Towards Clustering High-dimensional Gaussian Mixture Clouds in Linear Running Time

Dan Kushnir, Shirin Jalali, Iraj Saniee *

In this supplementary file, we provide the proofs of all the results provided in the main file. To keep the numbering of the theorems, lemmas, etc. consistent with the main paper, here, we also repeat the results in the paper that do not need a proof.

Theorem 1 *Consider two spherical Gaussian distributions $\mathcal{N}(\mathbf{m}_1, \sigma_1^2 I_p)$ and $\mathcal{N}(\mathbf{m}_2, \sigma_2^2 I_p)$ in $\mathbb{R}^p$. Consider projecting these two Gaussian distributions on $\mathbb{R}$, using $\mathbf{A} = (A_1, \ldots, A_p)$, where $A_1, \ldots, A_p$ are i.i.d. $\mathcal{N}(0, 1)$. Given $\gamma > 0$, let*

$$c \triangleq \frac{\|\mathbf{m}_1 - \mathbf{m}_2\|}{(\sigma_1 + \sigma_2)\sqrt{p}}.$$

Then, the probability that the separation of the projected Gaussians is larger than γ is larger than

$$2\left(1 - \mathrm{e}^{-\frac{p-1}{2}(\tau - \log(1+\tau))}\right) Q\left(\frac{\gamma}{c}\sqrt{\frac{(1-\frac{1}{p})}{(1-\frac{\gamma^2}{pc^2})}(1+\tau)}\right), \tag{1}$$

where $\tau > 0$ is a free parameter.

Proof 1 *Since the unitary vector $\mathbf{A}/\|\mathbf{A}\|$ is uniformly distributed over the unit sphere, we have*

$$\begin{aligned} \mathrm{P}\left(\left|\langle\frac{\mathbf{m}_1 - \mathbf{m}_2}{\|\mathbf{m}_1 - \mathbf{m}_2\|}, \frac{\mathbf{A}}{\|\mathbf{A}\|}\rangle\right| > \frac{\gamma(\sigma_1+\sigma_2)}{\|\mathbf{m}_1 - \mathbf{m}_2\|}\right) &= \mathrm{P}\left(\left|\langle(1,0,\ldots,0)^T, \frac{\mathbf{A}}{\|\mathbf{A}\|}\rangle\right| > \frac{\gamma(\sigma_1+\sigma_2)}{\|\mathbf{m}_1 - \mathbf{m}_2\|}\right) \\ &= \mathrm{P}\left(\frac{|A_1|}{\|\mathbf{A}\|} > \frac{\gamma(\sigma_1+\sigma_2)}{\|\mathbf{m}_1 - \mathbf{m}_2\|}\right). \end{aligned} \tag{2}$$

Therefore, we are interested in deriving a lower bound on

$$\mathrm{P}\left(\frac{|A_1|}{\|\mathbf{A}\|} > \frac{\gamma(\sigma_1+\sigma_2)}{\|\mathbf{m}_1 - \mathbf{m}_2\|}\right). \tag{3}$$

Note that, due to symmetry, we have

$$\mathrm{E}[\frac{A_1^2}{\|\mathbf{A}\|^2}] = \mathrm{E}[\frac{A_2^2}{\|\mathbf{A}\|^2}] = \ldots = \mathrm{E}[\frac{A_p^2}{\|\mathbf{A}\|^2}], \tag{4}$$

Moreover,

$$\sum_{i=1}^p \mathrm{E}[\frac{A_i^2}{\|\mathbf{A}\|^2}] = \mathrm{E}\left[\frac{\sum_{i=1}^p A_i^2}{\|\mathbf{A}\|^2}\right] = 1. \tag{5}$$

