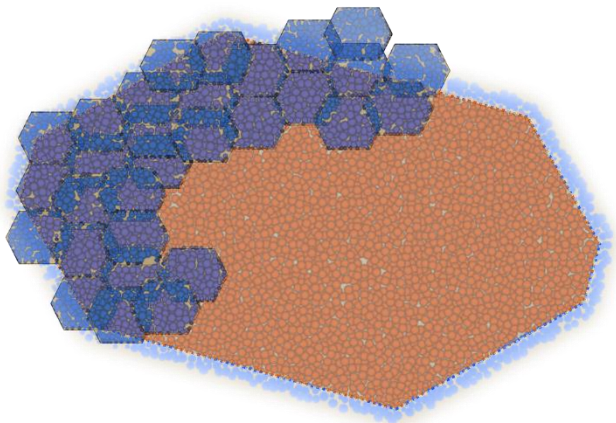
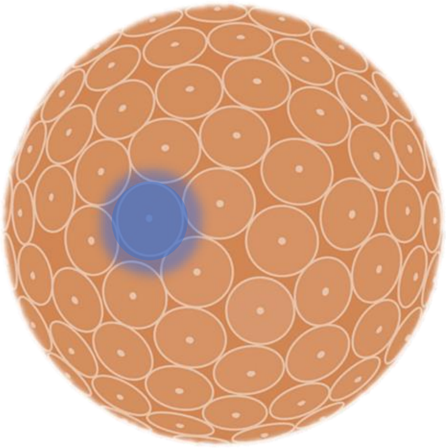
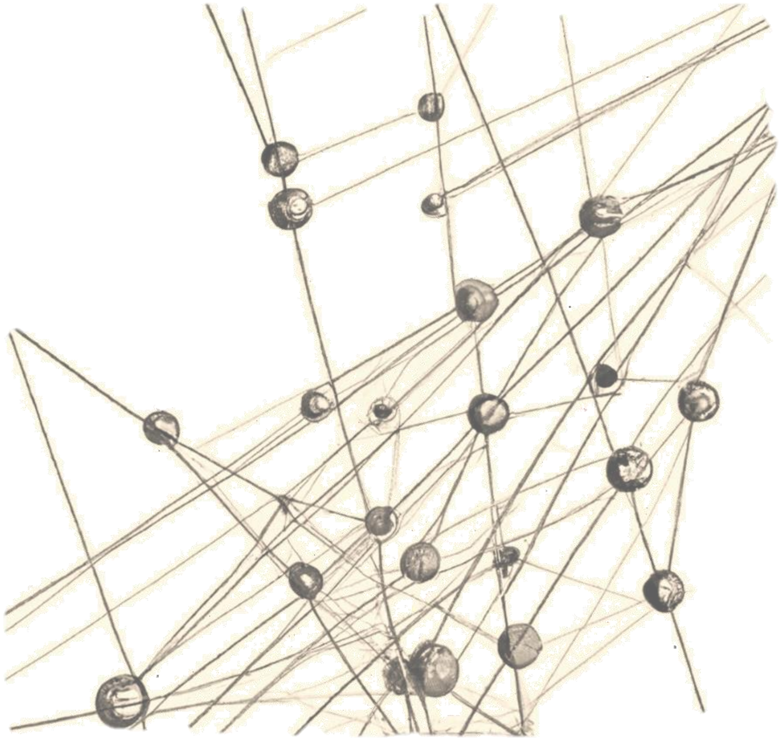
Sparsifying generalized linear models

Yang P. Liu
Institute for Advanced Study







Joint with Arun Jambulapati (Simons), James R. Lee (UW), and Aaron Sidford (Stanford)

Given functions $f_1, f_2, \dots, f_m : \mathbb{R}^n \to \mathbb{R}$, define $F : \mathbb{R}^n \to \mathbb{R}$ by

$$F(x) \coloneqq f_1(x) + f_2(x) + \cdots + f_m(x)$$

- $\tilde{F} : \mathbb{R}^n \to \mathbb{R}$ is an **ε-approximation** to F if

$$\left|F(x) - \tilde{F}(x)\right| \le \varepsilon\, F(x), \qquad \forall x \in \mathbb{R}^n$$

- $\tilde{F}$ is **s-sparse** (wrt F) if $\tilde{F} = c_1 f_1 + c_2 f_2 + \cdots + c_m f_m$ for weights $c_1, \dots, c_m \ge 0$ such that

$$\#\{\, i \in [m] : c_i \ne 0 \} \le s$$

Success:

$$s \le \frac{n}{\varepsilon^2} (\log n)^{O(1)}$$

Application: regression

Least squares regression: Given $a_1, \dots, a_m \in \mathbb{R}^n$ and $b_1, \dots, b_m \in \mathbb{R}$ with $m \gg n$, try to find $x \in \mathbb{R}^n$ such that $\langle a_i, x \rangle \approx b_i$ for every $i = 1, \dots, m$.

