JMLR: Workshop and Conference Proceedings vol 40:1–24, 2015

Minimax rates for memory-bounded sparse linear regression

Jacob Steinhardt JSTEINHARDT@CS.STANFORD.EDU
Stanford University, Department of Computer Science

John Duchi JDUCHI@STANFORD.EDU
Stanford University, Department of Statistics

Abstract

We establish a minimax lower bound of $\Omega\left(\frac{kd}{B\epsilon}\right)$ on the sample size needed to estimate parameters in a k-sparse linear regression of dimension d under memory restrictions to B bits, where ϵ is the ℓ_2 parameter error. When the covariance of the regressors is the identity matrix, we also provide an algorithm that uses $\tilde{O}(B+k)$ bits and requires $\tilde{O}(\frac{kd}{B\epsilon^2})$ observations to achieve error ϵ. Our lower bound holds in a more general communication-bounded setting, where instead of a memory bound, at most B bits of information are allowed to be (adaptively) communicated about each sample.

1. Introduction

The growth in size and scope of datasets underscores a need for techniques that balance between multiple practical desiderata in statistical inference and learning, as classical tools are insufficient for providing such tradeoffs. In this work, we build on a nascent theory of resource-constrained learning—one in which researchers have investigated privacy (Kasiviswanathan et al., 2011; Duchi et al., 2013), communication and memory (Zhang et al., 2013; Garg et al., 2014; Shamir, 2014), and computational constraints (Berthet and Rigollet, 2013)—and study bounds on the performance of regression estimators under memory and communication constraints. In particular, we provide minimax upper and lower bounds that exhibit a tradeoff between accuracy and available resources for a broad class of regression problems, which have to this point been difficult to characterize.

To situate our work, we briefly review existing results on statistical learning with resource constraints. There is a considerable literature on computation and learning, beginning with Valiant's work (1984) on probably approximately correct (PAC) learning, which separates concept learning in poynomial versus non-polynomial settings. More recent work shows (under natural complexity-theoretic assumptions) that restriction to polynomial-time procedures increases sample complexity for several problems, including sparse principal components analysis (Berthet and Rigollet, 2013) classification problems (Daniely et al., 2013, 2014), and, pertinent to this work, sparse linear regression (Zhang et al., 2014; Natarajan, 1995). Showing fine-grained tradeoffs in this setting has been challenging, however. Researchers have given more explicit guarantees in other resource-constrained learning problems; Duchi et al. (2013) study estimation under a *privacy constraint* that the statistician should never be able to infer more than a small amount about any individual in a sample. For several tasks, including mean estimation and regression, they establish matching upper and lower bounds exhibiting a tradeoff between privacy and statistical accuracy. In the communication-limited setting, Zhang et al. (2013) study problems where data are stored on multiple machines, and the goal is to minimize communication between machines and a centralized fusion center. In the full version of that paper (Duchi et al., 2014), they show that mean-squared error for d-dimensional mean estimation has lower-bound $\Omega(\frac{d^2}{Bmn\log(m)})$ for m machines, each with n observations, and B bits of communication per machine. In concurrent work, Garg et al. (2014) provide identical results.

Perhaps most related to our work, Shamir (2014) considers communication constraints in the online learning setting, where at most B bits of information about an example can be transmitted to the algorithm before it moves on to the next example. In this case, any memory-bounded algorithm is communication-bounded, and so communication lower bounds imply memory lower bounds (Alon et al., 1999). For d-dimensional 1-sparse (i.e. only a single non-zero component) mean estimation problems, Shamir (2014) shows that $\Omega(d/B)$ observations are necessary for parameter recovery, while without memory constraints, a sample of size $O(\log d)$ is sufficient. He also shows that for certain principal component analyses, a sample of size $\Omega\left(d^4/B\right)$ is necessary for estimation, while (without constraints) $O(d^2\log^3(d))$ observations suffice.

1.1. Outline of results

In this work, we also focus on the online setting. We work in a regression model, in which we wish to estimate an unknown d-dimensional parameter vector $w^* \in \mathbb{R}^d$, and we observe an i.i.d. sequence $(X^{(i)}, Y^{(i)}) \in \mathbb{R}^d \times \mathbb{R}$ such that

$$Y^{(i)} = \langle w^*, X^{(i)}\rangle + \epsilon^{(i)}, \qquad (1)$$

where $\epsilon^{(i)}$ is mean-zero noise independent of $X^{(i)}$ with $\mathrm{Var}(\epsilon^{(i)}) \le \sigma^2$. We focus on the case that w^* is k-*sparse*, meaning that it has at most k non-zero entries, that is, $\|w^*\|_0 \le k$.

We consider resource-constrained procedures that may store information about the ith observation in a B_i-bit vector $Z^{(i)} \in \{0,1\}^{B_i}$. After n observations, the procedure outputs an estimate $\widehat{w}$ of w^*. We have two types of resource constraints:

Communication constraints $Z^{(i)}$ is a measurable function of $X^{(i)}$, $Y^{(i)}$, and the history $Z^{(1:i-1)}$, and the estimate $\widehat{w}$ is a measurable function of $Z^{(1:n)} = (Z^{(1)}, \dots, Z^{(n)})$.

Memory constraints $Z^{(i)}$ is a measurable function of $X^{(i)}$, $Y^{(i)}$, and $Z^{(i-1)}$, and $\widehat{w}$ is a measurable function of only $Z^{(n)}$.

Note that the communication constraints do not allow for arbitrary communication, but rather only "one-pass" communication protocols in which all information is communicated from left to right.

Any memory-constrained procedure is also communication-constrained, so we prove our lower bounds in the weaker (less restrictive) communication-constrained setting, later providing an algorithm for the memory-constrained setting (which then carries through to the communication-constrained setting). Our two main results (Theorems 2 and 3) are that, for a variety of regression problems with observation noise σ, the mean-squared error satisfies (subject to certain assumptions on the covariates X and noise ϵ)

$$\Omega(1)\cdot\sigma^2\min\left\{\frac{kd}{Bn}, 1\right\} \le \inf_{\widehat{w}}\sup_{w^*\in\mathcal{W}} \mathbb{E}\left[\|\widehat{w}-w^*\|_2^2\right] \le \widetilde{O}(1)\cdot\sigma^2\max\left\{\sqrt{\frac{kd}{Bn}}, \frac{kd}{Bn}\right\}, \qquad (2)$$

where the infimum is over all communication/memory-constrained procedures using B bits per iteration, and the supremum is over a set $\mathcal{W}$ of k-sparse vectors to be defined later. The lower bound (2) shows that any B-bit communication-constrained procedure achieving squared error ϵ

requires $\Omega(\frac{\sigma^2}{\epsilon}\frac{kd}{B})$ observations. In the absence of resource constraints, standard results on sparse regression (e.g. Wainwright, 2009) show that $\Theta(\frac{1}{\epsilon}k\log d)$ observations suffice; moreover, algorithms which require only $\tilde{O}(d)$ memory can essentially match this bound (Agarwal et al., 2012; Steinhardt et al., 2014). Thus, for any communication budget $B \le d^\alpha$ with $\alpha < 1$, we lose exponentially in the dimension. The upper bound also has sharp dependence on B and d, though worse dependence on accuracy and variance: it shows that $\tilde{O}(\frac{\sigma^4}{\epsilon^2}\frac{kd}{B})$ observations suffice.

1.2. Summary of techniques

Before continuing, we summarize the techniques used to prove the bounds (2). The high-level structure of the lower bound follows that of Arias-Castro et al. (2013), who show lower bounds for sparse regression with an adaptive design matrix. They adapt Assouad's method (1983) to reduce the analysis to bounding the mutual information between a single observation and the parameter vector w^*. In our case, even this is challenging; rather than choosing a design matrix, the procedure chooses an arbitrary B-bit message. We control such procedures using two ideas. First, we focus on the fully sparse ($k = 1$) case, showing a *quantitative data processing inequality*, based off of techniques developed by Duchi et al. (2013), Zhang et al. (2013), and Shamir (2014). By analyzing certain carefully constructed likelihood ratio bounds, we show, roughly, that if the pair $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ provides information I^* about the parameter w^*, then any B-bit message Z based on (X, Y) provides information at most BI^*/d about w^*; see Section 2 for these results. After this data processing bound for $k = 1$, in Section 3 we develop a *direct sum* theorem, based on the ideas of Garg et al. (2014) and a common tool in the communication complexity literature, which shows that finding a k-sparse vector w^* is as hard as finding k independent 1-sparse vectors.

While our lower bound builds off of several prior results, the regression setting introduces a subtle and challenging technical issue: both the direct sum technique and prior quantitative data processing inequalities in communication settings rely strongly on the assumption that each observation consists of d independent random variables. As Y necessarily depends on X in the model (1), we have strong coupling of our observations. To address this, we develop a technique that allows us to treat Y as "part of the communication"; instead of Z communicating information about X and Y, now Y and Z communicate information about X. The effective number of bits of communication increases by (an analogue of) the entropy of Y.

For the upper bound (2), in Section 4 we show how to implement ℓ_1-regularized dual averaging (Xiao, 2010) using a count sketch data structure (Charikar et al., 2002), which space-efficiently counts elements in a data stream. We use the count sketch structure to maintain a coarse estimate of model parameters, while also exactly storing a small active set of at most k coordinates. These combined estimates allow us to implement dual averaging when the ℓ_1-regularization is sufficiently large; the necessary amount of ℓ_1-regularization is inversely proportional to the amount of memory needed by the count sketch structure, leading to a tradeoff between memory and statistical efficiency.

Notation We perpetrate several abuses. Superscripts index iterations and subscripts dimensions, so $X_j^{(i)}$ denotes the coordinate j of the ith X vector. We let $D_{\rm kl}\left(P \,\|\, Q\right)$ denote the KL-divergence between P and Q, and use upper case to denote unbound variables and lower case to denote fixed conditioning variables, e.g., $D_{\rm kl}\left(P(X \mid z) \,\|\, Q(X \mid z)\right)$ denotes the KL divergence of the distributions of X under P and Q conditional on $Z = z$, and $I(X; Z)$ is the mutual information between X and Z. We use standard big-Oh notation, where $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ denote bounds holding up to polylogarithmic factors. We let $[n] = \{1, \ldots, n\}$.

2. Lower bound when $k = 1$

We begin our analysis in a restricted and simpler setting than the full one we consider in the sequel, providing a lower bound for estimation in the regression model (1) when w^* is 1-sparse. Our lower bound holds even in the presence of certain types of side information, which is important for our later analysis.

