Supplement for "Sharp Analysis of Expectation-Maximization for Weakly Identifiable Models"

Contents

A Proofs of main results

In this section, we present the proofs for our main results while deferring some technical results to the appendices.

A.1 Proof of Theorem 1

Our result makes use of the following corollary (proven in Appendix C.1):

Corollary 1. *Given constants $\delta \in (0,1)$ and $\beta \in (0, 1/12]$, suppose that we generate the the sample-level EM sequence $\theta_n^{t+1} = M_{n,1}(\theta_n^t)$ starting from an initialization $|\theta_n^0| \in I'_\beta$, and using a sample size n lower bounded as*

$n \gtrsim \log^{1/(12\beta)}(\log(1/\beta)/\delta)$. *Then for all iterations* $t \geq n^{1/2-6\beta}\log(n)\log(1/\beta)$, *we have*

$$|\theta_n^t - \theta^*| \leq c_1 \left(\frac{1}{n}\log\frac{\log(1/\beta)}{\delta}\right)^{\frac{1}{12}-\beta}, \tag{15}$$

with probability at least $1-\delta$.

Remark: We note that the sub-optimal bound (15) obtained from Corollary 1 is not an artifact of the localization argument and arises due to the definition of the operator operator $\widetilde{M}_{n,1}$ (11a). As we have alluded to earlier, indeed a finer analysis with the population EM operator $\overline{M}_1$ is required to prove the rate of $n^{-1/8}$ stated in Theorem 1. However, a key assumption in the further derivation is that the sample EM iterates θ_n^t can converge to a ball of radius $r \precsim n^{-1/16}$ around θ^* in a finite number of steps, for which Corollary 1 comes in handy.

We now begin with a sketch the two stage-argument, and then provide a rigorous proof for Theorem 1.

A.1.1 Proof sketch

As mentioned earlier, the pseudo-population operator $\widetilde{M}_{n,1}$ is not sufficient to achieve the sharp rate of EM iterates under the univariate symmetric Gaussian mixture fit. Therefore, we make use of corrected-population operator $\overline{M}_1$ to get a sharp statistical rate of EM. Our proof for the tight convergence rate of sample EM updates relies on a novel two-stage localization argument that we are going to sketch.

First stage argument: Plugging in $\beta = 1/84$ in Corollary 1, we obtain that for $t \gtrsim \sqrt{n}\log(n)$, with probability at least $1-\delta$ we have that

$$\left|\theta_n^t - \theta^*\right| \leq cn^{-\frac{1}{14}}\log^{\frac{1}{14}}\frac{\log(1/\beta)}{\delta} \leq n^{-\frac{1}{16}}, \tag{16}$$

where the second inequality follows from the large sample condition $n \geq c'\log^8\frac{\log 84}{\delta}$. All the following claims are made conditional on the event (16).

Second stage argument: In order to keep the presentation of the proof sketch simple, we do not track constant and logarithmic factors in the arguments to follow. In epoch ℓ, for any iteration t the EM iterates satisfy $\theta_n^t \in [n^{-a_{\ell+1}}, n^{-a_\ell}]$ where $a_{\ell+1} > a_\ell$ and $a_\ell \leq 1/16$. Applying Lemma 1 for such iterations, we find that with high probability

$$\left|\overline{M}_1(\theta_n^t)\right| \precsim \underbrace{(1-n^{-6a_{\ell+1}})}_{=:\gamma_\ell}\left|\theta_n^t\right| \quad \text{and} \quad \left|M_{n,1}(\theta_n^t) - \overline{M}_1(\theta_n^t)\right| \precsim \frac{n^{-3a_\ell}}{\sqrt{n}},$$

where the first bound follows from the $1-c\theta^6$ contraction bound (12b) and the second bound follows from the cubic-type Rademacher bound (12d). Invoking the basic triangle inequality T times, we obtain that

$$\left|\theta_n^{t+T}\right| \overset{(i)}{\precsim} e^{-Tn^{-6a_{\ell+1}}}n^{-a_\ell} + \frac{1}{1-\gamma_\ell}\cdot\frac{n^{-3a_\ell}}{\sqrt{n}} \overset{(ii)}{\precsim} \frac{1}{1-\gamma_\ell}\cdot\frac{n^{-3a_\ell}}{\sqrt{n}} = n^{6a_{\ell+1}-3a_\ell-1/2},$$

where in step (ii) we have used the fact that for large enough T, the first term is dominated by the second term in the RHS of step (i). To obtain a recursion for the sequence a_ℓ, we set the RHS equal to $n^{-a_{\ell+1}}$. Doing so yields the recursion

$$a_{\ell+1} = \frac{3a_\ell}{7} + \frac{1}{14}, \quad \text{where} \quad a_0 = 1/16. \tag{17a}$$

Solving for the limit $a_{\ell+1} = a_\ell = a_\star$, we find that $a_\star = 1/8$. Thus, we can conclude that sample EM iterates in the univariate setting converge to a ball of radius $n^{-1/8}$ as claimed in the theorem statement.

A.1.2 Formal proof of sample EM convergence rate

We now turn to providing a formal proof for the preceding arguments.

Notations: To make the proof comprehensible, some additional notations are necessary which we collect here. Let $\ell_\star = \lceil \log(8/\beta)/\log(7/3) \rceil$ so that $a_{\ell_\star} \leq 1/8 - \beta$. We define the following shorthand:

$$\omega := \frac{n}{c_{n,\delta}}, \quad \text{where} \quad c_{n,\delta} := \log^{10}(10n(\ell_\star + 1)/\delta). \tag{17b}$$

For $\ell = 0, \ldots, \ell_\star$, we define the time sequences t_ℓ and T_ℓ as follows:

$$t_0 = \sqrt{n}, \quad t_\ell = \left\lceil 10\omega^{6a_\ell} \log \omega \right\rceil, \quad \text{and} \quad T_\ell = \sum_{j=0}^{\ell} t_j. \tag{17c}$$

Direct computation leads to

$$T_{\ell_\star} \leq \sqrt{n} + \ell_\star t_{\ell_\star} \precsim \log\left(\frac{n \log \frac{1}{\beta}}{c_{n,\delta}\delta}\right)\left(\frac{n}{c_{n,\delta}}\right)^{3/4-6\beta} \precsim n^{3/4}. \tag{17d}$$

In order to facilitate the proof argument later, we define the following set

$$\mathcal{R} := \left\{\omega^{-a_1}, \ldots, \omega^{-a_{\ell_\star}}, c'\omega^{-a_1}, \ldots, c'\omega^{-a_{\ell_\star}}\right\}, \tag{17e}$$

where $c' := (5c_2 + 1)$. Here, c_2 is the universal constant from Lemma 1.

Formal argument: We show that with probability at least $1 - \delta$ the following holds:

$$\left|\theta_n^t\right| \leq \left(\frac{c_{n,\delta}}{n}\right)^{a_\ell} = \omega^{-a_\ell}, \quad \text{for all } t \geq T_\ell, \text{ and } \ell \leq \ell_\star. \tag{18}$$

As a consequence of this claim and the definitions (17a)-(17d) of $a_{\ell_\star}$ and $T_{\ell_\star}$, we immediately obtain that

$$|\theta_n^t - \theta^*| \precsim \left(\frac{c_{n,\delta}}{n}\right)^{1/8-\beta} \precsim \left(\frac{1}{n}\log^{10}\frac{10n\log(8/\beta)}{\delta}\right)^{1/8-\beta},$$

for all number of iterates $t \succsim n^{3/4-6\beta}\log(n)\log(1/\beta)$ with probability at least $1 - \delta$ as claimed in Theorem 1.

We now define the high probability event that is crucial for our proof. For any $r \in \mathcal{R}$, define the event E_r as follows

$$E_r := \left\{ \sup_{\theta \in \mathbb{B}(0,r)} \left|M_{n,1}(\theta) - \overline{M}_1(\theta)\right| \leq c_2 r^3 \sqrt{\frac{\log^{10}(5n\,|\mathcal{R}|\,/\delta)}{n}} \right\}.$$

Then, for the event

$$\mathcal{E} := \bigcap_{r \in \mathcal{R}} E_r \cap \{\text{Event (16) holds }\}, \tag{19}$$

applying the union bound with Lemma 1 yields that $\mathbb{P}[\mathcal{E}] \geq 1 - \delta$. All the arguments that follow are conditional on the event $\mathcal{E}$ and hence hold with the claimed high probability.

In order to prove the claim (18), we make use of the following intermediate claim:

Lemma 2. *Conditional on the event $\mathcal{E}$, if $|\theta| \leq \omega^{-a_\ell}$, then $|M_{n,1}(\theta)| \leq \omega^{-a_\ell}$ for any $\ell \leq \ell_\star$.*

Deferring the proof of Appendix C.2, we now establish the claim (18) conditional on the event $\mathcal{E}$ only for $t = T_\ell$ and when $|\theta_n^t| \in [\omega^{-a_{\ell+1}}, \omega^{-a_\ell}]$ in which we now prove using induction.

Proof of base case $\ell = 0$: Note that we have $a_0 = 1/16$ and that $n^{-1/16} \leq \omega^{1/16}$. Also, by the definition (19) we have that the event (16) $\subseteq \mathcal{E}$. Hence, under the event $\mathcal{E}$ we have that $|\theta_n^t| \leq n^{-1/16}$, for $t \succsim \sqrt{n}\log(n)$. Putting all the pieces together, we find that under the event $\mathcal{E}$, we have $|\theta_n^t| \leq n^{-1/16} \leq \omega^{1/16}$ and the base case follows.

Proof of inductive step: We now establish the inductive step. Note that Lemma 2 implies that we need to show the following: if $|\theta_n^t| \leq \omega^{-a_\ell}$ for all $t \in \{T_\ell, T_\ell + 1, \ldots, T_{\ell+1} - 1\}$ for any given $\ell \leq \ell_\star$, then $|\theta_n^{T_{\ell+1}}| \leq \omega^{-a_{\ell+1}}$. We establish this claim in two steps:

$$\theta_n^{T_\ell + t_\ell/2} \leq c' \omega^{-a_{\ell+1}}, \quad \text{and,} \tag{20a}$$

$$\theta_n^{T_{\ell+1}} \leq \omega^{-a_{\ell+1}}, \tag{20b}$$

where $c' = (5c_2 + 1) \geq 1$ is a universal constant. Note that the inductive claim follows from the bound (20b). It remains to establish the two claims (20a) and (20b) which we now do one by one.

Proof of claim (20a): Let $\Theta_\ell = \{\theta : |\theta| \in [\omega^{-a_{\ell+1}}, \omega^{-a_\ell}]\}$. Now, conditional on the event $\mathcal{E}$, Lemma 1 implies that

$$\sup_{\theta \in \Theta_\ell} \left| M_{n,1}(\theta) - \overline{M}_1(\theta) \right| \leq c_2 \omega^{-3a_\ell - 1/2}, \quad \text{and} \quad \sup_{\theta \in \Theta_\ell} \left| \overline{M}_1(\theta)/\theta \right| \leq (1 - \omega^{-6a_{\ell+1}}/5) =: \gamma_\ell.$$

We can check that $\gamma_\ell \leq e^{-\omega^{6a_{\ell+1}}/5}$. Unfolding the basic triangle inequality $t_\ell/2$ times and noting that $\theta_n^t \in \Theta_\ell$ for all $t \in \{T_\ell, \ldots, T_\ell + t_\ell/2\}$, we obtain that

$$\begin{aligned}
\left| \theta_n^{T_\ell + t_\ell/2} \right| &\leq \gamma_\ell^{t_\ell/2} \left| \theta_n^{T_\ell} \right| + (1 + \gamma_\ell + \ldots + \gamma_\ell^{t_\ell/2 - 1}) c_2 \omega^{-3a_\ell - 1/2} \\
&\leq e^{-t_\ell \omega^{-6a_{\ell+1}}/10} \omega^{-a_\ell} + \frac{1}{1 - \gamma_\ell} c_2 \omega^{-3a_\ell - 1/2} \\
&\overset{(i)}{\leq} (1 + 5c_2) \omega^{6a_{\ell+1} - 3a_\ell - 1/2} \\
&\overset{(ii)}{=} (5c_2 + 1) \omega^{-a_{\ell+1}}
\end{aligned}$$

where step (i) follows from plugging in the value of γ_ℓ and invoking the definition (17c) of t_ℓ, which leads to

$$e^{-t_\ell \omega^{6a_{\ell+1}}/10} \omega^{-a_\ell} \leq \omega^{6a_{\ell+1} - 3a_\ell - 1/2}.$$

Moreover, step (ii) is a direct consequence of the definition (17a) of the sequence a_ℓ. Therefore, we achieve the conclusion of claim (20a).

