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 $1-\delta$,

$$\|Q^s(w) - w\|_p \le 4\|w\|_1\left(\frac{1}{s}\right)^{1-\frac{1}{p}} + \|w\|_1\left(\frac{8\log(1/\delta)}{s}\right)^{\frac{1}{2}} \le \|w\|_1\left(\frac{24\log(1/\delta)}{s}\right)^{1-\frac{1}{p}}. \tag{7}$$

Moreover, the following in-expectation guarantee holds:

$$\mathbb{E}\|Q^s(w) - w\|_p \le \left(\mathbb{E}\|Q^s(w) - w\|_p^p\right)^{1/p} \le 4\|w\|_1\left(\frac{1}{s}\right)^{1-\frac{1}{p}}. \tag{8}$$

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

$$\begin{aligned}\|Q^s(w) - w\|_p &\le \frac{2}{s}\cdot\mathbb{E}_Z\left(\sum_{t=1}^{s}\|Z_t\|_p^p\right)^{1/p} + B\sqrt{\frac{8\log(1/\delta)}{s}}\\ &\le \frac{4B}{s^{1-\frac{1}{p}}} + B\sqrt{\frac{8\log(1/\delta)}{s}}.\end{aligned}$$

□

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 $\|Z_t\|_p \le B$ almost surely and $\mathbb{E}[Z_t] = 0$. Then with probability at least $1-\delta$,

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

Furthermore, a sharper guarantee holds in expectation:

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

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 & Ben-David, 2014)), which states that with probability at least $1-\delta$,

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

where $\epsilon \in \{\pm1\}^n$ are Rademacher random variables. Conditioning on $Z_1,\ldots,Z_n$, we have

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

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

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

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

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

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

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

□

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 $\|w\|_1 \le B$ and $\|x\|_\infty \le R$ almost surely, one has $|Z_\tau| \le 2BR$ 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

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

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

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

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

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

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

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

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

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. □

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 $q \ge 2$. Then Algorithm 1 guarantees that

$$\mathbb{E}\, F(Q^s(w)) \le F(w) + \frac{\beta_q \|w\|_1^2}{s}. \tag{9}$$

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

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

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

$$\begin{aligned}\mathbb{E}_{Z_s} F\left(w + \frac{1}{s}\sum_{\tau=s}^{s} Z_\tau\right) &\le F\left(w + \frac{1}{s}\sum_{\tau=s}^{s-1} Z_\tau\right) + \mathbb{E}_{Z_s}\left\langle \nabla F\left(w + \frac{1}{s}\sum_{\tau=s}^{s-1} Z_\tau\right), \frac{Z_s}{s}\right\rangle + \frac{\beta_q}{2s^2}\,\mathbb{E}_{Z_s}\|Z_s\|_p^2 \\ &\le F\left(w + \frac{1}{s}\sum_{\tau=s}^{s-1} Z_\tau\right) + \frac{\beta_q}{s^2}\|w\|_1^2.\end{aligned}$$

Proceeding backwards in the same fashion, we get

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

□

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 $\|a\|_1 \vee \|b\|_1 \vee \|c\|_1 \le B$. Then it holds that

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

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 $\|x\| \vee \|y\| \le B$, we have

$$\|x\|^2 - \|y\|^2 \le \|x - y\|^2 + 2\|x - y\|\|y\| \le 4B\|x - y\|. \tag{10}$$

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

$$\nabla\mathcal{R}(\theta) = \|\theta\|_p^{2-p} \cdot \left(|\theta_1|^{p-1}\mathrm{sgn}(\theta_1), \ldots, |\theta_d|^{p-1}\mathrm{sgn}(\theta_d)\right). \tag{11}$$

Proposition 6. For any vector θ,

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

Proof of Proposition 6. Expanding the expression in (11), 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

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

□

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 (10) 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.

□

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

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

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

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

Putting everything together, this establishes that

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

□

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

$$\|\nabla\mathcal{R}(a) - \nabla\mathcal{R}(b)\|_\infty \leq 2B^{2-p}\|a-b\|_\infty^{p-1} + \|a-b\|_p, \tag{12}$$

and

$$\|\nabla\mathcal{R}(a) - \nabla\mathcal{R}(b)\|_q \leq 2B^{2-p}\|a-b\|_p^{p-1} + \|a-b\|_p. \tag{13}$$

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

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

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:

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

Using the triangle inequality:

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

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

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

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

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

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

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

For the ℓ_∞ bound, the result follows immediately by using that $\left|\bar{b}_i\right| \leq \left\|\bar{b}\right\|_p \leq 1$. 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 $\left\|\bar{b}\right\|_p \leq 1$.