Minimize $\|Ax - b\|_2^2$ over $x \in \mathbb{R}^n$

$$A = \begin{pmatrix} - \; a_1 \; - \\ - \; a_2 \; - \\ \vdots \\ - \; a_m \; - \end{pmatrix} \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

$$f_i(x) := |\langle a_i, x \rangle - b_i|^2$$

$$F(x) = \|Ax - b\|_2^2 \qquad A \in \mathbb{R}^{m \times n}$$

$$\tilde{F}(x) = \|SAx - Sb\|_2^2 \qquad SA \in \mathbb{R}^{s \times n}$$

$$S = \begin{pmatrix} c_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & c_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & c_5 \end{pmatrix}$$

$$\left| \|Ax - b\|_2^2 - \|SAx - Sb\|_2^2 \right| \le \varepsilon \|Ax - b\|_2^2, \qquad \forall x \in \mathbb{R}^n$$

Application: graph sparsification

Sparsification of graphs: Given a weighted undirected graph $G = (V, E, w)$, find a graph $\tilde{G} = (V, \tilde{E}, \widetilde{w})$ with $\tilde{E} \subseteq E$ such that $|\tilde{E}| \leq s$, and:

Cut sparsifiers [Benczur-Karger]: $f_{uv}(x) = w_{uv}|x_u - x_v|$

$$\sum_{uv \in E} w_{uv} \, |x_u - x_v| = (1 \pm \varepsilon) \sum_{uv \in \tilde{E}} \widetilde{w}_{uv} |x_u - x_v| , \qquad \forall x \in \mathbb{R}^n$$

Spectral sparsifiers [Spielman-Teng]: $f_{uv}(x) = w_{uv}(x_u - x_v)^2$

$$\sum_{uv \in E} w_{uv} \, (x_u - x_v)^2 = (1 \pm \varepsilon) \sum_{uv \in \tilde{E}} \widetilde{w}_{uv} (x_u - x_v)^2 , \qquad \forall x \in \mathbb{R}^n$$

$F(x) := f_1(x) + f_2(x) + \cdots + f_m(x)$

Generalized linear models

$$f_i(x) = \varphi(\langle a_i, x\rangle - b_i), \quad i = 1, \dots, m$$

$\varphi(y) = \lvert y\rvert^2$	$s \lesssim n/\varepsilon^2$	[Batson-Spielman-Srivastava 2014]
$\varphi(y) = \lvert y\rvert$	$s \lesssim n \log n / \varepsilon^2$	[Talagrand 1991]
$\varphi(y) = \lvert y\rvert^p$, $1 < p < 2$	$s \lesssim n \log n\, (\log\log n)^2/\varepsilon^2$	[Talagrand 1995]
$\varphi(y) = \lvert y\rvert^p$, $0 < p < 1$	$s \lesssim n(\log n)^3/\varepsilon^2$	[Schechtman-Zvavich 2001]
$\varphi(y) = \lvert y\rvert^p$, $p > 2$	$s \lesssim n^{p/2}/\varepsilon^2$	[Bourgain-Lindenstrauss-Milman 89, Ledoux-Talagrand 91]

$F(x) \coloneqq f_1(x) + f_2(x) + \cdots + f_m(x)$

Generalized linear models

$$f_i(x) = \varphi(\langle a_i, x\rangle - b_i), \quad i = 1, \dots, m$$

$$\varphi(y) \approx \min(|y|, |y|^2)$$

Huber loss

$$\gamma_p(y) \approx \min(|y|^p, |y|^2)$$

$0 < p \leq 2$

[Bubeck-Cohen-Lee-Li 2018]

$$s \lesssim n^{1.17}/\varepsilon^2$$

[Musco-Musco-Woodruff-Yasuda 2022]

$$\varphi(y) = \min(1, |y|^2)$$

Tukey loss

$F(x) := f_1(x) + f_2(x) + \cdots + f_m(x)$

Generalized linear models

$$f_i(x) = \varphi(\langle a_i, x\rangle - b_i), \quad i = 1, \dots, m$$

$$\varphi(y) \approx \min(|y|, |y|^2)$$

Huber loss

$$\gamma_p(y) \approx \min(|y|^p, |y|^2)$$

$0 < p \leq 2$

[Bubeck-Cohen-Lee-Li 2018]

$s \lesssim n^{1.17}/\varepsilon^2$

[Musco-Musco-Woodruff-Yasuda 2022]

$$\varphi(y) = \max(0, |y| - 0.1)$$

ReLU requires $s \geq 2^{\Omega(n)}$

$F(x) := f_1(x) + f_2(x) + \cdots + f_m(x)$

Generalized linear models

$$f_i(x) = \varphi(\langle a_i, x\rangle - b_i), \quad i = 1, \dots, m$$

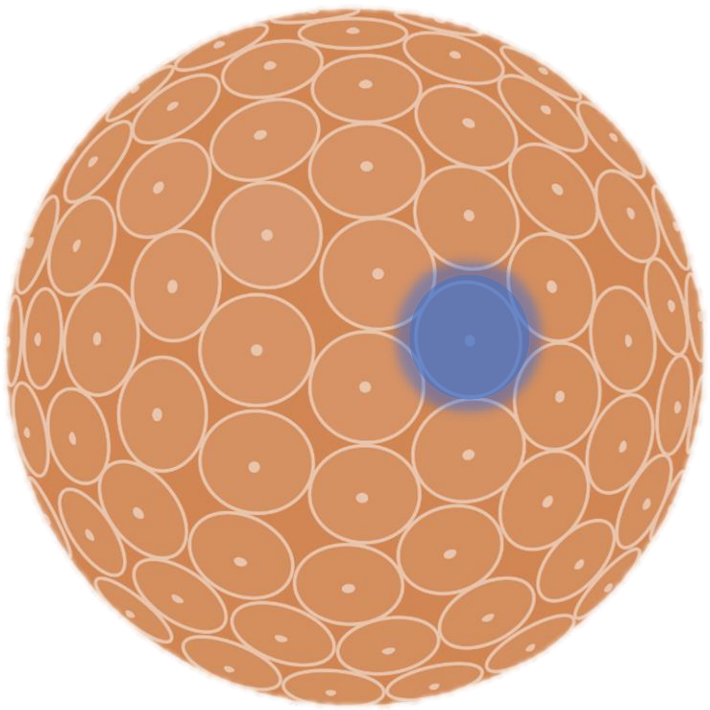
$$\varphi(y) \approx \min(|y|, |y|^2)$$

Huber loss

$$\gamma_p(y) \approx \min(|y|^p, |y|^2)$$

$0 < p \leq 2$

[Bubeck-Cohen-Lee-Li 2018]