Proof of claim (20b): The proof of this step is very similar to the previous step, except that we now use the set $\Theta'_\ell = \{\theta : |\theta| \in [\omega^{-a_{\ell+1}}, c'\omega^{-a_{\ell+1}}]\}$ for our arguments. Applying Lemma 1, we have

$$\sup_{\theta \in \Theta'_\ell} \left| M_{n,1}(\theta) - \overline{M}_1(\theta) \right| \leq c_2 (c')^3 \omega^{-3a_{\ell+1} - 1/2}, \quad \text{and} \quad \sup_{\theta \in \Theta'_\ell} \left| \overline{M}_1(\theta)/\theta \right| \leq \gamma_\ell.$$

Using the similar argument as that from the previous case, we find that

$$\begin{aligned}
\left| \theta_n^{T_\ell + t_\ell/2 + t_\ell/s2} \right| &\leq e^{-t_\ell \omega^{6a_{\ell+1}}/10} c' \omega^{-a_{\ell+1}} + \frac{1}{1 - \gamma_\ell} c_2 (c')^3 \omega^{-3a_{\ell+1} - 1/2} \\
&\leq (5c_2 + 1)(c')^3 \omega^{4a_{\ell+1} - 1/2} \cdot \omega^{-a_{\ell+1}} \\
&\overset{(i)}{\leq} \omega^{-a_{\ell+1}}
\end{aligned}$$

where step (i) follows from the inequality $e^{-t_\ell \omega^{6a_{\ell+1}}/10} \leq \omega^{4a_{\ell+1} - 1/2}$ and the inequality

$$\omega^{4a_{\ell+1} - 1/2} \leq \omega^{4a_{\ell_\star} - 1/2} \leq \omega^{-4\beta} \leq 1/(c')^4,$$

since $n \geq (c')^{1/\beta} c_{n,\delta}$. The claim now follows.

A.2 Proof of Theorem 2

Before proceeding further, we first derive the convergence rates for the scale parameter σ_n^t using Theorem 2. Noting that $(\theta^*, \sigma^*) = (0, 1)$, we obtain the following relation

$$\left| (\sigma_n^t)^2 - (\sigma^*)^2 \right| = \left| \frac{\sum_{i=1}^n \|X_i\|_2^2}{dn} - (\sigma^*)^2 - \frac{\|\theta_n^t - \theta^*\|_2^2}{d} \right|.$$

Using standard chi-squared bounds, we obtain that

$$\left| \frac{\sum_{i=1}^n \|X_i\|_2^2}{dn} - (\sigma^*)^2 \right| \precsim (nd)^{-\frac{1}{2}},$$

with high probability. From the bound (8), we also have $\|\theta_n^t - \theta^*\|_2^2/d \precsim (nd)^{-\frac{1}{2}}$. Putting the pieces together, we conclude that the statistical error for the scale parameter satisfies

$$|(\sigma_n^t)^2 - (\sigma^*)^2| \precsim (nd)^{-\frac{1}{2}} \quad \text{for all } t \succsim \left(\tfrac{n}{d}\right)^{\frac{1}{2}}, \tag{21}$$

with high probability. Consequently, in the sequel, we focus primarily on the convergence rate for the EM estimates θ_n^t of the location parameter, as the corresponding guarantee for the scale parameter σ_n^t is readily implied by it.

The proof of Theorem 2 is based on the population-to-sample analysis and follows a similar road-map as of the proofs in the paper [Dwivedi et al., 2020]. We first analyze the population-level EM operator and then using epoch-based-localization argument derive the statistical rates (8). We make use of the following d-dimensional analog of the pseudo-population operator (cf. equation (11a)):

$$\widetilde{M}_{n,d}(\theta) := \mathbb{E}_{Y \sim \mathcal{N}(0, I_d)} \left[Y \tanh \left(\frac{Y^\top \theta}{\sum_{j=1}^n \|X_j\|_2^2/(nd) - \|\theta\|^2/d} \right) \right]. \tag{22}$$

In the next lemma, we establish the contraction properties and the perturbation bounds for $\widetilde{M}_{n,d}$:

Lemma 3. *The operator $\widetilde{M}_{n,d}$ satisfies*

$$\left(1 - \frac{3\|\theta\|_2^2}{4}\right) \leq \frac{\|\widetilde{M}_{n,d}(\theta)\|_2}{\|\theta\|_2} \leq \left(1 - \frac{(1-1/d)\|\theta\|_2^2}{4}\right), \quad \text{for all } \|\theta\|_2 \in I_\beta, \tag{23a}$$

with probability at least $1-\delta$. Moreover, there exists a universal constant c_2 such that for any fixed $\delta \in (0,1)$, $\beta \in (0, \frac{1}{4}]$, and $r \in (0, \frac{1}{8})$ we have

$$\mathbb{P}\left[\sup_{\theta \in \mathbb{B}(0,r)} \|M_{n,d}(\theta) - \widetilde{M}_{n,d}(\theta)\|_2 \leq c_2 r \sqrt{\frac{d \log(1/\delta)}{n}} \right] \geq 1 - \delta - e^{-(nd)^{4\beta}/8}. \tag{23b}$$

See Appendix C.3 for the proof.

Lemma 3 shows that the operator $\widetilde{M}_{n,d}$ has a faster contraction (order $1 - \|\theta\|_2^2$) towards zero, when compared to its univariate-version (order $1 - \theta^6$ cf. (12a)). This difference between the univariate and the multivariate case had already been highlighted in Section 1.2 in Figure 2. Indeed substituting $d = 1$ in the bound (23a) gives us a vacuous bound for the univariate case, providing further evidence for the benefit of sharing variance among different dimensions in multivariate setting of symmetric fit (1). With Lemma 3 at hand, the proof of Theorem 2 follows by using the localization argument from the paper [Dwivedi et al., 2020]. Mimicking the arguments similar to equation (13b), we obtain the following statistical rate:[4]

$$\frac{\epsilon \cdot r/\sqrt{n}}{1 - \gamma(\epsilon)} = \epsilon \quad \Longrightarrow \frac{\epsilon r/\sqrt{n}}{\epsilon^2} = \epsilon \quad \Longrightarrow \quad \epsilon \sim n^{-\frac{1}{4}}. \tag{24}$$

Much of the work in the proof of Theorem 2 is to establish Lemma 3. With the bounds (23a) and (23b) at hand, using the localization argument (in a manner similar to the proof of Theorem 1), easily leads to the statistical rate of order $(d/n)^{1/4}$ as claimed in Theorem 2. The detailed proof is thereby omitted.

[4] Moreover, similar to the arguments made in the paper [Dwivedi et al., 2020], localization argument is necessary to derive a sharp rate. Indeed, a direct application of the framework introduced by Balakrishnan et al. [Balakrishnan et al., 2017] for our setting implies a sub-optimal rate of order $(d/n)^{1/6}$ for the Euclidean error $\|\theta_n^t - \theta^*\|$ (cf. (13a) and (13b)).

A.3 Proof of Lemma 1

We now prove Lemma 1 which provides the basis for the two-staged proof of Theorem 1.

The proof for the contraction property (12b) of the corrected population operator $\overline{M}_1$ is similar to that of the property (12a) pseudo-population operator $\widetilde{M}_{n,1}$ (albeit with a few high probability arguments replaced by deterministic arguments). Hence, while we provide a complete proof of the bound (12a) (in Section A.3.1), we only provide a proof sketch for the bound (12b) at its end. Moreover the proofs of bounds (12c) and (12d) are provided in Sections A.3.2 and A.3.3 respectively.

A.3.1 Contraction bound for population operator $\widetilde{M}_{n,1}$

We begin by defining some notation. For $\beta \in (0, 1/12]$ and $\alpha \geq 1/2 - 6\beta$, we define the event $\mathcal{E}_\alpha$ and the interval $I_{\alpha,\beta}$ as follows

$$\mathcal{E}_\alpha = \left\{ \left| \sum_{j=1}^{n} X_j^2/n - 1 \right| \leq n^{-\alpha} \right\}, \quad \text{and,} \tag{25}$$

$$I_{\alpha,\beta} = [3n^{-1/12+\beta}, \sqrt{9/400 - n^{-\alpha}}], \tag{26}$$

where in the above notations we have omitted the dependence on n, as it is clear from the context. We also use the scalars a and b to denote the following:

$$a := 1 - n^{-\alpha} \quad \text{and} \quad b := 1 + n^{-\alpha}.$$

With the above notation in place, observe that standard chi-squared tail bounds yield that $\mathbb{P}[\mathcal{E}_\alpha] \geq 1 - e^{-n^{1-2\alpha}/8} \geq 1 - \delta$. Moreover, invoking the lower bound on n in Theorem 1, we have that $[3n^{-1/12+\beta}, 1/10] \subseteq I_{\alpha,\beta}$. Now conditional on the high probability event $\mathcal{E}_\alpha$, the population EM update $\widetilde{M}_{n,1}(\theta)$, in absolute value, can be upper and lower bounded as follows:

$$\left|\widetilde{M}_{n,1}(\theta)\right| \leq \mathbb{E}_Y \left[Y \tanh\left(\frac{Y\,|\theta|}{a - \theta^2} \right) \right] = |\theta| \underbrace{\mathbb{E}_Y \left[\frac{Y}{|\theta|} \tanh\left(\frac{|\theta|\, X}{a - \theta^2} \right) \right]}_{=:\overline{\gamma}(\theta)}, \quad \text{and,}$$

$$\left|\widetilde{M}_{n,1}(\theta)\right| \geq \mathbb{E}_Y \left[Y \tanh\left(\frac{X\,|\theta|}{b - \theta^2} \right) \right] = |\theta| \underbrace{\mathbb{E}_Y \left[\frac{Y}{|\theta|} \tanh\left(\frac{|\theta|\, Y}{b - \theta^2} \right) \right]}_{=:\underline{\gamma}(\theta)},$$

where the last two inequalities follows directly from the definition of $\widetilde{M}_{n,1}(\theta)$ in equation (11a), and from the fact that for any fixed $y, \theta \in \mathbb{R}$, the function $w \mapsto y \tanh(y\,|\theta|\,/(w - \theta^2))$ is non-increasing in w for $w > \theta^2$. Consequently, in order to complete the proof, it suffices to establish the following bounds:

$$1 - 3\theta^6/2 \leq \underline{\gamma}(\theta), \quad \text{and} \quad \overline{\gamma}(\theta) \leq (1 - \theta^6/5). \tag{27}$$

The following properties of the hyperbolic function $x \mapsto x \tanh(x)$ are useful for our proofs:

Lemma 4. *For any $x \in \mathbb{R}$, the following holds*

$$\textit{(Lower bound):} \quad x \tanh(x) \geq x^2 - \frac{x^4}{3} + \frac{2x^6}{15} - \frac{17x^8}{315},$$

$$\textit{(Upper bound):} \quad x \tanh(x) \leq x^2 - \frac{x^4}{3} + \frac{2x^6}{15} - \frac{17x^8}{315} + \frac{62x^{10}}{2835}.$$

See Appendix C.4 for its proof.

