Distributed Learning with Sublinear Communication

Jayadev Acharya
Cornell University
acharya@cornell.edu

Christopher De Sa
Cornell University
cdesa@cs.cornell.edu

Dylan J. Foster
MIT
dylanf@mit.edu

Karthik Sridharan
Cornell University
sridharan@cs.cornell.edu

Abstract

In distributed statistical learning, N samples are split across m machines and a learner wishes to use minimal communication to learn as well as if the examples were on a single machine. This model has received substantial interest in machine learning due to its scalability and potential for parallel speedup. However, in high-dimensional settings, where the number examples is smaller than the number of features ("dimension"), the speedup afforded by distributed learning may be overshadowed by the cost of communicating a single example. This paper investigates the following question: When is it possible to learn a d-dimensional model in the distributed setting with total communication sublinear in d?

Starting with a negative result, we observe that for learning ℓ_1-bounded or sparse linear models, no algorithm can obtain optimal error until communication is linear in dimension. Our main result is that that by slightly relaxing the standard boundedness assumptions for linear models, we can obtain distributed algorithms that enjoy optimal error with communication *logarithmic* in dimension. This result is based on a family of algorithms that combine mirror descent with randomized sparsification/quantization of iterates, and extends to the general stochastic convex optimization model.

1 Introduction

In statistical learning, a learner receives examples $z_1, \ldots, z_N$ i.i.d. from an unknown distribution $\mathcal{D}$. Their goal is to output a hypothesis $\hat{h} \in \mathcal{H}$ that minimizes the prediction error $L_{\mathcal{D}}(h) \coloneqq \mathbb{E}_{z \sim \mathcal{D}} \ell(h, z)$, and in particular to guarantee that *excess risk* of the learner is small, i.e.

$$L_{\mathcal{D}}(\hat{h}) - \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \le \varepsilon(\mathcal{H}, N), \tag{1}$$

where $\varepsilon(\mathcal{H}, N)$ is a decreasing function of N. This paper focuses on *distributed* statistical learning. Here, the N examples are split evenly across m machines, with $n \coloneqq N/m$ examples per machine, and the learner wishes to achieve an excess risk guarantee such as (1) with minimal overhead in computation or communication.

Distributed learning has been the subject of extensive investigation due to its scalability for processing massive data: We may wish to efficiently process datasets that are spread across multiple data-centers, or we may want to distribute data across multiple machines to allow for parallelization of learning procedures. The question of parallelizing computation via distributed learning is a well-explored problem (Bekkerman et al., 2011; Recht et al., 2011; Dekel et al., 2012; Chaturapruek et al., 2015). However, one drawback that limits the practical viability of these approaches is that the communication cost amongst machines may overshadow gains in parallel speedup (Bijral et al., 2016). Indeed, for high-dimensional statistical inference tasks where N could be much smaller than the dimension d, or in modern deep learning models where the number of

model parameters exceeds the number of examples (e.g. He et al. (2016)), communicating a single gradient or sending the raw model parameters between machines constitutes a significant overhead.

Algorithms with reduced communication complexity in distributed learning have received significant recent development (Seide et al., 2014; Alistarh et al., 2017; Zhang et al., 2017; Suresh et al., 2017; Bernstein et al., 2018; Tang et al., 2018), but typical results here take as a given that when gradients or examples live in d dimensions, communication will scale as $\Omega(d)$. Our goal is to revisit this tacit assumption and understand when it can be relaxed. We explore the question of *sublinear communication*:

Suppose a hypothesis class $\mathcal{H}$ has d parameters. When is it possible to achieve optimal excess risk for $\mathcal{H}$ in the distributed setting using $o(d)$ communication?

1.1   Sublinear Communication for Linear Models?

In this paper we focus on linear models, which are a special case of the general learning setup (1). We restrict to linear hypotheses of the form $h_w(x) = \langle w, x \rangle$ where $w, x \in \mathbb{R}^d$ and write $\ell(h_w, z) = \phi(\langle w, x \rangle, y)$, where $\phi(\cdot, y)$ is a fixed link function and $z = (x, y)$. We overload notation slightly and write

$$L_\mathcal{D}(w) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \, \phi(\langle w, x \rangle, y). \tag{2}$$

The formulation captures standard learning tasks such as square loss regression, where $\phi(\langle w, x \rangle, y) = (\langle w, x \rangle - y)^2$, logistic regression, where $\phi(\langle w, x \rangle, y) = \log(1 + e^{-y\langle w,x \rangle})$, and classification with surrogate losses such as the hinge loss, where $\phi(\langle w, x \rangle, y) = \max\{1 - \langle w, x \rangle \cdot y, 0\}$.

Our results concern the communication complexity of learning for linear models in the ℓ_p/ℓ_q-*bounded setup*: weights belong to $\mathcal{W}_p := \{w \in \mathbb{R}^d \mid \|w\|_p \le B_p\}$ and feature vectors belong to $\mathcal{X}_q := \{x \in \mathbb{R}^d \mid \|x\|_q \le R_q\}$.[1] This setting is a natural starting point to investigate sublinear-communication distributed learning because *learning is possible even when $N \ll d$.*

Consider the case where p and q are dual, i.e. $\frac{1}{p} + \frac{1}{q} = 1$, and where ϕ is 1-Lipschitz. Here it is well known (Zhang, 2002; Kakade et al., 2009) that whenever $q \ge 2$, the optimal sample complexity for learning, which is achieved by choosing the learner's weights $\widehat{w}$ using empirical risk minimization (ERM), is

$$L_\mathcal{D}(\widehat{w}) - \inf_{w \in \mathcal{W}_p} L_\mathcal{D}(w) = \Theta\left(\sqrt{\frac{B_p^2 R_q^2 C_q}{N}}\right), \tag{3}$$

where $C_q = q - 1$ for finite q and $C_\infty = \log d$, or in other words

$$L_\mathcal{D}(\widehat{w}) - \inf_{w \in \mathcal{W}_1} L_\mathcal{D}(w) = \Theta\left(\sqrt{\frac{B_1^2 R_\infty^2 \log d}{N}}\right). \tag{4}$$

We see that when $q < \infty$ the excess risk for the dual ℓ_p/ℓ_q setting is *independent of dimension* so long as the norm bounds B_p and R_q are held constant, and that even in the ℓ_1/ℓ_∞ case there is only a mild logarithmic dependence. Hence, we can get nontrivial excess risk even when the number of examples N is arbitrarily small compared to the dimension d. This raises the intriguing question: *Given that we can obtain nontrivial excess risk when $N \ll d$, can we obtain nontrivial excess risk when* communication *is sublinear in d?*

To be precise, we would like to develop algorithms that achieve (3)/(4) with total bits of communication $\mathrm{poly}(N, m, \log d)$, permitting also $\mathrm{poly}(B_p, R_q)$ dependence. The prospect of such a guarantee is exciting because—in light of the discussion above—as this would imply that we can obtain nontrivial excess risk with fewer bits of total communication than are required to naively send a *single feature vector*.

---

[1]Recall the definition of the ℓ_p norm: $\|w\|_p = \left(\sum_{i=1}^d |w_i|^p\right)^{1/p}$.

1.2 Contributions

We provide new communication-efficient distributed learning algorithms and lower bounds for ℓ_p/ℓ_q-bounded linear models, and more broadly, stochastic convex optimization. We make the following observations:

- For ℓ_2/ℓ_2-bounded linear models, sublinear communication *is* achievable, and is obtained by using a derandomized Johnson-Lindenstrauss transform to compress examples and weights.

- For ℓ_1/ℓ_∞-bounded linear models, no distributed algorithm can obtain optimal excess risk until communication is *linear* in dimension.

These observations lead to our main result. We show that by relaxing the ℓ_1/ℓ_∞-boundedness assumption and instead learning ℓ_1/ℓ_q-bounded models for a constant $q < \infty$, one unlocks a plethora of new algorithmic tools for sublinear distributed learning:

1. We give an algorithm with optimal rates matching (3), with communication $\mathrm{poly}(N, m^q, \log d)$.

2. We extend the sublinear-communication algorithm to give refined guarantees, including instance-dependent *small loss* bounds for smooth losses, *fast rates* for strongly convex losses, and optimal rates for matrix learning problems.

Our main algorithm is a distributed version of mirror descent that uses randomized sparsification of weight vectors to reduce communication. Beyond learning in linear models, the algorithm enjoys guarantees for the more general distributed stochastic convex optimization model.

To elaborate on the fast rates mentioned above, another important case where learning is possible when $N \ll d$ is the sparse high-dimensional linear model setup central to compressed sensing and statistics. Here, the standard result is that when ϕ is strongly convex and the benchmark class consists of k-sparse linear predictors, i.e. $\mathcal{W}_0 := \left\{ w \in \mathbb{R}^d \mid \|w\|_0 \le k \right\}$, one can guarantee

$$L_{\mathcal{D}}(\widehat{w}) - \inf_{w \in \mathcal{W}_0} L_{\mathcal{D}}(w) = \Theta\left(\frac{k \log(d/k)}{N} \right). \tag{5}$$

With ℓ_∞-bounded features, no algorithm can obtain optimal excess risk for this setting until communication is linear in dimension, even under compressed sensing-style assumptions. When features are ℓ_q-bounded however, our general machinery gives optimal fast rates matching (5) under Lasso-style assumptions, with communication $\mathrm{poly}(N^q, \log d)$.

The remainder of the paper is organized as follows. In Section 2 we develop basic upper and lower bounds for the ℓ_2/ℓ_2 and ℓ_1/ℓ_∞-bounded settings. Then in Section 3 we shift to the ℓ_1/ℓ_q-bounded setting, where we introduce the family of sparsified mirror descent algorithms that leads to our main results and sketch the analysis.

1.3 Related Work

Much of the work in algorithm design for distributed learning and optimization does not explicitly consider the number of bits used in communication per messages, and instead tries to make communication efficient via other means, such as decreasing the communication frequency or making learning robust to network disruptions (Duchi et al., 2012; Zhang et al., 2012). Other work reduces the number of bits of communication, but still requires that this number be linear in the dimension d. One particularly successful line of work in this vein is low-precision training, which represents the numbers used for communication and elsewhere within

the algorithm using few bits (Alistarh et al., 2017; Zhang et al., 2017; Seide et al., 2014; Bernstein et al., 2018; Tang et al., 2018; Stich et al., 2018; Alistarh et al., 2018). Although low-precision methods have seen great success and adoption in neural network training and inference, low-precision methods are fundamentally limited to use bits proportional to d; once they go down to one bit per number there is no additional benefit from decreasing the precision. Some work in this space tries to use sparsification to further decrease the communication cost of learning, either on its own or in combination with a low-precision representation for numbers (Alistarh et al., 2017; Wangni et al., 2018; Wang et al., 2018). While the majority of these works apply low-precision and sparsification to gradients, a number of recent works apply sparsification to model parameters (Tang et al., 2018; Stich et al., 2018; Alistarh et al., 2018); We also adopt this approach. The idea of sparsifying weights is not new (Shalev-Shwartz et al., 2010), but our work is the first to provably give communication logarithmic in dimension. To achieve this, our assumptions and analysis are quite a bit different from the results mentioned above, and we crucially use mirror descent, departing from the gradient descent approaches in Tang et al. (2018); Stich et al. (2018); Alistarh et al. (2018).

Lower bounds on the accuracy of learning procedures with limited memory and communication have been explored in several settings, including mean estimation, sparse regression, learning parities, detecting correlations, and independence testing (Shamir, 2014; Duchi et al., 2014; Garg et al., 2014; Steinhardt and Duchi, 2015; Braverman et al., 2016; Steinhardt et al., 2016; Acharya et al., 2018a,b; Raz, 2018; Han et al., 2018; Sahasranand and Tyagi, 2018; Dagan and Shamir, 2018; Dagan et al., 2019). In particular, the results of Steinhardt and Duchi (2015) and Braverman et al. (2016) imply that optimal algorithms for distributed sparse regression need communication much larger than the sparsity level under various assumptions on the number of machines and communication protocol.

2 Linear Models: Basic Results

In this section we develop basic upper and lower bounds for communication in ℓ_2/ℓ_2- and ℓ_1/ℓ_∞-bounded linear models. Our goal is to highlight some of the counterintuitive ways in which the interaction between the geometry of the weight vectors and feature vectors influences the communication required for distributed learning. In particular, we wish to underscore that the communication complexity of distributed learning and the statistical complexity of centralized learning do not in general coincide, and to motivate the ℓ_1/ℓ_q-boundedness assumption under which we derive communication-efficient algorithms in Section 3.

2.1 Preliminaries

We formulate our results in a distributed communication model following Shamir (2014). Recalling that $n = N/m$, the model is as follows.

- For machine $i = 1, \ldots, m$:
  - Receive n i.i.d. examples $S_i \coloneqq z_1^i, \ldots, z_n^i$.
  - Compute message $W_i = f_i(S_i\,;\,W_1, \ldots, W_{i-1})$, where W_i is at most b_i bits.
- Return $W = f(W_1, \ldots, W_m)$.

We refer to $\sum_{i=1}^m b_i$ as the *total communication*, and we refer to any protocol with $b_i \leq b \; \forall i$ as a (b, n, m) *protocol*. As a special case, this model captures a serial distributed learning setting where machines proceed one after another: Each machine does some computation on their data $z_1^i, \ldots, z_n^i$ and previous messages $W_1, \ldots, W_{i-1}$, then broadcasts their own message W_i to all subsequent machines, and the final model in

(1) is computed from W, either on machine m or on a central server. The model also captures protocols in which each machine independently computes a local estimator and sends it to a central server, which aggregates the local estimators to produce a final estimator (Zhang et al., 2012). All of our upper bounds have the serial structure above, and our lower bounds apply to any (b, n, m) protocol.

2.2 ℓ_2/ℓ_2-Bounded Models

In the ℓ_2/ℓ_2-bounded setting, we can achieve sample optimal learning with sublinear communication by using dimensionality reduction. The idea is to project examples into $k = \tilde{O}(N)$ dimensions using the Johnson-Lindenstrauss transform, then perform a naive distributed implementation of any standard learning algorithm in the projected space. Here we implement the approach using stochastic gradient descent.

The first machine picks a JL matrix $A \in \mathbb{R}^{k \times d}$ and communicates the identity of the matrix to the other $m-1$ machines. The JL matrix is chosen using the derandomized sparse JL transform of Kane and Nelson (2010), and its identity can be communicated by sending the random seed, which takes $O(\log(k/\delta) \cdot \log d)$ bits for confidence parameter δ. The dimension k and parameter δ are chosen as a function of N.

Now, each machine uses the matrix A to project its features down to k dimensions. Letting $x'_t = Ax_t$ denote the projected features, the first machine starts with a k-dimensional weight vector $u_1 = 0$ and performs the online gradient descent update (Zinkevich, 2003; Cesa-Bianchi and Lugosi, 2006) over its n projected samples as:

$$u_t \leftarrow u_{t-1} - \eta \nabla \phi(\langle u_t, x'_t \rangle, y_t),$$

where $\eta > 0$ is the learning rate. Once the first machine has passed over all its samples, it broadcasts the last iterate u_{n+1} as well the average $\sum_{s=1}^{n} u_s$, which takes $\tilde{O}(k)$ communication. The next machine machine performs the same sequence of gradient updates on its own data using u_{n+1} as the initialization, then passes its final iterate and the updated average to the next machine. This repeats until we arrive at the mth machine. The mth machine computes the k-dimensional vector $\widehat{u} := \frac{1}{N} \sum_{t=1}^{N} u_t$, and returns $\widehat{w} = A^\top \hat{u}$ as the solution.

