A. Proofs of Lemma 1 and 2

Let X be a random variable which is distributed according to $\mathcal{M}$ and suppose we obtain T samples $y_1, y_2, \dots, y_T \sim \mathcal{M}$. We will divide these T samples into $B := \lceil T/t \rceil$ batches each of size t. In that case let us denote $S^j_{1,t}$ and $S^j_{2,t}$ to be the sample mean and the sample variance of the j^{th} batch i.e.

$$S^j_{1,t} = \sum_{i \in \text{Batch } j} \frac{y_i}{t} \qquad \text{and} \qquad S^j_{2,t} = \frac{1}{t-1} \sum_{i \in \text{Batch } j} (y_i - (S^j_{1,t}))^2.$$

We will estimate the true mean $\mathbb{E}X$ and the true variance $\text{var}X$ by computing $\hat{M}_1$ and $\hat{M}_2$ respectively (See Algorithm 2) where

$$\hat{M}_1 \triangleq \mathsf{median}(\{S^j_{1,t}\}_{j=1}^B) \quad \text{and} \quad \hat{M}_2 \triangleq \mathsf{median}(\{S^j_{2,t}\}_{j=1}^B).$$

Proof of Lemma 1. For a fixed batch j, we can use Chebychev's inequality to say that

$$\Pr\left(\left|S^j_{1,t} - \mathbb{E}X\right| \ge \epsilon_1\right) \le \frac{\text{var}X}{t\epsilon_1^2}$$

We have

$$\text{var}X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{1}{2}\Big(2\sigma^2 + \mu_1^2 + \mu_2^2\Big) - \frac{1}{4}(\mu_1 + \mu_2)^2 = \sigma^2 + \frac{(\mu_1 - \mu_2)^2}{4}$$

Noting that we must have $t \ge 1$ as well, we obtain

$$\Pr\left(\left|S^j_{1,t} - \mathbb{E}X\right| \ge \epsilon_1\right) \le \frac{\sigma^2 + (\mu_1 - \mu_2)^2/4}{t\epsilon_1^2} \le \frac{1}{3}$$

for $t = O(\lceil(\sigma^2 + (\mu_1 - \mu_2)^2)/\epsilon_1^2\rceil)$. Therefore for each batch j, we define an indicator random variable $Z_j = \mathbb{1}[\left|S^j_{1,t} - \mathbb{E}X\right| \ge \epsilon_1]$ and from our previous analysis we know that the probability of Z_j being 1 is less than $1/3$. It is clear that $\mathbb{E}\sum_{j=1}^B Z_j \le B/3$ and on the other hand $|\hat{M}_1 - \mathbb{E}X| \ge \epsilon_1$ iff $\sum_{j=1}^B Z_j \ge B/2$. Therefore, using the Chernoff bound, we have

$$\Pr\left(\left|\hat{M}_1 - \mathbb{E}X\right| \ge \epsilon_1\right) \le Pr\Bigg(\left|\sum_{j=1}^B Z_j - \mathbb{E}\sum_{j=1}^B Z_j\right| \ge \frac{\mathbb{E}\sum_{j=1}^B Z_j}{2}\Bigg) \le 2e^{-B/36}.$$

Hence, for $B = 36\log\eta^{-1}$, the estimate $\hat{M}_1$ is atmost ϵ_1 away from the true mean $\hat{M}_1$ with probability at least $1 - 2\eta$. Therefore the total sample complexity required is $T = O(\log\eta^{-1}\lceil(\sigma^2 + (\mu_1 - \mu_2)^2)/\epsilon_1^2\rceil)$ proving the lemma. □

Proof of Lemma 2. We have

$$\begin{aligned}
\mathbb{E}S^j_{2,t} &= \mathbb{E}\frac{1}{t-1} \sum_{i \in \text{Batch } j} (y_i - (S^j_{1,t}))^2 \\
&= \mathbb{E}\frac{1}{t(t-1)} \sum_{\substack{i_1, i_2 \in \text{Batch } j \\ i_1 < i_2}} (y_{i_1} - y_{i_2})^2 \\
&= \frac{1}{t(t-1)} \sum_{\substack{i_1, i_2 \in \text{Batch } j \\ i_1 < i_2}} \mathbb{E}(y_{i_1} - y_{i_2})^2 \\
&= \frac{1}{t(t-1)} \sum_{\substack{i_1, i_2 \in \text{Batch } j \\ i_1 < i_2}} \mathbb{E}y_{i_1}^2 + \mathbb{E}y_{i_2}^2 - 2\mathbb{E}\left[y_{i_1} y_{i_2}\right]
\end{aligned}$$

$$= \frac{1}{t(t-1)} \sum_{\substack{i_1, i_2 \in \text{Batch } j \\ i_1 < i_2}} 2\sigma^2 + \mu_1^2 + \mu_2^2 - \frac{(\mu_1+\mu_2)^2}{2}$$
$$= \sigma^2 + \frac{(\mu_1-\mu_2)^2}{4} = \text{var}X.$$

Hence the estimator $S_{2,t}^j$ is an unbiased estimator since it's expected value is the true variance of X. Again, we must have

$$\mathbb{E}(S_{2,t}^j)^2 = \mathbb{E}\frac{1}{t^2(t-1)^2}\Big(\sum_{\substack{i_1, i_2 \in \text{Batch } j \\ i_1 < i_2}} (y_{i_1} - y_{i_2})^2\Big)^2.$$

Claim 4. *We have*

$$\mathbb{E}\left[(y_{i_1} - y_{i_2})^2 (y_{i_3} - y_{i_4})^2\right] \le 48\Big(\sigma^2 + \frac{(\mu_1-\mu_2)^2}{4}\Big)^2$$

for any i_1, i_2, i_3, i_4 such that $i_1 < i_2$ and $i_3 < i_4$.

