Sparsifying generalized linear models

Arun Jambulapati* James R. Lee† Yang P. Liu‡ Aaron Sidford§

Abstract

We consider the sparsification of sums $F : \mathbb{R}^n \to \mathbb{R}_+$ where $F(x) = f_1(\langle a_1, x\rangle) + \cdots + f_m(\langle a_m, x\rangle)$ for vectors $a_1, \ldots, a_m \in \mathbb{R}^n$ and functions $f_1, \ldots, f_m : \mathbb{R} \to \mathbb{R}_+$. We show that $(1+\varepsilon)$-approximate sparsifiers of F with support size $\frac{n}{\varepsilon^2}(\log \frac{n}{\varepsilon})^{O(1)}$ exist whenever the functions $f_1, \ldots, f_m$ are symmetric, monotone, and satisfy natural growth bounds. Additionally, we give efficient algorithms to compute such a sparsifier assuming each f_i can be evaluated efficiently.

Our results generalize the classic case of ℓ_p sparsification, where $f_i(z) = |z|^p$, for $p \in (0, 2]$, and give the first near-linear size sparsifiers in the well-studied setting of the Huber loss function and its generalizations, e.g., $f_i(z) = \min\{|z|^p, |z|^2\}$ for $0 < p \leqslant 2$. Our sparsification algorithm can be applied to give near-optimal reductions for optimizing a variety of generalized linear models including ℓ_p regression for $p \in (1, 2]$ to high accuracy, via solving $(\log n)^{O(1)}$ sparse regression instances with $m \leqslant n(\log n)^{O(1)}$, plus runtime proportional to the number of nonzero entries in the vectors $a_1, \ldots, a_m$.

*University of Washington, `jmblpati@uw.edu`, †University of Washington, `jrl@cs.washington.edu`
‡Institute for Advanced Study, `yangpliu@ias.edu`, § Stanford University, `sidford@stanford.edu`

1 Introduction

Empirical risk minimization (ERM) is a widely studied problem in learning theory and statistics (see, e.g., [LSZ19a], for relevant references to the expansive literature on this topic). A prominent special case is the problem of optimizing a *generalized linear model (GLM)*, i.e.,

$$\min_{x\in\mathbb{R}^n} F(x) \quad \text{for} \quad F(x) := \sum_{i=1}^{m} f_i(\langle a_i, x\rangle - b_i)\,, \tag{1.1}$$

where the *total loss* $F : \mathbb{R}^n \to \mathbb{R}$, is defined by vectors $a_1, \ldots, a_m \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, and *loss functions* $f_1, \ldots, f_m : \mathbb{R} \to \mathbb{R}$. Different choices of the loss functions $\{f_i\}$ capture important problems, including linear regression, logistic regression, and ℓ_p regression [BCLL18, AKPS19b].

Recently, efficient algorithms for solving (1.1) to high-accuracy have been developed in many settings [BLSS20, BLL⁺21, GPV21] such as linear programming and ℓ_1-regression, where $f_i(x) = |x|$. For example, when m is on the order of n, it is known how to solve linear programs and some GLMs in roughly (up to logarithmic factors) the time it currently takes to multiply two general $n \times n$ matrices [AKPS19b, CLS21, LSZ19a, JSWZ21] which is, up to logarithmic factors, the best-known, running time for solving a single linear system in a dense $n \times n$ matrix.

When $m \gg n$, a natural approach for fast algorithms is to apply sparsification techniques to reduce the value of m, while maintaining a good multiplicative approximation of the objective value. More precisely, say that the objective F *admits an s-sparse ε-approximation* if there are non-negative weights $w_1, \ldots, w_m \in \mathbb{R}_+^m$, at most s of which are non-zero, and such that

$$|F(x) - \tilde{F}(x)| \leqslant \varepsilon\, F(x) \quad \text{for all } x \in \mathbb{R}^n, \text{ where } \tilde{F}(x) := \sum_{i=1}^{m} w_i f_i(\langle a_i, x\rangle - b_i)\,.$$

When $f_i(z) = |z|^p$ are ℓ_p losses, near-optimal sparsification results are known: If $p > 0$, then F admits an s-sparse ε-approximation for $s \leqslant \tilde{O}(n^{\max\{1,p/2\}}\varepsilon^{-2})$;[1] this sparsity bound is known to be optimal up to polylogarithmic factors [BLM89, Tal90, Tal95, SZ01]. In particular, for $p \in (0, 2]$, the size is $\tilde{O}(n\varepsilon^{-2})$, near-linear in the underlying dimension n. The $p = 2$ case has been especially influential in the development of several fast algorithms for linear programming and graph optimization over the last two decades [ST14, SS11, BLN⁺20].

However, as far as the authors know, ℓ_p losses are the only class of natural loss functions for which linear-size sparsification results are known for GLMs. For instance, for the widely-studied class of Huber loss functions (see (1.2)) and related variants, e.g., $f_i(z) = \min\{|z|, |z|^2\}$, the best known sparsity bound was $\tilde{O}(n^{4-2\sqrt{2}}\varepsilon^{-2})$ [MMWY22]. Improving this bound to near-linear (in n) is an established important open problem that has potential applications to regression for Huber and ℓ_p losses [AS20, ABKS21, GPV21, MMWY22, WY23].

The main result of this paper is near-optimal sparsification for a large family of loss functions $\{f_i\}$ that include the Huber losses, ℓ_p losses, and generalizations. Informally, we show that if the loss functions $\{f_i\}$ are nonnegative, symmetric, and grow at most quadratically, then there exists an s-sparse ε-approximation of F with $s \leqslant \tilde{O}(n\varepsilon^{-2})$. Moreover, the sparse approximation can be

[1] Throughout, we use $\tilde{O}(f)$ to suppress polylogarithmic quantities in m, n, ε^{-1}, and f.

found very efficiently, in time proportional to the time used for $\tilde{O}(1)$ instances of ℓ_2-sparsification (Theorem 1.1). A particularly nice application of our result is an algorithm that solves ℓ_p-regression to high accuracy for $1 < p \leqslant 2$ by reducing to $\tilde{O}_p(1)$ instances of ℓ_p-regression with $m = \tilde{O}(n)$ (Theorem 1.2). Our framework can also be applied to minimizing sums of γ_p functions for $p \in (1, 2]$ (see (1.2)) to high accuracy, and to approximate Huber regression.

The main technical hurdle in obtaining these results is that the loss functions are not necessarily homogeneous and they can exhibit different behaviors at different scales. Note that this hurdle arises already for losses like $f_i(z) = \min\{|z|, |z|^2\}$, even though the loss function only has two different scaling regimes. To overcome this hurdle we develop a multiscale notion of "importance scores" for appropriately down-sampling F into a sparse representation.

1.1 Hypotheses and results for sparsification

Consider a generalized linear model as in (1.1), with loss functions $f_1, \ldots, f_m : \mathbb{R} \to \mathbb{R}_+$ and vectors $a_1, \ldots, a_m \in \mathbb{R}^n$. For simplicity, we assume that $b = 0$ in (1.1). This is without loss of generality, as $\langle a_i, x\rangle - b_i = \langle (a_i, b_i), (x, -1)\rangle$, and $(a_i, b_i), (x, -1) \in \mathbb{R}^{n+1}$, so we can re-encode the problem in $n + 1$ dimensions with $b = 0$.

We will often think of the case $f_i(z) = h_i(z)^2$ for some $h_i : \mathbb{R} \to \mathbb{R}_+$, as the assumptions we need are stated more naturally in terms of $\sqrt{f_i}$. To that end, consider a function $h : \mathbb{R}^k \to \mathbb{R}_+$ and the following two properties, where $L \geqslant 1$ and $c, \theta > 0$ are some positive constants.[2]

(P1) (L-auto-Lipschitz) $|h(z) - h(z')| \leqslant L\, h(z - z')$ for all $z, z' \in \mathbb{R}^k$.

(P2) (Lower θ-homogeneous) $h(\lambda z) \geqslant c\lambda^\theta h(z)$ for all $z \in \mathbb{R}^k$ and $\lambda \geqslant 1$.

Note that if $h : \mathbb{R} \to \mathbb{R}$ is concave and symmetric, then it is 1-auto-Lipschitz (see Lemma 3.15).

We can now state our main theorem, whose proof appears in Section 3.3.

Theorem 1.1. *Consider $f_1, \ldots, f_m : \mathbb{R} \to \mathbb{R}_+$, and suppose there are numbers $L \geqslant 1, c, \theta > 0$ such that each $\sqrt{f_i}$ is L-auto-Lipschitz and lower θ-homogeneous (with constant c). Then for any $a_1, \ldots, a_m \in \mathbb{R}^n$, and numbers $0 < \varepsilon < \frac{1}{2}$ and $s_{\max} > s_{\min} \geqslant 0$, there are nonnegative weights $w_1, \ldots, w_m \geqslant 0$ such that*

$$\left| F(x) - \sum_{i=1}^{m} w_i f_i(\langle a_i, x\rangle) \right| \leqslant \varepsilon F(x)\,, \quad \forall x \in \mathbb{R}^n \ \text{ s.t. } \ s_{\min} \leqslant F(x) \leqslant s_{\max}\,,$$

where $F(x) := f_1(\langle a_1, x\rangle) + \cdots + f_m(\langle a_m, x\rangle)$, and

$$|i \in \{1, \ldots, m\} : w_i > 0| \lesssim_{L,c,\theta} \frac{n}{\varepsilon^2} \log\left(\frac{n}{\varepsilon}\frac{s_{\max}}{s_{\min}}\right) (\log S)^3\,, \quad \text{where} \quad S := \frac{n}{\varepsilon}\log\left(\frac{2 s_{\max}}{s_{\min}}\right).$$

Moreover, with high probability, the weights $\{w_i\}$ can be computed in time

$$\tilde{O}_{L,c,\theta}\big((\mathsf{nnz}(a_1, \ldots, a_m) + n^\omega + m\mathcal{T}_{\text{eval}}) \log(m s_{\max}/s_{\min})\big)\,.$$

[2]The setting $k = 1$ suffices for the present work, though we state them for general $k \geqslant 1$.

Here, $\mathcal{T}_{\text{eval}}$ is the maximum time needed to evaluate each f_i, $\mathsf{nnz}(a_1, \ldots, a_m)$ is the total number of non-zero entries in the vectors $a_1, \ldots, a_m$, and ω is the matrix multiplication exponent. "High probability" means that the failure probability can be made less than $n^{-\ell}$ for any $\ell > 1$ by increasing the running time by an $O(\ell)$ factor.

We use the notation $O_{L,c,\theta}$ and $\tilde{O}_{L,c,\theta}(\cdot)$ to indicate an implicit dependence on the parameters L, c, θ, and $A \lesssim_{L,c,\theta} B$ is shorthand for $A \leqslant O_{L,c,\theta}(B)$. The constant hidden by the $O_{L,c,\theta}(\cdot)$ notation is about $(L/c)^{O(\theta^{-2})}$, though we made no significant effort to optimize this dependence.

It is not difficult to see that for $0 < p \leqslant 2$, the function $f_i(z) = |z|^p$ satisfies the required hypotheses of Theorem 1.1. In Section 3.4, we show that γ_p functions, defined as

$$\gamma_p(z) := \begin{cases} \frac{p}{2} z^2 & \text{for} \quad |z| \leqslant 1 \\ |z|^p - (1 - \frac{p}{2}) & \text{for} \quad |z| \geqslant 1, \end{cases} \tag{1.2}$$

for $p \in (0, 2]$, also satisfy the conditions. The special case of γ_1 is known as the Huber loss. (See Section 3.4 for a generalization to general thresholds.)

The γ_p functions were introduced in [BCLL18] and have since been used in several works on high-accuracy ℓ_p regression [AKPS19b, ABKS21, GPV21]. Due to these connections, the works [GPV21, MMWY22] studied sparsification with γ_p losses, providing sparsity bounds of $\tilde{O}(n^3)$ and $\tilde{O}(n^{4-2\sqrt{2}}) \approx \tilde{O}(n^{1.172})$, respectively. More precisely, [MMWY22] establish a bound of $\tilde{O}(n^{1+\delta(p)})$ for $p \in [1, 2]$ with $\delta(1) = 3 - 2\sqrt{2}$, and $\delta(p) \to 0$ as $p \to 2$.

1.2 Fast ℓ_p regression

Combining our sparsification theorem with iterative refinement [AKPS19b] yields near-optimal reductions for solving ℓ_p regression to *high accuracy*. More specifically, we show that ℓ_p regression for matrices $A \in \mathbb{R}^{m \times n}$ can be reduced to a sequence of $\tilde{O}_p(1)$ instances with $\widetilde{A} \in \mathbb{R}^{\tilde{O}(n) \times n}$. It is known how to solve such instances in time n^{ω_0} for $\omega_0 := 2 + \max\left\{\frac{1}{6}, \omega - 2, \frac{1-\alpha}{2}\right\}$ [LSZ19a], where α is the dual matrix multiplication exponent. Alternatively, they can each be solved in roughly $n^{1/3}$ iterations and time $n^{\max\{\omega, 2+1/3\}}$ [AKPS19b], where an "iteration" refers to an operation that is dominated by the cost of solving a particular $n \times n$ linear system.

Theorem 1.2 (Fast ℓ_p regression)**.** *There is an algorithm that given any $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $p \in (1, 2]$ computes an x satisfying*

$$\|Ax - b\|_p^p \leqslant (1 + \varepsilon) \min_{x \in \mathbb{R}^n} \|Ax - b\|_p^p$$

in either $\tilde{O}_p(n^{\frac{2-p}{p+2}})$ iterations and $\tilde{O}_p(\mathsf{nnz}(A) + n^{\max\{\omega, 2+1/3\}})$ time, or $\tilde{O}_p(\sqrt{n})$ iterations and $\tilde{O}_p(\mathsf{nnz}(A) + n^{\omega_0})$ time, with high probability.

It is standard to turn a high accuracy algorithm for an optimization problem into one that solves a corresponding dual problem. We present such an argument for ℓ_p-regression in Section 4.2.2.

Theorem 1.3 (Dual of ℓ_p regression)**.** *There is an algorithm that given $A \in \mathbb{R}^{m \times n}$, $c \in \mathbb{R}^m$, and $q \in [2, \infty)$ computes a $y \in \mathbb{R}^m$ satisfying $A^\top y = c$ and*

$$\|y\|_q^q \leqslant (1 + \varepsilon) \min_{A^\top y = c} \|y\|_q^q$$

in either $\tilde{O}_q(n^{\frac{q-2}{3q-2}})$ *iterations and* $\tilde{O}_q(\mathsf{nnz}(A) + n^{\max\{\omega, 2+1/3\}})$ *time, or* $\tilde{O}_q(\sqrt{n})$ *iterations and* $\tilde{O}_q(\mathsf{nnz}(A) + n^{\omega_0})$ *time, with high probability.*

Prior work [JLS22] shows that ℓ_p-regression can be solved in $\tilde{O}_p(n^{1/3})$ iterations of solving a linear system for $p \in [2, \infty)$. Combining that result with Theorem 1.2 shows that ℓ_p-regression can be solved using $\tilde{O}_p(n^{1/3})$ linear systems for all $p > 1$.

1.3 Discussion of the hypotheses

Let us now discuss various hypotheses and the extent to which they are necessary for sparsifiers of nearly-linear size to exist. In addition to the properties (P1) and (P2), let us consider three others that we will use frequently. In what follows, $C, u > 0$ are positive constants and $h : \mathbb{R}^n \to \mathbb{R}_+$.

(P3) (C-symmetric) $h(z) \leqslant Ch(-z)$ for all $z \in \mathbb{R}^n$.

(P4) (C-monotone) $h(z) \leqslant Ch(\lambda z)$ for $\lambda \geqslant 1$.

(P5) (Upper u-homogeneous) $h(\lambda z) \leqslant C\lambda^u h(z)$ for all $z \in \mathbb{R}^n$ and $\lambda \geqslant 1$.

First, note that (P1) and (P2) imply (P3)–(P5).

Lemma 1.4. *The following implications hold:*

1. *h is L-auto-Lipschitz $\implies$ h is L-symmetric.*
2. *h is lower θ-homogeneous with constant c $\implies$ h is $1/c$-monotone.*
3. *h is L-auto-Lipschitz and C-monotone $\implies$ h is upper 1-homogeneous with constant $2CL$.*

Proof. Since $h(0) = 0$, applying the definition of L-auto-Lipschitz with $z = 0$ gives $h(-z) \leqslant Lh(z)$ for any $z \in \mathbb{R}^n$. The second implication is immediate. For the third, note that for a positive integer k, we have $h(kz) \leqslant \sum_{j=0}^{k-1} |h((j+1)z) - h(jz)| \leqslant kLh(z)$. Using the L-auto-Lipschitz property again gives

$$h(\lambda z) \leqslant h(\lceil \lambda \rceil z) + Lh((\lceil \lambda \rceil - \lambda) z) \leqslant \lceil \lambda \rceil L \cdot h(z) + LCh(z) \leqslant 2CL\lambda h(z),$$

where the penultimate inequality uses C-monotonicity. □

Symmetry (P3). To illustrate the need for approximate symmetry, let us consider gluing together two functions that are otherwise "nice" in our framework:

$$f(z) := \begin{cases} |z|^2, & z \geqslant 0 \\ |z|, & z < 0 . \end{cases}$$

Suppose that $f_1 = \cdots = f_m = f$. Consider unit vectors $\hat{a}_1, \ldots, \hat{a}_m \in \mathbb{R}^n$ such that $\delta_{ij} := |\langle \hat{a}_i, \hat{a}_j \rangle| < \frac{1}{2}$ for $i \neq j$. A basic volume computation shows that one can choose $m \geqslant 2^{\Omega(n)}$. Denote $a_i := (\hat{a}_i, 1) \in \mathbb{R}^{n+1}$ for $i = 1, \ldots, m$.

Then for $\lambda > 0$ and $x := \lambda(\hat{a}_i, -\frac{1}{2})$, we have

$$f_j(\langle a_j, x\rangle) = f(\langle a_j, \lambda(\hat{a}_i, -\tfrac{1}{2})\rangle) = \begin{cases} f(\lambda/2) \asymp \lambda^2 & i = j\,, \\ f(\lambda(\delta_{ij} - \frac{1}{2})) \lesssim \lambda & \text{otherwise.} \end{cases}$$

Thus in any approximate sparsifier $\tilde{F} = w_1 f_1 + \cdots + w_m f_m$, it must be that either $w_i > 0$, or $\sum_{j \neq i} w_j \gtrsim \lambda$. Sending $\lambda \to \infty$ shows that the latter is impossible.

Lower growth and monotonicity (P2), (P4). We consider these properties together since monotonicity is a weaker property than lower homogeneity. A natural function that does not satisfy lower homogeneity is the *Tukey loss* which, for the sake of the present discussion, one can take as $f_i(z) := \min\{1, |z|^2\}$, which is a natural analog of γ_p (recall (1.2)) for $p = 0$.

For sparsifying GLMs with the Tukey loss, previous works have made additional assumptions. For example, that one only ensures sparsification when $\|a_i\|_2 \leqslant n^{O(1)}$, and for inputs $x \in \mathbb{R}^n$ satisfying $\|x\|_2 \leqslant n^{O(1)}$; see [CWW19, Assumption 2] and the discussion afterwards, and [MMWY22, §8.3]. In Section 3.4.2, we show how to achieve a $\tilde{O}(n^{1+o(1)}\varepsilon^{-2})$-sparse ε-approximations under these assumptions. At a high level, the simple idea is to consider the proxy loss functions $\hat{f}_i(z) := \min\{|z|^p, |z|^2\}$ with p sufficiently small.

Upper quadratic growth (P5). Note that, by Lemma 1.4, if $\sqrt{f_i}$ satisfies (P1), then f_i is upper 2-homogeneous. For near-linear size sparsifiers, 2-homogeneity is a natural condition, since sparsifying with loss functions $f_i(x) = |x|^p$ and $p > 2$ requires the sparsifier to have at least $\Omega(n^{p/2})$ terms [BLM89].

The auto-Lipschitz property (P1). As Lemma 1.4 shows, this property gives us approximate symmetry (P3) and upper 1-homogeneity (P5). Crucially, this property also allows us to exploit the geometry of the vectors $a_1, \ldots, a_m \in \mathbb{R}^n$. Note that (P1) implies

$$(f_i(z) - f_i(z'))^2 = (f_i(z)^{1/2} - f_i(z')^{1/2})^2\, (f_i(z)^{1/2} + f_i(z')^{1/2})^2 \leqslant 2L^2 f_i(z - z')(f_i(z) + f_i(z'))\,.$$

In particular, we have

$$\Big(f_i(\langle a_i, x\rangle) - f_i(\langle a_i, y\rangle)\Big)^2 \leqslant 2L^2\, \underbrace{f_i(\langle a_i, x - y\rangle)}\, (f_i(\langle a_i, x\rangle + f_i(\langle a_i, y\rangle)\,.$$

The braced term is what us allows to access the linear structure of the vectors in our analysis.

Comparison to M-estimators. The works [CW15, MMWY22] consider regression and sparsification for what they call general M-estimators. Essentially, this corresponds to the special case of our framework where all the loss functions are the same: $f_1 = \cdots = f_m = M$, and one assumes $M(0) = 0$, monotonicity, and upper and lower growth lower bounds. They additionally assume that M is p-subadditive (for $p = 1/2$) in the sense that $M(x + y)^p \leqslant M(x)^p + M(y)^p$, which is a stronger condition than the auto-Lipschitz property (P1) for $h = f_i^{1/2}$.

Under this stronger set of assumptions, the authors of [MMWY22] achieve approximations with sparsity $\tilde{O}(n^{\max\{2, p/2+1\}})$, which is a factor n larger than what one might hope for. In the regime $p \leqslant 2$ of possible near-linear-sized sparsifiers, we close this gap: Theorem 1.1 gives sparsity $\tilde{O}(n)$.

1.3.1 Discussion of the $s_{\max}/s_{\min}$ dependence

Note that Theorem 1.1 only achieves an approximation for $s_{\min} \leqslant F(x) \leqslant s_{\max}$, and there is a logarithmic dependence on $s_{\max}/s_{\min}$ in the sparsity bound. Intuitively, some dependence on $s_{\max}/s_{\min}$ is necessary in the generality of Theorem 1.1 because nothing in our assumptions precludes the functions f_i from behaving nearly independently on different scales (at least if the scales are sufficiently well separated).

In the case that each of the functions $f_1, \ldots, f_m$ is p-homogeneous, in the sense that $f_i(\lambda z) = |\lambda|^p f_i(z)$, then F and the sparsifier $\tilde{F}$ are both p-homogeneous, and therefore the guarantee $|F(x) - \tilde{F}(x)| \leqslant \varepsilon$ for $F(x) = 1$ already suffices to obtain $|F(x) - \tilde{F}(x)| \leqslant \varepsilon F(x)$ for all $x \in \mathbb{R}^n$, meaning there is no scale dependence.