Theorem 1. *When ϕ is L-Lipschitz and $k = \Omega(N \log(dN))$, the strategy above guarantees that*

$$\mathbb{E}_S \mathbb{E}_A \left[L_{\mathcal{D}}(\widehat{w}) \right] - \inf_{w \in \mathcal{W}_2} L_{\mathcal{D}}(w) \leq O\left(\sqrt{\frac{L^2 B_2^2 R_2^2}{N}} \right),$$

where $\mathbb{E}_S$ denotes expectation over samples and $\mathbb{E}_A$ denotes expectation over the algorithm's randomness. The total communication is $O(mN \log(dN) \log(LB_2R_2N) + m \log(dN) \log d)$ bits.

2.3 ℓ_1/ℓ_∞-Bounded Models: Model Compression

While the results for the ℓ_2/ℓ_2-bounded setting are encouraging, they are not useful in the common situation where features are dense. When features are ℓ_∞-bounded, Equation (4) shows that one can obtain nearly dimension-independent excess risk so long as they restrict to ℓ_1-bounded weights. This ℓ_1/ℓ_∞-bounded setting is particularly important because it captures the fundamental problem of learning from a finite hypothesis class, or *aggregation* (Tsybakov, 2003): Given a class $\mathcal{H}$ of $\{\pm 1\}$-valued predictors with $|\mathcal{H}| < \infty$ we can set $x = (h(z))_{h \in \mathcal{H}} \in \mathbb{R}^{|\mathcal{H}|}$, in which case (4) turns into the familiar finite class bound $\sqrt{\log|\mathcal{H}|/N}$ (Shalev-Shwartz and Ben-David, 2014). Thus, algorithms with communication sublinear in dimension for the ℓ_1/ℓ_∞ setting would lead to positive results in the general setting (1).

As first positive result in this direction, we observe that by using the well-known technique of *randomized sparsification* or *Maurey sparsification*, we can compress models to require only logarithmic communication while preserving excess risk.[2] The method is simple: Suppose we have a weight vector w that lies on the simplex Δ_d. We sample s elements of $[d]$ i.i.d. according to w and return the empirical distribution, which we will denote $Q^s(w)$. The empirical distribution is always s-sparse and can be communicated using at most $O(s \log (ed/s))$ bits when $s \le d$,[3] and it follows from standard concentration tools that by taking s large enough the empirical distribution will approximate the true vector w arbitrarily well.

The following lemma shows that Maurey sparsification indeed provides a dimension-independent approximation to the excess risk in the ℓ_1/ℓ_∞-bounded setting. It applies to a version of the Maurey technique for general vectors, which is given in Algorithm 1.

Lemma 1. Let $w \in \mathbb{R}^d$ be fixed and suppose features belong to $\mathcal{X}_\infty$. When ϕ is L-Lipschitz, Algorithm 1 guarantees that

$$\mathbb{E}\, L_{\mathcal{D}}(Q^s(w)) \le L_{\mathcal{D}}(w) + \left(\frac{2L^2 R_\infty^2 \|w\|_1^2}{s} \right)^{1/2}, \tag{6}$$

where the expectation is with respect to the algorithm's randomness. Furthermore, when ϕ is β-smooth[4] Algorithm 1 guarantees:

$$\mathbb{E}\, L_{\mathcal{D}}(Q^s(w)) \le L_{\mathcal{D}}(w) + \frac{\beta R_\infty^2 \|w\|_1^2}{s}. \tag{7}$$

The number of bits required to communicate $Q^s(w)$, including sending the scalar $\|w\|_1$ up to numerical precision, is at most $O(s \log (ed/s) + \log(LB_1 R_\infty s))$. Thus, if any single machine is able to find an estimator $\widehat{w}$ with good excess risk, they can communicate it to any other machine while preserving the excess risk with sublinear communication. In particular, to preserve the optimal excess risk guarantee in (4) for a Lipschitz loss such as absolute or hinge, the total bits of communication required is only $O(N + \log (LB_1 R_\infty N))$, which is indeed sublinear in dimension! For smooth losses (square, logistic), this improves further to only $O(\sqrt{N \log (ed/N)} + \log (LB_1 R_\infty N))$ bits.

Algorithm 1 (Maurey Sparsification).
Input: Weight vector $w \in \mathbb{R}^d$. Sparsity level s.
- Define $p \in \Delta_d$ via $p_i \propto |w_i|$.
- For $\tau = 1, \ldots, s$:
    - Sample index $i_\tau \sim p$.
- Return $Q^s(w) := \frac{\|w\|_1}{s} \sum_{\tau=1}^s \mathrm{sgn}(w_{i_\tau}) e_{i_\tau}$.

2.4 ℓ_1/ℓ_∞-Bounded Models: Impossibility

Alas, we have only shown that *if* we happen to find a good solution, we can send it using sublinear communication. If we have to start from scratch, is it possible to use Maurey sparsification to coordinate between

---

[2]We refer to the method as Maurey sparsification in reference to Maurey's early use of the technique in Banach spaces (Pisier, 1980), which predates its long history in learning theory (Jones, 1992; Barron, 1993; Zhang, 2002).

[3]That $O(s \log (ed/s))$ bits rather than, e.g., $O(s \log d)$ bits suffice is a consequence of the usual "stars and bars" counting argument. We expect one can bring the expected communication down further using an adaptive scheme such as Elias coding, as in Alistarh et al. (2017).

[4]A scalar function is said to be β-smooth if it has β-Lipschitz first derivative.

all machines to find a good solution?

Unfortunately, the answer is no: For the ℓ_1/ℓ_∞ bounded setting, in the extreme case where each machine has a single example, *no algorithm* can obtain a risk bound matching (4) until the number of bits b allowed per machine is (nearly) linear in d.

Theorem 2. *Consider the problem of learning with the linear loss in the $(b, 1, N)$ model, where risk is $L_\mathcal{D}(w) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[-y\langle w, x \rangle]$. Let the benchmark class be the ℓ_1 ball $\mathcal{W}_1$, where $B_1 = 1$. For any algorithm $\widehat{w}$ there exists a distribution $\mathcal{D}$ with $\|x\|_\infty \le 1$ and $|y| \le 1$ such that*

$$\Pr\left(L_\mathcal{D}(\widehat{w}) - \inf_{w \in \mathcal{W}_1} L_\mathcal{D}(w) \ge \tfrac{1}{16}\sqrt{\tfrac{d}{b} \cdot \tfrac{1}{N}} \wedge \tfrac{1}{2}\right) \ge \tfrac{1}{2}.$$

The lower bound also extends to the case of multiple examples per machine, albeit with a less sharp tradeoff.

Proposition 1. *Let m, n, and $\varepsilon > 0$ be fixed. In the setting of Theorem 2, any algorithm in the (b, n, m) protocol with $b \le O(d^{1-\varepsilon/2}/\sqrt{N})$ has excess risk at least $\Omega(\sqrt{d^\varepsilon/N})$ with constant probability.*

This lower bound follows almost immediately from reduction to the "hide-and-seek" problem of Shamir (2014). The weaker guarantee from Proposition 1 is a consequence of the fact that the lower bound for the hide-and-seek problem from Shamir (2014) is weaker in the multi-machine case.

The value of Theorem 2 and Proposition 1 is to rule out the possibility of obtaining optimal excess risk with communication polylogarithmic in d in the ℓ_1/ℓ_∞ setting, even when there are many examples per machine. This motivates the results of the next section, which show that for ℓ_1/ℓ_q-bounded models it is indeed possible to get polylogarithmic communication for any value of m.

One might hope that it is possible to circumvent Theorem 2 by making compressed sensing-type assumptions, e.g. assuming that the vector $w^\star$ is sparse and that restricted eigenvalue or a similar property is satisfied. Unfortunately, this is not the case.

Proposition 2. *Consider square loss regression in the $(b, 1, N)$ model. For any algorithm $\widehat{w}$ there exists a distribution $\mathcal{D}$ with the following properties:*

- *$\|x\|_\infty \le 1$ and $|y| \le 1$ with probability 1.*

- *$\Sigma \coloneqq \mathbb{E}[xx^\top] = I$, so that the population risk is 1-strongly convex, and in particular has restricted strong convexity constant 1.*

- *$w^\star \coloneqq \arg\min_{w:\|w\|_1 \le 1} L_\mathcal{D}(w)$ is 1-sparse.*

- *Until $b = \Omega(d)$, $\Pr\left(L_\mathcal{D}(\widehat{w}) - L_\mathcal{D}(w^\star) \ge \tfrac{1}{256}\left(\tfrac{d}{b} \cdot \tfrac{1}{N}\right) \wedge \tfrac{1}{4}\right) \ge \tfrac{1}{2}.$*

Moreover, any algorithm in the (b, n, m) protocol with $b \le O(d^{1-\varepsilon/2}/\sqrt{N})$ has excess risk at least $\Omega(d^\varepsilon/N)$ with constant probability.

That $\Omega(d)$ communication is required to obtain optimal excess risk for $m = N$ was proven in Steinhardt and Duchi (2015). The lower bound for general m is important here because it serves as a converse to the algorithmic results we develop for sparse regression in Section 3. It follows by reduction to hide-and-seek.[5]

The lower bound for sparse linear models does not rule out that sublinear learning is possible using additional statistical assumptions, e.g. that there are many examples on each machine *and* support recovery is possible. See Appendix B.2 for detailed discussion.

---

[5]Braverman et al. (2016) also prove a communication lower bound for sparse regression. Their lower bound applies for all values of m and for more sophisticated interactive protocols, but does not rule out the possibility of $\mathrm{poly}(N, m, \log d)$ communication.

3 Sparsified Mirror Descent

We now deliver on the promise outlined in the introduction and give new algorithms with logarithmic communication under an assumption we call ℓ_1/ℓ_q-*boundness*. The model for which we derive algorithms in this section is more general than the linear model setup (2) to which our lower bounds apply. We consider problems of the form

$$\operatorname*{minimize}_{w \in \mathcal{W}} \quad L_{\mathcal{D}}(w) \coloneqq \mathbb{E}_{z \sim \mathcal{D}}\, \ell(w, z), \tag{8}$$

where $\ell(\cdot, z)$ is convex, $\mathcal{W} \subseteq \mathcal{W}_1 = \{w \in \mathbb{R}^d \mid \|w\|_1 \leq B_1\}$ is a convex constraint set, and subgradients $\partial \ell(w, z)$ are assumed to belong to $\mathcal{X}_q = \{x \in \mathbb{R}^d \mid \|x\|_q \leq R_q\}$. This setting captures linear models with ℓ_1-bounded weights and ℓ_q-bounded features as a special case, but is considerably more general, since the loss can be any Lipschitz function of w.

We have already shown that one cannot expect sublinear-communication algorithms for ℓ_1/ℓ_∞-bounded models, and so the ℓ_q-boundedness of subgradients in (8) may be thought of as strengthening our assumption on the data generating process. That this is stronger follows from the elementary fact that $\|x\|_q \geq \|x\|_\infty$ for all q.

Statistical complexity and nontriviality. For the dual ℓ_1/ℓ_∞ setup in (2) the optimal rate is $\Theta(\sqrt{\log d/N})$. While our goal is to find minimal assumptions that allow for distributed learning with sublinear communication, the reader may wonder at this point whether we have made the problem easier *statistically* by moving to the ℓ_1/ℓ_q assumption. The answer is "yes, but only slightly." When q is constant the optimal rate for ℓ_1/ℓ_q-bounded models is $\Theta(\sqrt{1/N})$,[6] and so the effect of this assumption is to shave off the $\log d$ factor that was present in (4).

3.1 Lipschitz Losses

Our main algorithm is called *sparsified mirror descent* (Algorithm 2). The idea behind the algorithm is to run the online mirror descent algorithm (Ben-Tal and Nemirovski, 2001; Hazan, 2016) in serial across the machines and sparsify the iterates whenever we move from one machine to the next.

In a bit more detail, Algorithm 2 proceeds from machine to machine sequentially. On each machine, the algorithm generates a sequence of iterates $w_1^i, \dots, w_n^i$ by doing a single pass over the machine's n examples $z_1^i, \dots, z_n^i$ using the mirror descent update with regularizer $\mathcal{R}(w) = \frac{1}{2}\|w\|_p^2$, where $\frac{1}{p} + \frac{1}{q} = 1$, and using stochastic gradients $\nabla_t^i \in \partial \ell(w_t^i, z_t^i)$. After the last example is processed on machine i, we compress the last iterate using Maurey sparsification (Algorithm 1) and send it to the next machine, where the process is repeated.

To formally describe the algorithm, we recall the definition of the *Bregman divergence*. Given a convex regularization function $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$, the Bregman divergence with respect to $\mathcal{R}$ is defined as

$$D_{\mathcal{R}}(w \| w') = \mathcal{R}(w) - \mathcal{R}(w') - \langle \nabla \mathcal{R}(w'), w - w' \rangle.$$

For the ℓ_1/ℓ_q setting we exclusively use the regularizer $\mathcal{R}(w) = \frac{1}{2}\|w\|_p^2$, where $\frac{1}{p} + \frac{1}{q} = 1$.

The main guarantee for Algorithm 2 is as follows.

---

[6] The upper bound follows from (3) and the lower bound follows by reduction to the one-dimensional case.

Algorithm 2 (Sparsified Mirror Descent).
Input:
   Constraint set $\mathcal{W}$ with $\|w\|_1 \le B_1$.
   Gradient norm parameter $q \in [2, \infty)$.
   Gradient ℓ_q norm bound R_q.
   Learning rate η, Initial point $\bar{w}$, Sparsity $s, s_0 \in \mathbb{N}$.

Define $p = \frac{q}{q-1}$ and $\mathcal{R}(w) = \frac{1}{2}\|w - \bar{w}\|_p^2$.

For machine $i = 1, \ldots, m$:

- Receive $\widehat{w}^{i-1}$ from machine $i-1$ and set $w_1^i = \widehat{w}^{i-1}$ (if machine 1 set $w_1^1 = \bar{w}$).

- For $t = 1, \ldots, n$: `// Mirror descent step.`
    - Get gradient $\nabla_t^i \in \partial\ell(w_t^i; z_t^i)$.
    - $\nabla\mathcal{R}(\theta_{t+1}^i) \leftarrow \nabla\mathcal{R}(w_t^i) - \eta\nabla_t^i$.
    - $w_{t+1}^i \leftarrow \arg\min_{w \in \mathcal{W}} D_{\mathcal{R}}(w\|\theta_{t+1}^i)$.

- Let $\widehat{w}^i \leftarrow Q^s(w_{n+1}^i)$. `// Sparsification.`

- Send $\widehat{w}^i$ to machine $i+1$.

Sample $i \in [m]$, $t \in [n]$ uniformly at random and return $\widehat{w} := Q^{s_0}(w_t^i)$.

Theorem 3. *Let $q \ge 2$ be fixed. Suppose that subgradients belong to $\mathcal{X}_q$ and that $\mathcal{W} \subseteq \mathcal{W}_1$. If we run [Algorithm 2](#) with $\eta = \frac{B_1}{R_q}\sqrt{\frac{1}{C_q N}}$ and initial point $\bar{w} = 0$, then whenever $s = \Omega(m^{2(q-1)})$ and $s_0 = \Omega(N^{\frac{q}{2}})$ the algorithm guarantees*

$$\mathbb{E}[L_{\mathcal{D}}(\widehat{w})] - L_{\mathcal{D}}(w^\star) \le O\left(\sqrt{\frac{B_1^2 R_q^2 C_q}{N}}\right),$$

where $C_q = q - 1$ is a constant depending only on q.