Therefore, combining (4) and (5), we have

$$\mathrm{E}[\frac{A_1^2}{\|\mathbf{A}\|^2}] = \frac{1}{p}.$$

On the other hand, replacing $\gamma^2(\sigma_1+\sigma_2)^2/\|\mathbf{m}_1 - \mathbf{m}_2\|^2$ by α/p in (3), where

$$\alpha \triangleq \frac{\gamma^2(\sigma_1+\sigma_2)^2 p}{\|\mathbf{m}_1 - \mathbf{m}_2\|^2},$$

*first.last@nokia-bell-labs.com

we have

$$\mathrm{P}\left(\frac{A_1^2}{\|\mathbf{A}\|^2} > \frac{\gamma^2(\sigma_1+\sigma_2)^2}{\|\mathbf{m}_1-\mathbf{m}_2\|^2}\right) = \mathrm{P}\left(A_1^2 > \frac{\alpha}{p}\sum_{i=1}^{p} A_i^2\right)$$
$$= \mathrm{P}\left((1-\frac{\alpha}{p})A_1^2 > \frac{\alpha}{p}\sum_{i=2}^{p} A_i^2\right). \tag{6}$$

But since $A_1,\ldots,A_p \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $\sum_{i=2}^{p} A_i^2$ *has a chi-square distribution of order* p*. Then, for any* $\tau>0$*, by Lemma 2 in [2],*

$$\mathrm{P}\left(\frac{1}{p-1}\sum_{i=2}^{p} A_i^2 \geq 1+\tau\right) \leq \mathrm{e}^{-\frac{p-1}{2}(\tau-\log(1+\tau))}. \tag{7}$$

Given $\tau>0$*, define event* $\mathcal{E}$ *as*

$$\mathcal{E} \triangleq \{\frac{1}{p-1}\sum_{i=2}^{p} A_i^2 < 1+\tau\}.$$

By the law of total probability,

$$\mathrm{P}\left((1-\frac{\alpha}{p})A_1^2 > \frac{\alpha}{p}\sum_{i=2}^{p} A_i^2\right) = \mathrm{P}\left((1-\frac{\alpha}{p})A_1^2 > \frac{\alpha}{p}\sum_{i=2}^{p} A_i^2, \mathcal{E}\right) + \mathrm{P}\left((1-\frac{\alpha}{p})A_1^2 > \frac{\alpha}{p}\sum_{i=2}^{p} A_i^2, \mathcal{E}^c\right)$$
$$\geq \mathrm{P}\left((1-\frac{\alpha}{p})A_1^2 > \frac{\alpha}{p}\sum_{i=2}^{p} A_i^2, \mathcal{E}\right)$$
$$\overset{(2)}{\geq} \mathrm{P}\left((1-\frac{\alpha}{p})A_1^2 > \frac{\alpha(p-1)}{p}(1+\tau) > \frac{\alpha}{p}\sum_{i=2}^{p} A_i^2\right)$$
$$= \mathrm{P}\left((1-\frac{\alpha}{p})A_1^2 > \frac{\alpha(p-1)}{p}(1+\tau), 1+\tau > \frac{1}{p-1}\sum_{i=2}^{p} A_i^2\right)$$
$$= \mathrm{P}\left((1-\frac{\alpha}{p})A_1^2 > \alpha(1-\frac{1}{p})(1+\tau)\right)\mathrm{P}\left(\frac{1}{p-1}\sum_{i=2}^{p} A_i^2 < 1+\tau\right),$$

where inequality (2) results from the fact that if a random variable Y *is smaller than c, w.p. 1, then* $P(X>Y) \geq P(X>c)$*. The last equality follows from the independence of* A_1 *and* $(A_2,\ldots,A_p)$*.*

Lemma 1 *Consider points in* $\mathbb{R}$ *drawn from a mixture of two Gaussian distributions* $w\mathcal{N}(m_1,\sigma_1)+(1-w)\mathcal{N}(m_2,\sigma_2)$ *. Assume that the two components of the mixture are c-separated. Then, the error probability of the optimal Bayesian classifier is smaller than* $Q\left(\frac{c}{2}\right)$*. In the special case where* $\sigma_1=\sigma_2=\sigma$*, the error probability of the optimal Bayesian classifier is smaller than* $Q(c)$*.*