$$\varphi(y) = \max(0, |y| - 0.1)$$

ReLU requires $s \geq 2^{\Omega(n)}$

$a_1, \dots, a_m \in \mathbb{S}^{n-1}$

$\{\, x \in \mathbb{S}^{n-1}: \langle a_i, x\rangle > 0.1 \,\}$ pairwise disjoint

$$f_i(x) = \varphi(\langle a_i, x\rangle) > 0 \Leftrightarrow \langle a_i, x\rangle > 0.1$$

Motivation: ℓ_p-regression

Minimize $\|Ax - b\|_p^p$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{\boldsymbol{m} \times n}$, $1 < p \leq 2$

$F(x) := f_1(x) + f_2(x) + \cdots + f_m(x)$

Motivation: ℓ_p-regression

Minimize $\|Ax - b\|_p^p$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{\boldsymbol{m} \times n}$, $1 < p \leq 2$

When $p = 2$:

$$\left| \|Ax - b\|_2^2 - \|SAx - Sb\|_2^2 \right| \leq \varepsilon \|Ax - b\|_2^2, \qquad \forall x \in \mathbb{R}^n$$

$SA \in \mathbb{R}^{\boldsymbol{s} \times n}$, $\boldsymbol{s} \lesssim \tilde{O}(n/\varepsilon^2)$

$F(x) := f_1(x) + f_2(x) + \cdots + f_m(x)$

Motivation: ℓ_p-regression

Minimize $\|Ax - b\|_p^p$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{\boldsymbol{m} \times n}$, $1 < p \leq 2$

When $p = 2$:

$$\left| \|Ax - b\|_2^2 - \|SAx - Sb\|_2^2 \right| \leq \varepsilon \|Ax - b\|_2^2, \qquad \forall x \in \mathbb{R}^n$$

$SA \in \mathbb{R}^{\boldsymbol{s} \times n}$, $\boldsymbol{s} \lesssim \tilde{O}(n/\varepsilon^2)$

Seems: $(1 + \varepsilon)$-approximate least squares regression requires runtime/samples proportional to ε^{-2}?

$F(x) := f_1(x) + f_2(x) + \cdots + f_m(x)$

Basic iterative refinement

Based on work of [Adil-Kyng-Peng-Sachdeva 2019]

Minimize $\|Ax - b\|_2^2$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$

Basic iterative refinement

Minimize $\|Ax - b\|_2^2$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{\boldsymbol{m} \times n}$

Basic iterative refinement

Minimize $\|Ax - b\|_2^2$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$

$$\|A(x_0 + \Delta) - b\|_2^2 - \|Ax_0 - b\|_2^2 = \langle g, \Delta \rangle + \|A\Delta\|_2^2$$
$$g = 2(Ax_0 - b)$$

Basic iterative refinement

Minimize $\|Ax - b\|_2^2$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$

$$\|A(x_0 + \Delta) - b\|_2^2 - \|Ax_0 - b\|_2^2 = \langle g, \Delta \rangle + \|A\Delta\|_2^2$$
$$g = 2(Ax_0 - b)$$

Minimizing $\langle g, \Delta \rangle + \|A\Delta\|_2^2$ to a factor 2 error reduces function error by 1/2

$$\|A(x_0 + \Delta) - b\|_2^2 - \|Ax^* - b\|_2^2 \leq \frac{1}{2}(\|Ax_0 - b\|_2^2 - \|Ax^* - b\|_2^2)$$

Basic iterative refinement

Minimize $\|Ax - b\|_2^2$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{\boldsymbol{m} \times n}$

$$\|A(x_0 + \Delta) - b\|_2^2 - \|Ax_0 - b\|_2^2 = \langle g, \Delta\rangle + \|A\Delta\|_2^2$$
$$g = 2(Ax_0 - b)$$

Minimizing $\langle g, \Delta\rangle + \|A\Delta\|_2^2$ to a factor 2 error reduces function error by 1/2

$$\|A(x_0 + \Delta) - b\|_2^2 - \|Ax^* - b\|_2^2 \leq \frac{1}{2}\left(\|Ax_0 - b\|_2^2 - \|Ax^* - b\|_2^2\right)$$

Plan: sparsify $\|A\Delta\|_2^2 \approx_2 \|SA\Delta\|_2^2$ to take step. $\boldsymbol{s} \lesssim \tilde{O}(n)$

Repeat $O(\log(1/\varepsilon))$ times to get $(1+\varepsilon)$-approximate solution (high accuracy!)

