A ℓ_p sparsification

Our approach is based on techniques for extracting low-dimensional sketches of small subspaces in high dimensions. The usual ℓ_2 norm uses much simpler underlying techniques, and we describe it first. The extension to ℓ_p norms for $p \neq 2$ is obtained via *Lewis weights* [Lewis, 1978].

A.1 Euclidean sparsifiers

The kind of sketches we need originate in the work of Batson et al. [2012]. Specifically, it will be convenient to start from the following variant due to Boutsidis et al. [2014]:

Lemma 10 (BSS weights [Boutsidis et al., 2014]) *Let $[u] \in \mathbb{R}^{d\times t}$ $(t < d)$ be the matrix with rows $\vec{u}_1, \ldots, \vec{u}_d$ such that $\sum_{i=1}^d u_i u_i^\top = I_t$. Then given an integer $r \in (t, d]$, there exist $s_1, \ldots, s_d \geq 0$ such that at most r of the s_i are nonzero and for the $d \times r$ matrix $[s]$ with ith column $\sqrt{s_i}\vec{e}_i$,*

$$\lambda_t([u]^\top[s][s]^\top[u]) \geq (1 - \sqrt{t/r})^2 \text{ and}$$
$$\lambda_1([u]^\top[s][s]^\top[u]) \leq (1 + \sqrt{t/r})^2$$

where λ_j denotes the jth largest eigenvalue.

In particular, taking $r = t/\gamma^2$ for some $\gamma \in (0, 1)$, we obtain that for the $[s]$ guaranteed to exist by Lemma 10, $\|[s]^\top[u]\vec{v}\|_2^2 = ([s]^\top[u]\vec{v})^\top([s]^\top[u]\vec{v})$ for any $\vec{v}$, and hence by Lemma 10, $(1-\gamma)\|\vec{v}\|_2 \leq \|[s][u]\vec{v}\|_2 \leq (1+\gamma)\|\vec{v}\|_2$. Furthermore, we can bound the magnitude of the entries of $[s]$ for orthonormal $[u]$ as follows:

Lemma 11 *Suppose the rows of $[u]$ are orthonormal. Then the matrix $[s]$ obtained by Lemma 10 has entries of magnitude at most $(1 + \sqrt{t/r})\sqrt{d}$.*

Proof: Observe that since the each ith row of $[u]$ has unit norm, it must have an entry u_{i,j^*} that is at least $1/\sqrt{d}$ in magnitude. By the above argument,

$$\|[s]^\top[u]\vec{e}_{j^*}\|_2^2 \leq (1 + \sqrt{t/r})^2\|\vec{e}_{j^*}\|_2^2 = (1 + \sqrt{t/r})^2$$

where notice in particular, the ith row of $[s]^\top$ contributes at least $(\sqrt{s_i}u_{i,j^*})^2 \geq s_i/d$ to the norm. Thus, $s_i \leq (1 + \sqrt{t/r})^2 d$. ■

A.2 Sparsifiers for non-Euclidean norms

It is possible to obtain an analogue of the BSS weights for $p \neq 2$ using techniques based on *Lewis weights* [Lewis, 1978]. Lewis weights are a general way to reduce problems involving ℓ_p norms to analogous ℓ_2 computations. Cohen and Peng [2015] applied this to sparsification to obtain the following family of sparsifiers:

Theorem 12 (ℓ_p weights [Cohen and Peng, 2015]) *Given a $d \times t$ matrix $[u]$ there exists a set of $r(p, t, \gamma)$ weights $s_1, \ldots, s_r$ such that for the $d \times r$ matrix $[s]$ which has as its ith column $s_i\vec{e}_i$,*

$$(1 - \gamma)\|[u]\vec{v}\|_p \leq \|[s]^\top[u]\vec{v}\|_p \leq (1 + \gamma)\|[u]\vec{v}\|_p$$

where $r(p, t, \gamma)$ is asymptotically bounded as in Table 4.

Table 4: Dimension required for $(1 \pm \gamma)$-approximate ℓ_p sparsification of t-dimensional subspaces. $p = 2$ uses BSS weights. (Table 1)

p	Required dimension r
$p = 1$	$\frac{t \log t}{\gamma^2}$
$1 < p < 2$	$\frac{1}{\gamma^2} t \log(t/\gamma) \log^2 \log(t/\gamma)$
$p = 2$	t/γ^2
$p > 2$	$\frac{\log 1/\gamma}{\gamma^5} t^{p/2} \log t$

Cohen and Peng also show how to construct the sparsifiers for a given matrix efficiently, but we won't be able to make use of this, since we will be searching for the sparsifier for an unknown subset of the rows.