We begin with a formal description of the setting. Let $r > 0$ and let W be uniformly random in the set $\{-r, 0, r\}^d$, where we constrain $\|W\|_0 = 1$, i.e. only a single coordinate of W is non-zero. At iteration i, the procedure observes the triple $(X^{(i)}, \xi^{(i)}, Y^{(i)}) \in \{-1, 1\}^d \times \Xi \times \mathbb{R}$. Here $X^{(i)}$ is an i.i.d. sequence of vectors uniform on $\{-1, 1\}^d$, $\xi^{(i)}$ is side information independent of $X^{(1:n)}$ and $Z^{(1:i-1)}$ (which is necessary to consider for our direct sum argument later), and we observe

$$Y^{(i)} = W^\top X^{(i)} + \xi^{(i)} + \epsilon^{(i)}, \tag{3}$$

where $\epsilon^{(i)}$ are i.i.d. mean-zero Laplace distributed with $\mathrm{Var}(\epsilon) = \sigma^2$. Additionally, we let P_0 be a "null" distribution in which we set $Y^{(i)} = \xi^{(i)} + rs^{(i)} + \epsilon^{(i)}$, where $s^{(i)} \in \{-1, 1\}$ is an i.i.d. sequence of Rademacher variables; note that P_0 is constructed to have the same marginal over $Y^{(i)}$ as the true distribution. We present our lower bound in terms of *average communication*, defined as

$$\widetilde{B}_0 \stackrel{\mathrm{def}}{=} \frac{1}{n} \sum_{i=1}^{n} I_{P_0}(X^{(i)}; Y^{(i)}, Z^{(i)} \mid Z^{(1:i-1)}), \tag{4}$$

where I_{P_0} denotes mutual information under P_0. With this setup in place, we have:

Theorem 1 *Let $\delta = 2\sqrt{2}r/\sigma$. Under the setting above, for any estimator $\widehat{W}(Z^{(1)}, \ldots, Z^{(n)})$ of the random vector W, we have*

$$\mathbb{E}\left[\|\widehat{W} - W\|_2^2\right] \geq \frac{r^2}{2}\left(\frac{1}{2} - \sqrt{\frac{e^\delta(e^\delta - 1)^2 \widetilde{B}_0 n}{d}}\right).$$

To interpret Theorem 1, we study $\widetilde{B}_0$ under the assumption that the procedure simply observes pairs $(X^{(i)}, Y^{(i)})$ from the regression model (1), that is, $\xi^{(i)} \equiv 0$, and we may store/communicate only B bits. By the chain rule for information and the fact that conditioning reduces entropy, we then have

$$\begin{aligned} I_{P_0}(X^{(i)}; Y^{(i)}, Z^{(i)} \mid Z^{(1:i-1)}) &= I_{P_0}(X^{(i)}; Y^{(i)} \mid Z^{(1:i-1)}) + I_{P_0}(X^{(i)}; Z^{(i)} \mid Y^{(i)}, Z^{(1:i-1)}) \\ &\stackrel{(i)}{=} I_{P_0}(X^{(i)}; Z^{(i)} \mid Y^{(i)}, Z^{(1:i-1)}) \leq H_{P_0}(Z^{(i)}) \leq B, \end{aligned}$$

where (i) uses the fact that $X^{(i)}$ and $Y^{(i)}$ are independent under P_0. Now consider the signal-to-noise ratio $\delta = 2\sqrt{2}r/\sigma$, which we may vary by scaling the radius r. Noting that $e^\delta(e^\delta - 1)^2 \leq 2\delta^2$ for $\delta \leq 1/3$, Theorem 1 implies the lower bound

$$\mathbb{E}\left[\|\widehat{W} - W\|_2^2\right] \geq \sup_{0 \leq \delta \leq 1/3} \frac{\sigma^2\delta^2}{16}\left(\frac{1}{2} - \sqrt{\frac{2\delta^2 Bn}{d}}\right) \geq \frac{\sigma^2}{64} \min\left\{\frac{1}{32}\frac{d}{Bn}, \frac{1}{9}\right\},$$

where we have chosen $\delta = \min\{1/3, \sqrt{d/(32Bn)}\}$. Thus, to achieve mean-squared error smaller than σ^2, we require $n \gtrsim d/B$ observations. This contrasts with the unbounded memory case, where n scales only logarithmically in the dimension d if we use the Lasso algorithm (Wainwright, 2009).

2.1. Proof of Theorem 1

We prove the theorem using an extension of Assouad's method, which transforms the minimax lower bounding problem into one of multiple hypothesis tests (Assouad, 1983; Arias-Castro et al., 2013; Shamir, 2014). We extend a few results of Arias-Castro et al. and Shamir to apply in our slightly more complex setting.

We introduce a bit of additional notation before continuing. Let $J \in \{\pm 1, \pm 2, \ldots, \pm d\}$ be a random index (and sign) corresponding to the vector W, so that $J = j$ means that $W_j = r$ and $J = -j$ means that $W_j = -r$. Let P_{+j}, P_{-j} denote the joint distribution of all variables conditioned on $J = +j, -j$ respectively, and let P_0 be the null distribution as defined earlier. Throughout, we use lower-case p to denote the density of P with respect to a fixed base measure.

Overview. We prove Theorem 1 in three steps. First, we show that the minimax estimation error $\|\hat{W} - W\|_2^2$ can be lower-bounded in terms of the recovery error $\mathbb{P}[\hat{J} \neq J]$ (Lemma 1). Next, we show that recovering J is difficult unless $Z^{(1:n)}$ contains substantial information about J (Lemma 2). Finally, we reach the crux of our argument, which is to establish a *strong data processing inequality* (Lemma 4 and equation (9)), which shows that, for the Markov chain $W \to (X, Y) \to Z$, the mutual information $I(W; Z)$ degrades by a factor of d/B from the information $I(X, Y; Z)$; this shows that classical minimax bounds increase by a factor of d/B in our setting.

Estimation to testing to information. We begin by bounding squared error by testing error:

Lemma 1 *For any estimator $\hat{W}$, there is an estimator $\hat{J}$ such that*

$$\mathbb{E}\left[\|\hat{W} - W\|_2^2\right] \geq \frac{r^2}{2}\mathbb{P}\left[\hat{J} \neq J\right]. \tag{5}$$

We next state a lower bound on the probability of error in a hypothesis test.

Lemma 2 *Let $J \sim \mathrm{Uniform}(\{\pm 1, \ldots, \pm d\})$ and $K_{\pm j} \overset{\text{def}}{=} D_{\text{kl}}\left(P_0(Z^{(1:n)}) \parallel P_{\pm j}(Z^{(1:n)})\right)$. Then for any estimator $\widehat{J}(Z^{(1:n)})$*

$$\mathbb{P}(\widehat{J}(Z^{(1:n)}) \neq J) \geq \left(1 - \frac{1}{2d}\right) - \sqrt{\frac{1}{4d}\sum_{j=1}^{d}(K_{-j} + K_{+j})}. \tag{6}$$

The proofs of Lemmas 1 and 2 are in Sec. A.1 and A.2, respectively. We next provide upper bounds on $K_{\pm j}$ from Lemma 2, focusing on the $+j$ term as the $-j$ term is symmetric. By the chain rule,

$$D_{\text{kl}}\left(P_0(Z^{(1:n)}) \parallel P_{+j}(Z^{(1:n)})\right) = \sum_{i=1}^{n} \int D_{\text{kl}}\left(P_0(Z^{(i)} \mid z^{(1:i-1)}) \parallel P_{+j}(Z^{(i)} \mid z^{(1:i-1)})\right) dP_0(z^{(1:i-1)}).$$

For notational simplicity, introduce the shorthand $Z = Z^{(i)}$ and $\widehat{z} = z^{(1:i-1)}$; we will focus on bounding $D_{\text{kl}}\left(P_0(Z \mid \widehat{z}) \parallel P_{+j}(Z \mid \widehat{z})\right)$ for a fixed i and $\widehat{z} = z^{(1:i-1)}$.

A strong data-processing inequality. We now relate the divergence between $P_0(Z)$ and $P_{+j}(Z)$ to that for the distributions of (X, Y) (i.e., the divergence if there were no memory or communication constraints). As in Theorem 1, let $\delta = 2\sqrt{2}r/\sigma$ be the signal to noise ratio. Note that

$$p(x, \xi, y \mid \widehat{z}) = p(x, \xi \mid \widehat{z})p(y \mid x, \xi, \widehat{z}) = p(x, \xi)p(y \mid x, \xi, \widehat{z})$$

for $p = p_0, p_{\pm j}$, as the pair (x, ξ) is independent of the index J and the past (stored) data $\widehat{z}$. Thus

$$|\log p_{+j}(x, \xi, y \mid \widehat{z}) - \log p_0(x, \xi, y \mid \widehat{z})| = |\log p_{+j}(y \mid x, \xi, \widehat{z}) - \log p_0(y \mid x, \xi, \widehat{z})| \le \delta, \quad (7)$$

as the distribution of $p_{+j}(y|x, \xi, \widehat{z}, s)$ is a mean shift of at most $2r$ relative to $p_0(y|x, \xi, s)$, and both distributions are $\mathrm{Laplace}(\sigma/\sqrt{2})$ about their mean (recall that s is the Rademacher variable in the definition of p_0; marginalizing it out as in (7) can only bring the densities closer). Leveraging (7), we obtain the following lemma, which bounds the KL-divergence in terms of the χ^2-divergence.

Lemma 3 *For any index j and past $\widehat{z}$,*

$$D_{\mathrm{kl}}\left(P_0(Z \mid \widehat{z}) \,\|\, P_{+j}(Z \mid \widehat{z})\right) \le \int \frac{|p_{+j}(z \mid \widehat{z}) - p_0(z \mid \widehat{z})|^2}{p_{+j}(z \mid \widehat{z})} d\mu(z) \le e^{\delta} \int \frac{|p_{+j}(z \mid \hat{z}) - p_0(z \mid \hat{z})|^2}{p_0(z \mid \widehat{z})} d\mu(z).$$

Proof The first inequality is the standard bound of KL-divergence in terms of χ^2-divergence (cf. Tsybakov, 2009, Lemma 2.7). The second follows from inequality (7), as

$$\begin{aligned} p_{+j}(z \mid \widehat{z}) = \int p_{+j}(z \mid \xi, x, y, \widehat{z}) dP_{+j}(x, \xi, y \mid \widehat{z}) &= \int p_0(z \mid \xi, x, y, \widehat{z}) dP_{+j}(x, \xi, y \mid \widehat{z}) \\ &\ge e^{-\delta} \int p_0(z \mid \xi, x, y, \widehat{z}) dP_0(x, \xi, y \mid \widehat{z}) = e^{-\delta} p_0(z \mid \widehat{z}), \end{aligned}$$

which gives the desired result. ■

Picking up from Lemma 3, we analyze $|p_{+j}(z \mid \hat{z}) - p_0(z \mid \hat{z})|$ in Lemma 4, which is the key technical lemma in this section (all results so far, while non-trivial, are standard results in the literature).

Lemma 4 (Information contraction) *For any $\hat{z}$ and z, we have*

$$|p_{+j}(z \mid \widehat{z}) - p_0(z \mid \widehat{z})| \le (e^{\delta} - 1) p_0(z \mid \widehat{z}) \int \sqrt{2 D_{\mathrm{kl}}\left(P_0(X_j \mid y, z, \widehat{z}) \,\|\, P_0(X_j \mid y, \widehat{z})\right)} dP_0(y \mid z, \widehat{z}).$$

Lemma 4 is proved in Sec. A.3. The intuition behind the proof is that, since $p_{+j}(y|\hat{z}) = p_0(y|\hat{z})$, the only way for z to distinguish between 0 and $+j$ is by storing information about x_j (information about $x_{\neg j}$ is useless since it doesn't affect the distribution over y). The amount of information about x_j is measured by the KL divergence term in the bound; the $e^{\delta} - 1$ term appears because even when $x_{\neg j}$ is known, the distributions over y differ by a factor of at most e^{δ}.