Proof. In order to prove this claim consider three cases:

Case 1 (i_1, i_2, i_3, i_4 are distinct): In this case, we have that $y_{i_1} - y_{i_2}$ and $y_{i_3} - y_{i_4}$ are independent and therefore,

$$\mathbb{E}\left[(y_{i_1} - y_{i_2})^2 (y_{i_3} - y_{i_4})^2\right] = \mathbb{E}\left[(y_{i_1} - y_{i_2})^2\right]\mathbb{E}\left[(y_{i_3} - y_{i_4})^2\right] \le 4\Big(\sigma^2 + \frac{(\mu_1-\mu_2)^2}{4}\Big)^2.$$

Case 2 ($i_1 = i_3, i_2 = i_4$): In this case, we have

$$\mathbb{E}\left[(y_{i_1} - y_{i_2})^2 (y_{i_3} - y_{i_4})^2\right] = \mathbb{E}\left[(y_{i_1} - y_{i_2})^4\right].$$

Notice that

$$y_{i_1} - y_{i_2} \sim \frac{1}{2}\mathcal{N}(0, 2\sigma^2) + \frac{1}{4}\mathcal{N}(\mu_1 - \mu_2, 2\sigma^2) + \frac{1}{4}\mathcal{N}(\mu_2 - \mu_1, 2\sigma^2)$$

and therefore we get

$$\mathbb{E}\left[(y_{i_1} - y_{i_2})^4\right] = 48\sigma^4 + 12\sigma^2(\mu_1-\mu_2)^2 + \frac{(\mu_1-\mu_2)^4}{2} \le 48\Big(\sigma^2 + \frac{(\mu_1-\mu_2)^2}{4}\Big)^2.$$

Case 3 ($\{i_1, i_2, i_3, i_4\}$ has 3 unique elements): Without loss of generality let us assume that $i_1 = i_3$. In that case we have

$$\mathbb{E}\left[(y_{i_1} - y_{i_2})^2 (y_{i_1} - y_{i_4})^2\right] = \mathbb{E}_{y_{i_1}}\mathbb{E}\left[(y_{i_2} - y_{i_1})^2 (y_{i_4} - y_{i_1})^2 \mid y_{i_1}\right]$$

Notice that for a fixed value of y_{i_1}, we must have $y_{i_2} - y_{i_1}, y_{i_4} - y_{i_1}$ to be independent and identically distributed i.e.

$$y_{i_2} - y_{i_1}, y_{i_4} - y_{i_1} \sim \frac{1}{2}\mathcal{N}(\mu_1 - y_{i_1}, \sigma^2) + \frac{1}{2}\mathcal{N}(\mu_2 - y_{i_1}, \sigma^2).$$

Therefore,

$$\mathbb{E}\left[(y_{i_2} - y_{i_1})^2 (y_{i_4} - y_{i_1})^2 \mid y_{i_1}\right] = \mathbb{E}\left[(y_{i_2} - y_{i_1})^2 \mid y_{i_1}\right]\mathbb{E}\left[(y_{i_2} - y_{i_1})^2 \mid y_{i_1}\right]$$
$$= \frac{1}{4}\Big(2\sigma^2 + (\mu_1 - y_{i_1})^2 + (\mu_2 - y_{i_1})^2\Big)^2.$$

Again, we have

$$y_{i_1} - \mu_1 \sim \frac{1}{2}\mathcal{N}(0, \sigma^2) + \frac{1}{2}\mathcal{N}(\mu_2 - \mu_1, \sigma^2) \quad \text{and} \quad y_{i_1} - \mu_2 \sim \frac{1}{2}\mathcal{N}(0, \sigma^2) + \frac{1}{2}\mathcal{N}(\mu_1 - \mu_2, \sigma^2).$$

Hence,

$$\mathbb{E}\Big(2\sigma^2 + (\mu_1 - y_{i_1})^2 + (\mu_2 - y_{i_1})^2\Big)^2 = 4\sigma^4 + \mathbb{E}(\mu_1 - y_{i_1})^4 + \mathbb{E}(\mu_1 - y_{i_2})^4$$
$$+ 4\sigma^2(\mathbb{E}(\mu_1 - y_{i_1})^2 + \mathbb{E}(\mu_1 - y_{i_2})^2) + \mathbb{E}((\mu_1 - y_{i_1})^2(\mu_2 - y_{i_1})^2).$$

We have

$$\begin{aligned}
&\mathbb{E}(\mu_1 - y_{i_1})^4 = \mathbb{E}(\mu_2 - y_{i_1})^4 = 3\sigma^4 + \frac{(\mu_1-\mu_2)^4}{2} + 3\sigma^2(\mu_1-\mu_2)^2 \\
&\mathbb{E}(\mu_1 - y_{i_1})^2 = \mathbb{E}(\mu_2 - y_{i_1})^2 = \sigma^2 + \frac{(\mu_1-\mu_2)^2}{2} \\
&\mathbb{E}((\mu_1 - y_{i_1})^2(\mu_2 - y_{i_1})^2) = \mathbb{E}((\mu_1 - y_{i_1})^2(\mu_2 - \mu_1 + \mu_1 - y_{i_1})^2) \\
&= \mathbb{E}[(\mu_1 - y_{i_1})^4 + (\mu_1-\mu_2)^2(\mu_1 - y_{i_1})^2 + 2(\mu_2-\mu_1)(\mu_1 - y_{i_1})^3] \\
&= 3\sigma^4 + \frac{(\mu_1-\mu_2)^4}{2} + 5\sigma^2(\mu_1-\mu_2)^2.
\end{aligned}$$

Plugging in, we get

$$\mathbb{E}\Big(2\sigma^2 + (\mu_1 - y_{i_1})^2 + (\mu_2 - y_{i_1})^2\Big)^2 = 17\sigma^4 + \frac{3(\mu_1-\mu_2)^4}{2} + 13\sigma^2(\mu_1-\mu_2)^2.$$

Hence, we obtain

$$\mathbb{E}\left[(y_{i_1} - y_{i_2})^2(y_{i_1} - y_{i_4})^2\right] \le 7\Big(\sigma^2 + \frac{(\mu_1-\mu_2)^2}{2}\Big)^2.$$

which proves the claim. □

From Claim 4, we can conclude that

$$\mathbb{E}(S^j_{2,t})^2 \le 12\Big(\sigma^2 + \frac{(\mu_1-\mu_2)^2}{4}\Big)^2.$$

From this point onwards, the analysis in this lemma is very similar to Lemma 1. We can use Chebychev's inequality to say that

$$\Pr\left(\left|S^j_{2,t} - \mathrm{var}X\right| \ge \epsilon_2\right) \le \frac{\mathrm{var}S^j_{2,t}}{t\epsilon_2^2} \le \frac{\mathbb{E}(S^j_{2,t})^2}{t\epsilon_2^2}.$$

Therefore, we obtain by noting that $t \ge 1$ as well,

$$\Pr\left(\left|S^j_{2,t} - \mathrm{var}X\right| \ge \epsilon_2\right) \le \frac{12(\sigma^2 + (\mu_1-\mu_2)^2/4)^2}{t\epsilon_2^2} \le \frac{1}{3}$$

for $t = O(\left\lceil(\sigma^2 + (\mu_1-\mu_2)^2)^2/\epsilon_2^2\right\rceil)$. At this point, doing the same analysis as in Lemma 1 shows that $B = 36\log\eta^{-1}$ batches of batchsize t is sufficient to estimate the variance within an additive error of ϵ_2 with probability at least $1-2\eta$. Therefore the total sample complexity required is $T = O(\log\eta^{-1}\left\lceil(\sigma^2 + (\mu_1-\mu_2)^2)^2/\epsilon_2^2\right\rceil)$ thus proving the lemma.