The total number of bits sent by each machine—besides communicating the final iterate $\widehat{w}$—is at most $O(m^{2(q-1)}\log(d/m) + \log(B_1 R_q N))$, and so the total number of bits communicated globally is at most

$$O\left(N^{\frac{q}{2}}\log(d/N) + m^{2q-1}\log(d/m) + m\log(B_1 R_q N)\right).$$

In the linear model setting [(2)](#) with 1-Lipschitz loss ϕ it suffices to set $s_0 = \Omega(N)$, so that the total bits of communication is

$$O(N\log(d/N) + m^{2q-1}\log(d/m) + m\log(B_1 R_q N)).$$

We see that the communication required by sparsified mirror descent is exponential in the norm parameter q. This means that whenever q is constant, the overall communication is polylogarithmic in dimension. It is helpful to interpret the bound when q is allowed to grow with dimension. An elementary property of ℓ_q norms is that for $q = \log d$, $\|x\|_q \approx \|x\|_\infty$ up to a multiplicative constant. In this case the communication from [Theorem 3](#) becomes polynomial in dimension, which we know from [Section 2.4](#) is necessary.

The guarantee of [Algorithm 2](#) extends beyond the statistical learning model to the first-order stochastic convex optimization model, as well as the online convex optimization model.

Proof sketch. They basic premise behind the algorithm and analysis is that by using the same learning rate across all machines, we can pretend as though we are running a single instance of mirror descent on a

centralized machine. The key difference from the usual analysis is that we need to bound the error incurred by sparsification between successive machines. Here, the choice of the regularizer is crucial. A fundamental property used in the analysis of mirror descent is *strong convexity* of the regularizer. In particular, to give convergence rates that do not depend on dimension (such as (3)) it is essential that the regularizer be $\Omega(1)$-strongly convex. Our regularizer $\mathcal{R}$ indeed has this property.

Proposition 3 (Ball et al. (1994)). *For $p \in (1, 2]$, $\mathcal{R}$ is $(p-1)$-strongly convex with respect to $\|\cdot\|_p$. Equivalently, $D_{\mathcal{R}}(w\|w') \geq \frac{p-1}{2} \cdot \|w - w'\|_p^2 \quad \forall w, w' \in \mathbb{R}^d$.*

On the other hand, to argue that sparsification has negligible impact on convergence, our analysis leverages *smoothness* of the regularizer. Strong convexity and smoothness are at odds with each other: It is well known that in infinite dimension, any norm that is both strongly convex and smooth is isomorphic to a Hilbert space (Pisier, 2011). What makes our analysis work is that while the regularizer $\mathcal{R}$ is not smooth, it is *Hölder-smooth* for any finite q. This is sufficient to bound the approximation error from sparsification. To argue that the excess risk achieved by mirror descent with the ℓ_p regularizer $\mathcal{R}$ is optimal, however, it is essential that the gradients are ℓ_q-bounded rather than ℓ_∞-bounded.

In more detail, the proof can be broken into three components:

- *Telescoping.* Mirror descent gives a regret bound that telescopes across all m machines up to the error introduced by sparsification. To argue that we match the optimal centralized regret, all that is required is to bound m error terms of the form

$$D_{\mathcal{R}}(w^\star \| Q^s(w_{n+1}^i)) - D_{\mathcal{R}}(w^\star \| w_{n+1}^i).$$

- *Hölder-smoothness.* We prove (Theorem 7) that the difference above is of order

$$B_1 \left\| Q^s(w_{n+1}^i) - w_{n+1}^i \right\|_p + B_1^{3-p} \left\| Q^s(w_{n+1}^i) - w_{n+1}^i \right\|_\infty^{p-1}.$$

- *Maurey for ℓ_p norms.* We prove (Theorem 6) that $\left\| Q^s(w_{n+1}^i) - w_{n+1}^i \right\|_p \lesssim \left(\frac{1}{s}\right)^{1-1/p}$ and likewise that $\left\| Q^s(w_{n+1}^i) - w_{n+1}^i \right\|_\infty \lesssim \left(\frac{1}{s}\right)^{1/2}$.

With a bit more work these inequalities yield Theorem 3. We close this section with a few more notes about Algorithm 2 and its performance.

Remark 1. *We can modify Algorithm 2 so that it enjoys a high-probability excess risk bound by changing the final step slightly. Instead of subsampling (i, t) randomly and returning $Q^s(w_t^i)$, have each machine i average all its iterates $w_1^i, \ldots, w_n^i$, then sparsify the average and send it to the final machine, which averages the averaged iterates from all machines and returns $\widehat{w}$ as the result.*

There appears to be a tradeoff here: The communication of the high probability algorithm is $\tilde{O}(m^{2q-1} + mN^q)$, while Algorithm 2 has communication $\widetilde{O}(m^{2q-1} + N^q)$. We leave a comprehensive exploration of this tradeoff for future work.

Remark 2. *For the special case of ℓ_1/ℓ_q-bounded linear models, it is not hard to show that the following strategy also leads to sublinear communication: Truncate each feature vector to the top $\Theta(N^{q/2})$ coordinates, then send all the truncated examples to a central server, which returns the empirical risk minimizer. This strategy matches the risk of Theorem 3 with total communication $\tilde{O}(N^{q/2+1})$, but has two deficiencies. First, it scales as $N^{O(q)}$, which is always worse than $m^{O(q)}$. Second, it does not appear to extend to the general optimization setting.*

3.2 Smooth Losses

We can improve the statistical guarantee and total communication further in the case where $L_{\mathcal{D}}$ is *smooth* with respect to ℓ_q rather than just Lipschitz. We assume that ℓ has β_q-Lipschitz gradients, in the sense that for all $w, w' \in \mathcal{W}_1$ for all z,

$$\left\| \nabla \ell(w,z) - \nabla \ell(w',z) \right\|_q \le \beta_q \| w - w' \|_p,$$

where p is such that $\frac{1}{p} + \frac{1}{q}$.

Theorem 4. *Suppose in addition to the assumptions of Theorem 3 that $\ell(\cdot, z)$ is non-negative and has β_q-Lipschitz gradients with respect to ℓ_q. Let $L^{\star} = \inf_{w \in \mathcal{W}} L_{\mathcal{D}}(w)$. If we run Algorithm 2 with learning rate $\eta = \sqrt{\frac{B_1^2}{C_q \beta_q L^{\star} N}} \wedge \frac{1}{4 C_q \beta_q}$ and $\bar{w} = 0$ then, if $s = \Omega(m^{2(q-1)})$ and $s_0 = \sqrt{\frac{\beta_q B_1^2 N}{C_q L^{\star}}} \wedge \frac{N}{C_q}$, the algorithm guarantees*

$$\mathbb{E}[L_{\mathcal{D}}(\widehat{w})] - L^{\star} \le O\left(\sqrt{\frac{C_q \beta_q B_1^2 L^{\star}}{N}} + \frac{C_q \beta_q B_1^2}{N} \right).$$

The total number of bits sent by each machine—besides communicating the final iterate $\widehat{w}$—is at most $O(m^{2(q-1)} \log(d/m))$, and so the total number of bits communicated globally is at most

$$O\left(\left(\sqrt{\frac{\beta_q B_1^2 N}{C_q L^{\star}}} \wedge \frac{N}{C_q} \right) \log(d/N) + m^{2q-1} \log(d/m) + m \log(\beta_q B_1 N) \right).$$

Compared to the previous theorem, this result provides a so-called "small-loss bound" (Srebro et al., 2010), with the main term scaling with the optimal loss $L^{\star}$. The dependence on N in the communication cost can be as low as $O(\sqrt{N})$ depending on the value of $L^{\star}$.

3.3 Fast Rates under Restricted Strong Convexity

So far all of the algorithmic results we have present scale as $O(N^{-1/2})$. While this is optimal for generic Lipschitz losses, we mentioned in Section 2 that for strongly convex losses the rate can be improved in a nearly-dimension independent fashion to $O(N^{-1})$ for sparse high-dimensional linear models. As in the generic lipschitz loss setting, we show that making the assumption of ℓ_1/ℓ_q-boundness is sufficient to get statistically optimal distributed algorithms with sublinear communication, thus providing a way around the lower bounds for fast rates in Section 2.4.

The key assumption for the results in this section is that the population risk satisfies a form of *restricted strong convexity* over $\mathcal{W}$:

Assumption 1. There is some constant γ_q such that

$$\forall w \in \mathcal{W}, \quad L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^{\star}) - \langle \nabla L_{\mathcal{D}}(w^{\star}), w - w^{\star} \rangle \ge \frac{\gamma_q}{2} \| w - w^{\star} \|_p^2.$$

In a moment we will show how to relate this property to the standard restricted eigenvalue property in high-dimensional statistics (Negahban et al., 2012) and apply it to sparse regression.

Our main algorithm for strongly convex losses is Algorithm 3. The algorithm does not introduce any new tricks for distributed learning over Algorithm 2; rather, it invokes Algorithm 2 repeatedly in an inner loop,

---

Algorithm 3 (Sparsified Mirror Descent for Fast Rates).
Input:

  Constraint set $\mathcal{W}$ with $\|w\|_1 \le B_1$.
  Gradient norm parameter $q \in [2, \infty)$.
  Gradient ℓ_q norm bound R_q.
  RSC constant γ_q. Constant $c > 0$.

  Let $\widehat{w}_0 = 0$, $B_k = 2^{-k/2}B$ and $N_{k+1} = C_q \cdot \left(\frac{4cR}{\gamma B_{k-1}}\right)^2$.

  Let $T = \max\left\{T \mid \sum_{k=1}^{T} N_k \le N\right\}$.

  Let examples have order: $z_1^1, \ldots, z_n^1, \ldots, z_1^m, \ldots, z_n^m$.

  For round $k = 1, \ldots, T$:

  Let $\widehat{w}_k$ be the result of running Algorithm 2 on N_k consecutive examples in the ordering above, with the following configuration:

      1. *The algorithm begins on the example immediately after the last one processed at round $k-1$.*

      2. *The algorithm uses parameters B_1, R_q, s, s_0, and η as prescribed in Proposition 8, with initialization $\bar{w} = \widehat{w}_{k-1}$ and radius $\bar{B} = B_{k-1}$.*

  Return $\widehat{w}_T$.

---

relying on these invocations to take care of communication. This reduction is based on techniques developed in Juditsky and Nesterov (2014), whereby restricted strong convexity is used to establish that error decreases geometrically as a function of the number of invocations to the sub-algorithm. We refer the reader to Appendix C for additional details.

The main guarantee for Algorithm 3 is as follows.

Theorem 5. *Suppose Assumption 1 holds, that subgradients belong to $\mathcal{X}_q$ for $q \ge 2$, and that $\mathcal{W} \subset \mathcal{W}_1$. When the parameter $c > 0$ is a sufficiently large absolute constant, Algorithm 3 guarantees that*

$$\mathbb{E}[L_{\mathcal{D}}(\widehat{w}_T)] - L_{\mathcal{D}}(w^\star) \le O\left(\frac{C_q R_q^2}{\gamma_q N}\right).$$

The total numbers of bits communicated is

$$O\left(\left(N^{2(q-1)}m^{2q-1}\left(\frac{\gamma_q^2 B_q^2}{C_q R_q^2}\right)^{2(q-1)} + N^q \left(\frac{\gamma_q B_1}{C_q R_q}\right)^q\right)\log d + m\log(B_1 R_q N)\right).$$

Treating scale parameters as constant, the total communication simplifies to $O\left(N^{2q-2}m^{2q-1}\log d\right)$.

Note that the communication in this theorem depends polynomially on the various scale parameters, which was not the case for Theorem 3.

Application: Sparse Regression. As an application of Algorithm 3, we consider the sparse regression setting (5), where $L_{\mathcal{D}}(w) = \mathbb{E}_{x,y}(\langle w, x\rangle - y)^2$. We assume $\|x\|_q \le R_q$ and $|y| \le 1$. We let $w^\star = \arg\min_{w \in \mathcal{W}_1} L_{\mathcal{D}}(w)$, so $\|w^\star\|_1 \le B_1$. We assume $w^\star$ is k-sparse, with support set $S \subset [d]$.

We invoke Algorithm 3 constraint set $\mathcal{W} := \left\{w \in \mathbb{R}^d \mid \|w\|_1 \le \|w^\star\|_1\right\}$ and let $\Sigma = \mathbb{E}[xx^\top]$. Our bound depends on the restricted eigenvalue parameter: $\gamma := \inf_{\nu \in \mathcal{W} - w^\star \setminus \{0\}} \|\Sigma^{1/2}\nu\|_2^2 / \|\nu\|_2^2$.

Proposition 4. *Algorithm 3, with constraint set $\mathcal{W}$ and appropriate choice of parameters, guarantees:*

$$\mathbb{E}[L_{\mathcal{D}}(\widehat{w}_T)] - L_{\mathcal{D}}(w^\star) \le O\left(C_q B_1^2 R_q^2 \cdot \frac{k}{\gamma N}\right).$$

Suppressing problem-dependent constants, total communication is of order $O((N^{2q-2} m^{2q-1} \log d)/k^{4q-4})$.

3.4 Extension: Matrix Learning and Beyond

The basic idea behind sparsified mirror descent—that by assuming ℓ_q-boundedness one can get away with using a Hölder-smooth regularizer that behaves well under sparsification—is not limited to the ℓ_1/ℓ_q setting. To extend the algorithm to more general geometry, all that is required is the following:

- The constraint set $\mathcal{W}$ can be written as the convex hull of a set of atoms $\mathcal{A}$ that has sublinear bit complexity.

- The data should be bounded in some norm $\|\cdot\|$ such that the dual $\|\cdot\|_\star$ admits a regularizer $\mathcal{R}$ that is strongly convex and Hölder-smooth with respect to $\|\cdot\|_\star$

- $\|\cdot\|_\star$ is preserved under sparsification. We remark in passing that this property and the previous one are closely related to the notions of type and cotype in Banach spaces (Pisier, 2011).

Here we deliver on this potential and sketch how to extend the results so far to *matrix learning* problems where $\mathcal{W} \subseteq \mathbb{R}^{d \times d}$ is a convex set of matrices. As in Section 3.1 we work with a generic Lipschitz loss $L_{\mathcal{D}}(W) = \mathbb{E}_z \, \ell(W, z)$. Letting $\|W\|_{S_p} = \mathrm{tr}((WW^\top)^{\frac{p}{2}})$ denote the Schatten p-norm, we make the following spectral analogue of the ℓ_1/ℓ_q-boundedness assumption: $\mathcal{W} \subseteq \mathcal{W}_{S_1} := \{W \in \mathbb{R}^{d \times d} \mid \|W\|_{S_1} \le B_1\}$ and subgradients $\partial \ell(\cdot, z)$ belong to $\mathcal{X}_{S_q} := \{X \in \mathbb{R}^{d \times d} \mid \|X\|_{S_q} \le R_q\}$, where $q \ge 2$. Recall that S_1 and S_∞ are the nuclear norm and spectral norm. The S_1/S_∞ setup has many applications in learning (Hazan et al., 2012).

We make the following key changes to Algorithm 2:

- Use the Schatten regularizer $\mathcal{R}(W) = \frac{1}{2}\|W\|_{S_p}^2$.