Minimize $\|Ax - b\|_p^p$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $1 < p \leq 2$

Minimize $\sum_{i=1}^{m} f_i(\langle a_i, x\rangle - b_i)$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $f_i(z) = |z|^p$

Minimize $\|Ax - b\|_p^p$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $1 < p \leq 2$

Minimize $\sum_{i=1}^m f_i(\langle a_i, x\rangle - b_i)$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $f_i(z) = |z|^p$

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

$$= \langle g, \Delta\rangle + \sum_{i=1}^m D^{f_i}_{\langle a_i, x_0\rangle - b_i}(\langle a_i, x_0 + \Delta\rangle - b_i)$$

g is the gradient

$D_y^f(z) = f(z) - f(y) - f'(y)(z - y)$ is the *Bregman divergence*

Minimize $\|Ax - b\|_p^p$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $1 < p \leq 2$

Minimize $\sum_{i=1}^m f_i(\langle a_i, x\rangle - b_i)$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $f_i(z) = |z|^p$

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

$$= \langle g, \Delta\rangle + \sum_{i=1}^m D^{f_i}_{\langle a_i, x_0\rangle - b_i}(\langle a_i, x_0 + \Delta\rangle - b_i)$$

g is the gradient

$D_y^f(z) = f(z) - f(y) - f'(y)(z - y)$ is the *Bregman divergence*

$$D_x^{f_i}(x + \delta) \approx \gamma_p^{|x|}(\delta) \approx \min\{x^{p-2}\delta^2, |\delta|^p\}$$

$F(x) := f_1(x) + f_2(x) + \cdots + f_m(x)$

Iterative refinement

Minimize $\|Ax - b\|_p^p$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $1 < p \leq 2$

Minimize $\sum_{i=1}^m f_i(\langle a_i, x\rangle - b_i)$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $f_i(z) = |z|^p$

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

$$\approx \langle g, \Delta\rangle + \sum_{i=1}^m \gamma_p^{|\langle a_i, x_0\rangle - b_i|}(\langle a_i, \Delta\rangle)$$

g is the gradient

$D_y^f(z) = f(z) - f(y) - f'(y)(z - y)$ is the *Bregman divergence*

$$D_x^{f_i}(x + \delta) \approx \gamma_p^{|x|}(\delta) \approx \min\{x^{p-2}\delta^2, |\delta|^p\}$$

$F(x) := f_1(x) + f_2(x) + \cdots + f_m(x)$

Iterative refinement

Minimize $\|Ax - b\|_p^p$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $1 < p \leq 2$

Minimize $\sum_{i=1}^m f_i(\langle a_i, x\rangle - b_i)$ over $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $f_i(z) = |z|^p$

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

$$\approx \langle g, \Delta\rangle + \sum_{i=1}^m c_i \gamma_p^{|\langle a_i, x_0\rangle - b_i|}(\langle a_i, \Delta\rangle)$$

c is s-sparse

g is the gradient

$D_y^f(z) = f(z) - f(y) - f'(y)(z - y)$ is the *Bregman divergence*

$$D_x^{f_i}(x + \delta) \approx \gamma_p^{|x|}(\delta) \approx \min\{x^{p-2}\delta^2, |\delta|^p\}$$

Theorem [JLLS 2023]: Suppose $h_1, \dots, h_m : \mathbb{R} \to \mathbb{R}$ satisfy

- $|h_i(u) - h_i(v)| \lesssim h_i(u - v)$ for all $u, v \in \mathbb{R}$
- $h_i(\lambda u) \gtrsim \lambda^\theta h_i(u)$ for some $\theta > 0$ and all $u \in \mathbb{R}, \lambda \geq 1$

Then for any $a_1, \dots, a_m \in \mathbb{R}^n$, the function

$$F(x) = h_1(\langle a_1, x\rangle)^2 + \cdots + h_m(\langle a_m, x\rangle)^2$$

admits an s-sparse $\tilde{F}$ such that

$$\left|F(x) - \tilde{F}(x)\right| \leq \varepsilon F(x) \quad \text{for all} \quad \alpha \leq F(x) \leq \beta$$

with $s \lesssim \frac{n}{\varepsilon^2} (\log n)^3 \log \frac{n\beta}{\alpha}$

Running time:

$$\tilde{O}(\mathrm{nnz}(A) + n^\omega + mT_{\mathrm{eval}}) \log \frac{n\beta}{\alpha}$$

Theorem [JLLS 2023]: Suppose $h_1, \dots, h_m : \mathbb{R} \to \mathbb{R}$ satisfy

- $|h_i(u) - h_i(v)| \lesssim h_i(u - v)$ for all $u, v \in \mathbb{R}$
- $h_i(\lambda u) \gtrsim \lambda^\theta h_i(u)$ for some $\theta > 0$ and all $u \in \mathbb{R}, \lambda \geq 1$

Then for any $a_1, \dots, a_m \in \mathbb{R}^n$, the function

$$F(x) = h_1(\langle a_1, x\rangle)^2 + \cdots + h_m(\langle a_m, x\rangle)^2$$

admits an s-sparse $\tilde{F}$ such that

$$\left|F(x) - \tilde{F}(x)\right| \leq \varepsilon\, F(x) \quad \text{for all} \quad \alpha \leq F(x) \leq \beta$$

with $s \lesssim \frac{n}{\varepsilon^2} (\log n)^3 \log \frac{n\beta}{\alpha}$

Applies to:

$h_i(u)^2 = |u|^{p_i}$, or

$h_i(u)^2 = \min\{|u|^2, |u|^{p_i}\}$

$0 < p_1, \dots, p_m \leq 2$

Theorem [JLLS 2023]: Suppose $h_1, \dots, h_m : \mathbb{R} \to \mathbb{R}$ satisfy

- $|h_i(u) - h_i(v)| \lesssim h_i(u - v)$ for all $u, v \in \mathbb{R}$
- $h_i(\lambda u) \gtrsim \lambda^\theta h_i(u)$ for some $\theta > 0$ and all $u \in \mathbb{R}, \lambda \geq 1$