We furthermore obtain an analogue of Lemma 11 for the ℓ_p weights, using essentially the same argument:

Lemma 13 *Suppose the rows of $[u]$ are orthonormal. Then the matrix $[s]$ obtained by Theorem 12 has entries of magnitude at most $(1 + \gamma)\sqrt{d}$.*

B Analysis of the algorithms

We now give the full analysis of our weighting algorithms. For convenience, we will recall the entire algorithm and the statements of the theorems. In the following, let $\Pi_{d_1,\ldots,d_s}$ denote the projection to coordinates $d_1, d_2, \ldots, d_s$.

In our analysis of this algorithm, we will find it convenient to use the Rademacher generalization bounds for linear predictors (note that $x \mapsto |x|^p$ is pb^{p-1}-Lipschitz on $[-b, b]$):

Theorem 14 (Bartlett et al. 2002, Kakade et al. 2009) *For $b > 0$, $p \geq 1$, random variables $(\vec{Y}, Z)$ distributed over $\{\vec{y} \in \mathbb{R}^d : \|y\|_2 \leq b\} \times [b, b]$, and any $\delta \in (0, 1)$, let $L_p(\vec{a})$ denote $\mathbb{E}[|\langle\vec{a}, \vec{Y}\rangle - Z|^p]$, and for an an i.i.d. sample of size m let $\hat{L}_p(\vec{a})$ be the empirical loss $\frac{1}{m}\sum_{j=1}^m |\langle\vec{a}, \vec{y}^{(j)}\rangle - z^{(j)}|^p$. We then have that with probability $1 - \delta$ for all $\vec{a}$ with $\|a\|_2 \leq b$,*

$$|L_p(\vec{a}) - \hat{L}_p(\vec{a})| \leq \frac{2pb^{p+1}}{\sqrt{m}} + b^p\sqrt{\frac{2\ln(4/\delta)}{m}}.$$

Note that although this bound is stated in terms of the ℓ_2 norm of the attribute and parameter vectors $\vec{y}$ and $\vec{a}$, we can obtain a bound in terms of the dimension s of the sparse rule if we are given a bound B on the magnitude of the entries: $b \leq \sqrt{s}B$.

input : Examples $(\vec{x}^{(1)}, \vec{y}^{(1)}, z^{(1)}), \ldots, (\vec{x}^{(m)}, \vec{y}^{(m)}, z^{(m)})$, target loss bound ϵ and fraction μ.
output : A k-DNF over $x_1, \ldots, x_n$ and linear predictor over $y_1, \ldots, y_d$, or INFEASIBLE if none exist.
subroutines : WtCond takes as inputs examples $(\vec{x}^{(1)}, \ldots, \vec{x}^{(m)})$, nonnegative weights $(w^{(1)}, \ldots, w^{(m)})$, and a bound μ, and returns a k-DNF $\hat{c}$ over $x_1, \ldots, x_n$ solving the weighted conditional distribution search task.

begin
  Let $m_0 = \lceil \frac{1}{\mu}(\frac{b^{2p}}{\gamma^{2p}\epsilon^{2p}}(2pb + \sqrt{2\ln(12/\delta)})^2 + \ln\frac{3}{\delta})\rceil$, r is as given in Table 4.
  forall $(d_1, \ldots, d_s) \in \binom{[d]}{s}$, $(q_1, \ldots, q_r) \in \{-\lceil\frac{1}{\gamma}(\ln r - \frac{1}{p}\ln\gamma)\rceil, \ldots, 0, \ldots, \lceil\ln(s+1)/2\gamma\rceil\}$ *and* $(j_1, \ldots, j_r) \in \binom{[m_0]}{r}$ **do**
    Let $\vec{a}$ be a solution to the following convex optimization problem: minimize $\sum_{\ell=1}^{r}((1+\gamma)^{q_\ell}(\langle\vec{a}, \Pi_{d_1,\ldots,d_s}\vec{y}^{(j_\ell)}\rangle - z^{(j_\ell)})^p$ subject to $\|\vec{a}\|_p \leq b$.
    Put $c \leftarrow$ the output of WtCond on $(\vec{x}^{(1)}, \ldots, \vec{x}^{(m)})$ with the weights $w^{(i)} = |\langle\vec{a}, \Pi_{d_1,\ldots,d_s}\vec{y}^{(i)}\rangle - z^{(i)}|^p$ and bound μ.
    if WtCond *did not return INFEASIBLE and* $\hat{\mathbb{E}}[(\langle\vec{a}, \Pi_{d_1,\ldots,d_s}\vec{Y}\rangle - Z)^p | c(\vec{X})]^{1/p} \leq \alpha\epsilon$ **then return** $\vec{a}$ *and* c.
  end
  return *INFEASIBLE.*
end