Given the bounds in Lemma 4, we derive the upper and lower bounds in the inequality (27) separately.

Upper bound for $\overline{\gamma}(\theta)$: Invoking the upper bound on $x\tanh(x)$ from Lemma 4, we find that

$$\overline{\gamma}(\theta) \leq \frac{a-\theta^2}{\theta^2}\bigg(\frac{\theta^2}{(a-\theta^2)^2}\mathbb{E}\left[Y^2\right] - \frac{\theta^4}{3(a-\theta^2)^4}\mathbb{E}\left[Y^4\right] + \frac{2\theta^6}{15(a-\theta^2)^6}\mathbb{E}\left[Y^6\right] - \frac{17\theta^8}{315(a-\theta^2)^8}\mathbb{E}\left[Y^8\right] + \frac{62\theta^{10}}{2835(a-\theta^2)^{10}}\mathbb{E}\left[Y^{10}\right]\bigg).$$

Recall that, for $Y \sim \mathcal{N}(0,1)$, we have $\mathbb{E}\left[Y^{2k}\right] = (2k-1)!!$ for all $k \geq 1$. Therefore, the last inequality can be simplified to

$$\overline{\gamma}(\theta) \leq \frac{1}{a-\theta^2} - \frac{\theta^2}{(a-\theta^2)^3} + \frac{2\theta^4}{(a-\theta^2)^5} - \frac{17\theta^6}{3(a-\theta^2)^7} + \frac{62\theta^8}{3(a-\theta^2)^9}. \tag{28}$$

When $n^{-\alpha} + \theta^2 \leq 9/400$, we can verify that the following inequalities hold:

$$\begin{aligned}
\frac{1}{1-n^{-\alpha}-\theta^2} &\leq 1 + (n^{-\alpha}+\theta^2) + (n^{-\alpha}+\theta^2)^2 + (n^{-\alpha}+\theta^2)^3 + 2(n^{-\alpha}+\theta^2)^4,\\
-\frac{\theta^2}{(1-n^{-\alpha}-\theta^2)^3} &\leq -\theta^2\left(1 + 3(n^{-\alpha}+\theta^2) + 6(n^{-\alpha}+\theta^2)^2 + 10(n^{-\alpha}+\theta^2)^3\right),\\
\frac{\theta^4}{(1-n^{-\alpha}-\theta^2)^5} &\leq \theta^4\left(1 + 5(n^{-\alpha}+\theta^2) + 16(n^{-\alpha}+\theta^2)^2\right),\\
-\frac{\theta^6}{(1-n^{-\alpha}-\theta^2)^7} &\leq -\theta^6\left(1 + 7(n^{-\alpha}+\theta^2)\right),\\
\frac{\theta^8}{(1-n^{-\alpha}-\theta^2)^9} &\leq 5\theta^8/4.
\end{aligned}$$

Substituting $a = 1 - n^{-\alpha}$ into the bound (28) and doing some algebra with the above inequalities and using the fact that $\max\{\theta, n^{-\alpha}\} \leq 1$ we have that

$$\overline{\gamma}(\theta) \leq 1 - \frac{2}{3}\theta^6 + \frac{61}{6}\theta^8 + 100n^{-\alpha} \leq 1 - \frac{2}{5}\theta^6 + 100n^{-\alpha} \leq 1 - \frac{1}{5}\theta^6.$$

The second last inequality above follows since $\theta \leq 3/20$, and the last inequality above utilizes the fact that if $\alpha \geq 1/2 - 6\beta$, then $\theta^6/5 \geq 100n^{-\alpha}$ for all $\theta \geq 3n^{-1/12+\beta}$. This completes the proof of the upper bound of $\overline{\gamma}(\theta)$.

Lower bound for $\underline{\gamma}(\theta)$: We start by utilizing the lower bound of $x\tanh(x)$ in the expression for $\underline{\gamma}(\theta)$, which yields:

$$\underline{\gamma}(\theta) \geq \frac{1}{b-\theta^2} - \frac{\theta^2}{(b-\theta^2)^3} + \frac{2\theta^4}{(b-\theta^2)^5} - \frac{17\theta^6}{3(b-\theta^2)^7}. \tag{29}$$

Since $|\theta| \in [3n^{-1/12+\beta}, \sqrt{9/400 - n^{-\alpha}}]$ by assumption, we have the following lower bounds:

$$\begin{aligned}
\frac{1}{1+n^{-\alpha}-\theta^2} &\geq 1 + (\theta^2 - n^{-\alpha}) + (\theta^2 - n^{-\alpha})^2 + (\theta^2 - n^{-\alpha})^3 + (\theta^2 - n^{-\alpha})^4,\\
-\frac{\theta^2}{(1+n^{-\alpha}\theta^2)^3} &\geq -\theta^2 - \left(1 + 3(\theta^2 - n^{-\alpha}) + 6(\theta^2 - n^{--\alpha})^2 + 11(\theta^2 - n^{-\alpha})^3\right),\\
\frac{\theta^4}{(1+n^{-\alpha}-\theta^2)^5} &\geq \theta^4\left(1 + 5(\theta^2 - n^{-\alpha}) + 15(\theta^2 - n^{-\alpha}\right),\\
-\frac{\theta^6}{(1+n^{-\alpha}-\theta^2)^7} &\geq -\theta^6\left(1 + 8(\theta^2 - n^{-\alpha})\right).
\end{aligned}$$

Substituting $b = 1 + n^{-\alpha}$ into the bound (29) and doing some algebra with the above inequalities and using the fact that $\max\{\theta, n^{-\alpha}\} \leq 1$ we have that

$$\underline{\gamma}(\theta) \geq 1 - \frac{2}{3}\theta^6 - \frac{76}{3}\theta^8 - 100n^{-\alpha} \geq 1 - \frac{5}{4}\theta^6 - 100n^{-\alpha} \geq 1 - \frac{3}{2}\theta^6,$$

The second last inequality above follows since $\theta \leq 3/20$, and the last inequality above utilizes the fact that if $\alpha \geq 1/2 - 6\beta$, then $\theta^6/4 \geq 100n^{-\alpha}$ for all $\theta \geq 3n^{-1/12+\beta}$. This completes the proof of the lower bound of $\underline{\gamma}(\theta)$.

Proof of contraction bound for $\overline{M}_1$: Note that it suffices to repeat the arguments with $a = 1$ and $b = 1$ in the RHS of the inequalities (28) and (29) respectively. Given the other computations, the remaining steps are straightforward algebra and are thereby omitted.

A.3.2 Proof of perturbation bound for $\widetilde{M}_{n,1}$

We now prove the bound (12c) which is based on standard arguments to derive Rademacher complexity bounds. We first symmetrize with Rademacher variables, and apply the Ledoux-Talagrand contraction inequality. We then invoke results on sub-Gaussian and sub-exponential random variables, and finally perform the associated Chernoff-bound computations to obtain the desired result.

To ease the presentation, we denote $\alpha := 1/2 - 2\beta$ and $\mathcal{I} := [1 - n^{-\alpha} - 1/64, 1 - n^{-\alpha}]$. Next we fix $r \in [0, 1/8]$ and define $\widetilde{r} := \frac{r}{1-n^{-\alpha}-1/64}$. For sufficiently large n, we have $\widetilde{r} \leq 2r$. Recall the definition (25) of the event: $\mathcal{E}_\alpha = \{|\sum_{j=1}^n X_j^2/n - 1| \leq n^{-\alpha}\}$. Conditional on the event $\mathcal{E}_\alpha$, the following inequalities hold

$$\begin{aligned}\left|M_{n,1}(\theta) - \widetilde{M}_{n,1}(\theta)\right| &\leq \sup_{\theta \in \mathbb{B}(0,r), \sigma^2 \in \mathcal{I}} \left| \frac{1}{n}\sum_{i=1}^n X_i \tanh\left(\frac{X_i\theta}{\sigma^2}\right) - \mathbb{E}\left[Y \tanh\left(\frac{Y\theta}{\sigma^2}\right)\right]\right| \\ &\leq \sup_{\widetilde{\theta} \in \mathbb{B}(0,\widetilde{r})} \left|\widehat{M}_n(\widetilde{\theta}) - \widehat{M}(\widetilde{\theta})\right|,\end{aligned}$$

with all them valid for any $\theta \in \mathbb{B}(0, r)$. Here Y denotes a standard normal variate $\mathcal{N}(0, 1)$ whereas the operators $\widehat{M}$ and $\widehat{M}_n$ are defined as

$$\widehat{M}(\widetilde{\theta}) := \mathbb{E}[Y \tanh(Y\widetilde{\theta})] \quad \text{and} \quad \widehat{M}_n(\widetilde{\theta}) := \frac{1}{n}\sum_{i=1}^n X_i \tanh(X_i\widetilde{\theta}).$$

To facilitate the discussion later, we define the unconditional random variable

$$Z := \sup_{\widetilde{\theta} \in \mathbb{B}(0,\widetilde{r})} \left|\widehat{M}_n(\widetilde{\theta}) - \widehat{M}(\widetilde{\theta})\right|.$$

Employing standard symmetrization argument from empirical process theory [van der Vaart and Wellner, 2000], we find that

$$\mathbb{E}[\exp(\lambda Z)] \leq \mathbb{E}\left[\exp\left(\sup_{\widetilde{\theta} \in \mathbb{B}(0,\widetilde{r})} \frac{2\lambda}{n}\sum_{i=1}^n \varepsilon_i \tanh(X_i\widetilde{\theta})X_i\right)\right],$$

where $\varepsilon_i, i \in [n]$ are i.i.d. Rademacher random variables independent of $\{X_i, i \in [n]\}$. Noting that, the following inequality with hyperbolic function $\tanh(x)$ holds

$$\left|\tanh(x\widetilde{\theta}) - \tanh(x\widetilde{\theta}')\right| \leq \left|(\widetilde{\theta} - \widetilde{\theta}')x\right| \qquad \text{for all } x.$$

Consequently for any given x, the function $\widetilde{\theta} \mapsto \tanh(x\widetilde{\theta})$ is Lipschitz. Invoking the Ledoux-Talagrand contraction result for Lipschitz functions of Rademacher processes [Ledoux and Talagrand, 1991] and following the proof argument from Lemma 1 in the paper [Dwivedi et al., 2020], we obtain that

$$Z \leq c\widetilde{r}\sqrt{\frac{\log(1/\delta)}{n}}, \qquad \text{with probability } \geq 1 - \delta,$$

for some universal constant c. Finally, using $\widetilde{r} \leq 2r$ for large n, we obtain that

$$\left|M_{n,1}(\theta) - \widetilde{M}_{n,1}(\theta)\right| \leq 2cr\sqrt{\frac{\log(1/\delta)}{n}}, \qquad \text{with probability } \geq 1 - \delta - e^{-n^{1-2\alpha}/8},$$

where we have also used the fact that $\mathbb{P}[\mathcal{E}_\alpha] \geq 1 - e^{-n^{1-2\alpha}/8}$ from standard chi-squared tail bounds. The bound (12c) follows and we are done.

