SUPPLEMENTARY MATERIAL:
Sparse Solutions to Nonnegative Linear Systems and Applications

A Auxiliary lemmas

We use the following technical lemmas in the proof.

Lemma A.1. *Let $\delta > 0$, and $f(t) := (1+\delta)^t - 1$. Then for any $t_1, t_2 \geq 0$, we have*

$$f(t_1) + f(t_2) \leq f(t_1 + t_2).$$

Proof. The proof follows immediately upon expansion:

$$f(t_1 + t_2) - f(t_1) - f(t_2) = \big((1+\delta)^{t_1} - 1\big)\big((1+\delta)^{t_2} - 1\big).$$

The term above is non-negative because δ, t_1, t_2 are all ≥ 0. □

Lemma A.2 (Averaging)**.** *Let $\{a_i, b_i\}_{i=1}^k$ be non-negative real numbers, such that*

$$\sum_{i=1}^k a_i = A \qquad \textit{and} \qquad \sum_{i=1}^k b_i = 1.$$

Then for any parameter $C > 1$, there exists an index i such that $b_i \geq 1/(Ck)$, and $a_i \leq b_i \cdot A/(1 - 1/C)$.

Proof. Let $S := \{i \ : \ b_i \geq 1/(Ck)\}$. Now since there are only k indices, we have $\sum_{i \in [k] \setminus S} b_i < k \cdot 1/(Ck) < 1/C$, and thus

$$\sum_{i \in S} b_i > (1 - 1/C). \tag{6}$$

Next, since all the a_i are non-negative, we get that

$$\sum_{i \in S} a_i \leq A.$$

Combining the two, we have

$$\frac{\sum_{i \in S} a_i}{\sum_{i \in S} b_i} < \frac{A}{1 - 1/C}.$$

Thus there exists an index $i \in S$ such that $a_i < b_i \cdot A/(1 - 1/C)$ (because otherwise, we have $a_i \geq b_i A/(1 - 1/C)$ for all i, thus summing over $i \in S$, we get a contradiction to the above). This proves the lemma. □

Lemma A.3. *For any $0 < x < 1$ and $\delta > 0$, we have*

$$(1+\delta)^x \leq 1 + \delta x \leq (1+\delta)^{x(1+\delta)}.$$

Proof. For any $0 < \theta < \delta$, we have

$$\frac{1}{1+\theta} < \frac{1}{1+\theta x} < \frac{1+\delta}{1+\theta}.$$

The first inequality is because $x < 1$, and the second is because the RHS is bigger than 1 while the LHS is smaller. Now integrating from $\theta = 0$ to $\theta = \delta$, we get

$$\log(1+\delta) < \frac{\log(1 + x\delta)}{x} < (1+\delta)\log(1+\delta).$$

Multiplying by x and exponentiating gives the desired claim. □

B Gaussian mixtures

B.1 Proof of Lemma 3.5

Let $\hat{f}^{\mathcal{S}'}$ be the empirical distribution over $\mathcal{S}'$. Since $|\mathcal{I}_i| = \frac{1}{\epsilon_1}$, the induced partition $\mathcal{S}$ satisfies $|\mathcal{S}| \leq \frac{1}{\epsilon_1^d}$. Hence by the Chernoff and union bounds, for $n \geq \frac{8}{\epsilon_1^d \epsilon^3} \log \frac{2}{\delta \epsilon_1^d}$, with probability $\geq 1-\delta$,

$$|f^{\mathcal{S}'}(S) - \hat{f}^{\mathcal{S}'}(S)| \leq \sqrt{f^{\mathcal{S}'}(S)\epsilon_1^d \epsilon^3}/2 + \epsilon_1^d \epsilon^3/2, \ \forall S \in \mathcal{S}. \tag{7}$$

For the set U,

$$\begin{aligned}
f^{\mathcal{S}'}(U) &= \sum_{S:\hat{f}^{\mathcal{S}'}(S) \leq \epsilon^d \epsilon} f^{\mathcal{S}'}(S) \\
&\leq \epsilon + \sum_{S \in \mathcal{S}} \sqrt{f^{\mathcal{S}'}(S)\epsilon_1^d \epsilon^3}/2 + \sum_{S \in \mathcal{S}} \epsilon_1^d \epsilon^3/2 \\
&\leq 2\epsilon,
\end{aligned}$$

where the second inequality follows from the concavity of $\sqrt{x}$. However $b(U) = 2\epsilon$ and hence $b(U) \geq f^{\mathcal{S}'}(U)(1-2\epsilon)$.