Proving Theorem 1. Combining Lemmas 3 and 4, we bound $D_{\mathrm{kl}}\left(P_0(Z \mid \hat{z}) \,\|\, P_{+j}(Z \mid \hat{z})\right)$ in terms of an averaged KL-divergence as follows; we have

$$\begin{aligned} D_{\mathrm{kl}}\left(P_0(Z \mid \widehat{z}) \,\|\, P_{+j}(Z \mid \widehat{z})\right) &\overset{(i)}{\le} e^{\delta} \int \frac{|p_{+j}(z \mid \widehat{z}) - p_0(z \mid \widehat{z})|^2}{p_0(z \mid \widehat{z})} d\mu(z) \\ &\overset{(ii)}{\le} e^{\delta}(e^{\delta} - 1)^2 \int \left(\int \sqrt{2 D_{\mathrm{kl}}\left(P_0(X_j \mid y, z, \widehat{z}) \,\|\, P_0(X_j \mid y, \widehat{z})\right)} dP_0(y \mid z, \widehat{z}) \right)^2 dP_0(z \mid \widehat{z}) \\ &\overset{(iii)}{\le} 2e^{\delta}(e^{\delta} - 1)^2 \int D_{\mathrm{kl}}\left(P_0(X_j \mid y, z, \widehat{z}) \,\|\, P_0(X_j \mid y, \widehat{z})\right) dP_0(y, z \mid \widehat{z}) \\ &= 2e^{\delta}(e^{\delta} - 1)^2 I_{P_0}(X_j; Z \mid Y, \widehat{Z} = \widehat{z}), \end{aligned}$$

where the final equality is the definition of conditional mutual information. Step (i) follows from Lemma 3, step (ii) by the strong information contraction of Lemma 4, and step (iii) as a consequence of Jensen. By noting that $I(A; B \mid C) + I(A; C) = I(A; B, C)$ for any random variables A, B, C, the final information quantity is bounded by $I(X_j; Z, Y \mid \widehat{Z} = \widehat{z})$, whence we obtain

$$D_{\mathrm{kl}}\left(P_0(Z \mid \widehat{z}) \,\|\, P_{+j}(Z \mid \widehat{z})\right) \le 2e^{\delta}(e^{\delta} - 1)^2 I_{P_0}(X_j; Z, Y \mid \widehat{Z} = \widehat{z}). \tag{8}$$

By construction, the X_j are independent (even given $\widehat{Z}$), yielding the joint information bound

$$\begin{aligned}
\frac{1}{2d}\sum_{j=1}^{d}\int D_{\mathrm{kl}}\left(P_0(Z \,|\, \widehat{z}) \,\|\, P_{+j}(Z \,|\, \widehat{z})\right) dP_0(\widehat{z}) &\le \frac{e^{\delta}(e^{\delta}-1)^2}{d}\sum_{j=1}^{d}\int I_{P_0}(X_j; Z, Y \mid \hat{Z} = \hat{z}) dP_0(\widehat{z}) \\
&= \frac{e^{\delta}(e^{\delta}-1)^2}{d}\sum_{j=1}^{d} I_{p_0}(X_j; Z, Y \mid \hat{Z}) \\
&\le \frac{e^{\delta}(e^{\delta}-1)^2}{d} I_{p_0}(X; Z, Y \mid \hat{Z}).
\end{aligned} \tag{9}$$

Returning to the full divergence in inequality (6), we sum over indices $i = 1, \ldots, n$ to obtain

$$\begin{aligned}
&\frac{1}{4d}\sum_{j=1}^{d} D_{\mathrm{kl}}\left(P_0(Z^{(1:n)}) \,\|\, P_{-j}(Z^{(1:n)})\right) + D_{\mathrm{kl}}\left(P_0(Z^{(1:n)}) \,\|\, P_{+j}(Z^{(1:n)})\right) \\
&\qquad \le \frac{e^{\delta}(e^{\delta}-1)^2}{d}\sum_{i=1}^{n} I_{P_0}(X^{(i)}; Z^{(i)}, Y^{(i)} \mid Z^{(1:i-1)}) = \frac{e^{\delta}(e^{\delta}-1)^2 \widetilde{B}_0 n}{d}.
\end{aligned}$$

Hence, by Lemma 2, $\mathbb{P}[\hat{J} \neq J] \ge \frac{1}{2} - \sqrt{\frac{e^{\delta}(e^{\delta}-1)^2 \widetilde{B}_0 n}{d}}$. Applying Lemma 1, we have

$$\mathbb{E}\left[\|\hat{W} - W\|_2^2\right] \ge \frac{r^2}{2}\left(\frac{1}{2} - \sqrt{\frac{e^{\delta}(e^{\delta}-1)^2 \widetilde{B}_0 n}{d}}\right),$$

which proves Theorem 1.

3. Lower bound for general k

Theorem 1 provides a lower bound on the memory-constrained minimax risk when $k = 1$. We can extend to general k using a so-called "direct-sum" approach (e.g. Braverman, 2012; Garg et al., 2014). To do so, we define a distribution that is a bit different from the standard model (1).

Let $W^* \in \{-r/\sqrt{k}, 0, r/\sqrt{k}\}^d$ be a d-dimensional vector, whose d coordinates we split into k contiguous blocks, each of size at least $\lfloor \frac{d}{k} \rfloor$. Within a block, with probability $\frac{1}{2}$ all coordinates are zero, and otherwise we choose a single coordinate uniformly at random (within the block) to have value $\pm r/\sqrt{k}$. We denote this distribution by P^*. As before, we let each $X^{(i)} \in \{-1, 1\}^d$ be a random sign vector. We now define the noise process. Let $\epsilon_l^{(i)} = 0$ for all i if $W_l^* \neq 0$, choose $\epsilon_l^{(i)}$ i.i.d. and uniformly from $\{\pm r/\sqrt{k}\}$ if $W_l^* = 0$, and set $\epsilon^{(i)} = \epsilon_0^{(i)} + \sum_{l=1}^{k} \epsilon_l^{(i)}$, where $\epsilon_0^{(i)} \sim \mathrm{Laplace}(\sigma/\sqrt{2})$. At iteration i, then, we observe

$$Y^{(i)} = {W^*}^{\top} X^{(i)} + \epsilon^{(i)}. \tag{10}$$

The key idea that allows us to extend our techniques from the previous section to obtain a lower bound for general $k \in \mathbb{N}$ (as opposed to $k = 1$) is that we can decompose Y as

$$Y^{(i)} = \left[W_l^{*\top} X_l^{(i)} + \epsilon_l^{(i)}\right] + \left[W_{\neg l}^{*\top} X_{\neg l}^{(i)} + \sum_{l' \neq l} \epsilon_{l'}^{(i)}\right] + \epsilon_0^{(i)}.$$

We can then reduce to the case when $k = 1$ by letting $W = W_l^*$, the lth block, $\xi^{(i)} = W_{\neg l}^{*\top} X_{\neg l}^{(i)} + \sum_{l' \neq l} \epsilon_{l'}^{(i)}$, and $\epsilon^{(i)} = \epsilon_0^{(i)}$. Of course, our procedure is not allowed to know $W_{\neg l}^*$ or $\epsilon_{\neg l}$, but any lower bound in which a procedure observes these is only stronger than one in which the procedure does not. By showing that estimation of the lth block is still challenging in this model, we obtain our *direct sum* result, that is, that estimating each of the k blocks is difficult in a memory-restricted setting. Thus, Theorem 1 gives us:

Proposition 1 *Let $d_l \ge \lfloor \frac{d}{k} \rfloor$ be the size of the lth block, let*

$$\widetilde{B}_l \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} I_{P^*}(X_l^{(i)}; Z^{(i)}, Y^{(i)} \mid Z^{(1:i-1)}, W_l^*, W_{\neg l}^*), \tag{11}$$

and let $\nu = 2\sqrt{2}r/\sigma$. Then for any communication-constrained estimator $\widehat{W}_l$ of W_l^*,

$$\mathbb{E}\left[\|\widehat{W}_l - W_l^*\|_2^2\right] \ge \frac{r^2}{4k}\left(\frac{1}{2} - \sqrt{\frac{2\widetilde{B}_l n}{d_l} e^{\nu/\sqrt{k}}(e^{\nu/\sqrt{k}} - 1)^2}\right). \tag{12}$$

Proof The key is to relate P^* to the distributions considered in Section 2, for which we already have results. While this may seem extraneous, using P^* is crucial for allowing our bounds to tensorize in Theorem 2.

First focusing on the setting $k = 1$ as in Theorem 1, let P_0 be the "null" distribution defined in the beginning of Section 2, and let $\widehat{P}$ be the joint distribution of W, Z, X, Y when W is drawn uniformly from the 1-sparse vectors in $\{-r, 0, r\}^d$. Letting $P = \frac{1}{2}(P_0 + \widehat{P})$, we have for any $i \in [n]$

$$\begin{aligned} I_P(\cdot \mid Z^{(1:i-1)}, W) &= \frac{1}{2} I_P(\cdot \mid Z^{(1:i-1)}, W = 0) + \frac{1}{2} I_P(\cdot \mid Z^{(1:i-1)}, W, W \neq 0) \\ &= \frac{1}{2} I_{P_0}(\cdot \mid Z^{(1:i-1)}) + \frac{1}{2} I_{\widehat{P}}(\cdot \mid Z^{(1:i-1)}, W) \ge \frac{1}{2} I_{P_0}(\cdot \mid Z^{(1:i-1)}). \end{aligned}$$

Thus, in the setting of Theorem 1, if we define

$$\widetilde{B} \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} I_P(X^{(i)}; Y^{(i)}, Z^{(i)} \mid Z^{(1:i-1)}, W),$$

we have $\widetilde{B}_0 \le 2\widetilde{B}$. Moreover, we have $P \ge \frac{1}{2}\widehat{P}$, and so we obtain that $\mathbb{E}_P[\|\widehat{W} - W\|_2^2] \ge \frac{1}{2}\mathbb{E}_{\widehat{P}}[\|\widehat{W} - W\|_2^2]$, where the second expectation is the risk bounded in Theorem 1. Coupled with the definition of $\widetilde{B}$, this implies that for $k = 1$, we have

$$\mathbb{E}_P\left[\|\widehat{W} - W\|_2^2\right] \ge \frac{1}{2}\frac{r^2}{2}\left(\frac{1}{2} - \sqrt{\frac{e^{\delta}(e^{\delta} - 1)^2 \widetilde{B}_0 n}{d}}\right) \ge \frac{r^2}{4}\left(\frac{1}{2} - \sqrt{\frac{2e^{\delta}(e^{\delta} - 1)^2 \widetilde{B} n}{d}}\right). \tag{13}$$

Now we show how to give a similar result, focusing on the $k > 1$ case, for a single block l under the distribution P^*. Indeed, we have that $P^*(\cdot \mid W^*_{\neg l}) = \frac{1}{2}P^*(\cdot \mid W^*_l, W^*_l \neq 0, W^*_{\neg l}) + \frac{1}{2}P^*(\cdot \mid W^*_{\neg l}, W^*_l = 0)$. Let $\widetilde{B}^*_l = \frac{1}{n}\sum_{i=1}^n I_{P^*}(X_l^{(i)}; Z^{(i)}, Y^{(i)} \mid Z^{(1:i-1)}, W^*_l, W^*_{\neg l} = w^*_{\neg l})$ be the average mutual information conditioned on the realization $W^*_{\neg l} = w^*_{\neg l}$. Then the lower bound (13), coupled with the discussion preceding this proposition, implies

$$\mathbb{E}_{P^*}\left[\|\widehat{W}_l - W^*_l\|_2^2 \mid W^*_{\neg l} = w^*_{\neg l}\right] \geq \frac{r^2}{4k}\left(\frac{1}{2} - \sqrt{\frac{2e^{\delta}(e^{\delta}-1)^2\widetilde{B}^*_l n}{d_l}}\right), \tag{14}$$

where we have used $\delta = \nu/\sqrt{k}$. We must remove the conditioning in (14). To that end, note that $\int \sqrt{\widetilde{B}^*_l} dP^*(w^*_{\neg l}) \leq \left(\int \widetilde{B}^*_l dP^*(w^*_{\neg l})\right)^{\frac{1}{2}} = \widetilde{B}_l^{\frac{1}{2}}$ by Jensen. Integrating (14) completes the proof. ■

Extending this proposition, we arrive at our final lower bound, which holds for any k.

Theorem 2 *Let $\nu = 2\sqrt{2}r/\sigma$ and assume that $\nu \leq \sqrt{k}/3$. Assume that $Z^{(i)}$ consists of $B_i \geq 1$ bits, and define $B = \frac{1}{n}\sum_{i=1}^n B_i$. For any communication-constrained estimator $\widehat{W}$ of W^*,*

$$\mathbb{E}\left[\|\widehat{W} - W^*\|_2^2\right] \geq \frac{r^2}{4}\left(\frac{1}{2} - \sqrt{\frac{9Bn\nu^2 + 4n\nu^2\log(1+\nu^2)}{2k^2\lfloor d/k \rfloor}}\right).$$

To prove Theorem 2, we sum (12) over l from 1 to k; the main work is to show that $\sum_{l=1}^k \tilde{B}_l$ from Proposition 1 is at most slightly larger than the bit constraint B. The intuition is that $Y^{(i)}$, being a single scalar, only adds a small amount of information on top of $Z^{(i)}$. The full proof is in Sec. A.4.