- Use the following spectral version of the Maurey operator $Q^s(W)$: Let W have singular value decomposition $W = \sum_{i=1}^d \sigma_i u_i v_i^\top$ with $\sigma_i \ge 0$ and define $P \in \Delta_d$ via $P_i \propto \sigma_i$.[7] Sample $i_1, \ldots, i_s$ i.i.d. from P and return $Q^s(W) = \frac{\|W\|_{S_1}}{s} \sum_{\tau=1}^s u_{i_\tau} v_{i_\tau}^\top$.

- Encode and transmit $Q^s(W)$ as the sequence $(u_{i_1}, v_{i_1}), \ldots, (u_{i_s}, v_{i_s})$, plus the scalar $\|W\|_{S_1}$. This takes $\tilde{O}(sd)$ bits.

Proposition 5. *Let $q \ge 2$ be fixed, and suppose that subgradients belong to $\mathcal{X}_{S_q}$ and that $\mathcal{W} \subseteq \mathcal{W}_{S_1}$. If we run the variant of Algorithm 2 described above with learning rate $\eta = \frac{B_1}{R_q}\sqrt{\frac{1}{C_q N}}$ and initial point $\bar{W} = 0$, then whenever $s = \Omega(m^{2(q-1)})$ and $s_0 = \Omega(N^{\frac{q}{2}})$, the algorithm guarantees*

$$\mathbb{E}[L_{\mathcal{D}}(\widehat{W})] - \inf_{W \in \mathcal{W}} L_{\mathcal{D}}(W) \le O\left(\sqrt{\frac{B_1^2 R_q^2 C_q}{N}}\right),$$

where $C_q = q - 1$. The total number of bits communicated globally is at most $\tilde{O}(m^{2q-1} d + N^{\frac{q}{2}} d)$.

---

[7]We may assume $\sigma_i \ge 0$ without loss of generality.

In the matrix setting, the number of bits required to naively send weights $W \in \mathbb{R}^{d \times d}$ or subgradients $\partial\ell(W, z) \in \mathbb{R}^{d \times d}$ is $O(d^2)$. The communication required by our algorithm scales only as $\tilde{O}(d)$, so it is indeed sublinear.

The proof of Proposition 5 is sketched in Appendix C. The key idea is that because the Maurey operator $Q^s(W)$ is defined in the same basis as W, we can directly apply approximation bounds from the vector setting.

4 Discussion

We hope our work will lead to further development of algorithms with sublinear communication. A few immediate questions:

- Can we get matching upper and lower bounds for communication in terms of m, N, $\log d$, and q?

- Currently all of our algorithms work serially. Can we extend the techniques to give parallel speedup?

- Returning to the general setting (1), what abstract properties of the hypothesis class $\mathcal{H}$ are required to guarantee that learning with sublinear communication is possible?

Acknowledgements Part of this work was completed while DF was a student at Cornell University and supported by the Facebook PhD fellowship.

References

Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Distributed simulation and distributed inference. *arXiv preprint arXiv:1804.06952*, 2018a.

Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Inference under information constraints I: lower bounds from chi-square contraction. *arXiv preprint arXiv:1812.11476*, 2018b.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems*, pages 5977–5987, 2018.

Keith Ball, Eric A Carlen, and Elliott H Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones mathematicae*, 115(1):463–482, 1994.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993.

Ron Bekkerman, Mikhail Bilenko, and John Langford. *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press, 2011.

Ahron Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2. Siam, 2001.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. Signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 559–568, 2018.

Avleen S Bijral, Anand D Sarwate, and Nathan Srebro. On data dependence in distributed stochastic optimization. *arXiv preprint arXiv:1603.04379*, 2016.

Mark Braverman, Ankit Garg, Tengyu Ma, Huy L Nguyen, and David P Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1011–1020. ACM, 2016.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Sorathan Chaturapruek, John C Duchi, and Christopher Ré. Asynchronous stochastic convex optimization: the noise is in the noise and sgd don't care. In *Advances in Neural Information Processing Systems*, pages 1531–1539, 2015.

Yuval Dagan and Ohad Shamir. Detecting correlations with little memory and communication. In *Conference On Learning Theory*, pages 1145–1198, 2018.

Yuval Dagan, Gil Kur, and Ohad Shamir. Space lower bounds for linear prediction. *arXiv preprint arXiv:1902.03498*, 2019.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(Jan):165–202, 2012.

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, 2012.

John C Duchi, Michael I Jordan, Martin J Wainwright, and Yuchen Zhang. Optimality guarantees for distributed statistical estimation. *arXiv preprint arXiv:1405.0782*, 2014.

Ankit Garg, Tengyu Ma, and Huy Nguyen. On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2014.

Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. In *Conference On Learning Theory*, pages 3163–3188, 2018.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4): 157–325, 2016.

Elad Hazan, Satyen Kale, and Shai Shalev-Shwartz. Near-optimal algorithms for online matrix prediction. *Conference on Learning Theory*, 2012.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Lee K Jones. A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training. *The annals of Statistics*, 20(1):608–613, 1992.

Anatoli Juditsky and Yuri Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1):44–80, 2014.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in neural information processing systems*, pages 793–800, 2009.

Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13(Jun):1865–1890, 2012.

Daniel M Kane and Jelani Nelson. A derandomized sparse johnson-lindenstrauss transform. *arXiv preprint arXiv:1006.3585*, 2010.

Po-Ling Loh, Martin J Wainwright, et al. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, Bin Yu, et al. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4): 538–557, 2012.

Gilles Pisier. Remarques sur un résultat non publié de b. maurey. *Séminaire Analyse fonctionnelle (dit)*, pages 1–12, 1980.

Gilles Pisier. Martingales in banach spaces (in connection with type and cotype). course ihp, feb. 2–8, 2011. 2011.

Ran Raz. Fast learning requires good memory: A time-space lower bound for parity learning. *Journal of the ACM (JACM)*, 66(1):3, 2018.

Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in neural information processing systems*, pages 693–701, 2011.

KR Sahasranand and Himanshu Tyagi. Extra samples can reduce the communication for independence testing. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 2316–2320. IEEE, 2018.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shai Shalev-Shwartz, Nathan Srebro, and Tong Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 20(6):2807–2832, 2010.

Ohad Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. In *Advances in Neural Information Processing Systems*, pages 163–171, 2014.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, pages 2199–2207, 2010.

Jacob Steinhardt and John Duchi. Minimax rates for memory-bounded sparse linear regression. In *Conference on Learning Theory*, pages 1564–1587, 2015.

Jacob Steinhardt, Gregory Valiant, and Stefan Wager. Memory, communication, and statistical queries. In *Conference on Learning Theory*, pages 1490–1516, 2016.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. In *Advances in Neural Information Processing Systems*, pages 4452–4463, 2018.

Ananda Theertha Suresh, Felix X Yu, Sanjiv Kumar, and H Brendan McMahan. Distributed mean estimation with limited communication. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3329–3337. JMLR. org, 2017.

Hanlin Tang, Shaoduo Gan, Ce Zhang, Tong Zhang, and Ji Liu. Communication compression for decentralized training. In *Advances in Neural Information Processing Systems*, pages 7663–7673, 2018.

Alexandre B Tsybakov. Optimal rates of aggregation. In *Learning Theory and Kernel Machines*, pages 303–313. Springer, 2003.

Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, and Stephen Wright. Atomo: Communication-efficient learning via atomic sparsification. In *Advances in Neural Information Processing Systems*, pages 9872–9883, 2018.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, pages 1306–1316, 2018.

Hantian Zhang, Jerry Li, Kaan Kara, Dan Alistarh, Ji Liu, and Ce Zhang. Zipml: Training linear models with end-to-end low precision, and a little bit of deep learning. In *International Conference on Machine Learning*, pages 4035–4043, 2017.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.

Yuchen Zhang, Martin J Wainwright, and John C Duchi. Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems*, pages 1502–1510, 2012.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, pages 928–936, 2003.

A Basic Results

A.1 Sparsification

In this section we provide approximation guarantees for the Maurey sparsification operator Q^s defined in Algorithm 1.

Theorem 6. *Let $p \in [1, 2]$ be fixed. Then for any $w \in \mathbb{R}^d$, with probability at least $1 - \delta$,*

$$\|Q^s(w) - w\|_p \leq 4\|w\|_1 \left(\frac{1}{s}\right)^{1-\frac{1}{p}} + \|w\|_1 \left(\frac{8\log(1/\delta)}{s}\right)^{\frac{1}{2}} \leq \|w\|_1 \left(\frac{24\log(1/\delta)}{s}\right)^{1-\frac{1}{p}}. \tag{9}$$

Moreover, the following in-expectation guarantee holds:

$$\mathbb{E}\|Q^s(w) - w\|_p \leq \left(\mathbb{E}\|Q^s(w) - w\|_p^p\right)^{1/p} \leq 4\|w\|_1 \left(\frac{1}{s}\right)^{1-\frac{1}{p}}. \tag{10}$$

Proof of Theorem 6. Let $B = \|w\|_1$, and let $Z_\tau = \|w\|_1 \text{sgn}(w_{i_\tau})e_{i_\tau} - w$, and observe that $\mathbb{E}[Z_\tau] = 0$ and $Q^s(w) - w = \frac{1}{s}\sum_{\tau=1}^s Z_\tau$. Since $\|w\|_p \leq B$, we have $\|Z_\tau\|_p \leq 2B$, and so Lemma 2 implies that with probability at least $1 - \delta$,

$$\|Q^s(w) - w\|_p \leq \frac{2}{s} \cdot \mathbb{E}_Z\left(\sum_{t=1}^s \|Z_t\|_p^p\right)^{1/p} + B\sqrt{\frac{8\log(1/\delta)}{s}}$$

$$\leq \frac{4B}{s^{1-\frac{1}{p}}} + B\sqrt{\frac{8\log(1/\delta)}{s}}.$$

$\square$

Lemma 2. Let $p \in [1,2]$. Let $Z_1, \ldots, Z_s$ be a sequence of independent $\mathbb{R}^d$-valued random variables with $\|Z_t\|_p \leq B$ almost surely and $\mathbb{E}[Z_t] = 0$. Then with probability at least $1 - \delta$,

$$\left\|\frac{1}{s}\sum_{t=1}^s Z_t\right\|_p \leq \frac{2}{s} \cdot \mathbb{E}_Z\left(\sum_{t=1}^s \|Z_t\|_p^p\right)^{1/p} + B\sqrt{\frac{2\log(1/\delta)}{s}}$$

Furthermore, a sharper guarantee holds in expectation:

$$\mathbb{E}_Z\left\|\frac{1}{s}\sum_{t=1}^s Z_t\right\|_p \leq \left(\mathbb{E}_Z\left\|\frac{1}{s}\sum_{t=1}^s Z_t\right\|_p^p\right)^{1/p} \leq \frac{2}{s} \cdot \mathbb{E}_Z\left(\sum_{t=1}^s \|Z_t\|_p^p\right)^{1/p}.$$

Proof of Lemma 2. To obtain the high-probability statement, the first step is to apply the standard Mcdiarmid-type high-probability uniform convergence bound for Rademacher complexity (e.g. Shalev-Shwartz and Ben-David (2014)), which states that with probability at least $1 - \delta$,

$$\left\|\frac{1}{s}\sum_{t=1}^s Z_t\right\|_p \leq 2\,\mathbb{E}_Z\,\mathbb{E}_\epsilon\left\|\frac{1}{s}\sum_{t=1}^s \epsilon_t Z_t\right\|_p + B\sqrt{\frac{2\log(1/\delta)}{s}},$$

where $\epsilon \in \{\pm 1\}^n$ are Rademacher random variables. Conditioning on $Z_1, \ldots, Z_n$, we have

$$\mathbb{E}_\epsilon\left\|\frac{1}{s}\sum_{t=1}^s \epsilon_t Z_t\right\|_p \leq \left(\mathbb{E}_\epsilon\left\|\frac{1}{s}\sum_{t=1}^s \epsilon_t Z_t\right\|_p^p\right)^{1/p}.$$

On the other hand, for the in-expectation results, Jensen's inequality and the standard in-expectation symmetrization argument for Rademacher complexity directly yield

$$\mathbb{E}_Z\left\|\frac{1}{s}\sum_{t=1}^s Z_t\right\|_p \leq \left(\mathbb{E}_Z\left\|\frac{1}{s}\sum_{t=1}^s Z_t\right\|_p^p\right)^{1/p} \leq 2\left(\mathbb{E}_Z\,\mathbb{E}_\epsilon\left\|\frac{1}{s}\sum_{t=1}^s \epsilon_t Z_t\right\|_p^p\right)^{1/p}.$$

From here the proof proceeds in the same fashion for both cases. Let $Z_t[i]$ denote the ith coordinate of Z_t and let $z_i = (Z_1[i], \ldots, Z_s[i]) \in \mathbb{R}^s$. We have

$$\mathbb{E}_\epsilon\left\|\frac{1}{s}\sum_{t=1}^s \epsilon_t Z_t\right\|_p^p = \sum_{i=1}^d \mathbb{E}_\epsilon\left(\frac{1}{s}\sum_{t=1}^s \epsilon_t Z_t[i]\right)^p \leq \sum_{i=1}^d \left(\mathbb{E}_\epsilon\left(\frac{1}{s}\sum_{t=1}^s \epsilon_t Z_t[i]\right)^2\right)^{p/2},$$

18

where the inequality follows from Jensen's inequality since $p \leq 2$. We now use that cross terms in the square vanish, as well as the standard inequality $\|x\|_2 \leq \|x\|_p$ for $p \leq 2$:

$$\sum_{i=1}^{d} \left(\mathbb{E}_\epsilon \left(\frac{1}{s} \sum_{t=1}^{s} \epsilon_t Z_t[i] \right)^2 \right)^{p/2} = \sum_{i=1}^{d} \left(\frac{1}{s^2} \|z_i\|_2^2 \right)^{p/2} = \frac{1}{s^p} \sum_{i=1}^{d} \|z_i\|_2^p \leq \frac{1}{s^p} \sum_{i=1}^{d} \|z_i\|_p^p = \frac{1}{s^p} \sum_{t=1}^{s} \|Z_t\|_p^p.$$

$\square$

Proof of Lemma 1. We first prove the result for the smooth case. Let x and y be fixed. Let $B = \|w\|_1$, and let us abbreviate $R := R_\infty$. Let $Z_\tau = \langle \|w\|_1 \mathrm{sgn}(w_{i_\tau}) e_{i_\tau} - w, x \rangle$, and observe that $\mathbb{E}[Z_\tau] = 0$ and $\langle Q^s(w) - w, x \rangle = \frac{1}{s} \sum_{\tau=1}^{s} Z_\tau$. Since we have $\|w\|_1 \leq B$ and $\|x\|_\infty \leq R$ almost surely, one has $|Z_\tau| \leq 2BR$ almost surely. We can write

$$\phi(\langle Q^s(w), x \rangle, y) = \phi \left(\langle w, x \rangle + \frac{1}{s} \sum_{\tau=1}^{s} Z_\tau, y \right).$$

Using smoothness, we can write

$$\phi \left(\langle w, x \rangle + \frac{1}{s} \sum_{\tau=1}^{s} Z_\tau, y \right) \leq \phi \left(\langle w, x \rangle + \frac{1}{s} \sum_{\tau=1}^{s-1} Z_\tau, y \right) + \phi' \left(\langle w, x \rangle + \frac{1}{s} \sum_{\tau=1}^{s-1} Z_\tau, y \right) \cdot \frac{Z_s}{s} + \frac{\beta}{2s^2} (Z_s)^2.$$

Since $\mathbb{E}[Z_s \mid Z_1, \ldots, Z_{s-1}] = 0$, and since Z_s is bounded, taking expectation gives

$$\mathbb{E}_{Z_s} \left[\phi \left(\langle w, x \rangle + \frac{1}{s} \sum_{\tau=1}^{s} Z_\tau, y \right) \mid Z_1, \ldots, Z_{s-1} \right] \leq \phi \left(\langle w, x \rangle + \frac{1}{s} \sum_{\tau=1}^{s-1} Z_\tau, y \right) + \frac{\beta B^2}{s^2} \|x\|_\infty^2.$$

Proceeding backwards in the, fashion, we arrive at the inequality

$$\mathbb{E}_Z \, \phi \left(\langle w, x \rangle + \frac{1}{s} \sum_{\tau=1}^{s} Z_\tau, y \right) \leq \phi(\langle w, x \rangle, y) + \frac{\beta B^2}{s} \|x\|_\infty^2.$$

The final result follows by taking expectation over x and y.

