_1, \ldots, \mathbf{u}_N \in \mathbb{R}^d$ satisfying $\|\mathbf{u}_i\|_2 \le R, i \in [N]$. Let $U^0 = U^{-1} = [\mathbf{u}_1, \ldots, \mathbf{u}_N]^\top, \overline{\mathbf{u}} = \frac{1}{N}\sum_{i=1}^N \mathbf{u}_i^0, \overline{U} = \mathbf{1}\overline{\mathbf{u}}^\top$. Suppose accelerated gossip is run for L iterations with update $U^{k+1} = (1+\theta)PU^k - \theta U^{k-1}$ for $k \ge 0$. Then, for every agent $i \in [N]$, $\|\mathbf{u}_i^L - \overline{\mathbf{u}}\|_2 \le \frac{2R}{T^2N^2}$.

Proof. By Proposition 1 of Ye et al. (2023), the accelerated gossip iterations satisfy[3]

$$\|U^L - \overline{U}\|_F \le \sqrt{14}\left(1 - \left(1 - \frac{1}{\sqrt{2}}\right)\sqrt{1-\sigma_2(P)}\right)^L \|U^0 - \overline{U}\|_F.$$

By the definition of L and the inequality $\log(1-x) \le -x$ for $x \in (0,1)$, with $a = \left(1 - \frac{1}{\sqrt{2}}\right)\sqrt{1-\sigma_2(P)}$, we have $L \ge \log(T^2\sqrt{14}N^{5/2})/a$, and hence

$$\sqrt{14}\left(1 - \left(1 - \frac{1}{\sqrt{2}}\right)\sqrt{1-\sigma_2(P)}\right)^L = \sqrt{14}(1-a)^L \le \sqrt{14}e^{-aL} \le \frac{1}{T^2N^{5/2}}.$$

Moreover, since $\|\mathbf{u}_i\|_2 \le R$ for all $i \in [N]$, we also have

$$\|\overline{\mathbf{u}}\|_2 = \left\|\frac{1}{N}\sum_{i=1}^N \mathbf{u}_i\right\|_2 \le \frac{1}{N}\sum_{i=1}^N \|\mathbf{u}_i\|_2 \le R.$$

Hence, for every $i \in [N]$,

$$\|\mathbf{u}_i^0 - \overline{\mathbf{u}}\|_2 \le \|\mathbf{u}_i^0\|_2 + \|\overline{\mathbf{u}}\|_2 = \|\mathbf{u}_i\|_2 + \|\overline{\mathbf{u}}\|_2 \le 2R.$$

Therefore,

$$\|U^0 - \overline{U}\|_F = \left(\sum_{i=1}^N \|\mathbf{u}_i^0 - \overline{\mathbf{u}}\|_2^2\right)^{1/2} \le 2R\sqrt{N}.$$

Combining the previous inequalities gives

$$\|U^L - \overline{U}\|_F \le \frac{1}{T^2N^{5/2}} \cdot 2R\sqrt{N} = \frac{2R}{T^2N^2}.$$

Since the Euclidean norm of each row is upper bounded by the Frobenius norm, we obtain

$$\|\mathbf{u}_i^L - \overline{\mathbf{u}}\|_2 \le \|U^L - \overline{U}\|_F \le \frac{2R}{T^2N^2}, \qquad i \in [N]. \tag{11}$$

This completes the proof. □

B Omitted Details from Section 3

In this section, we provide the omitted details from Section 3. We begin by introducing two key lemmas for the proof of Theorem 1. We then combine the two lemmas to prove Theorem 1, and finally prove the lower bound in Theorem 2.

Lemma 5 reduces the decentralized regret of any agent to a sum of block-level surrogate losses. It shows that, after accelerated gossip, the regret can be controlled by linearized gradients plus an exp-concave quadratic correction term, up to lower-order consensus errors. The key role of this lemma is to connect the decentralized online exp-concave problem with the delayed ONS analysis used later.

Compared with Wan (2025, Theorem 1), which uses the uniform quadratic curvature of strongly convex losses, the reduction below retains the gradient-dependent rank-one correction of exp-concavity. The averaging and Jensen steps are standard; their role here is to preserve this correction in the expression passed to Lemma 2.

[3] In the proof of Proposition 1 of Ye et al. (2023), the upper-right block of the transition matrix should be $-\eta_w I$, rather than $-\eta_w W$. With this correction, the projection identities still hold, so the proof requires only this minor adjustment and the proposition remains valid.

Lemma 5. *Under the same conditions as Algorithm 1. For any comparator $\mathbf{x} \in \mathcal{X}$, the regret of any agent $i \in [N]$ satisfies*

$$\mathrm{Reg}_{T,i}(\mathbf{x}) \leq \sum_{z=1}^{T/L} \sum_{j=1}^{N} \left(\langle \mathbf{g}_j(z), \widehat{\mathbf{x}}_j(z) - \mathbf{x} \rangle - \frac{\beta}{2L} \langle \mathbf{g}_j(z), \mathbf{x}_j(z) - \mathbf{x} \rangle^2 \right) + \mathcal{O}\left(\frac{GD}{NT} + \frac{\beta G^2 D^2}{NT} \right).$$

Proof. Before proving the lemma, we first record two consensus bounds implied by Lemma 1. For each block $z \in [T/L]$, recall

$$\overline{\mathbf{x}}(z) = \frac{1}{N} \sum_{j=1}^{N} \widehat{\mathbf{x}}_j(z) \text{ and } \overline{\mathbf{g}}(z) = \frac{1}{N} \sum_{t \in \mathcal{T}_z} \sum_{j=1}^{N} \nabla f_{t,j}\left(\mathbf{x}_j(z)\right).$$

By Assumption 2, the set $\mathcal{X}$ contains the origin and has diameter at most D. Since $\widehat{\mathbf{x}}_i(z) \in \mathcal{X}$, we have $\|\widehat{\mathbf{x}}_i(z)\|_2 \leq D$. Hence, after L accelerated gossip steps in Algorithm 1, and using the non-expansiveness of projection onto a convex set (Orabona, 2026, Proposition 2.11), the played decision satisfies

$$\|\mathbf{x}_i(z) - \overline{\mathbf{x}}(z)\|_2 \leq \left\|\mathbf{x}_i^L(z) - \overline{\mathbf{x}}(z)\right\|_2 \leq \frac{2D}{T^2 N^2}. \tag{12}$$

By Assumption 1,

$$\left\| \sum_{t \in T_z} \nabla f_{t,i}(\mathbf{x}_i(z)) \right\|_2 \leq LG.$$

Applying Lemma 1 to the gradient-gossip process gives

$$\|\mathbf{g}_i(z) - \overline{\mathbf{g}}(z)\|_2 \leq \frac{2LG}{T^2 N^2}. \tag{13}$$

Moreover, by the definition of $\overline{\mathbf{g}}(z)$ and Assumption 1,

$$\|\overline{\mathbf{g}}(z)\|_2 \leq LG.$$

Therefore, by Eq. (13),

$$\|\mathbf{g}_i(z)\|_2 \leq \|\overline{\mathbf{g}}(z)\|_2 + \|\mathbf{g}_i(z) - \overline{\mathbf{g}}(z)\|_2 \leq LG + \frac{2LG}{T^2 N^2} \leq 3LG.$$

For any $\mathbf{x} \in \mathcal{X}$, by Assumption 1, we have

$$\begin{aligned}
f_{t,j}\left(\mathbf{x}_i(z)\right) - f_{t,j}(\mathbf{x}) &\leq f_{t,j}\left(\mathbf{x}_j(z)\right) - f_{t,j}(\mathbf{x}) + G\left\|\mathbf{x}_j(z) - \mathbf{x}_i(z)\right\|_2 \\
&\leq \left\langle \nabla f_{t,j}\left(\mathbf{x}_j(z)\right), \mathbf{x}_j(z) - \mathbf{x} \right\rangle - \frac{\beta}{2} \left\langle \nabla f_{t,j}\left(\mathbf{x}_j(z)\right), \mathbf{x}_j(z) - \mathbf{x} \right\rangle^2 \\
&\quad + G\left\|\mathbf{x}_j(z) - \overline{\mathbf{x}}(z) + \overline{\mathbf{x}}(z) - \mathbf{x}_i(z)\right\|_2 \qquad \text{(Assumption 3 and Lemma 4)} \\
&\leq \left\langle \nabla f_{t,j}\left(\mathbf{x}_j(z)\right), \overline{\mathbf{x}}(z) - \mathbf{x} \right\rangle - \frac{\beta}{2} \left\langle \nabla f_{t,j}\left(\mathbf{x}_j(z)\right), \mathbf{x}_j(z) - \mathbf{x} \right\rangle^2 \\
&\quad + \left\langle \nabla f_{t,j}\left(\mathbf{x}_j(z)\right), \mathbf{x}_j(z) - \overline{\mathbf{x}}(z) \right\rangle + \frac{4GD}{T^2 N^2} \\
&\leq \left\langle \nabla f_{t,j}\left(\mathbf{x}_j(z)\right), \overline{\mathbf{x}}(z) - \mathbf{x} \right\rangle - \frac{\beta}{2} \left\langle \nabla f_{t,j}\left(\mathbf{x}_j(z)\right), \mathbf{x}_j(z) - \mathbf{x} \right\rangle^2 \\
&\quad + G\left\|\mathbf{x}_j(z) - \overline{\mathbf{x}}(z)\right\|_2 + \frac{4GD}{T^2 N^2} \\
&\leq \left\langle \nabla f_{t,j}\left(\mathbf{x}_j(z)\right), \overline{\mathbf{x}}(z) - \mathbf{x} \right\rangle - \frac{\beta}{2} \left\langle \nabla f_{t,j}\left(\mathbf{x}_j(z)\right), \mathbf{x}_j(z) - \mathbf{x} \right\rangle^2 + \frac{6GD}{T^2 N^2},
\end{aligned} \tag{14}$$

where the third and the last inequalities use Eq. (12). Summing over $z \in [T/L]$, $t \in \mathcal{T}_z$, and $j \in [N]$, we have

$$\sum_{z=1}^{T/L} \sum_{t \in \mathcal{T}_z} \sum_{j=1}^{N} \left(f_{t,j}\left(\mathbf{x}_i(z)\right) - f_{t,j}(\mathbf{x}) \right)$$

$$\leq \sum_{z=1}^{T/L} \sum_{t\in\mathcal{T}_z} \sum_{j=1}^{N} \left(\langle \nabla f_{t,j}\left(\mathbf{x}_j(z)\right), \overline{\mathbf{x}}(z) - \mathbf{x} \rangle - \frac{\beta}{2} \langle \nabla f_{t,j}\left(\mathbf{x}_j(z)\right), \mathbf{x}_j(z) - \mathbf{x} \rangle^2 \right) + \frac{6GD}{TN}$$