Proof 2 *Without loss of generality assume that* $m_1 \leq m_2$*. Consider the a sub-optimal classifier that assigns all points to the left of* $\frac{m_1+m_2}{2}$ *to Class 1 and everything else to class 2. Then the error probability of the optimal Bayesian classifier is upper bounded by the error achieved by the described classifier, which is equal to*

$$wQ(\frac{m_2-m_1}{2\sigma_1}) + (1-w)Q(\frac{m_2-m_1}{2\sigma_2}). \tag{8}$$

But, by assumption, $\frac{m_2-m_1}{\sigma_1+\sigma_2} \geq c$*. Therefore,*

$$\min(\frac{m_2-m_1}{2\sigma_1}, \frac{m_2-m_1}{2\sigma_2}) \geq \frac{c}{2}.$$

Therefore, since Q *is a decreasing function of its argument,*

$$wQ(\frac{m_2-m_1}{2\sigma_1}) + (1-w)Q(\frac{m_2-m_1}{2\sigma_2}) \leq wQ(\frac{c}{2}) + (1-w)Q(\frac{c}{2}) = Q(\frac{c}{2}).$$

In the case of $\sigma_1=\sigma_2=\sigma$*, (8) simplifies to* $Q(\frac{m_2-m_1}{2\sigma})$ *which is smaller than* $Q(c)$*, because* $(m_2-m_1)/2\sigma \geq c$*.*

Lemma 2 *Consider the same setup as Theorem 1. Then,*

$$\mathrm{E}\left[\frac{|\langle \mathbf{A}, \mathbf{m}_1 - \mathbf{m}_2 \rangle|^2}{(\sigma_1+\sigma_2)^2\|\mathbf{A}\|^2}\right] = c^2. \tag{9}$$

Proof 3 *Let $\mathbf{A} = (A_1, \ldots, A_p)$ be generated i.i.d. according to $\mathcal{N}(0,1)$. Then the separation of the two projected Gaussians under $\mathbf{A}$ is equal to*

$$\gamma = \frac{|\langle \mathbf{A}, \mathbf{m}_1 \rangle - \langle \mathbf{A}, \mathbf{m}_2 \rangle|}{(\sigma_1+\sigma_2)\|\mathbf{A}\|}.$$

Therefore,

$$\begin{aligned}\gamma^2 &= \frac{|\langle \mathbf{A}, \mathbf{m}_1 - \mathbf{m}_2 \rangle|^2}{(\sigma_1+\sigma_2)^2\|\mathbf{A}\|^2} \\ &= \frac{\|\mathbf{m}_1 - \mathbf{m}_2\|^2}{(\sigma_1+\sigma_2)^2}\left|\left\langle \frac{\mathbf{A}}{\|\mathbf{A}\|}, \frac{\mathbf{m}_1 - \mathbf{m}_2}{\|\mathbf{m}_1 - \mathbf{m}_2\|}\right\rangle\right|^2.\end{aligned}$$

Since $\frac{\mathbf{A}}{\|\mathbf{A}\|}$ is uniformly distributed under the unit sphere in $\mathbb{R}^p$, in evaluating $\mathrm{E}\,\gamma^2$, without loss of generality we can assume that $\frac{\mathbf{m}_1-\mathbf{m}_2}{\|\mathbf{m}_1-\mathbf{m}_2\|} = (1, 0, \ldots, 0)^T$. Therefore,

$$\begin{aligned}\mathrm{E}[\gamma^2] &= \frac{\|\mathbf{m}_1 - \mathbf{m}_2\|^2}{(\sigma_1+\sigma_2)^2}\,\mathrm{E}\left[\left|\left\langle \frac{\mathbf{A}}{\|\mathbf{A}\|}, \frac{\mathbf{m}_1 - \mathbf{m}_2}{\|\mathbf{m}_1 - \mathbf{m}_2\|}\right\rangle\right|^2\right] \\ &= \frac{\|\mathbf{m}_1 - \mathbf{m}_2\|^2}{(\sigma_1+\sigma_2)^2}\,\mathrm{E}\left[\frac{A_1^2}{\|\mathbf{A}\|^2}\right].\end{aligned}$$