By Equation (7),

$$\begin{aligned}
|f^{\mathcal{S}'}(S) - \hat{f}^{\mathcal{S}'}(S)| &\leq \sqrt{f^{\mathcal{S}'}(S)\epsilon_1^d \epsilon^3}/2 + \epsilon_1^d \epsilon^3/2 \\
&\leq \sqrt{\hat{f}^{\mathcal{S}'}(S)\epsilon_1^d \epsilon^3}/2 + \epsilon_1^d \epsilon^3/2 \\
&\leq \hat{f}^{\mathcal{S}'}(S)\left(\sqrt{\epsilon^2}/2 + \epsilon^2/2\right) \\
&\leq \hat{f}^{\mathcal{S}'}(S)\epsilon.
\end{aligned}$$

The penultimate inequality follows from the fact that $\hat{f}^{\mathcal{S}'}(S) \geq \epsilon_1^d \epsilon$. Hence $\hat{f}^{\mathcal{S}'}(S) \geq f^{\mathcal{S}'}(S)(1-\epsilon)$. Furthermore by construction $b(S) \geq \hat{f}^{\mathcal{S}'}(S)(1-2\epsilon)$. Hence $b(S) \geq f^{\mathcal{S}'}(S)(1-3\epsilon) \, \forall S \in \mathcal{S}'$.

For the second part of the lemma observe that b and $f^{\mathcal{S}'}$ are distributions over $\mathcal{S}'$. Hence

$$\begin{aligned}
\sum_{S \in \mathcal{S}} |b(S) - f^{\mathcal{S}'}(S)| &= 2 \sum_{S: b(S) \leq f^{\mathcal{S}'}(S)} f^{\mathcal{S}'}(S) - b(S) \\
&\leq 2 \sum_{S \in S} f^{\mathcal{S}'}(S) \cdot 3\epsilon = 6\epsilon.
\end{aligned}$$

B.2 Proof of Lemma 3.7

If p and $\mathcal{I}$ are simultaneously scaled or translated, then the value of $\left\|p - p^{\mathcal{I}}\right\|_1$ remains unchanged. Hence proving the lemma for $p = N(0,1)$ is sufficient. We first divide $\mathcal{I}$ into $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$ depending on the minimum and maximum values of $p(x)$ in the corresponding intervals.

$$I \in \begin{cases} \mathcal{I}_1 & \text{if } \min_{x \in I} p(x) \geq \sqrt{\epsilon/(2\pi)}, \\ \mathcal{I}_2 & \text{if } \max_{x \in I} p(x) \leq \sqrt{\epsilon/(2\pi)}, \\ \mathcal{I}_3 & \text{else.} \end{cases}$$

The ℓ_1 distance between p and $p^{\mathcal{I}}$ is

$$\left\|p - p^{\mathcal{I}}\right\|_1 = \sum_{I \in \mathcal{I}} \int_{x \in I} |p(x) - p^{\mathcal{I}}(x)| dx.$$

We bound the above summation by breaking it into terms corresponding to $\mathcal{I}_1, \mathcal{I}_2$, and $\mathcal{I}_3$ respectively. Observe that $|\mathcal{I}_3| \le 2$ and $p(I) \le \epsilon \,\forall\, I \in \mathcal{I}_3$. Hence,

$$\sum_{I \in \mathcal{I}_3} \int_{x \in I} |p(x) - p^{\mathcal{I}}(x)| dx \le 4\epsilon.$$

Since $\max_{x \in I} p(x)$ for every interval in $I \in \mathcal{I}_2$ is $\le \sqrt{\epsilon/(2\pi)}$, by Gaussian tail bounds

$$\begin{aligned}\sum_{I \in \mathcal{I}_2} \int_{x \in I} |p(x) - p^{\mathcal{I}}(x)| dx &\le 2 \sum_{I \in \mathcal{I}_2} \int_{x \in I} p(x) dx \\ &\le 2\sqrt{\epsilon}.\end{aligned}$$