More generally, for F satisfying the hypotheses of Theorem 1.1, the growth assumptions on $f_1, \ldots, f_m$ allow one to obtain weak guarantees even for $F(x) \notin [s_{\min}, s_{\max}]$. For tamer functions with only a constant number of different scaling regimes, this allows one to avoid the $s_{\max}/s_{\min}$ dependence by applying such scaling arguments and a simple reduction. For the sake of concreteness, we demonstrate this for the Huber loss (the γ_1 function as in (1.2)). A similar argument applies for all the γ_p functionals.

Lemma 1.5. *Consider $a_1, \ldots, a_m \in \mathbb{R}^n$ for $m \geqslant 2$, and $1/m < \varepsilon < 1$. Denote*

$$F(x) := w_1\gamma_1(\langle a_1, x\rangle) + \cdots + w_m\gamma_m(\langle a_m, x\rangle)$$
$$\tilde{F}(x) := \tilde{w}_1\gamma_1(\langle a_1, x\rangle) + \cdots + \tilde{w}_m\gamma_m(\langle a_m, x\rangle)$$

for some nonnegative weights $w, \tilde{w} \in \mathbb{R}^m_+$. Suppose that

$$|F(x) - \tilde{F}(x)| \leqslant \varepsilon F(x) \ \text{ for } x \in \mathbb{R}^n \text{ such that } w_{\min} \leqslant F(x) \leqslant 4m^2 w_{\max}\,,$$

where $w_{\max} := \max(\max(w), \max(\tilde{w}))$ and $w_{\min} := \min(w)$. Then $\tilde{F}$ is a 2ε-approximation to F.

Combining this with an analysis of the weights produced by our construction and the guarantee of Theorem 1.1 yields the following consequence. The proof of Lemma 1.5 and the next result are presented in Section 3.4.1.

Corollary 1.6. *For every $\varepsilon > 0$, the function $F(x) := \gamma_1(\langle a_1, x\rangle) + \cdots + \gamma_1(\langle a_m, x\rangle)$ admits an s-sparse ε-approximation for*

$$s \lesssim \frac{n}{\varepsilon^2}(\log m)\left(\log\left(\frac{n}{\varepsilon}\log m\right)\right)^3.$$

Note that our sparsity bound has an m dependence, as opposed to the classical cases of ℓ_p sparsification, where sparsity bounds depend only on n and ε. However, some m dependence is not surprising, as [MMWY22, §4.5] present vectors $a_1, \ldots, a_m \in \mathbb{R}^n$ for which the sum of the sensitivities (see (1.6)) can grow doubly-logarithmically with m:

$$\sum_{i=1}^{m} \max_{0 \neq x \in \mathbb{R}^n} \frac{\gamma_1(\langle a_i, x\rangle)}{F(x)} \gtrsim n \log\log\frac{m}{n}\,.$$

[MMWY22] also shows that $\sum_{i=1}^m \max_{0\neq x\in\mathbb{R}^n} \frac{\gamma_p(\langle a_i, x\rangle)}{F(x)} \gtrsim n\log\frac{m}{n}$ is possible for $p \in [0, 1)$.

1.4 Importance sampling and multiscale weights

Given $F(x) = f_1(\langle a_1, x\rangle) + \cdots + f_m(\langle a_m, x\rangle)$, our approach to sparsification is via importance sampling. Given a probability vector $\rho \in \mathbb{R}^m$ with $\rho_1, \ldots, \rho_m > 0$ and $\rho_1 + \cdots + \rho_m = 1$, we sample $M \geqslant 1$ coordinates $\nu_1, \ldots, \nu_M$ i.i.d. from ρ, and define our potential approximator by

$$\tilde{F}(x) := \frac{1}{M} \sum_{j=1}^{M} \frac{f_{\nu_j}(\langle a_{\nu_j}, x\rangle)}{\rho_{\nu_j}} .$$

One can easily check that this gives an unbiased estimator for every $x \in \mathbb{R}^n$, i.e., $\mathbb{E}[\tilde{F}(x)] = F(x)$.

Since we want an approximation guarantee to hold simultaneously for many $x \in \mathbb{R}^n$, it is natural to analyze expressions of the form

$$\mathbb{E} \max_{F(x) \leqslant s} \left| F(x) - \tilde{F}(x) \right| .$$

Analysis of this expression involves the size of discretizations of the set $B_F(s) := \{x \in \mathbb{R}^n : F(x) \leqslant s\}$ at various granularities, as explained in Section 1.6.2. The key consideration (via Dudley's entropy inequality, Lemma 1.13) is how well $B_F(s)$ can be covered by cells on which we have uniform control on how much the terms $f_i(\langle a_i, x\rangle)/\rho_i$ vary within each cell.

The ℓ_2 case. Let's consider the case $f_i(z) = |z|^2$ so that $F(x) = |\langle a_1, x\rangle|^2 + \cdots + |\langle a_m, x\rangle|^2$. Here, $B_F(s) = \{x \in \mathbb{R}^n : \|Ax\|_2^2 \leqslant s\}$, where A is the matrix with $a_1, \ldots, a_m$ as rows.

A cell at scale 2^j looks like

$$\mathsf{K}_j := \left\{ x \in \mathbb{R}^n : \max_{i \in [m]} \frac{|\langle a_i, x\rangle|^2}{\rho_i} \leqslant 2^j \right\},$$

and the pertinent question is how many translates of K_j it takes to cover $B_F(s)$. In the ℓ_2 case, this is the well-studied problem of covering Euclidean balls by ℓ_∞ balls.

If N_j denotes the minimum number of such cells required, then the dual-Sudakov inequality (see Lemma 1.12 and Corollary 3.3) tells us that

$$\log N_j \lesssim \frac{s}{2^j} \log(m) \max_{i \in [m]} \frac{\|(A^\top A)^{-1/2} a_i\|_2^2}{\rho_i} .$$

Choosing $\rho_i := \frac{1}{n} \|(A^\top A)^{-1/2} a_i\|_2^2$, i.e., normalized leverage scores, yields uniform control on the size of the coverings:

$$\log N_j \lesssim \frac{s}{2^j} n \log m .$$

The ℓ_p case, $1 \leqslant p < 2$. Consider the case $f_i(z) = |z|^p$ so that $F(x) = \|Ax\|_p^p$. A cell at scale 2^j now looks like

$$\mathsf{K}_j := \left\{ x \in \mathbb{R}^n : \max_{i \in [m]} \frac{|\langle a_i, x\rangle|^p}{\rho_i} \leqslant 2^j \right\},$$

To cover $B_F(s)$ by translates of K_j, we again employ Euclidean balls, and use ℓ_p Lewis weights to relate the ℓ_p structure to an ℓ_2 structure.

A classical result of Lewis [Lew79] (see also [BLM89, CP15]) establishes that there are nonnegative weights $w_1, \ldots, w_m \geqslant 0$ such that if $W = \mathrm{diag}(w_1, \ldots, w_m)$ and $U := (A^\top W A)^{1/2}$, then

$$w_i = \frac{\|U^{-1}a_i\|_2^p}{\|U^{-1}a_i\|_2^2} = \frac{f_i(\|U^{-1}a_i\|_2)}{\|U^{-1}a_i\|_2^2}\,. \tag{1.3}$$

Assuming that A has full rank, a straightforward calculation gives $\sum_{i=1}^m w_i \|U^{-1}a_i\|_2^2 = \mathrm{tr}(U^2 U^{-2}) = n$.

Therefore, we can choose $\rho_i := \frac{1}{n} w_i \|U^{-1}a_i\|_2^2$ for $i = 1, \ldots, m$, and our cells become

$$\mathsf{K}_j := \left\{ x \in \mathbb{R}^n : \max_{i \in [m]} |\langle a_i, x\rangle|^p \leqslant \frac{2^j}{n} w_i \|U^{-1}a_i\|_2^2 \right\}.$$

(Note that the values $\{w_i \|U^{-1}a_i\|_2^2 : i = 1, \ldots, m\}$ are typically referred to as the "ℓ_p Lewis weights".)

If we are trying to use ℓ_2-ℓ_∞ covering bounds, we face an immediate problem: Unlike in the ℓ_2 case, we don't have prior control on $\|Ux\|_2$ for $x \in B_F(s)$. One can obtain an initial bound using the structure of $U = (A^\top W A)^{1/2}$:

$$\begin{aligned}\|Ux\|_2^2 = \sum_{i=1}^m w_i \langle a_i, x\rangle^2 &\overset{(1.3)}{=} \sum_{i=1}^m \|U^{-1}a_i\|_2^{p-2} \langle a_i, x\rangle^2 \\ &= \sum_{i=1}^m \left(\frac{|\langle a_i, x\rangle|}{\|U^{-1}a_i\|_2} \right)^{2-p} |\langle a_i, x\rangle|^p \leqslant \|Ux\|_2^{2-p} \sum_{i=1}^m |\langle a_i, x\rangle|^p\,,\end{aligned} \tag{1.4}$$

where the last inequality is Cauchy-Schwarz: $|\langle a_i, x\rangle| = |\langle U^{-1}a_i, Ux\rangle| \leqslant \|U^{-1}a_i\|_2 \|Ux\|_2$. This gives the bound $\|Ux\|_2 \leqslant \|Ax\|_p \leqslant s^{1/p}$ for $x \in B_F(s)$.

Problematically, this uniform ℓ_2 bound is too weak, but there is a straightforward solution: Suppose we cover $B_F(s)$ by translates of K_{j_0}. This gives an ℓ_∞ bound on the elements of each cell, meaning that we can apply (1.4) and obtain a better upper bound on $\|Ux\|_2$ for $x \in \mathsf{K}_{j_0}$. Thus to cover $B_F(s)$ by translates of K_j with $j < j_0$, we will cover first by translates of K_{j_0}, then cover each translate $(x + \mathsf{K}_{j_0}) \cap B_F(s)$ by translates of $\mathsf{K}_{j_0 - 1}$, and so on.

The standard approach in this setting (see [BLM89] and [LT11, §15.19]) is to instead use interpolation inequalities and duality of covering numbers for a cleaner analytic version of such an iterated covering bound. However, the iterative covering argument can be adapted to the non-homogeneous setting, as we discuss next.

Generalized linear models. When we move to more general loss functions $f_i : \mathbb{R} \to \mathbb{R}$, we lose the homogeneity property $f_i(\lambda x) = \lambda^p f_i(x), \lambda > 0$ that holds for ℓ_p losses. Because of this, we need to replace the single Euclidean structure present in (1.3) (given by the linear operator U) with a family of structures, one for every relevant scale.

Definition 1.7 (Approximate weights)**.** Fix $a_1, \ldots, a_m \in \mathbb{R}^n$ and loss functions $f_1, \ldots, f_m : \mathbb{R} \to \mathbb{R}_+$. We say that a vector $w \in \mathbb{R}_+^m$ is an *α-approximate weight at scale s* if

$$\frac{s}{\alpha} \leqslant \frac{f_i(\|M_w^{-1/2}a_i\|_2)}{w_i \|M_w^{-1/2}a_i\|_2^2} \leqslant \alpha s\,, \quad i = 1, \ldots, m\,, \text{ where } M_w := \sum_{j=1}^m w_j a_j a_j^\top\,. \tag{1.5}$$

To motivate this definition, let us define scale-specific sentivities:

$$\xi_i(s) := \max\left\{\frac{f_i(\langle a_i, x\rangle)}{F(x)} : x \in \mathbb{R}^n, F(x) \in [s/2, s]\right\}, \quad i = 1, \dots, m\,. \tag{1.6}$$

As shown in Corollary 2.3, if the functions $\{f_i\}$ are lower θ-homogeneous, upper 2-homogeneous, and $O(1)$-symmetric (in the sense of (P3)), then an α-approximate weight at scale s allows us to upper bound sensitivies by leverage scores:

$$\xi_i(s) \lesssim \sigma_i(W^{1/2}A)\,, \tag{1.7}$$

where $W = \mathrm{diag}(w_1, \dots, w_m)$, and the implicit constant depends on α and the homogeneity parameters. Here, $\sigma_i(V)$ denotes the ith leverage score of a matrix V with rows $v_1, \dots, v_m$:

$$\sigma_i(V) := \langle v_i, (V^\top V)^+ v_i\rangle\,,$$

where $(V^\top V)^+$ denotes the Moore-Penrose pseudoinverse. Notably, one always has $\sigma_1(V) + \cdots + \sigma_m(V) = \mathrm{rank}(V)$, and therefore (1.7) gives an upper bound $\xi_1(s) + \cdots + \xi_m(s) \lesssim n$.

In order to generalize the iterated covering argument for ℓ_p losses, we need there to be a relationship between weights at different scales.

Definition 1.8 (Weight schemes)**.** Let $\mathcal{J} \subseteq \mathbb{Z}$ be a contiguous interval. A family $\{w^{(j)} \in \mathbb{R}^m_+ : j \in \mathcal{J}\}$ is an *α-approximate weight scheme* if each $w^{(j)}$ is an α-approximate weight at scale 2^j and, furthermore, for every pair $j, j+1 \in \mathcal{J}$ and $i \in \{1, \dots, m\}$,

$$w_i^{(j+1)} \leqslant \alpha w_i^{(j)}\,. \tag{1.8}$$

Given a weight scheme, we choose sampling probabilities

$$\rho_i \propto \max_{j \in \mathcal{J}} w_i^{(j)} \|M_{w^{(j)}}^{-1/2} a_i\|_2^2 = \max_{j \in \mathcal{J}} \sigma_i(W_j^{1/2}A)\,, \quad i = 1, \dots, m\,,$$

where $W_j = \mathrm{diag}(w_1^{(j)}, \dots, w_m^{(j)})$. In our setting, $|\mathcal{J}| \leqslant O(\log(m s_{\max}/s_{\min}))$, which results in the sparsity increasing by a corresponding factor.

In Section 2, we establish the existence of approximate weight schemes for general families of loss functions satisfying certain growth bounds, along with efficient algorithms to compute the corresponding weights.

1.5 Regression via iterative refinement

Previous works have observed that combining *iterative refinement* with sparsification of γ_p-functions (recall (1.2)) leads to improved algorithms for ℓ_p-regression [AKPS19b, ABKS21, GPV21]. For the benefit of the reader, we give a description of these ideas in somewhat more generality.

Recall that our goal is to find a point $x \in \mathbb{R}^n$ that computes an approximate minimizer of $F(x) := \sum_{i=1}^m f_i(\langle a_i, x\rangle - b_i)$, up to high accuracy. For now, we assume that F is a differentiable convex function and denote $F_* := \inf_{x \in \mathbb{R}^n} F(x)$. Later, we will introduce additional conditions that allow for iterative refinement to succeed.

Broadly, iterative refinement minimizes $F(x)$ be repeatedly solving sub-problems, each of which make multiplicative progress in reducing the error of the current solution. Given a current point x_0, prior works on iterative refinement define a local approximation of F suitably symmetrized and centered around x_0 such that approximately minimizing this local approximation yields the desired decrease in function error. One way to derive such local approximations is through Bregman divergences, which give a natural way of recentering convex functions.

Definition 1.9 (The F-divergence)**.** For $x, y \in \mathbb{R}^n$, use $T_x^F(y) := F(x) + \nabla F(x)^\top (y - x)$ to denote the first order Taylor approximation of F at x, and define the *F-induced Bregman divergence* by $D_x^F(y) := F(y) - T_x^F(y)$.

Note that for convex F, the function $D_x^F(y)$ is convex and minimized at x. Consequently, given a point x_0, the function $D_{x_0}^F(y)$ is a natural function induced by F and minimized at x_0. Note that minimizing $F(x)$ is the same as minimizing $\langle \nabla F(x_0), x - x_0 \rangle + D_{x_0}^F(x)$ which in turn is the same as minimizing $\langle \nabla F(x), \Delta \rangle + D_{x_0}^F(x_0 + \Delta)$ over Δ and adding the minimizer to x_0.

Iterative refinement strategies approximately minimize $\langle \nabla F(x), \Delta \rangle + r(\Delta)$, where $r(\Delta)$ is a suitable approximation of $D_{x_0}^F(x_0 + \Delta)$. One step of refinement moves to $x_1 := x_0 + \eta \tilde{\Delta}$, where η is a suitably chosen step-size and $\tilde{\Delta}$ is the approximate minimizer. As motivation for our approach, here we consider the scheme suggested by prior work, where $r(\Delta)$ is a sparsification of a simple approximation to the divergence.

Informally, one can show that if the square root of the Bregman divergence of each f_i is L-auto-Lipschitz (P1) and lower θ-homogeneous for some $\theta > 1$ (P2), then one step of sparsification/refinement decreases the error in the objective value multiplicatively by an absolute constant. Interestingly, auto-Lipschitzness is only required for sparsification and not for refinement. However, we critically need the Bregman divergence to be lower θ-homogeneous for $\theta > 1$ for iterative refinement, while our sparsification results (Theorem 1.1) only require $\theta > 0$.

Lemma 1.10 (Refinement Lemma)**.** *Suppose $r : \mathbb{R}^n \to \mathbb{R}_+$ is lower θ-homogeneous with constant $c < 1$ for $\theta > 1$, and for $x_0 \in \mathbb{R}^n$ and all $\Delta \in \mathbb{R}^n$ and $\eta \in [0, 1]$,*

$$r(\Delta) \leqslant D_{x_0}^F(x_0 + \Delta) \leqslant \alpha\, r(\Delta)$$

where $\alpha \geqslant 1$ is fixed. Then $\inf_{\Delta \in \mathbb{R}^n} \left\{ T_{x_0}^F(x_0 + \hat{\Delta}) + r(\hat{\Delta}) \right\} \leqslant F_*$ and if $\hat{\Delta} \in \mathbb{R}^n$ satisfies

$$T_{x_0}^F(x_0 + \hat{\Delta}) + r(\hat{\Delta}) \leqslant F_* \,, \tag{1.9}$$

then

$$F(x_0 + \hat{\eta}\hat{\Delta}) - F_* \leqslant (1 - \hat{\eta})\,(f(x_0) - F_*)\,, \quad \text{where } \hat{\eta} := (\alpha/c)^{-1/(q-1)}\,. \tag{1.10}$$

Proof. First note that, for all $\Delta \in \mathbb{R}^n$,

$$F(x_0 + \Delta) = T_{x_0}^F(x_0 + \Delta) + D_{x_0}^F(x_0 + \Delta)\,. \tag{1.11}$$

Since, $D_{x_0}^F(x_0 + \Delta) \geqslant r(\Delta)$ this implies the desired bound

$$\inf_{\Delta \in \mathbb{R}^n} \left\{ T_{x_0}^F(x_0 + \Delta) + r(\Delta) \right\} \leqslant \inf_{\Delta \in \mathbb{R}^n} F(x_0 + \Delta) = F_*\,.$$

Next, note that for all $\eta \in [0,1]$ and $\Delta \in \mathbb{R}^n$,

$$T_{x_0}^F(x_0 + \eta\Delta) = F(x_0) + \eta \nabla F(x_0)^\top \Delta = (1-\eta)F(x_0) + \eta\, T_{x_0}^F(x_0 + \Delta) \tag{1.12}$$

$$D_{x_0}^F(x_0 + \eta\Delta) \leqslant \alpha r(\eta\Delta) \leqslant \alpha/c \cdot \eta^\theta r(\Delta) \tag{1.13}$$

Suppose that $\hat{\Delta} \in \mathbb{R}^n$ satisfies (1.9). Then plugging (1.12) and (1.13) into (1.11) with $\Delta = \hat{\eta}\hat{\Delta}$ yields

$$\begin{aligned} F(x_0 + \hat{\eta}\hat{\Delta}) &\leqslant (1-\hat{\eta})F(x_0) + \hat{\eta}\, T_{x_0}^F(x_0 + \hat{\Delta}) + \alpha/c \cdot \hat{\eta}^\theta r(\hat{\Delta}) \\ &= (1-\hat{\eta})F(x_0) + \hat{\eta}\left(T_{x_0}^F(x_0 + \hat{\Delta}) + r(\hat{\Delta})\right) \leqslant (1-\hat{\eta})F(x_0) + \hat{\eta} F_* \end{aligned}$$

where we used that $\alpha/c \cdot \hat{\eta}^{\theta-1} = 1$ and $\hat{\eta} \in [0,1]$. Rearranging yields (1.10). □

To apply Lemma 1.10 to ERM for general linear models, note that

$$D_{x_0}^F(x_0 + \Delta) = \sum_{i=1}^m D_{\langle a_i, x_0 \rangle - b_i}^{f_i}(\langle a_i, \Delta \rangle)\,.$$

Now, if the square root of each divergence $D_z^{f_i}$ is L-auto-Lipschitz and lower θ-homogeneous, then Theorem 1.1 gives weights $w \in \mathbb{R}_+^m$ with sparsity $\tilde{O}(n)$ such that

$$0.9 \cdot D_{x_0}^F(x_0 + \Delta) \leqslant r(\Delta) \leqslant D_{x_0}^F(x_0 + \Delta) \quad \text{where} \quad r(\Delta) := \sum_{i=1}^m w_i D_{\langle a_i, x_0 \rangle - b_i}^{f_i}(\langle a_i, \Delta \rangle).$$

Thus, Lemma 1.10 applies, and we can decrease the objective value error by a multiplicative factor by minimizing $r(\Delta)$. Since $r(\Delta)$ only has $\tilde{O}(n)$ nonzero terms, one can apply previous solvers for $m = \tilde{O}(n)$ [LSZ19a, AKPS19b] to obtain the desired runtimes.

In Section 4.2, we verify that the f-divergence of $f(z) = |z|^p$ is the γ_p function, which is lower p-homogeneous, and has an auto-Lipschitz square root. This yields our algorithm for ℓ_p regression (Theorem 1.2). A formal version of the argument is presented in Section 4. The main difference is that some technical work is needed because Theorem 1.1 only provides sparsification for a range of inputs $\{x \in \mathbb{R}^n : s_{\min} \leqslant F(x) \leqslant s_{\max}\}$. Moreover, our algorithm uses an approximate oracle for GLMs (Definition 4.2) with the functions $\{f_i\}$ (rather than with the approximate divergence $r(\cdot)$), and we show that the oracle does not need to be solved to high accuracy to make sufficient progress.

1.6 Preliminaries

Throughout the paper, we denote $[n] := \{1, 2, \ldots, n\}$. We use the notation $a \lesssim b$ to denote that there is a universal constant C such that $a \leqslant Cb$, and $a \lesssim_L b$ to denote that C may depend on L. We use $a \asymp b$ to denote the conjunction of $a \lesssim b$ and $b \lesssim a$, and $a \asymp_L b$ analogously. We also denote $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geqslant 0\}$ and $\mathbb{R}_{++} := \{x \in \mathbb{R} : x > 0\}$.