But, as we showed in the proof of Theorem 1,

$$\mathrm{E}\left[\frac{A_1^2}{\|\mathbf{A}\|^2}\right] = \frac{1}{p}.$$

Therefore, in summary,

$$\mathrm{E}[\gamma^2] = \frac{\|\mathbf{m}_1 - \mathbf{m}_2\|^2}{(\sigma_1+\sigma_2)^2 p} = c^2.$$

Corollary 1 *Consider the same setup as in Theorem 1. Then,*

$$\lim_{p\to\infty} d(\gamma) \le \frac{1}{2Q(\frac{\gamma}{c})}.$$

Proof 4 *Note that $d(\gamma)$ is equal to one over the probability that the separation of two projected Gaussians is larger than γ. Therefore, by Theorem 1, we have*

$$d(\gamma) \le \left(2\left(1 - \mathrm{e}^{-\frac{p-1}{2}(\tau - \log(1+\tau))}\right) Q\left(\frac{\gamma}{c}\sqrt{\frac{(1-\frac{1}{p})}{(1-\frac{\gamma^2}{pc^2})}(1+\tau)}\right)\right)^{-1}, \tag{10}$$

where $\tau > 0$ is a free parameter. On the other hand, for any $\tau > 0$,

$$\lim_{p\to\infty} \frac{(1-\frac{1}{p})}{(1-\frac{\gamma^2}{pc^2})} = \lim_{p\to\infty} (1 - \mathrm{e}^{-\frac{p-1}{2}(\tau - \log(1+\tau))}) = 1.$$

Therefore,

$$\lim_{p\to\infty} d(\gamma) \le \frac{1}{2Q(\frac{\gamma}{c}\sqrt{1+\tau})}.$$

Since τ is a free parameter, letting $\tau \to 0$ yields the desired result.

Corollary 2 *Consider the same setup as in Theorem 1. If γ is such that $\gamma \le c(\ln\ln p)^{\frac{1-\eta}{2}}$, where $\eta > 0$ is a free parameter, then $d(\gamma) = o(\ln p)$.*

Proof 5 *As argued in the proof of Corollary 1, $d(\gamma)$ satisfies (10). Hence, choosing p large enough such that $\mathrm{e}^{-\frac{p-1}{2}(\tau-\log(1+\tau))} \leq \frac{1}{2}$, it follows that*

$$d(\gamma) \leq \frac{1}{Q\left(\sqrt{\frac{(1-\frac{1}{p})\gamma^2}{(c^2-\frac{\gamma^2}{p})}(1+\tau)}\right)}.$$

On the other hand, for all $x > 0$, we have

$$\frac{x}{1+x^2}\phi(x) < Q(x), \tag{11}$$

where $\phi(x)$ denotes the pdf of a standard normal distribution. Therefore,

$$d(\gamma) \leq \frac{\sqrt{2\pi}(1+x^2)}{x}\mathrm{e}^{\frac{x^2}{2}},$$

where $x = \sqrt{\frac{(1-\frac{1}{p})\gamma^2}{(c^2-\frac{\gamma^2}{p})}(1+\tau)}$. The desired result follows by noting that $\frac{\gamma^2}{p} = o(1)$, and by assumption, $\gamma \leq c(\ln\ln p)^{\frac{1-\eta}{2}}$, where $\eta > 0$.