For Lipschitz losses, we use Lipschitzness and Jensen's inequality to write

$$\mathbb{E} \, L_{\mathcal{D}}(Q^s(w)) - L_{\mathcal{D}}(w) \leq L \sqrt{\mathbb{E} \, \mathbb{E}_x \langle Q^s(w) - w, x \rangle^2}.$$

The result now follows by appealing to the result for the smooth case to bound $\mathbb{E}_x \langle Q^s(w) - w, x \rangle^2$, since we can interpret this as the expectation of new linear model loss $\mathbb{E}_{x,y} \, \tilde{\phi}(\langle w', x \rangle, y) := \mathbb{E}_x (\langle w', x \rangle - \langle w, x \rangle)^2$, where $y = \langle w, x \rangle$. This loss is 2-smooth with respect to the first argument, which leads to the final bound. $\square$

Lemma 3. Let $w \in \mathbb{R}^d$ be fixed and let $F : \mathbb{R}^d \to \mathbb{R}$ have β_q-Lipschitz gradient with respect to ℓ_q, where $q \geq 2$. Then Algorithm 1 guarantees that

$$\mathbb{E} \, F(Q^s(w)) \leq F(w) + \frac{\beta_q \|w\|_1^2}{s}. \tag{11}$$

Proof of Lemma 3. The assumed gradient Lipschitzness implies that for any w, w'

$$F(w) \leq F(w') + \langle \nabla F(w'), w - w' \rangle + \frac{\beta_q}{2} \|w - w'\|_p^2,$$

19

where $\frac{1}{p} + \frac{1}{q} = 1$. As in the other Maurey lemmas, we write $Z_\tau = (\|w\|_1 \mathrm{sgn}(w_{i_\tau}) e_{i_\tau} - w)$, so that $\mathbb{E}[Z_\tau] = 0$ and $Q^s(w) - w = \frac{1}{s} \sum_{\tau=1}^{s} Z_\tau$. We can now write

$$\mathbb{E}\, F(Q^s(w)) = \mathbb{E}\, F\left(w + \frac{1}{s}\sum_{\tau=s}^{s} Z_\tau\right)$$

Using smoothness, we have

$$\mathbb{E}_{Z_s} F\left(w + \frac{1}{s}\sum_{\tau=s}^{s} Z_\tau\right) \leq F\left(w + \frac{1}{s}\sum_{\tau=s}^{s-1} Z_\tau\right) + \mathbb{E}_{Z_s}\left\langle \nabla F\left(w + \frac{1}{s}\sum_{\tau=s}^{s-1} Z_\tau\right), \frac{Z_s}{s}\right\rangle + \frac{\beta_q}{2s^2}\mathbb{E}_{Z_s}\|Z_s\|_p^2$$

$$\leq F\left(w + \frac{1}{s}\sum_{\tau=s}^{s-1} Z_\tau\right) + \frac{\beta_q}{s^2}\|w\|_1^2.$$

Proceeding backwards in the same fashion, we get

$$\mathbb{E}\, F(Q^s(s)) = \mathbb{E}_{Z_1, \ldots, Z_s} F\left(w + \frac{1}{s}\sum_{\tau=s}^{s} Z_\tau\right) \leq \frac{\beta_q \|w\|_1^2}{s}.$$

$\square$

A.2 Approximation for ℓ_p Norms

In this section we work with the regularizer $\mathcal{R}(\theta) = \frac{1}{2}\|\theta\|_p^2$, where $p \in [1, 2]$, and we let q be such that $\frac{1}{p} + \frac{1}{q} = 1$. The main structural result we establish is a form of Hölder smoothness of $\mathcal{R}$, which implies that ℓ_1 bounded vectors can be sparsified while preserving Bregman divergences for $\mathcal{R}$, with the quality degrading as $p \to 1$.

Theorem 7. *Suppose that $a, b, c \in \mathbb{R}^d$ have $\|a\|_1 \vee \|b\|_1 \vee \|c\|_1 \leq B$. Then it holds that*

$$D_{\mathcal{R}}(c\|a) - D_{\mathcal{R}}(c\|b) \leq 5B\|a - b\|_p + 4B^{3-p}\|a - b\|_\infty^{p-1}.$$

The remainder of this section is dedicated to proving Theorem 7.

We use the following generic fact about norms; all other results in this section are specific to the ℓ_p norm regularizer. For any norm and any x, y with $\|x\| \vee \|y\| \leq B$, we have

$$\|x\|^2 - \|y\|^2 \leq \|x - y\|^2 + 2\|x - y\|\|y\| \leq 4B\|x - y\|. \tag{12}$$

To begin, we need some basic approximation properties. We have the following expression:

$$\nabla \mathcal{R}(\theta) = \|\theta\|_p^{2-p} \cdot \left(|\theta_1|^{p-1}\mathrm{sgn}(\theta_1), \ldots, |\theta_d|^{p-1}\mathrm{sgn}(\theta_d)\right). \tag{13}$$

Proposition 6. *For any vector θ,*

$$\|\nabla \mathcal{R}(\theta)\|_q = \|\theta\|_p.$$

Proof of Proposition 6. Expanding the expression in (13), we have

$$\|\nabla \mathcal{R}(\theta)\|_q = \|\theta\|_p^{2-p} \cdot \left(\sum_{i=1}^{d}|\theta_i|^{q(p-1)}\right)^{1/q}.$$

Using that $q = \frac{p}{p-1}$, this simplifies to

$$\|\theta\|_p^{2-p} \cdot \|\theta\|_p^{p-1} = \|\theta\|_p.$$

$\square$

Lemma 4. Suppose that $\|a\|_p \vee \|b\|_p \leq B$. Then

$$D_{\mathcal{R}}(a\|b) \leq 3B\|a-b\|_p.$$

Proof of Lemma 4. We write

$$D_{\mathcal{R}}(a\|b) = \mathcal{R}(a) - \mathcal{R}(b) - \langle \nabla\mathcal{R}(b), a-b \rangle.$$

Using (12) and the expression for $\mathcal{R}$, it follows that

$$D_{\mathcal{R}}(a\|b) \leq 2B\|a-b\|_p - \langle \nabla\mathcal{R}(b), a-b \rangle.$$

This is further upper bounded by

$$D_{\mathcal{R}}(a\|b) \leq 2B\|a-b\|_p + \|\nabla\mathcal{R}(b)\|_q \|a-b\|_p.$$

The result follows by using that $\|\nabla\mathcal{R}(b)\|_q = \|b\|_p \leq B$, by Proposition 6.

$\square$

Lemma 5. Let $p \in [1,2]$ and let $h(x) = |x|^{p-1}\mathrm{sgn}(x)$. Then h is Hölder-continuous:

$$|h(x) - h(y)| \leq 2|x-y|^{p-1} \quad \forall x, y \in \mathbb{R}.$$

Proof of Lemma 5. Fix any $x, y \in \mathbb{R}$ and assume $|x| \geq |y|$ without loss of generality. We have two cases. First, when $\mathrm{sgn}(x) = \mathrm{sgn}(y)$ we have

$$|h(x) - h(y)| = \left||x|^{p-1} - |y|^{p-1}\right| = |x|^{p-1} - |y|^{p-1} \leq (|x|-|y|)^{p-1} \leq |x-y|^{p-1},$$

where we have used that $p-1 \in [0,1]$ and subadditivity of $x \mapsto x^{p-1}$ over $\mathbb{R}_+$, as well as triangle inequality. On the other hand if $\mathrm{sgn}(x) \neq \mathrm{sgn}(y)$, we have

$$|h(x) - h(y)| = \left||x|^{p-1} + |y|^{p-1}\right| = |x|^{p-1} + |y|^{p-1} \leq 2^{2-p}\big||x| + |y|\big|^{p-1}.$$

Now, using that $\mathrm{sgn}(x) \neq \mathrm{sgn}(y)$, we have

$$\big||x| + |y|\big|^{p-1} = \big||x| \cdot \mathrm{sgn}(x) + |y| \cdot \mathrm{sgn}(x)\big|^{p-1} = \big||x| \cdot \mathrm{sgn}(x) - |y| \cdot \mathrm{sgn}(y)\big|^{p-1} = |x-y|^{p-1}.$$

Putting everything together, this establishes that

$$|h(x) - h(y)| \leq 2^{2-p}|x-y|^{p-1} \leq 2|x-y|^{p-1}.$$

$\square$

Lemma 6. Suppose that $\|a\|_p \vee \|b\|_p \leq B$. Then it holds that

$$\|\nabla\mathcal{R}(a) - \nabla\mathcal{R}(b)\|_\infty \leq 2B^{2-p}\|a-b\|_\infty^{p-1} + \|a-b\|_p, \tag{14}$$

and

$$\|\nabla\mathcal{R}(a) - \nabla\mathcal{R}(b)\|_q \leq 2B^{2-p}\|a-b\|_p^{p-1} + \|a-b\|_p. \tag{15}$$

Proof of Lemma 6. Let $h(x) = |x|^{p-1}\operatorname{sgn}(x)$, so that

$$\nabla\mathcal{R}(\theta) = \|\theta\|_p^{2-p} \cdot (h(\theta_1), \dots, h(\theta_d)).$$

Fix vectors $a, b \in \mathbb{R}^d$. Assume without loss of generality that $\|a\|_p \geq \|b\|_p > 0$; if $\|b\|_p = 0$ the result follows immediately from Proposition 6. We work with the following normalized vectors: $\bar{a} := a/\|b\|_p$ and $\bar{b} := b/\|b\|_p$. Our assumptions on the norms imply $\|\bar{a}\|_p \geq \|\bar{b}\|_p = 1$.

Fix a coordinate $i \in [d]$. We establish the following chain of elementary inequalities:

$$\left|\nabla\mathcal{R}(\bar{a})_i - \nabla\mathcal{R}(\bar{b})_i\right| = \left|\|\bar{a}\|_p^{2-p}h(\bar{a}_i) - \|\bar{b}\|_p^{2-p}h(\bar{b}_i)\right|$$
$$= \left|\|\bar{a}\|_p^{2-p}h(\bar{a}_i) - \|\bar{a}\|_p^{2-p}h(\bar{b}_i) + \|\bar{a}\|_p^{2-p}h(\bar{b}_i) - \|\bar{b}\|_p^{2-p}h(\bar{b}_i)\right|$$

Using the triangle inequality:

$$\leq \|\bar{a}\|_p^{2-p} \cdot \left|h(\bar{a}_i) - h(\bar{b}_i)\right| + \left|\bar{b}_i\right|^{p-1} \cdot \left|\|\bar{a}\|_p^{2-p} - \|\bar{b}\|_p^{2-p}\right|$$

Using the Hölder-continuity of h established in Lemma 5:

$$\leq 2\|\bar{a}\|_p^{2-p} \cdot \left|\bar{a}_i - \bar{b}_i\right|^{p-1} + \left|\bar{b}_i\right|^{p-1} \cdot \left|\|\bar{a}\|_p^{2-p} - \|\bar{b}\|_p^{2-p}\right|$$

Using that $\|\bar{a}\|_p \geq \|\bar{b}\|_p = 1$:

$$\leq 2\|\bar{a}\|_p^{2-p} \cdot \left|\bar{a}_i - \bar{b}_i\right|^{p-1} + \left|\bar{b}_i\right|^{p-1} \cdot \left(\|\bar{a}\|_p^{2-p} - 1\right).$$

Finally, since $\|\bar{a}\|_p \geq 1$ and $2 - p \leq 1$, we can drop the exponent:

$$\leq 2\|\bar{a}\|_p^{2-p} \cdot \left|\bar{a}_i - \bar{b}_i\right|^{p-1} + \left|\bar{b}_i\right|^{p-1} \cdot \left(\|\bar{a}\|_p - 1\right).$$

To finish the proof, we rescale both sides of the inequality by $\|b\|_p$. Observe that $\nabla\mathcal{R}(\theta)$ is homogeneous in the following sense: For any $r \geq 0$,

$$\nabla\mathcal{R}(r\theta) = r \cdot \nabla\mathcal{R}(\theta).$$

Along with this observation, the inequality we just established implies

$$\left|\nabla\mathcal{R}(a)_i - \nabla\mathcal{R}(b)_i\right| \leq 2\|b\|_p\|\bar{a}\|_p^{2-p} \cdot \left|\bar{a}_i - \bar{b}_i\right|^{p-1} + \left|\bar{b}_i\right|^{p-1} \cdot \left(\|a\|_p - \|b\|_p\right)$$
$$\leq 2\|b\|_p\|\bar{a}\|_p^{2-p} \cdot \left|\bar{a}_i - \bar{b}_i\right|^{p-1} + \left|\bar{b}_i\right|^{p-1} \cdot \|a - b\|_p$$
$$= 2\left(\|\bar{a}\|_p\|b\|_p\right)^{2-p} \cdot \left|\bar{a}_i\|b\|_p - \bar{b}_i\|b\|_p\right|^{p-1} + \left|\bar{b}_i\right|^{p-1} \cdot \|a - b\|_p$$
$$= 2\|a\|_p^{2-p} \cdot |a_i - b_i|^{p-1} + \left|\bar{b}_i\right|^{p-1} \cdot \|a - b\|_p.$$

For the ℓ_∞ bound, the result follows immediately by using that $\left|\bar{b}_i\right| \leq \|\bar{b}\|_p \leq 1$. For the ℓ_q bound, we use that for any vector z, $\left\|(z_i^{p-1})_{i \leq d}\right\|_q = \|z\|_p^{p-1}$, and that $\|\bar{b}\|_p \leq 1$. $\qquad\qquad\qquad\square$

Proof of Theorem 7. Throughout this proof we use that $\|x\|_p \leq \|x\|_1$ for all $p \geq 1$. To start, expanding the definition of the Bregman divergence we have

$$D_\mathcal{R}(c\|a) - D_\mathcal{R}(c\|b) = D_\mathcal{R}(b\|a) + \langle\nabla\mathcal{R}(a) - \nabla\mathcal{R}(b), b - c\rangle.$$

Using Lemma 4, this is at most

$$= 3B\|a - b\|_p + \langle \nabla\mathcal{R}(a) - \nabla\mathcal{R}(b), b - c \rangle.$$

Now, applying Hölder's inequality, this is upper bounded by

$$\leq 3B\|a - b\|_p + \|\nabla\mathcal{R}(a) - \nabla\mathcal{R}(b)\|_\infty \|b - c\|_1$$
$$\leq 3B\|a - b\|_p + 2B\|\nabla\mathcal{R}(a) - \nabla\mathcal{R}(b)\|_\infty.$$

To conclude, we plug in the bound from Lemma 6. $\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

B Proofs from Section 2

B.1 Proofs from Section 2.2

Proof of Theorem 1. Let $A \in \mathbb{R}^{k \times d}$ be the derandomized JL matrix constructed according to Kane and Nelson (2010), Theorem 2. Let $x'_t = Ax_t$ denote the projected feature vector and $w^\star = \arg\min_{w:\|w\|_2 \leq 1} L_\mathcal{D}(w)$.