For simplicity of presentation, we assume that the vectors $a_1, \ldots, a_m \in \mathbb{R}^n$ in (1.1) span $\mathbb{R}^n$ and all are nonzero. In particular, this means that the matrix A with rows $a_1, \ldots, a_m$ has rank n and $A^\top W A$ is invertible for any diagonal matrix W with strictly positive entries on the diagonal.

1.6.1 Covering numbers and chaining

Consider a metric space (T, d). For $x \in T$ and $r > 0$, define the ball $B(x,r) := \{y \in T : d(x,y) \leqslant r\}$.

Definition 1.11 (Covering numbers)**.** For a radius $r > 0$, we define the *covering number* $\mathcal{N}(T, d, r)$ as the smallest number of balls of radius r (in the distance d) that are required to cover T. For $S, S' \subseteq \mathbb{R}^n$, we overload notation and use $\mathcal{N}(S, S')$ to denote the smallest number of translates of S' needed to cover S.

We require the following "dual Sudakov inequality" (see [PT85] and [LT11, (3.15)]) which gives bounds for covering the Euclidean ball using balls in an arbitrary norm.

Lemma 1.12 (Dual Sudakov inequality)**.** *Let B_2^n denote the unit ball in n dimensions, and $\|\cdot\|_X$ an arbitrary norm on $\mathbb{R}^n$. If $\boldsymbol{g}$ is a standard n-dimensional Gaussian, then*

$$\sqrt{\log \mathcal{N}(B_2^n, B_X)} \lesssim \mathbb{E}\, \|\boldsymbol{g}\|_X\,,$$

where $B_X := \{y \in \mathbb{R}^n : \|y\|_X \leqslant 1\}$.

We recall Talagrand's generic chaining functional [Tal14, Def. 2.2.19]:

$$\gamma_2(T, d) := \inf_{\{\mathcal{A}_h\}} \sup_{x \in T} \sum_{h=0}^{\infty} 2^{h/2} \mathrm{diam}(\mathcal{A}_h(x), d)\,,$$

where the infimum runs over all sequences $\{\mathcal{A}_h : h \geqslant 0\}$ of partitions of T satisfying $|\mathcal{A}_h| \leqslant 2^{2^h}$ for each $h \geqslant 0$. Note that we use the notation $\mathcal{A}_h(x)$ for the unique set of $\mathcal{A}_h$ that contains x.

The chaining functional is used to control the maximum of subgaussian processes (see, e.g., the discussion in [JLLS23, §2.2] where it is applied precisely in the setting of sparsification). Our use of the functional occurs only in the statement of Lemma 1.14 below, and in this paper we will only require the following classical upper bound. (See, eg., [Tal14, Prop 2.2.10].)

Lemma 1.13 (Dudley's entropy bound)**.** *For any metric space (T, d), it holds that*

$$\gamma_2(T, d) \lesssim \sum_{j \in \mathbb{Z}} 2^j \sqrt{\log \mathcal{N}(T, d, 2^j)}\,.$$

The interested reader will note that this is (up to constants) precisely the upper bound one obtains by choosing $\mathcal{A}_h$ as a uniform discretization of (T, d), i.e., to minimize $\sup\{\mathrm{diam}(\mathcal{A}_h(x), d) : x \in T\}$ over all partitions satisfying $|\mathcal{A}_h| \leqslant 2^{2^h}$.

1.6.2 Sparsification via subgaussian processes

We discuss sparsification via subgaussian processes. Consider $\varphi_1, \varphi_2, \ldots, \varphi_m : \mathbb{R}^n \to \mathbb{R}$, and define

$$F(x) := \sum_{j=1}^{m} \varphi_j(x)\,.$$

Given a strictly positive probability vector $\rho \in \mathbb{R}^m_{++}$, and an integer $s \geqslant 1$ and $\nu = (\nu_1, \ldots, \nu_s) \in [m]^s$, define the distance

$$d_{\rho,\nu}(x,y) := \left(\sum_{j=1}^{s} \left(\frac{\varphi_{\nu_j}(x) - \varphi_{\nu_j}(y)}{\rho_{\nu_j}\, s} \right)^2 \right)^{1/2}.$$

and the function $\tilde{F}_{\rho,\nu} : \mathbb{R}^n \to \mathbb{R}$

$$\tilde{F}_{\rho,\nu}(x) := \frac{1}{s} \sum_{j=1}^{s} \frac{\varphi_{\nu_j}(x)}{\rho_{\nu_j}}.$$

We require the following lemma which employs a variant of a standard symmetrization argument to control $\mathbb{E} \max_{x \in \Omega} |F(x) - \tilde{F}_{\rho,\boldsymbol{\nu}}(x)|$ using an associated Bernoulli process (see, for example, [Tal14, Lem 9.1.11]). For a subset $\Omega \subseteq \mathbb{R}^n$, denote $\|F\|_{C(\Omega)} := \sup_{x \in \Omega} |F(x)|$. The reader will note our typical application of the lemma to sets of the form $\Omega = \{x \in \mathbb{R}^n : F(x) \leqslant \lambda\}$ for some parameter $\lambda > 0$.

Lemma 1.14 ([JLLS23, Lemma 2.6]). *Consider $s \geqslant 1$, a subset $\Omega \subseteq \mathbb{R}^n$, and a probability vector $\rho \in \mathbb{R}^m_+$. Assume that*

$$\exists x_0 \in \Omega \quad \text{s.t.} \quad \varphi_1(x_0) = \cdots = \varphi_m(x_0) = 0.$$

Suppose, further, that for some $0 < \delta \leqslant 1$, and every $\nu \in [m]^s$, it holds that

$$\gamma_2(\Omega, d_{\rho,\nu}) \leqslant \delta \left(\|F\|_{C(\Omega)} \|\tilde{F}_{\rho,\nu}\|_{C(\Omega)} \right)^{1/2}.$$

If $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_s$ are sampled independently from ρ, then

$$\mathbb{E} \max_{x \in \Omega} \left| F(x) - \tilde{F}_{\rho,\boldsymbol{\nu}}(x) \right| \lesssim \mathbb{E} \left[\gamma_2(\Omega, d_{\rho,\boldsymbol{\nu}}) \right] \leqslant 8\delta \|F\|_{C(\Omega)}.$$

If it also holds that, for all $\nu \in [m]^s$,

$$\operatorname{diam}(\Omega, d_{\rho,\nu}) \leqslant \hat{\delta} \left(\|F\|_{C(\Omega)} \|\tilde{F}_{\rho,\nu}\|_{C(\Omega)} \right)^{1/2},$$

then there is a universal constant $K > 0$ such that for all $0 \leqslant t \leqslant \frac{1}{2K\hat{\delta}}$,

$$\mathbb{P} \left(\max_{x \in \Omega} \left| F(x) - \tilde{F}_{\rho,\boldsymbol{\nu}}(x) \right| > K(\delta + t\hat{\delta}) \|F\|_{C(\Omega)} \right) \leqslant e^{-Kt^2/4}.$$

2 Multiscale importance scores

Recall the definitions of approximate weights (Definition 1.7) and weight schemes (Definition 1.8). In the present section, we prove the following two results.

Theorem 2.1. *Suppose that $f_1, \ldots, f_m : \mathbb{R} \to \mathbb{R}_+$ are lower θ-homogeneous and upper u-homogeneous with $u > \theta > 0$ and uniform constants $c, C > 0$. Then there is some $\alpha = \alpha(\theta, c, u, C) > 1$ such that for every choice of vectors $a_1, \ldots, a_m \in \mathbb{R}^n$ and $s > 0$, there is an α-approximate weight at scale s.*

This is proved in Section 2.2 by considering critical points of the functional $U \mapsto \det(U)$ subject to the constraint $G(U) \leqslant s$, where $G(U) := f_1(\|Ua_1\|_2) + \cdots + f_m(\|Ua_m\|_2)$, which can be seen as a generalization of Lewis' original method.

Single-scale sensitivities. Let us now observe that, in the case $u \leqslant 2$, Theorem 2.1 allows us to bound sensitivities (recall (1.6)). The next lemma is a generalization of (1.4).

Lemma 2.2. *Suppose $f_1, \ldots, f_m : \mathbb{R} \to \mathbb{R}_+$ satisfy the assumptions of Theorem 2.1 with $u \leqslant 2$, and they are additionally K-symmetric in the sense of (P3). If $w \in \mathbb{R}^m_+$ is an α-approximate weight at scale s, then for any $x \in \mathbb{R}^n$, it holds that*

$$\|M_w^{1/2} x\|_2^\theta \leqslant \max\left(1, \alpha \frac{CK}{c} \frac{F(x)}{s}\right).$$

Proof. We may clearly assume that $\|M_w^{1/2} x\|_2 \geqslant 1$. Then using the Cauchy-Schwarz inequality $|\langle a_i, x\rangle| \leqslant \|M_w^{-1/2} a_i\|_2 \|M_w^{1/2} x\|_2$ together with the upper quadratic growth assumption gives

$$\frac{f_i(|\langle a_i, x\rangle|)}{|\langle a_i, x\rangle|^2} \geqslant \frac{1}{C} \frac{f_i(\|M_w^{1/2} x\|_2 \|M_w^{-1/2} a_i\|_2)}{\|M_w^{1/2} x\|_2^2 \|M_w^{-1/2} a_i\|_2^2} \geqslant \frac{c}{C} \frac{\|M_w^{1/2} x\|_2^\theta f_i(\|M_w^{-1/2} a_i\|_2)}{\|M_w^{1/2} x\|_2^2 \|M_w^{-1/2} a_i\|_2^2}, \tag{2.1}$$

where the last inequality uses the lower growth assumption.

Using $M_w = \sum_{i=1}^m w_i a_i a_i^\top$, we can bound

$$\begin{aligned} \|M_w^{1/2} x\|_2^2 = \sum_{i=1}^m w_i \langle a_i, x\rangle^2 &\overset{(2.1)}{\leqslant} \frac{C}{c} \|M_w^{1/2} x\|_2^{2-\theta} \sum_{i=1}^m w_i f_i(|\langle a_i, x\rangle|) \frac{\|M_w^{-1/2} a_i\|_2^2}{f_i(\|M_w^{-1/2} a_i\|_2)} \\ &\leqslant \alpha \frac{CK}{c} \frac{1}{s} \|M_w^{1/2} x\|_2^{2-\theta} \sum_{i=1}^m f_i(\langle a_i, x\rangle), \end{aligned}$$

where the last inequality uses the defining property of an α-approximate weight at scale s, along with the assumption of K-symmetry: $f_i(|\langle a_i, x\rangle|) \leqslant K f_i(\langle a_i, x\rangle)$ for all $i = 1, \ldots, m$. □

Corollary 2.3 (Sensitivity upper bound)**.** *Under the assumptions of Lemma 2.2, it holds that*

$$\xi_1(s) + \cdots + \xi_m(s) \lesssim n,$$

where the implicit constant depends on the parameters α, θ, C, c, K.

Proof. From Lemma 2.2, if $F(x) \leqslant s$, then $\|M_w^{1/2} x\|_2 \leqslant (\alpha CK/c)^{1/\theta}$. By Cauchy-Schwarz, this gives $|\langle a_i, x\rangle| \leqslant (\alpha CK/c)^{1/\theta} \|M_w^{-1/2} a_i\|_2$, and therefore

$$\begin{aligned} f_i(\langle a_i, x\rangle) \leqslant K f_i(|\langle a_i, x\rangle|) &\leqslant \frac{K}{c} f_i\left((\alpha CK/c)^{1/\theta} \|M_w^{-1/2} a_i\|_2\right) \\ &\leqslant \frac{CK}{c} \left(\frac{\alpha CK}{c}\right)^{2/\theta} f_i(\|M_w^{-1/2} a_i\|_2) \leqslant \frac{\alpha CK}{c} \left(\frac{\alpha CK}{c}\right)^{2/\theta} s \cdot w_i \|M_w^{-1/2} a_i\|_2^2, \end{aligned}$$

where the first inequality uses K-symmetry, the second uses $1/c$-monotonicity (which holds by Lemma 1.4), the third inequality uses upper homogeneity, and the last inequality uses that w is an α-approximate weight at scale s. Finally, one notes that $\sum_{i=1}^m w_i \|M_w^{-1/2} a_i\|_2^2 = \operatorname{tr}\left(M_w^{-1} M_w\right) = n$, completing the proof. □

We are only able to establish the existence of entire weight schemes (where the weights at adjacent scales are related) for $u < 4$, which suffices our applications, as $u \leqslant 2$ is a requirement for Theorem 1.1. The following theorem is proved in Section 2.1, based on the contractive iteration method introduced by Cohen and Peng [CP15].

Theorem 2.4. *Suppose that $f_1, \ldots, f_m : \mathbb{R} \to \mathbb{R}_+$ are lower θ-homogeneous and upper u-homogeneous with $4 > u > \theta > 0$ and uniform constants $c, C > 0$. Then there is some $\alpha = \alpha(u, C, \theta, c)$ such that for every choice of vectors $a_1, \ldots, a_m \in \mathbb{R}^n$, there is an α-approximate weight scheme $\{w_i^{(j)} : j \in \mathbb{Z}\}$.*

In the next section, we show how to compute an approximate weight scheme $\{w_i^{(j)} : j \in \mathcal{J}\}$ using $\tilde{O}(|\mathcal{J}|)$ computations of leverage scores $(\sigma_1(V), \ldots, \sigma_m(V))$ for matrices of the form $V = A^\top W A$.

2.1 Contractive algorithm

For a weight $w \in \mathbb{R}_+^m$ and $i \in \{1, \ldots, m\}$, define

$$\tau_i(w) := \frac{\sigma_i(W^{1/2}A)}{w_i} = \langle a_i, (A^\top W A)^{-1} a_i \rangle\,, \quad W := \mathrm{diag}(w_1, \ldots, w_m)\,,$$

and denote $\tau(w) := (\tau_1(w), \ldots, \tau_m(w))$.

Fix a scale parameter $s > 0$ and define the iteration $\varphi_s : \mathbb{R}_+^m \to \mathbb{R}_+^m$ by

$$(\varphi_s(w))_i := \frac{1}{s} \frac{f_i(\sqrt{\tau_i(w)})}{\tau_i(w)}\,. \tag{2.2}$$

Write $\varphi^k := \varphi \circ \cdots \circ \varphi$ for the k-fold composition of φ. In this case where $f_i(z) = |z|^p$ and $1 \leqslant p \leqslant 2$, it is known, for $s = 1$, starting from any $w_0 \in \mathbb{R}_+^m$, the sequence $\{\varphi_1^k(w_0) : k \geqslant 1\}$ converges to the unique fixed point of φ, which are the corresponding ℓ_p Lewis weights (1.3).

Define now a metric d on $\mathbb{R}_+^m$ by

$$d(u, w) := \max\left\{\left|\log \frac{u_i}{w_i}\right| : i = 1, \ldots, m\right\}.$$

We note the following characterization.

Fact 2.5. *A vector $w \in \mathbb{R}_+^m$ is an α-approximate weight at scale s if and only if*

$$d(w, \varphi_s(w)) \leqslant \log \alpha\,.$$

First, we observe that τ is 1-Lipschitz on $(\mathbb{R}_+^m, d)$. In the next proof, $\preceq$ denotes the ordering of two real, symmetric matrices in the Loewner order, i.e., $A \preceq B$ if and only if $B - A$ is positive semi-definite.

Lemma 2.6. *For any $w, w' \in \mathbb{R}_+^m$, it holds that $d(\tau(w), \tau(w')) \leqslant d(w, w')$.*

Proof. Denote $W = \mathrm{diag}(w), W' = \mathrm{diag}(w')$, and $\alpha := \exp(d(w, w'))$. Then $\alpha^{-1} W \preceq W' \preceq \alpha W$, therefore $\alpha^{-1} A^\top W A \preceq A^\top W' A \preceq \alpha A^\top W A$, and by monotonicity of the matrix inverse in the Loewner order, $\alpha^{-1}(A^\top W A)^{-1} \preceq (A^\top W' A)^{-1} \preceq \alpha (A^\top W A)^{-1}$. This implies $d(\tau(w), \tau(w')) \leqslant \log \alpha$, completing the proof. □

Proof of Theorem 2.4. Consider the map $\psi : \mathbb{R}_+^m \to \mathbb{R}_+^m$ whose i-th coordinate is defined as

$$\psi_i(x) := \frac{f_i(\sqrt{x_i})}{x_i}\,.$$

Our assumptions on lower and upper-homogeneity give, for all $y_i \geqslant x_i$,

$$c\left(\frac{y_i}{x_i}\right)^{\theta/2-1} \leqslant \frac{f_i(\sqrt{y_i})/y_i}{f_i(\sqrt{x_i})/x_i} \leqslant C\left(\frac{y_i}{x_i}\right)^{u/2-1},$$

yielding, for $C_1 := \max\{C, 1/c\}$,

$$d(\psi(x),\psi(y)) \leqslant \max\left(\left|\frac{\theta}{2}-1\right|,\left|\frac{u}{2}-1\right|\right) d(x,y) + \log(C_1)\,. \tag{2.3}$$

Fix $s > 0$ and consider the mapping $\varphi : \mathbb{R}_+^m \to \mathbb{R}_+^m$ defined in (2.2). Then for $u < 4$ and $\delta := \max\left(\left|\frac{\theta}{2}-1\right|,\left|\frac{u}{2}-1\right|\right) < 1$, (2.3) in conjunction with Lemma 2.6, shows that

$$d(\varphi_s(w),\varphi_s(w')) < \delta\, d(w,w') + \log(C_1)\,. \tag{2.4}$$

Applying this bound inductively, for any weight $w \in \mathbb{R}_+^m$ and $k \geqslant 1$, we have

$$d\left(\varphi_s^k(w),\varphi_s^{k+1}(w)\right) \leqslant \frac{\delta^k d(\varphi_s(w),w) + \log C_1}{1-\delta}\,, \tag{2.5}$$

Now define

$$w^{(0)} := \varphi_1^k(1,\ldots,1)\,,$$

where $k \geqslant 1$ is chosen large enough so that $d(w^{(0)},\varphi_1(w^{(0)})) \leqslant \frac{2\log C_1}{1-\delta}$. From Fact 2.5, one sees that $w^{(0)}$ is an α-approximate weight at scale 1 for $\alpha = C_1^{2/(1-\delta)}$.

Define inductively, for $j = 1,2,\ldots$,

$$\begin{aligned} w^{(j)} &:= \varphi_{2^j}(w^{(j-1)}) \\ w^{(-j)} &:= \varphi_{2^{-j}}(w^{(1-j)})\,. \end{aligned}$$

Note that

$$\begin{aligned} d(\varphi_{2^j}(w^{(j)}),w^{(j)}) &= d(\varphi_{2^j}^2(w^{(j-1)}),\varphi_{2^j}(w^{(j-1)})) \\ &\leqslant \delta d(\varphi_{2^j}(w^{(j-1)}),w^{(j-1)}) + \log(C_1) \\ &\leqslant \delta d(\varphi_{2^{j-1}}(w^{(j-1)}),w^{(j-1)}) + \delta\log(2) + \log(C_1)\,, \end{aligned}$$

where the last inequality uses $\varphi_{2s}(w) = 2\varphi_s(w)$ for all $w \in \mathbb{R}_+^m$.

Therefore, by induction, $d(\varphi_{2^j}(w^{(j)}),w^{(j)}) \leqslant \frac{2\log(C_1)+\log 2}{1-\delta}$ for all $j > 0$. To see that the family of weights $\{w^{(j)} : j \in \mathbb{Z}\}$ forms a weight scheme, note that

$$d(w^{(j)},w^{(j-1)}) = d(\varphi_{2^j}(w^{(j-1)}),w^{(j-1)}) \leqslant d(\varphi_{2^j}(w^{(j-1)}),w^{(j-1)}) + \log 2\,,$$

thus $\{w^{(j)} : j \in \mathbb{Z}\}$ is an α-approximate weight scheme for $\alpha = \frac{2\log(2C_1)}{1-\delta}$, completing the proof. $\square$

2.1.1 Efficient implementation of the iteration

In this section we give an efficient algorithm for implementing the iteration (2.2). The primary difficult is that we need to exhibit convergence even when the iterates are only computed approximately. For convenience, we use the notation $x \approx_\alpha y$ to denote that $\alpha^{-1} \leqslant |x/y| \leqslant \alpha$.

Theorem 2.7 (Algorithm for weight construction)**.** *Algorithm 2 takes as input functions $f_1, \ldots, f_m : \mathbb{R} \to \mathbb{R}_+$ are lower θ-homogeneous with constant c (P2) and upper u-homogeneous with constant C (P5), for some $\theta > 0$ and $u < 4$, vectors $a_1, \ldots, a_m \in \mathbb{R}^n$, integers $j_{\min} < j_{\max}$, and $w^\circ \in \mathbb{R}_+^m$ that satisfies*

$$d(\varphi_{2^{j_{\max}}}(w^\circ), w^\circ) \leqslant \beta. \tag{2.6}$$

For some $\alpha = \alpha(\theta, u, c, C)$, the algorithm returns an α-approximate weight scheme $\{w_i^{(j)} : j \in \mathcal{J}\}$ with $\mathcal{J} = \mathbb{Z} \cap [j_{\min}, j_{\max}]$, and succeeds with high probability in time

$$\tilde{O}_{\theta,u,c,C}\left((\mathsf{nnz}(a_1, \ldots, a_m) + n^\omega + m\mathcal{T}_{\mathrm{eval}})\left(|\mathcal{J}| + \log\max\{\beta, 1\}\right)\right) .$$

The algorithm proceeds along an iterative procedure akin to (2.2). At each step, the weights are updated by computing approximate leverage scores of a matrix $A^\top W A$.

Theorem 2.8 (Leverage score approximation, [SS11, LMP13, CLM⁺15])**.** *There is an algorithm* LevApprox(A, W, ε) *that takes a matrix $A \in \mathbb{R}^{m\times n}$, a non-negative diagonal matrix W, and $\varepsilon > 0$, and produces $(1+\varepsilon)$-approximations $\widetilde{\sigma}_i \approx_{1+\varepsilon} w_i a_i^\top (A^\top W A)^{-1} a_i$ for all $i \in [m]$, in $\tilde{O}(\varepsilon^{-2}(\mathsf{nnz}(A) + n^\omega)\log(1/\delta))$ time, with probability at least $1-\delta$.*

Algorithm 1: Iterate$(\{f_1, \ldots, f_m\}, \{a_1, \ldots, a_m\}, w, s, \varepsilon)$

---

input: Functions $f_1, \ldots, f_m : \mathbb{R} \to \mathbb{R}_+$, vectors $a_1, \ldots, a_m \in \mathbb{R}^n$, weights $w \in \mathbb{R}_+^m$, a scale $s \in \mathbb{R}_+$, and $\varepsilon \in \mathbb{R}_+$.

$\widetilde{\sigma} \leftarrow$ LevApprox(A, W, ε).