For every interval $I \in \mathcal{I}_1$ we first bound its interval length and maximum value of $p'(x)$. Note that

$$p(I) \ge |I| \min_{y \in I} p(y).$$

In particular since $p(I) \le \epsilon$ and $\min_{y \in I} p(y) \ge \sqrt{\epsilon/(2\pi)}$, $|I| \le \sqrt{2\pi\epsilon}$. Let $s = \max_{x \in I} |p'(x)|$.

$$s = \max_{x \in I} |p'(x)| = \max_{x \in I} \frac{|x|}{\sqrt{2\pi}} e^{-x^2/2} \le \max_{x \in I} |x| \cdot \max_{x \in I} p(x).$$

Since $\min_{y \in I} p(y) \ge \sqrt{\epsilon/(2\pi)}$, we have $\max_{y \in I} |y| \le \sqrt{\log 1/\epsilon}$. Let $y_1 = \text{argmax}_{y \in I} p(y)$ and $y_2 = \text{argmin}_{y \in I} p(y)$, then

$$\begin{aligned}\frac{\max_{y \in I} p(y)}{\min_{y \in I} p(y)} &= \frac{p(y_2)}{p(y_1)} \\ &= e^{(y_1^2 - y_2^2)/2} \\ &= e^{(y_1 - y_2)(y_2 + y_1)/2} \\ &\le e^{|I|\sqrt{\log \frac{1}{\epsilon}}} \\ &\le e^{\sqrt{2\pi\epsilon \log \frac{1}{\epsilon}}}.\end{aligned}$$

Since $p^{\mathcal{I}}(x) = p(I)/|I|$, by Rolle's theorem $\exists x_0$ such that $p^{\mathcal{I}}(x) = p(x_0) \,\forall x$. By first order Taylor's expansion,

$$\begin{aligned}\int_{x \in I} |p(x) - p^{\mathcal{I}}(x)| dx &\le \int_{x \in I} |(x - x_0) \max_{y \in [x_0, x]} |p'(y)| dx \\ &\le s \int_{x \in I} |x - x_0| dx \\ &\le s|I|^2/2 \\ &\le \frac{s}{2} \left(\frac{p(I)}{\min_{y \in I} p(y)} \right) \sqrt{2\pi\epsilon} \\ &\le \sqrt{2\pi\epsilon} p(I) \max_{x \in I} |x| \cdot \frac{\max_{y \in I} p(y)}{2 \min_{y \in I} p(y)} \\ &\le e^{\sqrt{2\pi\epsilon \log \frac{1}{\epsilon}}} \cdot \sqrt{\pi\epsilon/2} \cdot p(I) \max_{x \in I} |x|,\end{aligned}$$

where the last three inequalities follow from the bounds on $|I|$, s, and $\frac{\max_{y \in I} p(y)}{\min_{y \in I} p(y)}$ respectively. Thus,

$$\begin{aligned}\int_{x \in I} |p(x) - p^{\mathcal{I}}(x)| dx &\le e^{\sqrt{2\pi\epsilon \log \frac{1}{\epsilon}}} \sqrt{\pi\epsilon/2} \cdot p(I) \max_{x \in I} |x| \\ &\le e^{\sqrt{2\pi\epsilon \log \frac{1}{\epsilon}}} \sqrt{\pi\epsilon/2} \cdot \int_{x \in I} p(x)(|x| + \sqrt{2\pi\epsilon}) dx.\end{aligned}$$

Summing over $I \in \mathcal{I}_1$, we get the above summation is $\le e^{\sqrt{2\pi\epsilon \log \frac{1}{\epsilon}}} \sqrt{\pi\epsilon/2}(1 + \sqrt{2\pi\epsilon})$. Adding terms corresponding to $\mathcal{I}_1, \mathcal{I}_2$, and $\mathcal{I}_3$ we get

$$\left\| p - p^{\mathcal{I}} \right\|_1 \le e^{\sqrt{2\pi\epsilon \log \frac{1}{\epsilon}}} \cdot \sqrt{\pi\epsilon/2} \cdot (1 + \sqrt{2\pi\epsilon}) + 2\sqrt{\epsilon} + 4\epsilon < 30\sqrt{\epsilon}.$$

B.3 Proof of Theorem 3.8

Using Lemma 3.7 we first show that for every d-dimensional $(\mathcal{I}_1, \mathcal{I}_2 \dots \mathcal{I}_d), \epsilon$-flat Gaussian is close to the unflattened one.