We end with a few remarks on the implications of Theorem 2 for asymptotic rates of convergence. Fixing the variance parameter σ and number of observations n, we choose the size parameter r to optimize ν. To satisfy the assumptions of the theorem, we must have $r^2 \leq \sigma^2 k/72$. Choosing $r^2 = \frac{\sigma^2}{8}\min\left\{\frac{k^2\lfloor d/k \rfloor}{36Bn}, \frac{k}{9}, e^{\frac{9}{4}B} - 1\right\}$, we are guaranteed that $4\log(1+\nu^2) \leq 9B$, and also that $\frac{9Bn\nu^2 + 4n\nu^2\log(1+\nu^2)}{2k^2\lfloor d/k \rfloor} \leq \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$, whence Theorem 2 implies the lower bound

$$\mathbb{E}\left[\|\widehat{W} - W^*\|_2^2\right] \geq \frac{\sigma^2}{128}\min\left\{\frac{k^2\lfloor d/k \rfloor}{36Bn}, \frac{k}{9}, e^{\frac{9}{4}B} - 1\right\} \gtrsim \sigma^2 \min\left(\frac{kd}{Bn}, 1\right).$$

That is, we require at least an average of $B = \Omega(d/\log(d))$ bits of communication (or memory) per round of our procedure to achieve the optimal (unconstrained) estimation rate of $\Theta\left(\sigma^2 k \log(d)/n\right)$.

4. An algorithm and upper bound

We now provide an algorithm for the setting when the memory budget satisfies $B \geq \Omega(1) \cdot \max\{k\log d, k\log n\}$. It is no loss of generality to assume that the budget is at least this high, as otherwise we cannot even represent the optimal vector $w^* \in \mathbb{R}^d$ to high accuracy. Before giving the procedure, we enumerate the (admittedly somewhat restrictive) assumptions under which it operates. As before, all proofs are in the supplement.

Assumption A *The vectors $X^{(i)}$ and noise variables $\epsilon^{(i)}$ are independent and are drawn i.i.d. Additionally, they satisfy $\mathbb{E}[X] = 0$, $\mathbb{E}[\epsilon] = 0$, and $\mathrm{Cov}(X) = I_{d\times d}$, and X satisfies $\|X\|_\infty \leq \rho$ with probability 1. Moreover, ϵ is σ-sub-exponential, meaning that $\mathbb{E}[|\epsilon|^k] \leq (k!)\sigma^k$ for all k.*

In addition, we assume without further mention—as in Theorems 1 and 2—that w^* is k-sparse, meaning $\|w^*\|_0 \le k$, and that we know a bound r such that $\|w^*\|_2 \le r$. In the construction we gave in the lower bound, we assumed $\rho = 1$ and $r = \nu\sigma = O(\min\{1, \sqrt{kd/Bn}\}\sigma)$. The strongest of our assumptions is that $\mathrm{Cov}(X) = I_{d\times d}$; letting $U = \mathrm{supp}\, w^*$, we can weaken this to $\mathbb{E}[X_U X_{\neg U}^\top] = 0$ and $\mathrm{Cov}(X_U) \succeq \gamma I$ for some $\gamma > 0$, but we omit this for simplicity in exposition. Further weakenings of this assumption in online settings appear possible but challenging (Agarwal et al., 2012; Steinhardt et al., 2014).

We now describe a memory-bounded procedure for performing an analogue of regularized dual averaging (RDA; Xiao (2010)). In RDA, one receives a sequence f_i of loss functions, maintaining a vector $\theta^{(i)}$ of gradient sums, and at iteration i, performs the following two-step update:

$$w^{(i)} = \operatorname*{arg\,min}_{w\in\mathcal{W}} \left\{ \langle \theta^{(i)}, w\rangle + \psi(w) \right\} \quad \text{and} \quad \theta^{(i+1)} = \theta^{(i)} + g^{(i)}, \;\; \text{where } g^{(i)} \in \partial f_i(w^{(i)}). \qquad (15)$$

The set $\mathcal{W}$ is a closed convex constraint set, and the function ψ is a strongly-convex regularizing function; Xiao (2010) establishes convergence of this procedure for several functions ψ.

We apply a variant of RDA (15) to the sequence of losses $f_i(w) = \frac{1}{2}\left(y^{(i)} - w^\top x^{(i)}\right)^2$. Our procedure separates the coordinates into two sets; most coordinates are in the first set, stored in a compressed representation admitting approximate recovery. The accuracy of this representation is too low for accurate optimization but is high enough to determine which coordinates are in $\mathrm{supp}(w^*)$. We then track these (few) important coordinates more accurately. For compression we use a *count sketch* (CS) data structure (Charikar et al., 2002), which has two parameters: an accuracy ϵ, and a failure probability δ. The CS stores an approximation $\widehat{x}$ to a vector x by maintaining a low-dimensional projection Ax of x and supports two operations:

- $\texttt{Update}(v)$, which replaces x with $x + v$.
- $\texttt{Query}()$, which returns an approximation $\hat{x}$ to x.

Let $\mathcal{C}(\epsilon, \delta)$ denote the CS data structure with parameters ϵ and δ. It satisfies the following:

Proposition 2 (Gilbert and Indyk (2010), Theorem 2) *Let $0 < \epsilon, \delta < 1$ and $k \le \frac{1}{\epsilon}$. The data structure $\mathcal{C}(\epsilon, \delta)$ can perform $\texttt{Update}(v)$ and $\texttt{Query}()$ in $O(d \log \frac{n}{\delta})$ time and stores $O(\frac{\log(n/\delta)}{\epsilon})$ real numbers. If $\hat{x}$ is the output of $\texttt{Query}()$, then $\|\hat{x} - x\|_\infty^2 \le \epsilon \|x - x_{\mathrm{top\,k}}\|_2^2$ holds uniformly over the course of all n updates with probability $1 - \delta$, where $x_{\mathrm{top\,k}}$ denotes x with only its k largest entries (in absolute value) kept non-zero.*

Based on this proposition, we implement a version of ℓ_1-regularized regularized dual averaging, which we present in Algorithm 1. We show subsequently that this procedure (with high probability) correctly implements dual averaging, so we can leverage known convergence guarantees to give an upper bound on the minimax rate of convergence for memory-bounded procedures.

Letting $\tilde{w}^{(i)}$ be the parameter vector in iteration i, the algorithm tracks three quantities: first, a count-sketch approximation $\tilde{\theta}^{(i)}_{\mathrm{coarse}}$ to $\theta^{(i)} \stackrel{\mathrm{def}}{=} \sum_{i'<i} \partial f_{i'}(\tilde{w}^{(i')})$, which requires $O(\frac{1}{\epsilon}\log\frac{n}{\delta})$ memory (we specify ϵ presently); second, a set $\tilde{U}$ of "active" coordinates; and third, an approximation $\tilde{\theta}^{(i)}_{\mathrm{fine}}$ to $\theta^{(i)}$ supported on $\tilde{U}$. We also track a running average $\hat{w}$ of $\tilde{w}^{(1:i)}$, which we use at the end as a parameter estimate; this requires at most $|\widetilde{U}|$ numbers. In the algorithm, we use the soft-thresholding and projection operators, given by

$$\mathsf{T}_c(x) \stackrel{\mathrm{def}}{=} \left[\mathrm{sign}(x_j)\left[|x_j| - c\right]_+\right]_{j=1}^d \quad \text{and} \quad \mathsf{P}_r(x) \stackrel{\mathrm{def}}{=} \begin{cases} x & \text{if } \|x\|_2 \le r \\ rx/\|x\|_2 & \text{otherwise.} \end{cases}$$

Algorithm 1 Low-memory ℓ_1-RDA for regression

```
Algorithm parameters: $c$, $\Delta$, $R$, $\epsilon$, $\delta$. Initialize $\mathcal{C}(\epsilon, \delta)$, $\hat{w} \leftarrow 0$, $\tilde{U} \leftarrow \emptyset$
for $i = 1$ to $n$ do
    $\tilde{w}^{(i)} \leftarrow -\mathsf{P}_r(\eta \mathsf{T}_{c\sqrt{n}}(\tilde{\theta}^{(i)}_{\text{fine}}))$ and $\hat{w} \leftarrow \frac{1}{i}\tilde{w}^{(i)} + \frac{i-1}{i}\hat{w}$
    Predict $\tilde{y}^{(i)} = (\tilde{w}^{(i)})^\top x^{(i)}$ and compute gradient $g^{(i)} = (y^{(i)} - \tilde{y}^{(i)})x^{(i)}$
    Call Update($g^{(i)}$) and set $\tilde{\theta}^{(i+1)}_{\text{coarse}} \leftarrow$ Query()
    for $j \in \tilde{U}$ do
        $\tilde{\theta}^{(i+1)}_{\text{fine},j} \leftarrow \tilde{\theta}^{(i)}_{\text{fine},j} + g^{(i)}_j$
    end for
    for $j \notin \tilde{U}$ do
        if $|\tilde{\theta}^{(i+1)}_{\text{coarse},j}| \geq (c - 2\Delta)\sqrt{n}$ then
            Add $j$ to $\tilde{U}$ and set $\tilde{\theta}^{(i+1)}_{\text{fine},j} \leftarrow \tilde{\theta}^{(i+1)}_{\text{coarse},j}$
        else
            $\tilde{\theta}^{(i+1)}_{\text{fine},j} \leftarrow 0$
        end if
    end for
end for
return $\hat{w}$
```

We remark that in Algorithm 1, we need track only $\widehat{w}$, $\widetilde{\theta}^{(i)}_{\text{fine}}$, and the count sketch data structure, so the memory usage (in real numbers) is bounded by $\|\hat{w}\|_0 + \|\tilde{\theta}^{(i)}_{\text{fine}}\|_0$, plus the size of the count sketch structure. We will see later that the size of this structure is roughly inversely proportional to the degree of ℓ_1-regularization. Our main result concerns the convergence of Algorithm 1.

Theorem 3 *Let Assumption A and the model (1) hold. With appropriate setting of the constants c, Δ, r, ϵ (specified in the proof), for budget $B \in [k, d]$, Algorithm 1 uses at most $\widetilde{O}(B)$ bits and achieves risk*

$$\mathbb{E}\left[\|\widehat{w} - w^*\|_2^2\right] = \widetilde{O}\left(\max\left\{r\rho\sigma\sqrt{\frac{kd}{Bn}}, r^2\rho^2\frac{kd}{Bn}\right\}\right).$$

To compare Theorem 3 with our lower bounds, assume that $\frac{kd}{Bn} \leq 1$ and set $\rho = 1$ and $r = \nu\sigma$; this matches the setting of our lower bounds in Theorems 1 and 2. Then $r\rho\sigma = \nu\sigma^2 = \sqrt{\frac{kd}{bn}}\sigma^2$, and we have $\Omega\left(\sigma^2\frac{kd}{Bn}\right) \leq \inf_{\widehat{w}} \mathbb{E}\left[\|\widehat{w} - w^*\|_2^2\right] \leq \tilde{O}\left(\sigma^2\frac{kd}{Bn}\right)$. In particular, at least for some non-trivial regimes, our upper and lower bounds match to polylogarithmic factors. This does *not* imply that our algorithm is optimal: if the radius r is fixed as n grows, we expect the optimal error to decay as $\frac{1}{n}$ rather than the $\frac{1}{\sqrt{n}}$ rate in Theorem 3. One reason to believe the optimal rate is $\frac{1}{n}$ is that it is attainable when $k = 1$: simply split the coordinates into batches of size $\tilde{O}(B)$ and process the batches one at a time; since $\text{supp}(w^*)$ has size 1, it is always fully contained in one of the batches.