$\overline{w}_i \leftarrow \frac{1}{s}\frac{f_i(\sqrt{\widetilde{\sigma}_i/w_i})}{\widetilde{\sigma}_i/w_i}$ for $i = 1, \ldots, m$.

return $\overline{w}$.

Algorithm 1 called with $\varepsilon = 0$ is able to directly implement the iteration (2.2). We now show that Iterate remains approximately contracting for $\varepsilon > 0$.

Let $\varphi_1, \ldots, \varphi_m$ and $a_1, \ldots, a_m \in \mathbb{R}^n$ be given as in Theorem 2.7, and define the function $\tilde{\varphi}_{s,\varepsilon} : \mathbb{R}_+^m \to \mathbb{R}_+^m$ by

$$\tilde{\varphi}_{s,\varepsilon}(w) := \frac{1}{s}\frac{f_i(\sqrt{\tilde{\sigma}_i/w_i})}{\tilde{\sigma}_i/w_i}\,, \text{ where } \tilde{\sigma}_i \approx_{1+\varepsilon} w_i a_i^\top (A^\top W A)^{-1} a_i \text{ for all } i = 1, \ldots, m\,, \tag{2.7}$$

i.e., $\tilde{\sigma}_i$ are arbitrary approximate leverage scores.

Lemma 2.9. *For any $w, w' \in \mathbb{R}_+^m$ and $0 < \varepsilon < 1/3$, it holds that*

$$d(\tilde{\varphi}_{s,\varepsilon}(w), \tilde{\varphi}_{s,\varepsilon}(w')) \leqslant \delta d(w, w') + \log(2C_1),$$

with $\delta := \max\left(\left|\frac{\theta}{2} - 1\right|, \left|\frac{u}{2} - 1\right|\right)$, and $C_1 := \max(C, 1/c)$.

Algorithm 2: FindWeights($\{f_1,\dots,f_m\},\{a_1,\dots,a_m\},j_{\min},j_{\max},w^\circ,\beta,\theta,u,c,C$)

input: Functions $f_1,\dots,f_m:\mathbb{R}\to\mathbb{R}_+$, vectors $a_1,\dots,a_m\in\mathbb{R}^n$, integers $j_{\min}<j_{\max}$, an initial weight $w^\circ\in\mathbb{R}^m_+$ satisfying (2.6) for $\beta>1$, parameters θ,u,c,C

$\varepsilon := 0.1$

$T \leftarrow \left\lceil \dfrac{\log\left(\frac{1+\beta}{\log\max(2C,2/c)}\right)}{\log\min(|\frac{2}{\theta-2}|,|\frac{2}{u-2}|)} \right\rceil$

$w_0^{(j_{\max})} \leftarrow w^\circ$

for $i=0,1,\dots,T$ **do** $w_{i+1}^{(j_{\max})} \leftarrow$ Iterate($\{f_1,\dots,f_m\},\{a_1,\dots,a_m\},w_i^{(j_{\max})},2^{j_{\max}},\varepsilon$).

$w^{(0)} \leftarrow w_T^{(0)}$

for $i=j_{\max},j_{\max}-1,\dots,j_{\min}$ **do** $w^{(i)} \leftarrow$ Iterate($\{f_1,\dots,f_m\},\{a_1,\dots,a_m\},w^{(i+1)},2^i,\varepsilon$).

return $\{w^{(j)}: j_{\min}\leqslant j\leqslant j_{\max}\}$.

Proof. Define $W := \mathrm{diag}(w_1,\dots,w_m)$ and $W' := \mathrm{diag}(w'_1,\dots,w'_m)$. Suppose that $\rho,\rho'\in\mathbb{R}^m$ are such that $\rho_i \approx_{1+\varepsilon} a_i^\top(A^\top WA)^{-1}a_i$ and $\rho'_i \approx_{1+\varepsilon} a_i^\top(A^\top W'A)^{-1}a_i$ for each $i=1,\dots,m$, and

$$\tilde{\varphi}_{s,\varepsilon}(w)_i = \frac{1}{s}\frac{f_i(\sqrt{\rho_i})}{\sqrt{\rho_i}} \quad\text{and}\quad \tilde{\varphi}_{s,\varepsilon}(u)_i = \frac{1}{s}\frac{f_i(\sqrt{\rho'_i})}{\sqrt{\rho'_i}}\,.$$

From (2.3), we have

$$d(\tilde{\varphi}_{s,\varepsilon}(w),\tilde{\varphi}_{s,\varepsilon}(w')) = d\left(\psi(\rho),\psi(\rho')\right) \leqslant \delta d(\rho,\rho') + \log(C_1).$$

Because $\rho_i \approx_{1+\varepsilon} \tau_i(w)$ and $\rho'_i \approx_{1+\varepsilon} \tau_i(w')$,

$$d(\rho,\rho') = \max_{i\in[m]}\left|\log\frac{\rho_i}{\rho'_i}\right| \leqslant \max_{i\in[m]}\left|\log\frac{\tau_i(w)}{\tau_i(w')}\right| + \log\left(\frac{1+\varepsilon}{1-\varepsilon}\right) = d(\tau(w),\tau(w')) + \log\left(\frac{1+\varepsilon}{1-\varepsilon}\right).$$

Combined with Lemma 2.6, this gives

$$d(\tilde{\varphi}_{s,\varepsilon}(w),\tilde{\varphi}_{s,\varepsilon}(w')) \leqslant \delta d(w,w') + \log(C_1) + \log\left(\frac{1+\varepsilon}{1-\varepsilon}\right) \leqslant \delta d(w,w') + \log(2C_1)\,.$$ □

Proof of Theorem 2.7. Because we use Algorithm 2 to find the desired weights, the claimed running time bound follows from Theorem 2.8 for the choice $\delta = (m(|\mathcal{J}|+\log\max\{\beta,1\}))^{-O(1)}$. Thus it suffices to argue that the output weights $\{w^{(j)}: j\in\mathcal{J}\}$ form an approximate weight scheme. For this choice of δ, taking $\tilde{\varphi}_{s_{\max},0.1}(w) :=$ Iterate($\{f_1,\dots,f_m\},\{a_1,\dots,a_m\},w,s,0.1$) satisfies, with high probability, (2.7) for all $T+|\mathcal{J}|$ calls to Algorithm 1 from Algorithm 2.

For ease of notation, let us denote $s_{\max} := 2^{j_{\max}}$. The analysis is identical to that in the proof of Theorem 2.4 (recall (2.5)), except that (2.4) is replaced by Lemma 2.9, and the initial weight bound is replaced by

$$\begin{aligned} d\left(w_0^{(j_{\max})},\tilde{\varphi}_{s_{\max},0.1}\left(w_0^{(j_{\max})}\right)\right) &\leqslant d\left(\varphi_{s_{\max}}\left(w_0^{(j_{\max})}\right),w_0^{(j_{\max})}\right) + d\left(\tilde{\varphi}_{s_{\max},0.1}\left(w_0^{(j_{\max})}\right),\tilde{\varphi}_{s_{\max},0}\left(w_0^{(j_{\max})}\right)\right) \\ &\leqslant \log\beta + \log(2C_1)\,, \end{aligned}$$

where the last inequality follows from (2.6) and Lemma 2.6. □

2.1.2 Constructing initial weights

A mild problem arises when applying Theorem 2.7, which is that it may be computationally non-trivial to locate an initial weight $w^\circ \in \mathbb{R}_+^m$ satisfying (2.6) with β sufficiently small. In this section, we show how to efficiently compute small perturbations $\hat{f}_1, \ldots, \hat{f}_m : \mathbb{R}^n \to \mathbb{R}_+$ of the functions $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}_+$ along with good initial weights w° for $\{\hat{f}_i\}$.

Fix $0 < s_{\min} < s_{\max}$ and vectors $a_1, \ldots, a_m \in \mathbb{R}^n$. Consider $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}_+$ such that $f_1^{1/2}, \ldots, f_m^{1/2}$ are L-auto-Lipschitz (P1) and θ-lower homogeneous with constant c (P2). In that case, each f_i is continuous, and therefore for any $0 < \gamma \leqslant 1$ there exist numbers $\hat{z}_1, \ldots, \hat{z}_m > 0$ such that

$$\gamma s_{\max} \leqslant f_i(\hat{z}_i) \leqslant s_{\max}\,, \quad i = 1, \ldots, m\,. \tag{2.8}$$

Define the matrix $U := \sum_{i=1}^m \hat{z}_i^{-2} a_i a_i^\top$, and let $\tilde{\tau}_i \approx_2 \langle a_i, U^{-1} a_i \rangle$ for $i = 1, \ldots, m$. Note that these values can computed using a single call to $\textsc{LevApprox}(A, (\hat{z}_1^{-2}, \ldots, \hat{z}_m^{-2}), 1/2)$. Define $w_i := \delta/\tilde{\tau}_i$ for some $\delta > 0$ and $i = 1, \ldots, m$, and finally define

$$\hat{f}_i(z) := f_i(z) + s_{\max} w_i z^2\,, \quad i = 1, \ldots, m\,, \tag{2.9}$$

The following fact is straightforward.

Fact 2.10. *If $f_1^{1/2}, \ldots, f_m^{1/2}$ are L-auto-Lipschitz and θ-lower homogeneous with constant c, then $\hat{f}_1^{1/2}, \ldots, \hat{f}_m^{1/2}$ are $\max\{1, L\}$-auto-Lipschitz and θ-lower homogeneous with constant c.*

Proof. Lower homogeneity is clear. Note that for numbers $a, b, c, d \in \mathbb{R}$, it holds that

$$|(a^2+b^2)^{1/2} - (c^2+d^2)^{1/2}| = |\|(a,b)\|_2 - \|(c,d)\|_2| \leqslant \|(a,b)-(c,d)\|_2 = ((a-c)^2 + (b-d)^2)^{1/2}\,.$$

Employ this to write

$$\begin{aligned} |\hat{f}_i(z)^{1/2} - \hat{f}_i(z')^{1/2}| &\leqslant ((f_i(z)^{1/2} - f_i(z')^{1/2})^2 + s_{\max} w_i (z-z')^2)^{1/2} \\ &\leqslant (L^2 f_i(z-z') + s_{\max} w_i (z-z')^2)^{1/2} \leqslant \max\{1, L\} \hat{f}_i(z-z')^{1/2}. \end{aligned}$$

Thus each $\hat{f}_i^{1/2}$ is $\max\{1, L\}$-auto-Lipschitz for $i = 1, 2, \ldots, m$. ☐

Theorem 2.11. *Let $F(x) := \sum_{i=1}^m f_i(\langle a_i, x\rangle)$ and $\hat{F}(x) := \sum_{i=1}^m \hat{f}_i(\langle a_i, x\rangle)$ for $\delta > 0$ and any $\hat{z}_i > 0$ satisfying (2.8). For all $x \in \mathbb{R}^n$,*

$$F(x) \leqslant s_{\max} \implies 0 \leqslant \hat{F}(x) - F(x) \leqslant 2\delta m^2 s_{\max} (L/(\gamma c))^{2/\theta}\,. \tag{2.10}$$

Moreover, w is an $O((L/c)^2 m/\delta)$-approximate weight at scale $s_{\max}$ for $\{\hat{f}_i\}$ and $\{a_i\}$.

Proof. Define $M_w := A^\top W A$, where $W := \mathrm{diag}(w_1, \ldots, w_m)$. We first claim that

$$\frac{\delta}{2} U \preceq M_w \preceq 2\delta m U\,. \tag{2.11}$$

To prove the upper bound in (2.11), note that

$$a_i a_i^\top \preceq \langle a_i, U^{-1} a_i \rangle U \preceq 2\tilde{\tau}_i U\,,$$

Summing over $i = 1, \ldots, m$ indeed gives $M_w = \delta \sum_{i=1}^m \frac{1}{\tilde{\tau}_i} a_i a_i^\top \preceq 2\delta m U$. For the lower bound, note that $U \succeq \hat{z}_i^{-2} a_i a_i^\top$ and hence $U^{-1} \preceq (\hat{z}_i^2/\|a_i\|_2^2) a_i a_i^\top$. Therefore, $\tilde{\tau}_i \leqslant 2\langle a_i, U^{-1} a_i\rangle \leqslant 2\hat{z}_i^2$ which implies that $w_i \geqslant (\delta/2)\hat{z}_i^{-2}$ and indeed gives the lower bound in (2.11).

Next we prove (2.10). The lower bound of (2.10) is trivial. To prove the upper bound, we first show that if $f_i(\langle a_i, x\rangle) \leqslant s_{\max}$, then it holds that $\langle a_i, x\rangle^2 \leqslant (L/(\gamma c))^{2/\theta} \hat{z}_i^2$ for $i = 1, \ldots, m$. Indeed, if $|\langle a_i, x\rangle| = \lambda \hat{z}_i$ for $\lambda \geqslant 1$, then lower homogeneity, symmetry, and the definition of $\hat{z}_i$ in (2.8),

$$s_{\max} \geqslant f(\langle a_i, x\rangle) \geqslant \frac{1}{L} f(|\langle a_i, x\rangle|) \geqslant \frac{c\lambda^\theta}{L} f(\hat{z}_i) \geqslant \frac{\gamma c \lambda^\theta}{L} s_{\max}\,.$$

Thus $\lambda \leqslant (L/(\gamma c))^{1/\theta}$, as desired. Now note that

$$\hat{F}(x) - F(x) = s_{\max}\langle x, M_w x\rangle \overset{(2.11)}{\leqslant} 2\delta m \langle x, Ux\rangle = 2\delta m s_{\max} \sum_{i=1}^m \hat{z}_i^{-2} \langle a_i, x\rangle^2 \leqslant 2\delta m^2 s_{\max} (L/(\gamma c))^{2/\theta}\,.$$

Finally, we establish that w are suitable approximate weights. Observe that

$$d(\varphi_{s_{\max}}(w), w) = \max_{i \in [m]} \log \left| \frac{1}{s_{\max}} \frac{\hat{f}_i((a_i^\top M_w^{-1} a_i)^{1/2})}{w_i \cdot a_i^\top M_w^{-1} a_i} \right|.$$

Then by the definition (2.9), we have

$$\frac{1}{s_{\max}} \frac{\hat{f}_i((a_i^\top M_w^{-1} a_i)^{1/2})}{w_i \cdot a_i^\top M_w^{-1} a_i} = \frac{1}{s_{\max}} \left(\frac{f_i((a_i^\top M_w^{-1} a_i)^{1/2})}{w_i \cdot a_i^\top M_w^{-1} a_i} + \frac{s_{\max}\delta/\tilde{\tau}_i}{w_i} \right) = 1 + \frac{1}{s_{\max}} \frac{f_i((a_i^\top M_w^{-1} a_i)^{1/2})}{w_i \cdot a_i^\top M_w^{-1} a_i}\,.$$

Now (2.11) and $\tilde{\tau}_i \leqslant 2\hat{z}_i^2$ together give

$$a_i^\top M_w^{-1} a_i \leqslant \frac{2}{\delta} a_i^\top U^{-1} a_i \leqslant \frac{4}{\delta} \tilde{\tau}_i \leqslant \frac{8}{\delta} \hat{z}_i^2\,, \quad i = 1, \ldots, m\,.$$

If $(a_i^\top M_w^{-1} a_i)^{1/2} \leqslant \hat{z}_i$, then by monotonicity, $f_i((a_i^\top M_w^{-1} a_i)^{1/2}) \leqslant (1/c)^2 f_i(\hat{z}_i) = (1/c)^2 s_{\max}$. Otherwise, the auto-Lipschitz and lower homogeneous properties for $f_i^{1/2}$ give $f_i(\lambda x) \leqslant (2L/c)^2 \lambda^2 f_i(x)$ for every $\lambda \geqslant 1$ by Lemma 1.4, and this implies

$$f_i((a_i^\top M_w^{-1} a_i)^{1/2}) \leqslant 4(L/c)^2 \left(\frac{(a_i^\top M_w^{-1} a_i)^{1/2}}{\hat{z}_i} \right)^2 f_i(\hat{z}_i) \lesssim (L/c)^2 \frac{s_{\max}}{\delta}\,.$$

In either case, $f_i((a_i^\top M_w^{-1} a_i)^{1/2}) \lesssim (L/c)^2 s_{\max}/\delta$. For the denominator, use (2.11) to write

$$w_i \cdot a_i^\top M_w^{-1} a_i \geqslant \frac{1}{2\delta m} w_i a_i^\top U^{-1} a_i \geqslant \frac{1}{4m} \tilde{\tau}_i \tilde{\tau}_i^{-1} = \frac{1}{4m}\,.$$

Combining everything and using that f_i is non-negative, we arrive at

$$d(\varphi_{s_{\max}}(w), w) \leqslant \log\left(1 + \frac{1}{s_{\max}} \frac{O((L/c)^2 \delta^{-1} s_{\max})}{1/(4m)}\right) \leqslant \log(1 + O((L/c)^2 \delta^{-1} m))\,. \qquad \square$$

2.2 A variational approach to approximate weights

In this section we show that even if each f_i is upper u-homogeneous for $u \geqslant 4$, approximate weights still exist. Our sparsification analysis relies on the existence of weight schemes, where there is a relationship between weights at different scales. But the following existence proof is instructive.

Theorem 2.12. *Suppose $f_1, \ldots, f_m : \mathbb{R}_+ \to \mathbb{R}_+$ are lower θ-homogeneous with constant c and upper u-homogeneous with constant C with $u > \theta > 0$. Then there is a constant $\alpha = \alpha(\theta, c, u, C)$ such that for every choice of vectors $a_1, \ldots, a_m \in \mathbb{R}^n$ and $s > 0$, there is an α-approximate weight at scale s.*

The idea behind the proof is to set up a variational problem whose critical points produce approximate weights at a given scale. As observed in [SZ01], this analysis technique does not require convexity.

Lemma 2.13. *Suppose $g_1, \ldots, g_m : \mathbb{R}_+ \to \mathbb{R}_+$ are monotone increasing, continuously differentiable, and satisfy $g_1(0) = \cdots = g_m(0) = 0$. Then for every $\beta > 0$, there are weights $\{w_i \geqslant 0 : i = 1, \ldots, m\}$ such that*

$$w_i = \gamma \frac{g_i'(\|M_w^{-1/2} a_i\|_2)}{\|M_w^{-1/2} a_i\|_2} \quad i = 1, \ldots, m\,, \tag{2.12}$$

$$\gamma = n \left(\sum_{i=1}^m g_i'(\|M_w^{-1/2} a_i\|_2) \|M_w^{-1/2} a_i\|_2 \right)^{-1}, \tag{2.13}$$

$$\beta = \sum_{i=1}^m g_i(\|M_w^{-1/2} a_i\|_2)\,.$$

Proof. For a linear operator $U : \mathbb{R}^n \to \mathbb{R}^n$, define

$$G(U) := \sum_{i=1}^m g_i(\|U a_i\|_2)\,,$$

and consider the optimization

$$\text{maximize } \{\det(U) : G(U) \leqslant \beta\}\,. \tag{2.14}$$

Since $G(0) = 0$ and each g_i is monotone increasing, it holds that $G(cI) = \beta$ for some $c > 0$. Therefore for any maximizer U^*, it holds that $\det(U^*) > 0$, i.e., U^* is invertible, and $G(U^*) = \beta$.

For U invertible, we have

$$\nabla \det(U) = \det(U) U^{-\top}. \tag{2.15}$$

Let us also calculate

$$\mathsf{d}G(U) = \sum_{i=1}^m g_i'(\|U a_i\|_2) \frac{\mathsf{d}\|U a_i\|_2^2}{2\|U a_i\|_2}\,,$$

and use $\|U a_i\|_2^2 = \operatorname{tr}(U^\top U a_i a_i^\top)$ to write

$$\frac{1}{2}\mathsf{d}\|U a_i\|_2^2 = \operatorname{tr}\left((\mathsf{d}U)^\top U a_i a_i^\top\right),$$

so that

$$\nabla_U G(U) = UA^\top D_U A\,,$$

where D_U is the $m \times m$ diagonal matrix with $(D_U)_{ii} = \frac{g_i'(\|Ua_i\|_2)}{\|Ua_i\|_2}$ and $A \in \mathbb{R}^{m\times n}$ is the matrix with rows $a_1, \ldots, a_m$.

Combined with (2.15), we see that if U is an optimal solution to (2.14), then for some Lagrange multiplier $\gamma > 0$, we have

$$(U^\top U)^{-1} = \gamma A^\top D_U A\,.$$

Take $V := (U^\top U)^{-1}$ so that $V = \sum_{i=1}^m w_i a_i a_i^\top$ with $w_i := \gamma \frac{g_i'(\|V^{-1/2}a_i\|_2)}{\|V^{-1/2}a_i\|_2}$. To compute the value of γ, calculate

$$n = \mathrm{tr}(VV^{-1}) = \gamma \sum_{i=1}^m \frac{g_i'(\|V^{-1/2}a_i\|_2)}{\|V^{-1/2}a_i\|_2}\,\mathrm{tr}(V^{-1}a_i a_i^\top) = \gamma \sum_{i=1}^m g_i'(\|V^{-1/2}a_i\|_2)\|V^{-1/2}a_i\|_2\,. \qquad \square$$

The preceding lemma assumes that the functions g_i are monotone increasing. However, the functions f_i only satisfy lower homogeneity, which is a weaker condition. To prove Theorem 2.12, one can take monotone approximations of functions f_i by averaging over intervals.

Lemma 2.14. *Suppose f is lower θ-homogeneous with constant c and upper u-homogeneous with constant C for $u \geqslant 1$. Define $K := \max(e, (2/c)^{1/\theta})$ and $g : \mathbb{R}_+ \to \mathbb{R}_+$ by*

$$g(x) := \int_x^{Kx} \frac{f(t)}{t}\,dt\,.$$

Then g is continuously differentiable, monotone increasing, and satisfies $g(0) = 0$. Moreover, for all $x > 0$,

$$\frac{f(x)}{x} \leqslant g'(x) \leqslant CK^u\,\frac{f(x)}{x}\,, \quad \forall x \geqslant 0$$
$$c\,f(x) \leqslant g(x) \leqslant CK^u\,f(x)\,.$$

Proof. Note that $g'(x) = \frac{f(Kx)-f(x)}{x}$, and $f(Kx) \geqslant cK^\theta f(x) \geqslant 2f(x)$ by lower θ-homogeneity. Thus g is monotone increasing. Moreover, we have

$$g(x) = \int_x^{Kx} \frac{f(t)}{t} dt \geqslant cf(x)\int_x^{Kx} \frac{1}{t} dt \geqslant c\cdot f(x),$$

and for $u \geqslant 1$,

$$g(x) = \int_x^{Kx} \frac{f(t)}{t} dt \leqslant Cf(x)\int_x^{Kx} \frac{(t/x)^u}{t} dt \leqslant CK^u f(x).$$

$\square$

Proof of Theorem 2.12. Throughout the proof, we use $\asymp$ in place of $\asymp_{\theta,c,u,C}$. Let K be as in Lemma 2.14 and define $g_i(x) := \int_x^{Kx} \frac{f_i(t)}{t}\,dt$. By Lemma 2.14, we have $g_i'(x) \asymp f_i(x)/x$ and $g(x) \asymp f_i(x)$ for $i = 1, \ldots, m$.