Then for any $a_1, \dots, a_m \in \mathbb{R}^n$, the function

$$F(x) = h_1(\langle a_1, x\rangle)^2 + \cdots + h_m(\langle a_m, x\rangle)^2$$

admits an s-sparse $\tilde{F}$ such that

$$\left|F(x) - \tilde{F}(x)\right| \leq \varepsilon\, F(x) \quad \text{for all } \alpha \leq F(x) \leq \beta$$

with $s \lesssim \frac{n}{\varepsilon^2} (\log n)^3 \log \frac{n\beta}{\alpha}$

Applies to:

$h_i(u)^2 = |u|^{p_i}$, or

$h_i(u)^2 = \min\{|u|^2, |u|^{p_i}\}$

$0 < p_1, \dots, p_m \leq 2$

Algorithms for ℓ_p regression, $1 < p \leq 2$

Can solve in time $\tilde{O}(\mathrm{nnz}(A) + n^\omega)$

Intuition for properties

(1) $|h_i(u) - h_i(v)| \lesssim h_i(u - v)$ for all $u, v \in \mathbb{R}$

(2) $h_i(\lambda u) \gtrsim \lambda^\theta h_i(u)$ for some $\theta > 0$ and all $u \in \mathbb{R}, \lambda \geq 1$

$f_i(z) = h_i(z)^2$

Intuition for properties

(1) $|h_i(u) - h_i(v)| \lesssim h_i(u - v)$ for all $u, v \in \mathbb{R}$

(2) $h_i(\lambda u) \gtrsim \lambda^\theta h_i(u)$ for some $\theta > 0$ and all $u \in \mathbb{R}, \lambda \geq 1$

(3) $h_i(u) \approx h_i(-u)$ for all $u \in \mathbb{R}$

$f_i(z) = h_i(z)^2$

Intuition for properties

(1) $|h_i(u) - h_i(v)| \lesssim h_i(u - v)$ for all $u, v \in \mathbb{R}$

(2) $h_i(\lambda u) \gtrsim \lambda^\theta h_i(u)$ for some $\theta > 0$ and all $u \in \mathbb{R}, \lambda \geq 1$

(3) $h_i(u) \approx h_i(-u)$ for all $u \in \mathbb{R}$

(4) $h_i(\lambda u) \lesssim \lambda h_i(u)$, so $f_i(\lambda u) \lesssim \lambda^2 f_i(u)$

Intuition for properties

(1) $|h_i(u) - h_i(v)| \lesssim h_i(u - v)$ for all $u, v \in \mathbb{R}$

(2) $h_i(\lambda u) \gtrsim \lambda^\theta h_i(u)$ for some $\theta > 0$ and all $u \in \mathbb{R}, \lambda \geq 1$

(3) $h_i(u) \approx h_i(-u)$ for all $u \in \mathbb{R}$

(4) $h_i(\lambda u) \lesssim \lambda h_i(u)$, so $f_i(\lambda u) \lesssim \lambda^2 f_i(u)$

(4) says f_i grows subquadratically.

Essential for sparsification: $s \lesssim n^{p/2}/\varepsilon^2$ when $f_i(z) = |z|^p,\ p > 2$

$f_i(z) = h_i(z)^2$

Intuition for properties

(1) $|h_i(u) - h_i(v)| \lesssim h_i(u - v)$ for all $u, v \in \mathbb{R}$

(2) $h_i(\lambda u) \gtrsim \lambda^\theta h_i(u)$ for some $\theta > 0$ and all $u \in \mathbb{R}, \lambda \geq 1$

(3) $h_i(u) \approx h_i(-u)$ for all $u \in \mathbb{R}$

(4) $h_i(\lambda u) \lesssim \lambda h_i(u)$, so $f_i(\lambda u) \lesssim \lambda^2 f_i(u)$

(4) says f_i grows subquadratically.

Essential for sparsification: $s \lesssim n^{p/2}/\varepsilon^2$ when $f_i(z) = |z|^p,\ p > 2$

(2) does not hold for Tukey loss: $f(z) = \min\{1, z^2\}$

By "smoothing" to $\tilde{f}(z) = \min\{|z|^\delta, z^2\}$, and $\delta \to 0$, show $s \lesssim n^{1+o(1)}/\varepsilon^2$

Comparison to [MMWY22]

(1) $|h_i(u) - h_i(v)| \lesssim h_i(u - v)$ for all $u, v \in \mathbb{R}$

(2) $h_i(\lambda u) \gtrsim \lambda^\theta h_i(u)$ for some $\theta > 0$ and all $u \in \mathbb{R}, \lambda \geq 1$

(3) $h_i(u) \approx h_i(-u)$ for all $u \in \mathbb{R}$

(4) $h_i(\lambda u) \lesssim \lambda h_i(u)$, so $f_i(\lambda u) \lesssim \lambda^2 f_i(u)$

$f_i(z) = h_i(z)^2$

Comparison to [MMWY22]

(1) $|h_i(u) - h_i(v)| \lesssim h_i(u - v)$ for all $u, v \in \mathbb{R}$

(2) $h_i(\lambda u) \gtrsim \lambda^\theta h_i(u)$ for some $\theta > 0$ and all $u \in \mathbb{R}, \lambda \geq 1$

(3) $h_i(u) \approx h_i(-u)$ for all $u \in \mathbb{R}$

(4) $h_i(\lambda u) \lesssim \lambda h_i(u)$, so $f_i(\lambda u) \lesssim \lambda^2 f_i(u)$

In [MMWY22], (1) is replaced with $h_i(v + w) \leq h_i(v) + h_i(w)$.