A.3.3 Proof of perturbation bound for $\overline{M}_1$

We now prove the bound (12d). Note that it suffices to establish the following point-wise result:

$$\big|\overline{M}_1(\theta) - M_{n,1}(\theta)\big| \precsim \frac{|\theta|^3 \log^{10}(5n/\delta)}{\sqrt{n}} \quad \text{for all} \quad |\theta| \precsim n^{-1/16},$$

with probability at least $1-\delta$ for any given $\delta > 0$. For the reader's convenience, let us recall the definition of these operators

$$\overline{M}_1(\theta) = \mathbb{E}\Big[X \tanh(X\theta/(1-\theta^2))\Big], \tag{30a}$$

$$M_{n,1}(\theta) = \frac{1}{n}\sum_{i=1}^{n} X_i \tanh\Big(X_i\theta/(a_n - \theta^2)\Big), \tag{30b}$$

where $a_n := \sum_{i=1}^n X_i^2/n$. We further denote $\mu_k := \mathbb{E}_{X\sim\mathcal{N}(0,1)}[X^k]$, and $\widehat{\mu}_k := \frac{1}{n}\sum_{i=1}^n X_i^k$. From known results on Gaussian moments, we have $\mu_{2k} = (2k-1)!!$ for each integer $k = 1, 2, \ldots$.

For any given x and scalar b, consider the map $\theta \mapsto x\tanh(x\theta/(b-\theta^2))$. The 9-th order Taylor series for this function around $\theta = 0$ is given by

$$\begin{aligned} x\tanh(x\theta/(b-\theta^2)) = \frac{\theta x^2}{b} - \frac{\theta^3(x^4 - 3bx^2)}{3b^3} + \theta^5\left(\frac{2x^6}{15b^5} - \frac{x^4}{b^4} + \frac{x^2}{b^3}\right) \\ + \theta^7\left(-\frac{17x^8}{315b^7} + \frac{2x^6}{3b^6} - \frac{2x^4}{b^5} + \frac{x^2}{b^4}\right) \\ + \theta^9\left(\frac{62x^{10}}{2835b^9} - \frac{17x^8}{45b^8} + \frac{2x^6}{b^7} - \frac{10x^4}{3b^6} + \frac{x^2}{b^5}\right) + \varepsilon, \end{aligned} \tag{31}$$

where the remainder ε satisfies $\varepsilon \leq \mathcal{O}\left(\theta^{11}\right)$. Plugging in this expansion with $b = 1$ on RHS of equation (30a) and taking expectation over $X \sim \mathcal{N}(0,1)$, we obtain

$$\overline{M}_1(\theta) = \theta + \theta^3\Big(\sum_{k=1}^{2} c_{3,k}\mu_{2k}\Big) + \theta^5\Big(\sum_{k=1}^{3} c_{5,k}\mu_{2k}\Big) + \theta^7\Big(\sum_{k=1}^{4} c_{7,k}\mu_{2k}\Big) + \theta^9\Big(\sum_{k=1}^{5} c_{9,k}\mu_{2k}\Big) + \varepsilon, \tag{32a}$$

where we have used the notation $\mu_k := \mathbb{E}_{X\sim\mathcal{N}(0,1)}[X^k]$ and $c_{j,k}$ denote universal constants. Furthermore, plugging in the same expansion (31) with $b = a_n$ on RHS of equation (30b), we obtain the following expansion for the sample EM operator

$$M_{n,1}(\theta) = \theta + \theta^3\Big(\sum_{k=1}^{2} c_{3,k}\frac{\widehat{\mu}_{2k}}{a_n^{1+k}}\Big) + \theta^5\Big(\sum_{k=1}^{3} c_{5,k}\frac{\widehat{\mu}_{2k}}{a_n^{2+k}}\Big) + \theta^7\Big(\sum_{k=1}^{4} c_{7,k}\frac{\widehat{\mu}_{2k}}{a_n^{3+k}}\Big) + \theta^9\Big(\sum_{k=1}^{5} c_{9,k}\frac{\widehat{\mu}_{2k}}{a_n^{4+k}}\Big) + \varepsilon_n, \tag{32b}$$

where $\widehat{\mu}_k$ denotes the sample mean of X^k, i.e., $\widehat{\mu}_k := \frac{1}{n}\sum_{i=1}^n X_i^k$. In order to lighten the notation, we introduce the following convenient shorthand:

$$\beta_j = \sum_{k=1}^{\frac{j+1}{2}} c_{j,k}\mu_{2k} \quad \text{and} \quad \widehat{\beta}_j = \sum_{k=1}^{\frac{j+1}{2}} c_{j,k}\frac{\widehat{\mu}_{2k}}{a_n^{\frac{j-1}{2}+k}} \quad \text{for } j \in \{3,5,7,9\} =: \mathcal{J}. \tag{33}$$

A careful inspection reveals that $\beta_3 = \beta_5 = 0$. With the above notations in place, we find that

$$\begin{aligned} \big|\overline{M}_1(\theta) - M_{n,1}(\theta)\big| &= \Big|\sum_{j\in\mathcal{J}} \theta^j(\beta_j - \widehat{\beta}_j)\Big| + \varepsilon \\ &=: U_1 + U_2. \end{aligned}$$

Therefore, it remains to establish that

$$U_1 \precsim \frac{|\theta|^3\log^5(5n/\delta)}{\sqrt{n}} \quad \text{and} \quad U_2 \precsim \frac{|\theta|^3\log^5(5n/\delta)}{\sqrt{n}}, \tag{34}$$

with probability at least $1-\delta$ for any given $\delta > 0$. Since the remainder term is of order θ^{11}, the assumption $|\theta| \precsim n^{-1/16}$ ensures that the remainder term is bounded by a term of order $\theta^3/\sqrt{n}$ and thus the bound (34) on the second term U_2 follows.

We now use concentration properties of Gaussian moments in order to prove the bound (34) on the first term U_1. Since $|\theta| \leq 1$, it suffices to show that

$$\sup_{j \in \mathcal{J}} \left| \beta_j - \widehat{\beta}_j \right| \precsim \frac{\log^5(5n/\delta)}{\sqrt{n}} \tag{35}$$

with probability at least $1-\delta$. Using the relation (33), we find that

$$\begin{aligned} \left| \beta_j - \widehat{\beta}_j \right| = \Big| \sum_{k=1}^{\frac{j+1}{2}} \left(c_{j,k}\mu_{2k} - c_{j,k} \frac{\widehat{\mu}_{2k}}{a_n^{\frac{j-1}{2}+k}} \right) \Big| &\leq \sum_{k=1}^{\frac{j+1}{2}} \frac{c_{j,k}}{a_n^{\frac{j-1}{2}+k}} \left| \mu_{2k} - \widehat{\mu}_{2k} \right| + c_{j,k}(1 - a_n^{-\frac{j-1}{2}-k})\mu_{2k} \\ &\leq C \sum_{k=1}^{\frac{j+1}{2}} \left(\left| \mu_{2k} - \widehat{\mu}_{2k} \right| + \frac{\mu_{2k}}{\sqrt{n}} \right), \end{aligned} \tag{36}$$

for any $j \in \mathcal{J}$. Here in the last step we have used the following bounds:

$$\max_{j \in \mathcal{J}, k \leq \frac{j+1}{2}} c_{j,k} \leq C \quad \text{and} \quad \max_{j \in \mathcal{J}, k \leq \frac{j+1}{2}} (1 - a_n^{-\frac{j-1}{2}-k}) \leq \frac{C}{\sqrt{n}}$$

for some universal constant C. Thus a lemma for the $1/\sqrt{n}$-concentration[5] of higher moments of Gaussian random variable is now useful:

Lemma 5. *Let $X_1, \ldots, X_n$ are i.i.d. samples from $\mathcal{N}(0,1)$ and let $\mu_{2k} := \mathbb{E}_{X \sim \mathcal{N}(0,1)}[X^{2k}]$ and $\widehat{\mu}_{2k} := \frac{1}{n}\sum_{i=1}^n X_i^{2k}$. Then, we have*

$$\mathbb{P}\left(|\widehat{\mu}_{2k} - \mu_{2k}| \leq \frac{C_k \log^k(n/\delta)}{\sqrt{n}} \right) \geq 1 - \delta \quad \textit{for any} \quad k \geq 1,$$

where C_k denotes a universal constant depending only on k.

See the Appendix C.5 for the proof.

For any $\delta > 0$, consider the event

$$\mathcal{E} := \left\{ \left| \mu_{2k} - \widehat{\mu}_{2k} \right| \leq \frac{C_k \log^k(5n/\delta)}{\sqrt{n}} \quad \text{for all } k \in \{2, 4, \ldots, 10\} \right\}. \tag{37}$$

Straightforward application of union bound with Lemma 5 yields that $\mathbb{P}[\mathcal{E}] \geq 1 - \delta$. conditional on the event $\mathcal{E}$ inequality (35) implies that

$$\begin{aligned} \sup_{j \in \mathcal{J}} \left| \beta_j - \widehat{\beta}_j \right| &\leq C \sup_{j \in \mathcal{J}} \sum_{k=1}^{\frac{j+1}{2}} \left(\left| \mu_{2k} - \widehat{\mu}_{2k} \right| + \frac{\mu_{2k}}{\sqrt{n}} \right) \\ &\leq C \sup_{j \in \{3,5,7,9\}} \frac{j+1}{2} \left(\left| \mu_{j+1} - \widehat{\mu}_{j+1} \right| + \frac{(j+1)!!}{\sqrt{n}} \right) \\ &\overset{(i)}{\leq} C \sup_{j \in \{3,5,7,9\}} (j-1) \left(\left| C_{\frac{j+1}{2}} \frac{\log^{\frac{j+1}{2}}(5n/\delta)}{\sqrt{n}} \right| + \frac{(j+1)!!}{\sqrt{n}} \right) \\ &\overset{(ii)}{\leq} C \frac{\log^5(5n/\delta)}{\sqrt{n}}, \end{aligned} \tag{38}$$

where step (i) follows from the definition of the event (37) and in step (ii) using the fact that $j \leq 9$ is bounded we absorbed all the constants into a single constant. Since the event $\mathcal{E}$ has probability at least $1-\delta$, the claim (35) now follows.

[5]The bound from Lemma 5 is sub-optimal for $k = 1$ but is sharper than the standard tail bounds for Gaussian polynomials of degree $2k$ for $k \geq 2$. The $1/\sqrt{n}$ concentration of higher moments is necessary to derive the sharp rates stated in our results.

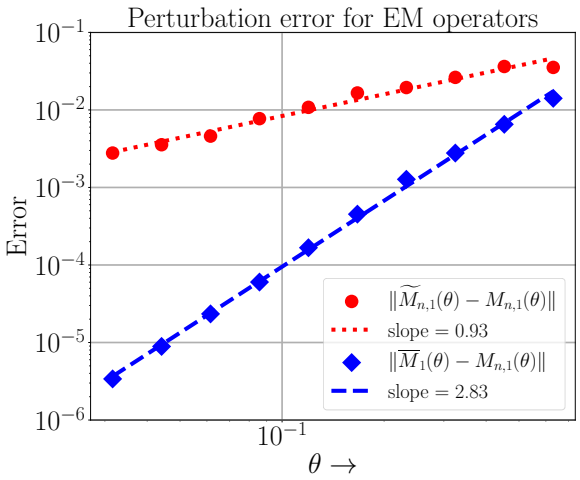

Figure 4. Plots of the perturbation errors for the pseudo-population operator $\widetilde{M}_{n,1}$ (11a) and the corrected population operator $\overline{M}_1$ (11b) with respect to the sample EM operator $M_{n,1}$ (4), as a function of θ. From the least-squares fit on the log-log scale, we see that the error $\|\widetilde{M}_{n,1}(\theta) - M_{n,1}(\theta)\|$ scales linearly with θ, the error $\|\overline{M}_1(\theta) - M_{n,1}(\theta)\|$ has a cubic dependence on θ, in accordance with Lemma 1.

A.3.4 Sharpness of bounds of Lemma 1

In Figure 4, we numerically verify the linear and cubic scaling of the bounds stated in Lemma 1.