Let $\{w_i\}$ and M_w be as in Lemma 2.13 when applied to $g_1, \ldots, g_m$ with $\beta = n/\lambda$. Then,

$$\sum_{i=1}^m g_i'(\|M_w^{-1/2}a_i\|_2)\|M_w^{-1/2}a_i\|_2 \asymp \sum_{i=1}^m f_i(\|M_w^{-1/2}a_i\|_2) \asymp \sum_{i=1}^m g_i(\|M_w^{-1/2}a_i\|_2) = \beta = \frac{n}{\lambda}\,.$$

Therefore $\gamma \asymp \lambda$ (recall (2.13)), and thus (2.12) gives the desired result. $\square$

3 Covering number bounds

Consider loss functions $f_1, \ldots, f_m : \mathbb{R} \to \mathbb{R}_+$, vectors $a_1, \ldots, a_m \in \mathbb{R}^n$, and a contiguous interval $\mathcal{J} \subseteq \mathbb{Z}$. Define $h_i(x) := f_i(x)^{1/2}$ for $i = 1, \ldots, m$, $\ell := \max(\mathcal{J}) + 1$, and

$$F(x) := \sum_{i=1}^{m} f_i(\langle a_i, x\rangle) = \sum_{i=1}^{m} h_i(\langle a_i, x\rangle)^2 .$$

For $s > 0$, denote $B_F(s) := \{x \in \mathbb{R}^n : F(x) \leqslant s\}$.

Suppose that $\{w^{(j)} : j \in \mathcal{J}\}$ is an α-approximate weight scheme (Definition 1.8) for the families $\{f_i\}, \{a_i\}$, and recall that

$$M_{w^{(j)}} = \sum_{i=1}^{m} w_i^{(j)} a_i a_i^\top .$$

Define $U_j := M_{w^{(j)}}^{1/2}$ for $j \in \mathcal{J}$ so that

$$\|U_j x\|_2^2 = \sum_{i=1}^{m} w_i^{(j)} \langle a_i, x\rangle^2 , \quad x \in \mathbb{R}^n . \tag{3.1}$$

Analogously to Section 1.4, define the sets

$$\begin{aligned} \mathsf{K}_j &:= \left\{x \in \mathbb{R}^n : h_i(\langle a_i, x\rangle)^2 \leqslant 2^j w_i^{(j)} \|U_j^{-1} a_i\|_2^2 , i \in [m]\right\} , \quad j \in \mathcal{J} \\ \mathsf{K}_\ell &:= B_F(2^\ell) . \end{aligned}$$

Our primary technical goal is an estimate on the covering numbers $\mathcal{N}(B_F(2^\ell), \mathsf{K}_j)$ when the functions $h_1, \ldots, h_m$ are sufficiently nice.

Recall our assumption in Theorem 1.1 that $h_i = f_i^{1/2}$ is auto-Lipschitz (property (P1)) and lower θ-homogeneous (property (P2)). The following properties (H1)–(H5) follow from these two assumptions, but we label them specifically as they will be employed numerous times in our arguments.

Assumption 3.1. Consider $h_1, \ldots, h_m : \mathbb{R} \to \mathbb{R}_+$ for some $L, C \geqslant 1$ and $\theta > 0, c < 1$:

(H1) $h_i(\lambda x) \geqslant c h_i(\pm x)$ for all $x \in \mathbb{R}$ and $\lambda \geqslant 1$.

(H2) $|h_i(x) - h_i(y)| \leqslant L h_i(x - y)$ for all $x, y \in \mathbb{R}$.

(H3) $h_i(\lambda x) \leqslant C\lambda h_i(\pm x)$ for all $x \in \mathbb{R}$ and $\lambda \geqslant 1$.

(H4) $h_i(\lambda x) \geqslant c\lambda^\theta h_i(\pm x)$ for all $x \in \mathbb{R}$ and $\lambda \geqslant 1$.

(H5) $h_i(x \pm y) \leqslant L\,(h_i(x) + h_i(y))$ for all $x, y \in \mathbb{R}$.

Compared to the assumptions discussed previously, we have added (H5), as it is used many times. Note that (H5) follows from (H2): $h(x + y) \leqslant h(y) + Lh(x)$ and $h(x - y) \leqslant h(-y) + Lh(x) \leqslant L(h(x) + h(y))$, and (H1) is an immediate consequence of (H4).

Let us now state our primary covering estimate, which is established over the next two sections.

Theorem 3.2. *There is a number $\hat{C} = \hat{C}(L, C, \theta, c, \alpha)$ such that for any $\{h_1, \ldots, h_m\}$ satisfying Assumption 3.1, it holds that*

$$\log \mathcal{N}\Big(B_F(2^\ell), \mathsf{K}_j\Big) \leqslant \hat{C} 2^{\ell - j} \log m\,, \quad \forall j \in \mathcal{J}\,.$$

3.1 Iterative covering

In order to prove Theorem 3.2, we will relate K_j to ℓ_2 and ℓ_∞ balls with respect to an appropriate inner product structure. To this end, let us define the norms, for $j \in J$,

$$\begin{aligned} \mathsf{N}_j^\infty(x) &:= \max_{i \in [m]} \frac{|\langle U_j^{-1} a_i, U_j x\rangle|}{\|U_j^{-1} a_i\|_2}\,, \\ \mathsf{N}_j^2(x) &:= \|U_j x\|_2\,, \end{aligned}$$

and let $\mathsf{B}_j^\infty := \{x \in \mathbb{R}^n : \mathsf{N}_j^\infty(x) \leqslant 1\}$ and $\mathsf{B}_j^2 := \{x \in \mathbb{R}^n : \mathsf{N}_j^2(x) \leqslant 1\}$ denote the corresponding unit balls. We observe the following consequence of Lemma 1.12.

Corollary 3.3. *For every $j \in \mathcal{J}$ and $\eta > 0$, it holds that*

$$\sqrt{\log \mathcal{N}(\mathsf{B}_j^2, \eta \mathsf{B}_j^\infty)} \lesssim \frac{\sqrt{\log m}}{\eta}\,.$$

Proof. By scaling, we may assume that $\eta = 1$. Then Lemma 1.12 gives

$$\sqrt{\log \mathcal{N}(\mathsf{B}_j^2, \mathsf{B}_j^\infty)} \lesssim \mathbb{E} \max_{i \in [m]} \left| \left\langle \frac{U_j^{-1} a_i}{\|U_j^{-1} a_i\|_2}, g \right\rangle \right| \lesssim \sqrt{\log m}\,,$$

using the fact that if $g_1, \ldots, g_k$ are random variables that each have law $N(0,1)$, then $\mathbb{E} \max_{i \in [k]} g_i \lesssim \sqrt{\log k}$ since $\mathbb{P}[|g_i| > t] \leqslant 2e^{-t^2/2}$. □

Thus our goal will be a pair of containment results for translates of the sets K_j. These are proved in the next section.

Lemma 3.4 (ℓ_∞ control)**.** *For $j \in \mathcal{J}$ and $c_0 := (c^4/(4\alpha))^{1/2\theta}$ it holds that*

$$c_0 \mathsf{B}_j^\infty \subseteq \mathsf{K}_j \subseteq \frac{1}{c_0} \mathsf{B}_j^\infty\,.$$

Lemma 3.5 (ℓ_2 control)**.** *For any $j \in \mathcal{J}$ and $z \in B_F(2^\ell)$, it holds that*

$$B_F(2^\ell) \cap (z + \mathsf{K}_{j+1}) \subseteq z + (4L^2 C_0 2^{\ell - j})^{1/2} \mathsf{B}_j^2\,,$$

where $C_0 := \max\{(2\alpha C/c)^{1/\theta}, (\alpha C/c)^2 (\alpha/c^2)^{1/\theta}\}$.

Note that the preceding lemma relates K_{j+1} to B_j^2. This is the one place we will employ the key property of weight schemes (Definition 1.8), which is the containment $\mathsf{B}_{j+1}^2 \subseteq \alpha \mathsf{B}_j^2$: For $j, j+1 \in \mathcal{J}$ and $x \in \mathbb{R}^n$, (3.1) gives

$$\|U_j x\|_2^2 = \sum_{i=1}^m w_i^{(j)} \langle a_i, x\rangle^2 \leqslant \alpha \sum_{i=1}^m w_i^{(j+1)} \langle a_i, x\rangle^2 = \alpha \|U_{j+1} x\|_2^2 .$$

With these two results in hand, we prove Theorem 3.2.

Proof of Theorem 3.2. First, note that for any $j \in \mathcal{J}$ and $z \in B_F(2^\ell)$, we have

$$\begin{aligned}\log \mathcal{N}\Big(B_F(2^\ell) \cap (z + \mathsf{K}_{j+1}), \mathsf{K}_j\Big) &\leqslant \log \mathcal{N}\Big((4L^2 C_0 2^{\ell-j})^{1/2} \mathsf{B}_j^2, c_0 \mathsf{B}_j^\infty\Big) \\ &\leqslant C_1 2^{\ell-j} \log m \end{aligned} \tag{3.2}$$

where $C_1 := 4L^2 C_0 / c_0^2$, the first inequality follows from Lemma 3.4 and Lemma 3.5, and the second inequality is a consequence of Corollary 3.3.

An application of (3.2) with $j = \ell - 1$ and $z = 0$ gives vectors $x_1, \ldots, x_{T_1} \in B_F(2^\ell)$ with $T_1 \leqslant 2C_1 \log m$, and such that

$$B_F(2^\ell) \subseteq \bigcup_{t=1}^{T_1} \Big(B_F(2^\ell) \cap (x_t + \mathsf{K}_{\ell-1})\Big) .$$

Now apply (3.2) with $j = \ell - 2$ and $z = x_1, \ldots, x_{T_1}$ to find vectors $x_{(t_1,t_2)} \in B_F(2^\ell)$ for $1 \leqslant t_1 \leqslant T_1, 1 \leqslant t_2 \leqslant T_2$ with $T_2 \leqslant 4C_1 \log m$, and such that

$$B_F(2^\ell) \cap (x_{t_1} + \mathsf{K}_{\ell-1}) \subseteq \bigcup_{t_2=1}^{T_2} \Big(B_F(2^\ell) \cap (x_{(t_1,t_2)} + \mathsf{K}_{\ell-2})\Big) , \quad t_1 = 1, \ldots, T_1$$

Continuing inductively, we cover $B_F(2^\ell)$ by $T_1 \cdot T_2 \cdots T_r$ translates of $\mathsf{K}_{\ell-r}$, and $\log |T_r| \leqslant C_1 2^r \log m$.

We conclude that for $j \in \mathcal{J}$,

$$\log \mathcal{N}(B_F(2^\ell), \mathsf{K}_j) \leqslant \log |T_1| + \cdots + \log |T_{\ell-j}| \lesssim \Big(2 + 2^2 + \cdots + 2^{\ell-j}\Big) C_1 \log m . \qquad \square$$

3.2 Norm control

Our goal now is to prove Lemma 3.4 and Lemma 3.5. We will frequently use the following consequence of the Cauchy-Schwarz inequality

$$|\langle a_i, x\rangle| = |\langle U_j^{-1} a_i, U_j x\rangle| \leqslant \|U_j^{-1} a_i\|_2 \|U_j x\|_2 , \quad j \in \mathcal{J}, i = 1, \ldots, m, x \in \mathbb{R}^n .$$

We also restate the guarantees of our weight scheme $\{w^{(j)} : j \in \mathcal{J}\}$ (recall (1.5)):

$$\alpha^{-1} h_i(\|U_j^{-1} a_i\|_2)^2 \leqslant 2^j w_i^{(j)} \|U_j^{-1} a_i\|_2^2 \leqslant \alpha h_i(\|U_j^{-1} a_i\|_2)^2 , \quad i = 1, \ldots, m, \ j \in \mathcal{J} . \tag{3.3}$$

Lemma 3.4 follows immediately from the next two lemmas.

Lemma 3.6. *If $j \in \mathcal{J}$ and $|\langle a_i, x\rangle| \leqslant c_0 \|U_j^{-1} a_i\|_2$ for each $i \in \{1, \ldots, m\}$, then $x \in \mathsf{K}_j$.*

Proof. Suppose that $|\langle a_i, x\rangle| \leqslant \delta \|U_j^{-1} a_i\|_2$ for some $0 < \delta < 1$. Then

$$\frac{h_i(\langle a_i, x\rangle)^2}{2^j w_i^{(j)} \|U_j^{-1} a_i\|_2^2} \overset{\text{(H1)}}{\leqslant} \frac{1}{c^2} \frac{h_i(\delta \|U_j^{-1} a_i\|_2)^2}{2^j w_i^{(j)} \|U_j^{-1} a_i\|_2^2} \overset{(3.3)}{\leqslant} \frac{\alpha}{c^2} \frac{h_i(\delta \|U_j^{-1} a_i\|_2)^2}{h_i(\|U_j^{-1} a_i\|_2)^2} \overset{\text{(H4)}}{\leqslant} \frac{\alpha}{c^4} \delta^{2\theta}.$$

Taking $\delta := (c^4/(4\alpha))^{1/2\theta}$ gives $x \in \mathsf{K}_j$. ◻

Lemma 3.7. *If $j \in \mathcal{J}$ and $x \in \mathsf{K}_j$, then*

$$|\langle a_i, x\rangle| \leqslant (\alpha/c^2)^{1/2\theta} \|U_j^{-1} a_i\|_2, \quad \forall i \in \{1, \ldots, m\}. \tag{3.4}$$

Proof. Fix $i \in \{1, \ldots, m\}$. Clearly we may assume that $|\langle a_i, x\rangle| \geqslant \|U_j^{-1} a_i\|_2$. In that case, we can bound

$$c^2 \left(\frac{|\langle a_i, x\rangle|}{\|U_j^{-1} a_i\|_2}\right)^{2\theta} \overset{\text{(H4)}}{\leqslant} \frac{h_i(\langle a_i, x\rangle)^2}{h_i(\|U_j^{-1} a_i\|_2)^2} \overset{x \in \mathsf{K}_j}{\leqslant} \frac{2^j w_i^{(j)} \|U_j^{-1} a_i\|_2^2}{h_i(\|U_j^{-1} a_i\|_2)^2} \overset{(3.3)}{\leqslant} \alpha,$$

establishing (3.4). ◻

Let us now move to the proof of Lemma 3.5. The next lemma follows from an argument identical to that of Lemma 2.2.

Lemma 3.8. *For any $j \in \mathcal{J}$, it holds that*

$$\|U_j x\|_2^{2\theta} \leqslant \max\left(1, \alpha \frac{C}{c} 2^{-j} F(x)\right).$$

Lemma 3.9. *If $j, j+1 \in \mathcal{J}$ and $x \in \mathsf{K}_{j+1}$, then*

$$\|U_j x\|_2^2 \leqslant C_0 2^{-j} F(x).$$

Proof. Since $j+1 \in \mathcal{J}$, from Lemma 3.7 we have $|\langle a_i, x\rangle| \leqslant (\alpha/c^2)^{1/(2\theta)}$ for each $i = 1, 2 \ldots, m$. Therefore

$$\frac{h_i(\langle a_i, x\rangle)^2}{\langle a_i, x\rangle^2} \overset{\text{(H3)}}{\geqslant} \frac{1}{C^2} \frac{h_i((\alpha/c^2)^{1/2\theta} \|U_{j+1}^{-1} a_i\|_2)^2}{(\alpha/c^2)^{1/\theta} \|U_{j+1}^{-1} a_i\|_2^2} \overset{\text{(H1)}}{\geqslant} \frac{c^2}{C^2 (\alpha/c^2)^{1/\theta}} \frac{h_i(\|U_{j+1}^{-1} a_i\|_2)^2}{\|U_{j+1}^{-1} a_i\|_2^2} \overset{(3.3)}{\geqslant} \frac{c^2}{\alpha} \frac{2^{j+1} w_i^{(j+1)}}{C^2 (\alpha/c^2)^{1/\theta}}.$$

It follows that

$$\|U_j x\|^2 \overset{(3.1)}{=} \sum_{i=1}^m w_i^{(j)} \langle a_i, x\rangle^2 \overset{(1.8)}{\leqslant} \alpha \sum_{i=1}^m w_i^{(j+1)} \langle a_i, x\rangle^2 \leqslant \frac{(\alpha C/c)^2 (\alpha/c^2)^{1/\theta}}{2^{j+1}} \sum_{i=1}^m h_i(\langle a_i, x\rangle)^2.$$ ◻

Corollary 3.10. *For any $\gamma \geqslant 1$, it holds that if $j \in \mathcal{J}$ and $x \in B_F(\gamma 2^\ell) \cap \mathsf{K}_{j+1}$, then*

$$\|U_j x\|_2^2 \leqslant \gamma C_0 2^{\ell - j}.$$

Proof. If $j, j+1 \in \mathcal{J}$, the result follows from Lemma 3.9. Otherwise, $j = \ell - 1$ and $\mathsf{K}_{j+1} = \mathsf{K}_\ell = B_F(2^\ell)$, and it follows from Lemma 3.8. ☐

Proof of Lemma 3.5. Note that $B_F(2^\ell) \cap (z + \mathsf{K}_{j+1}) = z + \big((B_F(2^\ell) - z) \cap \mathsf{K}_{j+1}\big)$. We claim that $B_F(2^\ell) - z \subseteq B_F(4L^2 2^\ell)$. Indeed, for $x \in B_F(2^\ell)$, we have

$$F(x-z) = \sum_{i=1}^m h_i(\langle a_i, x - z\rangle)^2 \overset{\text{(H5)}}{\leqslant} \sum_{i=1}^m (L(h_i(\langle a_i, x\rangle) + h_i(\langle a_i, z\rangle)))^2 \leqslant 2L^2(F(x) + F(z)) \leqslant 4L^2 2^\ell .$$

Thus Corollary 3.10 gives

$$B_F(2^\ell) \cap (z + \mathsf{K}_{j+1}) \subseteq z + (4L^2 C_0 2^{\ell - j})^{1/2} \mathsf{B}_j^2 . \qquad ☐$$

3.3 Sparsification analysis

Consider now weights $\tau \in \mathbb{R}_+^m$ satisfying

$$\tau_i \geqslant \max_{j \in \mathcal{J}} w_i^{(j)} \|U_j^{-1} a_i\|_2^2 , \quad i = 1, \ldots, m , \tag{3.5}$$

Define $\rho_i := \tau_i / \|\tau\|_1$ for $i = 1, \ldots, m$. Let us analyze the error from taking M independent samples according to ρ, following the sparsification framework of Section 1.6.2 with $\varphi_i(x) := f_i(\langle a_i, x\rangle)$.

In this case, our potential sparsifier is given by

$$\tilde{F}_{\rho,\nu}(x) := \sum_{j=1}^M \frac{f_{\nu_j}(\langle a_{\nu_j}, x\rangle)}{M \rho_{\nu_j}} , \tag{3.6}$$

and our approximation guarantee will be derived from Lemma 1.14, by analyzing covering numbers in the following family of metrics: For a sequence $\nu \in [m]^d$ and $x, y \in \mathbb{R}^n$,

$$\begin{aligned} d_{\rho,\nu}(x,y) &:= \left(\sum_{j=1}^M \left(\frac{f_{\nu_j}(\langle a_{\nu_j}, x\rangle) - f_{\nu_j}(\langle a_{\nu_j}, y\rangle)}{M\rho_{\nu_j}} \right)^2 \right)^{1/2} \\ &= \left(\sum_{j=1}^M \frac{(h_{\nu_j}(\langle a_{\nu_j}, x\rangle) - h_{\nu_j}(\langle a_{\nu_j}, y\rangle))^2}{M\rho_{\nu_j}} \frac{(h_{\nu_j}(\langle a_{\nu_j}, x\rangle) + h_{\nu_j}(\langle a_{\nu_j}, y\rangle))^2}{M\rho_{\nu_j}} \right)^{1/2} . \end{aligned}$$

Let us first observe the bound: For $x, y \in B_F(2^\ell)$,

$$\begin{aligned} d_{\rho,\nu}(x,y) &\overset{\text{(H2)}}{\leqslant} \left(2L^2 \sum_{j=1}^M \frac{h_{\nu_j}(\langle a_{\nu_j}, x - y\rangle)^2}{M\rho_{\nu_j}} \frac{h_{\nu_j}(\langle a_{\nu_j}, x\rangle)^2 + h_{\nu_j}(\langle a_{\nu_j}, y\rangle)^2}{M\rho_{\nu_j}} \right)^{1/2} \\ &\lesssim d_\infty(x,y) L \left(\frac{\|\tau\|_1}{M} \right)^{1/2} \left(\max_{x \in B_F(2^\ell)} \tilde{F}_{\rho,\nu}(x) \right)^{1/2} , \end{aligned} \tag{3.7}$$

where we define

$$d_\infty(x, y) := \max_{i \in [m]} \frac{h_i(\langle a_i, x - y\rangle)}{\sqrt{\tau_i}}.$$

In particular, we have

$$\gamma_2(B_F(2^\ell), d_{\rho,\nu}) \lesssim L\gamma_2(B_F(2^\ell), d_\infty) \left(\frac{\|\tau\|_1}{M}\right)^{1/2} \left(\max_{x \in B_F(2^\ell)} \tilde{F}_{\rho,\nu}(x)\right)^{1/2}. \tag{3.8}$$

Let us first bound the d_∞ diameter of $B_F(2^\ell)$.

Lemma 3.11. *For any $j \in \mathcal{J}$, it holds $\operatorname{diam}(B_F(2^{j+1}), d_\infty) \leqslant C_2 2^{j/2}$ for $C_2 := 4(\alpha C_0 C^2)^{1/2}$.*

Proof. Using (3.5) and Cauchy-Schwarz yields

$$\frac{h_i(\langle a_i, x\rangle)^2}{\tau_i} \leqslant \frac{h_i(\langle a_i, x\rangle)^2}{w_i^{(j)}\|U_j^{-1}a_i\|_2^2} \overset{\text{(H1)}}{\leqslant} \frac{1}{c}\frac{h_i(\|U_j^{-1}a_i\|_2\|U_j x\|_2)^2}{w_i^{(j)}\|U_j^{-1}a_i\|_2^2}. \tag{3.9}$$

Now from Lemma 3.8, we know that for $x \in B_F(2^{j+1})$, we have $\|U_j x\|_2^2 \leqslant 2C_0$, and therefore

$$h_i(\|U_j^{-1}a_i\|_2\|U_j x\|_2)^2 \overset{\text{(H3)}}{\leqslant} 2C_0C^2 h_i(\|U_j^{-1}a_i\|_2)^2$$

In conjunction with (3.9) and (3.3), we conclude that

$$\frac{h_i(\langle a_i, x\rangle)^2}{\tau_i} \leqslant 2\alpha C_0 C^2 2^j, \quad i = 1, \dots, m,$$

which implies the desired bound. □

Combining this with (3.7) yields a diameter bound.