(1) Is a bit weaker: consider $w = u - v$.

$f_i(z) = h_i(z)^2$

Comparison to [MMWY22]

(1) $|h_i(u) - h_i(v)| \lesssim h_i(u - v)$ for all $u, v \in \mathbb{R}$

(2) $h_i(\lambda u) \gtrsim \lambda^\theta h_i(u)$ for some $\theta > 0$ and all $u \in \mathbb{R}, \lambda \geq 1$

(3) $h_i(u) \approx h_i(-u)$ for all $u \in \mathbb{R}$

(4) $h_i(\lambda u) \lesssim \lambda h_i(u)$, so $f_i(\lambda u) \lesssim \lambda^2 f_i(u)$

In [MMWY22], (1) is replaced with $h_i(v + w) \leq h_i(v) + h_i(w)$.

(1) Is a bit weaker: consider $w = u - v$.

[MMWY22] requires that $f_1 = f_2 = \cdots = f_m$. Our theorem does not.

Given functions $f_1, f_2, \dots, f_m : \mathbb{R}^n \to \mathbb{R}$, define $F : \mathbb{R}^n \to \mathbb{R}$ by

$$F(x) \coloneqq f_1(x) + f_2(x) + \cdots + f_m(x)$$

Let $\rho = (\rho_1, \dots, \rho_m) \in \mathbb{R}^m_{++}$ be a probability distribution: $\rho_1 + \cdots + \rho_m = 1$

Algorithm: Sample indices $\boldsymbol{\nu_1}, \boldsymbol{\nu_2}, \dots, \boldsymbol{\nu_s} \in \{1, \dots, m\}$ i.i.d. from ρ

And define: $\tilde{F}(x) \coloneqq \frac{1}{s}\left(\frac{f_{\boldsymbol{\nu_1}}(x)}{\rho_{\boldsymbol{\nu_1}}} + \cdots + \frac{f_{\boldsymbol{\nu_s}}(x)}{\rho_{\boldsymbol{\nu_s}}}\right)$

$$\mathbb{E}[\tilde{F}(x)] = \mathbb{E}\left[\frac{f_{\boldsymbol{\nu_1}}(x)}{\rho_{\boldsymbol{\nu_1}}}\right] = \sum_{i=1}^{m} \rho_i \cdot \frac{f_i(x)}{\rho_i} = F(x), \qquad \forall x \in \mathbb{R}^n$$

Given functions $f_1, f_2, \dots, f_m : \mathbb{R}^n \to \mathbb{R}$, define $F : \mathbb{R}^n \to \mathbb{R}$ by

$$F(x) \coloneqq f_1(x) + f_2(x) + \cdots + f_m(x)$$

Let $\rho = (\rho_1, \dots, \rho_m) \in \mathbb{R}^m_{++}$ be a probability distribution: $\rho_1 + \cdots + \rho_m = 1$

Algorithm: Sample indices $\boldsymbol{\nu_1}, \boldsymbol{\nu_2}, \dots, \boldsymbol{\nu_s} \in \{1, \dots, m\}$ i.i.d. from ρ

And define: $\tilde{F}(x) \coloneqq \frac{1}{s}\left(\frac{f_{\boldsymbol{\nu_1}}(x)}{\rho_{\boldsymbol{\nu_1}}} + \cdots + \frac{f_{\boldsymbol{\nu_s}}(x)}{\rho_{\boldsymbol{\nu_s}}}\right)$

Need to establish: $\mathbb{E} \max\limits_{F(x) \le \lambda} \left|F(x) - \tilde{F}(x)\right| \le \varepsilon\lambda$ for s chosen large enough

Sampling weights

Main challenge: non-homogeneity of f_i

Main challenge: non-homogeneity of f_i

When $f_i(z) = |z|^p$ are homogeneous: Lewis weights

$$w_i = w_i^{1-2/p} a_i^\top \left(A^\top W^{1-2/p} A\right)^{-1} a_i$$

Main challenge: non-homogeneity of f_i

When $f_i(z) = |z|^p$ are homogeneous: Lewis weights

$$w_i = w_i^{1-2/p} a_i^\top \left(A^\top W^{1-2/p} A\right)^{-1} a_i$$

$$w_1 + \cdots + w_m = n, \qquad \rho_i = w_i / n$$

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

Sampling weights

Main challenge: non-homogeneity of f_i

When $f_i(z) = |z|^p$ are homogeneous: Lewis weights

$$w_i = w_i^{1-2/p} a_i^\top \left(A^\top W^{1-2/p} A\right)^{-1} a_i$$

$$w_1 + \cdots + w_m = n, \qquad \rho_i = w_i/n$$

Fix a scale $\lambda \in \mathbb{R}_+$, intuitively handles $\{x \ : \ F(x) \in [\lambda/2, \lambda]\}$

(Approximate weight) $\dfrac{f_i\left(\sqrt{a_i^\top M^{-1} a_i}\right)}{d_i a_i^\top M^{-1} a_i} \approx \lambda, \ \ M = \sum_{i=1}^m d_i a_i a_i^\top$ for $i = 1, \ldots, m.$

Main challenge: non-homogeneity of f_i

When $f_i(z) = |z|^p$ are homogeneous: Lewis weights

$$w_i = w_i^{1-2/p} a_i^\top \left(A^\top W^{1-2/p} A\right)^{-1} a_i$$

$$w_1 + \cdots + w_m = n, \qquad \rho_i = w_i/n$$

Fix a scale $\lambda \in \mathbb{R}_+$, intuitively handles $\{x \;:\; F(x) \in [\lambda/2, \lambda]\}$

(Approximate weight) $\dfrac{f_i\left(\sqrt{a_i^\top M^{-1} a_i}\right)}{d_i a_i^\top M^{-1} a_i} \approx \lambda, \;\; M = \sum_{i=1}^m d_i a_i a_i^\top$ for $i = 1, \ldots, m$.

