A. Preliminaries

Bernstein's inequality will be used multiple times:

Proposition A.1. *[Bernstein's Inequality](Maurer, 2003) Let $\{X_i\}$ be independent random variables with $\mathbb{E}(X_i^2) < \infty$ and $X_i \geq 0$. Set $X = \sum_i X_i$ and $\lambda > 0$. Then,*

$$Pr[X \leq \mathbb{E}(X) - \lambda] \leq \exp\left(\frac{-\lambda^2}{2\sum_i \mathbb{E}(X_i^2)}\right).$$

If $X_i - \mathbb{E}(X_i) \leq \Delta$ for all i, then with $\sigma_i^2 = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2$ we have

$$Pr[X \geq \mathbb{E}(X) + \lambda] \leq \exp\left(\frac{-\lambda^2}{2\sum_i \sigma_i^2 + \lambda\Delta}\right).$$

B. Omitted details from Section 3

For technical reasons we make the following assumptions on the parameters:

Assumption B.1. *We assume that the following inequalities hold for N, m, b, δ:*

$$\begin{aligned}
&m\varepsilon^2 \geq \max\left\{-4\ln(\delta),\ 3\ln\left(\beta m \log_2\left(\frac{m}{\varepsilon}\right)\right),\right.\\
&\left.2\ln\left(\log_2\left(\frac{n}{m\varepsilon}\right)+1\right), 2d\ln\left(1+\frac{n}{m\varepsilon}\right) - \ln(\delta)\right\}\\
&b \geq \max\left\{m,\ \delta^{-1}\right\}\\
&N = \Omega(bm^2d^2\varepsilon^{-1}\delta^{-1})
\end{aligned}$$

B.1. Contraction bounds

We set $Q_1 = \{q \in \{1, \ldots, q_{\max}\} \mid |W_q^+| \geq \beta m \wedge \|W_q^+\|_1 \geq \frac{\varepsilon}{q_{\max}}\}$ and $Q_2 = \{q \leq \log_2(\frac{m}{\varepsilon})\} \setminus Q_1$. We set $Q^* = Q_1 \cup Q_2$ to be the set of important weight classes. The following lemma shows that the weight of the remaining weight classes is negligible:

Lemma B.1. *It holds that $\sum_{q\in Q^*} \|W_q^+\|_1 \geq (1 - 5\varepsilon')G^+(z)$.*

Proof. The total weight W of those classes with $q \notin Q_1 \cup Q_2$ is at most

$$\begin{aligned}
W &\leq \sum_{q\leq q_{\max}, q\notin Q^*} \|W_q^+\|_1 \ + \sum_{q>q_{\max}} \|W_q^+\|_1\\
&\leq \frac{\varepsilon}{m}\beta m \sum_{q=0}^{\infty} 2^{-q} \ + \ 2^{-q_{\max}} n\\
&= 4\varepsilon + \frac{\varepsilon}{n} n = 5\varepsilon
\end{aligned}$$

as $\beta \leq 2$. Recall that $G^+(z) \geq \frac{1}{\mu'}$. Combining these two facts gives us

$$\begin{aligned}
\sum_{q\in Q^*} \|W_q^+\|_1 &= G^+(z) - \sum_{q\notin Q^*} \|W_q^+\|_1\\
&\geq G^+(z) - 5\varepsilon\\
&\geq G^+(z) - 5G^+(z)\mu'\varepsilon\\
&= G^+(z) - 5\varepsilon' G^+(z).
\end{aligned}$$

□

In the following we show that the important weight classes have at least the same contribution, up to a small error, after sketching. First we consider the weight classes with $q \in Q_2$, where the individual entries have a notable contribution themselves. Then we consider the weight classes with $q \in Q_1$ which consist of a large number of entries. In both cases we will show that for each important weight class W_q^+ there exists a subset $W_q^* \subset W_q^+$ with $\|W_q^*\|_1 \geq (1 - 7\varepsilon)\frac{\|W_q^+\|_1}{b^{h(q)}\beta}$ and where each entry $z_p \in W_q^+$ is much larger than the sum of all other entries in its bucket.

Heavy-hitters. This section is dedicated to showing that the large entries of z are well separated among the buckets and that they contribute about the same value after sketching. For $A \in \mathbb{R}^{n\times d}$, let $u \in \mathbb{R}^n$ denote the ℓ_1-leverage score vector of A, i.e.,

$$u_i = \max_{x\in\mathbb{R}^d\setminus\{0\}} \frac{|(Ax)_i|}{\|Ax\|_1}.$$

We start by showing that the rows of A with the largest ℓ_1-leverage scores are well separated and that only coordinates z_p with high ℓ_1-leverage can be large coordinates of z. To this end we need two lemmas:

Lemma B.2. *(Clarkson & Woodruff, 2015) For N_1, N_2 with $N_2 \geq N_1$ and with $N_1N_2 \leq \kappa N$, for $\kappa \in (0, 1/2)$, let Y_1 and Y_2 denote the sets of indices of the N_1 and N_2 coordinates with the largest leverage scores, so that $Y_1 \subset Y_2$. Then with probability at least $1 - 2\kappa$ each member of Y_1 is in a bucket containing no other member of Y_2.*

Lemma B.3. *If u_p is the k-th largest coordinate of u, then for z in the subspace spanned by the columns of A it holds that $|z_p| \leq \frac{d}{k}G(z)$.*

Proof. By (Dasgupta et al., 2009) there exists a so-called *Auerbach* basis Q of A with the following properties. It holds that $G(Qx) = \|Qx\|_1 \geq \|x\|_\infty$ for all $x \in \mathbb{R}^d$ and $\sum_{ij} |Q_{ij}| \leq d$. Note that by a change of basis

$$u_i = \max_{x\in\mathbb{R}^d\setminus\{0\}} \frac{|(Ax)_i|}{\|Ax\|_1} = \max_{x\in\mathbb{R}^d\setminus\{0\}} \frac{|(Qx)_i|}{\|Qx\|_1}.$$

Thus $|z_i| = |Q_i x| \leq \|Q_i\|_1\|x\|_\infty \leq \|Q_i\|_1\|Qx\|_1$ and it follows that $\sum_i u_i \leq \sum_i \|Q_i\|_1 = \sum_{ij} |Q_{ij}| \leq d$. Consequently the k-th largest coordinate of z can be at most $|z_p| \leq u_p G(Qx) \leq \frac{d}{k}G(Qx) = \frac{d}{k}G(z)$. □

We apply Lemma B.2 and Lemma B.3 as follows: set $N_1 = \frac{d\beta m}{\varepsilon}$ and $N_2 = d\beta m^2$, and let Y_1 (resp. Y_2) be the set of coordinates with the N_1 (resp. N_2) largest leverage scores. We denote by $\mathcal{E}_1$ the event that all coordinates in Y_1 are in a bucket with no other member of Y_2. By Lemma B.2 and Assumption B.1, $\mathcal{E}_1$ holds with probability at least $1-\delta$. Then by Lemma B.3, for any entry $z_p \geq v_1 := \frac{\varepsilon}{\beta m}$ we have $p \in Y_1$ and for any entry $p \in Y_2$ we have $z_p < \frac{1}{\beta m^2} = \frac{\varepsilon}{\beta m} \cdot \frac{\varepsilon}{m\varepsilon^2} = \frac{\varepsilon}{m\varepsilon^2} v_1$. It remains to show that the remaining entries in the buckets containing a heavy hitter only have a small contribution. To show this we use Bernstein's inequality. For a coordinate $p \in [n]$ we denote by B_p the bucket that contains p.