Corollary 3.12. *For any $j \in \mathcal{J}$, it holds that*

$$\operatorname{diam}(B_F(2^{j+1}), d_{\rho,\nu}) \lesssim 2^{j/2} L C_2 \left(\frac{\|\tau\|_1}{M}\right)^{1/2} \left(\max_{x \in B_F(2^\ell)} \tilde{F}_{\rho,\nu}(x)\right)^{1/2}.$$

We now handle small scales. For this, define $B_\infty(r) := \{x \in \mathbb{R}^n : d_\infty(x, 0) \leqslant r\}$.

Lemma 3.13. *For any $0 < r < R$, we have*

$$\sqrt{\log \mathcal{N}(B_\infty(R), B_\infty(r))} \lesssim \sqrt{\frac{m}{\theta}\log\left(\frac{2R}{cr}\right)}.$$

Proof. Denote $z_i := \sup\{|z| : h_i(z) \leqslant R\sqrt{\tau_i}\}$ for $i \in [m]$ and $\lambda := (CR/(cr))^{1/\theta}$. For $x \in \mathbb{R}^n$, define the vector $a(x) := (\langle a_1, x\rangle, \dots, \langle a_m, x\rangle)$. Then

$$|a(x)_i| \leqslant z_i, \quad \forall x \in B_\infty(R),\ i = 1, \dots, m.$$

By the Pigeonhole principle, there is a set $S \subseteq B_\infty(R)$ with $|S| \leqslant (2\lambda+1)^m$, and such that for all $x \in B_\infty(R)$, there is $\hat{x} \in S$ with $|\langle a_i, x-\hat{x}\rangle| \leqslant z_i/\lambda$ for every $i \in \{1,\ldots,m\}$. Thus for each i, we have

$$\frac{h_i(\langle a_i, x-\hat{x}\rangle)}{\sqrt{\tau_i}} \overset{\text{(H4)}}{\leqslant} \frac{1}{c\lambda^\theta}\frac{h_i(z_i)}{\sqrt{\tau_i}} \leqslant r,$$

One concludes that $\sqrt{\log \mathcal{N}(B_\infty(R), B_\infty(r))} \leqslant \sqrt{\log |S|}$, completing the proof. □

Theorem 3.14. *There is a number $K_1 = K_1(L, C, c, \theta, \alpha)$ such that if $|\mathcal{J}| \geqslant 4\log m$, then*

$$\gamma_2\left(B_F(2^\ell), d_{\rho,\nu}\right) \leqslant K_1 2^{\ell/2}(\log m)^{3/2}\left(\frac{\|\tau\|_1}{M}\right)^{1/2}\left(\max_{x\in B_F(2^\ell)} \tilde{F}_{\rho,\nu}(x)\right)^{1/2}$$

Proof. First, the assumption (3.5) guarantees that

$$\mathsf{K}_j \subseteq B_\infty(2^{j/2}), \quad \forall j \in \mathcal{J},$$

and therefore Theorem 3.2 gives

$$\left(\log \mathcal{N}\left(B_F(2^\ell), B_\infty(2^{j/2})\right)\right)^{1/2} \leqslant 2^{(\ell-j)/2}\sqrt{\hat{C}\log m}, \quad \forall j \in \mathcal{J}.$$

We now use Dudley's inequality (Lemma 1.13) to write

$$\gamma_2(B_F(2^\ell), d_\infty) \lesssim \sum_{j\in\mathbb{Z}} 2^{j/2}\sqrt{\log \mathcal{N}(B_F(2^\ell), B_\infty(2^{j/2}))}.$$

Define $\hat{\ell} := \lceil \log_2 \operatorname{diam}(B_F(2^\ell), d_\infty)\rceil$ and consider any $\ell_0 \leqslant \ell$. Splitting the sum into three pieces, we use (3.3) to bound the first two:

$$\sum_{j>\ell} 2^{j/2}\sqrt{\log \mathcal{N}(B_F(2^\ell), B_\infty(2^{j/2}))} \lesssim \operatorname{diam}(B_F(2^\ell), d_\infty)\sqrt{\hat{C}\log m}$$

$$\sum_{j=\ell_0}^{\ell} 2^{j/2}\sqrt{\log \mathcal{N}(B_F(2^\ell), B_\infty(2^{j/2}))} \lesssim 2^{\ell/2}(\ell-\ell_0)\sqrt{\hat{C}\log m}.$$

For the third piece, use Lemma 3.11 to bound $B_F(2^\ell) \subseteq B_\infty(2^{\hat{\ell}})$ so that

$$\sum_{j<\ell_0} 2^{j/2}\sqrt{\log \mathcal{N}(B_F(2^\ell), B_\infty(2^{j/2}))} \leqslant \sum_{j<\ell_0} 2^{j/2}\sqrt{\log \mathcal{N}(B_\infty(2^{\hat{\ell}}), B_\infty(2^{j/2}))}$$

$$\lesssim 2^{\ell_0/2}\sqrt{\frac{m}{\theta}\left((\hat{\ell}-\ell_0)+\log\frac{2}{c}\right)}.$$

As long as we can choose $\ell_0 \leqslant \ell - 4\log m$, we have therefore bounded

$$\gamma_2(B_F(2^\ell), d_\infty) \leqslant K_1 2^{\ell/2}(\log m)^{3/2}, \tag{3.10}$$

for some K_1 depending on the indicated parameters. Combining this with (3.8) yields the desired bound. □

Finally, we can prove Theorem 1.1.

Proof of Theorem 1.1. Let us denote $\mathcal{J} := \{j_{\min}, \dots, j_{\max}\}$, where

$$j_{\max} := \lceil \log_2 s_{\max} \rceil, \; j_{\min} := \lfloor \log_2 s_{\min} - 4 \log m \rfloor .$$

Theorem 2.4 yields the existence of an α-approximate weight scheme $\{w^{(j)} : j \in \mathcal{J}\}$ with $\alpha \lesssim_{L,c,\theta} 1$. This yields τ as in (3.5) satisfying

$$\|\tau\|_1 \lesssim_{L,c,\theta} n|\mathcal{J}| \lesssim n \log(m s_{\max}/s_{\min}) .$$

Now let us choose an integer $M := C\varepsilon^{-2}\|\tau\|_1(\log m)^3$ for a sufficiently large constant $C \geqslant 1$, and denote $\Omega_j := \{x \in \mathbb{R}^n : \hat{F}(x) \leqslant 2^j\}$ for $j \in \mathcal{J}$. Then Theorem 3.14 gives us the following, in relation to the assumptions of Lemma 1.14:

$$\delta \lesssim 2^{j/2} \log(m) \sqrt{C_1 \log m} \left(\frac{\|\tau\|_1}{M} \right)^{1/2} \leqslant \varepsilon 2^{j/2},$$

and Corollary 3.12 gives

$$\hat{\delta} \lesssim 2^{j/2} L\, C_2 \left(\frac{\|\tau\|_1}{M} \right)^{1/2} \leqslant \frac{\varepsilon 2^{j/2}}{(\log m)^{3/2}} .$$

Thus Lemma 1.14 shows that for some constant $c_0 \gtrsim_{L,c,\theta} 1$, with probability at least $1 - e^{-c_0(\log m)^3}|\mathcal{J}|$ over the choice of $\boldsymbol{\nu} \in [m]^M$,

$$\max_{x \in \Omega_j} |\tilde{F}_{\rho,\boldsymbol{\nu}}(x) - F(x)| \lesssim \varepsilon 2^j \quad \forall j \in \mathcal{J} , \tag{3.11}$$

where $\tilde{F}_{\rho,\nu}$ is defined as in (3.6). Moreover, $\tilde{F}_{\rho,\nu}$ is manifestly s-sparse with

$$s \leqslant M \lesssim_{L,c,\theta} \frac{n}{\varepsilon^2} \log(m s_{\max}/s_{\min})(\log m)^3 . \tag{3.12}$$

Observe that (3.11) gives

$$|\tilde{F}(x) - F(x)| \lesssim \varepsilon F(x) , \quad \forall x \in \mathbb{R}^n \text{ with } F(x) \in [s_{\min}, s_{\max}] , \tag{3.13}$$

with $\tilde{F} = \tilde{F}_{\rho,\nu}$. We remark that (3.12) is slightly worse than the bound claimed in Theorem 1.1. We address this at the end of the proof.

The algorithmic argument. To produce a sparsifier algorithmically, we need to be slightly more careful. Let $\hat{f}_1, \dots, \hat{f}_m : \mathbb{R}^n \to \mathbb{R}_+$ be the perturbations guaranteed by Theorem 2.11 with $\delta \asymp_{L,c,\theta} \varepsilon s_{\min}/(m^3 s_{\max})$ so that (2.10) gives

$$|F(x) - \hat{F}(x)| \leqslant \varepsilon s_{\min} , \quad \forall x \in \mathbb{R}^n \text{ with } F(x) \leqslant s_{\max} , \tag{3.14}$$

where $\hat{F}(x) := \sum_{i=1}^m \hat{f}_i(\langle a_i, x \rangle)$.

From a combination of Theorem 2.11 and Theorem 2.7, we can algorithmically produce an α-approximate weight scheme $\{w^{(j)} : j \in \mathcal{J}\}$ for $\{\hat{f}_i\}$ and $\{a_i\}$ with $\alpha \lesssim_{L,c,\theta} 1$, in time

$$\tilde{O}_{L,c,\theta}\left((\mathsf{nnz}(a_1, \dots, a_m) + n^\omega + m\mathcal{T}_{\text{eval}}) \log \frac{m s_{\max}}{\varepsilon s_{\min}} \right) .$$

By Fact 2.10, the preceding analysis applies to $\hat{f}_1, \ldots, \hat{f}_m$, yielding, with probability at least $1 - e^{-c_0(\log m)^3}|\mathcal{J}|$, a sparsifier $\tilde{F}$ of $\hat{F}$ with

$$|\tilde{F}(x) - \hat{F}(x)| \lesssim \varepsilon \hat{F}(x), \quad \forall x \in \mathbb{R}^n \text{ with } \hat{F}(x) \in [\tfrac{1}{2}s_{\min}, 2s_{\max}],$$

Combined with (3.14) and the triangle inequality, this gives

$$|\tilde{F}(x) - F(x)| \lesssim \varepsilon F(x), \quad \forall x \in \mathbb{R}^n \text{ with } F(x) \in [s_{\min}, s_{\max}],$$

completing the proof.

Improving the m dependence. The preceding arguments yield a sparsity bound of the form (3.12), together with an approximation guarantee (3.13). Define $\tilde{F}_0 := F$ and $m_0 := m$. Suppose, inductively, that $\tilde{F}_i$ has m_i non-zero terms. Let $\varepsilon_i > 0$ be a given, and let $\tilde{F}_{i+1}$ be such that (3.13) holds with $F = \tilde{F}_i$ and $\tilde{F} = \tilde{F}_{i+1}$, and such that $\tilde{F}_{i+1}$ has at most m_{i+1} non-zero terms, where

$$m_{i+1} \lesssim \frac{n}{\varepsilon_i^2} \log(m_i \hat{s}_{\max}/\hat{s}_{\min})(\log m_i)^3.$$

By scaling $\tilde{F}_i$ and adjusting ε_i by a constant factor, we may assume that

$$(1 - \varepsilon_i)\tilde{F}_i(x) \leqslant \tilde{F}_{i+1}(x) \leqslant \tilde{F}_i(x), \quad \forall x \in \mathbb{R}^n \text{ with } \tilde{F}_i(x) \in [\hat{s}_{\min}, \hat{s}_{\max}].$$

Then the triangle inequality gives, for any $h \geqslant 1$, and all $x \in \mathbb{R}^n$ satisfying $\tilde{F}_0(x), \ldots, \tilde{F}_{h-1}(x) \in [\hat{s}_{\min}, \hat{s}_{\max}]$,

$$|F(x) - \tilde{F}_h(x)| \leqslant \sum_{i=0}^{h-1} |\tilde{F}_i(x) - \tilde{F}_{i+1}(x)| \leqslant \sum_{i=0}^{h-1} \varepsilon_i \tilde{F}_i(x) \leqslant F(x) \sum_{i=0}^{h-1} \varepsilon_i. \tag{3.15}$$

It is straightforward that one can choose a geometrically increasing sequence $\varepsilon_0 < \varepsilon_1 < \cdots < \varepsilon_{h-1} = \varepsilon$ such that $m_h \lesssim \frac{n}{\varepsilon^2} \log(S \frac{\hat{s}_{\max}}{\hat{s}_{\min}})(\log S)^3$, with $S := \frac{n}{\varepsilon} \log \frac{2\hat{s}_{\max}}{\hat{s}_{\min}}$. Since the sequence $\{\varepsilon_i\}$ increases geometrically, the last expression in (3.15) is at most $C\varepsilon F(x)$ for some $C > 1$.

Let us now take $\hat{s}_{\min} := \frac{1}{2}s_{\min}$, $\hat{s}_{\max} := s_{\max}$. We may assume that $\varepsilon < 1/(2C)$, meaning that if $F_0(x) = F(x) \in [s_{\min}, s_{\max}]$, then $\tilde{F}_0(x), \ldots, \tilde{F}_{h-1}(x) \in [\hat{s}_{\min}, \hat{s}_{\max}]$, and therefore $\tilde{F}_h$ yields the desired sparsifier. □

3.4 The γ_p losses

First, let us verify that the γ_p losses satisfy the assumptions of Theorem 1.1. Define a generalization with variable thresholds: For $t \geqslant 0$,

$$\gamma_p(t, z) := \begin{cases} \frac{p}{2} t^{p-2} z^2, & |z| \leqslant t \\ |z|^p - (1 - \frac{p}{2}) t^p, & |z| \geqslant t, \end{cases} \tag{3.16}$$

Lemma 3.15. *Suppose $h : \mathbb{R} \to \mathbb{R}$ is continuous, symmetric and differentiable at all but finitely many points. If h is additionally approximately concave in the sense that $h'(x) \leqslant L\, h'(y)$ for $0 < x < y$, then h is L-auto-Lipschitz.*

Proof. Consider $0 < x < y$. Then,

$$h(y) - h(x) = \int_x^y h'(z)\, dz \leqslant L \int_0^{y-x} h'(z)\, dz = L\, h(y-x)\,. \qquad \square$$

Lemma 3.16. *For every $p \in (0, 2]$ and $t > 0$, the function $z \mapsto \gamma_p(t, z)^{1/2}$ is 1-auto-Lipschitz (P1) and lower $p/2$-homogeneous (P2) with constant 1.*

Proof. Fix $p \in (0, 2]$ and $t > 0$, and define $h(z) := \gamma_p(t, z)^{1/2}$. By Lemma 3.15, to show that h is 1-auto-Lipschitz, it suffices to show that $h'(z)$ is non-increasing for $z > 0$. This is clear for $z < t$, since $h'(z)$ is independent of z. For $z > t$, we calculate

$$\begin{aligned} h''(z) &= \frac{p(p-1)z^{p-2}}{2(z^p + (1-p/2)t^p)^{1/2}} - \frac{p^2 z^{2p-2}}{4(z^p + (1-p/2)t^p)^{3/2}} \\ &= \frac{p(2-p)z^{p-2}}{4(z^p + (1-p/2)t^p)^{3/2}} \left((p-1)t^p - z^p\right) \leqslant 0\,. \end{aligned}$$

To see that $h(z)^2$ is lower p-homogeneous with constant 1, we first check this for $x < y \leqslant t$. Then, $h(y)/h(x) = (y/x)^2 \geqslant (y/x)^p$. If $t \leqslant x < y$, then

$$\frac{h(y)^2}{h(x)^2} = \frac{y^p - (1-\frac{p}{2})t^p}{x^p - (1-\frac{p}{2})t^p} \geqslant (y/x)^p.$$

For $x < t < y$, we write

$$\frac{h(y)^2}{h(x)^2} = \frac{h(y)^2}{h(t)^2} \cdot \frac{h(t)^2}{h(x)^2} \geqslant (y/t)^p (t/x)^p = (y/x)^p. \qquad \square$$

3.4.1 The Huber Loss

Let us now prove Lemma 1.5 and Corollary 1.6.

Proof of Lemma 1.5. By scaling, we may assume that $w_{\max} = 1$. Accordingly, let us define $s_{\min} := w_{\min}$ and $s_{\max} := 6m^2$. Consider $x \in \mathbb{R}^n$ with $F(x) \leqslant w_{\min}$. In that case, $\gamma_1(\langle a_i, x\rangle) \leqslant 1$, and therefore $|\langle a_i, x\rangle| \leqslant 1$ for all $i = 1, \dots, m$. In particular, this implies that $\gamma_1(\langle a_i, x\rangle) = \langle a_i, x\rangle^2/2$ for each i so that $F(\lambda x) = \lambda^2 F(x)$ and $\tilde{F}(\lambda x) = \lambda^2 \tilde{F}(x)$ for $\lambda := (F(x)/w_{\min})^{-1/2}$. If we take $y := \lambda x$, then $F(y) = w_{\min} \in [s_{\min}, s_{\max}]$, so it holds that $|F(y) - \tilde{F}(y)| \leqslant \varepsilon F(y)$. Scaling by λ recovers the same guarantee for x.

Now consider $x \in \mathbb{R}^n$ with $F(x) \geqslant s_{\max}$. Note that $|\langle a_i, y\rangle| - \frac{1}{2} \leqslant \gamma_1(\langle a_i, y\rangle) \leqslant |\langle a_i, y\rangle|$ holds for any $y \in \mathbb{R}^n$ and $i \in \{1, \dots, m\}$, therefore

$$F(x) \leqslant \sum_{i=1}^m |\langle a_i, x\rangle| \leqslant F(x) + m/2\,.$$

Denote $\lambda := s_{\max}/(2F(x))$ and $y := \lambda x$, and note that

$$F(y) = \sum_{i=1}^m \gamma_1(\langle a_i, y\rangle) \leqslant \lambda \sum_{i=1}^m |\langle a_i, x\rangle| \leqslant \lambda(F(x) + m/2) \leqslant s_{\max}$$

$$F(y) \geqslant w_{\min}\left(-m/2 + \lambda \sum_{i=1}^{m} |\langle a_i, x\rangle|\right) \geqslant w_{\min}\left(\lambda F(x) - m/2\right) \geqslant s_{\min}\,.$$

Therefore $|F(y) - \tilde{F}(y)| \leqslant \varepsilon F(y)$ by assumption.

Now use the fact that $|\gamma_1(\beta z) - \beta\gamma_1(z)| \leqslant \beta$ for all $\beta > 1$ and $z \in \mathbb{R}$, implying

$$|F(x) - \lambda^{-1}F(y)| \leqslant \frac{1}{\lambda}\sum_{i=1}^{m} w_i \leqslant \frac{m}{\lambda}\,,$$
$$|\tilde{F}(x) - \lambda^{-1}\tilde{F}(y)| \leqslant \frac{1}{\lambda}\sum_{i=1}^{m} \tilde{w}_i \leqslant \frac{m}{\lambda}\,.$$

The triangle inequality gives

$$|F(x) - \tilde{F}(x)| \leqslant \frac{2m}{\lambda} + \frac{|F(y) - \tilde{F}(y)|}{\lambda} \leqslant \frac{4m}{s_{\max}}F(x) + \frac{\varepsilon}{\lambda}F(y) \leqslant 2\varepsilon F(x)\,,$$

where we have used $s_{\max} \geqslant 4m^2$ and $\varepsilon > 1/m$. □

Proof of Corollary 1.6. Note that the sparsifier $\tilde{F} = \tilde{F}_{\rho,\nu}$ satisfying the guarantee (3.13) is produced via importance sampling with respect to the distribution specified by $\rho \in \mathbb{R}_+^m$, and if the ith term has probability ρ_i to be sampled, then it is weighted by $(\rho_i M)^{-1}$ (recalling (3.6)). Moreover, $\tilde{F}_{\rho,\boldsymbol{\nu}}$ satisfies (3.13) with probability $1 - o(1)$ as $m \to \infty$.

By a union bound, $\mathbb{P}(\min_{j\in[M]} \rho_{\boldsymbol{\nu}_j} \leqslant \delta) \leqslant \delta M m$. Therefore with probability at least $1/2$, it holds $\min_{j\in[M]} \rho_{\boldsymbol{\nu}_j} \geqslant \frac{1}{2Mm}$. If we start with a uniform-weighted function $F(x) = \gamma_1(\langle a_1, x\rangle) + \cdots + \gamma_1(\langle a_m, x\rangle)$, then with probability at least $1/2$, the maximum weight in $\tilde{F}$ will be $2m$.

Thus if $\tilde{F}$ satisfies (3.13) with $s_{\min} = 1/2$ and $s_{\max} := 8m^3$, then Lemma 1.5 shows that that $\tilde{F}$ is a genuine $O(\varepsilon)$-approximation to F. Now the sparsity guarantee of Theorem 1.1 (or see (3.12)) completes the proof. □

3.4.2 The Tukey Loss

Let us denote the Tukey loss $\mathsf{T}(z) := \min\{z^2, 1\}$, and use $h(z) := \mathsf{T}(z)^{1/2} = \min\{|z|, 1\}$. We use the notation from the beginning of the section.

Theorem 3.17. *Suppose $a_1, \ldots, a_m \in \mathbb{R}^n$ satisfy $\|a_i\|_2 \leqslant n^{O(1)}$ for $i = 1, \ldots, m$, and denote $F(x) := \sum_{i=1}^m \mathsf{T}(\langle a_i, x\rangle)$. Then there are weights $w \in \mathbb{R}_+^m$ with $|\mathrm{supp}(w)| \leqslant n^{1+o(1)}\varepsilon^{-2}(\log m)^4$, and such that*

$$(1-\varepsilon)F(x) \leqslant \sum_{i=1}^{m} w_i \mathsf{T}(\langle a_i, x\rangle) \leqslant (1+\varepsilon)F(x)\,, \quad \forall x \in \mathbb{R}^n \text{ with } \|x\|_2 \leqslant n^{O(1)}\,. \tag{3.17}$$

The weights w can be computed in time $\tilde{O}(\mathsf{nnz}(a_1, \ldots, a_m) + n^{\omega} + m)$ with high probability.

The remainder of the section is sketching a proof of Theorem 3.17. Note that the Tukey loss satisfies (P1) and (P3)–(P5), but not (P2), which is why Theorem 1.1 is not directly applicable. This motivates us to define the function $\overline{h}(x) := \min\{|x|, |x|^\eta\}$ for some small $\eta > 0$ to be chosen later, and $\overline{f_i}(x) := \overline{h}(\langle a_i, x\rangle)^2$ for $i = 1, \ldots, m$, along with $\overline{F}(x) := \sum_{i=1}^m \overline{f_i}(\langle a_i, x\rangle)$.