□

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

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

To conclude, we plug in the bound from Lemma 6. □

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 & Nelson (2010), Theorem 2. Let $x'_t = Ax_t$ denote the projected feature vector and $w^\star = \arg\min_{w:\|w\|_2 \le 1} L_{\mathcal{D}}(w)$.

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

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

and so

$$L_{\mathcal{D}}(A^\top A w^\star) - L_{\mathcal{D}}(w^\star) \le L \cdot \mathbb{E}_x \, |\langle Ax, Aw^\star\rangle - \langle x, w^\star\rangle| . \tag{14}$$

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 $u \in \mathbb{R}^k$:

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

Equivalently, we have

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

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 & Lugosi, 2006) yields the following guarantee for any vector u:

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

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

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

or in other words,

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

We now relate this bound to the risk relative to the benchmark $L_{\mathcal{D}}(w^\star)$. Using (14) we have

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

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

$$\mathbb{E}_S\mathbb{E}_A\left[L_{\mathcal{D}}(\hat{w})\right] - L_{\mathcal{D}}(w^\star) \le \mathbb{E}_x\left[\mathbb{E}_A\left[\frac{1}{2\eta N}\|Aw^\star\|_2^2 + \frac{\eta L^2}{2}\|Ax\|_2^2 + L\,|\langle Ax, Aw^\star\rangle - \langle x, w^\star\rangle|\right]\right]. \tag{15}$$

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 (15) is taken. The derandomized JL transform guarantees (Kane & Nelson (2010), Theorem 2) that for any $\delta > 0$ and any fixed vectors $x, w^\star$, if we pick $k = O\left(\log(1/\delta)/\varepsilon^2\right)$, then with probability at least $1 - \delta$,

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

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

$$\|Ax\|_2 \le O(R_2), \quad \|Aw^\star\|_2 \le O(B_2), \quad \text{and} \quad |\langle Ax, Aw^\star\rangle - \langle x, w^\star\rangle| \le O\bigg(\frac{B_2R_2}{\sqrt{N}}\bigg).$$

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

$$\begin{aligned}&\mathbb{E}_A\left[\frac{1}{2\eta N}\|Aw^\star\|_2^2 + \frac{\eta L^2}{2}\|Ax\|_2^2 + L\,|\langle Ax, Aw^\star\rangle - \langle x, w^\star\rangle|\right]\\ &\le (1-\delta)\cdot O\left(\frac{B_2^2}{2\eta N} + \frac{\eta L^2R_2^2}{2} + \frac{LB_2R_2}{\sqrt{N}}\right) + \delta\cdot O\left(\mathrm{poly}(d)\cdot\left(\frac{B_2^2}{2\eta N} + \frac{\eta L^2R_2^2}{2} + LB_2R_2\right)\right).\end{aligned}$$

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

$$\mathbb{E}_A\left[\frac{1}{2\eta N}\|Aw^\star\|_2^2 + \frac{\eta L^2}{2}\|Ax\|_2^2 + L\,|\langle Ax, Aw^\star\rangle - \langle x, w^\star\rangle|\right] \le O(LB_2R_2/\sqrt{N}).$$

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

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

Using this inequality to bound the right-hand side in (15) yields the claimed excess risk bound. Recall that we have $k = O\big(\log(1/\delta)/\varepsilon^2\big) = O\left(N\log(Nd)\right)$, and so the communication cost to send a single iterate (taking into account numerical precision) is upper bounded by $O(N\log(Nd)\cdot\log(LB_2R_2N))$. □

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 $\{\pm1\}^d$ 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

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

Proof of Theorem 2. Recall that $\mathcal{W}_1 = \left\{w\in\mathbb{R}^d \mid \|w\|_1 \le 1\right\}$. 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

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

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

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

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

□

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). □

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 $i \neq k$ 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 $w^\star = \arg\min_{w\in\|w\|_1\leq 1} L_{\mathcal{D}_j}(w)$. It is clear from the expression above $w_i^\star = 0$ for all $i \neq j$. For coordinate j we have $w_j^\star = \arg\min_{-1\leq\alpha\leq 1}\{\alpha^2 - 4\rho\alpha\}$. Whenever $\rho \leq 1/2$ 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). □

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 $\widehat{w}^i := w_{n+1}^i$.