We first bound the regret of gradient descent in the projected space in terms of certain quantities that depend on A, then show how the JL matrix construction guarantees that these quantities are appropriately bounded.

Since ϕ is L-Lipschitz, we have the preliminary error estimate

$$\phi(\langle Ax, Aw^\star \rangle, y) - \phi(\langle x, w^\star \rangle, y) \leq L\ |\langle Ax, Aw^\star \rangle - \langle x, w^\star \rangle|,$$

and so

$$L_\mathcal{D}(A^\top A w^\star) - L_\mathcal{D}(w^\star) \leq L \cdot \mathbb{E}_x |\langle Ax, Aw^\star \rangle - \langle x, w^\star \rangle|. \tag{16}$$

Now recall that the m machines are simply running online gradient descent in serial over the k-dimensional projected space, and the update has the form $u_t \leftarrow u_{t-1} - \nabla\phi(\langle u_t, x'_t \rangle, y_t)$, where η is the learning rate parameter. The standard online gradient descent regret guarantee (Hazan, 2016) implies that for any vector $u \in \mathbb{R}^k$:

$$\frac{1}{N}\sum_{t=1}^N \phi(\langle u_t, x'_t \rangle, y_t) - \frac{1}{N}\sum_{t=1}^N \phi(\langle u, x'_t \rangle, y_t) \leq \frac{1}{2\eta N}\|u\|_2^2 + \frac{\eta}{2N}\sum_{t=1}^N \|x'_t\|_2^2.$$

Equivalently, we have

$$\frac{1}{N}\sum_{t=1}^N \phi(\langle A^\top u_t, x_t \rangle, y_t) - \frac{1}{N}\sum_{t=1}^N \phi(\langle A^\top u, x_t \rangle, y_t) \leq \frac{1}{2\eta N}\|u\|_2^2 + \frac{\eta}{2N}\sum_{t=1}^N \|Ax_t\|_2^2$$

Since the pairs (x_t, y_t) are drawn i.i.d., the standard online-to-batch conversion lemma for online convex optimization (Cesa-Bianchi and Lugosi, 2006) yields the following guarantee for any vector u:

$$\frac{1}{N}\sum_{t=1}^N \mathbb{E}_S\left[L_\mathcal{D}(A^\top u_t) \right] - L_\mathcal{D}(A^\top u) \leq \frac{1}{2\eta N}\|u\|_2^2 + \frac{\eta}{2N}\sum_{t=1}^N \mathbb{E}_S \|Ax_t\|_2^2$$
$$= \frac{1}{2\eta N}\|u\|_2^2 + \frac{\eta L^2}{2}\mathbb{E}_x \|Ax\|_2^2.$$

Applying Jensen's inequality to the left-hand side and choosing $u = u^* := Aw^*$, we conclude that

$$\mathbb{E}_S\left[L_{\mathcal{D}}\left(\frac{1}{N}\sum_{t=1}^{N}A^\top u_t\right)\right] - L_{\mathcal{D}}(A^\top u^*) \le \frac{1}{2\eta N}\|Aw^*\|_2^2 + \frac{\eta L^2}{2}\mathbb{E}_x\|Ax\|_2^2,$$

or in other words,

$$\mathbb{E}_S\left[L_{\mathcal{D}}\left(\hat{w}\right)\right] - L_{\mathcal{D}}(A^\top Aw^*) \le \frac{1}{2\eta N}\|Aw^*\|_2^2 + \frac{\eta L^2}{2}\mathbb{E}_x\|Ax\|_2^2.$$

We now relate this bound to the risk relative to the benchmark $L_{\mathcal{D}}(w^*)$. Using (16) we have

$$\mathbb{E}_S\left[L_{\mathcal{D}}\left(\hat{w}\right)\right] - L_{\mathcal{D}}(w^*) \le \frac{1}{2\eta N}\|Aw^*\|_2^2 + \frac{\eta L^2}{2}\mathbb{E}_x\|Ax\|_2^2 + L\mathbb{E}_x\left|\langle Ax, Aw^*\rangle - \langle x, w^*\rangle\right|.$$

Taking expectation with respect to the draw A, we get that

$$\mathbb{E}_S\mathbb{E}_A\left[L_{\mathcal{D}}\left(\hat{w}\right)\right] - L_{\mathcal{D}}(w^*) \le \mathbb{E}_x\left[\mathbb{E}_A\left[\frac{1}{2\eta N}\|Aw^*\|_2^2 + \frac{\eta L^2}{2}\|Ax\|_2^2 + L\left|\langle Ax, Aw^*\rangle - \langle x, w^*\rangle\right|\right]\right]. \quad (17)$$

It remains to bound the right-hand side of this expression. To begin, we condition on the vector x with respect to which the outer expectation in (17) is taken. The derandomized JL transform guarantees (Kane and Nelson (2010), Theorem 2) that for any $\delta > 0$ and any fixed vectors x, w^*, if we pick $k = O\left(\log(1/\delta)/\varepsilon^2\right)$, then we are guaranteed that with probability at least $1 - \delta$,

$$\|Ax\|_2 \le (1+\varepsilon)\|x\|_2, \quad \|Aw^*\|_2 \le (1+\varepsilon)\|w^*\|_2 \quad \text{and} \quad \left|\langle Ax, Aw^*\rangle - \langle x, w^*\rangle\right| \le \frac{\varepsilon}{4}\|x\|_2\|w^*\|_2.$$

We conclude that by picking $\varepsilon = O\left(1/\sqrt{N}\right)$, with probability $1 - \delta$,

$$\|Ax\|_2 \le O(R_2), \quad \|Aw^*\|_2 \le O(B_2), \quad \text{and} \quad \left|\langle Ax, Aw^*\rangle - \langle x, w^*\rangle\right| \le O\left(\frac{B_2 R_2}{\sqrt{N}}\right).$$

To convert this into an in-expectation guarantee, note that since entries in A belong to $\{-1, 0, +1\}$, the quantities $\|Ax\|_2$, $\|Aw^*\|_2$, and $\langle Ax, Aw^*\rangle$ all have magnitude $O(\text{poly}(d))$ with probability 1 (up to scale factors B_2 and R_2). Hence,

$$\mathbb{E}_A\left[\frac{1}{2\eta N}\|Aw^*\|_2^2 + \frac{\eta L^2}{2}\|Ax\|_2^2 + L\left|\langle Ax, Aw^*\rangle - \langle x, w^*\rangle\right|\right]$$

$$\le (1-\delta)\cdot O\left(\frac{B_2^2}{2\eta N} + \frac{\eta L^2 R_2^2}{2} + \frac{LB_2 R_2}{\sqrt{N}}\right) + \delta\cdot O\left(\text{poly}(d)\cdot\left(\frac{B_2^2}{2\eta N} + \frac{\eta L^2 R_2^2}{2} + LB_2 R_2\right)\right).$$

Picking $\delta = 1/\sqrt{\text{poly}(d)N}$ and using the step size $\eta = \sqrt{\frac{B_2^2}{L^2 R_2^2 N}}$, we get the desired bound:

$$\mathbb{E}_A\left[\frac{1}{2\eta N}\|Aw^*\|_2^2 + \frac{\eta L^2}{2}\|Ax\|_2^2 + L\left|\langle Ax, Aw^*\rangle - \langle x, w^*\rangle\right|\right] \le O(LB_2 R_2/\sqrt{N}).$$

Since this in-expectation guarantee holds for any fixed x, it also holds in expectation over x:

$$\mathbb{E}_x\,\mathbb{E}_A\left[\frac{1}{2\eta N}\|Aw^*\|_2^2 + \frac{\eta L^2}{2}\|Ax\|_2^2 + L\left|\langle Ax, Aw^*\rangle - \langle x, w^*\rangle\right|\right] \le O(L/\sqrt{N}).$$

Using this inequality to bound the right-hand side in (17) yields the claimed excess risk bound. Recall that we have $k = O\left(\log(1/\delta)/\varepsilon^2\right) = O\left(N\log(Nd)\right)$, and so the communication cost to send a single iterate (taking into account numerical precision) is upper bounded by $O(N\log(Nd)\cdot\log(LB_2 R_2 N))$. $\square$

B.2 Proofs from Section 2.4

Our lower bounds are based on reduction to the so-called "hide-and-seek" problem introduced by Shamir (2014).

Definition 1 (Hide-and-seek problem). *Let $\{\mathbb{P}_j\}_{j=1}^d$ be a set of product distributions over $\{\pm 1\}^d$ defined via $\mathbb{E}_{\mathbb{P}_j}[z_i] = 2\rho \mathbb{1}\{j = i\}$. Given N i.i.d. instances from $\mathbb{P}_{j^\star}$, where $j^\star$ is unknown, detect $j^\star$.*

Theorem 8 (Shamir (2014)). *Let $W \in [d]$ be the output of a $(b, 1, N)$ protocol for the hide-and-seek problem. Then there exists some $j^\star \in [d]$ such that*

$$\Pr_{j^\star}(W = j^\star) \le \frac{3}{d} + \sqrt{\frac{Nb\rho^2}{d}}.$$

Proof of Theorem 2. Recall that $\mathcal{W}_1 = \{w \in \mathbb{R}^d \mid \|w\|_1 \le 1\}$. We create a family of d statistical learning instances as follows. Let the hide-and seek parameter $\rho \in [0, 1/2]$ be fixed. Let $\mathcal{D}_j$ have features drawn from the be the jth hide-and-seek distribution $\mathbb{P}_j$ and have $y = 1$, and set $\phi(\langle w, x\rangle, y) = -\langle w, x\rangle y$, so that $L_{\mathcal{D}_j}(w) = -2\rho w_j$. Then we have $\min_{w \in \mathcal{W}_1} L_{\mathcal{D}_j}(w) = -2\rho$. Consequently, for any predictor weight vector w we have

$$L_{\mathcal{D}_j}(w) - L_{\mathcal{D}_j}(w^\star) = 2\rho(1 - w_j).$$

If $L_{\mathcal{D}_j}(\widehat{w}) - L_{\mathcal{D}_j}(w^\star) < \rho$, this implies (by rearranging) that $\widehat{w}_j > \frac{1}{2}$. Since $\widehat{w} \in \mathcal{W}_1$ and thus $\sum_{i=1}^d |\widehat{w}_j| \le 1$, this implies $j = \arg\max_i \widehat{w}_i$. Thus, if we define $W = \arg\max_i \widehat{w}$ as our decision for the hide-and-seek problem, we have

$$\Pr_j\big(L_{\mathcal{D}_j}(\widehat{w}) - L_{\mathcal{D}_j}(w^\star) < \rho\big) \le \Pr_j(W = j).$$

Appealing to Theorem 8, this means that for every algorithm $\widehat{w}$ there exists an index j for which

$$\Pr_j\big(L_{\mathcal{D}_j}(\widehat{w}) - L_{\mathcal{D}_j}(w^\star) < \rho\big) \le \frac{3}{d} + \sqrt{\frac{Nb\rho^2}{d}}.$$

To conclude the result we choose $\rho = \frac{1}{16}\sqrt{\frac{d}{bN}} \wedge \frac{1}{2}$.

$\square$

Proof of Proposition 1. This result is an immediate consequence of the reductions to the hide-and-seek problem established in Theorem 2. All that changes is which lower bound for the hide-and-seek problem we invoke. We set $\rho \propto \frac{d}{bN}$ in the construction in Theorem 2, then appeal to Theorem 3 in Shamir (2014). $\square$

Proof of Proposition 2. We create a family of d statistical learning instances as follows. Let the hide-and seek parameter $\rho \in [0, 1/2]$ be fixed. Let $\mathbb{P}_j$ be the jth hide-and-seek distribution. We create distribution $\mathcal{D}_j$ via: 1) Draw $x \sim \mathbb{P}_j$ 2) set $y = 1$. Observe that $\mathbb{E}[x_i x_k] = 0$ for all $i \ne k$ and $\mathbb{E}[x_i^2] = 1$, so $\Sigma = I$. Consequently, we have

$$L_{\mathcal{D}_j}(w) = \mathbb{E}_{x \sim \mathbb{P}_j}(\langle w, x\rangle - y)^2 = w^\top \Sigma w - 4\rho w_j + 1 = \|w\|_2^2 - 4\rho w_j + 1.$$

Let $w^\star = \arg\min_{w \in \|w\|_1 \le 1} L_{\mathcal{D}_j}(w)$. It is clear from the expression above $w_i^\star = 0$ for all $i \ne j$. For coordinate j we have $w_j^\star = \arg\min_{-1 \le \alpha \le 1}\{\alpha^2 - 4\rho\alpha\}$. Whenever $\rho \le 1/2$ the solution is 2ρ, so we can write $w^\star = 2\rho e_j$, which is clearly 1-sparse.

We can now write the excess risk for a predictor w as

$$L_{\mathcal{D}_j}(w) - L_{\mathcal{D}_j}(w^\star) = \|w\|_2^2 - 4\rho w_j + 4\rho^2 = \sum_{i \neq j} w_i^2 + (w_j - 2\rho)^2.$$

Now suppose that the excess risk for w is at most ρ^2. Dropping the sum term in the excess risk, this implies

$$(w_j - 2\rho)^2 < \rho^2.$$

It follows that $w_j \in (\rho, 3\rho)$. On the other hand, we also have

$$\sum_{i \neq j} w_i^2 < \rho^2,$$

and so any $i \neq j$ must have $|w_i| < \rho$. Together, these facts imply that if the excess risk for w is less than ρ^2, then $j = \arg\max_i w_i$.

Thus, for any algorithm output $\widehat{w}$, if we define $W = \arg\max_i \widehat{w}_i$ as our decision for the hide-and-seek problem, we have

$$\Pr_j\big(L_{\mathcal{D}_j}(\widehat{w}) - L_{\mathcal{D}_j}(w^\star) < \rho^2\big) \leq \Pr_j(W = j).$$

The result follows by appealing to Theorem 2 and Theorem 3 in Shamir (2014). $\qquad\square$

B.3 Discussion: Support Recovery

Our lower bound for the sparse regression setting (5) does not rule out the possibility of sublinear-communication distributed algorithms for well-specified models. Here we sketch a strategy that works for this setting if we significantly strengthen the statistical assumptions.

Suppose that we work with the square loss and labels are realized as $y = \langle w^\star, x \rangle + \varepsilon$, where ε is conditionally mean-zero and $w^\star$ is k-sparse. Suppose in addition that the population covariance Σ has the restricted eigenvalue property, and that $w^\star$ satisfies the so-called "β-min" assumption: All non-zero coordinates of $w^\star$ have magnitude bounded below.

In this case, if $N/m = \Omega(k \log d)$ and the smallest non-zero coefficients of $w^\star$ are at least $\widetilde{\Omega}(\sqrt{m/N})$ the following strategy works: For each machine, run Lasso on the first half of the examples to exactly recover the support of $w^\star$ (e.g. Loh et al. (2017)). On the second half of examples, restrict to the recovered support and use the strategy from Zhang et al. (2012): run ridge regression on each machine locally with an appropriate choice of regularization parameter, then send all ridge regression estimators to a central server that averages them and returns this as the final estimator.