Corollary 3 *Consider the same setup as in Theorem 1. If γ is such that $\gamma \leq c(\ln p)^{\frac{1-\eta}{2}}$, where $\eta > 0$ is a free parameter, then $d(\gamma) = o(p)$.*

Theorem 2 *Consider $\mathbf{m}_1, ..., \mathbf{m}_k \in \mathbb{R}^p$ and $\sigma_1, ..., \sigma_k \in \mathbb{R}^+$. Assume that $\mathbf{A} = (A_1, \ldots, A_p)$ are generated i.i.d. according to $\mathcal{N}(0,1)$. Given $\gamma_{\min} > 0$, and $i, j \in \{1, \ldots, k\}$ let*

$$c_{(i,j)} = \frac{\|\mathbf{m}_i - \mathbf{m}_j\|}{\sqrt{p}(\sigma_i + \sigma_j)}.$$

Let $c_{\min} \triangleq \min_{i,j} c_{(i,j)}$. Define event $\mathcal{B}$ as having separation larger than $\gamma_{\min}$ by all pairs of projected Gaussians. Thats is,

$$\mathcal{B} \triangleq \left\{\left|\left\langle \mathbf{m}_i - \mathbf{m}_j, \frac{\mathbf{A}}{\|\mathbf{A}\|}\right\rangle\right| \geq \gamma_{\min}(\sigma_i + \sigma_j) : \forall (i,j) \in \{1, \ldots, k\}^2, i \neq j\right\}. \tag{12}$$

Then,

$$\mathrm{P}(\mathcal{B}^c) \leq \frac{k^2}{2}\left(1 - 2Q\left(\frac{\gamma_{\min}}{c_{\min}}\sqrt{\frac{1.1}{1-\frac{\gamma_{\min}^2}{c_{\min}^2 p}}}\right)(1-\mathrm{e}^{-0.002p})\right). \tag{13}$$

Proof 6 (Proof of Lemma 2) *Define*

$$\phi_{(i,j)} \triangleq 2Q\left(\frac{\gamma_{\min}}{c_{(i,j)}}\sqrt{\frac{1-\frac{1}{p}}{1-\frac{\gamma_{\min}^2}{c_{(i,j)}^2 p}}(1+\tau)}\right)(1-\mathrm{e}^{-\frac{p-1}{2}(\tau-\log(1+\tau))}).$$

By the union bound,

$$\begin{aligned}
\mathrm{P}(\mathcal{B}^c) &\leq \sum_{i=1}^{k}\sum_{j=i+1}^{k} \mathrm{P}\left(\left\{\left|\left\langle \mathbf{m}_i - \mathbf{m}_j, \frac{\mathbf{A}}{\|\mathbf{A}\|}\right\rangle\right| \leq \gamma_{\min}(\sigma_i + \sigma_j)\right\}\right) \\
&\stackrel{(a)}{\leq} \sum_{i=1}^{k}\sum_{j=i+1}^{k} (1 - \phi_{(i,j)}) \\
&\leq \frac{k^2}{2}\left(1 - \max_{(i,j)}\{\phi_{(i,j)}\}\right),
\end{aligned} \tag{14}$$

where (a) follows from Theorem 1 and the fact that for $i, j \in \{1, \ldots, k\}$

$$\frac{\gamma_{\min}^2(\sigma_i + \sigma_j)^2 p}{\|\mathbf{m}_i - \mathbf{m}_j\|^2} = \frac{\gamma_{\min}^2}{c_{(i,j)}^2}.$$