Analysis of Algorithm 1 Define s_j to be the iteration where j is added to $\tilde{U}$ (or ∞ if this never happens). Also let i_{bad} be the first iteration where $\|\tilde{\theta}^{(i)}_{\text{coarse}} - \theta^{(i)}\|_\infty > \Delta\sqrt{n}$, and define

$$\bar{a}_j = \begin{cases} \tilde{\theta}^{(s_j)}_{\text{coarse},j} - \theta^{(s_j)}_j & : \quad s_j < i_{\text{bad}}, \\ 0 & : \quad s_j \geq i_{\text{bad}}. \end{cases}$$

The vector $\overline{a}$ tracks the "offset" between $\tilde{\theta}_{\text{fine}}$ and θ, while being clipped to ensure that $\|\overline{a}\|_\infty \leq \Delta\sqrt{n}$. Our key result is that Algorithm 1 implements an instantiation of RDA:

Lemma 5 *Let $(\overline{\theta}^{(i)}, \overline{w}^{(i)})$ be the sequence of iterates produced by RDA (15) with regularizer*

$$\psi(w) = \frac{1}{2\eta}\|w\|_2^2 + c\sqrt{n}\|w\|_1 + \overline{a}^\top w,$$

where $\operatorname{supp}(w)$ is constrained to lie in U and have ℓ_2-norm at most r. Suppose that for some G,

$$\mathsf{G} \geq \frac{1}{\sqrt{n}} \max_{i=1}^{n} \max_{j \notin U} |\overline{\theta}_j^{(i)}|, \quad \Delta \geq \sqrt{\epsilon(d-k)}\mathsf{G}, \quad \text{and} \quad c \geq 2\Delta + \mathsf{G}.$$

Also assume $\epsilon \leq \frac{1}{k}$. Then, with probability $1-\delta$, $\overline{w} = \tilde{w}$.

Let $\mathsf{Reg} \overset{\text{def}}{=} \sum_{i=1}^n f_i(\overline{w}^{(i)}) - f_i(w^*)$ denote the regret of the RDA procedure in Lemma 5, and let $E \overset{\text{def}}{=} \sum_{i=1}^n f_i(w^*) = \sum_{i=1}^n (\epsilon^{(i)})^2$ be the empirical loss of w^*. A mostly standard analysis yields:

Lemma 6 *Assume $\epsilon \leq \frac{1}{k}$. Also suppose that*

$$\max_{i=1}^{n} \max_{j \notin U} |\overline{\theta}_j^{(i)}| \leq 2\rho\sqrt{(\mathsf{Reg} + E)\log(2d/\delta)}. \tag{16}$$

Then, letting $V \overset{\text{def}}{=} 2\left(1 + 3\sqrt{\epsilon(d-k)}\right)\sqrt{\log(2d/\delta)}$ for short-hand, the regret Reg satisfies

$$\mathsf{Reg} \leq 2R\rho\sqrt{k}\left(2\sqrt{2} + V\right)\sqrt{E} + kR^2\rho^2\left(4 + 4V^2\right) \tag{17}$$

for appropriately chosen c and Δ, which moreover satisfy conditions of Lemma 5.

Lemma 6 is useful because $\frac{1}{n}\mathbb{E}\left[\mathsf{Reg}\right]$ can be shown to upper-bound $\mathbb{E}\left[\|\hat{w} - w^*\|_2^2\right]$. To wrap up, we need to deal with a few details. First, we need to show that (16) holds with high probability:

Lemma 7 *With probability $1-\delta$, $|\overline{\theta}_j^{(i)}| \leq 2\rho\sqrt{(\mathsf{Reg} + E)\log(2d/\delta)}$ for all $i \leq n$ and $j \notin U$.*

Second, we need a high probability bound on E; we prove the following Lemma in Section A.10.

Lemma 8 *Let $p \geq 1$. There are constants K_p satisfying $K_p \leq 6p$ for $p \geq 5$ and $K_p/p \to 2$ as $p \to \infty$ such that for any $t \geq 0$,*

$$\mathbb{P}\left[\sum_{i=1}^n (\epsilon^{(i)})^2 \geq \sigma^2 n + 4K_p\sigma^2 t\right] \leq \left(\frac{3\sqrt{n} + 2n^{1/p}p^2}{t}\right)^p.$$

Substituting $p = \max\{5, \log\frac{1}{\delta}\}$, we have, for a constant $C \leq 72e$ and with probability $1-\delta$,

$$2E = \sum_{i=1}^n \left(\epsilon^{(i)}\right)^2 \leq n\sigma^2 + C\sigma^2\log\frac{1}{\delta}\left[\sqrt{n} + n^{1/5}\log^2\frac{1}{\delta}\right]. \tag{18}$$

Combining Lemmas 7 and 8 with Lemma 6, we then have the following with probability $1-3\delta$:

$$\mathsf{Reg} \leq \sqrt{2}R\rho\sigma\sqrt{k}\left(2\sqrt{2} + V\right)\sqrt{n + C\log(1/\delta)[\sqrt{n} + n^{1/5}\log^2(1/\delta)]} + kR^2\rho^2\left(4 + 4V^2\right).$$

To interpret this, remember that $V = 2\left(1 + 3\sqrt{\epsilon(d-k)}\right)\sqrt{\log(2d/\delta)}$ and that the count-sketch structure stores $O\left(\log(n/\delta)/\epsilon\right)$ bits; also, we need $\epsilon \leq \frac{1}{k}$ for Proposition 2 to hold. As long as $B \geq k$, we can take $\epsilon = O\left(1/B\right)$ (using $\tilde{O}(B)$ bits) and have $V = \tilde{O}(\sqrt{d/B})$. The entire first term above is thus $\tilde{O}\left(R\rho\sigma\sqrt{\frac{dkn}{B}}\right)$, while the second is $\tilde{O}\left(R^2\rho^2\frac{dk}{B}\right)$. This essentially yields Theorem 3; the full proof is in Section A.5.

From real numbers to bits. In the above, we analyzed a procedure that stores $\tilde{O}(B)$ real numbers. In fact, each number only requires $\tilde{O}(1)$ bits of precision for the algorithm to run correctly. A detailed argument for this is given in Section B.

Acknowledgments. The anonymous reviewers were instrumental in improving this paper. JS was supported by an NSF Graduate Research Fellowship and a Fannie & John Hertz Fellowship.

References

A. Agarwal, S. Negahban, and M. Wainwright. Stochastic optimization and sparse statistical recovery: An optimal algorithm for high dimensions. In *Advances in Neural Information Processing Systems 25*, 2012.

N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.

E. Arias-Castro, E. J. Candes, and M. A. Davenport. On the fundamental limits of adaptive sensing. *Information Theory, IEEE Transactions on*, 59(1):472–481, 2013.

P. Assouad. Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 296(23):1021–1024, 1983.

Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Proceedings of the Twenty Sixth Annual Conference on Computational Learning Theory*, 2013.

M. Braverman. Interactive information complexity. In *Proceedings of the Fourty-Fourth Annual ACM Symposium on the Theory of Computing*, 2012.

L. Breiman. *Probability*. Society for Industrial and Applied Mathematics, 1992.

M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. In *Automata, Languages and Programming*, pages 693–703. Springer, 2002.

A. Daniely, N. Linial, and S. Shalev-Shwartz. More data speeds up training time in learning halfspaces over sparse vectors. In *Proceedings of the Twenty Sixth Annual Conference on Computational Learning Theory*, 2013.

A. Daniely, N. Linial, and S. Shalev-Shwartz. From average case complexity to improper learning complexity. In *Proceedings of the Fourty-Sixth Annual ACM Symposium on the Theory of Computing*, 2014.

V. H. de la Peña and E. Giné. *Decoupling: From Dependence to Independence*. Springer, 1999.

J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 429–438. IEEE, 2013.

J. C. Duchi, M. I. Jordan, M. J. Wainwright, and Y. Zhang. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. *arXiv:1405.0782 [cs.IT]*, 2014.

A. Garg, T. Ma, and H. Nguyen. On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems 27*, pages 2726–2734, 2014.

A. Gilbert and P. Indyk. Sparse recovery using sparse matrices. *Proceedings of the IEEE*, 98(6): 937–947, 2010.

S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24 (2):227–234, 1995.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

O. Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. In *Neural Information Processing Systems 28*, 2014.

J. Steinhardt, S. Wager, and P. Liang. The statistics of streaming sparse regression. *arXiv preprint arXiv:1412.4182*, 2014.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5): 2183–2202, 2009.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

Y. Zhang, J. Duchi, M. Jordan, and M. J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems*, pages 2328–2336, 2013.

Y. Zhang, M. J. Wainwright, and M. I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory (COLT)*, 2014.

Appendix A. Deferred proofs

A.1. Proof of Lemma 1

Given an estimator $\hat{W}$, define an estimator $\hat{J}$ by letting $|\hat{J}|$ be the index of the largest coordinate of $|\hat{W}|$ and letting $\operatorname{sign}(\hat{J})$ be the sign of that coordinate (ties can be broken arbitrarily). We claim that $\|\hat{W} - W\|_2^2 \geq \frac{r^2}{2}\mathbb{I}[\hat{J} \neq J]$. If $|\hat{J}| = |J|$, then either $\operatorname{sign}(J) = \operatorname{sign}(\hat{J})$, in which case $\mathbb{I}[\hat{J} \neq J] = 0$, or else $\operatorname{sign}(J) \neq \operatorname{sign}(\hat{J})$, in which case $\|\hat{W} - W\|_2^2 \geq W_{|J|}^2 = r^2$. In either case, the result holds.

Therefore, we turn out attention to the case that that $|\hat{J}| \neq |J|$. Then

$$\begin{aligned}
\|\hat{W} - W\|_2^2 &\geq (\hat{W}_{|\hat{J}|} - W_{|\hat{J}|})^2 + (\hat{W}_{|J|} - W_{|J|})^2 \\
&\geq \hat{W}_{|\hat{J}|}^2 + (|\hat{W}_{|J|}| - r)^2 \\
&\overset{(i)}{\geq} \hat{W}_{|J|}^2 + (|\hat{W}_{|J|}| - r)^2 \\
&\geq \frac{r^2}{2},
\end{aligned}$$

where (i) is because $\hat{J}$ indexes the largest coordinate of $\hat{W}$ by construction. So, the claimed inequality holds, and the desired result follows by taking expectations.

A.2. Proof of Lemma 2

We note that

$$\begin{aligned}
\mathbb{P}[\hat{J} \neq J] &= 1 - \mathbb{P}[\hat{J} = J] \\
&= 1 - \frac{1}{2d}\sum_{j=1}^{d} P_{+j}(\hat{J} = +j) + P_{-j}(\hat{J} = -j) \\
&\overset{(i)}{=} \left(1 - \frac{1}{2d}\right) - \frac{1}{2d}\sum_{j=1}^{d} \Big(P_{+j}(\hat{J} = +j) - P_0(\hat{J} = +j)\Big) + \Big(P_{-j}(\hat{J} = -j) - P_0(\hat{J} = -j)\Big) \\
&\geq \left(1 - \frac{1}{2d}\right) - \frac{1}{2d}\sum_{j=1}^{d} |P_{+j}(\hat{J} = +j) - P_0(\hat{J} = +j)| + |P_{-j}(\hat{J} = -j) - P_0(\hat{J} = -j)| \\
&\overset{(ii)}{\geq} \left(1 - \frac{1}{2d}\right) - \frac{1}{2d}\sum_{j=1}^{d} \|P_{+j} - P_0\|_{TV} + \|P_{-j} - P_0\|_{TV} \\
&\overset{(iii)}{\geq} \left(1 - \frac{1}{2d}\right) - \sqrt{\frac{1}{2d}\sum_{j=1}^{d} \|P_{+j} - P_0\|_{TV}^2 + \|P_{-j} - P_0\|_{TV}^2} \\
&\overset{(iv)}{\geq} \left(1 - \frac{1}{2d}\right) - \sqrt{\frac{1}{4d}\sum_{j=1}^{d} D_{\mathrm{kl}}\left(P_0 \parallel P_{+J}\right) + D_{\mathrm{kl}}\left(P_0 \parallel P_{-j}\right)},
\end{aligned}$$

as was to be shown. Here (i) uses the fact that $\sum_{j=1}^{d} P_0(\hat{J} = +j) + P_0(\hat{J} = -j) = 1$, (ii) uses the variational form of TV distance, (iii) is Cauchy-Schwarz, and (iv) is Pinsker's inequality.