(Eq. (14))

$$= \sum_{z=1}^{T/L} \left(N \langle \overline{\mathbf{g}}(z), \overline{\mathbf{x}}(z) - \mathbf{x} \rangle - \frac{\beta}{2} \sum_{t\in\mathcal{T}_z} \sum_{j=1}^{N} \langle \nabla f_{t,j}\left(\mathbf{x}_j(z)\right), \mathbf{x}_j(z) - \mathbf{x} \rangle^2 \right) + \frac{6GD}{TN}$$

$$(\overline{\mathbf{g}}(z) = \tfrac{1}{N} \textstyle\sum_{t\in\mathcal{T}_z} \sum_{j=1}^{N} \nabla f_{t,j}\left(\mathbf{x}_j(z)\right))$$

$$\leq \sum_{z=1}^{T/L} \left(N \langle \overline{\mathbf{g}}(z), \overline{\mathbf{x}}(z) - \mathbf{x} \rangle - \frac{\beta}{2NL} \left(\sum_{t\in\mathcal{T}_z} \sum_{j=1}^{N} \langle \nabla f_{t,j}\left(\mathbf{x}_j(z)\right), \mathbf{x}_j(z) - \mathbf{x} \rangle \right)^2 \right) + \frac{6GD}{TN}, \quad (15)$$

where the last inequality holds because of Jensen's inequality. Now we focus on

$$- \frac{\beta}{2NL} \left(\sum_{t\in\mathcal{T}_z} \sum_{j=1}^{N} \langle \nabla f_{t,j}\left(\mathbf{x}_j(z)\right), \mathbf{x}_j(z) - \mathbf{x} \rangle \right)^2$$

$$= -\frac{\beta}{2NL} \left(\sum_{t\in\mathcal{T}_z} \sum_{j=1}^{N} \langle \nabla f_{t,j}\left(\mathbf{x}_j(z)\right), \mathbf{x}_j(z) - \overline{\mathbf{x}}(z) \rangle + \sum_{t\in\mathcal{T}_z} \sum_{j=1}^{N} \langle \nabla f_{t,j}\left(\mathbf{x}_j(z)\right), \overline{\mathbf{x}}(z) - \mathbf{x} \rangle \right)^2$$

$$\leq -\frac{N\beta}{2L} \langle \overline{\mathbf{g}}(z), \overline{\mathbf{x}}(z) - \mathbf{x} \rangle^2 + \beta L G^2 D \sum_{j=1}^{N} \|\mathbf{x}_j(z) - \overline{\mathbf{x}}(z)\|_2$$

$$= -\frac{\beta}{2L} \sum_{j=1}^{N} \langle \overline{\mathbf{g}}(z), \overline{\mathbf{x}}(z) - \mathbf{x} \rangle^2 + \beta L G^2 D \sum_{j=1}^{N} \|\mathbf{x}_j(z) - \overline{\mathbf{x}}(z)\|_2$$

$$\leq -\frac{\beta}{2L} \sum_{j=1}^{N} \langle \mathbf{g}_j(z), \overline{\mathbf{x}}(z) - \mathbf{x} \rangle^2 + \beta L G^2 D \sum_{j=1}^{N} \|\mathbf{x}_j(z) - \overline{\mathbf{x}}(z)\|_2 + 2\beta G D^2 \sum_{j=1}^{N} \|\mathbf{g}_j(z) - \overline{\mathbf{g}}(z)\|_2$$

$$\leq -\frac{\beta}{2L} \sum_{j=1}^{N} \langle \mathbf{g}_j(z), \mathbf{x}_j(z) - \mathbf{x} \rangle^2 + 10\beta L G^2 D \sum_{j=1}^{N} \|\mathbf{x}_j(z) - \overline{\mathbf{x}}(z)\|_2 + 2\beta G D^2 \sum_{j=1}^{N} \|\mathbf{g}_j(z) - \overline{\mathbf{g}}(z)\|_2$$

$$\leq -\frac{\beta}{2L} \sum_{j=1}^{N} \langle \mathbf{g}_j(z), \mathbf{x}_j(z) - \mathbf{x} \rangle^2 + \frac{24\beta G^2 D^2 L}{T^2 N} \quad (16)$$

Here, the first inequality is obtained by decomposing $\mathbf{x}_j(z) - \mathbf{x} = \left(\overline{\mathbf{x}}(z) - \mathbf{x}\right) + \left(\mathbf{x}_j(z) - \overline{\mathbf{x}}(z)\right)$, and applying $-(a+b)^2 \leq -a^2 + 2|a||b|$, Cauchy–Schwarz, Assumption 1, and Assumption 2. The equality follows from the fact that $\langle \overline{\mathbf{g}}(z), \overline{\mathbf{x}}(z) - \mathbf{x} \rangle^2$ is independent of j. For the second inequality, we use the identity $|a^2 - b^2| = |a-b||a+b|$ with $a = \langle \overline{\mathbf{g}}(z), \overline{\mathbf{x}}(z) - \mathbf{x} \rangle$, and $b = \langle \mathbf{g}_j(z), \overline{\mathbf{x}}(z) - \mathbf{x} \rangle$,together with Cauchy–Schwarz, Assumption 1, and Assumption 2. The third inequality is proved analogously by taking $a = \langle \mathbf{g}_j(z), \overline{\mathbf{x}}(z) - \mathbf{x} \rangle$ and $b = \langle \mathbf{g}_j(z), \mathbf{x}_j(z) - \mathbf{x} \rangle$.The last inequality follows from Eq. (12) and Eq. (13).

Combining Eq. (15) and Eq. (16), we have

$$\sum_{z=1}^{T/L} \sum_{t\in\mathcal{T}_z} \sum_{j=1}^{N} \left(f_{t,j}\left(\mathbf{x}_i(z)\right) - f_{t,j}(\mathbf{x}) \right)$$

$$\leq \sum_{z=1}^{T/L} \left(N \langle \overline{\mathbf{g}}(z), \overline{\mathbf{x}}(z) - \mathbf{x} \rangle - \frac{\beta}{2L} \sum_{j=1}^{N} \langle \mathbf{g}_j(z), \mathbf{x}_j(z) - \mathbf{x} \rangle^2 \right) + \frac{6GD}{TN} + \frac{24\beta G^2 D^2}{TN}$$

$$= \sum_{z=1}^{T/L} \sum_{j=1}^{N} \left(\langle \overline{\mathbf{g}}(z), \hat{\mathbf{x}}_j(z) - \mathbf{x} \rangle - \frac{\beta}{2L} \langle \mathbf{g}_j(z), \mathbf{x}_j(z) - \mathbf{x} \rangle^2 \right) + \frac{6GD}{TN} + \frac{24\beta G^2 D^2}{TN}$$

$$(\overline{\mathbf{x}}(z) = \tfrac{1}{N} \textstyle\sum_{j=1}^{N} \hat{\mathbf{x}}_j(z))$$

$$\leq \sum_{z=1}^{T/L}\sum_{j=1}^{N}\left(\langle \mathbf{g}_j(z), \hat{\mathbf{x}}_j(z)-\mathbf{x}\rangle - \frac{\beta}{2L}\langle \mathbf{g}_j(z), \mathbf{x}_j(z)-\mathbf{x}\rangle^2\right) + \frac{6GD}{TN} + \frac{24\beta G^2D^2}{TN}$$
$$+ D\sum_{z=1}^{T/L}\sum_{j=1}^{N}\|\mathbf{g}_j(z)-\overline{\mathbf{g}}(z)\|_2$$
$$= \sum_{z=1}^{T/L}\sum_{j=1}^{N}\left(\langle \mathbf{g}_j(z), \hat{\mathbf{x}}_j(z)-\mathbf{x}\rangle - \frac{\beta}{2L}\langle \mathbf{g}_j(z), \mathbf{x}_j(z)-\mathbf{x}\rangle^2\right) + \mathcal{O}\left(\frac{GD}{NT}+\frac{\beta G^2D^2}{NT}\right),$$

where the last inequality follows from Eq. (13). □

It remains to control the block-level surrogate regrets appearing in Lemma 5. This step is more delicate than in the strongly convex case because the linear term and the quadratic correction are evaluated at two different points: the preparatory decision $\widehat{\mathbf{x}}_j(z)$ and the executed decision $\mathbf{x}_j(z)$. To handle this mismatch, we compare the delayed FTRL iterate $\widehat{\mathbf{x}}_j(z)$, which only uses feedback up to block $z-3$, with a current regularized leader that uses feedback up to block z. The comparison is carried out in the adaptive second-order norm induced by the accumulated gradient matrix.