Lemma B.4. *Assume $\mathcal{E}_1$ holds. Then with failure probability at most $e^{-m\varepsilon^2}$ for any p with $z_p \geq v_1$ and $z_p \in W_q$ for some $q \in Q_2$, we have $\|B_p \setminus \{p\}\|_1 \leq 3\varepsilon|z_p|$.*

Proof. Let p with $z_p \geq v_1$. Then by Lemma B.3 it holds that $p \in Y_1$. For each $i \in [n] \setminus \{p\}$ we define a random variable X_i by $X_i = z_i$ if i is put in bucket B_p and $X_i = 0$ otherwise. By our assumption, $X_i = 0$ if $i \in Y_2$. Otherwise we have $P(X_i \neq 0) \leq \frac{1}{\beta N}$ since the probability that any coordinate is put in any bucket is at most $\frac{1}{\beta N}$. Further we have $\mathbb{E}(\|B_p \setminus \{p\}\|_1) \leq \frac{1}{\beta N}$ since $G(z) = 1$. For the variance we have

$$\begin{aligned}\sigma_i^2 &= (z_i - \mathbb{E}(X_i))^2 P(X_i = z_i) + \mathbb{E}(X_i)^2 P(X_i = 0) \\ &\leq z_i^2 \frac{1}{\beta N} + \frac{z_i^2}{(\beta N)^2} < \frac{2z_i^2}{\beta N}\end{aligned}$$

Since $z_i \leq \frac{\varepsilon}{m\varepsilon^2} v_1$ for $i \notin Y_2$ this implies

$$\begin{aligned}\sum_{i\in[n]} \sigma_i^2 = \sum_{i\notin Y_2} \sigma_i^2 &< \sum_{i\notin Y_2} \frac{2z_i^2}{\beta N} \\ &< \frac{\varepsilon}{m\varepsilon^2 \beta N} v_1 \sum_{i\in[n]} z_i \\ &= \frac{\varepsilon}{m\varepsilon^2 \beta N} v_1 \\ &\leq \frac{\varepsilon^2}{\beta^2 m^3 q_{\max}^2}\end{aligned}$$

since $N \geq \varepsilon^{-1}$. Thus, applying Bernstein's inequality with $\lambda = \varepsilon v_1$ and $\Delta = \frac{\varepsilon}{m\varepsilon^2} v_1 = \frac{\lambda}{m\varepsilon^2}$ we get

$$\begin{aligned}Pr[X \geq \frac{1}{\beta N} + 2\lambda] &\leq \exp\left(\frac{-4\lambda^2}{2\sum_i \sigma_i^2 + \lambda\Delta}\right) \\ &\leq \exp\left(\frac{-4\lambda^2}{2\lambda^2/(m\varepsilon^2) + \lambda^2/(m\varepsilon^2)}\right) \\ &\leq e^{-4m\varepsilon^2/3}.\end{aligned}$$

Using that $\frac{1}{\beta N} + 2\lambda \leq 3\varepsilon z_p$ and using the union bound for at most $\beta m \log_2\left(\frac{m}{\varepsilon}\right)$ coordinates z_p with $z_p \in W_q^+$ for some $q \in Q_2$ concludes the proof, as the total failure probability P is bounded by

$$\begin{aligned}P &= \beta m \log_2\left(\frac{m}{\varepsilon}\right) \exp(-4m\varepsilon^2/3) \\ &= \exp\left(\ln\left(\beta m \log_2\left(\frac{m}{\varepsilon}\right)\right) - 4m\varepsilon^2/3\right) \\ &\leq \exp(-m\varepsilon^2)\end{aligned}$$

by Assumption B.1. □

Other important weight classes. Let $q \in Q_1$. Then we have $|W_q^+| \geq \beta m$. We show that we can find a set of representatives for W_q where each representative z_p is in a bucket with no other entry larger than $\frac{|z_p|}{\varepsilon m}$. We set $Y_q = \{p \in [n] \mid h_p = h(q) \text{ and } |z_p| \geq \frac{2^{q-1}}{m\varepsilon}\}$.

Lemma B.5. *Let $q \in Q_1$ and $h = h(q)$. Then with failure probability at most $\exp(-m\varepsilon^2)$ there exists a subset $W_q^* \subset W_q^+$ such that $\|W_q^*\|_1 \geq (1-7\varepsilon)\frac{\|W_q^+\|_1}{b^h \beta}$ and every $z_p \in W_q^*$ is in a bucket with no other element of Y_q.*

Proof. For $z_i \in W_q$ define $X_i = 1$ if $h_i = h$ and $X_i = 0$ otherwise, and set $X = \sum X_i$. The expected number of entries from W_q^+ at level h is $\mathbb{E}(X) = \frac{|W_q^+|}{\beta b^h} \in [m, bm)$ by definition of $h = h(q)$. Using Bernstein's inequality we get that with failure probability at most $P_1 = \exp\left(\frac{-(2\varepsilon\mathbb{E}(X))^2}{2\mathbb{E}(X)}\right) \leq \exp(-2\varepsilon^2 m)$, there are at least $\frac{|W_q^+|}{\beta b^h}(1-2\varepsilon)$ entries of W_q^+ at level h. We denote a uniform random subset of size $y = \frac{|W_q^+|}{\beta b^h}(1-2\varepsilon)$ of such entries by W_q'.