A.3. Proof of Lemma 4

We first make three observations, each of which relies on the specific structure of our problem. First, we have

$$p_{+j}(z \mid x_j, y, \widehat{z}) = p_0(z \mid x_j, y, \widehat{z}) \tag{19a}$$

since in both cases $X_{\neg j}$ are i.i.d. random sign vectors, $Z^{(i)}$ is $(X^{(i)}, \xi^{(i)}, Y^{(i)}, Z^{(1:i-1)})$-measurable, and $\xi^{(i)}$ is independent of X. Secondly, we have

$$p_{+j}(y \mid \widehat{z}) = p_0(y \mid \widehat{z}) \tag{19b}$$

by construction of p_0. Finally, we have the inequality

$$|p_{+j}(x_j, y \mid \widehat{z}) - p_0(x_j, y \mid \widehat{z})| = \left| \frac{p_{+j}(x_j, y \mid \widehat{z})}{p_0(x_j, y \mid \widehat{z})} - 1 \right| p_0(x_j, y \mid \widehat{z}) \le (e^{\delta} - 1) p_0(x_j, y \mid \widehat{z}), \tag{19c}$$

as the ratio between the quantities—as noted by inequality (7)—is bounded by e^{δ}. Therefore, expanding the distance between $p_{+j}(z)$ and $p_0(z)$, we have

$$\begin{aligned} |p_{+j}(z \mid \widehat{z}) - p_0(z \mid \widehat{z})| &= \left| \int p_{+j}(z \mid x_j, y, \widehat{z}) dP_{+j}(x_j, y \mid \widehat{z}) - p_0(z \mid x_j, y, \widehat{z}) dP_0(x_j, y \mid \widehat{z})) \right| \\ &\overset{(i)}{=} \left| \int p_0(z \mid x_j, y, \widehat{z})(dP_{+j}(x_j, y \mid \widehat{z}) - dP_0(x_j, y \mid \widehat{z})) \right| \\ &\overset{(ii)}{=} \left| \int (p_0(z \mid x_j, y, \widehat{z}) - p_0(z \mid y, \widehat{z}))(dP_{+j}(x_j, y \mid \widehat{z}) - dP_0(x_j, y \mid \widehat{z})) \right|, \end{aligned} \tag{20}$$

where step (i) follows from the independence equality (19a) and step (ii) because $p_0(z \mid y, \widehat{z})$ is constant with respect to x_j and $p_{+j}(y \mid \widehat{z}) = p_0(y \mid \widehat{z})$ by (19b). Next, by inequality (19c) we have that

$$|dP_{+j}(x_j, y \mid \widehat{z}) - dP_0(x_j, y \mid \widehat{z}))| \le (e^{\delta} - 1) dP_0(x_j, y \mid \widehat{z}), \tag{21}$$

whence we have the further upper bound

$$\begin{aligned} &|p_{+j}(z \mid \widehat{z}) - p_0(z \mid \widehat{z})| \\ &\quad \overset{(i)}{\le} (e^{\delta} - 1) \int |p_0(z \mid x_j, y, \widehat{z}) - p_0(z \mid y, \widehat{z})| dP_0(x_j, y \mid \widehat{z}) \\ &\quad \overset{(ii)}{=} (e^{\delta} - 1) \int \left| \frac{dP_0(x_j, y \mid z, \widehat{z}) p_0(z \mid \widehat{z})}{dP_0(x_j, y \mid \widehat{z})} - \frac{dP_0(y \mid z, \widehat{z}) p_0(z \mid \widehat{z})}{dP_0(y \mid \widehat{z})} \right| dP_0(x_j, y \mid \widehat{z}) \\ &\quad = (e^{\delta} - 1) p_0(z \mid \widehat{z}) \int \left| dP_0(x_j, y \mid z, \widehat{z}) - dP_0(x_j, y \mid \widehat{z}) \frac{dP_0(y \mid z, \widehat{z})}{dP_0(y \mid \widehat{z})} \right| \\ &\quad = (e^{\delta} - 1) p_0(z \mid \widehat{z}) \int \left| dP_0(x_j \mid y, z, \widehat{z}) - dP_0(x_j \mid y, \widehat{z}) \right| dP_0(y \mid z, \widehat{z}) \\ &\quad \overset{(iii)}{\le} (e^{\delta} - 1) p_0(z \mid \widehat{z}) \int \sqrt{2 D_{\rm kl} \left(P_0(X_j \mid y, z, \widehat{z}) \,\|\, P_0(X_j \mid y, \widehat{z}) \right)} dP_0(y \mid z, \widehat{z}), \end{aligned}$$

where (i) is by (20) and (21), (ii) is by Bayes' rule, and (iii) is by Pinsker's inequality.

A.4. Proof of Theorem 2

First, we observe that $e^{\nu/\sqrt{k}}(e^{\nu/\sqrt{k}}-1)^2 \le 2\nu^2/k$ for $\nu \le \sqrt{k}/3$, so we may replace the lower bound in Proposition 1 with

$$\mathbb{E}\left[\|\widehat{W}_l - W_l^*\|_2^2\right] \ge \frac{r^2}{4k}\left(\frac{1}{2} - \sqrt{\frac{4\widetilde{B}_l n\nu^2}{kd_l}}\right). \tag{22}$$

Now, by inequality (22) and Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}_{P^*}\left[\|\widehat{W} - W^*\|_2^2\right] = \sum_{l=1}^{k}\mathbb{E}_{P^*}\left[\|\widehat{W}_l - W_l^*\|_2^2\right] &\ge \frac{r^2}{4k}\sum_{k=1}^{l}\left(\frac{1}{2} - \sqrt{\frac{4n\nu^2}{k\lfloor d/k\rfloor}}\sqrt{\widetilde{B}_l}\right)\\ &\ge \frac{r^2}{4}\left(\frac{1}{2} - \sqrt{\frac{4n\nu^2}{k^3\lfloor d/k\rfloor}}\sum_{l=1}^{k}\sqrt{\widetilde{B}_l}\right)\\ &\ge \frac{r^2}{4}\left(\frac{1}{2} - \sqrt{\frac{4n\nu^2}{k^2\lfloor d/k\rfloor}}\sqrt{\sum_{l=1}^{k}\widetilde{B}_l}\right).\end{aligned}$$

We would thus like to bound $\sum_{l=1}^{k}\widetilde{B}_l$.

Next note that by the independence of the $X_l^{(i)}$ and the chain rule for mutual information, we have that

$$\begin{aligned}\sum_{l=1}^{k} I_{P^*}(X_l^{(i)}; Z^{(i)}, Y^{(i)}\, |Z^{(1:i-1)}, W^*) &\le I_{P^*}(X^{(i)}; Z^{(i)}, Y^{(i)} \mid Z^{(1:i-1)}, W^*)\\ &= I_{P^*}(X^{(i)}; Y^{(i)} \mid Z^{(1:i-1)}, W^*) + I_{P^*}(X^{(i)}; Z^{(i)} \mid Y^{(i)}, Z^{(1:i-1)}, W^*).\end{aligned} \tag{23}$$

We can upper bound the final term in expression (23) by $H(Z^{(i)}) \le B_i$. In addition, the first term on the right hand side of (23) satisfies

$$\begin{aligned}I_{P^*}(X^{(i)}; Y^{(i)} \mid Z^{(1:i-1)}, W^*) &= h(Y^{(i)} \mid Z^{(1:i-1)}, W^*) - h(Y^{(i)} \mid X^{(i)}, Z^{(1:i-1)}, W^*)\\ &\le \frac{1}{2}\log(2\pi e\,\mathrm{Var}(Y^{(i)})) - h(\mathrm{Laplace}(\sigma/\sqrt{2})),\end{aligned}$$

where h denotes differential entropy and we have used that the normal distribution maximizes entropy for a given variance. Using that $h(\mathrm{Laplace}(\sigma/\sqrt{2})) = 1 + \log(\sqrt{2}\sigma)$, inequality (23) thus implies

$$\begin{aligned}I_{P^*}(X^{(i)}; Z^{(i)}, Y^{(i)} \mid Z^{(1:i-1)}, W^*) &\le B_i + \frac{1}{2}\log(2\pi e\,\mathrm{Var}[Y^{(i)}]) - 1 - \log(\sqrt{2}\sigma)\\ &= B_i + \frac{1}{2}\log(2\pi e(\sigma^2 + R^2)) - \frac{1}{2}\log(2e^2\sigma^2)\\ &= B_i + \frac{1}{2}\log\frac{\pi\left(\sigma^2 + R^2\right)}{e\sigma^2} \le \frac{9}{8}B_i + \frac{1}{2}\log\left(1+\nu^2\right).\end{aligned}$$

In the last step we use $\frac{1}{2}\log(\pi/e) \le \frac{1}{8} \le \frac{1}{8}B_i$ and $R^2/\sigma^2 \le 8R^2/\sigma^2 = \nu^2$. Using the definition (11) of $\widetilde{B}_l$ and the preceding bound, we obtain that $\sum_{l=1}^{k}\widetilde{B}_l \le (9/8)B + \frac{1}{2}\log\left(1+\nu^2\right)$.

Using $\sum_{l=1}^{k}\widetilde{B}_l \le (9/8)B + \frac{1}{2}\log(1+\nu^2)$ gives the result.

A.5. Proof of Theorem 3

To prove Theorem 3, we first state the following more precise theorem, whose proof is given in Sec. A.6:

Theorem A *Let $\epsilon \leq \frac{1}{k}$. With probability $1 - 3\delta$, we have*

$$\mathsf{Reg} \leq \sqrt{2}R\rho\sigma\sqrt{k}\left(2\sqrt{2} + V\right)\sqrt{n + C\log(1/\delta)[\sqrt{n} + n^{1/5}\log^2(1/\delta)]} + kR^2\rho^2\left(4 + 4V^2\right). \tag{24}$$

In particular, let:

$$\delta = \frac{k^2R^2\rho^2}{12\sigma^2n^2d}$$
$$\hat{w} \overset{\text{def}}{=} \frac{1}{n}\left(\tilde{w}^{(1)} + \cdots + \tilde{w}^{(n)}\right).$$

Then we have the minimax bound:

$$\begin{aligned}&\mathbb{E}\left[\|\hat{w} - w^*\|_2^2\right] \\ &\quad\leq \frac{2}{n}\left(\sqrt{2}R\rho\sigma\sqrt{k}\left(2\sqrt{2} + V\right)\sqrt{n + C\log(1/\delta)[\sqrt{n} + n^{1/5}\log^2(1/\delta)]} + kR^2\rho^2\left(6 + 4V^2\right)\right).\end{aligned}$$

Based on this theorem, we have

$$\mathbb{E}\left[\|\hat{w} - w^*\|_2^2\right] \leq \tilde{O}(R\rho\sigma(1 + V)\sqrt{k/n} + R^2\rho^2(1 + V^2)k/n).$$

Also, by Proposition 2, the count sketch structure stores $\tilde{O}(1/\epsilon)$ numbers; in addition, only $O(k)$ numbers are needed to store $\hat{w}$ and $\hat{\theta}_{\text{fine}}$. Thus as long as $b \geq \tilde{\Omega}(k)$ (so that $\epsilon \leq \frac{1}{k}$), we have $\epsilon \leq \tilde{O}(1/b)$. Next recall that $V \overset{\text{def}}{=} 2\left(1 + 3\sqrt{\epsilon(d-k)}\right)\sqrt{\log(2d/\delta)} = \tilde{O}(1 + \sqrt{d/b})$. We assume that $b \leq d$, so we therefore have

$$\begin{aligned}\mathbb{E}\left[\|\hat{w} - w^*\|_2^2\right] &\leq \tilde{O}\left(R\rho\sigma\sqrt{kd/bn} + R^2\rho^2kd/bn\right) \\ &= \tilde{O}\left(\max\left(R\rho\sigma\sqrt{\frac{kd}{bn}}, R^2\rho^2\frac{kd}{bn}\right)\right),\end{aligned}$$

as was to be shown.

A.6. Proof of Theorem A

The count-sketch guarantee (Proposition 2), as well as Lemmas 7 and 8, each hold with probability $1 - \delta$, so with probability $1 - 3\delta$ all 3 results will hold, in which case Lemma 6 establishes the stated regret bound by plugging in the bound on E from Lemma 8 to (17).