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 $\|w\|_1 \leq B$. Then $D_{\mathcal{R}}(a\|b) \geq \frac{p-1}{2}\|a-b\|_p^2$ and if $\|a\|_1 \vee \|b\|_1 \vee \|c\|_1 \leq B$ it holds that

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

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 $\|a - \bar{w}\|_1 \leq 2B$, and likewise for b and c. □

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:

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

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 & Nemirovski, 2001) implies that for every time t, we have

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

Using Proposition 7, we have an upper bound of

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

Using Hölder's inequality and AM-GM:

$$\begin{aligned}
&\leq \frac{\eta}{2(p-1)}\left\|\nabla_t^i\right\|_q^2 + \frac{p-1}{2\eta}\left\|w_t^i - \theta_{t+1}^i\right\|_p^2 + \frac{1}{\eta}\left(D_{\mathcal{R}}(w^\star\|w_t^i) - D_{\mathcal{R}}(w^\star\|w_{t+1}^i) - \frac{p-1}{2}\left\|w_t^i - \theta_{t+1}^i\right\|_p^2\right) \\
&= \frac{\eta}{2(p-1)}\left\|\nabla_t^i\right\|_q^2 + \frac{1}{\eta}\left(D_{\mathcal{R}}(w^\star\|w_t^i) - D_{\mathcal{R}}(w^\star\|w_{t+1}^i)\right).
\end{aligned}$$

The result follows by summing across time and observing that the Bregman divergences telescope. □

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

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

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

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

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 $\mathbb{E}\left\|\widehat{w}^i - \widetilde{w}^i\right\|_\infty^{p-1} \leq O\left(B^{p-1}\left(\frac{1}{s}\right)^{\frac{p-1}{2}}\right)$.[9] 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

$$\mathbb{E}[L_{\mathcal{D}}(\widetilde{w})] - L_{\mathcal{D}}(w^\star) \leq \frac{\eta C_q}{2N}\sum_{i=1}^{m}\sum_{t=1}^{n}\mathbb{E}\left\|\nabla_t^i\right\|_q^2 + \frac{D_{\mathcal{R}}(w^\star\|\bar{w})}{\eta N} + O\left(\frac{B^2 m}{\eta N}\left(\frac{1}{s}\right)^{\frac{p-1}{2}}\right). \tag{16}$$

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

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

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.

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

In the linear model case, Lemma 1 directly implies that

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

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

□

Proof of Theorem 4. We begin from (16) in the proof of Theorem 3 which, once $s = \Omega(m^{\frac{2}{p-1}})$, implies

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

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

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

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 $\mathbb{E}\,L_{\mathcal{D}}(\widehat{w}) \leq L_{\mathcal{D}}(\widetilde{w}) + \frac{\beta_q B^2}{s_0}$. 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 $\widetilde{w}$. □

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 $\mathbb{E}\|\bar{w} - w^\star\|_p^2 \leq \bar{B}^2$. Suppose that subgradients belong to $\mathcal{X}_q$ for $q \geq 2$, and that $\mathcal{W} \subseteq \mathcal{W}_1$. Then, using learning rate $\eta := \frac{\bar{B}}{R}\sqrt{\frac{1}{C_qN}}$, $s = \Omega\left(m^{2(q-1)}(B/\bar{B})^{4(q-1)}\right)$, and $s_0 = \Omega((N/C_q)^{\frac{q}{2}} \cdot (B/\bar{B})^q)$, the algorithm guarantees

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

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

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

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

```
**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 $\ell_q$ norm bound $R_q$.
    RSC constant $\gamma_q$. 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 $T = \max\{T \mid \sum_{k=1}^{T} N_k \le N\}$.
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 $\eta$ as prescribed in Proposition 8, with initialization $\bar{w} = \widehat{w}_{k-1}$
           and radius $\bar{B} = B_{k-1}$.
Return $\widehat{w}_T$.
```

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

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

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 $O\left(\log d \cdot \left(\frac{R}{\gamma B}\right)^{q} 2^q\right) \cdot \sum_{k=1}^{T} 2^{qk}$. 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

$$O\left(\log d \cdot \left(\frac{\gamma^2 B^2}{C_q R^2}\right)^{2(q-1)} m^{2q-1} N^{2(q-1)} + \log d \cdot \left(\frac{\gamma B N}{C_q R}\right)^{q}\right). \tag{17}$$

□

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 $\gamma_q = \frac{\gamma}{4k}$.

□

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 $\|(w - w^\star)_{S^c}\|_1 \le \|(w - w^\star)_S\|_1$.

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

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

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

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

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

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 $p \le 2$. 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 $\|W\|_{S_1} \vee \|W^\star\|_{S_1} \le B$,

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

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

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

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

$$\mathbb{E}\|\sigma - \widehat{\sigma}\|_p \le 4B\left(\frac{1}{s}\right)^{1-1/p} \quad \text{and} \quad \sqrt{\mathbb{E}\|\sigma - \widehat{\sigma}\|_\infty^2} \le 4B\left(\frac{1}{s}\right)^{1/2}. \tag{18}$$

Returning to the Bregman divergence, we write

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

It follows immediately using Lemma 4 that

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

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 (11), 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 (18). □