This strategy has $O(mk)$ communication by definition, but the assumptions on sparsity and β-min depend on the number of machines. How far can these assumptions be weakened?

C Proofs from Section 3

Throughout this section of the appendix we adopt the shorthand $B := B_1$ and $R := R_q$. Recall that $\frac{1}{p} + \frac{1}{q} = 1$.

To simplify expressions throughout the proofs in this section we use the convention $\widehat{w}^0 := \bar{w}$ and $\widetilde{w}^i := w_{n+1}^i$.

We begin the section by stating a few preliminary results used to analyze the performance of Algorithm 2 and Algorithm 3. We then proceed to prove the main theorems.

For the results on fast rates we need the following intermediate fact, which states that centering the regularizer $\mathcal{R}$ at $\bar{w}$ does not change the strong convexity from Proposition 3 or smoothness properties established in Appendix A.2.

Proposition 7. *Let $\mathcal{R}(w) = \frac{1}{2}\|w - \bar{w}\|_p^2$, where $\|w\|_1 \le B$. Then $D_\mathcal{R}(a\|b) \ge \frac{p-1}{2}\|a - b\|_p^2$ and if $\|a\|_1 \vee \|b\|_1 \vee \|c\|_1 \le B$ it holds that*

$$D_\mathcal{R}(c\|a) - D_\mathcal{R}(c\|b) \le 10B\|a - b\|_p + 16B^{3-p}\|a - b\|_\infty^{p-1}.$$

Proof of Proposition 7. Let $\mathcal{R}_0(w) = \frac{1}{2}\|w\|_p^2$. The result follows from Proposition 3 and Theorem 7 by simply observing that $\nabla\mathcal{R}(w) = \nabla\mathcal{R}_0(w - \bar{w})$ so that $D_\mathcal{R}(w\|w') = D_{\mathcal{R}_0}(w - \bar{w}\|w' - \bar{w})$. To invoke Theorem 7 we use that $\|a - \bar{w}\|_1 \le 2B$, and likewise for b and c. $\square$

Lemma 7. Algorithm 2 guarantees that for any adaptively selected sequence ∇_t^i and all $w^\star \in \mathcal{W}$, any individual machine $i \in [m]$ deterministically satisfies the following guarantee:

$$\sum_{t=1}^{n}\langle\nabla_t^i, w_t^i - w^\star\rangle \le \frac{\eta C_q}{2}\sum_{t=1}^{n}\|\nabla_t^i\|_q^2 + \frac{1}{\eta}\big(D_\mathcal{R}(w^\star\|w_1^i) - D_\mathcal{R}(w^\star\|w_{n+1}^i)\big)$$

Proof of Lemma 7. This is a standard argument. Let $w^\star \in \mathcal{W}$ be fixed. The standard Bregman divergence inequality for mirror descent (Ben-Tal and Nemirovski, 2001) implies that for every time t, we have

$$\langle\nabla_t^i, w_t^i - w^\star\rangle \le \langle\nabla_t^i, w_t^i - \theta_{t+1}^i\rangle + \frac{1}{\eta}\big(D_\mathcal{R}(w^\star\|w_t^i) - D_\mathcal{R}(w^\star\|w_{t+1}^i) - D_\mathcal{R}(w_t^i\|\theta_{t+1}^i)\big).$$

Using Proposition 7, we have an upper bound of

$$\langle\nabla_t^i, w_t^i - \theta_{t+1}^i\rangle + \frac{1}{\eta}\Big(D_\mathcal{R}(w^\star\|w_t^i) - D_\mathcal{R}(w^\star\|w_{t+1}^i) - \frac{p-1}{2}\|w_t^i - \theta_{t+1}^i\|_p^2\Big).$$

Using Hölder's inequality and AM-GM:

$$\le \frac{\eta}{2(p-1)}\|\nabla_t^i\|_q^2 + \frac{p-1}{2\eta}\|w_t^i - \theta_{t+1}^i\|_p^2 + \frac{1}{\eta}\Big(D_\mathcal{R}(w^\star\|w_t^i) - D_\mathcal{R}(w^\star\|w_{t+1}^i) - \frac{p-1}{2}\|w_t^i - \theta_{t+1}^i\|_p^2\Big)$$

$$= \frac{\eta}{2(p-1)}\|\nabla_t^i\|_q^2 + \frac{1}{\eta}\big(D_\mathcal{R}(w^\star\|w_t^i) - D_\mathcal{R}(w^\star\|w_{t+1}^i)\big).$$

The result follows by summing across time and observing that the Bregman divergences telescope. $\square$

Proof of Theorem 3. To begin, the guarantee from Lemma 7 implies that for any fixed machine i, deterministically,

$$\sum_{t=1}^{n}\langle\nabla_t^i, w_t^i - w^\star\rangle \le \frac{\eta C_q}{2}\sum_{t=1}^{n}\|\nabla_t^i\|_q^2 + \frac{1}{\eta}\big(D_\mathcal{R}(w^\star\|w_1^i) - D_\mathcal{R}(w^\star\|w_{n+1}^i)\big).$$

We now use the usual reduction from regret to stochastic optimization: since w_t^i does not depend on ∇_t^i, we can take expectation over ∇_t^i to get

$$\mathbb{E}\Big[\sum_{t=1}^{n}\langle\nabla L_\mathcal{D}(w_t^i), w_t^i - w^\star\rangle\Big] \le \frac{\eta C_q}{2}\sum_{t=1}^{n}\mathbb{E}\|\nabla_t^i\|_q^2 + \frac{1}{\eta}\mathbb{E}\big[D_\mathcal{R}(w^\star\|w_1^i) - D_\mathcal{R}(w^\star\|w_{n+1}^i)\big]$$

and furthermore, $L_{\mathcal{D}}$ is convex, this implies

$$\mathbb{E}\left[\sum_{t=1}^{n} L_{\mathcal{D}}(w_t^i) - L_{\mathcal{D}}(w^\star)\right] \leq \frac{\eta C_q}{2} \sum_{t=1}^{n} \mathbb{E}\left\|\nabla_t^i\right\|_q^2 + \frac{1}{\eta}\mathbb{E}\left[D_{\mathcal{R}}(w^\star\|w_1^i) - D_{\mathcal{R}}(w^\star\|w_{n+1}^i)\right].$$

While the regret guarantee implies that this holds for each machine i conditioned on the history up until the machine begins working, it suffices for our purposes to interpret the expectation above as with respect to all randomness in the algorithm's execution except for the randomness in sparsification for the final iterate $\widehat{w}$.

We now sum this guarantee across all machines, which gives

$$\mathbb{E}\left[\sum_{i=1}^{m}\sum_{t=1}^{n} L_{\mathcal{D}}(w_t^i) - L_{\mathcal{D}}(w^\star)\right] \leq \frac{\eta C_q}{2} \sum_{i=1}^{m}\sum_{t=1}^{n} \mathbb{E}\left\|\nabla_t^i\right\|_q^2 + \frac{1}{\eta}\sum_{i=1}^{m}\mathbb{E}\left[D_{\mathcal{R}}(w^\star\|w_1^i) - D_{\mathcal{R}}(w^\star\|w_{n+1}^i)\right].$$

Rewriting in terms of $\widetilde{w}^i$ and its sparsified version $\widehat{w}^i$ and using that $w_1^1 = \bar{w}$, this is upper bounded by

$$\leq \frac{\eta C_q}{2} \sum_{i=1}^{m}\sum_{t=1}^{n} \mathbb{E}\left\|\nabla_t^i\right\|_q^2 + \frac{D_{\mathcal{R}}(w^\star\|\bar{w})}{\eta} + \frac{1}{\eta}\sum_{i=1}^{m-1}\mathbb{E}\left[D_{\mathcal{R}}(w^\star\|\widehat{w}^i) - D_{\mathcal{R}}(w^\star\|\widetilde{w}^i)\right].$$

We now bound the approximation error in the final term. Using Proposition 7, we get

$$\sum_{i=1}^{m-1}\mathbb{E}\left[D_{\mathcal{R}}(w^\star\|\widehat{w}^i) - D_{\mathcal{R}}(w^\star\|\widetilde{w}^i)\right] \leq O\left(\sum_{i=1}^{m-1} B\,\mathbb{E}\left\|\widehat{w}^i - \widetilde{w}^i\right\|_p + B^{3-p}\,\mathbb{E}\left\|\widehat{w}^i - \widetilde{w}^i\right\|_\infty^{p-1}\right).$$

Theorem 6 implies that $\mathbb{E}\left\|\widehat{w}^i - \widetilde{w}^i\right\|_p \leq O\left(B\left(\frac{1}{s}\right)^{1-\frac{1}{p}}\right)$ and $\mathbb{E}\left\|\widehat{w}^i - \widetilde{w}^i\right\|_\infty^{p-1} \leq O\left(B^{p-1}\left(\frac{1}{s}\right)^{\frac{p-1}{2}}\right)$.[8] In particular, we get

$$\sum_{i=1}^{m-1}\mathbb{E}\left[D_{\mathcal{R}}(w^\star\|\widehat{w}^i) - D_{\mathcal{R}}(w^\star\|\widetilde{w}^i)\right] \leq O\left(\sum_{i=1}^{m-1} B^2\left(\frac{1}{s}\right)^{1-\frac{1}{p}} + B^{3-p}\cdot B^{p-1}\left(\frac{1}{s}\right)^{\frac{1}{2}}\right) = O\left(B^2\sum_{i=1}^{m-1}\left(\frac{1}{s}\right)^{1-\frac{1}{p}} + \left(\frac{1}{s}\right)^{\frac{p-1}{2}}\right).$$

Since $p \leq 2$, the second summand dominates, leading to a final bound of $O\left(B^2 m\left(\frac{1}{s}\right)^{\frac{p-1}{2}}\right)$. To summarize, our developments so far (after normalizing by N) imply

$$\mathbb{E}\left[\frac{1}{mn}\sum_{i=1}^{m}\sum_{t=1}^{n} L_{\mathcal{D}}(w_t^i) - L_{\mathcal{D}}(w^\star)\right] \leq \frac{\eta C_q}{2N} \sum_{i=1}^{m}\sum_{t=1}^{n} \mathbb{E}\left\|\nabla_t^i\right\|_q^2 + \frac{D_{\mathcal{R}}(w^\star\|\bar{w})}{\eta N} + O\left(\frac{B^2 m}{\eta N}\left(\frac{1}{s}\right)^{\frac{p-1}{2}}\right).$$

Let $\widetilde{w}$ denote w_t^i for the index (i,t) selected uniformly at random in the final line of Algorithm 2. Interpreting the left-hand-side of this expression as a conditional expectation over $\widetilde{w}$, we get

$$\mathbb{E}[L_{\mathcal{D}}(\widetilde{w})] - L_{\mathcal{D}}(w^\star) \leq \frac{\eta C_q}{2N} \sum_{i=1}^{m}\sum_{t=1}^{n} \mathbb{E}\left\|\nabla_t^i\right\|_q^2 + \frac{D_{\mathcal{R}}(w^\star\|\bar{w})}{\eta N} + O\left(\frac{B^2 m}{\eta N}\left(\frac{1}{s}\right)^{\frac{p-1}{2}}\right). \tag{18}$$

Note that our boundedness assumptions imply $\left\|\nabla_t^i\right\|_q^2 \leq R^2$ and $D_{\mathcal{R}}(w^\star\|\bar{w}) = D_{\mathcal{R}}(w^\star\|0) \leq \frac{B^2}{2}$, so when $s = \Omega(m^{\frac{2}{p-1}})$ this is bounded by

$$\mathbb{E}[L_{\mathcal{D}}(\widetilde{w})] - L_{\mathcal{D}}(w^\star) \leq \frac{\eta C_q R^2}{2} + O\left(\frac{B^2}{\eta N}\right) \leq O(\sqrt{C_q B^2 R^2/N}),$$

---

[8]The second bound follows by appealing to the ℓ_2 case in Theorem 6 and using that $\|x\|_\infty \leq \|x\|_2$.

where the second inequality uses the choice of learning rate.

From here we split into two cases. In the general loss case, since $L_{\mathcal{D}}$ is R-Lipschitz with respect to ℓ_p (implied by the assumption that subgradients lie in $\mathcal{X}_q$ via duality), we get

$$L_{\mathcal{D}}(\widehat{w}) - L_{\mathcal{D}}(w^\star) \leq L_{\mathcal{D}}(\widetilde{w}) - L_{\mathcal{D}}(w^\star) + R\|\widehat{w} - \widetilde{w}\|_p.$$

We now invoke Theorem 6 once more, which implies that

$$\mathbb{E}\|\widehat{w} - \widetilde{w}\|_p \leq O\left(B\left(\frac{1}{s_0}\right)^{1-\frac{1}{p}}\right).$$

We see that it suffices to take $s_0 = \Omega((N/C_q)^{\frac{p}{2(p-1)}})$ to ensure that this error term is of the same order as the original excess risk bound.

In the linear model case, Lemma 1 directly implies that

$$\mathbb{E}\, L_{\mathcal{D}}(\widehat{w}) \leq L_{\mathcal{D}}(\widetilde{w}) + O(\sqrt{B^2 R^2/s_0}),$$

and so $s_0 = \Omega(N/C_q)$ suffices.

$\square$

Proof of Theorem 4. We begin from (18) in the proof of Theorem 3 which, once $s = \Omega(m^{\frac{2}{p-1}})$, implies

$$\mathbb{E}[L_{\mathcal{D}}(\widetilde{w})] - L_{\mathcal{D}}(w^\star) \leq \frac{\eta C_q}{2N} \sum_{i=1}^{m} \sum_{t=1}^{n} \mathbb{E}\|\nabla_t^i\|_q^2 + O\left(\frac{B^2}{\eta N}\right),$$

where $\widetilde{w}$ is the iterate w_t^i selected uniformly at random at the final step and the expectation is over all randomness except the final sparsification step. Since the loss $\ell(\cdot, z)$ is smooth, convex, and non-negative, we can appeal to Lemma 3.1 from Srebro et al. (2010), which implies that

$$\|\nabla_t^i\|_q^2 = \|\nabla \ell(w_t^i, z_t^i)\|_q^2 \leq 4\beta_q \ell(w_t^i, z_t^i).$$

Using this bound we have

$$\mathbb{E}[L_{\mathcal{D}}(\widetilde{w})] - L_{\mathcal{D}}(w^\star) \leq \frac{4\eta C_q \beta_q}{2N} \sum_{i=1}^{m} \sum_{t=1}^{n} \mathbb{E}\, \ell(w_t^i, z_t^i) + O\left(\frac{B^2}{\eta N}\right) = 2\eta C_q \beta_q \cdot \mathbb{E}[L_{\mathcal{D}}(\widetilde{w})] + O\left(\frac{B^2}{\eta N}\right).$$

Let $\varepsilon := 2\eta C_q \beta_q$. Rearranging, we write

$$(1-\varepsilon)\, \mathbb{E}[L_{\mathcal{D}}(\widetilde{w})] - L_{\mathcal{D}}(w^\star) \leq O\left(\frac{B^2}{2\eta N}\right).$$

When $\varepsilon < 1/2$, this implies $\mathbb{E}[L_{\mathcal{D}}(\widetilde{w})] - (1 + 2\varepsilon)L_{\mathcal{D}}(w^\star) \leq O\left(\frac{B^2}{2\eta N}\right)$, and so, by rearranging,

$$\mathbb{E}[L_{\mathcal{D}}(\widetilde{w})] - L_{\mathcal{D}}(w^\star) \leq O\left(\eta C_q \beta_q L^\star + \frac{B^2}{2\eta N}\right).$$