To prove the bound on $\mathbb{E}\left[\|\hat{w} - w^*\|_2^2\right]$, first note that

$$\begin{aligned}\mathbb{E}\left[f_i(\tilde{w}^{(i)})\right] &= \frac{1}{2}\left(\epsilon^{(i)} + (\tilde{w}^{(i)} - w^*)^\top x^{(i)}\right)^2 \\ &= \frac{1}{2}\left(\mathbb{E}\left[\left(\epsilon^{(i)}\right)^2\right] + \|\tilde{w}^{(i)} - w^*\|_2^2\right),\end{aligned}$$

where the second equality follows because $\mathrm{Cov}[X^{(i)}] = I$.

Also note that, by convexity, $\|\hat{w} - w^*\|_2^2 \le \frac{1}{n}\sum_{i=1}^n \|\tilde{w}^{(i)} - w^*\|_2^2$. Therefore, any bound on $\mathbb{E}\left[\sum_{i=1}^n f_i(\tilde{w}^{(i)})\right]$ implies a bound on $\mathbb{E}[\|\hat{w} - w^*\|_2^2]$; in particular,

$$\mathbb{E}[\|\hat{w} - w^*\|_2^2] \le \frac{2}{n}\mathbb{E}\left[\sum_{i=1}^n f_i(\tilde{w}^{(i)}) - \left(\epsilon^{(i)}\right)^2\right].$$

The main difficulty is that we have a bound that holds with high probability, and we need to make it hold in expectation. To accomplish this, we simply need to show that $\sum_i f_i$ is not too large even when the high probability bound fails.

To do this, first note that we have the bound

$$\begin{aligned} f_i(\tilde{w}^{(i)}) &= \frac{1}{2}\left(\epsilon^{(i)} + (\tilde{w}^{(i)} - w^*)^\top x^{(i)}\right)^2 \\ &\overset{(i)}{\le} \frac{1+\delta_0}{2}\left(\epsilon^{(i)}\right)^2 + \frac{1}{2}\left(1 + \frac{1}{\delta_0}\right)\left((\tilde{w}^{(i)} - w^*)^\top x^{(i)}\right)^2 \\ &\overset{(ii)}{\le} \frac{1+\delta_0}{2}\left(\epsilon^{(i)}\right)^2 + 2\left(1 + \frac{1}{\delta_0}\right)R^2\rho^2 d, \end{aligned}$$

where (i) is by Young's inequality and (ii) is by Cauchy-Schwarz and the fact that $\|\tilde{w}^{(i)} - w^*\|_2 \le 2R$. Thus, we have the bound

$$\sum_{i=1}^n f_i(\tilde{w}^{(i)}) \le 2n\left(1 + \frac{1}{\delta_0}\right)R^2\rho^2 d + \frac{1+\delta_0}{2}\sum_{i=1}^n \left(\epsilon^{(i)}\right)^2.$$

We also straightforwardly have the bound

$$\sum_{i=1}^n f_i(\tilde{w}^{(i)}) \le \mathsf{Reg} + \frac{1}{2}\sum_{i=1}^n \left(\epsilon^{(i)}\right)^2.$$

Combining these yields

$$\sum_{i=1}^n f_i(\tilde{w}^{(i)}) \le \min\left(\mathsf{Reg}, 2n\left(1 + \frac{1}{\delta_0}\right)R^2\rho^2 d\right) + \frac{1+\delta_0}{2}\sum_{i=1}^n \left(\epsilon^{(i)}\right)^2. \tag{25}$$

Let $\sigma_0 \overset{\mathrm{def}}{=} \mathbb{E}\left[\left(\epsilon^{(i)}\right)^2\right]$ and let Reg_0 denote the right-hand-side of (24). Then, taking expectations of (25) and using the fact that $\mathsf{Reg} \le \mathsf{Reg}_0$ with probability $1 - 3\delta$, we have

$$\frac{1}{2}\left(\sigma_0^2 + \mathbb{E}\left[\|\hat{w} - w^*\|_2^2\right]\right) \le \frac{(1-3\delta)}{n}\mathsf{Reg}_0 + 6\delta\left(1 + \frac{1}{\delta_0}\right)R^2\rho^2 d + \frac{1+\delta_0}{2}\sigma_0^2,$$

which implies that

$$\begin{aligned}
\mathbb{E}\left[\|\hat{w} - w^*\|_2^2\right] &\leq \frac{2(1-3\delta)}{n}\mathsf{Reg}_0 + 12\delta\left(1 + \frac{1}{\delta_0}\right)R^2\rho^2 d + \delta_0\sigma_0^2 \\
&\leq \frac{2}{n}\mathsf{Reg}_0 + 24\frac{\delta}{\delta_0}R^2\rho^2 d + \delta_0\sigma_0^2 \\
&= \frac{2}{n}\Bigg(\sqrt{2}R\rho\sigma\sqrt{k}\left(2\sqrt{2} + V\right)\sqrt{n + C\log(1/\delta)[\sqrt{n} + n^{1/5}\log^2(1/\delta)]} \\
&\qquad\qquad + kR^2\rho^2\left(4 + 4V^2\right)\Bigg) + 24\frac{\delta}{\delta_0}R^2\rho^2 d + \delta_0\sigma_0^2.
\end{aligned}$$

Note also that $\sigma_0 \leq \sqrt{2}\sigma$. Setting $\delta_0 = \frac{kR^2\rho^2}{\sigma^2 n}$ and $\delta = \frac{k\delta_0}{12nd} = \frac{k^2R^2\rho^2}{12\sigma^2 n^2 d}$, we obtain the desired bound.

A.7. Proof of Lemma 5

We proceed by contradiction. Let i be the minimal index where $\overline{w}^{(i)} \neq \tilde{w}^{(i)}$, and let j be a particular coordinate where the relation fails. Note that, by minimality of i, we have $\theta^{(i)} = \overline{\theta}^{(i)}$. We split into cases based on how i relates to s_j.

Case 1: $i < s_j$. By construction, we have $\tilde{w}_j^{(i)} = 0$ for $i < s_j$. Therefore, we must have $\overline{w}_j^{(i)} \neq 0$. In particular, this implies that $|\overline{\theta}_j^{(i)}| + |\overline{a}_j| > c\sqrt{n}$. But $\|\overline{a}\|_\infty \leq \Delta\sqrt{n}$ by construction, so we must have $|\overline{\theta}_j^{(i)}| > (c - \Delta)\sqrt{n}$. Since $\overline{\theta}_j^{(i)} = \theta_j^{(i)}$ and $\|\tilde{\theta}_{\text{coarse}}^{(i)} - \theta^{(i)}\|_\infty \leq \Delta\sqrt{n}$, we have $|\tilde{\theta}_{\text{coarse},j}^{(i)}| > (c - 2\Delta)\sqrt{n}$; thus $i \geq s_j$, which is a contradiction.

Case 2: $i \geq s_j$. Note that

$$\begin{aligned}
\tilde{w}_j^{(i)} &= -S_R(\eta T_{c\sqrt{n}}(\tilde{\theta}_{\text{fine},j}^{(i)})) \\
&= -S_R(\eta T_{c\sqrt{n}}(\theta_j^{(i)} + (\tilde{\theta}_{\text{coarse},j}^{(s_j)} - \theta_j^{(s_j)}))) \\
&= -S_R(\eta T_{c\sqrt{n}}(\overline{\theta}_j^{(i)} + (\tilde{\theta}_{\text{coarse},j}^{(s_j)} - \theta_j^{(s_j)}))),
\end{aligned}$$

while $\overline{w}_j^{(i)} = -S_R(\eta T_{c\sqrt{n}}(\overline{\theta}_j^{(i)} + \overline{a}_j))$. Therefore, $\tilde{w}_j^{(i)} \neq \overline{w}_j^{(i)}$ implies that $\overline{a}_j \neq \tilde{\theta}_{\text{coarse},j}^{(s_j)} - \theta_j^{(s_j)}$, which means that $s_j \geq i_{\text{bad}}$. Hence, $i \geq i_{\text{bad}}$ as well. This implies that there is an iteration where $\|\tilde{\theta}_{\text{coarse}}^{(i_{\text{bad}})} - \theta^{(i_{\text{bad}})}\|_\infty > \Delta\sqrt{n}$, such that the RDA algorithm and Algorithm 1 make the same predictions for all $i' < i_{\text{bad}}$. We will focus the rest of the proof on showing this is impossible.

Since RDA and Algorithm 1 make the same predictions for $i' < i_{\text{bad}}$, we in particular have $\theta^{(i_{\text{bad}})} = \overline{\theta}^{(i_{\text{bad}})}$. Now, using the count sketch guarantee, we have

$$\|\tilde{\theta}_{\text{coarse}}^{(i_{\text{bad}})} - \theta^{(i_{\text{bad}})}\|_\infty^2 \leq \epsilon\|\theta^{(i_{\text{bad}})} - \theta_{\text{top k}}^{(i_{\text{bad}})}\|_2^2 = \epsilon\|\overline{\theta}^{(i_{\text{bad}})} - \overline{\theta}_{\text{top k}}^{(i_{\text{bad}})}\|_2^2.$$

But since U has size k, we have the bound

$$\|\overline{\theta}^{(i_{\text{bad}})} - \overline{\theta}_{\text{top k}}^{(i_{\text{bad}})}\|_2^2 \leq \sum_{j \notin U} |\overline{\theta}_j^{(i_{\text{bad}})}|^2 \leq \mathsf{G}^2 n(d-k).$$

Putting these together, we have

$$\|\tilde{\theta}^{(i_{\text{bad}})}_{\text{coarse}} - \theta^{(i_{\text{bad}})}\|_\infty^2 \le \epsilon(d-k)\mathsf{G}^2 n \le \Delta^2 n,$$

which contradicts the assumption that $\|\tilde{\theta}^{(i_{\text{bad}})}_{\text{coarse}} - \theta^{(i_{\text{bad}})}\|_\infty > \Delta\sqrt{n}$.

Since both cases lead to a contradiction, Algorithm 1 and the RDA procedure must match. Moreover, the condition $c \ge 2\Delta + \mathsf{G}$ ensures that no coordinate outside of U is added to the active set $\tilde{U}$, which completes the proof.

A.8. Proof of Lemma 6

By the general regret bound for mirror descent (Theorem 2.15[1] of Shalev-Shwartz (2011)),

$$\begin{aligned}
\mathsf{Reg} = \sum_{i=1}^n f_i(w) - f_i(w^*) &\le \frac{1}{2\eta}\|w^*\|_2^2 + c\sqrt{n}\|w^*\|_1 + \bar{a}^\top w^* + \frac{\eta}{2}\sum_{i=1}^n \|z_U^{(i)}\|_2^2 \\
&\le \frac{1}{2\eta}R^2 + (c\sqrt{n} + \|\bar{a}\|_\infty)\|w^*\|_1 + \frac{\eta}{2}\sum_{i=1}^n \|x_U^{(i)}\sqrt{2f_i(w^{(i)})}\|_2^2 \\
&\le \frac{1}{2\eta}R^2 + (c+\Delta)\sqrt{nk}R + k\eta\rho^2\sum_{i=1}^n f(w^{(i)}) \\
&\le \frac{1}{2\eta}R^2 + (c+\Delta)\sqrt{nk}R + k\eta\rho^2\left(E + \mathsf{Reg}\right).
\end{aligned}$$

Now note that, by the condition of the proposition, we can take $\mathsf{G} = \rho\sqrt{\frac{2(\mathsf{Reg}+E)\log(2d/\delta)}{n}}$; hence setting $\Delta = \rho\sqrt{\frac{2\epsilon(d-k)(\mathsf{Reg}+E)\log(2d/\delta)}{n}}$, $c = \rho\sqrt{\frac{2(\mathsf{Reg}+E)\log(2d/\delta)}{n}}\left(1 + 2\sqrt{\epsilon(d-k)}\right)$ will satisfy the conditions of Lemma 5. Recalling the definition of V above, we thus have the bound

$$\mathsf{Reg} \le \frac{1}{2\eta}R^2 + \sqrt{k}R\rho V\sqrt{\mathsf{Reg}+E} + k\eta\rho^2(\mathsf{Reg}+E).$$

Re-arranging, we have:

$$\left(1 - k\eta\rho^2\right)\mathsf{Reg} \le \frac{R^2}{2\eta} + \sqrt{k}R\rho V\sqrt{\mathsf{Reg}+E} + k\eta\rho^2 E.$$

Now set $\eta = \min\left(\frac{R}{\rho\sqrt{2kE}}, \frac{1}{2k\rho^2}\right)$; note that the second term in the $\min$ allows us to divide through by $1 - k\eta\rho^2$ while only losing a factor of 2. We then have

$$\mathsf{Reg} \le 2R\rho\sqrt{k}\left(\sqrt{2E} + V\sqrt{\mathsf{Reg}+E}\right) + 2kR^2\rho^2,$$

where the first term is the bound when $\eta = R/\rho\sqrt{2kE}$ and the second term upper-bounds $\frac{R^2}{\eta}$ in the case that $\eta = 1/2k\rho^2$. Solving the quadratic,[2] we have

$$\mathsf{Reg} \le 2R\rho\sqrt{k}\left(2\sqrt{2} + V\right)\sqrt{E} + kR^2\rho^2\left(4 + 4V^2\right),$$

which completes the proof.