Algorithm 3: Weighted Sparse Regression (Algorithm 1)

Theorem 15 (Theorem 6) *For any constant s and $\gamma > 0$, r as given in Table 4 for $t = (s+1)$, and $m = m_0 + \Theta\left(\frac{((1+\gamma)b)^3}{\mu\epsilon\eta^2}(n^k + s\log d + r\log\frac{m_0\log(\gamma^{1/p}s/r)}{\gamma} + \log\frac{1}{\delta})\right)$ examples, Algorithm 3 runs in polynomial time and solves the conditional s-sparse ℓ_p regression task with $\alpha = \tilde{O}((1+\gamma)\sqrt{n^k}(\log b + \log\frac{1}{\eta} + \log\log\frac{1}{\delta})\epsilon)$.*

Proof: Given that we are directly checking the empirical ℓ_p loss before returning $\vec{a}$ and c, for the quoted number of examples m it is immediate by a union bound over the iterations that any $\vec{a}$ and c we return are satisfactory with probability $1-\delta$. All that needs to be shown is that the algorithm will find a pair that passes this final check.

By Theorem 14, we note that it suffices to have $\frac{b^{2p}}{\gamma^{2p}\epsilon^{2p}}(2pb + \sqrt{2\ln(12/\delta)})^2$ examples from the distribution conditioned on the unknown k-DNF event c^* to obtain that the ℓ_p loss of each candidate for $\vec{a}$ is estimated to within an additive $\gamma\epsilon$ with probability $1 - \delta/3$. By Hoeffding's inequality, therefore when we draw m_0 examples, there is a sufficiently large subset satisfying c^* with probability $1 - \delta/3$.

We let $[u]$ be an orthonormal basis for $\text{span}\{(\Pi_{d_1,\ldots,d_s}\vec{y}^{(j)}, z^{(j)}) : c^*(\vec{x}^{(j)}) = 1, j \leq m_0\}$ and invoke Lemma 10 for ℓ_2 or Theorem 12 for $p \neq 2$. In either case, there is some set of weights $s_1, \ldots, s_{r_0}$ for a subset of r_0 coordinates $j_1, \ldots, j_{r_0}$ such that for any $\vec{v}$ in the column span of $[u]$, $[s]^\top[u]\vec{v}$ has ℓ_p norm that is a $1 \pm \gamma$-approximation to the ℓ_p norm of $[u]\vec{v}$. In particular, for any $\vec{a}$, observing $\vec{v} = [y]\vec{a} - \vec{z}$ is in the column span of $[u]$ by construction, we obtain

$$(1-\gamma)\|[y]\vec{a}-\vec{z}\|_p \leq \|[s]^\top([y]\vec{a}-\vec{z})\|_p \leq (1+\gamma)\|[y]\vec{a}-\vec{z}\|_p.$$

Now, we observe that we can discard weights (and dimensions) from $s_1, \ldots, s_{r_0}$ of magnitude smaller than $\gamma/r_0^{1/p}$, since for any unit vector $\vec{v}$, the contribution of such entries to $\|[s]^\top[u]\vec{v}\|_p^p$ (recalling there are at most r_0 nonzero entries) is at most γ^p. So we may assume the $r \leq r_0$ remaining weights all have magnitude at least $\gamma/r^{1/p}$. Furthermore, if we round each weight to the nearest power of $(1+\gamma)$, this only changes $\|[s]^\top[u]\vec{v}|_p^p$ by an additional $(1 \pm \gamma)$ factor. Finally, we note that since $(\Pi_{d_1,\ldots,d_s}\vec{y}^{(j)}, z^{(j)})$ has dimension $s+1$, Lemmas 11 and 13 guarantee that the magnitude is also at most $(1+\gamma)\sqrt{s+1}$. Thus it indeed suffices to find the powers $(q_1, \ldots, q_r)$ for our r examples $j_1, \ldots, j_r$ such that $(1+\gamma)^{q_\ell}$ is within $(1+\gamma)$ of s_ℓ, and the resulting set of weights will approximate the ℓ_p-norm of every $\vec{v}$ in the column span to within a $1+3\gamma$-factor.