B Minimax bound

We now show that the error of order $n^{-\frac{1}{8}}$ (up to logarithmic factors) is, in fact, tight in the standard minimax sense. Given a compact set $\Omega \subset \mathbb{R} \times (0, \infty)$, and a set of true parameters $(\theta^*, \sigma^*) \in \Omega$, suppose that we draw n i.i.d. samples $\{X_i\}_{i=1}^n$ from a two-Gaussian mixture of the form $\frac{1}{2}\mathcal{N}(\theta^*, (\sigma^*)^2) + \frac{1}{2}\mathcal{N}(-\theta^*, (\sigma^*)^2)$. Let $(\widehat{\theta}_n, \widehat{\sigma}_n) \in \Omega$ denote any estimates—for the respective parameters—measurable with respect to the observed samples $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f_{\theta^*,\sigma^*}$ and let $\mathbb{E}_{(\theta^*,\sigma^*)}$ denote the corresponding expectation.

Proposition 1. *There exists a universal constant $c_\Omega > 0$ (depending only on Ω), such that*

$$\inf_{(\widehat{\theta}_n, \widehat{\sigma}_n)} \sup_{(\theta^*,\sigma^*)} \mathbb{E}_{(\theta^*,\sigma^*)} \left[\left(|\widehat{\theta}_n| - |\theta^*|\right)^2 + \left|(\widehat{\sigma}_n)^2 - (\sigma^*)^2\right| \right] \geq c_\Omega n^{-\frac{1}{4}-\delta} \qquad \text{for any } \delta > 0.$$

See Appendix C.1 for the proof.

Based on the connection between location parameter θ_n^t and scale parameter σ_n^t in the EM updates (cf. Equation (3c)), the minimax lower bound in Proposition 1 shows that the (non-squared) error of EM location updates $||\theta_n^t| - |\theta^*||$ is lower bounded by a term (arbitrarily close to) $n^{-\frac{1}{8}}$.

B.1 Proof of Proposition 1

We now present the proof of the minimax bound. We introduce the shorthand $v := \sigma^2$ and $\eta := (\theta, v)$. First of all, we claim the following key upper bound of Hellinger distance between mixture densities f_{η_1}, f_{η_2} in terms of the distances among their corresponding parameters η_1 and η_2:

$$\inf_{\eta_1,\eta_2 \in \Omega} \frac{h\left(f_{\eta_1}, f_{\eta_2}\right)}{\left(\left(|\theta_1| - |\theta_2|\right)^2 + |v_1 - v_2|\right)^r} = 0 \qquad \text{for any } r \in (1, 4). \tag{39}$$

Moreover, for any two densities p and q, we denote the total variation distance between p and q by $V(p, q) := (1/2) \int |p(x) - q(x)|\, dx$. Similarly, the squared Hellinger distance between p and q is given as $h^2(p, q) = (1/2) \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx$.

Taking the claim (39) as given for the moment, let us complete the proof of Proposition 1. Our proof relies on Le Cam's lemma for establishing minimax lower bounds. In particular, for any $r \in (1,4)$ and for any $\epsilon > 0$ sufficiently small, according to the result in equation (39), there exist $\eta_1 = (\theta_1, v_1)$ and $\eta_2 = (\theta_2, v_2)$ such that $(|\theta_1| - |\theta_2|)^2 + |v_1 - v_2| = 2\epsilon$ and $h\left(f_{\eta_1}, f_{\eta_2}\right) \leq c\epsilon^r$ for some universal constant c. From Lemma 1 from Yu [Yu, 1997], we obtain that

$$\sup_{\eta \in \{\eta_1, \eta_2\}} \mathbb{E}_\eta \left[\left(\left| \widehat{\theta}_n \right| - |\theta| \right)^2 + \left| (\widehat{\sigma}_n)^2 - (\sigma)^2 \right| \right] \gtrsim \epsilon \left(1 - V(f^n_{\eta_1}, f^n_{\eta_2})\right),$$

where f^n_η denotes the product of mixture densities f_η of the data $X_1, \ldots, X_n$. A standard relation between total variation distance and Hellinger distance leads to

$$V(f^n_{\eta_1}, f^n_{\eta_2}) \leq h(f^n_{\eta_1}, f^n_{\eta_2}) = \sqrt{1 - [1 - h^2(f_{\eta_1}, f_{\eta_2})]^n} \leq \sqrt{1 - [1 - c\epsilon^r]^n}.$$

By choosing $c\epsilon^r = 1/n$, we can verify that

$$\sup_{\eta \in \{\eta_1, \eta_2\}} \mathbb{E}_\eta \left[\left(\left| \widehat{\theta}_n \right| - |\theta| \right)^2 + \left| (\widehat{\sigma}_n)^2 - (\sigma)^2 \right| \right] \gtrsim \epsilon \asymp n^{-1/r},$$

which establishes the claim of Proposition 1.

B.1.1 Proof of claim (39)

In order to prove claim (39), it is sufficient to construct sequences $\eta_{1,n} = (\theta_{1,n}, v_{1,n})$ and $\eta_{2,n} = (\theta_{2,n}, v_{2,n})$ such that

$$h\left(f_{\eta_{1,n}}, f_{\eta_{2,n}}\right) / \left(\left(|\theta_{1,n}| - |\theta_{2,n}|\right)^2 + |v_{1,n} - v_{2,n}| \right)^r \to 0$$

as $n \to \infty$. Indeed, we construct these sequences as follows: $\theta_{2,n} = 2\theta_{1,n}$ and $v_{1,n} - v_{2,n} = 3\left(\theta_{1,n}\right)^2$ for all $n \geq 1$ while $\theta_{1,n} \to 0$ as $n \to \infty$. Direct computation leads to

$$f_{\eta_{1,n}}(x) - f_{\eta_{2,n}}(x) = \frac{1}{2} \underbrace{\left(\phi(x; -\theta_{1,n}, v_{1,n}) - \phi(x; -\theta_{2,n}, v_{2,n})\right)}_{T_{1,n}} + \frac{1}{2} \underbrace{\left(\phi(x; \theta_{1,n}, v_{1,n}) - \phi(x; \theta_{2,n}, v_{2,n})\right)}_{T_{2,n}}.$$

Invoking Taylor expansion up to the third order, we obtain that

$$\begin{aligned}
T_{1,n} &= \sum_{|\alpha| \leq 3} \frac{(\theta_{2,n} - \theta_{1,n})^{\alpha_1} (v_{1,n} - v_{2,n})^{\alpha_2}}{\alpha_1! \alpha_2!} \frac{\partial^{|\alpha|} \phi}{\partial \theta^{\alpha_1} \partial v^{\alpha_2}}(x; -\theta_{2,n}, v_{2,n}) + R_1(x), \\
T_{2,n} &= \sum_{|\alpha| \leq 3} \frac{(\theta_{1,n} - \theta_{2,n})^{\alpha_1} (v_{1,n} - v_{2,n})^{\alpha_2}}{\alpha_1! \alpha_2!} \frac{\partial^{|\alpha|} \phi}{\partial \theta^{\alpha_1} \partial v^{\alpha_2}}(x; \theta_{2,n}, v_{2,n}) + R_2(x)
\end{aligned}$$

where $|\alpha| = \alpha_1 + \alpha_2$ for $\alpha = (\alpha_1, \alpha_2)$. Here, $R_1(x)$ and $R_2(x)$ are Taylor remainders that have the following explicit representations

$$\begin{aligned}
R_1(x) &:= 4 \sum_{|\beta| = 4} \frac{(\theta_{2,n} - \theta_{1,n})^{\beta_1} (v_{1,n} - v_{2,n})^{\beta_2}}{\beta_1! \beta_2!} \\
&\quad \times \int_0^1 (1-t)^3 \frac{\partial^4 \phi}{\partial \theta^{\beta_1} \partial v^{\beta_2}} \left(x; -\theta_{2,n} + t(\theta_{2,n} - \theta_{1,n}), v_{2,n} + t(v_{1,n} - v_{2,n})\right) dt, \\
R_2(x) &:= 4 \sum_{|\beta| = 4} \frac{(\theta_{1,n} - \theta_{2,n})^{\beta_1} (v_{1,n} - v_{2,n})^{\beta_2}}{\beta_1! \beta_2!} \\
&\quad \times \int_0^1 (1-t)^3 \frac{\partial^4 \phi}{\partial \theta^{\beta_1} \partial v^{\beta_2}} \left(x; \theta_{2,n} + t(\theta_{1,n} - \theta_{2,n}), v_{2,n} + t(v_{1,n} - v_{2,n})\right) dt.
\end{aligned}$$

Recall from equation (2) that the univariate location-scale Gaussian distribution has the PDE structure of the following form

$$\frac{\partial^2 \phi}{\partial \theta^2}(x;\theta,\sigma^2) = 2\frac{\partial \phi}{\partial \sigma^2}(x;\theta,\sigma^2).$$

Therefore, we can write the formulations of $T_{1,n}$ and $T_{2,n}$ as follows:

$$T_{1,n} = \sum_{|\alpha|\leq 3} \frac{(\theta_{2,n}-\theta_{1,n})^{\alpha_1}(v_{1,n}-v_{2,n})^{\alpha_2}}{2^{\alpha_2}\alpha_1!\alpha_2!}\frac{\partial^{\alpha_1+2\alpha_2}\phi}{\partial\theta^{\alpha_1+2\alpha_2}}(x;-\theta_{2,n},v_{2,n}) + R_1(x),$$
$$T_{2,n} = \sum_{|\alpha|\leq 3} \frac{(\theta_{1,n}-\theta_{2,n})^{\alpha_1}(v_{1,n}-v_{2,n})^{\alpha_2}}{2^{\alpha_2}\alpha_1!\alpha_2!}\frac{\partial^{\alpha_1+2\alpha_2}\phi}{\partial\theta^{\alpha_1+2\alpha_2}}(x;\theta_{2,n},v_{2,n}) + R_2(x).$$

Via a Taylor series expansion, we find that

$$\frac{\partial^{\alpha_1+2\alpha_2}\phi}{\partial\theta^{\alpha_1+2\alpha_2}}(x;\theta_{2,n},v_{2,n}) = \sum_{\tau=0}^{3-|\alpha|}\frac{(2\theta_{2,n})^{\tau}}{\tau!}\frac{\partial^{\alpha_1+2\alpha_2+\tau}\phi}{\partial\theta^{\alpha_1+2\alpha_2+\tau}}(x;-\theta_{2,n},v_{2,n}) + R_{2,\alpha}(x)$$

for any $\alpha=(\alpha_1,\alpha_2)$ such that $1\leq|\alpha|\leq 3$. Here, $R_{2,\alpha}$ is Taylor remainder admitting the following representation

$$R_{2,\alpha}(x) = \sum_{\tau=4-|\alpha|}\frac{\tau\,(2\theta_{2,n})^{\tau}}{\tau!}\int_0^1(1-t)^{\tau-1}\frac{\partial^4\phi}{\partial\theta^{\alpha_1+\tau}\partial v^{\alpha_2}}\left(x;-\theta_{2,n}+2t\theta_{2,n},v_{2,n}\right)dt.$$

Governed by the above results, we can rewrite $f_{\eta_{1,n}}(x) - f_{\eta_{2,n}}(x)$ as

$$f_{\eta_{1,n}}(x) - f_{\eta_{2,n}}(x) = \sum_{l=1}^{6} A_{l,n}\frac{\partial^l\phi}{\partial\theta^l}(x;-\theta_{2,n},v_{2,n}) + R(x)$$

where the explicit formulations of $A_{l,n}$ and $R(x)$ are given by

$$A_{l,n} := \frac{1}{2}\sum_{\alpha_1,\alpha_2}\frac{1}{2^{\alpha_2}}\frac{(\theta_{2,n}-\theta_{1,n})^{\alpha_1}(v_{1,n}-v_{2,n})^{\alpha_2}}{\alpha_1!\alpha_2!}$$
$$+\frac{1}{2}\sum_{\alpha_1,\alpha_2,\tau}\frac{1}{2^{\alpha_2}}\frac{2^{\tau}(\theta_{2,n})^{\tau}(\theta_{1,n}-\theta_{2,n})^{\alpha_1}(v_{1,n}-v_{2,n})^{\alpha_2}}{\tau!\alpha_1!\alpha_2!},$$
$$R(x) := \frac{1}{2}R_1(x) + \frac{1}{2}R_2(x) + \sum_{|\alpha|\leq 2}\frac{1}{2^{\alpha_2}}\frac{(\theta_{1,n}-\theta_{2,n})^{\alpha_1}(v_{1,n}-v_{2,n})^{\alpha_2}}{\alpha_1!\alpha_2!}R_{2,\alpha}(x)$$

for any $l\in[6]$ and $x\in\mathbb{R}$. Here the ranges of α_1,α_2 in the first sum of $A_{l,n}$ satisfy $\alpha_1+2\alpha_2=l$ and $1\leq|\alpha|\leq 3$ while the ranges of α_1,α_2,τ in the second sum of $A_{l,n}$ satisfy $\alpha_1+2\alpha_2+\tau=l$, $0\leq\tau\leq 3-|\alpha|$, and $1\leq|\alpha|\leq 3$.