Lemma B.1. *For every $(\mathcal{I}_1, \mathcal{I}_2, \dots \mathcal{I}_d), \epsilon$-good axis-aligned Gaussian distribution $p = p_1 \times p_2 \times \dots p_d$, we have*

$$\left\| p - p^{\mathcal{S}} \right\|_1 \le 30d\sqrt{\epsilon}.$$

Proof. By triangle inequality, the distance between any two product distributions is upper bounded by the sum of distances in each coordinate. Hence,

$$\left\| p - p^{\mathcal{S}} \right\|_1 \le \sum_{i=1}^{d} \left\| p_i - p_i^{\mathcal{I}_i} \right\|_1 \le 30d\sqrt{\epsilon},$$

where the second inequality follows from Lemma 3.7. □

We now have all the tools to prove the main result on Gaussian mixtures.

Proof of Theorem 3.8. By triangle inequality,

$$\left\| \hat{f} - f \right\|_1 \le \left\| \hat{f}^{\mathcal{S}} - f^{\mathcal{S}} \right\|_1 + \left\| \hat{f}^{\mathcal{S}} - \hat{f} \right\|_1 + \left\| f^{\mathcal{S}} - f \right\|_1.$$

We now bound each of the terms above. By Lemma 3.6, the first term is $\le 74\epsilon$. By triangle inequality for $\hat{f} = \sum_{r=1}^{k'} \hat{w}_r \hat{p}_r$,

$$\left\| \hat{f}^{\mathcal{S}} - \hat{f} \right\|_1 \le \sum_{r=1}^{k'} \hat{w}_r \left\| \hat{p}_r^{\mathcal{S}} - \hat{p}_r \right\|_1 \le 30\sqrt{2}\epsilon,$$

where the last inequality follows from the fact that the allowed distributions in $A^{\mathcal{S}'}$ are $(\mathcal{I}_1, \mathcal{I}_2, \dots \mathcal{I}_d), 2\epsilon^2/d^2$-good and by Lemma B.1. By triangle inequality,

$$\begin{aligned}
\left\| f^{\mathcal{S}} - f \right\|_1 &\le \sum_{r=1}^{k} w_r \left\| p_r^{\mathcal{S}} - p_r \right\|_1 \\
&\le \sum_{r: w_r \ge \epsilon/k} w_r \left\| p_r^{\mathcal{S}} - p_r \right\|_1 + \sum_{r: w_r < \epsilon/k} w_r \left\| p_r^{\mathcal{S}} - p_r \right\|_1 \\
&\le \sum_{r: w_r \ge \epsilon/k} w_r \left\| p_r^{\mathcal{S}} - p_r \right\|_1 + 2\epsilon \\
&\le 30\sqrt{2}\epsilon + 2\epsilon.
\end{aligned}$$

where the last inequality follows from the proof of Lemma 3.6, where we showed that heavy components are $(\mathcal{I}_1, \mathcal{I}_2, \dots \mathcal{I}_d), 2\epsilon^2/d^2$-good and by Lemma B.1. Summing over the terms corresponding to $\left\| \hat{f}^{\mathcal{S}} - f^{\mathcal{S}} \right\|_1$, $\left\| \hat{f}^{\mathcal{S}} - \hat{f} \right\|_1$, and $\left\| f^{\mathcal{S}} - f \right\|_1$, we get the total error as $74\epsilon + 30\sqrt{2}\epsilon + 30\sqrt{2}\epsilon + 2\epsilon \le 170\epsilon$. The error probability and the number of samples necessary are same as that of Lemma 3.6. The run time follows from the comments in Section 3.2 and the bound on number of samples. □

C Lower bounds

We now detail the proofs of Theorems 4.1 and 4.3.