Lemma 2. *Under the same conditions as Algorithm 1. For any $j \in [N]$ and every comparator $\mathbf{x} \in \mathcal{X}$,*

$$\sum_{z=1}^{T/L}\left(\langle \mathbf{g}_j(z), \hat{\mathbf{x}}_j(z)-\mathbf{x}\rangle - \frac{\beta}{2L}\langle \mathbf{g}_j(z), \mathbf{x}_j(z)-\mathbf{x}\rangle^2\right) \leq \mathcal{O}\Big(\eta\|\mathbf{x}\|_2^2 + \frac{d(1+\beta GD)L\log\left(1+\frac{\beta G^2T}{\eta d}\right)}{\beta}\Big).$$

Proof. For all $j \in [N]$, define $F_{z,j}^\star(\mathbf{x}) \triangleq \sum_{s=1}^{z}\langle \mathbf{g}_j(s), \mathbf{x}\rangle + \psi_{z,j}^\star(\mathbf{x})$ and let $\mathbf{x}_j^\star(z) \in \arg\min_{\mathbf{x}\in\mathcal{X}} F_{z,j}^\star(\mathbf{x})$, where $\psi_{z,j}^\star(\mathbf{x}) \triangleq \frac{\beta}{2L}\sum_{s=1}^{z}\langle \mathbf{g}_j(s), \mathbf{x}_j(s)-\mathbf{x}\rangle^2 + \frac{\eta}{2}\|\mathbf{x}\|_2^2$. For all $j \in [N]$, we have

$$\sum_{z=1}^{T/L}\langle \mathbf{g}_j(z), \hat{\mathbf{x}}_j(z)-\mathbf{x}\rangle = \sum_{z=1}^{T/L}\langle \mathbf{g}_j(z), \hat{\mathbf{x}}_j(z)-\mathbf{x}_j^\star(z)\rangle + \sum_{z=1}^{T/L}\langle \mathbf{g}_j(z), \mathbf{x}_j^\star(z)-\mathbf{x}\rangle\ . \tag{17}$$

Using Lemma 3, we have

$$\sum_{z=1}^{T/L}\langle \mathbf{g}_j(z), \mathbf{x}_j^\star(z)-\mathbf{x}\rangle \leq \psi_{T/L,j}^\star(\mathbf{x}) - \sum_{z=1}^{T/L}\Big(\psi_{z,j}^\star\big(\mathbf{x}_j^\star(z)\big) - \psi_{z-1,j}^\star\big(\mathbf{x}_j^\star(z)\big)\Big) \leq \psi_{T/L,j}^\star(\mathbf{x}), \tag{18}$$

where the last inequality follows from $\psi_{z,j}^\star(\mathbf{x}) - \psi_{z-1,j}^\star(\mathbf{x}) = \frac{\beta}{2L}\langle \mathbf{g}_j(z), \mathbf{x}_j(z)-\mathbf{x}\rangle^2 \geq 0$.

It remains to bound

$$\sum_{z=1}^{T/L}\langle \mathbf{g}_j(z), \hat{\mathbf{x}}_j(z)-\mathbf{x}_j^\star(z)\rangle.$$

For notational convenience, for any integer $z \leq 0$, we set $A_j(z) \triangleq \eta I_d$, For any integer $z \geq 1$, define

$$A_j(z) \triangleq \eta I_d + \frac{\beta}{L}\sum_{s=1}^{z}\mathbf{g}_j(s)\mathbf{g}_j(s)^\top.$$

For each $z \geq 1$, We define

$$\hat{F}_{z,j}(\mathbf{x}) \triangleq \sum_{s=1}^{z-3}\left(\langle \mathbf{g}_j(s), \mathbf{x}\rangle + \frac{\beta}{2L}\langle \mathbf{g}_j(s), \mathbf{x}_j(s)-\mathbf{x}\rangle^2\right) + \frac{\eta}{2}\|\mathbf{x}\|_2^2$$

and $\hat{\mathbf{x}}_j(z) \in \arg\min_{\mathbf{x}\in\mathcal{X}}\hat{F}_{z,j}(\mathbf{x})$. Since $F_{z,j}^\star$ is 1-strongly convex with respect to $\|\cdot\|_{A_j(z)}$, and $\mathbf{x}_j^\star(z)$ minimizes $F_{z,j}^\star$ over $\mathcal{X}$, we have

$$F_{z,j}^\star(\hat{\mathbf{x}}_j(z)) - F_{z,j}^\star(\mathbf{x}_j^\star(z)) \geq \frac{1}{2}\|\hat{\mathbf{x}}_j(z)-\mathbf{x}_j^\star(z)\|_{A_j(z)}^2. \tag{19}$$

Likewise, since $\hat{F}_{z,j}$ is 1-strongly convex with respect to $\|\cdot\|_{A_j(z-3)}$, and $\hat{\mathbf{x}}_j(z)$ minimizes $\hat{F}_{z,j}$ over $\mathcal{X}$, we have

$$\hat{F}_{z,j}(\mathbf{x}_j^\star(z)) - \hat{F}_{z,j}(\hat{\mathbf{x}}_j(z)) \geq \frac{1}{2}\|\hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\|^2_{A_j(z-3)}. \tag{20}$$

Adding Eq. (19) and Eq. (20) is equivalent to applying the stability lemma of Qiu et al. (2025, Lemma A.2) to these two regularized objectives. This is a standard strong-convexity argument, not a new ingredient; we include the derivation to make the treatment of the two decisions explicit. Summing Eq. (19) and Eq. (20), and expanding the definitions of $F^\star_{z,j}$ and $\hat{F}_{z,j}$, yields

$$\begin{aligned}
&\frac{1}{2}\|\hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\|^2_{A_j(z)} + \frac{1}{2}\|\hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\|^2_{A_j(z-3)} \\
&\leq \sum_{s=z-2}^{z} \langle \mathbf{g}_j(s), \hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\rangle \\
&\quad + \frac{\beta}{2L}\sum_{s=z-2}^{z}\Big(\langle \mathbf{g}_j(s), \mathbf{x}_j(s) - \hat{\mathbf{x}}_j(z)\rangle^2 - \langle \mathbf{g}_j(s), \mathbf{x}_j(s) - \mathbf{x}_j^\star(z)\rangle^2\Big).
\end{aligned}$$

Using $a^2 - b^2 = (a-b)(a+b)$, with $a = \langle \mathbf{g}_j(s), \mathbf{x}_j(s) - \hat{\mathbf{x}}_j(z)\rangle$ and $b = \langle \mathbf{g}_j(s), \mathbf{x}_j(s) - \mathbf{x}_j^\star(z)\rangle$ we obtain

$$\begin{aligned}
&\frac{1}{2}\|\hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\|^2_{A_j(z)} + \frac{1}{2}\|\hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\|^2_{A_j(z-3)} \\
&\leq \sum_{s=z-2}^{z} \big|\langle \mathbf{g}_j(s), \hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\rangle\big| \\
&\quad + \frac{\beta}{2L}\sum_{s=z-2}^{z}\big|\langle \mathbf{g}_j(s), \hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\rangle\big| \cdot \big|\langle \mathbf{g}_j(s), \hat{\mathbf{x}}_j(z) + \mathbf{x}_j^\star(z) - 2\mathbf{x}_j(s)\rangle\big|.
\end{aligned}$$

Using Assumption 1, the definition of $\overline{\mathbf{g}}(s)$, and Eq. (13), we have

$$\|\mathbf{g}_j(s)\|_2 \leq \|\overline{\mathbf{g}}(s)\|_2 + \|\mathbf{g}_j(s) - \overline{\mathbf{g}}(s)\|_2 \leq LG + \frac{2LG}{T^2N^2} \leq 3LG.$$

Moreover, by Assumption 2,

$$\|\hat{\mathbf{x}}_j(z) + \mathbf{x}_j^\star(z) - 2\mathbf{x}_j(s)\|_2 \leq \|\hat{\mathbf{x}}_j(z) - \mathbf{x}_j(s)\|_2 + \|\mathbf{x}_j^\star(z) - \mathbf{x}_j(s)\|_2 \leq 2D.$$

Therefore,

$$\big|\langle \mathbf{g}_j(s), \hat{\mathbf{x}}_j(z) + \mathbf{x}_j^\star(z) - 2\mathbf{x}_j(s)\rangle\big| \leq 6LGD.$$

and therefore

$$\frac{1}{2}\|\hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\|^2_{A_j(z)} + \frac{1}{2}\|\hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\|^2_{A_j(z-3)} \leq (1 + 3\beta GD)\sum_{s=z-2}^{z}\big|\langle \mathbf{g}_j(s), \hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\rangle\big|.$$

Applying Cauchy–Schwarz in the pair of dual norms induced by $A_j(z)$,

$$\big|\langle \mathbf{g}_j(s), \hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\rangle\big| \leq \|\mathbf{g}_j(s)\|_{A_j(z)^{-1}}\|\hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\|_{A_j(z)},$$

we get

$$\begin{aligned}
&\frac{1}{2}\|\hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\|^2_{A_j(z)} + \frac{1}{2}\|\hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\|^2_{A_j(z-3)} \\
&\leq (1 + 3\beta GD)\left(\sum_{s=z-2}^{z}\|\mathbf{g}_j(s)\|_{A_j(z)^{-1}}\right)\|\hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\|_{A_j(z)}.
\end{aligned}$$

Dropping the second nonnegative term on the left-hand side gives

$$\|\hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\|_{A_j(z)} \leq 2\,(1 + 3\beta GD)\sum_{s=z-2}^{z}\|\mathbf{g}_j(s)\|_{A_j(z)^{-1}}.$$

Consequently,

$$\langle \mathbf{g}_j(z), \hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z) \rangle \le \|\mathbf{g}_j(z)\|_{A_j(z)^{-1}} \|\hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z)\|_{A_j(z)} \le 2\,(1+3\beta GD)\,\|\mathbf{g}_j(z)\|_{A_j(z)^{-1}} \sum_{s=z-2}^{z} \|\mathbf{g}_j(s)\|_{A_j(z)^{-1}}. \tag{21}$$

Summing Eq. (21) over $z = 1, \dots, T/L$, we obtain

$$\begin{aligned}
\sum_{z=1}^{T/L} \langle \mathbf{g}_j(z), \hat{\mathbf{x}}_j(z) - \mathbf{x}_j^\star(z) \rangle &\le 2\,(1+3\beta GD) \sum_{z=1}^{T/L} \left(\|\mathbf{g}_j(z)\|_{A_j(z)^{-1}} \sum_{s=z-2}^{z} \|\mathbf{g}_j(s)\|_{A_j(z)^{-1}} \right) \\
&\le (1+3\beta GD) \sum_{z=1}^{T/L} \left(3\|\mathbf{g}_j(z)\|^2_{A_j(z)^{-1}} + \sum_{s=z-2}^{z} \|\mathbf{g}_j(s)\|^2_{A_j(z)^{-1}} \right) \\
&\qquad \text{(AM-GM inequality)} \\
&\le (1+3\beta GD) \sum_{z=1}^{T/L} \left(3\|\mathbf{g}_j(z)\|^2_{A_j(z)^{-1}} + \sum_{s=z-2}^{z} \|\mathbf{g}_j(s)\|^2_{A_j(s)^{-1}} \right) \\
&\qquad (A_j(s) \preceq A_j(z)) \\
&\le 6\,(1+3\beta GD) \sum_{z=1}^{T/L} \|\mathbf{g}_j(z)\|^2_{A_j(z)^{-1}} \\
&\le \frac{6d\,(1+3\beta GD)\,L}{\beta} \log\left(1 + \frac{9\beta TG^2}{\eta d}\right).
\end{aligned} \tag{22}$$

where the last inequality holds by using Orabona (2026, Lemma 7.29).