Now, when the loop in Algorithm 3 considers (i) the dimensions $d_1^*, \ldots, d_s^*$ contained in the optimal s-sparse regression rule $\vec{a}^*$ (ii) the set of examples $j_1^*, \ldots, j_r^*$ used for the sparse approximation for these coordinates and (iii) the appropriate weights $(1+\gamma)^{q_1^*}, \ldots, (1+\gamma)^{q_r^*}$, the algorithm will obtain a vector $\vec{a}$ that achieves a $(1+3\gamma)$-approximation to the empirical ℓ_p-loss of $\vec{a}^*$ on the same s coordinates.

It then follows from Theorem 3 that with probability at least $1 - \delta/3$ over the data, WtCond will in turn return to us a k-DNF c with probability $(1-\eta)\mu$ that selects a subset of the data on which $\vec{a}$ achieves an $\alpha\epsilon = \tilde{O}((1+\gamma)\sqrt{n^k}(\log b + \log 1/\eta + \log\log 1/\delta)\epsilon)$ approximation to the empirical ℓ_p loss of $\vec{a}^*$ on c^*. This choice of $\vec{a}$ and c passes the final check and is thus sufficient. ■

The extension to reference class ℓ_p-norm regression proceeds by replacing the weighted condition search algorithm with a variant of the tolerant elimination

algorithm from Juba [2016], given in Algorithm 4.

input : Examples $(\vec{x}^{(1)}, w^{(1)}), \ldots, (\vec{x}^{(m)}, w^{(m)})$, query point $\vec{x}^*$, minimum fraction μ_0, minimum loss target ϵ_0, approximation parameter η.
output : A k-DNF over $x_1, \ldots, x_n$.
begin
  Initialize $\mu \leftarrow 1$, $\hat{c} \leftarrow \bot$, $\hat{\epsilon} \leftarrow \max_j w^{(j)}$
  while $\mu \geq \mu_0$ **do**
    Initialize $\epsilon \leftarrow \hat{\epsilon}$
    while $\epsilon \geq \epsilon_0/(1+\eta)$ **do**
      Initialize c to be the empty disjunction
      forall *Terms T of at most k literals* **do**
        if $\sum_{j:T(\vec{x}^{(j)})=1} w^{(j)} \leq \epsilon\mu m$ **then** Add T to c.
      end
      Put $\epsilon \leftarrow \epsilon/(1+\eta)$
    end
    if $c(\vec{x}^*) = 1$, $\sum_{j=1}^m c(\vec{x}^{(j)}) \geq \mu m$, *and* $\epsilon < \hat{\epsilon}$ **then** Put $\hat{c} \leftarrow c$, $\hat{\epsilon} \leftarrow \epsilon$
    Put $\mu \leftarrow \mu/(1+\eta)$
  end
  return $\hat{c}$
end

Algorithm 4: Reference Class Search (Algorithm 2)

Lemma 16 (Lemma 8) *If* $m \geq \Omega(\frac{b^3}{\eta^2(\epsilon_0+\epsilon^*)\mu_0}(k\log n + \log\frac{1}{\eta\delta} + \log\log\frac{1}{\mu_0} + \log\log\frac{b}{\epsilon_0}))$ *where* $W \in [0,b]$, *then Algorithm 4 returns a k-DNF $\hat{c}$ such that with probability $1-\delta$, 1. $\hat{c}(\vec{x}^*) = 1$ 2. $\Pr[\hat{c}(\vec{X})] \geq \mu_0/(1+\eta)$ 3. $\mathbb{E}[W|\hat{c}(\vec{X})] \leq O((1+\eta)^4 n^k(\epsilon_0+\epsilon^*))$ where ϵ^* is the minimum $\mathbb{E}[W|c^*(\vec{X})]$ over k-DNF c^* such that $c^*(\vec{x}^*) = 1$ and $\Pr[c^*(\vec{X})] \geq (1+\eta)\mu_0$.*