$\rho_i = d_i a_i^\top M^{-1} a_i / n$ for $i = 1, \ldots, m$.

Repeat for all $\lambda = 2^k, \;\; \alpha/m^{O(1)} \le \lambda \le \beta$

Existence of weights

Fix a scale $\lambda \in \mathbb{R}_+$, intuitively handles $\{x \ : \ F(x) \in [\lambda/2, \lambda]\}$

(Approximate weight) $\dfrac{f_i\left(\sqrt{a_i^\top M^{-1} a_i}\right)}{d_i a_i^\top M^{-1} a_i} \approx \lambda, \;\; M = \sum_{i=1}^m d_i a_i a_i^\top$ for $i = 1, \dots, m$.

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

Existence of weights

Fix a scale $\lambda \in \mathbb{R}_+$, intuitively handles $\{x : F(x) \in [\lambda/2, \lambda]\}$

(Approximate weight) $\frac{f_i\left(\sqrt{a_i^\top M^{-1} a_i}\right)}{d_i a_i^\top M^{-1} a_i} \approx \lambda,\ \ M = \sum_{i=1}^m d_i a_i a_i^\top$ for $i = 1, \ldots, m$.

Use a contractive map / algorithm [Cohen-Peng 2015].

$$\psi(d) := \frac{1}{\lambda} \frac{f_i\left(\sqrt{a_i^\top M_d^{-1} a_i}\right)}{a_i^\top M_d^{-1} a_i},\ \ M_d := \sum_{i=1}^m d_i a_i a_i^\top \text{ for } i = 1, \ldots, m.$$

Existence of weights

Fix a scale $\lambda \in \mathbb{R}_+$, intuitively handles $\{x \ : \ F(x) \in [\lambda/2, \lambda]\}$

(Approximate weight) $\dfrac{f_i\left(\sqrt{a_i^\top M^{-1} a_i}\right)}{d_i a_i^\top M^{-1} a_i} \approx \lambda, \ \ M = \sum_{i=1}^m d_i a_i a_i^\top$ for $i = 1, \ldots, m.$

Use a contractive map / algorithm [Cohen-Peng 2015].

$$\psi(d) := \frac{1}{\lambda} \frac{f_i\left(\sqrt{a_i^\top M_d^{-1} a_i}\right)}{a_i^\top M_d^{-1} a_i}, \ \ M_d := \sum_{i=1}^m d_i a_i a_i^\top \ \text{ for } \ i = 1, \ldots, m.$$

$\psi(d) \approx d$ is equivalent to d being approximate weight

$$\psi(d) := \frac{1}{\lambda}\frac{f_i\left(\sqrt{a_i^\top M_d^{-1} a_i}\right)}{a_i^\top M_d^{-1} a_i}, \quad M_d := \sum_{i=1}^m d_i a_i a_i^\top \quad \text{for} \quad i = 1, \ldots, m.$$

$\psi(d) \approx d$ is equivalent to d being approximate weight

Lemma: If $d \approx_\gamma d'$ then $\psi(d) \approx_{\gamma^C} \psi(d')$ for $C = 1 - \frac{\theta}{2} < 1$

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

Existence of weights

$$\psi(d) := \frac{1}{\lambda}\frac{f_i\left(\sqrt{a_i^\top M_d^{-1} a_i}\right)}{a_i^\top M_d^{-1} a_i}, \quad M_d := \sum_{i=1}^m d_i a_i a_i^\top \quad \text{for} \quad i = 1, \dots, m.$$

$\psi(d) \approx d$ is equivalent to d being approximate weight

Lemma: If $d \approx_\gamma d'$ then $\psi(d) \approx_{\gamma^C} \psi(d')$ for $C = 1 - \frac{\theta}{2} < 1$

Constructing approximate weight:

$$d_t := \psi^{(t)}(d_0)$$

$$d_1 = \psi(d_0) \approx_\gamma d_0$$

$$d_2 = \psi(d_1) \approx_{\gamma^C} \psi(d_0) = d_1$$

$$\dots$$

$$d_{T+1} = \psi(d_T) \approx_{\gamma^{C^T}} \psi(d_{T-1}) = d_T$$

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

Existence of weights

$$\psi(d) := \frac{1}{\lambda} \frac{f_i\left(\sqrt{a_i^\top M_d^{-1} a_i}\right)}{a_i^\top M_d^{-1} a_i}, \quad M_d := \sum_{i=1}^m d_i a_i a_i^\top \quad \text{for} \quad i = 1, \ldots, m.$$

$\psi(d) \approx d$ is equivalent to d being approximate weight

Lemma: If $d \approx_\gamma d'$ then $\psi(d) \approx_{\gamma^C} \psi(d')$ for $C = 1 - \frac{\theta}{2} < 1$

Constructing approximate weight:

$$d_t := \psi^{(t)}(d_0)$$

$$d_1 = \psi(d_0) \approx_\gamma d_0$$

$$d_2 = \psi(d_1) \approx_{\gamma^C} \psi(d_0) = d_1$$

$$\ldots$$

$$d_{T+1} = \psi(d_T) \approx_{\gamma^{C^T}} \psi(d_{T-1}) = d_T$$

Take $T \to \infty$

Showing concentration

And define: $\tilde{F}(x) := \frac{1}{s}\left(\frac{f_{\boldsymbol{v}_1}(x)}{\rho_{\boldsymbol{v}_1}} + \cdots + \frac{f_{\boldsymbol{v}_s}(x)}{\rho_{\boldsymbol{v}_s}}\right)$