Substituting Eq. (18) and Eq. (22) into Eq. (17), we have

$$\begin{aligned}
&\sum_{z=1}^{T/L} \left(\langle \mathbf{g}_j(z), \hat{\mathbf{x}}_j(z) - \mathbf{x} \rangle - \frac{\beta}{2L} \langle \mathbf{g}_j(z), \mathbf{x}_j(z) - \mathbf{x} \rangle^2 \right) \\
&\le \psi^\star_{T/L,j}(\mathbf{x}) - \sum_{z=1}^{T/L} \frac{\beta}{2L} \langle \mathbf{g}_j(z), \mathbf{x}_j(z) - \mathbf{x} \rangle^2 + \frac{6d\,(1+3\beta GD)\,L}{\beta} \log\left(1 + \frac{9\beta TG^2}{\eta d}\right) \\
&= \frac{\eta}{2}\|\mathbf{x}\|_2^2 + \frac{6d\,(1+3\beta GD)\,L}{\beta} \log\left(1 + \frac{9\beta TG^2}{\eta d}\right) \\
&= \mathcal{O}\left(\eta\|\mathbf{x}\|_2^2 + \frac{d\,(1+3\beta GD)\,L}{\beta} \log\left(1 + \frac{\beta TG^2}{\eta d}\right) \right).
\end{aligned} \tag{23}$$

□

We now prove Theorem 1 by combining the regret decomposition with the delayed ONS bound.

Theorem 1. *Suppose that the sequence of loss functions $\{f_{t,i}\}_{t\in[T],i\in[N]}$ satisfy Assumption 1, 2, and 3. Let P be the communication matrix defined in Definition 1. If every agent $i \in [N]$ runs Algorithm 1 with $\beta = \frac{1}{2}\min\{\frac{1}{4GD}, \alpha\}$ and with parameters θ and L chosen in Eq. (3), then, for any agent $i \in [N]$, the regret is bounded by:*

$$\mathrm{Reg}_{T,i}(\mathbf{x}) = \mathcal{O}\left(\eta N\|\mathbf{x}\|_2^2 + \frac{Nd(1+\beta GD)L}{\beta} \log\left(1 + \frac{\beta G^2 T}{\eta d}\right) \right). \tag{5}$$

Proof. Fix any agent $i \in [N]$ and any comparator $\mathbf{x} \in \mathcal{X}$. By Lemma 5,

$$\mathrm{Reg}_{T,i} \le \sum_{z=1}^{T/L} \sum_{j=1}^{N} \left(\langle \mathbf{g}_j(z), \widehat{\mathbf{x}}_j(z) - \mathbf{x} \rangle - \frac{\beta}{2L} \langle \mathbf{g}_j(z), \mathbf{x}_j(z) - \mathbf{x} \rangle^2 \right) + \mathcal{O}\left(\frac{GD}{NT} + \frac{\beta G^2 D^2}{NT} \right).$$

Applying Lemma 2 to each $j \in [N]$ gives

$$\mathrm{Reg}_{T,i}(\mathbf{x}) = \mathcal{O}\left(\eta N \|\mathbf{x}\|_2^2 + \frac{Nd(1+\beta GD)L}{\beta} \log\left(1 + \frac{\beta G^2 T}{\eta d}\right)\right).$$

This proves Eq. (5). Substituting the choice of L in Eq. (3) gives the stated dependence on the spectral gap. □

B.1 Omitted Details in Section 3.3

We now prove Theorem 2. We begin with a lower bound for a centralized sequential prediction problem and then lift it to the decentralized setting by placing the informative losses sufficiently far from agent 1 on a cycle network.

Fix $B > 0$ and $\alpha > 0$, and let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \le B\}$. For $u \in [-B, B]$, define $\pi(u) \triangleq \frac{1}{2} + \frac{u}{4B} \in [\frac{1}{4}, \frac{3}{4}]$, and for $q \in (0,1)$ and $y \in \{0,1\}$ let $\ell(q,y) \triangleq -y \log q - (1-y)\log(1-q)$ denote the logarithmic loss. For $j \in [d]$ and $y \in \{0,1\}$, define

$$h_{j,y}(\mathbf{x}) \triangleq \frac{1}{\alpha}\,\ell\big(\pi(x_j), y\big) = -\frac{1}{\alpha}\log\Big(\frac{1}{2} + \frac{(2y-1)x_j}{4B}\Big), \qquad \mathbf{x} \in \mathcal{X}. \tag{24}$$

Since $|x_j| \le \|\mathbf{x}\|_2 \le B$, the argument of the logarithm lies in $[\frac{1}{4}, \frac{3}{4}]$. The set $\mathcal{X}$ is compact and convex, contains the origin, and has diameter $D = 2B$. Moreover, $h_{j,y}$ is the composition of $-\frac{1}{\alpha}\log(\cdot)$ with a positive affine function, so it is convex and differentiable. In particular, $\|\nabla h_{j,y}(\mathbf{x})\|_2 = \frac{1}{\alpha}\frac{1/(4B)}{\frac{1}{2}+\frac{(2y-1)x_j}{4B}} \le \frac{1}{\alpha B} \triangleq G$. Also, $\exp(-\alpha h_{j,y}(\mathbf{x})) = \frac{1}{2} + \frac{(2y-1)x_j}{4B}$ is affine and hence concave, so $h_{j,y}$ is α-exp-concave on $\mathcal{X}$. The zero function satisfies the same assumptions. Thus all losses used below satisfy Assumption 1–Assumption 3 with $G = \frac{1}{\alpha B}$, and $\alpha G D = 2$.

The next lemma gives the centralized lower bound that will be embedded into the network construction.

Lemma 6. *Let $d, n \ge 1$ be integers, let $M = nd$, and let $j_i \triangleq (i \bmod d) + 1$ for $i = 0, \dots, M-1$. Let ξ be a random variable, and let $\tilde{\mathbf{x}}(i) = \varphi_i(y_0, \dots, y_{i-1}, \xi) \in \mathcal{X}$, $i = 0, \dots, M-1$, be arbitrary measurable maps. Then there exist outcomes $y_0, \dots, y_{M-1} \in \{0,1\}$ such that*

$$\mathbb{E}_\xi\left[\sum_{i=0}^{M-1} h_{j_i,y_i}\big(\tilde{\mathbf{x}}(i)\big) - \min_{\mathbf{x}\in\mathcal{X}} \sum_{i=0}^{M-1} h_{j_i,y_i}(\mathbf{x})\right] \ge \frac{d}{2\alpha}\log\frac{n}{2\pi e\, d}.$$

Proof. We use the probabilistic method. Let $w \triangleq \frac{1}{4\sqrt{d}}$, draw $\theta_1, \dots, \theta_d$ independently and uniformly from $[\frac{1}{2} - w, \frac{1}{2} + w]$, and, conditionally on $\theta = (\theta_1, \dots, \theta_d)$, draw $Y_0, \dots, Y_{M-1}$ independently with $Y_i \sim \mathrm{Bernoulli}(\theta_{j_i})$. We couple ξ independently of (θ, Y). It is enough to prove that the expectation of the regret quantity in the statement, taken over (θ, Y, ξ), is at least $\frac{d}{2\alpha}\log\frac{n}{2\pi e\, d}$, because then some realization of Y attains at least this value after averaging only over ξ.