From the conditions that $\theta_{2,n}=2\theta_{1,n}$ and $v_{1,n}-v_{2,n}=3\left(\theta_{1,n}\right)^2$, we can check that $A_{l,n}=0$ for all $1\leq l\leq 3$. Additionally, we also have

$$\max\{|A_{4,n}|,|A_{5,n}|,|A_{6,n}|\}\precsim|\theta_{1,n}|^4.$$

Given the above results, we claim that

$$h\left(f_{\eta_{1,n}},f_{\eta_{2,n}}\right)\precsim|\theta_{1,n}|^8. \tag{40}$$

Assume that the claim (40) is given. From the formulations of sequences $\eta_{1,n}$ and $\eta_{2,n}$, we can verify that

$$\left(\left(|\theta_{1,n}|-|\theta_{2,n}|\right)^2+|v_{1,n}-v_{2,n}|\right)^r\asymp|\theta_{1,n}|^{2r}.$$

Since $1\leq r<4$ and $\theta_{1,n}\to 0$ as $n\to\infty$, the above results lead to

$$h\left(f_{\eta_{1,n}},f_{\eta_{2,n}}\right)/\left(\left(|\theta_{1,n}|-|\theta_{2,n}|\right)^2+|v_{1,n}-v_{2,n}|\right)^r\precsim|\theta_{1,n}|^{8-2r}\to 0.$$

As a consequence, we achieve the conclusion of the claim (39).

B.1.2 Proof of claim (40)

The definition of Hellinger distance leads to the following equations

$$\begin{aligned}
2h^2\left(f_{\eta_{1,n}}, f_{\eta_{2,n}}\right) &= \int \frac{\left(f_{\eta_{1,n}}(x) - f_{\eta_{2,n}}(x)\right)^2}{\left(\sqrt{f_{\eta_{1,n}}(x)} + \sqrt{f_{\eta_{2,n}}(x)}\right)^2} dx \\
&= \int \frac{(\sum_{l=4}^{6} A_{l,n} \frac{\partial^l \phi}{\partial \theta^l}(x; -\theta_{2,n}, v_{2,n}) + R(x))^2}{\left(\sqrt{f_{\eta_{1,n}}(x)} + \sqrt{f_{\eta_{2,n}}(x)}\right)^2} dx \\
&\precsim \int \frac{\sum_{l=4}^{6} (A_{l,n})^2 \left(\frac{\partial^l \phi}{\partial \theta^l}(x; -\theta_{2,n}, v_{2,n})\right)^2 + R^2(x)}{\left(\sqrt{f_{\eta_{1,n}}(x)} + \sqrt{f_{\eta_{2,n}}(x)}\right)^2} dx, \qquad (41)
\end{aligned}$$

where the last inequality is due to Cauchy-Schwarz's inequality. According to the structure of location-scale Gaussian density, the following inequalities hold

$$\int \frac{\left(\frac{\partial^l \phi}{\partial \theta^l}(x; -\theta_{2,n}, v_{2,n})\right)^2}{\left(\sqrt{f_{\eta_{1,n}}(x)} + \sqrt{f_{\eta_{2,n}}(x)}\right)^2} dx \precsim \int \frac{\left(\frac{\partial^l \phi}{\partial \theta^l}(x; -\theta_{2,n}, v_{2,n})\right)^2}{\phi(x; -\theta_{2,n}, v_{2,n})} dx < \infty \qquad (42)$$

for $4 \leq l \leq 6$. Note that, for any $\beta = (\beta_1, \beta_2)$ such that $|\beta| = 4$, we have

$$|\theta_{2,n} - \theta_{1,n}|^{\beta_1} |v_{1,n} - v_{2,n}|^{\beta_2} \asymp |\theta_{1,n}|^{4+\beta_2} \precsim |\theta_{1,n}|^4 .$$

With the above bounds, an application of Cauchy-Schwarz's inequality leads to

$$\begin{aligned}
&\int \frac{R_1^2(x)}{\left(\sqrt{f_{\eta_{1,n}}(x)} + \sqrt{f_{\eta_{2,n}}(x)}\right)^2} dx \\
&\precsim |\theta_{1,n}|^8 \sum_{|\beta|=4} \int \frac{\sup_{t\in[0,1]} \left(\frac{\partial^4 \phi}{\partial \theta^{\beta_1} \partial v^{\beta_2}}(x; -\theta_{2,n} + t(\theta_{2,n} - \theta_{1,n}), v_{2,n} + t(v_{1,n} - v_{2,n}))\right)^2}{\phi(x; -\theta_{2,n}, v_{2,n})} dx \precsim |\theta_{1,n}|^8 .
\end{aligned}$$

With a similar argument, we also obtain that

$$\int \frac{R_2^2(x)}{\left(\sqrt{f_{\eta_{1,n}}(x)} + \sqrt{f_{\eta_{2,n}}(x)}\right)^2} dx \precsim |\theta_{1,n}|^8 , \quad \max_{1\leq|\alpha|\leq 4} \int \frac{R_{2,\alpha}^2(x)}{\left(\sqrt{f_{\eta_{1,n}}(x)} + \sqrt{f_{\eta_{2,n}}(x)}\right)^2} dx \precsim |\theta_{1,n}|^8 .$$

Governed by the above bounds, another application of Cauchy-Schwarz's inequality implies that

$$\int \frac{R^2(x)}{\left(\sqrt{f_{\eta_{1,n}}(x)} + \sqrt{f_{\eta_{2,n}}(x)}\right)^2} dx \precsim |\theta_{1,n}|^8 . \qquad (43)$$

Combining the results from equations (41), (42), and (43), we achieve the conclusion of the claim (40).

C Proofs of auxiliary results

In this appendix, we collect the proofs of several auxiliary results stated throughout the paper.

C.1 Proof of Corollary 1

In order to ease the presentation, we only provide the proof sketch for the localization argument with this corollary. The detail proof argument for the corollary can be argued in similar fashion as that of Theorem 1. In particular,

we consider the iterations t such that $\theta_n^t \in [n^{-a_\ell}, n^{-a_r}]$ where $a_\ell > a_r$. For all such iterations with θ_n^t, invoking Lemma 1, we find that

$$\left|\widetilde{M}_{n,1}(\theta_n^t)\right| \lesssim \underbrace{(1 - n^{-6a_\ell})}_{=:\gamma_{a_\ell}} \left|\theta_n^t\right| \quad \text{and} \quad \left|M_{n,1}(\theta_n^t) - \widetilde{M}_{n,1}(\theta_n^t)\right| \lesssim n^{-a_r}/\sqrt{n}.$$

Therefore, we obtain that

$$\left|\theta_n^{t+T}\right| \leq \left|\widetilde{M}_{n,1}(\theta_n^{t+T-1})\right| + \left|\widetilde{M}_{n,1}(\theta_n^{t+T-1}) - M_{n,1}(\theta_n^{t+T-1})\right| \leq \gamma_{a_\ell}\theta_n^{t+T-1} + n^{-a_r}/\sqrt{n}.$$

Unfolding the above inequality T times, we find that

$$\begin{aligned}\left|\theta_n^{t+T}\right| \leq \gamma_{a_\ell}^2(\theta_n^{t+T-2}) + n^{-a_r}/\sqrt{n}(1+\gamma_m) &\leq \gamma_{a_\ell}^T\theta_n^t + (1 + \gamma_{a_\ell} + \ldots + \gamma_{a_\ell}^{T-1})n^{-a_r}/\sqrt{n} \\ &\leq e^{-Tn^{-6a_\ell}}n^{-a_r} + \frac{1}{1-\gamma_{a_\ell}} \cdot n^{-a_r}/\sqrt{n}.\end{aligned}$$

As T is sufficiently large such that the second term is the dominant term, we find that that

$$\left|\theta_n^{t+T}\right| \lesssim \frac{1}{1-\gamma_{a_\ell}} \cdot n^{-a_r}/\sqrt{n} = n^{6a_\ell - a_r - 1/2}.$$

Setting the RHS equal to n^{-a_ℓ}, we obtain the recursion that

$$a_\ell = \frac{a_r}{7} + \frac{1}{14}. \tag{44}$$

Solving for the limit $a_\ell = a_r = a_\star$ yields that $a_\star = 1/12$. It suggests that we eventually have $\theta_n^t \to \mathbb{B}(0, n^{-\frac{1}{12}})$. As a consequence, we achieve the conclusion of the corollary.

C.2 Proof of Lemma 2

Without loss of generality, we can assume that $|\theta| \in [\omega^{-a_{\ell+1}}, \omega^{-a_\ell}]$. Conditional on the event $\mathcal{E}$, we have that

$$\left|\overline{M}_1(\theta)\right| \leq (1 - \omega^{-6a_{\ell+1}}/5)\,|\theta| \quad \text{and} \quad \left|M_{n,1}(\theta) - \overline{M}_1(\theta)\right| \leq c_2\omega^{-3a_\ell}\omega^{-\frac{1}{2}}.$$

As a result, we have

$$\begin{aligned}|M_{n,1}(\theta)| \leq \left|M_{n,1}(\theta) - \overline{M}_1(\theta)\right| + \left|\overline{M}_1(\theta)\right| &\leq (1 - \omega^{-6a_{\ell+1}}/5)\,|\theta| + c_2\omega^{-\frac{1}{2}}\omega^{-3a_\ell} \\ &\leq (1 - \omega^{-6a_{\ell+1}}/5 + c_2\omega^{-\frac{1}{2}}\omega^{-2a_\ell})\omega^{-a_\ell} \\ &\leq \omega^{-a_\ell}.\end{aligned}$$

Here, to establish the last inequality, we have used the following observation: for $\omega = n/c_{n,\delta}$ and that $n \geq (c')^{1/\beta}c_{n,\delta}$, we have

$$5c_2\omega^{6a_{\ell+1}-2a_\ell-1/2} \leq 5c_2\omega^{4a_\ell-1/2} \leq c'\omega^{4a_{\ell_\star}-1/2} \leq c'\omega^{-4\beta} \leq 1/(c')^3 \leq 1,$$

which leads to $-\omega^{-6a_{\ell+1}}/5 + c_2\omega^{-\frac{1}{2}}\omega^{-2a_\ell} \leq 0$. As a consequence, we achieve the conclusion of the lemma.

C.3 Proof of Lemma 3

The proof of the perturbation bound (23b) is a standard extension of $d = 1$ case presented above in Section A.3.2, and thereby is omitted.