Proof: For convenience, let $N \leq (1+\log_{1+\eta} b/\epsilon_0)(1+\log_{1+\eta} 1/\mu_0)$ denote the total number of iterations. Consider first what happens when the loop considers the largest $\mu \leq \Pr[c^*(\vec{X})]/(1+\eta)$ and the smallest ϵ that is at least $(1+\eta)^2\epsilon^*$. On this iteration, for each term T of c^*, we observe that $\mathbb{E}[W \cdot T(\vec{X})] \leq \epsilon^* \Pr[c^*(\vec{X})]$—indeed,

$$\mathbb{E}[W \cdot T(\vec{X})] \leq \mathbb{E}[W \cdot c^*(\vec{X})] \leq \epsilon^* \Pr[c^*(\vec{X})].$$

So, since W is bounded by b, by a Chernoff bound, $\frac{1}{m}\sum_{j:T(x^{(j)})=1} w^{(j)} \leq (1+\eta)\epsilon^* \Pr[c^*(\vec{X})]$ with probability $1 - \frac{\delta}{2\binom{n}{\leq k}N}$. Since this is in turn at most $\epsilon\mu$, T will be included in c on this iteration. But similarly, for T not in c with $\mathbb{E}[W \cdot T(\vec{X})] > (1+\eta)\mu\epsilon$, the Chernoff bound also yields that $\sum_{j:T(x^{(j)})=1} w^{(j)} \geq \epsilon\mu m$ with probability $1 - \delta/2\binom{n}{\leq k}N$. By a union bound over all $T \in c^*$ and T not in c^* with such large error, we see that with probability at least $1 - \delta/2N$, all of the terms of c^* are included in c and only terms with $\mathbb{E}[W \cdot T(\vec{X})] \leq (1+\eta)\mu\epsilon$ are included in c. So,

$$\begin{aligned}\mathbb{E}[W \cdot c(\vec{X})] &\leq \sum_{T \text{ in } c} \mathbb{E}[W \cdot T(\vec{X})] \\ &\leq O((1+\eta)n^k\mu\epsilon).\end{aligned}$$

Furthermore, by yet another application of a Chernoff bound, c^* is true of at least μm examples with probability at least $1 - \delta/2N$. Thus, with probability $1 - \delta/N$, after this iteration $\hat{c}$ is set to some k-DNF and $\hat{\epsilon} \leq (1+\eta)^2 \max\{\epsilon^*, \epsilon_0\}$.

Now, furthermore, on every iteration, we see more generally that with probability $1-\delta/N$, only terms with $\mathbb{E}[W \cdot T(\vec{X})] \leq (1+\eta)\mu\epsilon$ are included in c, and $\hat{c}$ is only updated if $c(\vec{x}^*) = 1$ and $\Pr[c(\vec{X})] \geq \mu/(1+\eta)$, where $\mu \geq \mu_0$. Thus, for the $\hat{c}$ we return, since $\mathbb{E}[W \cdot c(\vec{X})] \leq O((1+\eta)n^k\mu\epsilon)$ and $\mathbb{E}[W|c(\vec{X})] = \mathbb{E}[W \cdot c(\vec{X})]/\Pr[c(\vec{X})]$, $\mathbb{E}[W|\hat{c}(\vec{X})] \leq O((1+\eta)^2 n^k\hat{\epsilon})$. Thus, with probability $1-\delta$ overall, since we found above that $\hat{\epsilon} \leq (1+\eta)^2 \max\{\epsilon^*, \epsilon_0\}$, we return a k-DNF $\hat{c}$ as claimed. ■

Now, as noted above, our algorithm for reference class regression is obtained essentially by substituting Algorithm 4 for the subroutine WtCond in Algorithm 3; the analysis, similarly, substitutes the guarantee of Lemma 8 for Theorem 3. In summary, we find

Theorem 17 *For any constant s and $\gamma > 0$, r as given in Table 4 for $t = (s+1)$, and*

$$m \geq m_0 + \Omega\left(\frac{(1+\gamma)^3 b^3}{\eta^2(\epsilon_0+\epsilon^*)\mu_0}\left(r\log\frac{m_0\log(\gamma^{1/p}s/r)}{\gamma} + \log\left(\frac{n^k d^s}{\eta\delta}\log\frac{1}{\mu_0}\log\frac{(1+\gamma)b}{\epsilon_0}\right)\right)\right)$$

examples, our modified algorithm runs in polynomial time and solves the reference class s-sparse ℓ_p regression task with $\alpha = O((1+\gamma)(1+\eta)^4 n^k)$.