For a given θ, define $\mathbf{x}_\theta \in \mathbb{R}^d$ coordinatewise by $(\mathbf{x}_\theta)_j = 4B(\theta_j - \frac{1}{2})$. Since $\|\mathbf{x}_\theta\|_2^2 = 16B^2\sum_{j=1}^d(\theta_j - \frac{1}{2})^2 \le 16B^2 d w^2 = B^2$, we have $\mathbf{x}_\theta \in \mathcal{X}$, and $\pi((\mathbf{x}_\theta)_j) = \theta_j$ for every $j \in [d]$. Hence, for every realization of (θ, Y),

$$\min_{\mathbf{x}\in\mathcal{X}} \sum_{i=0}^{M-1} h_{j_i,Y_i}(\mathbf{x}) \le \sum_{i=0}^{M-1} h_{j_i,Y_i}(\mathbf{x}_\theta) = \frac{1}{\alpha}\sum_{i=0}^{M-1} \ell(\theta_{j_i}, Y_i). \tag{25}$$

Next, let $q_i \triangleq \pi(\tilde{x}_{j_i}(i)) \in [\frac{1}{4}, \frac{3}{4}]$, where $\tilde{x}_{j_i}(i)$ is the j_i-th coordinate of $\tilde{\mathbf{x}}(i)$, and set $\bar{q}_i(y_0, \dots, y_{i-1}) \triangleq \mathbb{E}_\xi[q_i] \in [\frac{1}{4}, \frac{3}{4}]$. Because ξ is independent of the outcomes, $\bar{q}_i$ depends only on $(y_0, \dots, y_{i-1})$. Since $\ell(\cdot, y)$ is convex, Jensen's inequality implies, for every deterministic outcome sequence y,

$$\mathbb{E}_\xi\left[\sum_{i=0}^{M-1} h_{j_i,y_i}\big(\tilde{\mathbf{x}}(i)\big)\right] \ge \frac{1}{\alpha}\sum_{i=0}^{M-1} \ell(\bar{q}_i, y_i). \tag{26}$$

The sequential probabilities $\bar{q}_i$ define a probability mass function Q on $\{0,1\}^M$ by

$$Q(y) \triangleq \prod_{i=0}^{M-1} \bar{q}_i(y_0,\dots,y_{i-1})^{y_i}\big(1-\bar{q}_i(y_0,\dots,y_{i-1})\big)^{1-y_i},$$

whereas, for each θ, the conditional law of Y is $P_\theta(y) \triangleq \prod_{i=0}^{M-1} \theta_{j_i}^{y_i}(1-\theta_{j_i})^{1-y_i}$. Therefore $\sum_{i=0}^{M-1}(\ell(\bar{q}_i,y_i)-\ell(\theta_{j_i},y_i)) = \log\frac{P_\theta(y)}{Q(y)}$. Combining Eq. (25) and Eq. (26), and then taking expectations over (θ,Y), gives

$$\mathbb{E}\left[\sum_{i=0}^{M-1} h_{j_i,Y_i}\big(\tilde{\mathbf{x}}(i)\big) - \min_{\mathbf{x}\in\mathcal{X}}\sum_{i=0}^{M-1} h_{j_i,Y_i}(\mathbf{x})\right] \geq \frac{1}{\alpha}\,\mathbb{E}_\theta\big[\mathrm{KL}(P_\theta\|Q)\big] \geq \frac{1}{\alpha} I(\theta;Y),$$

where the last inequality follows from $\mathbb{E}_\theta[\mathrm{KL}(P_\theta\|Q)] = I(\theta;Y) + \mathrm{KL}(\bar{P}\|Q)$, with $\bar{P} \triangleq \mathbb{E}_\theta[P_\theta]$ the marginal law of Y.

For each $j \in [d]$, let $Y^{(j)} \triangleq (Y_i)_{i:\, j_i=j}$, which contains exactly n observations. The pairs $(\theta_j, Y^{(j)})$, $j \in [d]$, are mutually independent, so $I(\theta;Y) = \sum_{j=1}^d I(\theta_j;Y^{(j)})$. Since θ_j is uniform on an interval of length $2w$, $h(\theta_j) = \log(2w)$, where h denotes differential entropy. Let $\hat{\theta}_j$ be the empirical mean of $Y^{(j)}$. Conditional on θ_j, $\hat{\theta}_j$ is unbiased with variance $\theta_j(1-\theta_j)/n$, and hence $\mathbb{E}[(\theta_j - \hat{\theta}_j)^2] = \mathbb{E}[\theta_j(1-\theta_j)]/n \leq 1/(4n)$. Using the fact that the Gaussian maximizes differential entropy among distributions with a fixed variance, followed by Jensen's inequality and the optimality of conditional expectation in mean squared error, we obtain

$$h(\theta_j \mid Y^{(j)}) \leq \frac{1}{2}\log\Big(2\pi e\,\mathbb{E}\big[\mathrm{Var}(\theta_j \mid Y^{(j)})\big]\Big) \leq \frac{1}{2}\log\Big(2\pi e\,\mathbb{E}\big[(\theta_j-\hat{\theta}_j)^2\big]\Big) \leq \frac{1}{2}\log\frac{\pi e}{2n}.$$

Consequently,

$$I(\theta_j;Y^{(j)}) = h(\theta_j) - h(\theta_j \mid Y^{(j)}) \geq \log(2w) - \frac{1}{2}\log\frac{\pi e}{2n} = \frac{1}{2}\log\frac{n}{2\pi e\, d}.$$

Summing over $j \in [d]$ and substituting this bound into the previous display proves the claim. □

We now use Lemma 6 to construct the decentralized hard instance. For completeness, we repeat Theorem 2:

Theorem 2. *Let $N = 4K+2$ for some positive integer K, let $d \geq 1$, let $\alpha > 0$, fix any $B > 0$, and suppose that $T \geq (\pi e)^2 N d^3$. Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq B\}$, a compact convex set containing the origin with diameter $D = 2B$, and let $\mathcal{G} = ([N], E)$ be the cycle graph with $P = I - \frac{1}{2(d_{\max}+1)} L_{\mathcal{G}}$. Then, for any (possibly randomized) decentralized online algorithm, there exists a sequence of local losses $\{f_{t,i} : \mathcal{X} \to \mathbb{R}\}_{i\in[N],t\in[T]}$ satisfying Assumption 1–Assumption 3 with $G = \frac{1}{\alpha B}$, for which*

$$\mathbb{E}[\mathrm{Reg}_{T,1}] \geq \frac{Nd}{20\,\alpha\sqrt{1-\sigma_2(P)}}\log\frac{T}{Nd}.$$

Proof. Let $\mathcal{G}$ be the cycle with edges $\{v, v+1\}$ for $v \in [N-1]$ and $\{N,1\}$, and write $K = \frac{N-2}{4}$. Define $V_{\mathrm{mid}} \triangleq \{K+1,\dots,N-K+1\}$. This set contains $S \triangleq N-2K+1 = 2K+3$ agents, and every $v \in V_{\mathrm{mid}}$ satisfies $\mathrm{dist}(1,v) = \min\{v-1, N-v+1\} \geq K$. Let $Z \triangleq \lfloor T/K \rfloor$, $n \triangleq \lfloor Z/d \rfloor$, and $M \triangleq nd \leq Z$, and define $\mathcal{I}_i \triangleq \{iK+1,\dots,(i+1)K\}$ for $i = 0,\dots,M-1$. These are disjoint intervals of length K contained in $[T]$. As in Lemma 6, let $j_i = (i \bmod d) + 1$.

For outcomes $y_0,\dots,y_{M-1} \in \{0,1\}$ to be specified below, define, for every $i = 0,\dots,M-1$ and $t \in \mathcal{I}_i$,

$$f_{t,v} \triangleq \begin{cases} h_{j_i,y_i}, & v \in V_{\mathrm{mid}}, \\ 0, & v \notin V_{\mathrm{mid}}, \end{cases}$$

and set $f_{t,v} \triangleq 0$ for every $v \in [N]$ and $t > MK$. By the construction above, every local loss satisfies Assumption 1–Assumption 3 with $G = \frac{1}{\alpha B}$ and $D = 2B$. For brevity, write $h_i \triangleq h_{j_i,y_i}$.

Under the D-OCO protocol, an agent chooses its decision at round t using its past local feedback and the messages received before that decision, then observes its round-t local feedback and performs one round of communication. It follows inductively that feedback generated by agent v at round s can affect the decision of agent u at round t only if $s + \max\{\mathrm{dist}(u, v), 1\} \le t$. Fix $t \in \mathcal{I}_i$ and $v \in V_{\mathrm{mid}}$. Since $\mathrm{dist}(1, v) \ge K$, such feedback can affect $\mathbf{x}_1(t)$ only if $s \le t - K \le iK$. Thus, before agent 1 chooses any decision in block $\mathcal{I}_i$, it can only have received outcome-dependent information from the preceding blocks $\mathcal{I}_0, \dots, \mathcal{I}_{i-1}$. The agents outside V_{mid} have identically zero local losses, so they reveal no additional information about the current outcome. Hence, if ξ denotes all internal randomness of the decentralized algorithm, independently of the loss sequence, then for every $t \in \mathcal{I}_i$, the decision $\mathbf{x}_1(t)$ is a measurable function of $(y_0, \dots, y_{i-1}, \xi)$.

Define the block-averaged decision $\tilde{\mathbf{x}}(i) \triangleq K^{-1} \sum_{t \in \mathcal{I}_i} \mathbf{x}_1(t)$. By convexity of $\mathcal{X}$, $\tilde{\mathbf{x}}(i) \in \mathcal{X}$, and the preceding information constraint shows that it is measurable with respect to $(y_0, \dots, y_{i-1}, \xi)$. For $t \in \mathcal{I}_i$, the aggregate loss is $\sum_{v=1}^{N} f_{t,v} = Sh_i$, while it is zero after round MK. Therefore, by convexity of h_i,

$$\sum_{t=1}^{T} \sum_{v=1}^{N} f_{t,v}\big(\mathbf{x}_1(t)\big) = S \sum_{i=0}^{M-1} \sum_{t \in \mathcal{I}_i} h_i\big(\mathbf{x}_1(t)\big) \ge SK \sum_{i=0}^{M-1} h_i\big(\tilde{\mathbf{x}}(i)\big),$$

whereas for every fixed comparator $\mathbf{x} \in \mathcal{X}$, $\sum_{t=1}^{T} \sum_{v=1}^{N} f_{t,v}(\mathbf{x}) = SK \sum_{i=0}^{M-1} h_i(\mathbf{x})$. Consequently,

$$\mathrm{Reg}_{T,1} \ge SK \left(\sum_{i=0}^{M-1} h_i\big(\tilde{\mathbf{x}}(i)\big) - \min_{\mathbf{x} \in \mathcal{X}} \sum_{i=0}^{M-1} h_i(\mathbf{x}) \right).$$

The maps $(y_0, \dots, y_{i-1}, \xi) \mapsto \tilde{\mathbf{x}}(i)$ satisfy the hypotheses of Lemma 6. Hence there exist outcomes $y_0, \dots, y_{M-1}$, and therefore a deterministic sequence of local losses, such that

$$\mathbb{E}[\mathrm{Reg}_{T,1}] \ge SK \cdot \frac{d}{2\alpha} \log \frac{n}{2\pi e\, d}. \tag{27}$$