□

B. Proof of Lemma 4

Suppose, we use $O\Big(\left\lceil\frac{1}{\epsilon^2}\right\rceil\log\eta^{-1}\Big)$ samples to recover $\hat{\mu_1}$ and $\hat{\mu_2}$ using the method of moments. According to the guarantee provided in Theorem 3, we must have with probability at least $1-1/\eta$,

$$|\hat{\mu}_i - \mu| \le 2(\epsilon + \sqrt{\epsilon})\sqrt{\sigma^2 + (\mu_1-\mu_2)^2} \quad \text{for } i = 1, 2.$$

Therefore, we have

$$\begin{aligned}&|\mu_1-\mu_2|-|\mu_1-\hat{\mu}_1|-|\mu_2-\hat{\mu}_2| \le |\hat{\mu}_1-\hat{\mu}_2| \le |\mu_1-\mu_2|+|\mu_1-\hat{\mu}_1|+|\mu_2-\hat{\mu}_2| \\ &||\hat{\mu}_1-\hat{\mu}_2|-|\mu_1-\mu_2|| \le 4(\epsilon+\sqrt{\epsilon})\sqrt{\sigma^2+(\mu_1-\mu_2)^2}.\end{aligned}$$

We will substitute $\epsilon = 1/256$. In that case we have

$$||\hat{\mu}_1-\hat{\mu}_2|-|\mu_1-\mu_2|| \le \frac{17}{64}\sqrt{\sigma^2+(\mu_1-\mu_2)^2} \le \frac{17\sigma}{64}+\frac{17\,|\mu_1-\mu_2|}{64}.$$

and therefore

$$-\frac{17\sigma}{64}+\frac{47\,|\mu_1-\mu_2|}{64} \le |\hat{\mu}_1-\hat{\mu}_2| \le \frac{17\sigma}{64}+\frac{81\,|\mu_1-\mu_2|}{64}.$$

Hence, we have

$$-\frac{17\sigma}{81}+\frac{64\,|\hat{\mu}_1-\hat{\mu}_2|}{81} \le |\mu_1-\mu_2| \le \frac{17\sigma}{47}+\frac{64\,|\hat{\mu}_1-\hat{\mu}_2|}{47}.$$

Case 1 ($\sigma \ge \gamma$): This implies that if $|\hat{\mu}_1-\hat{\mu}_2| \le 15\sigma/32$, then $|\mu_1-\mu_2| \le \sigma$ and we will use the Method of Moments (Algorithm 3). On the other hand, if $|\hat{\mu}_1-\hat{\mu}_2| \ge 15\sigma/32$, then $|\mu_1-\mu_2| \ge 13\sigma/81$ and we will use EM algorithm (Algorithm 1). The sample complexity required is $O(\left\lceil \log \eta^{-1} \right\rceil)$ samples.

Case 2 ($\sigma \le \gamma$): This implies that if $|\hat{\mu}_1-\hat{\mu}_2| \le 15\gamma/32$, then $|\mu_1-\mu_2| \le \gamma$ and we will fit a single gaussian (Algorithm 4) to recover the means. On the other hand, if $|\hat{\mu}_1-\hat{\mu}_2| \ge 15\gamma/32$, then $|\mu_1-\mu_2| \ge 13\gamma/81$ and we will use EM algorithm (Algorithm 1). The sample complexity required is $O(\left\lceil \log \eta^{-1} \right\rceil)$ samples.

C. Proof of Lemma 5

We have

$$\begin{aligned}&\mu_{1,1}=\langle \mathbf{x}^1,\beta_1\rangle \quad \mu_{1,2}=\langle \mathbf{x}^1,\beta_2\rangle \quad \mu_{2,1}=\langle \mathbf{x}^2,\beta_1\rangle \quad \mu_{2,2}=\langle \mathbf{x}^2,\beta_2\rangle \\ &\mu_{\mathsf{sum},1}=\langle \mathbf{x}^1+\mathbf{x}^2,\beta_1\rangle \quad \mu_{\mathsf{sum},2}=\langle \mathbf{x}^1+\mathbf{x}^2,\beta_2\rangle \quad \mu_{\mathsf{diff},1}=\langle \mathbf{x}^1-\mathbf{x}^2,\beta_1\rangle \quad \mu_{\mathsf{diff},2}=\langle \mathbf{x}^1-\mathbf{x}^2,\beta_2\rangle.\end{aligned}$$

For a particular unknown mean $\mu_{\cdot,\cdot}$, we will denote the corresponding recovered estimate by $\hat{\mu}_{\cdot,\cdot}$ and moreover, let us assume without loss of generality that $\pi_1,\pi_2,\pi_{\mathsf{sum}},\pi_{\mathsf{diff}}$ are all same and the identity permutation itself (but note that this fact is unknown). If all the unknown parameters are recovered upto an additive error of γ, then we must have

$$|\hat{\mu}_{\mathsf{sum},1}-\hat{\mu}_{1,1}-\hat{\mu}_{2,1}| \le |\hat{\mu}_{\mathsf{sum},1}-\mu_{\mathsf{sum},1}|+|\mu_{1,1}-\hat{\mu}_{1,1}|+|\mu_{2,1}-\hat{\mu}_{2,1}| \le 3\gamma.$$

$$|\hat{\mu}_{\mathsf{diff},1}-\hat{\mu}_{1,1}+\hat{\mu}_{2,1}| \le |\hat{\mu}_{\mathsf{diff},1}-\mu_{\mathsf{diff},1}|+|\mu_{1,1}-\hat{\mu}_{1,1}|+|\mu_{2,1}-\hat{\mu}_{2,1}| \le 3\gamma.$$