C.1 Proof of Theorem 4.1

We reduce hard instance of the Max k-cover problem to our problem as follows. For each set S_i, we set A_i (column i in A) to be the indicator vector of set S_i. We also let b to be the vector with all entries equal to one. In the YES case, it is easy to construct a k-sparse x s.t. $Ax = b$, while in the NO case, finding a solution of sparsity $o(k \ln \frac{1}{\epsilon})$ contradicts the hardness result stated above.

In the YES case, we know there are k disjoint sets whose union is the universe, and we construct solution x^* as follows. We set x^*_i (the ith entry of x^*) to one if set S_i is one of these k sets, and zero otherwise. It is clear that Ax^* is equal to b, and therefore there exists a k-sparse solution in the YES case.

On the other hand, for every ϵ-approximate non-negative solution $\hat{x}$, we know that the number of non-zero entries of $A\hat{x}$ is at most ϵm by definition. Define C to be the sub-collection of sets with non-zero entry in $\hat{x}$, i.e. $\{S_i \mid \hat{x}_i > 0\}$. We know that each non-zero entry in $A\hat{x}$ is covered by some set in sub-collection C. In other words, the union of sets in C has size at least $(1-\epsilon)m$. We imply that the number of sets in collection C should be at least $\Omega(k \ln(\frac{1}{\epsilon+\delta}))$ since $(1-\frac{1}{k})^k$ is in range $[\frac{1}{4}, \frac{1}{e}]$. We can choose δ to be ϵ, and therefore the sparsest solution that one can find in the NO case is $\Omega(k \ln(\frac{1}{\epsilon}))$-sparse. Assuming $\mathcal{P} \neq \mathcal{NP}$, it is not possible to find a $o(k \ln \frac{1}{\epsilon})$-sparse ϵ-approximate solution when there exists a k-sparse solution, otherwise it becomes possible to distinguish between the YES and NO cases of the Max k-Cover problem in polynomial time.

C.2 Proof of Theorem 4.3

Let n (which is $O(m/\log m)$) be the number of sets in the set system. Let A be the $m \times n$ matrix whose i,jth entry is 1 if element i is in set j, and 0 otherwise. It is clear that in the YES case, there exists a solution to $Ax = \mathbf{1}$ of sparsity k. It suffices to show that in the NO case, there is no ϵ-approximate solution to $Ax = \mathbf{1}$ with fewer than $\Omega(k/\epsilon^2)$ entries.

Let us define $C = 1/\epsilon^2$, for convenience. The proof follows the standard template in random matrix theory (e.g. Rudelson and Vershynin [2010]): we show that for any *fixed* Ck-sparse vector x, the probability that $\|Ax - \mathbf{1}\|_1 < 1/(4\sqrt{C})$ is tiny, and then take a union bound over all x in a fine enough grid to conclude the claim for all k-sparse x.

Thus let us fix some Ck sparse vector x and consider the quantity $\|Ax - \mathbf{1}\|_1$. Let us then consider one row, which we denote by y, and consider $|\langle y, x\rangle - 1|$. Now each element of y is 1 with probability $1/k$ and 0 otherwise (by the way the set system was constructed). Let us define the mean-zero random variable W_i, $1 \le i \le n$, as follows:

$$W_i = \begin{cases} 1 - 1/k & \text{with probability } 1/k, \\ -1/k & \text{otherwise.} \end{cases}$$

We first note that $\mathbb{E}[|\langle y, x\rangle - 1|^2] \ge \mathbb{E}[(\sum_i W_i x_i)^2]$. This follows simply from the fact that for any random variable Z, we have $\mathbb{E}[|Z-1|^2] \ge \mathbb{E}[|Z - \mathbb{E}[Z]|^2]$ (i.e., the best way to "center" a distribution with respect to a least squares objective is at its mean). Thus let us consider

$$\mathbb{E}\left[\left(\sum_i W_i x_i\right)^2\right] = \sum_i x_i^2 \cdot \mathbb{E}[W_i^2] = \sum_i x_i^2 \cdot \frac{1}{k}\left(1 - \frac{1}{k}\right).$$