It remains to express this bound in terms of T, N, and the spectral gap. Since $N = 4K + 2 \ge 6$, $SK = (2K + 3)K = \frac{(N+4)(N-2)}{8} \ge \frac{N^2}{8}$. The assumption $T \ge (\pi e)^2 N d^3$ implies $T \ge Nd \ge N$. Since $K \le N/4$, we have $Z = \lfloor T/K \rfloor \ge T/K - 1 \ge 4T/N - 1 \ge 3T/N$, and therefore $n = \lfloor Z/d \rfloor \ge Z/d - 1 \ge 3T/(Nd) - 1 \ge 2T/(Nd)$. Writing $\kappa \triangleq T/(Nd)$, the same assumption gives $\kappa \ge (\pi e)^2 d^2$, hence $\pi e\, d \le \sqrt{\kappa}$ and $n/(2\pi e\, d) \ge \kappa/(\pi e\, d) \ge \sqrt{\kappa}$. Substituting these estimates into Eq. (27) yields

$$\mathbb{E}[\mathrm{Reg}_{T,1}] \ge \frac{N^2}{8} \cdot \frac{d}{2\alpha} \cdot \frac{1}{2} \log \frac{T}{Nd} = \frac{N^2 d}{32\alpha} \log \frac{T}{Nd}.$$

Finally, because the cycle has $d_{\max} = 2$, $P = I - \frac{1}{6} L_{\mathcal{G}}$. The eigenvalues of $L_{\mathcal{G}}$ are $2 - 2\cos(2\pi k/N)$ for $k = 0, \dots, N-1$, so the eigenvalues of P are $1 - \frac{1}{3}(1 - \cos(2\pi k/N)) \in [\frac{1}{3}, 1]$. Thus P is symmetric, positive semidefinite, and doubly stochastic, with $P_{uv} = \frac{1}{6}$ for $\{u, v\} \in E$ and $P_{uv} = 0$ for non-adjacent $u \ne v$, so it satisfies Definition 1. Its singular values therefore coincide with its eigenvalues, and

$$1 - \sigma_2(P) = \frac{1}{3}\Big(1 - \cos\frac{2\pi}{N}\Big) = \frac{2}{3} \sin^2 \frac{\pi}{N} \ge \frac{8}{3N^2},$$

where we used $\sin x \ge \frac{2x}{\pi}$ for $x \in [0, \frac{\pi}{2}]$. Hence $N \ge \sqrt{8/3}\,(1 - \sigma_2(P))^{-1/2}$. Combining this with the preceding regret lower bound and using $\sqrt{8/3}/32 \ge 1/20$ gives

$$\mathbb{E}[\mathrm{Reg}_{T,1}] \ge \frac{N^2 d}{32\alpha} \log \frac{T}{Nd} \ge \frac{Nd}{20\,\alpha\sqrt{1 - \sigma_2(P)}} \log \frac{T}{Nd}.$$

This proves the theorem. □

C Omitted Details from Section 4

In this section, we provide the omitted details from Section 4. We first introduce the notation used in the analysis, then establish three auxiliary lemmas, and finally combine the delay and gossip estimates to prove Theorem 3. As in Section 4, we assume that T/L is an integer.

Let $\mathcal{T}_z = \{(z-1)L+1, \dots, zL\}$ for $z \in [T/L]$. For every round t and agent j, define $f_{t,j}(\mathbf{x}) \triangleq \frac{1}{2}(\langle \mathbf{h}_{t,j}, \mathbf{x}\rangle - y_{t,j})^2$ and $\ell_t(\mathbf{x}) \triangleq \sum_{j=1}^N f_{t,j}(\mathbf{x})$. For $|y| \le Y$, let $\phi_y(u) \triangleq \frac{1}{2}(\Pi_{[-Y,Y]}(u) - y)^2$. For each block s, stack the LN feature vectors $\{\mathbf{h}_{\tau,j}\}_{\tau\in\mathcal{T}_s, j\in[N]}$, in an arbitrary fixed order, into the rows of $B_s \in \mathbb{R}^{LN\times d}$, and stack the corresponding labels into $\mathbf{y}_s \in \mathbb{R}^{LN}$. Then $\sum_{j=1}^N C_j^0(s) = B_s^\top B_s$, $\sum_{j=1}^N Y_j^0(s) = B_s^\top \mathbf{y}_s$, and $\sum_{\tau\in\mathcal{T}_s} \ell_\tau(\mathbf{x}) = \frac{1}{2}\|B_s\mathbf{x} - \mathbf{y}_s\|_2^2$.

For every integer $k \le T/L$, define

$$A_k \triangleq \lambda I_d + \sum_{s=1}^k B_s^\top B_s, \qquad \mathbf{b}_k \triangleq \sum_{s=1}^k B_s^\top \mathbf{y}_s, \qquad F_k(\mathbf{x}) \triangleq \frac{\lambda}{2}\|\mathbf{x}\|_2^2 + \sum_{s=1}^k \sum_{\tau\in\mathcal{T}_s} \ell_\tau(\mathbf{x}). \tag{28}$$

Empty sums are understood as zero, so $A_k = \lambda I_d$, $\mathbf{b}_k = \mathbf{0}$, and $F_k(\mathbf{x}) = \frac{\lambda}{2}\|\mathbf{x}\|_2^2$ for $k \le 0$. Since $A_k \succeq \lambda I_d \succ 0$, F_k is a strictly convex quadratic with unique minimizer $\mathbf{u}_k \triangleq A_k^{-1}\mathbf{b}_k$ and minimum $M_k \triangleq F_k(\mathbf{u}_k)$. Thus $\mathbf{u}_k = \mathbf{0}$ and $M_k = 0$ for $k \le 0$. The quadratic expansion around the minimizer is exact: $F_k(\mathbf{x}) = M_k + \frac{1}{2}\|\mathbf{x} - \mathbf{u}_k\|_{A_k}^2$. For $k \ge 1$, define $\gamma_k \triangleq \operatorname{tr}(B_k A_{k-1}^{-1} B_k^\top)$ and $\delta_k \triangleq \log(\det A_k / \det A_{k-1})$, and set $\gamma_k = \delta_k = 0$ for $k \le 0$. Both are nonnegative.

The exact delayed predictor of Section 4 is $\mathbf{x}'(z) = A_{z-2}^{-1}\mathbf{b}_{z-2} = \mathbf{u}_{z-2}$. The decentralized predictor is

$$\mathbf{x}_i(z) = \widehat{A}_i(z)^{-1}\widehat{\mathbf{b}}_i(z), \qquad \widehat{A}_i(z) \triangleq \lambda I_d + N\sum_{s=1}^{z-2} \Pi_{\mathbb{S}_+^d}(C_i(s)), \qquad \widehat{\mathbf{b}}_i(z) \triangleq N\sum_{s=1}^{z-2} Y_i(s). \tag{29}$$

For a symmetric matrix $C = \sum_{k=1}^d \mu_k \mathbf{v}_k\mathbf{v}_k^\top$, $\Pi_{\mathbb{S}_+^d}(C) = \sum_{k=1}^d \max\{\mu_k, 0\}\mathbf{v}_k\mathbf{v}_k^\top$ is the Frobenius-norm projection onto the positive semidefinite cone. The gossip iterates remain symmetric because they are real linear combinations of symmetric matrices.

For a block z and $\mathbf{v} \in \mathbb{R}^d$, let $\Phi_z(\mathbf{v}) \triangleq \sum_{t\in\mathcal{T}_z}\sum_{j=1}^N \phi_{y_{t,j}}(\langle \mathbf{h}_{t,j}, \mathbf{v}\rangle)$. Since agent i uses the same model throughout block z,

$$\mathrm{Reg}_{T,i}^{\mathrm{OLR}}(\mathbf{x}) = \sum_{z=1}^{T/L} \Phi_z(\mathbf{x}_i(z)) - \sum_{t=1}^T\sum_{j=1}^N f_{t,j}(\mathbf{x}). \tag{30}$$

C.1 Auxiliary Lemmas

The first lemma records the accelerated-gossip estimate at a general accuracy level.

Lemma 7. *Let P be the communication matrix defined in Definition 1, let θ be chosen according to Eq. (3), and let $\rho \triangleq (1 - \frac{1}{\sqrt{2}})\sqrt{1-\sigma_2(P)}$. Consider arbitrary vectors $\mathbf{w}_1, \dots, \mathbf{w}_N \in \mathbb{R}^p$ satisfying $\|\mathbf{w}_i\|_2 \le R$. Let $W^0 = W^{-1} = [\mathbf{w}_1, \dots, \mathbf{w}_N]^\top$, $\overline{\mathbf{w}} = N^{-1}\sum_{i=1}^N \mathbf{w}_i$, and $\overline{W} = \mathbf{1}\overline{\mathbf{w}}^\top$. If $W^{k+1} = (1+\theta)PW^k - \theta W^{k-1}$ is run for $n \ge \log(\sqrt{14}/\varepsilon)/\rho$ iterations, where $\varepsilon \in (0,1]$, then $\|\mathbf{w}_i^n - \overline{\mathbf{w}}\|_2 \le 2\varepsilon R\sqrt{N}$ for every $i \in [N]$.*

Proof. The contraction used in the proof of Lemma 1 holds for every integer $n \ge 0$:

$$\|W^n - \overline{W}\|_F \le \sqrt{14}(1-\rho)^n\|W^0 - \overline{W}\|_F.$$

Since $\rho \in (0,1)$, $(1-\rho)^n \le e^{-\rho n} \le \varepsilon/\sqrt{14}$. Moreover, $\|\overline{\mathbf{w}}\|_2 \le R$, hence $\|W^0 - \overline{W}\|_F \le 2R\sqrt{N}$. The norm of every row is bounded by the Frobenius norm, which gives $\|\mathbf{w}_i^n - \overline{\mathbf{w}}\|_2 \le 2\varepsilon R\sqrt{N}$. □

The next lemma collects the elementary properties of the clipped loss.