On the other hand, we must have

$$\begin{aligned}|\hat{\mu}_{\mathsf{sum},1}-\hat{\mu}_{1,1}-\hat{\mu}_{2,2}| &\ge |\hat{\mu}_{\mathsf{sum},1}-\mu_{\mathsf{sum},1}+\mu_{1,1}-\hat{\mu}_{1,1}+\mu_{2,1}-\mu_{2,2}+\mu_{2,2}-\hat{\mu}_{2,2}| \\ &\ge |\mu_{2,1}-\mu_{2,2}|-|\hat{\mu}_{\mathsf{sum},1}-\mu_{\mathsf{sum},1}|-|\mu_{1,1}-\hat{\mu}_{1,1}|-|\mu_{2,1}-\hat{\mu}_{2,1}| \\ &\ge |\mu_{2,1}-\mu_{2,2}|-3\gamma.\end{aligned}$$

$$\begin{aligned}|\hat{\mu}_{\mathsf{diff},1}-\hat{\mu}_{1,1}+\hat{\mu}_{2,2}| &\ge |\hat{\mu}_{\mathsf{diff},1}-\mu_{\mathsf{diff},1}+\mu_{1,1}-\hat{\mu}_{1,1}-\mu_{2,1}+\mu_{2,2}-\mu_{2,2}+\hat{\mu}_{2,2}| \\ &\ge |\mu_{2,1}-\mu_{2,2}|-|\hat{\mu}_{\mathsf{diff},1}-\mu_{\mathsf{diff},1}|-|\mu_{1,1}-\hat{\mu}_{1,1}|-|\mu_{2,1}-\hat{\mu}_{2,1}| \\ &\ge |\mu_{2,1}-\mu_{2,2}|-3\gamma.\end{aligned}$$

$$\begin{aligned}|\hat{\mu}_{\mathsf{sum},1}-\hat{\mu}_{1,2}-\hat{\mu}_{2,1}| &\geq |\hat{\mu}_{\mathsf{sum},1}-\mu_{\mathsf{sum},1}+\mu_{2,1}-\hat{\mu}_{2,1}+\mu_{1,1}-\mu_{1,2}+\mu_{1,2}-\hat{\mu}_{1,2}| \\ &\geq |\mu_{1,1}-\mu_{1,2}|-|\hat{\mu}_{\mathsf{sum},1}-\mu_{\mathsf{sum},1}|-|\mu_{2,1}-\hat{\mu}_{2,1}|-|\mu_{1,2}-\hat{\mu}_{1,2}| \\ &\geq |\mu_{1,1}-\mu_{1,2}|-3\gamma\end{aligned}$$

$$\begin{aligned}|\hat{\mu}_{\mathsf{diff},1}-\hat{\mu}_{1,2}+\hat{\mu}_{2,1}| &\geq |\hat{\mu}_{\mathsf{diff},1}-\mu_{\mathsf{diff},1}-\mu_{2,1}+\hat{\mu}_{2,1}+\mu_{1,1}-\mu_{1,2}+\mu_{1,2}-\hat{\mu}_{1,2}| \\ &\geq |\mu_{1,1}-\mu_{1,2}|-|\hat{\mu}_{\mathsf{diff},1}-\mu_{\mathsf{diff},1}|-|\mu_{2,1}-\hat{\mu}_{2,1}|-|\mu_{1,2}-\hat{\mu}_{1,2}| \\ &\geq |\mu_{1,1}-\mu_{1,2}|-3\gamma\end{aligned}$$

$$\begin{aligned}|\hat{\mu}_{\mathsf{sum},1}-\hat{\mu}_{1,2}-\hat{\mu}_{2,2}| &\geq |\hat{\mu}_{\mathsf{sum},1}-\mu_{\mathsf{sum},1}+\mu_{1,1}+\mu_{2,1}-\mu_{1,2}-\mu_{2,2}+\mu_{1,2}-\hat{\mu}_{1,2}+\mu_{2,2}-\hat{\mu}_{2,2}| \\ &\geq |\mu_{1,1}+\mu_{2,1}-\mu_{1,2}-\mu_{2,2}|-|\hat{\mu}_{\mathsf{sum},1}-\mu_{\mathsf{sum},1}|-|\mu_{1,2}-\hat{\mu}_{1,2}|-|\mu_{2,2}-\hat{\mu}_{2,2}| \\ &\geq |\mu_{1,1}+\mu_{2,1}-\mu_{1,2}-\mu_{2,2}|-3\gamma.\end{aligned}$$

$$\begin{aligned}|\hat{\mu}_{\mathsf{diff},1}-\hat{\mu}_{1,2}+\hat{\mu}_{2,2}| &\geq |\hat{\mu}_{\mathsf{diff},1}-\mu_{\mathsf{diff},1}+\mu_{1,1}-\mu_{2,1}-\mu_{1,2}+\mu_{2,2}+\mu_{1,2}-\hat{\mu}_{1,2}-\mu_{2,2}+\hat{\mu}_{2,2}| \\ &\geq |\mu_{1,1}-\mu_{2,1}-\mu_{1,2}+\mu_{2,2}|-|\hat{\mu}_{\mathsf{sum},1}-\mu_{\mathsf{sum},1}|-|\mu_{1,2}-\hat{\mu}_{1,2}|-|\mu_{2,2}-\hat{\mu}_{2,2}| \\ &\geq |\mu_{1,1}-\mu_{2,1}-\mu_{1,2}+\mu_{2,2}|-3\gamma.\end{aligned}$$