The choice $\eta = \sqrt{\frac{B^2}{C_q \beta_q L^\star N}} \wedge \frac{1}{4C_q \beta_q}$ ensures that $\varepsilon \leq 1/2$, and that

$$\eta C_q \beta_q L^\star + \frac{B^2}{2\eta N} = O\left(\sqrt{\frac{C_q \beta_q B^2 L^\star}{N}} + \frac{C_q \beta_q B^2}{N}\right).$$

Now, Lemma 3 implies that, conditioned on $\widetilde{w}$, we have $\mathbb{E}\,L_{\mathcal{D}}(\widehat{w}) \le L_{\mathcal{D}}(\widetilde{w}) + \frac{\beta_q B^2}{s_0}$. The choice $s_0 = \sqrt{\frac{\beta_q B^2 N}{C_q L^\star}} \wedge \frac{N}{C_q}$ guarantees that this approximation term is on the same order as the excess risk bound of $\widetilde{w}$. $\square$

Proposition 8. *Suppose we run Algorithm 2 with initial point $\bar{w}$ that is chosen by some randomized procedure independent of the data or randomness used by Algorithm 2. Suppose that we are promised that this selection procedure satisfies $\mathbb{E}\|\bar{w} - w^\star\|_p^2 \le \bar{B}^2$. Suppose that subgradients belong to $\mathcal{X}_q$ for $q \ge 2$, and that $\mathcal{W} \subseteq \mathcal{W}_1$. Then, using learning rate $\eta := \frac{\bar{B}}{R}\sqrt{\frac{1}{C_q N}}$, $s = \Omega\big(m^{2(q-1)}(B/\bar{B})^{4(q-1)}\big)$, and $s_0 = \Omega((N/C_q)^{\frac{q}{2}} \cdot (B/\bar{B})^q)$, the algorithm guarantees*

$$\mathbb{E}[L_{\mathcal{D}}(\widehat{w})] - L_{\mathcal{D}}(w^\star) \le O\left(\bar{B}R\sqrt{\frac{C_q}{N}}\right).$$

Proof of Proposition 8. We proceed exactly as in the proof of Theorem 3, which establishes that conditioned on $\bar{w}$,

$$\mathbb{E}[L_{\mathcal{D}}(\widehat{w})] - L_{\mathcal{D}}(w^\star) \le \frac{\eta C_q}{2N}\sum_{i=1}^{m}\sum_{t=1}^{n}\mathbb{E}\big\|\nabla_t^i\big\|_q^2 + \frac{D_{\mathcal{R}}(w^\star\|\bar{w})}{\eta N} + O\left(\frac{B^2 m}{\eta N}\left(\frac{1}{s}\right)^{\frac{p-1}{2}}\right) + O\left(BR\left(\frac{1}{s_0}\right)^{1-1/p}\right).$$

We now take the expectation over $\bar{w}$. We have that $\mathbb{E}\,D_{\mathcal{R}}(w^\star\|\bar{w}) = \frac{1}{2}\mathbb{E}\|\bar{w} - w^\star\|_p^2 \le \bar{B}^2/2$. It is straightforward to verify from here that the prescribed sparsity levels and learning rate give the desired bound. $\square$

Proof of Theorem 5. Let $\widehat{w}_0 = 0$, and let us use the shorthand $\gamma := \gamma_q$.

We will show inductively that $\mathbb{E}\|\widehat{w}_k - w^\star\|_p^2 \le 2^{-k}B^2 =: B_k^2$. Clearly this is true for $\widehat{w}_0$. Now assume the statement is true for $\widehat{w}_k$. Then, since $\mathbb{E}\|\widehat{w}_k - w^\star\|_p^2 \le B_k^2$, Proposition 8 guarantees that

$$\mathbb{E}[L_{\mathcal{D}}(\widehat{w}_{k+1})] - L_{\mathcal{D}}(w^\star) \le c \cdot B_k R\sqrt{\frac{C_q}{N_{k+1}}},$$

where $c > 0$ is some absolute constant. Since the objective satisfies the restricted strong convexity condition (Assumption 1), and since $L_{\mathcal{D}}$ is convex and $\mathcal{W}$ is also convex, we have $\langle \nabla L_{\mathcal{D}}(w^\star), w - w^\star\rangle \ge 0$ and so

$$\mathbb{E}\|\widehat{w}_{k+1} - w^\star\|_p^2 \le \frac{2c \cdot B_k R}{\gamma}\sqrt{\frac{C_q}{N_{k+1}}}.$$

Consequently, choosing $N_{k+1} = C_q \cdot \left(\frac{4cR}{\gamma B_k}\right)^2$ guarantees that

$$\mathbb{E}\|\widehat{w}_{k+1} - w^\star\|_p^2 \le \frac{1}{2}B_k^2,$$

so the recurrence indeed holds. In particular, this implies that

$$\mathbb{E}[L_{\mathcal{D}}(\widehat{w}_T)] - L_{\mathcal{D}}(w^\star) \le \frac{\gamma}{4}B_{T-1}^2 = 2^{-T}\cdot\frac{\gamma B^2}{2}.$$

The definition of T implies that

$$T \ge \log_2\left(\frac{N}{32 C_q}\left(\frac{\gamma B}{Rc}\right)^2\right),$$

30

and so

$$\mathbb{E}[L_{\mathcal{D}}(\widehat{w}_T)] - L_{\mathcal{D}}(w^\star) \le 2^{-T} \cdot \frac{\gamma B^2}{2} \le O\left(\frac{C_q R^2}{\gamma N}\right).$$

This proves the optimization guarantee.

To prove the communication guarantee, let m_k denote the number of consecutive machines used at round k. The total number of bits broadcasted—summing the sparsity levels from Proposition 8 over T rounds—is at most

$$\log d \cdot \sum_{k=1}^{T} (m_k)^{2q-1}\left(\frac{B}{B_{k-1}}\right)^{4(q-1)} + \left(\frac{N_k}{C_q}\right)^{\frac{q}{2}} \cdot \left(\frac{B}{B_{k-1}}\right)^q,$$

plus an additive $O(m \log(BRN))$ term to send the scalar norm for each sparsified iterate $\widehat{w}_i$. Note that we have $m_k = \frac{N_k}{n} \vee 1$, so this is at most

$$\log d \cdot \sum_{k=1}^{T} \left(\frac{N_k}{n}\right)^{2q-1}\left(\frac{B}{B_{k-1}}\right)^{4(q-1)} + \left(\frac{N_k}{C_q}\right)^{\frac{q}{2}} \cdot \left(\frac{B}{B_{k-1}}\right)^q.$$

The first term in this sum simplifies to $O\left(\log d \cdot \left(\frac{C_q R^2}{n\gamma^2 B^2}\right)^{2q-1}\right) \cdot \sum_{k=1}^{T} 2^{(4q-3)k}$, while the second simplifies to $O\left(\log d \cdot \left(\frac{R}{\gamma B}\right)^q 2^q\right) \cdot \sum_{t=1}^{T} 2^{qk}$. We use that $\sum_{t=1}^{T} \beta^t \le \beta^{T+1}$ for $\beta \ge 2$ to upper bound by

$$O\left(\log d \cdot \left(\frac{C_q R^2}{n\gamma^2 B^2}\right)^{2q-1} 2^q\right) \cdot 2^{(4q-3)T} + O\left(\log d\left(\frac{R}{\gamma B}\right)^q 2^q\right) \cdot 2^{qT}.$$

Substituting in the value of T and simplifying leads to a final bound of

$$O\left(\log d \cdot \left(\frac{\gamma^2 B^2}{C_q R^2}\right)^{2(q-1)} m^{2q-1} N^{2(q-1)} + \log d \cdot \left(\frac{\gamma BN}{C_q R}\right)^q\right). \tag{19}$$

□

Proof of Proposition 4. It immediately follows from the definitions in the proposition that Algorithm 3 guarantees

$$\mathbb{E}[L_{\mathcal{D}}(\widehat{w}_T)] - L_{\mathcal{D}}(w^\star) \le O\left(\frac{C_q B^2 R^2}{\gamma_q N}\right),$$

where γ_q is as in Assumption 1. We now relate γ_q and γ. From the optimality of $w^\star$ and strong convexity of the square loss with respect to predictions it holds that for all $w \in \mathcal{W}_p$,

$$\mathbb{E}[L_{\mathcal{D}}(w)] - L_{\mathcal{D}}(w^\star) - \langle \nabla L_{\mathcal{D}}(w^\star), w - w^\star \rangle \ge \mathbb{E}\langle x, w - w^\star \rangle^2.$$

Our assumption on γ implies

$$\mathbb{E}\langle x, w - w^\star \rangle^2 = \left\|\Sigma^{1/2}(w - w^\star)\right\|_2^2 \ge \gamma \|w - w^\star\|_2^2.$$

Using Proposition 9, we have

$$\|w - w^\star\|_p \le \|w - w^\star\|_1 \le 2\|(w - w^\star)_S\|_1 \le 2\sqrt{k}\|(w - w^\star)_S\|_2 \le 2\sqrt{k}\|w - w^\star\|_2$$

Thus, it suffices to take $\gamma_q = \frac{\gamma}{4k}$.

□

31

The following proposition is a standard result in high-dimensional statistics. For a given vector $w \in \mathbb{R}^d$, let $w_S \in \mathbb{R}^d$ denote the same vector with all coordinates outside $S \subseteq [d]$ set to zero.

Proposition 9. *Let $\mathcal{W}$, $w^\star$, and S be as in [Proposition 4](#). All $w \in \mathcal{W}$ satisfy the inequality $\|(w - w^\star)_{S^c}\|_1 \leq \|(w - w^\star)_S\|_1$.*

Proof of [Proposition 9](#). Let $\nu = w - w^\star$. From the definition of $\mathcal{W}$, we have that for all $w \in \mathcal{W}$,

$$\|w^\star\|_1 \geq \|w\|_1 = \|w^\star + \nu\|_1.$$

Applying triangle inequality and using that the ℓ_1 norm decomposes coordinate-wise:

$$\|w^\star + \nu\|_1 = \|w^\star + \nu_S + \nu_{S^c}\|_1 = \|w^\star + \nu_S\|_1 + \|\nu_{S^c}\|_1 \geq \|w^\star\|_1 - \|\nu_S\|_1 + \|\nu_{S^c}\|_1.$$

Rearranging, we get $\|\nu_{S^c}\|_1 \leq \|\nu_S\|_1$. □

Proof of [Proposition 5](#). To begin, we recall from [Kakade et al. (2012)](#) that the regularizer $\mathcal{R}(W) = \frac{1}{2}\|W\|_{S_p}^2$ is $(p-1)$-strongly convex for $p \leq 2$. This is enough to show under our assumptions that the centralized version of mirror descent (without sparsification) guarantees excess risk $O\left(\sqrt{\frac{C_q B_1^2 R_q^2}{N}}\right)$, with $C_q = q-1$, which matches the ℓ_1/ℓ_q setting.

What remains is to show that the new form of sparsification indeed preserves Bregman divergences as in the ℓ_1/ℓ_q setting. We now show that when W and $W^\star$ have $\|W\|_{S_1} \vee \|W^\star\|_{S_1} \leq B$,

$$\mathbb{E}[D_{\mathcal{R}}(W^\star \| Q^s(W)) - D_{\mathcal{R}}(W^\star \| W)] \leq O\left(B^2 \left(\frac{1}{s}\right)^{\frac{p-1}{2}}\right).$$

To begin, let $U \in \mathbb{R}^{d \times d}$ be the left singular vectors of W and $V \in \mathbb{R}^{d \times d}$ be the right singular vectors. We define $\widehat{\sigma} = \frac{\|W\|_{S_1}}{s} \sum_{\tau=1}^s e_{i_\tau}$, so that we can write $W = U\mathrm{diag}(\sigma)V^\top$ and $Q^s(W) = U\mathrm{diag}(\widehat{\sigma})V^\top$.

Now note that since the Schatten norms are unitarily invariant, we have

$$\|W - Q^s(W)\|_{S_p} = \|U\mathrm{diag}(\sigma - \widehat{\sigma})V^\top\|_{S_p} = \|\sigma - \widehat{\sigma}\|_p$$

for any p. Note that our assumptions imply that $\|\sigma\|_1 \leq B$, and that $\widehat{\sigma}$ is simply the vector Maurey operator applied to σ, so it follows immediately from [Theorem 6](#) that

$$\mathbb{E}\|\sigma - \widehat{\sigma}\|_p \leq 4B\left(\frac{1}{s}\right)^{1-1/p} \quad \text{and} \quad \sqrt{\mathbb{E}\|\sigma - \widehat{\sigma}\|_\infty^2} \leq 4B\left(\frac{1}{s}\right)^{1/2}. \tag{20}$$

Returning to the Bregman divergence, we write

$$\begin{aligned}
D_{\mathcal{R}}(W^\star \| Q^s(W)) - D_{\mathcal{R}}(W^\star \| W) &= D_{\mathcal{R}}(W \| Q^s(W)) + \langle \nabla\mathcal{R}(Q^s(W)) - \nabla\mathcal{R}(W), W - W^\star \rangle \\
&\leq D_{\mathcal{R}}(W \| Q^s(W)) + \|\nabla\mathcal{R}(Q^s(W)) - \nabla\mathcal{R}(W)\|_{S_\infty} \|W - W^\star\|_{S_1} \\
&\leq D_{\mathcal{R}}(W \| Q^s(W)) + 2B\|\nabla\mathcal{R}(Q^s(W)) - \nabla\mathcal{R}(W)\|_{S_\infty}.
\end{aligned}$$

It follows immediately using [Lemma 4](#) that

$$D_{\mathcal{R}}(W \| Q^s(W)) \leq 3B\|W - Q^s(W)\|_{S_p} = 3B\|\sigma - \widehat{\sigma}\|_p.$$

To make progress from here we use a useful representation for the gradient of $\mathcal{R}$. Define

$$g(\sigma) = \|\sigma\|_p^{2-p} \cdot \left(|\sigma_1|^{p-1}\mathrm{sgn}(\sigma_1), \ldots, |\sigma_d|^{p-1}\mathrm{sgn}(\sigma_d)\right).$$

Then using Theorem 30 from Kakade et al. (2012) along with (13), we have

$$\nabla\mathcal{R}(W) = U\mathrm{diag}(g(\sigma))V^\top, \quad \text{and} \quad \nabla\mathcal{R}(Q^s(W)) = U\mathrm{diag}(g(\widehat{\sigma}))V^\top.$$

For the gradient error term, unitary invariance again implies that

$$\|\nabla\mathcal{R}(Q^s(W)) - \nabla\mathcal{R}(W)\|_{S_\infty} = \left\|U\mathrm{diag}(g(\sigma) - g(\widehat{\sigma}))V^\top\right\|_{S_\infty} = \|g(\sigma) - g(\widehat{\sigma})\|_\infty.$$

Lemma 6 states that

$$\|g(\sigma) - g(\widehat{\sigma})\|_\infty \le 2B^{2-p}\|\sigma - \widehat{\sigma}\|_\infty^{p-1} + \|\sigma - \widehat{\sigma}\|_p.$$

Putting everything together, we get

$$D_{\mathcal{R}}(W^\star\|Q^s(W)) - D_{\mathcal{R}}(W^\star\|W) \le 5B\|\sigma - \widehat{\sigma}\|_p + 4B^{3-p}\|\sigma - \widehat{\sigma}\|_\infty^{p-1}.$$

The desired result follows by plugging in the bounds in (20). $\qquad\square$