We now present the proof of the contraction bound (23a), which has several similarities with the proofs of bounds (12a) and (12b) from Lemma 1. In order to simplify notation, we use the shorthand $Z_{n,d} := \frac{1}{nd}\sum_{j=1}^n \|X_j\|_2^2$. Recalling the definition (22) of operator $\widetilde{M}_{n,d}(\theta)$, we have

$$\|\widetilde{M}_{n,d}(\theta)\|_2 = \left\|\mathbb{E}_{Y\sim\mathcal{N}(0,1)}\left[Y\tanh\left(\frac{Y^\top\theta}{Z_{n,d} - \|\theta\|_2^2/d}\right)\right]\right\|_2. \tag{45}$$

We can find an orthonormal matrix R such that $R\theta = \|\theta\|_2 e_1$, where e_1 is the first canonical basis in $\mathbb{R}^d$. Define the random vector $V = RY$. Since $Y \sim \mathcal{N}(0, I_d)$, we have that $V \sim \mathcal{N}(0, I_d)$. On performing the change of variables $Y = R^\top V$, we find that

$$\begin{aligned}\|\mathbb{E}_Y\left[Y \tanh\left(\frac{Y^\top\theta}{Z_{n,d} - \|\theta\|_2^2/d}\right)\right]\|_2 &= \|\mathbb{E}_V\left[R^\top V \tanh\left(\frac{\|\theta\|_2 V_1}{Z_{n,d} - \|\theta\|_2^2/d}\right)\right]\|_2 \\ &= \left|\mathbb{E}_{V_1}\left[V_1 \tanh\left(\frac{\|\theta\|_2 V_1}{Z_{n,d} - \|\theta\|_2^2/d}\right)\right]\right|\end{aligned}$$

where the final equality follows from the fact that

$$\mathbb{E}[R^\top V f(V_1)] = R^\top \mathbb{E}[V f(V_1)] = R^\top(\mathbb{E}[V_1 f(V_1)], 0, \ldots, 0)^\top.$$

Furthermore, the orthogonality of the matrix R implies that $\|\mathbb{E}[R^\top V f(V_1)]\|_2^2 = |\mathbb{E}[V_1 f(V_1)]|^2$.

In order to simplify the notation, we define the scalars a, b and the event $\mathcal{E}_{\alpha,d}$ as follows:

$$a := 1 - (nd)^{-\alpha}, \quad b := 1 + (nd)^{-\alpha}, \quad \text{and} \quad \mathcal{E}_{\alpha,d} = \left\{|Z_{n,d} - 1| \le (nd)^{-\alpha}\right\}, \tag{46a}$$

where α is a suitable scalar to be specified later. Note that standard chi-squared tail bounds guarantee that

$$\mathbb{P}[\mathcal{E}_{\alpha,d}] \ge 1 - 2e^{-d^{2\alpha} n^{1-2\alpha}/8}. \tag{46b}$$

Now conditional on the event $\mathcal{E}_{\alpha,d}$, we have

$$\begin{aligned}\|\widetilde{M}_{n,d}(\theta)\|_2 &\le \left|\mathbb{E}_{V_1}\left[V_1 \tanh\left(\frac{\|\theta\|_2 V_1}{a - \|\theta\|_2^2/d}\right)\right]\right| = \|\theta\|_2 \underbrace{\mathbb{E}_{V_1}\left[\frac{V_1}{\|\theta\|_2}\tanh\left(\frac{\|\theta\|_2 V_1}{a - \|\theta\|_2^2/d}\right)\right]}_{=:\overline{\rho}(\theta)}, \quad \text{and,} \\ \|\widetilde{M}_{n,d}(\theta)\|_2 &\ge \left|\mathbb{E}_{V_1}\left[V_1 \tanh\left(\frac{\|\theta\|_2 V_1}{b - \|\theta\|_2^2/d}\right)\right]\right| = \|\theta\|_2 \underbrace{\mathbb{E}_{V_1}\left[\frac{V_1}{\|\theta\|_2}\tanh\left(\frac{\|\theta\|_2 V_1}{b - \|\theta\|_2^2/d}\right)\right]}_{=:\underline{\rho}(\theta)},\end{aligned}$$

where the above inequalities follow from the fact that for any fixed $y, \theta \in \mathbb{R}^d$, the function $w \mapsto y\tanh(y\|\theta\|_2/(w - \|\theta\|_2^2/d))$ is non-increasing in w for $w > \|\theta\|_2^2/d$.

Substituting $\alpha = 1/2 - 2\beta$ in the bound (46b) and invoking the large sample size assumption in the theorem statement, we obtain that $\mathbb{P}[\mathcal{E}_{\alpha,d}] \ge 1 - \delta$. Putting these observations together, it remains to prove that

$$\underline{\rho}(\theta) \ge \left(1 - \frac{3\|\theta\|_2^2}{4}\right)\|\theta\|_2^2, \quad \text{and} \quad \overline{\rho}(\theta) \le \left(1 - \left(1 - \frac{1}{d}\right)\frac{\|\theta\|_2^2}{4}\right)\|\theta\|_2^2, \tag{47}$$

for all $5(d/n)^{-1/4+\beta} \le \|\theta\|_2^2 \le (d-1)/(6d-1)$ conditional on the event $\mathcal{E}_{\alpha,d}$ for $\alpha = 1/2 - 6\beta$ to obtain the conclusion of the theorem.

The proof of the claims in equation (47) relies on the following bounds on the hyperbolic function $\tanh(x)$. For any $x \in \mathbb{R}$, the following bounds hold:

$$\text{(Upper bound)} \quad x^2 - \frac{x^4}{3} + \frac{2x^6}{15} \;\ge\; x\tanh(x) \;\ge\; x^2 - \frac{x^4}{3} \quad \text{(Upper bound)}. \tag{48}$$

We omit the proof of these bounds, as it is very similar to that of similar results stated and proven later in Lemma 4. We now turn to proving the bounds stated in equation (47) one-by-one.

Bounding $\overline{\rho}(\theta)$: Applying the upper bound (48) for $x\tanh(x)$, we obtain that

$$\overline{\rho}(\theta) \le \frac{a - \|\theta\|_2^2/d}{\|\theta\|_2^2}\left(\frac{\|\theta\|_2^2}{(a - \|\theta\|_2^2/d)^2}\mathbb{E}\left[V_1^2\right] - \frac{\|\theta\|_2^4}{3(a - \|\theta\|_2^2/d)^4}\mathbb{E}\left[V_1^4\right] + \frac{2\|\theta\|_2^6}{15(a - \|\theta\|_2^2/d)^6}\mathbb{E}\left[V_1^6\right]\right).$$

Substituting $\mathbb{E}\left[V_1^{2k}\right] = (2k-1)!!$ for $k = 1, 2, 3$ in the RHS above, we find that

$$\overline{\rho}(\theta) \leq \frac{1}{a - \|\theta\|_2^2/d} - \frac{\|\theta\|_2^2}{(a - \|\theta\|_2^2/d)^3} + \frac{2\|\theta\|_2^4}{(a - \|\theta\|_2^2/d)^5}. \tag{49}$$

The condition $\|\theta\|_2^2 + (nd)^{-\alpha} \leq \frac{d-1}{6d-4} < 1/6$ implies the following bounds:

$$\begin{aligned}
\frac{1}{1 - (nd)^{-\alpha} - \|\theta\|_2^2/d} &\leq 1 + \left((nd)^{-\alpha} + \|\theta\|_2^2/d\right) + 3/2 \cdot \left((nd)^{-\alpha} + \|\theta\|_2^2/d\right)^2, \\
\frac{1}{(1 - (nd)^{-\alpha} - \|\theta\|_2^2/d)^3} &\geq 1 + 3\left((nd)^{-\alpha} + \|\theta\|_2^2/d\right), \\
\frac{1}{(1 - (nd)^{-\alpha} - \|\theta\|_2^2/d)^5} &\leq 3/2.
\end{aligned}$$

Substituting the definitions (46a) of a and b and plugging the previous three bounds on the RHS of the inequality (49) yields that

$$\begin{aligned}
\overline{\rho}(\theta) &\leq 1 + \frac{\|\theta\|_2^2}{d} + \frac{3\|\theta\|_2^4}{2d^2} - \|\theta\|_2^2\left(1 + \frac{3\|\theta\|_2^2}{d}\right) + 3\|\theta\|_2^4 + \frac{11}{2}(nd)^{-\alpha} \\
&\leq 1 - \left(1 - \frac{1}{d}\right)\|\theta\|_2^2 + \left(3 - \frac{2}{d}\right)\|\theta\|_2^4 + \frac{11}{2}(nd)^{-\alpha} \\
&\leq 1 - \left(1 - \frac{1}{d}\right)\frac{\|\theta\|_2^2}{4}
\end{aligned}$$

where the last step follows from the following observations that

$$(3 - 2/d)\|\theta\|_2^4 \leq (1 - 1/d)\|\theta\|_2^2/2, \quad \text{for all } \|\theta\|_2 \leq (d-1)/(6d-4), \tag{50}$$

$$11(nd)^{-\alpha}/2 \leq (1 - 1/d)\|\theta\|_2^2/4, \quad \text{for all } \|\theta\|_2 \geq 5(d/n)^{-1/4+\beta} \text{ when } \alpha = 1/2 - 2\beta. \tag{51}$$

Therefore, the claim with an upper bound of $\overline{\rho}(\theta)$ now follows.

Bounding $\underline{\rho}(\theta)$: Using the lower bound (48) for $x\tanh(x)$, we find that

$$\underline{\rho}(\theta) \geq \frac{b - \|\theta\|_2^2/d}{\|\theta\|_2^2}\left(\frac{\|\theta\|_2^2}{(b - \|\theta\|_2^2/d)^2}\mathbb{E}\left[V_1^2\right] - \frac{\|\theta\|_2^4}{3(b - \|\theta\|_2^2/d)^4}\mathbb{E}\left[V_1^4\right]\right) \tag{52}$$

$$= \frac{1}{b - \|\theta\|_2^2/d} - \frac{\|\theta\|_2^2}{(b - \|\theta\|_2^2/d)^3}. \tag{53}$$

The condition $\|\theta\|_2 - (nd)^{-\alpha} \geq 0$ leads to

$$\begin{aligned}
\frac{1}{1 + (nd)^{-\alpha} - \|\theta\|_2^2/d} &\geq 1 + \left(\|\theta\|_2^2/d - (nd)^{-\alpha}\right) + \left(\|\theta\|_2^2/d - (nd)^{-\alpha}\right)^2, \\
\frac{1}{(1 + (nd)^{-\alpha} - \|\theta\|_2^2/d)^3} &\leq 1 + 4\left(\|\theta\|_2^2/d - (nd)^{-\alpha}\right).
\end{aligned}$$

Applying these inequalities to the bound (53), we obtain that

$$\begin{aligned}
\underline{\rho}(\theta) &\geq 1 + \frac{\|\theta\|_2^2}{d} + \frac{\|\theta\|_2^4}{d^2} - \|\theta\|_2^2\left(1 + \frac{4\|\theta\|_2^2}{d}\right) - 2(nd)^{-\alpha} \\
&\overset{(i)}{\geq} 1 - \|\theta\|_2^2\left(1 - \frac{1}{d}\right) - \frac{\|\theta\|_2^2}{6}\left(\frac{4}{d} - \frac{1}{d^2}\right) - \frac{\|\theta\|_2^2(1 - 1/d)}{11} \\
&\geq 1 - \frac{3\|\theta\|_2^2}{4}
\end{aligned}$$

where step (i) in the above inequalities follows from the observations (50)-(51) above. The lower bound (47) for $\underline{\rho}(\theta)$ now follows.