Similar guarantees also exist for $\mu_{\mathsf{sum},2}$ and $\mu_{\mathsf{diff},2}$ and therefore we must have

$$\begin{aligned}&|\hat{\mu}_{\mathsf{sum},2}-\hat{\mu}_{1,2}-\hat{\mu}_{2,2}| \leq 3\gamma \\ &|\hat{\mu}_{\mathsf{diff},2}-\hat{\mu}_{1,2}+\hat{\mu}_{2,2}| \leq 3\gamma \\ &|\hat{\mu}_{\mathsf{sum},2}-\hat{\mu}_{1,2}-\hat{\mu}_{2,1}| \geq |\mu_{2,1}-\mu_{2,2}|-3\gamma \\ &|\hat{\mu}_{\mathsf{diff},2}-\hat{\mu}_{1,2}+\hat{\mu}_{2,1}| \geq |\mu_{2,1}-\mu_{2,2}|-3\gamma \\ &|\hat{\mu}_{\mathsf{sum},2}-\hat{\mu}_{1,1}-\hat{\mu}_{2,2}| \geq |\mu_{1,1}-\mu_{1,2}|-3\gamma \\ &|\hat{\mu}_{\mathsf{diff},2}-\hat{\mu}_{1,1}+\hat{\mu}_{2,2}| \geq |\mu_{1,1}-\mu_{1,2}|-3\gamma \\ &|\hat{\mu}_{\mathsf{sum},2}-\hat{\mu}_{1,1}-\hat{\mu}_{2,1}| \geq |\mu_{1,1}+\mu_{2,1}-\mu_{1,2}-\mu_{2,2}|-3\gamma \\ &|\hat{\mu}_{\mathsf{diff},2}-\hat{\mu}_{1,1}-\hat{\mu}_{2,1}| \geq |\mu_{1,1}-\mu_{2,1}-\mu_{1,2}+\mu_{2,2}|-3\gamma\end{aligned}$$

Let us consider the case when $|\mu_{1,1}-\mu_{1,2}| \geq 9\gamma$ and $|\mu_{2,1}-\mu_{2,2}| \geq 9\gamma$. We will have

$$\begin{array}{lll}|\hat{\mu}_{\mathsf{sum},1}-\hat{\mu}_{1,1}-\hat{\mu}_{2,1}| \leq 3\gamma & |\hat{\mu}_{\mathsf{sum},1}-\hat{\mu}_{1,1}-\hat{\mu}_{2,2}| \geq 6\gamma & |\hat{\mu}_{\mathsf{sum},1}-\hat{\mu}_{1,2}-\hat{\mu}_{2,1}| \geq 6\gamma \\ |\hat{\mu}_{\mathsf{diff},1}-\hat{\mu}_{1,1}+\hat{\mu}_{2,1}| \leq 3\gamma & |\hat{\mu}_{\mathsf{diff},1}-\hat{\mu}_{1,1}+\hat{\mu}_{2,2}| \geq 6\gamma & |\hat{\mu}_{\mathsf{diff},1}-\hat{\mu}_{1,2}+\hat{\mu}_{2,1}| \geq 6\gamma \\ |\hat{\mu}_{\mathsf{sum},2}-\hat{\mu}_{1,2}-\hat{\mu}_{2,2}| \leq 3\gamma & |\hat{\mu}_{\mathsf{sum},2}-\hat{\mu}_{1,1}-\hat{\mu}_{2,2}| \geq 6\gamma & |\hat{\mu}_{\mathsf{sum},1}-\hat{\mu}_{1,2}-\hat{\mu}_{2,1}| \geq 6\gamma \\ |\hat{\mu}_{\mathsf{diff},2}-\hat{\mu}_{1,2}+\hat{\mu}_{2,2}| \leq 3\gamma & |\hat{\mu}_{\mathsf{diff},1}-\hat{\mu}_{1,1}+\hat{\mu}_{2,2}| \geq 6\gamma & |\hat{\mu}_{\mathsf{diff},1}-\hat{\mu}_{1,2}+\hat{\mu}_{2,1}| \geq 6\gamma\end{array}$$

Moreover, at least one of the following two must be true:

$$\begin{array}{c}|\hat{\mu}_{\mathsf{sum},1}-\hat{\mu}_{1,2}-\hat{\mu}_{2,2}| \geq 15\gamma \quad \text{and} \quad |\hat{\mu}_{\mathsf{sum},2}-\hat{\mu}_{1,1}-\hat{\mu}_{2,1}| \geq 15\gamma \\ \text{or} \quad |\hat{\mu}_{\mathsf{diff},1}-\hat{\mu}_{1,2}+\hat{\mu}_{2,2}| \geq 15\gamma \quad \text{and} \quad |\hat{\mu}_{\mathsf{diff},2}-\hat{\mu}_{1,1}+\hat{\mu}_{2,1}| \geq 15\gamma\end{array}$$

depending on whether $\mu_{1,1}-\mu_{1,2}$ and $\mu_{2,1}-\mu_{2,2}$ have the same sign or not. This shows that either for the sum query or for the difference query, only the correct set of means is closest to their corresponding value (sum or difference) and any wrong choice of means is away from that particular value (sum or difference). Hence the lemma is proved.

D. Proof of Lemma 6 and 7

Proof of Lemma 6. Notice that for a particular query $\mathbf{x}^i, i \in [m]$, the difference of the means $\mu_{i,1}-\mu_{i,2}$ is distributed according to

$$\mu_{i,1}-\mu_{i,2} \sim \mathcal{N}(0, ||\beta^1-\beta^2||_2^2).$$

Therefore, we have

$$\Pr(|\mu_{i,1} - \mu_{i,2}| \leq 13\gamma) = \int_{-13\gamma}^{13\gamma} \frac{e^{-x^2/2||\beta^1-\beta^2||_2^2}}{\sqrt{2\pi||\beta^1-\beta^2||_2^2}} dx \leq \frac{13\sqrt{2}\gamma}{\sqrt{\pi}||\beta^1-\beta^2||_2}. \tag{2}$$

where the upper bound is obtained by using $e^{-x^2/2||\beta^1-\beta^2||_2^2} \leq 1$. Therefore the probability that for all the m' queries $\{\mathbf{x}^i\}_{i=1}^m$, the difference between the means is less than 13γ must be

$$\Pr(\bigcup_{i=1}^{m'} \mu_{i,1} - \mu_{i,2} \leq 13\gamma) = \prod_{i=1}^{m'} \Pr(\mu_{i,1} - \mu_{i,2} \leq 13\gamma) \leq \Big(\frac{13\sqrt{2}\gamma}{\sqrt{\pi}||\beta^1-\beta^2||_2}\Big)^{m'} = e^{-m' \log \frac{\sqrt{\pi}||\beta^1-\beta^2||_2}{13\sqrt{2}\gamma}}$$

Therefore for $m' = \Big\lceil \log \eta^{-1} / \log \frac{\sqrt{\pi}||\beta^1-\beta^2||_2}{13\sqrt{2}\gamma} \Big\rceil$, we have that $\Pr(\bigcup_{i=1}^{m'} \mu_{i,1} - \mu_{i,2} \leq 13\gamma) \leq \eta$. $\square$