Since $q \in Q_1$ we have $|W_q^+| \geq \frac{\varepsilon/q_{\max}}{2^{1-q}}$ and thus also $h = \lfloor \log_b(\frac{|W_q^+|}{\beta m}) \rfloor \geq \log_b(\frac{\varepsilon 2^{q-1}}{\beta m q_{\max}}) - 1$. Since $G(z) = 1$ there are at most $2^{q-1} m\varepsilon$ entries larger or equal to $\frac{1}{m\varepsilon 2^{q-1}}$. The expected number of entries from Y_q at level h is thus bounded by $E = \frac{2^{q-1} m\varepsilon}{\beta b^h} \leq q_{\max} m^2 b$. To show that the number is not much larger than that, we define independent random variables $X_i = 1$ if $h_i = h$ and $X_i = 0$ otherwise, for $i \in Y_q$. The variance is bounded by

$$\sum_{i\in Y_q} \sigma_i^2 \leq \sum_{i\in Y_q} \frac{1}{\beta b^h} = E.$$

Thus using Bernstein's inequality we get

$$\begin{aligned}P_2 = P(|Y_q| \geq E + E) &\leq \exp\left(\frac{-E^2}{2E + E}\right) \\ &= \exp\left(\frac{-E}{3}\right) \\ &\leq \exp(-2m\varepsilon^2).\end{aligned}$$

Now for $z_i \in W'_q$ consider the random variable $X_i = 1$ if z_i is put into a bucket with another entry of Y_q, and $X_i = 0$ otherwise. Set $X = \sum X_i$. We have $P(X_i = 1) < \frac{2E}{N}$ and $\mathbb{E}(X) \leq \frac{2Ey}{N}$. The variance is bounded by $\sigma_i^2 = \frac{2E}{N} + \left(\frac{2E}{N}\right)^2 \leq \frac{4E}{N}$. Thus up to failure probability

$$\begin{aligned} P_3 = P(X \geq \frac{2Ey}{N} + \varepsilon y) &\leq \exp\left(-\frac{2y^2\varepsilon^2}{8Ey/N + y\varepsilon}\right) \\ &\leq \exp(-m(1-2\varepsilon)\varepsilon/2) \\ &\leq \exp(-2m\varepsilon^2) \end{aligned}$$

we have that at most $2(\frac{2Ey}{N} + \varepsilon y) \leq 4\varepsilon y$ entries in W'_q are in a bucket with another entry of Y_q. (Here the factor of 2 comes from the possibility that z_i is placed in a bucket with another entry of W'_q which was put into the bucket before z_i.) We denote by W^*_q the subset of W'_q that is in a bucket with no other entry of Y_q, and set $y' = |W^*_q|$. Note that with with failure probability at most P_3, we have $y' \geq \frac{|W^+_q|}{\beta b^h}(1-6\varepsilon)$.

Now for $z_i \in W^*_q$ consider the random variable $X_i = z_i$ if $z_i \in W^*_q$, and $X_i = 0$ otherwise. Again using Bernstein's inequality we get

$$\begin{aligned} P_4 &= P(\|W^*_q\|_1 \leq \frac{y'}{|W^+_q|}\|W^*_q\|_1 - 2y'\varepsilon 2^{-q}) \\ &\leq \exp\left(\frac{-(2\varepsilon y' 2^{-q})^2}{2y'2^{-2q}}\right) \\ &= \exp\left(-2y'\varepsilon^2\right) \leq \exp(-3m\varepsilon^2/2). \end{aligned}$$

Since $\varepsilon y' 2^{-q} \leq \varepsilon \frac{\|W^+_q\|_1}{b^h\beta}$ this shows that W^*_q has the desired properties with failure probability at most $P_1 + P_2 + P_3 + P_4 \leq 4\exp(-2m\varepsilon^2/2) \leq \exp(-m\varepsilon^2)$. □

Lemma B.6. *Let $q \in Q_1$ and assume W^*_q from Lemma B.5 exists. Then with failure probability at most $e^{-m\varepsilon^2}$, for any $z_p \in W^*_q$ we have $\|B_p \setminus \{p\}\|_1 \leq 3\varepsilon|z_p|$.*

The proof of Lemma B.6 is similar to the proof of Lemma B.4.

Contribution of important weight classes. By combining Lemma B.1, Lemma B.6, and Lemma B.4, we can prove Theorem 3:

Proof of Theorem 3. By assumption, $\mathcal{E}$ holds. Further, by a union bound, with failure probability at most

$$\begin{aligned} P &= q_{\max}\exp(-m\varepsilon^2) + (q_{\max}+1)\exp(-m\varepsilon^2) \\ &\leq \exp(\ln(2q_{\max}+1) - m\varepsilon^2) \\ &\leq \exp(-m\varepsilon^2/2) \end{aligned}$$

for each important weight class there exists W^*_q as in Lemma B.5, and the events of Lemmas B.4 and B.6 hold. For $q \leq \log_b\left(\frac{\varepsilon}{\beta m q_{\max}}\right)$ we set $W^*_q = W^+_q$. We have

$$\begin{aligned} G^+(Sz) &\geq \sum_{q\in Q^*, z_p \in W^*_q} (1-3\varepsilon) z_p b^{h_p}\beta \\ &\geq \sum_{q\in Q^*} (1-3\varepsilon)\|W^*_q\|_1 b^{h_p}\beta \\ &\geq \sum_{q\in Q^*} (1-3\varepsilon)(1-7\varepsilon)\frac{\|W^+_q\|_1}{b^{h(q)}\beta} b^{h(q)}\beta \\ &\geq (1-10\varepsilon)(1-5\varepsilon')G^+(z) \\ &\geq (1-60\varepsilon')G^+(z). \end{aligned}$$

The first inequality follows by Lemma B.6 and Lemma B.4. The second follows by Lemma B.5 and the third one by Lemma B.1. □

B.2. Net argument

Next we show that for all z we have $G^+(Sz) \geq (1-\varepsilon')G^+(z)$ with high probability. We need the following lemma:

Lemma B.7. *Let $z, e \in \mathbb{R}^n$ with $G(e) \leq \frac{\varepsilon}{b^{h_{\max}}}G(z)$. Then it holds that $G^+(z+e) = (1\pm\varepsilon')G^+(z)$ and $G^+(S(z+e)) = (1\pm\varepsilon')G^+(Sz)$. Further, if $G^+(Sz) \geq (1-\varepsilon')G^+(z)$, then $G^+(S(z+e)) \geq (1-4\varepsilon')G^+(z+e)$.*

Proof. A simple case distinction shows that for all $v, v' \in \mathbb{R}$ it holds that $|\max\{v',0\} - \max\{v,0\}| \leq |v'-v|$: If both v and v' have the same sign then either $|\max\{v',0\} - \max\{v,0\}| = 0$ or $|\max\{v',0\} - \max\{v,0\}| = |v'-v|$, and if v and v' have different signs then $|v'-v| = |v'| + |v| \geq |\max\{v',0\} - \max\{v,0\}|$. Thus we have

$$\begin{aligned} &|G^+(z+e) - G^+(z)| \\ &\leq \sum_{i\in[n]} |\max\{z_i+e_i,0\} - \max\{z_i,0\}| \\ &\leq G(e) \leq \varepsilon G(z) \\ &\leq \frac{\varepsilon'}{\mu'}G(z) \leq \varepsilon' G^+(z). \end{aligned}$$