C.4 Proof of Lemma 4

The proof of this lemma relies on an evaluation of coefficients with x^{2k} as $k \geq 1$. In particular, we divide the proof of the lemma into two key parts:

Upper bound: From the definition of hyperbolic function $\tanh(x)$, it is sufficient to demonstrate that

$$x\left(\exp(x) - \exp(-x)\right) \leq \left(x^2 - \frac{x^4}{3} + \frac{2x^6}{15} - \frac{17x^8}{315} + \frac{62x^{10}}{2835}\right)\left(\exp(x) + \exp(-x)\right).$$

Invoking the Taylor series of $\exp(x)$ and $\exp(-x)$, the above inequality is equivalent to

$$\sum_{k=0}^{\infty} \frac{2x^{2k+2}}{(2k+1)!} \leq \left(x^2 - \frac{x^4}{3} + \frac{2x^6}{15} - \frac{17x^8}{315} + \frac{62x^{10}}{2835}\right)\left(\sum_{k=0}^{\infty} \frac{2x^{2k}}{(2k)!}\right).$$

Our approach to solve the above inequality is to show that the coefficients of x^{2k} in the LHS is smaller than that of x^{2k} in the RHS for all $k \geq 1$. In fact, when $1 \leq k \leq 3$, we can quickly check that the previous observation holds. For $k \geq 4$, it suffices to validate that

$$\frac{2}{(2k)!} - \frac{2}{3(2k-2)!} + \frac{4}{15(2k-4)!} - \frac{34}{315(2k-6)!} + \frac{124}{2835(2k-8)!} - \frac{2}{(2k+1)!} \geq 0.$$

Direct computation with the above inequality leads to

$$(k-1)(k-2)(k-3)(k-4)(496k^4 - 1736k^3 + 1430k^2 + 446k - 381) \geq 0$$

for all $k \geq 4$, which is always true. As a consequence, we achieve the conclusion with the upper bound of the lemma.

Lower bound: For the lower bound of the lemma, it is equivalent to prove that

$$\sum_{k=0}^{\infty} \frac{2x^{2k+2}}{(2k+1)!} \geq \left(x^2 - \frac{x^4}{3} + \frac{2x^6}{15} - \frac{17x^8}{315}\right)\left(\sum_{k=0}^{\infty} \frac{2x^{2k}}{(2k)!}\right).$$

Similar to the proof technique with the upper bound, we only need to verify that

$$\frac{2}{(2k)!} - \frac{2}{3(2k-2)!} + \frac{4}{15(2k-4)!} - \frac{34}{315(2k-6)!} - \frac{2}{(2k+1)!} \leq 0$$

for any $k \geq 3$. The above inequality is identical to

$$(k-1)(k-2)(k-3)(4352k^3 - 4352k^2 - 512k + 1472) \geq 0$$

for all $k \geq 3$, which always holds. Therefore, we obtain the conclusion with the lower bound of the lemma.

C.5 Proof of Lemma 5

The proof of this lemma is based on appropriate truncation argument. More concretely, given any positive scalar τ, and the random variable $X \sim \mathcal{N}(0,1)$, consider the pair of truncated random variables (Y, Z) defined by:

$$Y := X^{2k}\mathbb{I}_{|X| \leq \tau} \quad \text{and} \quad Z := X^{2k}\mathbb{I}_{|X| \geq \tau}. \tag{54}$$

With the above notation in place, for n i.i.d. samples $X_1, \ldots, X_n$ from $\mathcal{N}(0,1)$, we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i^{2k} = \frac{1}{n}\sum_{i=1}^{n} Y_i + \frac{1}{n}\sum_{i=1}^{n} Z_i := S_{Y,n} + S_{Z,n}.$$

where $S_{Y,n}$ and $S_{Z,n}$, denote the averages of the random variables $Y_i's$ and $Z_i's$ respectively. Observe that $|Y_i| \leq \tau^{2k}$ for all $i \in [n]$; consequently, by standard sub-Gaussian concentration of bounded random variables, we have

$$\mathbb{P}\left(|S_{Y,n} - \mathbb{E}\left[Y\right]| \geq t_1\right) \leq 2\exp\left(-\frac{nt_1^2}{2\tau^{4k}}\right). \tag{55}$$

Next, applying Markov's inequality with the non-negative random variable $S_{Z,n}$, we find that

$$\mathbb{P}\left(S_{Z,n} \geq t_2\right) \leq \frac{\mathbb{E}\left[S_{Z,n}\right]}{t_2} = \frac{\mathbb{E}\left[Z\right]}{t_2}. \tag{56}$$

By definition of the truncated random variable Y, we have $\mathbb{E}[Y] \leq \mathbb{E}[X^{2k}]$; moreover, an application of Holder's inequality to $\mathbb{E}\left[Z\right]$ yields

$$\mathbb{E}\left[Z\right] = \mathbb{E}\left(X^{2k}\mathbb{I}_{|X|\geq\tau}\right) \leq \sqrt{\mathbb{E}\left[X^{4k}\right]}\sqrt{\mathbb{P}\left(|X| \geq \tau\right)} \leq \sqrt{2\mathbb{E}\left[X^{4k}\right]}\exp(-\tau^2/4).$$

Combining the bounds on $\mathbb{E}[Y]$ and $\mathbb{E}[Z]$ with the inequalities (55) and (56) we deduce that

$$\frac{\sum_{i=1}^n X_i^{2k}}{n} \leq \mathbb{E}\left[Y\right] + t_1 + t_2 \leq \mathbb{E}\left[X^{2k}\right] + t_1 + t_2, \quad \text{and,} \tag{57a}$$

$$\frac{\sum_{i=1}^n X_i^{2k}}{n} \geq \mathbb{E}\left[X^{2k}\right] - t_1 - t_2\sqrt{2\mathbb{E}\left[X^{4k}\right]}\exp(-\tau^2/4) \tag{57b}$$

with probability at least $1 - \exp\left(-\frac{nt_1^2}{2\tau^{4k}}\right) - \sqrt{2\mathbb{E}\left[X^{4k}\right]}\exp(-\tau^2/4)$. Finally, given any $\delta > 0$, choose the scalars τ, t_1, t_2 as follows:

$$\tau = 2\sqrt{\log\left(\frac{2\sqrt{2n\mathbb{E}\left[X^{4k}\right]}}{\delta}\right)}, \quad t_1 = \tau^2\sqrt{\frac{1}{n}\log\left(\frac{2}{\delta}\right)} \quad \text{and} \quad t_2 = \frac{1}{\sqrt{n}}.$$

Substituting the choice of t_1, t_2 and τ, in bounds (57a) and (57b) we conclude that with probability at least $1 - \delta$

$$\left|\frac{\sum_{i=1}^n X_i^{2k}}{n} - \mathbb{E}\left[X^{2k}\right]\right| \leq \frac{C_k \log^k(n/\delta)}{\sqrt{n}},$$

where C_k is a universal constant that depends only on k. This completes the proof of Lemma 5.

C.6 Proof of one step bound for population EM

We now describe a special one-step contraction property of the population operator.

Lemma 6. *For any vector θ^0 such that $\|\theta^0\| \leq \sqrt{d}$, we have $\|\widetilde{M}_{n,d}(\theta^0)\| \leq \sqrt{2/\pi}$ with probability at least $1 - \delta$.*

The proof of this lemma is a straightforward application of the proof argument in Lemma 3 in Appendix C.3. In order to simplify notations, we use the shorthand $Z_{n,d} = \sum_{j=1}^n \|X_j\|_2^2/(nd)$. Recalling the definition (22) of operator $\widetilde{M}_{n,d}$, we have

$$\|\widetilde{M}_{n,d}(\theta)\|_2 = \left\|\mathbb{E}_{Y\sim\mathcal{N}(0,1)}\left[Y\tanh\left(\frac{Y^\top\theta}{Z_{n,d} - \|\theta\|_2^2/d}\right)\right]\right\|_2.$$

As demonstrated in the proof of Theorem 2, we have the equivalence

$$\|\widetilde{M}_{n,d}(\theta)\|_2 = \mathbb{E}\left[V_1\tanh\left(\frac{\|\theta\|_2 V_1}{Z_{n,d} - \|\theta\|_2^2/d}\right)\right]$$

where $V_1 \sim \mathcal{N}(0,1)$. Since the function $x \tanh\left(\frac{\|\theta\|_2 x}{a - \|\theta\|_2^2/d}\right)$ is an even function in terms of x for any given a, we find that

$$\mathbb{E}\left[V_1 \tanh\left(\frac{\|\theta\|_2 V_1}{Z_{n,d} - \|\theta\|_2^2/d}\right)\right] = \mathbb{E}\left[|V_1| \tanh\left(\frac{\|\theta\|_2 |V_1|}{Z_{n,d} - \|\theta\|_2^2/d}\right)\right]$$
$$\leq \mathbb{E}\left[|V_1|\right] = \sqrt{\frac{2}{\pi}}$$

where the second inequality is due to the basic inequality $\tanh(x) \leq 1$ for all $x \in \mathbb{R}$. The inequality in the above display implies that regardless of the initialization θ^0, we always have $\|\widetilde{M}_{n,d}(\theta)\|_2 \leq \sqrt{2/\pi}$, as claimed.

D Wasserstein Distance

In Figures 1 and 3, we use EM to estimate all the parameters of the fitted Gaussian mixture (e.g., the parameters $\{w_i, \mu_i, \Sigma_i, i \in [k]\}$ if the fitted mixture were $\mathcal{G} = \sum_{i=1}^k w_i \mathcal{N}(\mu_i, \Sigma_i)$) and use first-order Wasserstein distance between the fitted model and the true model to measure the quality of the estimate. Here we briefly summarize the definition of the first-order Wasserstein distance and refer the readers to the book [Villani, 2008] and the paper [Ho and Nguyen, 2016b] for more details. Given two Gaussian mixture distributions of the form

$$\mathcal{G} = \sum_{i=1}^k w_i \mathcal{N}(\mu_i, \Sigma_i) \quad \text{and} \quad \mathcal{G}' = \sum_{j=1}^{k'} w_j \mathcal{N}(\mu'_j, \Sigma_j),$$

the first-order Wasserstein distance between the two is given by

$$W_1(\mathcal{G}, \mathcal{G}') = \inf_{q \in \mathcal{Q}} \sum_{i=1}^k \sum_{j=1}^{k'} q_{ij} \left(\|\theta_i - \theta'_j\|_2 + |||\Sigma_i - \Sigma'_j|||_{\mathbb{F}}\right), \tag{58}$$

where $|||A|||_{\mathbb{F}}$ denotes the Frobenius norm of the matrix A (which in turn is defined as $\sqrt{\sum_{ij} A_{ij}^2}$). Moreover, $\mathcal{Q}$ denotes the set of all couplings on $[k] \times [k']$ such that

$$q_{ij} \in [0,1], \qquad \sum_{i=1}^k q_{ij} = w'_j \quad \text{and} \quad \sum_{j=1}^{k'} q_{ij} = w_i \quad \text{for all} \quad i \in [k], j \in [k'].$$

We note that the optimization problem (58) is a linear program in the $k \times k'$ dimensional variable q and standard linear program solvers can be used for solving it. Also, we remark that here we have abused the notation slightly since the the definition of the Wasserstein distance above is typically used for the mixing measures which only depends on the parameters of the Gaussian mixture (and not the Gaussian density). Finally, applying definition (58), we can directly conclude that for the symmetric fit (1), we have

$$W_1\left(\frac{1}{2}\mathcal{N}(\theta, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(-\theta, \sigma^2 I_d), \mathcal{N}(\theta_\star, \sigma_\star^2 I_d)\right) = \|\theta - \theta_\star\|_2 + \sqrt{d}\sqrt{|\sigma^2 - \sigma_\star^2|}, \tag{59}$$

where we have assumed that $\min\{\|\theta - \theta_\star\|_2, \| - \theta - \theta_\star\|_2\} = \|\theta - \theta_\star\|_2$.