By Theorem 2.4, for some $\alpha = \alpha(\eta)$, there is an α-approximate weight scheme $\{w^{(j)} : j \in \mathcal{J}\}$ for $\{\overline{f_i}\}$ and $\{a_i\}$, where $\mathcal{J} := \{j \in \mathbb{Z} : |j| \leqslant C \log m\}$, with the constant $C > 1$ chosen sufficiently large. Define $\tau_i := \max_{j \in J} w_i^{(j)} \|U_j^{-1} a_i\|_2^2$ for $i = 1, \ldots, m$ so that $\|\tau\|_1 \lesssim n|\mathcal{J}| \lesssim n \log m$, along with sampling weights $\rho_i := \tau_i / \|\tau\|_1$ for $i = 1, \ldots, m$.

Take $\Omega := \{x : \|x\|_2 \leqslant n^{O(1)}\}$, with the notation of Section 1.6.2 and $\varphi_i(x) := \mathsf{T}(\langle a_i, x\rangle)$, and $\tilde{F}(x) = \sum_{i=1}^M \frac{1}{M\rho_{\nu_i}} \mathsf{T}(\langle a_{\nu_i}, x\rangle)$ as constructed in Section 1.6.2. Consider $\ell \in \mathbb{Z}$ with $|\ell| \leqslant 2 \log m$. We can estimate (as in (3.7)): for any $x, y \in \Omega \cap B_F(2^\ell)$,

$$d_{\rho,\nu}(x, y) \lesssim d_\infty(x, y) \left(\frac{\|\tau\|_1}{M}\right)^{1/2} \left(\max_{x \in \Omega \cap B_F(2^\ell)} \tilde{F}_{\rho,\nu}(x)\right)^{1/2} \quad \text{for} \quad d_\infty(x, y) := \max_{i \in [m]} \frac{h(\langle a_i, x - y\rangle)}{\sqrt{\tau_i}},$$

One should observe that (3.7) only requires (H2), which $\mathsf{T}(z)^{1/2}$ satisfies.

Let us also define the distance $\overline{d}_\infty(x, y) := \max_{i \in [m]} \frac{\overline{h}(\langle a_i, x - y\rangle)}{\sqrt{\tau_i}} \geqslant d_\infty(x, y)$. Note that $\Omega \cap B_F(2^\ell) \subseteq B_{\overline{F}}(n^{O(\eta)} 2^\ell)$. Therefore,

$$\begin{aligned}
\operatorname{diam}(\Omega \cap B_F(2^\ell), d_{\rho,\nu}) &\lesssim \left(\frac{\|\tau\|_1}{M}\right)^{1/2} \left(\max_{x \in \Omega \cap B_F(2^\ell)} \tilde{F}_{\rho,\nu}(x)\right)^{1/2} \cdot \operatorname{diam}(\Omega \cap B_F(2^\ell), \overline{d}_\infty) \\
&\leqslant \left(\frac{\|\tau\|_1}{M}\right)^{1/2} \left(\max_{x \in \Omega \cap B_F(2^\ell)} \tilde{F}_{\rho,\nu}(x)\right)^{1/2} \cdot \operatorname{diam}(B_{\overline{F}}(n^{O(\eta)} 2^\ell), \overline{d}_\infty) \\
&\leqslant \left(\frac{\|\tau\|_1}{M}\right)^{1/2} \left(\max_{x \in \Omega \cap B_F(2^\ell)} \tilde{F}_{\rho,\nu}(x)\right)^{1/2} \cdot C(\eta) n^{O(\eta)} 2^{\ell/2},
\end{aligned}$$

where the last inequality is from Lemma 3.11, and $C(\eta)$ is a number depending on $\eta > 0$.

Thus for some choice $M \lesssim C(\eta) \varepsilon^{-2} n^{1+O(\eta)} (\log m)^4$, we have

$$\operatorname{diam}(\Omega \cap B_F(2^\ell), d_{\rho,\nu}) \lesssim \frac{\varepsilon 2^{\ell/2}}{(\log m)^{3/2}} \left(\max_{x \in \Omega \cap B_F(2^\ell)} \tilde{F}_{\rho,\nu}(x)\right)^{1/2}.$$

Similarly,

$$\begin{aligned}
\gamma_2(\Omega \cap B_F(2^\ell), d_{\rho,\nu}) &\lesssim \left(\frac{\|\tau\|_1}{M}\right)^{1/2} \left(\max_{x \in \Omega \cap B_F(2^\ell)} \tilde{F}_{\rho,\nu}(x)\right)^{1/2} \cdot \gamma_2(\Omega \cap B_F(2^\ell), \overline{d}_\infty) \\
&\leqslant \left(\frac{\|\tau\|_1}{M}\right)^{1/2} \left(\max_{x \in \Omega \cap B_F(2^\ell)} \tilde{F}_{\rho,\nu}(x)\right)^{1/2} \cdot \gamma_2(B_{\overline{F}}(n^{O(\eta)} 2^\ell), \overline{d}_\infty) \\
&\overset{(3.10)}{\leqslant} \left(\frac{\|\tau\|_1}{M}\right)^{1/2} \left(\max_{x \in \Omega \cap B_F(2^\ell)} \tilde{F}_{\rho,\nu}(x)\right)^{1/2} \cdot C(\eta) n^{O(\eta)} 2^{\ell/2} (\log m)^{3/2} \\
&\leqslant \varepsilon 2^{\ell/2} \left(\max_{x \in \Omega \cap B_F(2^\ell)} \tilde{F}_{\rho,\nu}(x)\right)^{1/2},
\end{aligned}$$

again for $M \lesssim C(\eta)\varepsilon^{-2}n^{1+O(\eta)}(\log m)^4$ chosen sufficiently large. This implies $|F(x) - \tilde{F}(x)| \leqslant \varepsilon \cdot F(x)$ for all $x \in \Omega \cap B_F(2^\ell)$. This holds for all $|\ell| \leqslant 2\log m$, so $|F(x) - \tilde{F}(x)| \leqslant \varepsilon \cdot F(x)$ for all $x \in \Omega$.

Thus applying Lemma 1.14 shows that $\tilde{F}$ satisfies (3.17). The claimed result now follows by choosing $\eta = o(1)$ as $n \to \infty$.

We remark that if the proofs of Lemma 3.11 and (3.10) are unraveled, one can check that $C(\eta) \lesssim \exp(O(\eta^{-2}))$, so the optimal choice is $\eta \asymp (\log n)^{-1/3}$, yielding sparsifiers of size $\varepsilon^{-2}n^{1+O((\log n)^{-1/3})}(\log m)^4$.

4 Algorithms for generalized linear models

4.1 Optimizing generalized linear models

Here we describe an algorithm that optimizes a GLM (recall (1.1)) under certain assumptions on its divergence. The algorithm reduces optimizing a GLM to a few queries to what we call a *sparse GLM oracle* that optimizes convex GLMs induced by a limited number of functions plus a linear term. We start by defining convex GLMs (Definition 4.1), a sparse GLM oracle (Definition 4.2), the assumptions on the functions we consider (Assumption 4.3), and giving the main theorem on the algorithm for optimizing convex GLMs with a sparse GLM oracle (Theorem 4.4).

Definition 4.1 (Convex GLM)**.** We call a a family $\mathcal{F} = \{f_1, \dots, f_m\}$ of differentiable, convex functions, along with vectors $a_1, \dots, a_m \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$ a *convex GLM*. We let $F(x) \stackrel{\text{def}}{=} \sum_{i=1}^m f_i(\langle a_i, x\rangle - b_i)$ and let $F^* := \min_{x \in \mathbb{R}^n} F(x)$.

Definition 4.2 (Sparse GLM oracle)**.** We call an algorithm a *sparse GLM oracle* for a convex GLM (Definition 4.1) if when given as input $\varepsilon > 0$, $w \in \mathbb{R}^m_+$, and $y, x_{\text{in}} \in \mathbb{R}^n$, it outputs a vector $x_{\text{out}} \in \mathbb{R}^n$ such that

$$G(x_{\text{out}}) - G(x^*) \leqslant \varepsilon(G(x_{\text{in}}) - G(x^*)),$$
$$\text{where } G(x) := \langle y, x\rangle + \sum_{i=1}^m w_i f_i(\langle a_i, x\rangle - b_i), \quad \text{and} \quad x^* := \operatorname{argmin}_{x \in \mathbb{R}^n} G(x).$$

We use $\mathcal{T}^{\mathcal{F}}_{\text{GLM}}(s, \varepsilon)$ to denote the worst-case runtime of such a GLM oracle on inputs where w has at most s nonzero entries, over all convex GLMs with a given family of functions $\mathcal{F} = \{f_1, \dots, f_m\}$.

Assumption 4.3 (Approximate Divergence Sparsifiability)**.** Assume that, for some parameters, $\theta > 1$ and $\alpha, c, L > 0$, the function $f : \mathbb{R} \to \mathbb{R}$ is convex, differentiable, and satisfies the following property: For every $x_0 \in \mathbb{R}$, there is a convex function $r_{x_0} : \mathbb{R} \to \mathbb{R}$ that can be evaluated in $\tilde{O}(\mathcal{T}_{\text{eval}})$ time, and such that

- $r_{x_0}(\Delta) \leqslant D^f_{x_0}(x_0 + \Delta) \leqslant \alpha \cdot r_{x_0}(\Delta)$ for all $\Delta \in \mathbb{R}$,
- $r_{x_0}^{1/2}$ is L-auto-Lipschitz (property (P1)).
- r_{x_0} is lower θ-homogeneous with constant c (property (P2)).

Theorem 4.4. *Consider $x^{(0)} \in \mathbb{R}^n$ and any convex GLM (Definition 4.1) where each $f_i \in \mathcal{F}$ satisfies Assumption 4.3 with uniform constants $\theta > 1, \alpha, c, L > 0$. Provided an upper bound $F(x^{(0)}) - F^* \leqslant \Gamma$, with high probability* SolveGLM *(Algorithm 4) outputs a vector $y \in \mathbb{R}^n$ satisfying $F(y) - F^* \leqslant \delta$ in time*

$$\tilde{O}\left(\eta^{-1}(\mathsf{nnz}(A) + \mathcal{T}_{\mathrm{GLM}}^{\mathcal{F}}(\tilde{O}_{\alpha,c,\theta,L}(n), O(\eta/\alpha)) + m\mathcal{T}_{\mathrm{eval}})\log(\Gamma/\delta)\right),$$

where $\eta := (10\alpha^2/c)^{-1/(\theta-1)}$, and $\mathcal{T}_{\mathrm{eval}}$ is the worst-case time needed to evaluate a function f_i.

Under mild assumptions on the family $\mathcal{F} = \{f_1, \ldots, f_m\}$, it holds that, $\mathcal{T}_{\mathrm{GLM}}^{\mathcal{F}}(\tilde{O}(n), \varepsilon) \leqslant \tilde{O}(n^{\omega_0}\log(1/\varepsilon))$, where ω_0 is as defined as in Section 1.2 (see [LSZ19b, Theorem 4.2] for the formal conditions).

The algorithm establishing Theorem 4.4 is straightforward. In each iteration, we write our optimization problem as a divergence with respect to the current point $x^{(t)}$. Then we sparsify the divergence using Theorem 1.1. Finally, we call a GLM oracle (Definition 4.2), which is efficient, as we have sparsified the sum down to $\tilde{O}(n)$ terms. We use this to take a step, obtaining an improved point $x^{(t+1)}$, and repeat this iteration $\tilde{O}(1)$ times. The main technical point is in handling the dependence of Theorem 1.1 on the scale parameters $s_{\min}$ and $s_{\max}$.

Recall from Definition 1.9 that T_x^f is the first-order Taylor approximation to f at x, and $D_x^f = f - T_x^f$ is the associated divergence.

Algorithm 3: GLMIterate$(\{f_1, \ldots, f_m\}, \{a_1, \ldots, a_m\}, b, x, \tilde{\Gamma})$

---

input: Functions $f_i : \mathbb{R} \to \mathbb{R}$, vectors $a_i \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, initial point $x \in \mathbb{R}^n$. Performs one step of iterative refinement to decrease function value.

$y_i \leftarrow \langle a_i, x\rangle - b_i$ for $i \in [m]$.

Let r_{i,y_i} denote the approximation to $D_{y_i}^{f_i}$, as in Assumption 4.3.

$s_{\min} \leftarrow m^{-O(1)}\tilde{\Gamma}$, $s_{\max} \leftarrow m^{O(1)}\tilde{\Gamma}$.

Use Theorem 1.1 to find weights $w \in \mathbb{R}_{\geqslant 0}^m$ that induce an $\tilde{O}_{\alpha,c,\theta,L}(n)$-sparse $(1/10)$-approximation of $r(\Delta) := \sum_{i\in[m]} r_{i,y_i}(\langle a_i, \Delta\rangle)$ for all $s_{\min} \leqslant r(\Delta) \leqslant s_{\max}$.

Let $h(\Delta) := T_x^F(x + \Delta) + \frac{2}{3\alpha}\sum_{i\in[m]} w_i D_{y_i}^{f_i}(\langle a_i, x + \Delta\rangle - b_i)$.

$\eta \leftarrow (10\alpha^2/c)^{-1/(\theta-1)}$.

Find $\hat{\Delta}$ satisfying $h(\hat{\Delta}) \leqslant h(\Delta^*) + \tilde{\Gamma}/10$ by calling a GLM oracle (Definition 4.2) with $\mathcal{F} = \{f_1, \ldots, f_m\}$, $x_{\mathrm{in}} = 0$, and $\varepsilon = \eta/(30\alpha)$.

if $F(x + \eta\hat{\Delta}) \leqslant F(x)$ **then return** $x + \eta\hat{\Delta}$ **else return** x

---

Let $E = F(x) - F(x^*)$ denote the function value error. The analysis of this algorithm proceeds by showing that one step of GLMIterate decreases E as long as $C_0 s_{\min} \leqslant E \leqslant s_{\max}/C_0$ for sufficiently large constant C_0 (and otherwise, does not increase the error). Intuitively, as long as this is true, we will show that $\Delta^* := x^* - x$ and $\hat{\Delta}$ are points where we indeed have sparsification guarantees by Theorem 1.1. Towards this, we first establish that $r(\Delta^*) \asymp E$, where throughout this section we allow the implicit constants in the $\asymp, \lesssim, \gtrsim$ notation to depend on the parameters c, L, θ, α in Theorem 4.4.

Lemma 4.5. *In the setting of Theorem 4.4, for $\Delta^* := x^* - x$ we have $E/\alpha \lesssim r(\Delta^*) \leqslant 2E/\eta$.*

Algorithm 4: SolveGLM($\{f_1, \dots, f_m\}, \{a_1, \dots, a_m\}, b, x^{(0)}, \Gamma, \delta$)

input: Functions $f_i : \mathbb{R} \to \mathbb{R}$, vectors $a_i \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, initial point $x^{(0)} \in \mathbb{R}^n$. Output a point that is a δ-approximate minimizer of $\min_x \sum_{i \in [m]} f_i(\langle a_i, x \rangle - b_i)$.

```
τ ← ⌈2η^{-1} log(Γ/δ)⌉ for η = (10α²/c)^{-1/(θ-1)}.
for T = 0, 1, ..., τ - 1 do
    x^{(T+1)} ← GLMIterate({f_1, ..., f_m}, {a_1, ..., a_m}, b, x^{(T)}, (1 - η/2)^T Γ).
end
return x^{(τ)}.
```

Proof. Let $g = \nabla F(x)$. By optimality of Δ^*, and because (P5) holds by Lemma 1.4,

$$-E \leqslant \langle g, 2\Delta^* \rangle + D_x^F(x + 2\Delta^*) \leqslant 2 \langle g, \Delta^* \rangle + \alpha r(2\Delta^*) \overset{\text{(P5)}}{\leqslant} 2 \langle g, \Delta^* \rangle + O(\alpha \cdot r(\Delta^*)).$$

Because $\langle g, \Delta^* \rangle \leqslant -E$ by convexity, we deduce that $r(\Delta^*) \gtrsim -E/\alpha$.

For the upper bound, choose $\eta = (10\alpha/c)^{-1/(\theta-1)}$. By optimality of Δ^* we have

$$\begin{aligned} -E &\leqslant \langle g, \eta\Delta^* \rangle + D_x^F(x + \eta\Delta^*) \leqslant \eta \langle g, \Delta^* \rangle + \alpha r(\eta\Delta^*) \\ &\leqslant \eta \langle g, \Delta^* \rangle + \eta^\theta \alpha/c \cdot r(\Delta^*) \leqslant \eta \langle g, \Delta^* \rangle + \eta/2 \cdot r(\Delta^*) \leqslant -\eta/2 \cdot r(\Delta^*), \end{aligned}$$

where the final step uses $\langle g, \Delta^* \rangle + r(\Delta^*) \leqslant 0$. Thus, $r(\Delta^*) \leqslant 2E/\eta$, as desired. □

Define $\widetilde{r}(\Delta) = \frac{2}{3\alpha} \sum_{i \in [m]} w_i D_{y_i}^{f_i}(\langle a_i, x + \Delta \rangle - b_i)$. Note that $D_{y_i}^{f_i}(\langle a_i, x + \Delta \rangle - b_i) \leqslant \alpha r_{i, y_i}(\langle a_i, \Delta \rangle)$ under the assumptions of Theorem 4.4.

Lemma 4.6. *If $\tilde{\Gamma}/2 \leqslant E \leqslant \tilde{\Gamma}$, then $-E \lesssim \left\langle \nabla F(x), \hat{\Delta} \right\rangle + \widetilde{r}(\hat{\Delta}) \leqslant -0.8E$. Also, $r(\hat{\Delta}) \asymp E$.*

Proof. Note that $C_0 s_{\min} \leqslant E \leqslant s_{\max}/C_0$ for a large constant C_0 because $\tilde{\Gamma}/2 \leqslant E \leqslant \tilde{\Gamma}$, and the choice of $s_{\min}$ and $s_{\max}$. Theorem 1.1 tells us that $\widetilde{r}(\Delta) \leqslant r(\Delta) \leqslant 2\widetilde{r}(\Delta)$ whenever $s_{\min} \leqslant r(\Delta) \leqslant s_{\max}$, with high probability. Combining $C_0 s_{\min} \leqslant E \leqslant s_{\max}/C_0$ and Lemma 4.5 gives that $s_{\min} \leqslant r(\Delta^*) \leqslant s_{\max}$. Thus, we know that $\widetilde{r}(\Delta^*) \leqslant r(\Delta^*) \leqslant 2\widetilde{r}(\Delta^*)$. By the definition of $\hat{\Delta}$ as in line 7 of Algorithm 3,

$$\begin{aligned} h(\hat{\Delta}) &\leqslant h(\Delta^*) + \tilde{\Gamma}/10 = T_x^F(x + \Delta^*) + \widetilde{r}(\Delta^*) + \tilde{\Gamma}/10 \\ &\leqslant T_x^F(x + \Delta^*) + r(\Delta^*) + \tilde{\Gamma}/10 \leqslant T_x^F(x + \Delta^*) + D_x^F(x + \Delta^*) + \tilde{\Gamma}/10 \\ &= F(x^*) - F(x) + \tilde{\Gamma}/10 \leqslant -0.8E, \end{aligned}$$

as long as $E \geqslant \tilde{\Gamma}/2$. This demonstrates the upper bound. Let us proceed to the lower bound. Let $g = \nabla F(x)$. We first consider the case where $r(\hat{\Delta}) < s_{\min}$. Let β be minimal so that $r(\beta\hat{\Delta}) \geqslant s_{\min}$. By (P4), we have that $r(\beta\hat{\Delta}) \lesssim s_{\min}$ in fact. Then

$$\left\langle g, 2\beta\hat{\Delta} \right\rangle + \widetilde{r}(2\beta\hat{\Delta}) \leqslant 2\beta \left\langle g, \hat{\Delta} \right\rangle + r(2\beta\hat{\Delta}) \leqslant 2\beta \left\langle g, \hat{\Delta} \right\rangle + O(s_{\min}) < 1.5 \left\langle g, \hat{\Delta} \right\rangle,$$

contradicting the optimality of $\hat{\Delta}$ (as we have noted that $\left\langle g, \hat{\Delta} \right\rangle \leqslant -0.8E$). If $s_{\min} \leqslant r(\hat{\Delta}) \leqslant s_{\max}$, then for $\eta = (10\alpha^2/c)^{-1/(\theta-1)}$,

$$\left\langle g, \hat{\Delta} \right\rangle + \widetilde{r}(\hat{\Delta}) \geqslant \left\langle g, \hat{\Delta} \right\rangle + \frac{1}{2\alpha} r(\hat{\Delta}) \geqslant \eta^{-1}(\left\langle g, \eta\hat{\Delta} \right\rangle + \alpha r(\eta\hat{\Delta}))$$

$$\geqslant \eta^{-1}(\left\langle g, \eta\hat{\Delta}\right\rangle + D_x^F(x+\eta\hat{\Delta})) \geqslant -\eta^{-1}E.$$

Finally, if $r(\hat{\Delta}) > s_{\max}$, first pick η to be maximal so that $r(\eta\hat{\Delta}) \leqslant s_{\max}$. It is easy to see that $r(\eta\hat{\Delta}) \gtrsim s_{\max}$. Then $\widetilde{r}(\eta\hat{\Delta}) \gtrsim s_{\max}$. Because $\widetilde{r}$ lower θ-homogeneous with constant c, we deduce that

$$\eta | \left\langle g, \hat{\Delta}\right\rangle | \geqslant \eta\widetilde{r}(\hat{\Delta}) \gtrsim \widetilde{r}(\eta\hat{\Delta}) \gtrsim s_{\max}.$$

For $\eta' = (C\alpha/c)^{-1/(\theta-1)}$ for sufficiently large C, we have that

$$-E \leqslant \left\langle g, \eta\eta'\hat{\Delta}\right\rangle + D_x^F(x+\eta\eta'\hat{\Delta}) \leqslant \eta' \left\langle g, \eta\hat{\Delta}\right\rangle + (\eta')^\theta \alpha/c \cdot r(\eta\hat{\Delta}) \lesssim -s_{\max},$$

as $r(\eta\hat{\Delta}) \leqslant s_{\max}$ and $\eta \left\langle g, \hat{\Delta}\right\rangle \lesssim -s_{\max}$. This is a contradiction.

For the second point, we have shown $s_{\min} \leqslant r(\hat{\Delta}) \leqslant s_{\max}$, so $r(\hat{\Delta}) \asymp \widetilde{r}(\hat{\Delta}) \asymp E$, as desired. □

We now prove that taking one iteration makes sufficient progress.