It holds that $G(Se) \leq b^{h_{\max}}G(e)$ since all entries in S are bounded by $b^{h_{\max}}$ and each entry of e is multiplied by exactly one non-zero entry of S. Hence

$$\begin{aligned} |G^+(S(z+e)) - G^+(Sz)| &\leq G(Se) \\ &\leq b^{h_{\max}}G(e) \\ &\leq \frac{\varepsilon'}{\mu'}G(z) \leq \varepsilon' G^+(z). \end{aligned}$$

Finally if $G^+(Sz) \geq (1-\varepsilon')G^+(z)$ then by combining the previous inequalities we get

$$G^+(S(z+e))$$

$$
\begin{aligned}
&\geq G^+(Sz) - \varepsilon' G^+(z) \\
&\geq (1-\varepsilon')G^+(z) - \varepsilon' G^+(z) \\
&\geq (1-\varepsilon')(G^+(z+e) - \varepsilon' G^+(z)) - \varepsilon' G^+(z) \\
&\geq (1-\varepsilon')G^+(z+e) - 2\varepsilon' G^+(z) \\
&\geq (1-\varepsilon')G^+(z+e) - 2\varepsilon' G^+(z+e)/(1-\varepsilon') \\
&\geq (1-4\varepsilon')G^+(z+e)\,.
\end{aligned}
$$

□

Now we are ready to prove Theorem 4.

Proof of Theorem 4. With failure probability at most δ, we can assume that $\mathcal{E}$ holds. Since $G^+(az) = a(G^+(z))$ for all $z \in \mathbb{R}^d$ and $a \in \mathbb{R}_{\geq 0}$, it suffices to show that $G^+(Sz) \geq (1-\varepsilon')G^+(z)$ holds for all $z \in \mathbb{R}^d$ with $G(z) = 1$. We set $M = \lceil \frac{b^{h_{\max}}}{\varepsilon} \rceil$. Consider the set $N_\varepsilon = \{(n_1, \ldots, n_d)\frac{1}{M} \mid n_1 + \cdots + n_d = M\}$. The set N_ε consists of at most $\left(\left(1 + \frac{b^{h_{\max}}}{\varepsilon}\right)^d\right) = \left(\exp\left(d \ln\left(1 + \frac{n}{m\varepsilon}\right)\right)\right)$ elements as $h_{\max} = \ln_b\left(\frac{n}{m}\right)$. By Theorem 3 and a union bound we have that $G^+(Sz) \geq (1 - 60\varepsilon')G^+(z)$ holds for all $z \in N_\varepsilon$ with failure probability at most $\exp\left(d \ln\left(1 + \frac{n}{m\varepsilon}\right)\right) \exp(-m\varepsilon^2/2) < \delta$ since $d \ln(1 + \frac{n}{m\varepsilon}) - m\varepsilon^2/2 \leq \ln(\delta)$ by Assumption B.1. For each z with $G(z) = 1$ there exists $z' \in N_\varepsilon$ such that $G(z' - z) \leq \frac{\varepsilon}{b^{h_{\max}}}$. Thus we can apply Lemma B.7 and get $G^+(Sz) \geq (1 - 240\varepsilon')G^+(z)$. The total failure probability is at most 2δ using the union bound. □

B.3. Dilation bounds

Our first dilation bound is very simple but yields only an $h_{\max} = O(\log n)$ approximation.

Lemma B.8. *We have* $\mathbb{E}(G^+(Sz)) \leq h_{\max} G^+(z)$.

Proof. The expected contribution of z_i is less than 0 if $z_i < 0$. If $z_i \geq 0$ then the expected contribution is upper bounded by $\sum_{h=0}^{h_{\max}} \frac{1}{b^h \beta} b^h \beta z_i = \sum_{h=0}^{h_{\max}} z_i = h_{\max} z_i$ with equality if and only if all z_i are greater or equal to zero. Thus

$$\mathbb{E}(G^+(Sz)) \leq \sum_{z_i \geq 0} z_i h_{\max} = h_{\max} G^+(z).$$

□

We can achieve a better constant dilation by considering only the top buckets at each level. More precisely set $K = \beta m \log(\frac{m}{\varepsilon}) + \beta m b \log_2(b\varepsilon^{-1}) = O(mb \log_2(b\varepsilon^{-1}))$. We define

$$G_c^+(Sz) := \sum_h \beta b^h \sum_{i \in [K]} G^+(L_{h,i})$$

where $L_{h,i}$ denotes the level h bucket with the i-th largest sum of entries among all level h buckets. It is important here to take the buckets with the largest contributions to preserve the convexity of the objective as pointed out in (Clarkson & Woodruff, 2015), since the resulting function is related to the Ky Fan norm and is thus convex. The proof of the bounded contraction of $G_c^+(Sz)$, Theorem 5, only requires lower bounds on $G^+(L_{h,i})$ for those at most K buckets in level h containing some member of W_q^*: there are at most $\frac{bm}{\varepsilon}$ entries greater or equal to $\frac{\varepsilon}{bm}$. For other important weight classes, the cardinality of W_q^* is bounded by bm and the number of important weight classes W_q^+ with $h(q) = h$ is bounded by $\log_2(b\varepsilon^{-1})$, as it must hold that $|W_q^+| \in [\beta m b^h, \beta m b^{h+1}]$, $|W_q^+| 2^{-q} \leq \frac{\varepsilon}{q_{\max}}$ and $|W_q^+| 2^{1-q} \geq \|W_q^+\|_1 \geq \frac{\varepsilon}{q_{\max}}$ and thus

$$\beta m b^h \frac{q_{\max}}{\varepsilon} \leq 2^q \leq 2\beta m b^{h+1} \frac{q_{\max}}{\varepsilon}\,.$$

Thus if the estimator for $G^+(z)$ uses only the largest buckets with the largest sums, the proven bounds on contraction continue to hold, and in particular $G_c^+(Sz) \geq (1-\varepsilon')G^+(z)$. To show that the dilation of $G_c^+(Sz)$ is constant we need the following lemma, which shows that the probability that an important entry of W_q gets placed at a much higher level than $h(q)$ is low. This way we can bound the contribution that entries have at higher levels.

Lemma B.9. *Let* $q' = \log_2(nh_{max})$. *With failure probability at most* δ *the event* $\mathcal{E}'$ *holds that there is no entry* $z_p \in W_q$ *with* $q \leq q'$ *and* $h_p > h_q := h(q) + \log_b\left(\frac{q' bm}{\delta}\right)$.