Lemma 8. *Let $Y \geq 0$ and $y \in [-Y, Y]$. Then, for all $u, v \in \mathbb{R}$, $0 \leq \phi_y(u) \leq \min\{2Y^2, \frac{1}{2}(u-y)^2\}$ and $|\phi_y(u) - \phi_y(v)| \leq 2Y|u - v|$.*

Proof. Both $\Pi_{[-Y,Y]}(u)$ and y lie in $[-Y, Y]$, so $|\Pi_{[-Y,Y]}(u) - y| \leq 2Y$ and therefore $\phi_y(u) \leq 2Y^2$. Since projection onto $[-Y, Y]$ is non-expansive and $\Pi_{[-Y,Y]}(y) = y$, we also have $|\Pi_{[-Y,Y]}(u) - y| \leq |u - y|$, which implies $\phi_y(u) \leq \frac{1}{2}(u - y)^2$. Finally,

$$|\phi_y(u) - \phi_y(v)| = \frac{1}{2}|\Pi_{[-Y,Y]}(u) - \Pi_{[-Y,Y]}(v)||\Pi_{[-Y,Y]}(u) + \Pi_{[-Y,Y]}(v) - 2y| \leq 2Y|u - v|.$$

□

The following lemma describes the change in a ridge objective when a batch of squared losses is added.

Lemma 9. *Let $A \in \mathbb{R}^{d\times d}$ be symmetric positive definite, $\mathbf{u} \in \mathbb{R}^d$, $m \in \mathbb{R}$, and $F(\mathbf{v}) = m + \frac{1}{2}\|\mathbf{v} - \mathbf{u}\|_A^2$. Let $B \in \mathbb{R}^{n\times d}$, $\mathbf{y} \in \mathbb{R}^n$, and $G(\mathbf{v}) = F(\mathbf{v}) + \frac{1}{2}\|B\mathbf{v} - \mathbf{y}\|_2^2$. Define $K \triangleq BA^{-1}B^\top \succeq 0$ and $\mathbf{r} \triangleq B\mathbf{u} - \mathbf{y}$. Then:*

(a) G has the unique minimizer $\mathbf{u}^+ = \mathbf{u} - (A + B^\top B)^{-1}B^\top\mathbf{r}$, and

$$\min_{\mathbf{v}} G(\mathbf{v}) - m = \frac{1}{2}\mathbf{r}^\top(I_n + K)^{-1}\mathbf{r} \geq \frac{\|\mathbf{r}\|_2^2}{2(1 + \operatorname{tr} K)}.$$

(b) $\|\mathbf{u}^+ - \mathbf{u}\|_{A+B^\top B}^2 = \mathbf{r}^\top K(I_n + K)^{-1}\mathbf{r} \leq 2\operatorname{tr}(K)(\min_{\mathbf{v}} G(\mathbf{v}) - m)$.

(c) $\log(1 + \operatorname{tr} K) \leq \log\det(I_n + K) = \log(\det(A + B^\top B)/\det A)$.

Proof. Both $A + B^\top B$ and $I_n + K$ are positive definite. From $B^\top(I_n + K) = (A + B^\top B)A^{-1}B^\top$, we obtain $(A + B^\top B)^{-1}B^\top = A^{-1}B^\top(I_n + K)^{-1}$ and hence $B(A + B^\top B)^{-1}B^\top = K(I_n + K)^{-1} = I_n - (I_n + K)^{-1}$.

Writing $\mathbf{v} = \mathbf{u} + \mathbf{w}$ gives $G(\mathbf{u} + \mathbf{w}) - m = \frac{1}{2}\mathbf{w}^\top(A + B^\top B)\mathbf{w} + \mathbf{w}^\top B^\top\mathbf{r} + \frac{1}{2}\|\mathbf{r}\|_2^2$. The unique minimizer is $\mathbf{w}^+ = -(A + B^\top B)^{-1}B^\top\mathbf{r}$, and substituting it gives $\min G - m = \frac{1}{2}\mathbf{r}^\top(I_n + K)^{-1}\mathbf{r}$. Since $\lambda_{\max}(I_n + K) = 1 + \|K\|_{\mathrm{op}} \leq 1 + \operatorname{tr} K$, the lower bound in (a) follows.

Using the same push-through identity, $\|\mathbf{u}^+ - \mathbf{u}\|_{A+B^\top B}^2 = \mathbf{r}^\top K(I_n + K)^{-1}\mathbf{r}$. Because K and $(I_n + K)^{-1}$ share an eigenbasis, $K(I_n + K)^{-1} \preceq \|K\|_{\mathrm{op}}(I_n + K)^{-1} \preceq \operatorname{tr}(K)(I_n + K)^{-1}$. Combining this with part (a) proves (b).

Finally, if $\mu_1, \ldots, \mu_n \geq 0$ are the eigenvalues of K, then $\det(I_n + K) = \prod_k(1 + \mu_k) \geq 1 + \sum_k \mu_k = 1 + \operatorname{tr} K$. Sylvester's determinant identity gives $\det(I_n + BA^{-1}B^\top) = \det(I_d + A^{-1}B^\top B) = \det(A + B^\top B)/\det A$, proving (c). □

C.2 Delay Error

We next bound the regret of the exact delayed predictor. Relative to the ridge solution after block $z-1$, the predictor $\mathbf{x}'(z) = \mathbf{u}_{z-2}$ misses the statistics of block $z-1$; as a blockwise-fixed predictor, it also cannot use the labels in the current block. Both effects can be charged to increments of the regularized minimum.

Lemma 10. *For every comparator $\mathbf{x} \in \mathbb{R}^d$,*

$$\sum_{z=1}^{T/L} \Phi_z(\mathbf{x}'(z)) - \sum_{t=1}^{T}\sum_{j=1}^{N} f_{t,j}(\mathbf{x}) \leq \frac{\lambda}{2}\|\mathbf{x}\|_2^2 + (2 + 16\sqrt{2})Y^2LNd\log\left(1 + \frac{NTH^2}{\lambda d}\right).$$

Proof. Let $c_0 \triangleq 2Y^2LN$. Since each block contains LN examples, Lemma 8 gives $0 \leq \Phi_z(\mathbf{v}) \leq c_0$ for every $\mathbf{v}$ and z. Define $\Delta_k \triangleq M_k - M_{k-1}$ and set $\Delta_k = 0$ for $k \leq 0$. Since $F_k \geq F_{k-1}$

pointwise, $\Delta_k \geq 0$. Moreover, $\sum_{z=1}^{T/L} \Delta_z = M_{T/L} \leq F_{T/L}(\mathbf{x}) = \frac{\lambda}{2}\|\mathbf{x}\|_2^2 + \sum_{t=1}^{T}\sum_{j=1}^{N} f_{t,j}(\mathbf{x})$. Thus it remains to control $\Phi_z(\mathbf{u}_{z-2}) - \Delta_z$.

Applying Lemma 9(a) to the update from F_{z-1} to F_z gives $\Delta_z \geq \|B_z\mathbf{u}_{z-1} - \mathbf{y}_z\|_2^2/[2(1+\gamma_z)] \geq \Phi_z(\mathbf{u}_{z-1})/(1+\gamma_z)$. Consequently,

$$\Phi_z(\mathbf{u}_{z-1}) - \Delta_z \leq c_0 \frac{\gamma_z}{1+\gamma_z} \leq c_0 \log(1+\gamma_z) \leq c_0\delta_z,$$

where the last inequality follows from Lemma 9(c).

The update from F_{z-2} to F_{z-1} and Lemma 9(b) similarly give $\|\mathbf{u}_{z-1} - \mathbf{u}_{z-2}\|_{A_{z-1}}^2 \leq 2\gamma_{z-1}\Delta_{z-1}$; this is also valid for $z = 1$ because both sides vanish. By the Lipschitz property in Lemma 8 and Cauchy–Schwarz,

$$\begin{aligned}\Phi_z(\mathbf{u}_{z-2}) - \Phi_z(\mathbf{u}_{z-1}) &\leq 2Y\|\mathbf{u}_{z-2} - \mathbf{u}_{z-1}\|_{A_{z-1}} \sum_{t\in\mathcal{T}_z}\sum_{j=1}^{N} \|\mathbf{h}_{t,j}\|_{A_{z-1}^{-1}} \\ &\leq 2Y\|\mathbf{u}_{z-2} - \mathbf{u}_{z-1}\|_{A_{z-1}} \sqrt{LN\gamma_z} \\ &\leq 2Y\sqrt{2LN\gamma_z\gamma_{z-1}\Delta_{z-1}}.\end{aligned}$$

Combining the preceding two bounds, and also using $\Phi_z(\mathbf{u}_{z-2}) - \Delta_z \leq c_0$, yields

$$\Phi_z(\mathbf{u}_{z-2}) - \Delta_z \leq c_0\delta_z + \min\left\{c_0,\, 2Y\sqrt{2LN\gamma_z\gamma_{z-1}\Delta_{z-1}}\right\}. \tag{31}$$

To sum this bound without paying directly for a potentially large Δ_{z-1}, let $\mathcal{J} \triangleq \{z : \Delta_z \geq 2c_0\}$, $\mathcal{I} \triangleq \{z \notin \mathcal{J} : z-1 \in \mathcal{J}\}$, and $\mathcal{R} \triangleq [T/L] \setminus (\mathcal{J} \cup \mathcal{I})$. If $z \in \mathcal{J}$, then $\Phi_z(\mathbf{u}_{z-2}) - \Delta_z \leq -c_0$; if $z \in \mathcal{I}$, the same quantity is at most c_0. If $z \in \mathcal{R}$, then either $z = 1$ or $\Delta_{z-1} < 2c_0$, and Eq. (31) gives $\Phi_z(\mathbf{u}_{z-2}) - \Delta_z \leq \omega_z$, where

$$\omega_z \triangleq c_0\delta_z + \min\left\{c_0,\, 4\sqrt{2}\,Y^2LN\sqrt{\gamma_z\gamma_{z-1}}\right\}.$$