Proof of Lemma 7. The proof of Lemma 7 is very similar to the proof of Lemma 6. Again for a particular query $\mathbf{x}^i, i \in [m]$, the difference of the means $\mu_{i,1} - \mu_{i,2}$ is distributed according to

$$\mu_{i,1} - \mu_{i,2} \sim \mathcal{N}(0, ||\beta^1 - \beta^2||_2^2).$$

Therefore, we have for any constant $c_1 > 0$,

$$\Pr(|\mu_{i,1} - \mu_{i,2}| \leq c_1\sigma) = \int_{-c_1\sigma}^{c_1\sigma} \frac{e^{-x^2/2||\beta^1-\beta^2||_2^2}}{\sqrt{2\pi||\beta^1-\beta^2||_2^2}} dx \leq \frac{\sqrt{2}c_1\sigma}{\sqrt{\pi}||\beta^1-\beta^2||_2}.$$

where the upper bound is obtained by using $e^{-x^2/2||\beta^1-\beta^2||_2^2} \leq 1$. Hence the lemma is proved by substituting $c_2 = \sqrt{2}c_1/\sqrt{\pi}$. $\square$

E. Proof of Theorem 1 ($\gamma = \Omega(\left|\left|\beta^1 - \beta^2\right|\right|_2)$)

.

Algorithm 9 RECOVER UNKNOWN VECTORS $2(\sigma, \gamma)$ Recover the unknown vectors β^1 and β^2

Set $m = c_s k \log n$ and $T = O\Big(\Big\lceil \sigma^2 \log k / (\gamma - 0.8\left|\left|\beta^1 - \beta^2\right|\right|_2)^2 \Big\rceil\Big)$.
Sample $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^m \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ independently.
for $i = 1, 2, \ldots, m$ **do**
    Compute $\hat{\mu}_i$ by running Algorithm FIT A SINGLE GAUSSIAN $(\mathbf{x}^i, T)$.
end for
Set $\mathbf{u}$ to be the m-dimensional vector whose i^{th} element is $\hat{\mu}_i$.
Set $\mathbf{A}$ to be the $m \times n$ matrix such that its i^{th} row is $\mathbf{x}^i$, with each entry normalized by $\sqrt{m}$.
Set $\hat{\beta}$ to be the solution of the optimization problem $\min_{\mathbf{z} \in \mathbb{R}^n} ||\mathbf{z}||_1$ s.t. $||\mathbf{A}\mathbf{z} - \frac{1}{\sqrt{m}}\mathbf{u}||_2 \leq \gamma$
Return $\hat{\beta}$.

We will first assume $\gamma > 0.8\left|\left|\beta^1 - \beta^2\right|\right|_2$ to prove the claim, and later extend this to any $\gamma = \Omega(\left|\left|\beta^1 - \beta^2\right|\right|_2)$. The recovery procedure in the setting when $\gamma > 0.8\left|\left|\beta^1 - \beta^2\right|\right|_2$ is described in Algorithm 9. We will start by proving the following claim

Claim 5. *Algorithm 9 returns a vector* $\mathbf{u}$ *of length* m *using* $O\Big(m\Big\lceil \sigma^2 \log \eta^{-1}/\epsilon^2 \Big\rceil\Big)$ *queries such that*

$$\left|\mathbf{u}[i] - \langle \mathbf{x}^i, \beta^1 \rangle\right| \leq \epsilon + \frac{\left|\langle \mathbf{x}^i, \beta^1 - \beta^2 \rangle\right|}{2}$$
$$\left|\mathbf{u}[i] - \langle \mathbf{x}^i, \beta^2 \rangle\right| \leq \epsilon + \frac{\left|\langle \mathbf{x}^i, \beta^1 - \beta^2 \rangle\right|}{2}$$

for all $i \in [m]$ *with probability at least* $1 - m\eta$.

Proof. In Algorithm 9, for each query $\mathbf{x}^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, we can use a batchsize of $O\Big(\Big\lceil \sigma^2 \log \eta^{-1}/\epsilon^2 \Big\rceil\Big)$ to recover $\hat{\mu}_i$ such that

$$\left|\hat{\mu}_i - \frac{\langle \mathbf{x}^i, \beta^1 \rangle + \langle \mathbf{x}^i, \beta^2 \rangle}{2}\right| \le \epsilon$$

with probability at least $1-\eta$ according to the guarantees of Lemma 3. We therefore have

$$\begin{aligned}\left|\hat{\mu}_i - \langle \mathbf{x}^i, \beta^1 \rangle\right| &\le \left|\hat{\mu}^i - \frac{\langle \mathbf{x}^i, \beta^1 \rangle + \langle \mathbf{x}^i, \beta^2 \rangle}{2}\right| \\ &+ \frac{\left|\langle \mathbf{x}^i, \beta^1 - \beta^2 \rangle\right|}{2} \le \epsilon + \frac{\left|\langle \mathbf{x}^i, \beta^1 - \beta^2 \rangle\right|}{2}.\end{aligned}$$

where the last inequality follows by using the guarantees on $\hat{\mu}_i$. We can show a similar chain of inequalities for $\left|\hat{\mu}_i - \langle \mathbf{x}^i, \beta^2 \rangle\right|$ and finally take a union bound over all $i \in [m]$ to conclude the proof of the claim. ◻

Next, let us define the random variable $\omega_i \triangleq \left|\langle \mathbf{x}^i, \beta^1 - \beta^2 \rangle\right|$ where the randomness is over $\mathbf{x}^i$. Subsequently let us define the m-dimensional vector $\mathbf{b}$ whose i^{th} element is $\epsilon + \omega_i/2$. Again, for $m \ge c_s k \log n$, let $\mathbf{A}$ denote the matrix whose i^{th} row is $\mathbf{x}^i$ normalized by $\sqrt{m}$.

Claim 6. *We must have*

$$\left|\left|\mathbf{A}\beta^1 - \frac{\mathbf{u}}{\sqrt{m}}\right|\right|_2 \le \frac{||\mathbf{b}||_2}{\sqrt{m}} \;\&\; \left|\left|\mathbf{A}\beta^2 - \frac{\mathbf{u}}{\sqrt{m}}\right|\right|_2 \le \frac{||\mathbf{b}||_2}{\sqrt{m}}.$$

Proof. The proof of the claim is immediate from definition of $\mathbf{A}$ and $\mathbf{b}$. ◻

Next, we show high probability bounds on ℓ_2-norm of the vector $\mathbf{b}$ in the following claim.