Proof. Let $q \leq q'$. For any coordinate p and any level h, the probability that $h_p > h$ can be bounded by $\sum_{h'=h+1}^{h_{\max}} \frac{1}{\beta b^{h'}} \leq \frac{1}{b^h}$ since $b > 2$. Thus the expected number of coordinates $z_p \in W_q$ with $h_p > h_q$ is bounded by

$$\frac{|W_q|}{b^{h_q}} \leq \frac{b^{h_q+1} m}{b^{h_q}} \leq \frac{\delta}{q'}.$$

This also gives us an upper bound for the probability that there is no coordinate in W_q with $h_p > h_q$. Using the union bound for all $q \leq q'$ completes the proof. □

Proof of Theorem 5. Assume that $\mathcal{E}'$ holds. Note that the expected contribution of any entry z_p is at most $z_p h_{\max}$, and thus the expected contribution of all entries less than or equal to $\frac{1}{h_{\max} n}$ is at most 1. It remains to show that for each $q \leq q'$ the expected contribution of W_q^+ is bounded by $C\|W_q^+\|_1$. We consider the expected contribution of W_q^+ at level h and distinguish three cases:

Case 1: $h = h(q) - k$ for $k > \log_b(N \frac{\ln(h_{\max} N)}{m})$.
For $z_i \in W_q$ consider again the random variable $X_i = 1$ if $h_i = h$, and $X_i = 0$ otherwise, and set $X = \sum X_i$. The expected number of entries from W_q^+ at level h is

$\mathbb{E}(X) = \frac{|W_q^+|}{\beta b^h} \geq N \ln(h_{\max} N)$. For the variance we have $\sum \sigma_i \leq \mathbb{E}(X)$ as X is a sum of Bernoulli random variables. Using Bernstein's inequality we get that $|G(L_h) \cap W_q^+| \leq 2\beta^{-1} b^{-h} |W_q^+|$ with failure probability at most

$$P_1 = \exp\left(\frac{-(\mathbb{E}(X))^2}{2\mathbb{E}(X) + \mathbb{E}(X)}\right) \leq \exp(-N/3).$$

Hence we can assume that the number of entries in each bucket at level h is at most $\frac{2|W_q^+|}{\beta b^h N} \leq 2\frac{b^k m}{N} = \ln(h_{\max} N)$.

Next for each bucket and each $z_i \in W_q$ we consider the random variable $Z = \sum Z_i$ where $Z_i = 1$ if entry i is in the corresponding bucket, and $Z_i = 0$ otherwise. Then the variance is bounded by $\frac{1}{N} \cdot 1^2 + 1 \cdot \frac{1}{N^2} \leq \frac{2}{N}$. Applying Bernstein's inequality gives us

$$P(Z \geq 2\frac{b^k m}{N} + \lambda) \leq \exp\left(\frac{-\lambda^2}{2N \cdot \frac{2}{N} + 2\lambda/3}\right).$$

For $\lambda = \ln(h_{\max} N)$ this implies $P(Z \geq 2\frac{b^k m}{N} + \lambda) = O((h_{\max} N)^{-1})$. Using the union bound, the probability that there exists a bucket with at least $\frac{b^k m}{N} + \lambda$ coordinates can be bounded by $P_2 = O((h_{\max})^{-1})$ Further we have

$$\|W_q^+\|_1 \geq |W_q^+| 2^{-q} \geq 2^{-q} b^{h(q)} \beta m.$$

The expected contribution of W_q^+ at level h can thus be bounded by

$$\Lambda = P_2 \cdot 2^{1-q} 3 |W_q^+| b^h \beta + K\left(\frac{3|W_q^+|}{\beta b^h N}\right) 2^{1-q} b^h \beta$$
$$\leq \left(O(h_{\max}^{-1}) + \frac{3K}{N} 2\right) \|W_q^+\|_1 = O(h_{\max}^{-1}) \|W_q^+\|_1.$$

Summing over at most $h_{\max}$ levels we have that the contribution in this case is bounded by $O(1)\|W_q^+\|_1$.

Case 2: $h = h(q) + k$ for $k \geq \log_b\left(\frac{q' bm}{\delta}\right)$.
By $\mathcal{E}'$ (see Lemma B.9), the set $L_h \cap W_q^+$ is empty, and thus the expected contribution in this case is 0.

Case 3: $h \geq h(q) - \log_b(\frac{N \ln(h_{\max} N)}{m})$ and $h < h(q) + \log_b\left(\frac{q' bm}{\delta}\right)$.
Note that $\log_b(\frac{N \ln(h_{\max} N)}{m}) + \log_b\left(\frac{q' bm}{\delta}\right)$ is constant since $b \geq \max\left\{m, \delta^{-1}\right\}$ by Assumption B.1, and thus $N = b^{\mathcal{O}(1)}$, and the expected contribution of each level is at most constant. The total expected contribution is thus $O(1)\|W_q^+\|_1$. □

Proof of Theorem 2. The result follows by combining Theorem 4 and Theorem 5 and substituting $240\varepsilon'$ by ε. The $\mathrm{poly}(\mu d \log n)$ bound on the sketch size follows from $r = N h_{\max} = N h_{\max} = O(N \log n)$ and by using Assumption B.1 for bounding N. □

C. Omitted details from Section 4

To show Theorem 6 we will first define the notions of sensitivities, VC-dimension, and the range space induced by a set of functions.

Definition C.1. *(Langberg & Schulman, 2010) Consider a family of functions $\mathcal{F} = \{g_1, \ldots, g_n\}$ mapping from $\mathbb{R}^d$ to $[0, \infty)$ and weighted by $w \in \mathbb{R}^n_{>0}$. The sensitivity of g_i for $f_w(x) = \sum_{i=1}^n w_i g_i(x)$ is*

$$\varsigma_i = \sup \frac{w_i g_i(x)}{f_w(x)} \tag{1}$$

where sup *is over all $x \in \mathbb{R}^d$ with $f_w(x) > 0$. If this set is empty then $\varsigma_i = 0$. The total sensitivity is $\mathfrak{S} = \sum_{i=1}^n \varsigma_i$.*

The sensitivity of a point bounds the maximal relative contribution to the target function the point can have. Computing the sensitivities is often intractable and necessitates approximating the original optimization problem close to optimality. However, this is the problem that we want to solve, see (Braverman et al., 2016). Fortunately, for our applications it suffices to obtain a reasonable upper bound for the sensitivities.