[1] Note that Shalev-Shwartz (2011) refers to the procedure as online mirror descent rather than regularized dual averaging, but it is actually the same procedure.

[2] More precisely, we use the fact that if $x \le a + \sqrt{bx+c}$ for $a, b, c \ge 0$, then $x \le 2a + b + \sqrt{c}$.

A.9. Proof of Lemma 7

Note that

$$\begin{aligned}\overline{\theta}_j^{(i+1)} - \overline{\theta}_j^{(i)} &= z_j^{(i)} \\ &= (y^{(i)} - (\tilde{w}^{(i)})^\top x^{(i)})x_j^{(i)} \\ &= ((w^* - \tilde{w}^{(i)})^\top x^{(i)} + \epsilon^{(i)})x_j^{(i)}.\end{aligned}$$

Since $\mathrm{supp}(w^* - \tilde{w}^{(i)}) \subseteq U$ and $j \notin U$, this is a zero-mean random variable (since $\mathrm{Cov}[X^{(i)}] = I$), and so $\overline{\theta}_j^{(i)}$ is a martingale difference sequence. Moreover, letting $\mathcal{F}_i$ be the sigma-algebra generated by the first $i-1$ samples, we have that

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\lambda\left(\overline{\theta}_j^{(i+1)} - \overline{\theta}_j^{(i)}\right) - \rho^2\lambda^2 f_i(\overline{w}^{(i)})\right) \mid \mathcal{F}_i\right] \\ &= \mathbb{E}\left[\exp\left(\lambda x_j^{(i)}\left(y^{(i)} - (\overline{w}^{(i)})^\top x^{(i)}\right) - \rho^2\lambda^2 f_i(\overline{w}^{(i)})\right) \mid \mathcal{F}_i\right] \\ &= \mathbb{E}\left[\exp\left(\lambda x_j^{(i)}\left(y^{(i)} - (\overline{w}^{(i)})^\top x^{(i)}\right)\right) \mid \mathcal{F}_i\right] \\ &\overset{(i)}{\leq} \mathbb{E}\left[\exp\left(\frac{1}{2}\lambda^2\rho^2\left(y^{(i)} - (\overline{w}^{(i)})^\top x^{(i)}\right)^2 - \rho^2\lambda^2 f_i(\overline{w}^{(i)})\right) \mid \mathcal{F}_i\right] \\ &= 1.\end{aligned}$$

Here (i) is obtained by marginalizing over $x_j^{(i)}$, using the fact that $x_j^{(i)}$ is independent of both $y^{(i)}$ and $(\overline{w}^{(i)})^\top x^{(i)}$ (since $\mathrm{supp}(w^*) \subseteq U$ and $\mathrm{supp}(\overline{w}^{(i)}) \subseteq U$), together with the fact that $|x_j^{(i)}| \leq \rho$ and hence is sub-Gaussian.

Consequently, for any λ, $Y_i = \exp\left(\lambda\overline{\theta}_j^{(i+1)} - \rho^2\lambda^2\sum_{k=1}^{i} f_k(\overline{w}^{(k)})\right)$ is a supermartingale. For any t, define the stopping criterion $Y_i \geq t$. Then, by Doob's optional stopping theorem (Corollary 5.11 of Breiman (1992)) applied to this stopped process, together with Markov's inequality, we have $\mathbb{P}\left[\max_{i=1}^{n} Y_i \geq t\right] \leq \frac{1}{t}$.

Bounding $\sum_{k=1}^{i} f_k$ by $\sum_{k=1}^{n} f_k$ and inverting, for any fixed j and λ, with probability $1-\delta$, we have

$$\max_{i=1}^{n} \overline{\theta}_j^{(i)} \leq \lambda\rho^2 \sum_{i=1}^{n} f_i(\overline{w}^{(i)}) + \frac{1}{\lambda}\log(1/\delta).$$

Union-bounding over $j = 1, \dots, d$ and $\pm\overline{\theta}_j$ changes the $\log(\cdot)$ term to $\log(2d/\delta)$, yielding the high-probability-bound

$$\max_{i=1}^{n} \|\overline{\theta}^{(i)}\|_\infty \leq \lambda\rho^2 \sum_{i=1}^{n} f_i(\overline{w}^{(i)}) + \frac{1}{\lambda}\log(2d/\delta).$$

Using the equality $\sum_{i=1}^{n} f_i(\overline{w}^{(i)}) = E + \mathrm{Reg}$ and choosing the optimal value of λ yields the desired result.

Remark 1 *Note that λ is set adaptively based on $E + \mathrm{Reg}$, which depends on the randomness in $\overline{\theta}^{(1:n)}$ and thus could be problematic. However, later in our argument we end up with an upper bound on $E + \mathrm{Reg}$ that depends only on the problem parameters; if we set λ based on this upper bound, our proofs still go through, and we eliminate the dependence of λ on $\overline{\theta}$.*

A.10. Proof of Lemma 8

We prove the inequality using standard symmetrization inequalities. First, we use the Høffmann-Jorgensen inequality (e.g. de la Peña and Giné, 1999, Theorem 1.2.3)), which states that given symmetric random variables U_i and $S_n = \sum_{i=1}^n U_i$ and some t_0 is such that $\mathbb{P}\left[|S_n| \geq t_0\right] \leq 1/8$, then for any $p > 0$,

$$\mathbb{E}\left[\left|\sum_{i=1}^n U_i\right|^p\right]^{1/p} \leq K_p\left(t_0 + \mathbb{E}\left[\max_{i \leq n} |U_i|^p\right]^{1/p}\right), \text{ where } K_p \leq 2^{1+\frac{4}{p}} \exp\left(\frac{p+1}{p}\log(p+1)\right). \tag{26}$$

Notably, for any $p \geq 5$ we have $K_p \leq 6p$, for any $p \geq 1$ we have $K_p \leq 128p$, and $K_p/p \to 2$ as $p \to \infty$.

Now, recall that $\epsilon^{(i)}$ is sub-exponential, meaning that there is some σ such that $\mathbb{E}\left[|\epsilon^{(i)}|^p\right]^{1/p} \leq \sigma p$ for all $p \geq 1$. Letting $r_i \in \{\pm 1\}$ be i.i.d. Rademacher variables and $V_i = (\epsilon^{(i)})^2$ be shorthand, an immediate symmetrization inequality (e.g. de la Peña and Giné, 1999, Lemma 1.2.6) gives that

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^n V_i \geq \sigma^2 n + t\right] &\leq \mathbb{P}\left[\sum_{i=1}^n V_i \geq \sum_{i=1}^n \mathbb{E}\left[V_i\right] + t\right] \leq t^{-p}\mathbb{E}\left[\left|\sum_{i=1}^n (V_i - \mathbb{E}\left[V_i\right])\right|^p\right] \\ &\leq 2^p t^{-p}\mathbb{E}\left[\left|\sum_{i=1}^n r_i V_i\right|^p\right].\end{aligned}$$

Now we apply the Høffmann-Jorgensen inequality (26) to $U_i = r_i V_i$, which are symmetric and satisfy

$$\mathbb{P}\left[\left|\sum_{i=1}^n r_i V_i\right| \geq t_0\right] \leq \frac{1}{t_0^2}\mathbb{E}\left[\left|\sum_{i=1}^n r_i V_i\right|^2\right] = \frac{1}{t_0^2}\mathbb{E}\left[\sum_{i=1}^n V_i^2\right] \leq \frac{4n\sigma^4}{t_0^2},$$

so that taking $t_0 = 6\sigma^2\sqrt{n}$ and applying inequality (26) yields

$$\mathbb{E}\left[\left|\sum_{i=1}^n r_i V_i\right|^p\right] \leq K_p^p\left(6\sigma^2\sqrt{n} + \mathbb{E}\left[\max_{i \leq n} V_i^p\right]^{1/p}\right)^p.$$

Moreover, we have that

$$\mathbb{E}\left[\max_{i \leq n} V_i^p\right] \leq \sum_{i=1}^n \mathbb{E}\left[V_i^p\right] \leq n\mathbb{E}\left[{\epsilon^{(i)}}^{2p}\right] \leq n\sigma^{2p}(2p)^{2p}.$$

In particular, we have

$$\mathbb{P}\left[\sum_{i=1}^n V_i \geq \sigma^2 n + 4K_p\sigma^2 t\right] \leq \left(\frac{2}{4K_p\sigma^2 t}\right)^p K_p^p \sigma^{2p}\left(3\sqrt{n} + 2n^{1/p}p^2\right)^p = \left(\frac{3\sqrt{n} + 2n^{1/p}p^2}{t}\right)^p. \tag{27}$$

Taking $p = \log\frac{1}{\delta}$ and $t = e \cdot \left(3\sqrt{n} + 2n^{\log^{-1}\frac{1}{\delta}}\log^2\frac{1}{\delta}\right)$ gives the desired result.

Appendix B. From real numbers to bits

In Section 4, we analyzed a procedure that stored $\tilde{O}(B)$ real numbers. We now argue that each number only needs to be stored to polynomial precision, and so requires only $\tilde{O}(1)$ bits. Algorithm 1 stores the quantities $\tilde{U}$, $\hat{w}$, $\tilde{\theta}_{\text{fine}}$, and the count sketch structure ($\tilde{\theta}_{\text{coarse}}$). The set $\tilde{U}$ requires $O(k \log d)$ bits. To handle $\tilde{\theta}_{\text{coarse}}$ and $\tilde{\theta}_{\text{fine}}$, we randomly (and unbiasedly) round each $g_j^{(i)}$ to the nearest multiple of $\frac{1}{M}$ for some large integer M; for instance, a value of $\frac{4.1}{M}$ will be rounded to $\frac{4}{M}$ with probability 0.9 and to $\frac{5}{M}$ with probability 0.1. Since this yields an unbiased estimate of $g_j^{(i)}$, the RDA procedure will still work, with the overall regret bound increasing only slightly (essentially, by $\frac{kn}{M^2}$ in expectation).

If $M' \stackrel{\text{def}}{=} \left\lceil \max_{i \in [n], j \in [d]} |g_j^{(i)}| \right\rceil$, then the number of bits needed to represent each coordinate of $\tilde{\theta}_{\text{fine}}$ (as well as each number in the count sketch structure) is $O(\log(nMM'))$. But $|g_j^{(i)}| = |x_j^{(i)}| \sqrt{2 f_i(\tilde{w}^{(i)})}| \leq \rho \sqrt{2(E + \mathsf{Reg})}$, which is polynomial in n, so M' is polynomial in n. In addition, for M growing polynomially in n, the increase of $\frac{kn}{M^2}$ in the regret bound becomes negligibly small. Hence, we can take nMM' to be polynomial in n, thus requiring $\tilde{O}(1)$ bits per coordinate to represent $\tilde{\theta}_{\text{coarse}}$ and $\tilde{\theta}_{\text{fine}}$. Finally, to handle $\hat{w}$, we deterministically round to the nearest multiple of $\frac{1}{M}$. Since $\hat{w}$ does not affect any choices in the algorithm, the accumulated error per coordinate after n steps is at most $\frac{n}{M}$, which is again negligible for large M. Since each coordinate of $\hat{w}$ is at most R, we can store them each with $\tilde{O}(1)$ bits, as well.