Claim 7. *We must have* $\frac{||\mathbf{b}||_2^2}{m} \le 2\epsilon^2 + 0.64\left|\left|\beta^1 - \beta^2\right|\right|_2^2$ *with probability at least* $1 - O(e^{-m})$.

Proof. Notice that

$$\frac{||\mathbf{b}||_2^2}{m} \le \frac{1}{m}\sum_{i=1}^m \Big(\epsilon + \frac{\omega_i}{2}\Big)^2 \le \frac{2}{m}\sum_{i=1}^m \Big(\epsilon^2 + \frac{\omega_i^2}{4}\Big).$$

Using the fact that $\mathbf{x}^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ and by definition, we must have that ω_i is a random variable distributed according to $\mathcal{N}(0, \left|\left|\beta^1 - \beta^2\right|\right|_2)$. Therefore, we have (see Lemma 1.2 (Boche et al., 2015))

$$\begin{aligned}&\Pr\left(\left|\sum_{i=1}^m \omega_i^2 - m\left|\left|\beta^1 - \beta^2\right|\right|_2^2\right| \ge m\rho\left|\left|\beta^1 - \beta^2\right|\right|_2^2\right) \\ &\le 2\exp\Big(-\frac{m}{2}\Big(\frac{\rho^2}{2} - \frac{\rho^3}{3}\Big)\Big)\end{aligned}$$

for $0 < \rho < 1$. Therefore, by substituting $\rho = 0.28$, we get that

$$\frac{2}{m}\sum_{i=1}^m \Big(\epsilon^2 + \frac{\omega_i^2}{4}\Big) \le 2\epsilon^2 + 0.64\left|\left|\beta^1 - \beta^2\right|\right|_2^2$$

with probability at least $1 - O(e^{-m})$. ◻

From Claim 7, we get

$$\frac{||\mathbf{b}||_2}{\sqrt{m}} \le \sqrt{2}\epsilon + 0.8\left|\left|\beta^1 - \beta^2\right|\right|_2.$$

where we use the inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b > 0$. Subsequently we solve the following convex optimization problem

$$\min_{\mathbf{z} \in \mathbb{R}^n} ||\mathbf{z}||_1 \text{ s.t. } ||\mathbf{A}\mathbf{z} - \frac{\mathbf{u}}{\sqrt{m}}||_2 \le \gamma$$

where $\gamma = \sqrt{2}\epsilon + 0.8 \left|\left|\beta^1 - \beta^2\right|\right|_2$ in order to recover $\hat{\beta}$ and return it as estimate of both β^1, β^2. For $m = O(k \log n), \eta = (m \log n)^{-1}$ and $\sqrt{2}\epsilon = \gamma - 0.8 \left|\left|\beta^1 - \beta^2\right|\right|_2$, the number of queries required is $O\Big(k \log n \Big\lceil \sigma^2 \log k / (\gamma - 0.8 \left|\left|\beta^1 - \beta^2\right|\right|_2)^2 \Big\rceil\Big)$. Further, by using the theoretical guarantees provided in Theorem 1.6 in (Boche et al., 2015), we obtain the guarantees of the main theorem with error probability atmost $o(1)$. Again, by substituting the definition of the Noise Factor $\mathsf{NF} = \gamma/\sigma$ and the Signal to Noise ratio $\mathsf{SNR} = O(\left|\left|\beta^1 - \beta^2\right|\right|_2^2 / \sigma^2)$, we obtain the query complexity to be

$$O\Big(k \log n \Big\lceil \frac{\log k}{(\mathsf{NF} - 0.8\sqrt{\mathsf{SNR}})^2} \Big\rceil\Big).$$

Now let us assume any $\gamma = \Omega(\left|\left|\beta^1 - \beta^2\right|\right|_2)$. If the desired $\gamma < \left|\left|\beta^1 - \beta^2\right|\right|_2$, then one can just define $\gamma' = \left|\left|\beta^1 - \beta^2\right|\right|_2$ and obtain a precision γ' which is a constant factor within γ. Further, the query complexity also becomes independent of the noise factor since $\mathsf{NF} = \sqrt{\mathsf{SNR}}$ for this choice of γ' and thus we obtain the promised query complexity in Theorem 1.

F. Discussion on Noiseless Setting $\sigma = 0$

Step 1: In the noiseless setting, we obtain $m = O(k \log n)$ query vectors $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^m$ sampled i.i.d according to $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ and repeat each of them for $2 \log m$ times. For a particular query $\mathbf{x}_i$, the probability that we do not obtain any samples from β^1 or β^2 is at most $(1/2)^{2 \log m}$. We can take a union bound to conclude that for all queries, we obtain samples from both β^1 and β^2 with probability at least $1 - O(m^{-1})$. Further note that for each query $\mathbf{x}^i$, $\langle \mathbf{x}^i, \beta^1 - \beta^2 \rangle$ is distributed according to $\mathcal{N}(0, \left|\left|\beta^1 - \beta^2\right|\right|_2^2)$ and therefore, it must happen with probability 1 that $\langle \mathbf{x}^i, \beta^1 \rangle \ne \langle \mathbf{x}^i, \beta^1 \rangle$. Thus for each query $\mathbf{x}_i$, we can recover the tuple $(\langle \mathbf{x}^i, \beta^1 \rangle, \langle \mathbf{x}^i, \beta^2 \rangle)$ but we cannot recover the ordering i.e. we do not know which element of the tuple corresponds to β^1 and which one to β^2.