Need to establish: $\mathbb{E} \max_{F(x) \leq \lambda} \left|F(x) - \tilde{F}(x)\right| \leq \varepsilon\lambda$ for s chosen large enough

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

Showing concentration

And define: $\tilde{F}(x) := \frac{1}{s}\left(\frac{f_{\boldsymbol{v}_1}(x)}{\rho_{\boldsymbol{v}_1}} + \cdots + \frac{f_{\boldsymbol{v}_s}(x)}{\rho_{\boldsymbol{v}_s}}\right)$

Need to establish: $\mathbb{E} \max_{F(x) \leq \lambda} \left|F(x) - \tilde{F}(x)\right| \leq \varepsilon\lambda$ for s chosen large enough

Chaining: efficient way to union bound over all $x \in \mathbb{R}^n$

- Definition of *approximate weight* was exactly chosen to make a chaining proof work
- Critically use $|h_i(u) - h_i(v)| \lesssim h_i(u - v)$ for all $u, v \in \mathbb{R}$

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

Showing concentration

And define: $\tilde{F}(x) := \frac{1}{s}\left(\frac{f_{\boldsymbol{v}_1}(x)}{\rho_{\boldsymbol{v}_1}} + \cdots + \frac{f_{\boldsymbol{v}_s}(x)}{\rho_{\boldsymbol{v}_s}}\right)$

Need to establish: $\mathbb{E} \max_{F(x) \le \lambda} \left|F(x) - \tilde{F}(x)\right| \le \varepsilon\lambda$ for s chosen large enough

Chaining: efficient way to union bound over all $x \in \mathbb{R}^n$

- Definition of *approximate weight* was exactly chosen to make a chaining proof work
- Critically use $|h_i(u) - h_i(v)| \lesssim h_i(u - v)$ for all $u, v \in \mathbb{R}$

Weight schemes: relating weights between adjacent scales $\lambda, \lambda/2, \lambda/4, \ldots$

- $d^{(\lambda)}$: approximate weight at scale λ
- Chaining proof requires that $d^{(\lambda)} \approx d^{(\lambda/2)}$ for all λ
- True by the contractive proof

Summary

- Sparsifying $F(x) \coloneqq f_1(\langle a_i, x\rangle) + \cdots + f_m(\langle a_m, x\rangle)$ down to $\tilde{O}(n/\varepsilon^2)$ terms
- Natural assumptions on f_i: auto-Lipschitz + lower-growth

Summary

- Sparsifying $F(x) \coloneqq f_1(\langle a_i, x\rangle) + \cdots + f_m(\langle a_m, x\rangle)$ down to $\tilde{O}(n/\varepsilon^2)$ terms
- Natural assumptions on f_i: auto-Lipschitz + lower-growth
- Definition of approximate weights at each level of scale
- Existence of weights via contractive algorithm

Summary

- Sparsifying $F(x) \coloneqq f_1(\langle a_i, x\rangle) + \cdots + f_m(\langle a_m, x\rangle)$ down to $\tilde{O}(n/\varepsilon^2)$ terms
- Natural assumptions on f_i: auto-Lipschitz + lower-growth
- Definition of approximate weights at each level of scale
- Existence of weights via contractive algorithm
- Analyze sparsification via chaining
- Requires weight schemes: relations between weights at consecutive scales

Open problems / Future directions

- $p > 2$? (1) -> $|h_i(u) - h_i(v)| \lesssim h_i(u-v)$ for all $u, v \in \mathbb{R}$, for $h_i(u) \coloneqq f_i(u)^{1/p}$
  - Goal: sparsity $\tilde{O}(n^{p/2}/\varepsilon^2)$
  - We can show approximate weights exist. Weight schemes unclear.

Open problems / Future directions

- $p > 2$? (1) -> $|h_i(u) - h_i(v)| \lesssim h_i(u - v)$ for all $u, v \in \mathbb{R}$, for $h_i(u) \coloneqq f_i(u)^{1/p}$
  - Goal: sparsity $\tilde{O}(n^{p/2}/\varepsilon^2)$
  - We can show approximate weights exist. Weight schemes unclear.
- Higher dimensional functions?
  - $F(x) \coloneqq f_1(A_1 x) + \cdots + f_m(A_m x)$ where $A_i \in \mathbb{R}^{d \times n}$.
  - Properties? Natural goal: $\tilde{O}(dn/\varepsilon^2)$? Even $\tilde{O}(n/\varepsilon^2)$?

Open problems / Future directions

- $p > 2$? (1) -> $|h_i(u) - h_i(v)| \lesssim h_i(u - v)$ for all $u, v \in \mathbb{R}$, for $h_i(u) \coloneqq f_i(u)^{1/p}$
  - Goal: sparsity $\tilde{O}(n^{p/2}/\varepsilon^2)$
  - We can show approximate weights exist. Weight schemes unclear.
- Higher dimensional functions?
  - $F(x) \coloneqq f_1(A_1 x) + \cdots + f_m(A_m x)$ where $A_i \in \mathbb{R}^{d \times n}$.
  - Properties? Natural goal: $\tilde{O}(dn/\varepsilon^2)$? Even $\tilde{O}(n/\varepsilon^2)$?
- Other sparsification beyond norms? Coresets for clustering?