The map $z \mapsto z-1$ is injective from $\mathcal{I}$ into $\mathcal{J}$. Since $\omega_z \geq 0$, the positive contribution of $\mathcal{I}$ is canceled by the negative contribution of the corresponding indices in $\mathcal{J}$, and therefore

$$\sum_{z=1}^{T/L}\big(\Phi_z(\mathbf{u}_{z-2}) - \Delta_z\big) \leq \sum_{z=1}^{T/L}\omega_z. \tag{32}$$

For $a, b \geq 0$, $\min\{1, \sqrt{ab}\} \leq \min\{1,a\} + \min\{1,b\}$ and $\min\{1,a\} \leq 2\log(1+a)$. Together with $\log(1+\gamma_k) \leq \delta_k$ from Lemma 9(c), these inequalities imply

$$\min\left\{c_0,\, 4\sqrt{2}\,Y^2LN\sqrt{\gamma_z\gamma_{z-1}}\right\} \leq 8\sqrt{2}\,Y^2LN(\delta_z + \delta_{z-1}).$$

Hence, using $\delta_0 = 0$,

$$\sum_{z=1}^{T/L}\omega_z \leq (2+16\sqrt{2})Y^2LN\log\frac{\det A_{T/L}}{\lambda^d}. \tag{33}$$

Finally, $\operatorname{tr}(A_{T/L}) = \lambda d + \sum_{t=1}^{T}\sum_{j=1}^{N}\|\mathbf{h}_{t,j}\|_2^2 \leq \lambda d + NTH^2$. The arithmetic–geometric mean inequality for the eigenvalues of $A_{T/L}$ gives

$$\log\frac{\det A_{T/L}}{\lambda^d} \leq d\log\frac{\operatorname{tr}(A_{T/L})}{\lambda d} \leq d\log\left(1 + \frac{NTH^2}{\lambda d}\right).$$

Using $\mathbf{x}'(z) = \mathbf{u}_{z-2}$ and combining this with Eq. (32), Eq. (33), and the telescoping bound on $\sum_z \Delta_z$ proves the lemma. □

C.3 GOSSIP ERROR

We now control the difference between the decentralized predictor $\mathbf{x}_i(z)$ and the exact delayed predictor $\mathbf{x}'(z)$. Both the covariance and label-feature statistics are perturbed by gossip, so the model error contains a direct perturbation of the right-hand side and a matrix-inverse perturbation.

Lemma 11. *Under the conditions of Theorem 3, for every agent $i \in [N]$, block $z \in [T/L]$, round $t \in \mathcal{T}_z$, and agent $j \in [N]$,*

$$|\langle \mathbf{h}_{t,j}, \mathbf{x}_i(z) - \mathbf{x}'(z)\rangle| \le \frac{2YH}{\lambda TN}\|\mathbf{h}_{t,j}\|_2.$$

Consequently,

$$\sum_{z=1}^{T/L}\big(\Phi_z(\mathbf{x}_i(z)) - \Phi_z(\mathbf{x}'(z))\big) \le \frac{4Y^2H^2}{\lambda}.$$

Proof. For $z \le 2$, both predictors are zero. Fix $z \ge 3$ and $s \in [z-2]$. The statistics $C_j^0(s) = C_j^{-1}(s)$ and $Y_j^0(s) = Y_j^{-1}(s)$ are formed at the end of block s and undergo exactly L accelerated-gossip iterations during block $s+1$, so $C_i(s) = C_i^L(s)$ and $Y_i(s) = Y_i^L(s)$ are available before the model for every block $z \ge s+2$ is computed.

The local statistics satisfy $\|Y_j^0(s)\|_2 \le LYH$ and $\|C_j^0(s)\|_F \le LH^2$. Set $\varepsilon \triangleq [T^3N^{7/2}(1 + H^2/\lambda)]^{-1} \in (0,1]$. The choice of L in Theorem 3 ensures the iteration requirement in Lemma 7. With $\overline{Y}(s) \triangleq N^{-1}B_s^\top \mathbf{y}_s$ and $\overline{C}(s) \triangleq N^{-1}B_s^\top B_s$, the lemma gives

$$\|Y_i(s) - \overline{Y}(s)\|_2 \le \frac{2LYH}{T^3N^3(1+H^2/\lambda)}, \qquad \|C_i(s) - \overline{C}(s)\|_F \le \frac{2LH^2}{T^3N^3(1+H^2/\lambda)}. \tag{34}$$

Because $\overline{C}(s) \succeq 0$, non-expansiveness of the Frobenius projection onto $\mathbb{S}_+^d$ implies $\|\Pi_{\mathbb{S}_+^d}(C_i(s)) - \overline{C}(s)\|_F \le \|C_i(s) - \overline{C}(s)\|_F$.

By Eq. (28), $A_{z-2} = \lambda I_d + N\sum_{s=1}^{z-2}\overline{C}(s)$ and $\mathbf{b}_{z-2} = N\sum_{s=1}^{z-2}\overline{Y}(s)$. There are at most T/L completed blocks, so summing Eq. (34) and using $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_F$ gives

$$\|\widehat{\mathbf{b}}_i(z) - \mathbf{b}_{z-2}\|_2 \le \frac{2YH}{T^2N^2(1+H^2/\lambda)}, \qquad \|\widehat{A}_i(z) - A_{z-2}\|_{\mathrm{op}} \le \frac{2H^2}{T^2N^2(1+H^2/\lambda)}. \tag{35}$$

Also, $\|\mathbf{b}_{z-2}\|_2 \le NTYH$. Both $\widehat{A}_i(z)$ and A_{z-2} dominate λI_d, and hence both inverse operator norms are at most $1/\lambda$.

Using $\widehat{A}^{-1} - A^{-1} = -\widehat{A}^{-1}(\widehat{A} - A)A^{-1}$, we have

$$\mathbf{x}_i(z) - \mathbf{x}'(z) = \widehat{A}_i(z)^{-1}\big(\widehat{\mathbf{b}}_i(z) - \mathbf{b}_{z-2}\big) - \widehat{A}_i(z)^{-1}\big(\widehat{A}_i(z) - A_{z-2}\big)A_{z-2}^{-1}\mathbf{b}_{z-2}.$$

Therefore, by Eq. (35),

$$\begin{aligned}|\langle \mathbf{h}_{t,j}, \mathbf{x}_i(z) - \mathbf{x}'(z)\rangle| &\le \|\mathbf{h}_{t,j}\|_2\left(\frac{2YH}{\lambda T^2N^2(1+H^2/\lambda)} + \frac{2YH^3}{\lambda^2TN(1+H^2/\lambda)}\right)\\ &= \frac{2YH\|\mathbf{h}_{t,j}\|_2}{\lambda TN(1+H^2/\lambda)}\left(\frac{1}{TN} + \frac{H^2}{\lambda}\right) \le \frac{2YH}{\lambda TN}\|\mathbf{h}_{t,j}\|_2.\end{aligned} \tag{36}$$

The last inequality uses $TN \ge 1$.

Finally, Lemma 8, Eq. (36), and $\|\mathbf{h}_{t,j}\|_2 \le H$ give

$$\sum_{z=1}^{T/L}\big(\Phi_z(\mathbf{x}_i(z)) - \Phi_z(\mathbf{x}'(z))\big) \le 2Y\sum_{t=1}^{T}\sum_{j=1}^{N}|\langle \mathbf{h}_{t,j}, \mathbf{x}_i(z(t)) - \mathbf{x}'(z(t))\rangle| \le \frac{4Y^2H^2}{\lambda},$$

where $z(t)$ denotes the block containing round t. □

C.4 Proof of Theorem 3

We are now ready to prove Theorem 3 by combining Lemma 10 and Lemma 11. We include the statement of Theorem 3 for completeness.

Theorem 3. *Let P be the communication matrix defined in Definition 1. In the D-OLR problem, if every agent $i \in [N]$ executes Algorithm 2 with $\lambda > 0$, θ chosen according to Eq. (3), and $L = \left\lceil \frac{\sqrt{2}}{(\sqrt{2}-1)\sqrt{1-\sigma_2(P)}} \log\big(\sqrt{14}\,T^3 N^{7/2}(1+H^2/\lambda)\big) \right\rceil$, then for any agent $i \in [N]$ and comparator $\mathbf{x} \in \mathbb{R}^d$, the regret satisfies $\mathrm{Reg}_{T,i}^{\mathrm{OLR}}(\mathbf{x}) = \mathcal{O}\big(\lambda\|\mathbf{x}\|_2^2 + Y^2H^2/\lambda + Y^2LNd\log(1+NTH^2/(\lambda d))\big)$.*

Proof. Fix any agent $i \in [N]$ and comparator $\mathbf{x} \in \mathbb{R}^d$. By Eq. (30),

$$\mathrm{Reg}_{T,i}^{\mathrm{OLR}}(\mathbf{x}) = \left(\sum_{z=1}^{T/L} \Phi_z(\mathbf{x}'(z)) - \sum_{t=1}^{T}\sum_{j=1}^{N} f_{t,j}(\mathbf{x})\right) + \sum_{z=1}^{T/L}\big(\Phi_z(\mathbf{x}_i(z)) - \Phi_z(\mathbf{x}'(z))\big).$$

Applying Lemma 10 to the first term and Lemma 11 to the second gives

$$\mathrm{Reg}_{T,i}^{\mathrm{OLR}}(\mathbf{x}) \le \frac{\lambda}{2}\|\mathbf{x}\|_2^2 + \frac{4Y^2H^2}{\lambda} + (2+16\sqrt{2})Y^2LNd\log\left(1+\frac{NTH^2}{\lambda d}\right),$$

which proves Theorem 3. Substituting the stated value of L gives the claimed dependence on the spectral gap. □