Since x is Ck-sparse, and since $\|x\|_1 \ge 3k/4$, we have $\sum_i x_i^2 \ge \frac{1}{Ck} \cdot \|x\|_1^2 \ge k/2C$. Plugging this above and combining with our earlier observation, we obtain

$$\mathbb{E}[|\langle y, x\rangle - 1|^2] \ge \mathbb{E}\left[\left(\sum_i W_i x_i\right)^2\right] \ge \frac{1}{3C}. \tag{8}$$

Now we will use the Paley-Zygmund inequality,[5] with the random variable $Z := |\langle y, x\rangle - 1|^2$. For this we need to upper bound $\mathbb{E}[Z^2] = \mathbb{E}[|\langle y, x\rangle - 1|^4]$. We claim that we can bound it by a constant. Now since $\|x\|_1$ is between

[5] For any non-negative random variable Z, we have $\Pr(Z \ge \theta \mathbb{E}[Z]) \ge (1-\theta)^2 \cdot \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}$.

$3k/4$ and $5k/4$, we have $|\langle y, x\rangle - \sum_i W_i x_i| < 1/2$. This in turn implies that $\mathbb{E}[Z^2] \le 4(\mathbb{E}[(\sum_i W_i x_i)^4] + 4)$. We will show that $\mathbb{E}[(\sum_i W_i x_i)^4] = O(1)$.

$$\begin{aligned}\mathbb{E}[(\sum_i W_i x_i)^4] &= \sum_i W_i^4 x_i^4 + 3\sum_{i,j} W_i^2 W_j^2 x_i^2 x_j^2 \\ &\le \frac{1}{k}\cdot\sum_i x_i^4 + \frac{3}{k^2}\sum_{i,j} x_i^2 x_j^2 \\ &\le 1 + \frac{3}{k^2}\cdot(\sum_i x_i^2)^2 = O(1).\end{aligned}$$

Here we used the fact that we have $0 \le x_i \le 1$ for all i, and that $\sum_i x_i^2 \le \sum_i x_i \le 5k/4$.

This implies, by using the Paley-Zygmund inequality, that

$$\Pr\left[|\langle y, x\rangle - 1| < \frac{1}{4\sqrt{C}}\right] < 1 - 1/10. \tag{9}$$

Thus if we now look at the m rows of A, and consider the number of them that satisfy $|\langle y, x\rangle - 1| < 1/(4\sqrt{C})$, the expected number is $< 9m/10$, thus the probability that there are more than $19m/20$ such rows is $\exp(-\Omega(m))$. Thus we have that for any Ck-sparse x with $\|x\|_1 \in [3k/4, 5k/4]$ and $\|x\|_\infty \le 1$,

$$\Pr\left[\|Ax - \mathbf{1}\|_1 < \frac{1}{80\sqrt{C}}\right] < e^{-m/40}. \tag{10}$$

Now let us construct an ϵ'-net[6] for the set of all Ck-sparse vectors, with $\epsilon' = 1/m^2$. A simple way to do it is to first pick the non-zero coordinates, and take all integer multiples of ϵ'/m as the coordinates. It is easy to see that this set of points (call it $\mathcal{N}$) is an ϵ' net, and furthermore, it has size roughly

$$\binom{m}{Ck}\left(\frac{m}{\epsilon'}\right)^{Ck} = O\left(m^{4Ck}\right).$$

Thus as long as $m > 200Ck\log m$, we can take a union bound over all the vectors in the ϵ' net, to conclude that with probability $e^{-\Omega(m)}$, we have

$$\|Ax - \mathbf{1}\|_1 > \frac{1}{80\sqrt{C}} \qquad \text{for all } x \in \mathcal{N}.$$

In the event that this happens, we can use the fact that $\mathcal{N}$ is an ϵ' net (with $\epsilon' = 1/m^2$), to conclude that $\|Ax - \mathbf{1}\|_1 > \frac{1}{100\sqrt{C}}$ for *all* Ck-sparse vectors with coordinates in $[0, 1]$ and $\|x\|_1 \in [3k/4, 5k/4]$.

This completes the proof of the theorem, since $\frac{1}{100\sqrt{C}}$ is $\Omega(\epsilon)$, and $k < m/\log^2 m$.