Definition C.2. *A range space is a pair $\mathfrak{R} = (\mathcal{F}, \mathrm{ranges})$ where $\mathcal{F}$ is a set and* ranges *is a family of subsets of $\mathcal{F}$. The VC dimension $\Delta(\mathfrak{R})$ of $\mathfrak{R}$ is the size $|G|$ of the largest subset $G \subseteq \mathcal{F}$ such that G is shattered by* ranges*, i.e., $|\{G \cap R \mid R \in \mathrm{ranges}\}| = 2^{|G|}$.*

Definition C.3. *Let $\mathcal{F}$ be a finite set of functions mapping from $\mathbb{R}^d$ to $\mathbb{R}_{\geq 0}$. For every $x \in \mathbb{R}^d$ and $r \in \mathbb{R}_{\geq 0}$, let $\mathrm{range}_{\mathcal{F}}(x, r) = \{f \in \mathcal{F} \mid f(x) \geq r\}$, and $\mathrm{ranges}(\mathcal{F}) = \{\mathrm{range}_{\mathcal{F}}(x, r) \mid x \in \mathbb{R}^d, r \in \mathbb{R}_{\geq 0}\}$, and $\mathfrak{R}_{\mathcal{F}} = (\mathcal{F}, \mathrm{ranges}(\mathcal{F}))$ be the range space induced by $\mathcal{F}$.*

The VC-dimension can be thought of as something similar to the dimension of our problem. For example the VC-dimension of the set of hyperplane classifiers in $\mathbb{R}^d$ is $d + 1$. The sensitivity scores were combined with a theory on the VC dimension of range spaces in (Braverman et al., 2016). We employ a more recent version from (Feldman et al., 2020).

Proposition C.4. *(Feldman et al., 2020) Consider a family of functions $\mathcal{F} = \{f_1, \ldots, f_n\}$ mapping from $\mathbb{R}^d$ to $[0, \infty)$ and a vector of weights $w \in \mathbb{R}^n_{>0}$. Let $\varepsilon, \delta \in (0, 1/2)$. Let $s_i \geq \varsigma_i$. Let $S = \sum_{i=1}^n s_i \geq \mathfrak{S}$. Given s_i one can compute in time $O(|\mathcal{F}|)$ a set $R \subset \mathcal{F}$ of*

$$O\left(\frac{S}{\varepsilon^2}\left(\Delta \ln S + \ln\left(\frac{1}{\delta}\right)\right)\right)$$

weighted functions such that with probability $1 - \delta$, we have for all $x \in \mathbb{R}^d$ simultaneously

$$\left|\sum_{f_i \in \mathcal{F}} w_i f_i(x) - \sum_{f_i \in R} u_i f_i(x)\right| \leq \varepsilon \sum_{f_i \in \mathcal{F}} w_i f_i(x),$$

where each element of R is sampled i.i.d. with probability $p_j = \frac{s_j}{S}$ from $\mathcal{F}$, $u_i = \frac{Sw_j}{s_j|R|}$ denotes the weight of a function $f_i \in R$ that corresponds to $f_j \in \mathcal{F}$, and where Δ is an upper bound on the VC dimension of the range space $\mathfrak{R}_{\mathcal{F}^*}$ induced by $\mathcal{F}^*$ obtained by defining $\mathcal{F}^*$ to be the set of functions $f_j \in \mathcal{F}$, where each function is scaled by $\frac{Sw_j}{s_j|R|}$.

Now we show how Proposition C.4 can be used to approximate our loss function on the negative domain. We define $g_i(x) = \min\{g(a_i x), \ln(2)\}$.

Lemma C.5. *The range space induced by $\mathcal{F} = \{g_i \mid i \in [n]\}$ satisfies $\Delta(\mathcal{R}_{\mathcal{F}}) \leq d + 1$.*

Proof. Note that g is invertible and monotone. Let $G \subseteq \mathcal{F}$, $x \in \mathbb{R}^d$ and $r \in \mathbb{R}$. For $r > \ln(2)$ we have $\text{range}_G(x, r) = \emptyset$, otherwise (for $r \leq \ln(2)$) we have

$$\begin{aligned}\text{range}_G(x, r) &= \{g_i \in G \mid g_i(x) \geq r\} \\ &= \{g_i \in G \mid a_i x \geq g^{-1}(r)\}.\end{aligned}$$

Hence

$$\begin{aligned}&|\{\text{range}_G(x, r) \mid x \in \mathbb{R}^d, r \in \mathbb{R}_{\geq 0}\}| \\ &= \left|\{\{g_i \in G \mid a_i x - g^{-1}(r) \geq 0\} \mid x \in \mathbb{R}^d, r \leq \ln(2)\} \right. \\ &\quad \left. \cup \{\emptyset\}\right| \\ &\leq \left|\{\{g_i \in G \mid a_i x - \tau \geq 0\} \mid x \in \mathbb{R}^d, \tau \in \mathbb{R}\}\right|.\end{aligned}$$

The last set is the set of points that is shattered by the affine hyperplane classifier $a_i \mapsto \mathbf{1}_{a_i x - \tau \geq 0}$. Its VC dimension is thus $d + 1$ (Kearns & Vazirani, 1994), implying $|\{\text{range}_G(x, r) \mid x \in \mathbb{R}^d, r \in \mathbb{R}_{\geq 0}\}| = 2^{|G|}$ can only hold if $|G| \leq d + 1$, and thus the VC dimension of $\mathcal{F}$ is at most $d + 1$. □

We set $f_{\min}(x) = \frac{n}{\mu}$ for all $x \in \mathbb{R}^d$.

Corollary C.6. *The range space induced by $\mathcal{F} = \{g_i \mid i \in [n]\} \cup \{f_{\min}\}$ satisfies $\Delta(\mathcal{R}_{\mathcal{F}}) \leq d + 2$.*

Proof. Assume there exists $G \subset \mathcal{F}$ with $|G| \geq d + 3$ and $\{G \cap R \mid R \in \text{ranges}(\mathcal{F})\} = 2^G$. Then we have for $G' = G \setminus \{f_{\min}\}$ $|G'| \geq d + 2$ and $\{G' \cap R \mid R \in \text{ranges}(\mathcal{F})\} = 2^{G'}$ contradicting Lemma C.5. □

Now we are ready to prove Theorem 6.