Lemma 4.7. *If $\tilde{\Gamma}/2 \leqslant E \leqslant \tilde{\Gamma}$, then for $\eta = (10\alpha^2/c)^{-1/(\theta-1)}$, we have $F(x+\eta\hat{\Delta}) - F(x^*) \leqslant (1-\eta/2)(F(x) - F(x^*))$.*

Proof. Note that $r(\eta\hat{\Delta}) \asymp E$ by Lemma 4.6, so we deduce that

$$\begin{aligned} F(x+\eta\hat{\Delta}) - F(x^*) &= \eta \left\langle g, \hat{\Delta}\right\rangle + D_x^F(x+\eta\hat{\Delta}) \leqslant \eta \left\langle g, \hat{\Delta}\right\rangle + \alpha r(\eta\hat{\Delta}) \leqslant \eta \left\langle g, \hat{\Delta}\right\rangle + 2\alpha^2\widetilde{r}(\eta\hat{\Delta}) \\ &\leqslant \eta(\left\langle g, \hat{\Delta}\right\rangle + \widetilde{r}(\hat{\Delta})) \leqslant -0.8\eta E, \end{aligned}$$

where we have applied Lemma 4.6 again. □

Now Theorem 4.4 follows easily.

Proof of Theorem 4.4. We first bound the runtime. We consider a single iteration of GLMIterate. One part of the runtime is from sparsifying $r(\Delta)$, and is bounded by Theorem 1.1. The other part is from calling an GLM oracle in line 7. The sparsity of $w_1, \dots, w_m$ is at most $\tilde{O}_{\alpha,c,\theta,L}(n)$ by Theorem 1.1, and $\varepsilon = \eta/(30\alpha)$, so the runtime is $\mathcal{T}_{\mathrm{GLM}}^{\mathcal{F}}(\tilde{O}_{\alpha,c,\theta,L}(n), \eta/(30\alpha))$.

Now we verify that $h(\Delta)$ takes the form of the GLM oracle as described in Definition 4.2. Indeed,

$$\begin{aligned} h(\Delta) &= T_x^F(x+\Delta) + \frac{2}{3\alpha}\sum_{i\in[m]} w_i D_{y_i}^{f_i}(\langle a_i, x+\Delta\rangle - b_i) \\ &= F(x) + \langle\nabla F(x), \Delta\rangle + \frac{2}{3\alpha}\sum_{i\in[m]} w_i(f_i(\langle a_i, x+\Delta\rangle - b_i) - f_i(\langle a_i, x\rangle - b_i) - f_i'(\langle a_i, x\rangle - b_i)\langle a_i, \Delta\rangle). \end{aligned}$$

Now define $\overline{\Delta} := x + \Delta$. Upon setting $\Delta = \overline{\Delta} - x$ in the above expression, all terms become either reweighted versions of $f_i(\langle a_i, \overline{\Delta}\rangle - b_i)$, constants, or linear terms in $\overline{\Delta}$. Thus, $h(\Delta)$ has the form of the GLM oracle.

Now we check that setting $\varepsilon = \eta/(30\alpha)$ and $x_{\mathrm{in}} = 0$ suffices to achieve $h(\hat{\Delta}) \leqslant h(\Delta^*) + \tilde{\Gamma}/10$. It suffices to verify that $\varepsilon(h(0) - h(\Delta^*)) \leqslant \tilde{\Gamma}/10$. Indeed, a calculation yields

$$h(0) - h(\Delta^*) = F(x) - (T_x^F(x^*) + D_x^F(x^*)) + (D_x^F(x^*) - \tilde{r}(\Delta^*))$$

$$= F(x) - F(x^*) + (D_x^F(x^*) - \tilde{r}(\Delta^*)) \leqslant E + D_x^F(x^*) \overset{(i)}{\leqslant} E + \alpha r(\Delta^*) \overset{(ii)}{\leqslant} E + 2\alpha E/\eta \leqslant 3\alpha\tilde{\Gamma}/\eta,$$

where (i) follows by the first bullet of Assumption 4.3, and (ii) follows from Lemma 4.5.

Now we check correctness of the overall algorithm. We show by induction that $F(x^{(T)}) - F(x^*) \leqslant \tilde{\Gamma} = (1-\eta/2)^T\Gamma$. This holds for $T = 0$ by the hypothesis in Theorem 4.4 that $F(x^{(0)}) - F^* \leqslant \Gamma$. If $F(x^{(T)}) - F(x^*) \leqslant (1-\eta/2)^{T+1}\Gamma$, then there is nothing to show because $F(x^{(T+1)}) \leqslant F(x^{(T)})$, by line 8 of Algorithm 3. Otherwise, we conclude by Lemma 4.7, whose hypothesis that $\tilde{\Gamma}/2 \leqslant E \leqslant \tilde{\Gamma}$ holds for $\tilde{\Gamma} = (1-\eta/2)^T\Gamma$. ◻

4.2 Applications

4.2.1 ℓ_p-regression

To establish Theorem 1.2, it suffices to check that the function $g(x) = |x|^p$ satisfies the conditions in Theorem 4.4.

Proof of Theorem 1.2. Consider the function $g(z) = |z|^p$ for $p > 1$. From [AKPS19a, Lemma 4.5], it holds that

$$\gamma_p(|x_0|, \Delta) \asymp_p D_{x_0}^g(x_0 + \Delta) \quad \text{for any} \quad x_0, \Delta \in \mathbb{R}\,, \tag{4.1}$$

where $\gamma_p(t, z)$ is defined in (3.16). Therefore it suffices to check that the function $r_{x_0}(\Delta) := \gamma_p(|x_0|, \Delta)$ satisfies the required conditions, and this is the content of Lemma 3.16. ◻

4.2.2 Dual of ℓ_p-regression

We now establish Theorem 1.3 by reducing the dual problem to a primal problem and applying Theorem 1.2. Consider $p \in (1, 2]$ and the dual exponent $q := \frac{p}{p-1} \geqslant 2$, along with $A \in \mathbb{R}^{m\times n}$. Then,

$$\min_{A^\top y = c} \|y\|_q = \min_{A^\top y = c} \max_{\|z\|_p \leqslant 1} z^\top y = \max_{\|z\|_p \leqslant 1} \min_{A^\top y = c} z^\top y = \max_{\|Ax\|_p \leqslant 1} c^\top x = \left(\min_{c^\top x = 1} \|Ax\|_p\right)^{-1},$$

where the second equality uses convex duality (von Neumann's minimax theorem).

Denote $x^* := \operatorname{argmin}_{c^\top x = 1} \|Ax\|_p$. By the KKT conditions, it holds that

$$y^* := \operatorname{argmin}_{A^\top y = c} \|y\|_q = \frac{\operatorname{sign}(Ax^*)|Ax^*|^{p-1}}{\|Ax^*\|_p^p},$$

where we apply scalar functions to vectors in the straightforward way, e.g., $|Ax^*|^{p-1} = \left(|(Ax^*)_1|^{p-1}, \ldots, |(Ax^*)_m|^{p-1}\right)$.

Suppose that we have a vector $\overline{x} \in \mathbb{R}^n$ satisfying $c^\top \overline{x} = 1$ and $\|A\overline{x}\|_p \leqslant (1+\varepsilon_0)\|Ax^*\|_p$, for some choice of $\varepsilon_0 > 0$ that we will make momentarily. We can find $\overline{x}$ using Theorem 1.2 with the objective $\min_{x\in\mathbb{R}^n} K|\langle c, x\rangle - 1|^p + \|Ax\|_p^p$, for sufficiently large constant K depending on ε_0, A. Define

$$\overline{y} := \frac{\operatorname{sign}(A\overline{x})|A\overline{x}|^{p-1}}{\|A\overline{x}\|_p^p}.$$

Our output y will be the orthogonal projection of $\overline{y}$ onto the affine subspace $\{y : A^\top y = c\}$, which can be found by solving a single linear system in $A^\top A$. Denote $\delta_y := y - \overline{y}$. Clearly $\|y\|_q \leqslant \|\overline{y}\|_q + \|\delta_y\|_q$. Noting that

$$\|\overline{y}\|_q = \|A\overline{x}\|_p^{-1} \leqslant \|Ax^*\|_p^{-1} = \|y^*\|_q ,$$

our goal is to bound $\|\delta_y\|_q$.

The next lemma shows that choosing $\varepsilon_0 \asymp_p (\varepsilon m^{1/q-1/2})^{2q/p}$ suffices to obtain $\|y\|_q^q \leqslant (1+\varepsilon)\|y^*\|_q^q$.

Lemma 4.8. *It holds that* $\|\delta_y\|_q \lesssim_p m^{1/2-1/q}\varepsilon_0^{p/(2q)}\|y^*\|_q$.

Proof. Write

$$\|\delta_y\|_q \leqslant \|\delta_y\|_2 \leqslant \|y^* - \overline{y}\|_2 \leqslant m^{1/2-1/q}\|y^* - \overline{y}\|_q ,$$

where the first inequality uses $q \geqslant 2$, the second uses the fact that y is the orthogonal projection onto $\{y : A^\top y = c\}$, which contains y^*, and the third uses Hölder's inequality.

To bound the last expression, first write

$$\begin{aligned}\|y^* - \overline{y}\|_q &= \left\| \frac{\operatorname{sign}(Ax^*)|Ax^*|^{p-1}}{\|Ax^*\|_p^p} - \frac{\operatorname{sign}(A\overline{x})|A\overline{x}|^{p-1}}{\|A\overline{x}\|_p^p} \right\|_q \\ &\leqslant \left| \frac{1}{\|Ax^*\|_p^p} - \frac{1}{\|A\overline{x}\|_p^p} \right| \cdot \left\|\operatorname{sign}(Ax^*)|Ax^*|^{p-1}\right\|_q + \left\| \frac{\operatorname{sign}(Ax^*)|Ax^*|^{p-1} - \operatorname{sign}(A\overline{x})|A\overline{x}|^{p-1}}{\|A\overline{x}\|_p^p} \right\|_q .\end{aligned}$$

Because $\|A\overline{x}\|_p \leqslant (1+\varepsilon_0)\|Ax^*\|_p$, the first term is bounded by

$$O(\varepsilon_0) \cdot \|Ax^*\|_p^{-p} \cdot \|Ax^*\|_p^{p/q} \lesssim \varepsilon_0 \|Ax^*\|_p^{-1} \lesssim \varepsilon_0 \|y^*\|_q .$$

For the second term, we use 2-uniform convexity (with constant $p-1$) of the ℓ_p norm for $p \in (1,2]$ to obtain

$$\|A(\overline{x} - x^*)\|_p^2 \leqslant \frac{2}{p-1}(\|A\overline{x}\|_p^2 - \|Ax^*\|_p^2) \lesssim_p \varepsilon_0 \|Ax^*\|_p^2 .$$

Because the function $f(z) := \operatorname{sign}(z)|z|^{p-1}$ satisfies $|f(y) - f(z)| \lesssim f(y-z)$ for $1 < p \leqslant 2$, we get

$$\begin{aligned}&\left\| \frac{\operatorname{sign}(Ax^*)|Ax^*|^{p-1} - \operatorname{sign}(A\overline{x})|A\overline{x}|^{p-1}}{\|A\overline{x}\|_p^p} \right\|_q \lesssim_p \|A\overline{x}\|_p^{-p} \||A(x^* - \overline{x})|^{p-1}\|_q \\ &\leqslant \|A\overline{x}\|_p^{-p}\|A(x^* - \overline{x})\|_p^{p/q} \lesssim_p \varepsilon_0^{p/(2q)}\|Ax^*\|_p^{-p}\|Ax^*\|_p^{p/q} = \varepsilon_0^{p/(2q)}\|y^*\|_q .\end{aligned}$$

□

4.2.3 γ_p regression

Consider the case when the loss functions are of the form $f_1(z) = \cdots = f_m(z) = \gamma_p(z)$, and we with to minimize $F(x) := f_1(\langle a_1, x\rangle) + \cdots + f_m(\langle a_m, x\rangle)$. For $p = 1$ (Huber regression), one can compute a $(1+\varepsilon)$-approximate solution by first sparsifying down to $\frac{n}{\varepsilon^2}(\log m)^{O(1)}$ terms using Corollary 1.6, and then solving the resulting problem using a GLM oracle for the Huber loss on sparse instances.

For $p \in (1,2]$, our framework allows us to find a minimizer to high accuracy, in analogy with the case of ℓ_p regression. This follows because the divergence of γ_p around any point is, up to constants, equal to $\gamma_p(t,z)$ for some threshold $t \geqslant 0$ (defined in (3.16)), and thus we can apply Theorem 4.4.

Lemma 4.9. *For all $p \in (1,2]$, the following holds: If $|z| \leqslant 1$ then $D_z^{\gamma_p}(\Delta + z) \asymp_p \gamma_p(\Delta)$, and if $|z| \geqslant 1$ then $D_z^{\gamma_p}(\Delta + z) \asymp \gamma_p(|z|, \Delta)$.*

Proof. We will make use of (4.1) and the following fact: For any continuously differentiable function $h : \mathbb{R} \to \mathbb{R}$ that is twice-differentiable at all but finitely many points,

$$D_z^h(z + \Delta) = h(z + \Delta) - [h(z) + h'(z)\Delta] = \int_0^{\Delta} (\Delta - t)h''(z + t)\, dt \quad \text{for all} \quad z, \Delta \in \mathbb{R}\,. \tag{4.2}$$

We now prove the lemma by case analysis on $z, \Delta \in \mathbb{R}$. By symmetry, we may assume that $z \geqslant 0$.

Case (1): $z \in [0,1], |\Delta| \leqslant 4$**.** Here we have $z + \Delta \in [-4, 5]$, and on this interval $\gamma_p''(z + \Delta) \asymp 1$. From (4.2), this yields $D_z^{\gamma_p}(z + \Delta) \asymp_p \Delta^2 \asymp \gamma_p(\Delta)$.

Case (2): $z \in [0,1], |\Delta| > 4$**.** Since $|z + \Delta| > 1$ and $p > 1$, it holds that

$$D_z^{\gamma_p}(z + \Delta) = |z + \Delta|^p - (1 - p/2) - \frac{p}{2}z^2 - pz|\Delta| \asymp_p |\Delta|^p \asymp \gamma_p(\Delta)\,.$$

Case (3): $z > 1, |z + \Delta| \geqslant 1$**.** In this case, we have $D_z^{\gamma_p}(z + \Delta) = D_z^g(z + \Delta)$, where $g(z) := |z|^p$. From (4.1), we have $D_z^g(z + \Delta) \asymp_p \gamma_p(|z|, \Delta)$.

Case (4): $z > 1, z + \Delta \in [1/2, 1]$**.** Because $\gamma_p''(y) \asymp_p g''(y)$ for all $y \in [1/2, \infty) \setminus \{1\}$, (4.2) gives $D_z^{\gamma_p}(z + \Delta) \asymp_p D_z^g(z + \Delta)$.

Case (5): $z > 1, z + \Delta \in [-1, 1/2]$**.** Convexity of γ_p implies that

$$D_z^{\gamma_p}(-1) \leqslant D_z^{\gamma_p}(z + \Delta) \leqslant D_z^{\gamma_p}(1/2).$$

We have already argued that $D_z^{\gamma_p}(-1) \asymp_p \gamma_p(|z|, -1 - z)$ and $D_z^{\gamma_p}(1/2) \asymp_p \gamma_p(|z|, 1/2 - z)$. Because $-1-z \asymp 1/2-z$, we conclude that $D_z^{\gamma_p}(-1) \asymp_p D_z^{\gamma_p}(1/2) \asymp_p \gamma_p(|z|, \Delta)$, for any $\Delta \in [-1-z, 1/2-z]$. ◻

Acknowledgments

Part of this work was conducted while the authors were visiting the Simons Institute for the Theory of Computing. James R. Lee is supported in part by NSF CCF-2007079 and a Simons Investigator Award. Yang P. Liu is partially supported by the Google Research Fellowship and NSF DMS-1926686. Aaron Sidford is supported in part by a Microsoft Research Faculty Fellowship, NSF CCF-1844855, NSF CCF-1955039, a PayPal research award, and a Sloan Research Fellowship.

References

[ABKS21] Deeksha Adil, Brian Bullins, Rasmus Kyng, and Sushant Sachdeva. Almost-linear-time weighted ℓ_p-norm solvers in slightly dense graphs via sparsification. In Nikhil Bansal, Emanuela Merelli, and James Worrell, editors, *48th International Colloquium on Automata, Languages, and Programming, ICALP 2021, July 12-16, 2021, Glasgow, Scotland (Virtual Conference)*, volume 198 of *LIPIcs*, pages 9:1–9:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021. 2, 4, 10

[AKPS19a] Deeksha Adil, Rasmus Kyng, Richard Peng, and Sushant Sachdeva. Iterative refinement for ℓ_p-norm regression. 2019. Available at arXiv:1901.06765. 40

[AKPS19b] Deeksha Adil, Rasmus Kyng, Richard Peng, and Sushant Sachdeva. Iterative refinement for ℓp-norm regression. In Timothy M. Chan, editor, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 1405–1424. SIAM, 2019. 2, 4, 10, 12

[AS20] Deeksha Adil and Sushant Sachdeva. Faster p-norm minimizing flows, via smoothed q-norm problems. In Shuchi Chawla, editor, *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*, pages 892–910. SIAM, 2020. 2

[BCLL18] Sébastien Bubeck, Michael B. Cohen, Yin Tat Lee, and Yuanzhi Li. An homotopy method for ℓ_p regression provably beyond self-concordance and in input-sparsity time. In Ilias Diakonikolas, David Kempe, and Monika Henzinger, editors, *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 1130–1137. ACM, 2018. 2, 4

[BLL+21] Jan van den Brand, Yin Tat Lee, Yang P. Liu, Thatchaphol Saranurak, Aaron Sidford, Zhao Song, and Di Wang. Minimum cost flows, mdps, and ℓ_1-regression in nearly linear time for dense instances. In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 859–869. ACM, 2021. 2

[BLM89] J. Bourgain, J. Lindenstrauss, and V. Milman. Approximation of zonoids by zonotopes. *Acta Math.*, 162(1-2):73–141, 1989. 2, 6, 9

[BLN+20] Jan van den Brand, Yin Tat Lee, Danupon Nanongkai, Richard Peng, Thatchaphol Saranurak, Aaron Sidford, Zhao Song, and Di Wang. Bipartite matching in nearly-linear time on moderately dense graphs. In Sandy Irani, editor, *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, Durham, NC, USA, November 16-19, 2020*, pages 919–930. IEEE, 2020. 2

[BLSS20] Jan van den Brand, Yin Tat Lee, Aaron Sidford, and Zhao Song. Solving tall dense linear programs in nearly linear time. In Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy, editors, *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 775–788. ACM, 2020. 2

[CLM+15] Michael B. Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform sampling for matrix approximation. In *ITCS*, pages 181–190. ACM, 2015. 18

[CLS21] Michael B. Cohen, Yin Tat Lee, and Zhao Song. Solving linear programs in the current matrix multiplication time. *J. ACM*, 68(1):3:1–3:39, 2021. 2

[CP15] Michael B. Cohen and Richard Peng. l_p row sampling by lewis weights. In Rocco A. Servedio and Ronitt Rubinfeld, editors, *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 183–192. ACM, 2015. 9, 16

[CW15] Kenneth L. Clarkson and David P. Woodruff. Input sparsity and hardness for robust subspace approximation. In Venkatesan Guruswami, editor, *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 310–329. IEEE Computer Society, 2015. 6

[CWW19] Kenneth L. Clarkson, Ruosong Wang, and David P. Woodruff. Dimensionality reduction for tukey regression. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1262–1271. PMLR, 2019. 6

[GPV21] Mehrdad Ghadiri, Richard Peng, and Santosh S. Vempala. Faster p-norm regression using sparsity. abs/2109.11537, 2021. 2, 4, 10

[JLLS23] Arun Jambulapati, James R. Lee, Yang P. Liu, and Aaron Sidford. Sparsifying sums of norms. *arXiv preprint arXiv:2305.09049*, 2023. To appear in FOCS 2023. Available at arXiv:2305.09049. 13, 14

[JLS22] Arun Jambulapati, Yang P. Liu, and Aaron Sidford. Improved iteration complexities for overconstrained p-norm regression. In Stefano Leonardi and Anupam Gupta, editors, *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022*, pages 529–542. ACM, 2022. 5

[JSWZ21] Shunhua Jiang, Zhao Song, Omri Weinstein, and Hengjie Zhang. A faster algorithm for solving general lps. In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 823–832. ACM, 2021. 2

[Lew79] D. R. Lewis. Ellipsoids defined by Banach ideal norms. *Mathematika*, 26(1):18–29, 1979. 9

[LMP13] Mu Li, Gary L Miller, and Richard Peng. Iterative row sampling. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 127–136. IEEE, 2013. 18

[LSZ19a] Yin Tat Lee, Zhao Song, and Qiuyi Zhang. Solving empirical risk minimization in the current matrix multiplication time. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 2140–2157. PMLR, 2019. 2, 4, 12

[LSZ19b] Yin Tat Lee, Zhao Song, and Qiuyi Zhang. Solving empirical risk minimization in the current matrix multiplication time. 2019. Available at arXiv:1905.04447. 37

[LT11] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. Isoperimetry and processes, Reprint of the 1991 edition. 9, 13

[MMWY22] Cameron Musco, Christopher Musco, David P. Woodruff, and Taisuke Yasuda. Active linear regression for ℓ_p norms and beyond. In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022, Denver, CO, USA, October 31 - November 3, 2022*, pages 744–753. IEEE, 2022. 2, 4, 6, 7

[PT85] Alain Pajor and Nicole Tomczak-Jaegermann. Remarques sur les nombres d'entropie d'un opérateur et de son transposé. *C. R. Acad. Sci. Paris Sér. I Math.*, 301(15):743–746, 1985. 13

[SS11] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM J. Comput.*, 40(6):1913–1926, 2011. 2, 18

[ST14] Daniel A. Spielman and Shang-Hua Teng. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM J. Matrix Anal. Appl.*, 35(3):835–885, 2014. 2

[SZ01] Gideon Schechtman and Artem Zvavitch. Embedding subspaces of L_p into l_p^N, $0 < p < 1$. *Math. Nachr.*, 227:133–142, 2001. 2, 22

[Tal90] Michel Talagrand. Embedding subspaces of l_1 into ℓ_1^n. *Proceedings of the American Mathematical Society*, 108(2):363–369, 1990. 2

[Tal95] M. Talagrand. Embedding subspaces of L_p in l_p^N. In *Geometric aspects of functional analysis (Israel, 1992–1994)*, volume 77 of *Oper. Theory Adv. Appl.*, pages 311–325. Birkhäuser, Basel, 1995. 2

[Tal14] Michel Talagrand. *Upper and lower bounds for stochastic processes*, volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer, Heidelberg, 2014. Modern methods and classical problems. 13, 14

[WY23] David P. Woodruff and Taisuke Yasuda. Sharper bounds for ℓ_p sensitivity sampling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 37238–37272. PMLR, 2023. 2