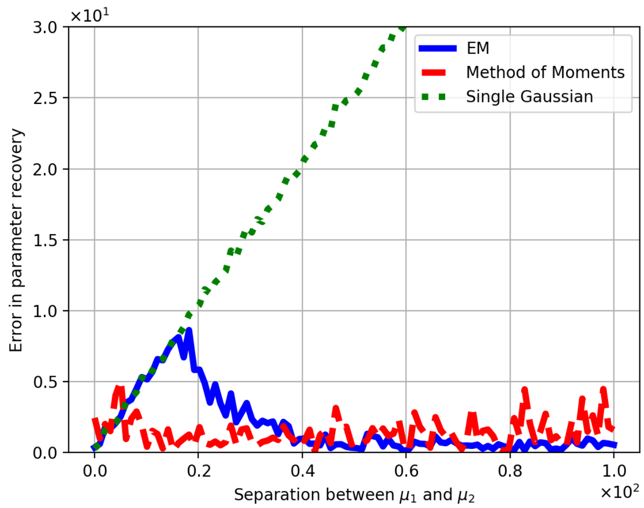
Step 2: Note that we are still left with the alignment step where we need to cluster the $2m$ recovered parameters $\{(\langle \mathbf{x}^i, \beta^1 \rangle, \langle \mathbf{x}^i, \beta^2 \rangle)\}_{i=1}^m$ into two clusters of size m each so that there exists exactly one element from each tuple in each of the two clusters and all the elements in the same cluster correspond to the same unknown vector. In order to complete this step, we use ideas from (Krishnamurthy et al., 2019). We query $\mathbf{x}_1 + \mathbf{x}_i$ and $\mathbf{x}_1 - \mathbf{x}_i$ for all $i \ne 1$ each for $2 \log m$ times to the oracle and recover the tuples $(\langle \mathbf{x}^1 + \mathbf{x}^i, \beta^1 \rangle, \langle \mathbf{x}^1 + \mathbf{x}^i, \beta^2 \rangle)$ and $(\langle \mathbf{x}^1 - \mathbf{x}^i, \beta^1 \rangle, \langle \mathbf{x}^1 - \mathbf{x}^i, \beta^2 \rangle)$ for all $i \ne 1$. For a particular $i \in [m] \setminus \{1\}$, we will choose two elements (say a and b) from the pairs $(\langle \mathbf{x}_1, \beta^1 \rangle, \langle \mathbf{x}_1, \beta^2 \rangle)$ and $(\langle \mathbf{x}_i, \beta^1 \rangle, \langle \mathbf{x}_i, \beta^2 \rangle)$ (one element from each pair) such that their sum belongs to the pair $\langle \mathbf{x}_1 + \mathbf{x}_i, \beta^1 \rangle, \langle \mathbf{x}_1 + \mathbf{x}_i, \beta^2 \rangle$ and their difference belongs to the pair $\langle \mathbf{x}_1 - \mathbf{x}_i, \beta^1 \rangle, \langle \mathbf{x}_1 - \mathbf{x}_i, \beta^2 \rangle$. In our algorithm, we will put a, b into the same cluster and the other two elements into the other cluster. From construction, we must put $(\langle \mathbf{x}_1, \beta^1 \rangle, \langle \mathbf{x}_i, \beta^1 \rangle)$ in one cluster and $(\langle \mathbf{x}_1, \beta^2 \rangle, \langle \mathbf{x}_i, \beta^2 \rangle)$ in other. Note that a failure in this step is not possible because the $2m$ recovered parameters are different from each other with probability 1.

Step 3: Once we have clustered the samples, we have reduced our problem to the usual compressed sensing setting (with only 1 unknown vector) and therefore we can run the well known convex optimization routine in order to recover the unknown vectors β^1 and β^2. The total query complexity is $O(k \log n \log k)$.

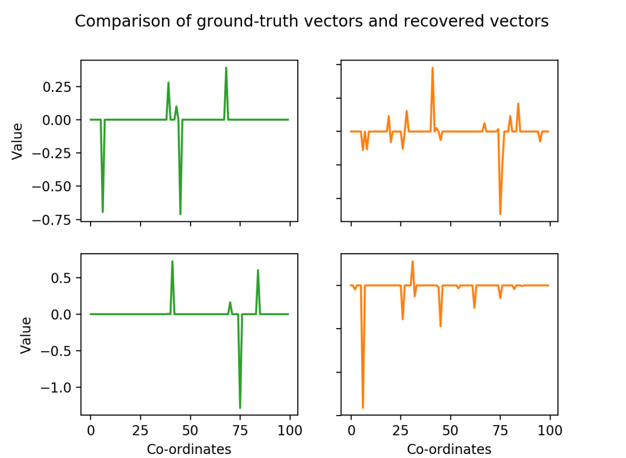
G. 'Proof of Concept' Simulations

The methods of parameter recovery in Gaussian mixtures are compared in Fig 1a. As claimed in Sec. 2.2, the EM starts performing better than the method of moments when the gap between the parameters is large.

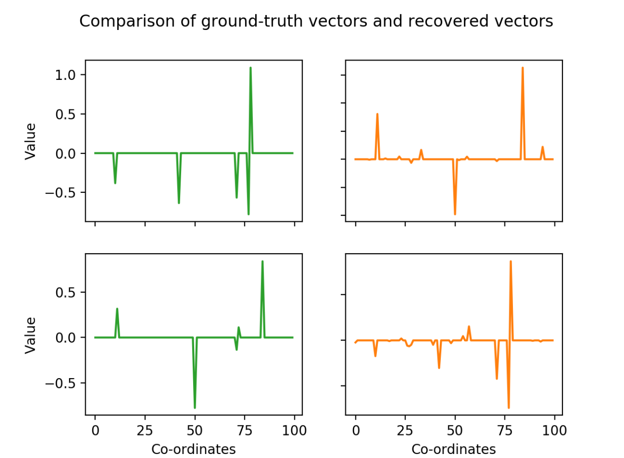
We have also run Algorithm 8 for different set of pairs of sparse vectors and example recovery results for visualization are shown in Figures 1b and 1c. Note that, while quite accurate reconstruction is possible the vectors are not reconstructed in order, as to be expected.



(a) Comparison of the three techniques for recovery of parameters of a Gaussian mixture with 1000 samples (see Algorithms 1,3 and 4). The error in parameter recovery is plotted with separation between μ_1 and μ_2 (by keeping μ_1 fixed at 0 and varying μ_2).



(b) The 100-dimensional ground truth vectors β^1 and β^2 with sparsity $k = 5$ plotted in green (left) and the recovered vectors (using Algorithm 8) $\hat{\beta}^1$ and $\hat{\beta}^2$ plotted in orange (right) using a batch-size ~ 100 for each of 150 random gaussian queries. The order of the recovered vectors and the ground truth vectors is reversed.



(c) The 100-dimensional ground truth vectors β^1 and β^2 with sparsity $k = 5$ plotted in green (left) and the recovered vectors (using Algorithm 8) $\hat{\beta}^1$ and $\hat{\beta}^2$ plotted in orange (right) using a batch-size ~ 600 for each of 150 random gaussian queries. The order of the recovered vectors and the ground truth vectors is reversed.

Figure 1. Simulation results of our techniques.