D Experiments

To get a better understanding of the accuracy of the analysis, we implemented our algorithm with some natural settings for matrices A. We now give a brief summary of the results.

Random A

The most natural choice for A are $m \times n$ matrices, with each entry distributed independently. We picked each entry uniformly in $(0, 1)$, and then normalized the columns.[7] The vector b is obtained by taking a random combination of the first k columns of A. Here we observed the following: as n grows, keeping m fixed, the

[6] An ϵ-net for a set S of points is a subset T such that for any $s \in S$, there is a $t \in T$ such that $\|s - t\|_2 < \epsilon$.

[7] Note that unlike the case of random Gaussian entries, the columns here are not incoherent, and nor does A possess the restricted isometry property.

number of non-zero entries in the solution (found by the algorithm) slowly grows, until it stabilizes. There is also a disparity between the number of iterations of the algorithm (which we used to bound the sparsity), and the actual sparsity. In order to allow for a large n relative to m, we fix a small value of m (15 below), and $k = 5$. We also show the behavior with three choices of ϵ.

In the table, the #iter refers to the average number of iterations for the ϵ, $\langle k' \rangle$ refers to the average sparsity of the solution obtained, and the $\pm$ denotes the standard deviation in the value of k'.

Table 1: Random A, with $m = 15$ rows, and n columns, varying ϵ

n	$\epsilon = 0.005$			$\epsilon = 0.002$			$\epsilon = 0.001$		
	#iter	$\langle k' \rangle$	$\pm$	#iter	$\langle k' \rangle$	$\pm$	#iter	$\langle k' \rangle$	$\pm$
10	75	5.78	2.36	169	5.78	2.36	391	6.22	3.68
15	92	7.11	4.11	274	8.33	4.69	552	8.33	4.69
20	116	8.56	7.23	325	9.22	7.72	694	9.56	8.84
40	129	11.6	9.81	430	13.8	7.85	919	14.2	8.34
60	155	13.7	15.4	648	15	18.1	1650	15.9	17.2
100	102	17.6	22.3	278	21.6	25.8	635	24.6	31.1
500	63	29.4	13.8	141	44.1	21.8	275	59.9	30.9
1000	60	30.2	9.46	164	52.2	25.6	279	64.6	27.9
2000	52	30.2	12.9	123	50.7	23.2	229	70.7	28.4

Gaussian mixture in one dimension

The next choice for A comes from our result on learning mixtures of Gaussians. Here we consider the one-dimensional version, in which we are given a mixture of k Gaussians on a line, whose components we wish to find. We picked $k = 4$, and unit variance for the Gaussians. The means were picked at random in the interval $(0, 20)$, which was discretized into 200 sub-intervals. We then considered 200 candidate Gaussians, with means at the centers of each sub-interval. The goal is to approximate the given mixture by a mixture of these candidate Gaussians (up to an error ϵ) using as few components as possible.

The following table gives the results for varying ϵ. We also give the true means, and the means of the Gaussians used in the approximation we find.

Table 2: Experiment with $k = 3$, true means: $\{6.6, 9.2, 11.1\}$

ϵ	# iterations	sparsity (k')	means found by algorithm
0.05	10	6	6.5, 6.7, 9.3, 11.0, 11.1, 11.4
0.02	18	11	6.5, 6.6, 6.7, 9.2, 9.3, 10.9, 11.0, 11.1, 11.2, 11.3, 11.4
0.01	68	12	6.5, 6.6, 6.7, 9.1, 9.2, 9.3, 10.9, 11.0, 11.1, 11.2, 11.3, 11.4
0.005	262	12	6.5, 6.6, 6.7, 9.1, 9.2, 9.3, 10.9, 11.0, 11.1, 11.2, 11.3, 11.4

We observe that the approximate solution we find uses Gaussians that are "close" to the component Gaussians. (E.g., we use Gaussians with means 6.5, 6.6, 6.7 to approximate the effect of the Gaussian with mean 6.6). The table below shows how the number of iterations and k' change when we set $k = 4$.

Table 3: Experiment with $k = 4$

ϵ	# iterations	sparsity (k')
0.05	16	10
0.02	26	11
0.01	61	18
0.005	305	19