Proof of Theorem 6. We want to apply Proposition C.4 to $\mathcal{F} = \{g_i \mid i \in [n]\} \cup \{f_{\min}\}$. By Corollary C.6 the VC-dimension of $\mathcal{F}$ is at most $d + 2$. Note that the sensitivity of any function, in particular of $f_{\min}$, is at most 1 and for the sensitivity of any function other than $f_{\min}$ we have

$$\sup_{x \in \mathbb{R}^d \setminus \{0\}} \frac{g_i(x)}{f((Ax)^-) + f_{\min}(x)} \leq \sup_x \frac{g_i(x)}{f_{\min}(x)} \leq \frac{\ln(2)\mu}{n}$$

and the total sensitivity is thus bounded by $\ln(2)\mu + 1$. Hence we can use Proposition C.4 and with failure probability at most δ_1, we get a subset r of $\mathcal{F} = \{g_i \mid i \in [n]\} \cup \{f_{\min}\}$ and a weight vector u such that

$$\left|\sum_{f_i \in \mathcal{F}} f_i(x) - \sum_{f_i \in R} u_i f_i(x)\right| \leq \varepsilon \sum_{f_i \in \mathcal{F}} f_i(x).$$

As it holds that $\sum_{f \in \mathcal{F}} f_i(x) = f((Ax)^-) + \frac{n}{\mu}$, and since the weight of $f_{\min}$ is 1, this implies for $R' = R \setminus \{f_{\min}\}$ that we have an error Δ of at most

$$\begin{aligned}\Delta &= \left|\sum_{f_i \in \mathcal{F} \setminus \{f_{\min}\}} f_i(x) - \sum_{f_i \in R'} u_i f_i(x)\right| \\ &\leq \varepsilon \sum_{f_i \in \mathcal{F}} f_i(x) \\ &\leq \varepsilon \left(\sum_{f_i \in \mathcal{F} \setminus \{f_{\min}\}} f_i(x) \; + \; \frac{n}{\mu}\right) \\ &\leq 3\varepsilon f(Ax).\end{aligned}$$

The last inequality follows from Lemma 2.2, and using that $f(Ax) \geq \sum_{f_i \in \mathcal{F} \setminus \{f_{\min}\}} f_i(x)$. This proves the first part of the theorem.

Observe that the expected contribution of row a_i is $P(g_i \in R) a_i u_i = \frac{k}{n} \frac{n}{k} g(a_i x) = g(a_i x)$. Thus the second statement follows by linearity of expectation. □

D. Omitted details from Section 5

Proof of Corollary 5.2. We need one streaming pass over the data in $O(\texttt{nnz}(A))$ time to draw a uniform random sample T from Theorem 6 and to compute $A' = SA$. Now compute the x^* minimizing the convex objective function $f(Tx^*) + G_c^+(A'x^*)$. This can be done using the ellipsoid method on the following convex program: we have one variable x_i for $i \in [d]$. For each row t_i of T, construct a variable v_i and a constraint $v_i \geq t_i x$. Similarly for each row a'_i of A' construct a variable $v'_i \geq 0$ and a constraint $v'_i \geq a'_i x$. The objective is to minimize the convex function

$$\sum g(v_i) + \sum v'_i.$$

The convex program has $\text{poly}(\mu d \log n)$ many variables and thus the running time is also $\text{poly}(\mu d \log n)$. Using the same analysis as in the previous proof shows that the solution x' we get satisfies

$$f(Ax') \leq O(1) \min_{x \in \mathbb{R}^d} f(Ax)$$

with constant probability. □

E. Omitted details from Section 6

Dataset	n	d	**Source**
Covertype	581012	54	https://archive.ics.uci.edu/ml/datasets/Covertype
Webspam	350000	128	https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#webspam
Kddcup	494021	33	https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
Synthetic	100000	2	-

Table 1. Summary of the datasets: d is given without intercept. Datasets are downloaded resp. generated automatically by our open code.

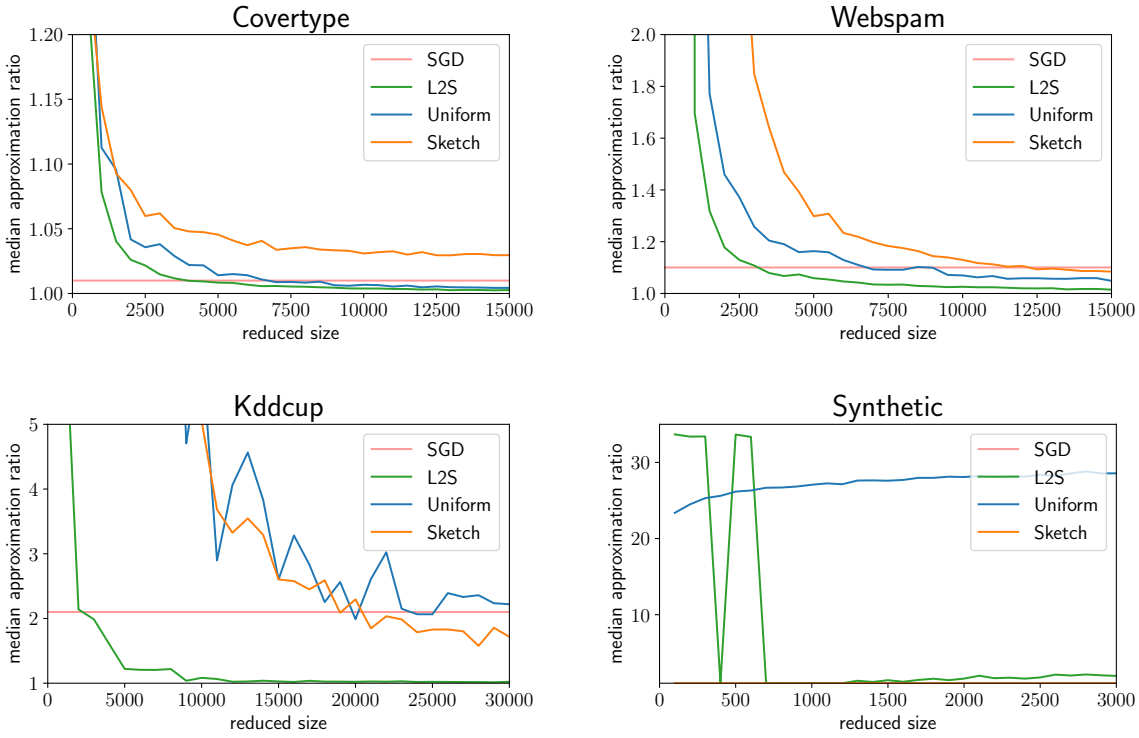

Figure 2. Comparison of the approximation ratios.

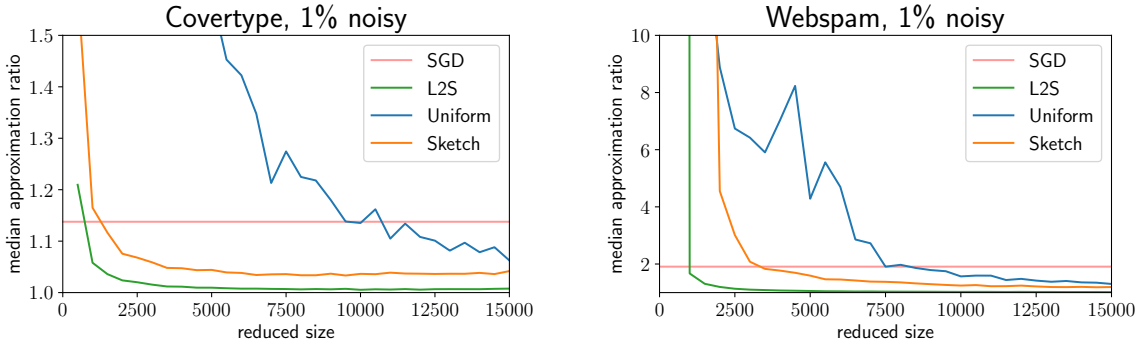

Figure 3. Comparison of the approximation ratios with added noise.

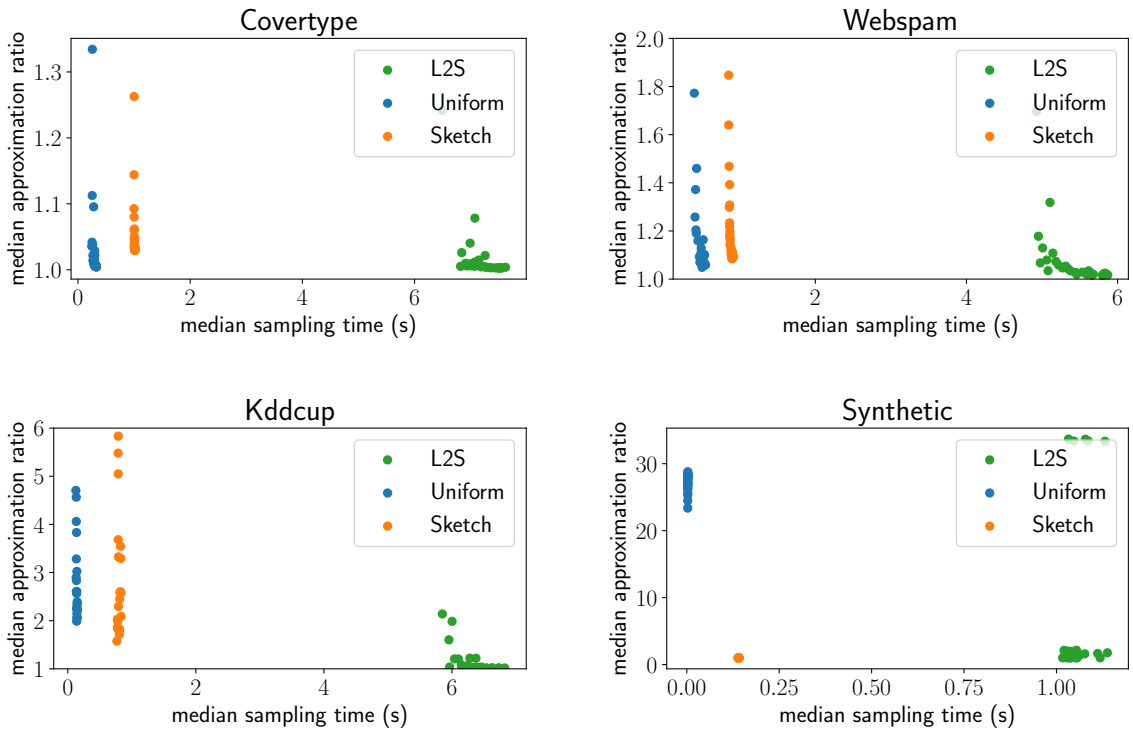

Figure 4. Comparison of sketching resp. sampling times vs. accuracy.

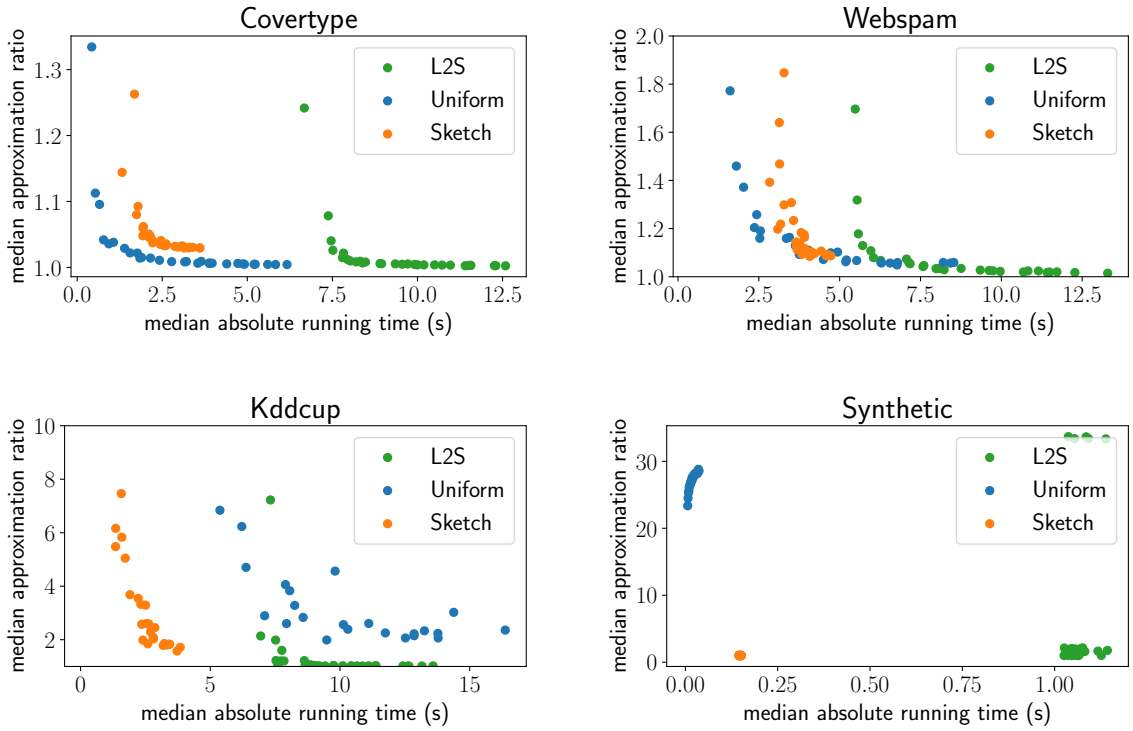

Figure 5. Comparison of total running times including optimization vs. accuracy.