For $\tau = 0.1$, $(\tau - \log(1+\tau))/2 \geq 0.002$. Therefore, setting $\tau = 0.1$ in (2) and noting that Q function is a monotonically decreasing function of its argument, it follows that

$$\phi_{(i,j)} \geq 2Q\left(\frac{\gamma_{\min}}{c_{(i,j)}}\sqrt{\frac{1.1(1-\frac{1}{p})}{(1-\frac{\gamma_{\min}^2}{c_{(i,j)}^2 p})}}\right)(1-\mathrm{e}^{-0.002p})$$

$$\geq 2Q\left(\frac{\gamma_{\min}}{c_{\min}}\sqrt{\frac{1.1}{1-\frac{\gamma_{\min}^2}{c_{\min}^2 p}}}\right)(1-\mathrm{e}^{-0.002p}), \tag{15}$$

where the last inequality holds because

$$\frac{(1-\frac{1}{p})}{(1-\frac{\gamma_{\min}^2}{c_{(i,j)}^2 p})} \leq \frac{1}{1-\frac{\gamma_{\min}^2}{c_{\min}^2 p}}.$$

Therefore, taking the maximum of the both sides of (15), it follows that

$$\max_{i,j} \phi_{(i,j)} \geq 2Q\left(\frac{\gamma_{\min}}{c_{\min}}\sqrt{\frac{1.1}{1-\frac{\gamma_{\min}^2}{c_{\min}^2 p}}}\right)(1-\mathrm{e}^{-0.002p}).$$

Corollary 4 *Consider the same setup as Theorem 2. Let $d(\gamma_{\min})$ denote the expected number of projections required to obtain separation $\gamma_{\min}$ between each pair of projected Gaussians. Then, if*

$$\gamma_{\min} \leq (1-\alpha)\sqrt{\frac{2\pi}{1.1}}\frac{c_{\min}}{k^2},$$

for some $\alpha \in (0,1)$, then $\limsup_{p\to\infty} d(\gamma_{\min}) \leq \frac{1}{\alpha}$.

Proof 7 *Consider event $\mathcal{B}$ defined in (12), which denotes the desired event where each pair of projected Gaussians satisfy the desired separation. But,*

$$d(\gamma_{\min}, p) = \frac{1}{\mathrm{P}(\mathcal{B})}, \tag{16}$$

where $\mathrm{P}(\mathcal{B}^c)$ is upper-bounded by Lemma 2. Taking the limit as p grows to infinity, it follows that

$$\limsup_p d(\gamma_{\min}, p) \leq \frac{1}{1-\frac{k^2}{2}\left(1-2Q\left(\frac{\gamma_{\min}\sqrt{1.1}}{c_{\min}}\right)\right)} \tag{17}$$

On the other hand, for $x > 0$,

$$1-2Q(x) = \frac{1}{\sqrt{2\pi}}\int_{-x}^{x}\mathrm{e}^{-\frac{u^2}{2}}du \leq \sqrt{\frac{2}{\pi}}x. \tag{18}$$

Combining (18) and (17), it follows that

$$\limsup_p d(\gamma_{\min}, p) \leq \frac{1}{1-k^2\left(\sqrt{\frac{1.1}{2\pi}}\frac{\gamma_{\min}}{c_{\min}}\right)} \leq \frac{1}{1-k^2\sqrt{\frac{1.1}{2\pi}}(1-\alpha)\sqrt{\frac{2\pi}{1.1}}\frac{1}{k^2}} = \frac{1}{\alpha},$$

where the last inequality follows from our assumption about $\gamma_{\min}$.

Theorem 3 *Consider $\mathbf{m}_1, \mathbf{m}_2 \in \mathbb{R}^p$ and semi-positive definite matrices Σ_1 and Σ_2. Assume that the entries of $\mathbf{A} = (A_1, \ldots, A_p)$ are generated i.i.d. according to $\mathcal{N}(0,1)$. Let $\lambda_{\max}$ denote the maximum eigenvalue of $\Sigma_1 + \Sigma_2$. Also, given $\gamma > 0$, let*

$$\beta \triangleq \frac{2\gamma^2\lambda_{\max}p}{\|\mathbf{m}_1 - \mathbf{m}_2\|^2}.$$

Then, for any $\tau > 0$, the probability that the 1-dimensional projected Gaussians using a uniformly random direction are γ-separated, i.e., $\mathrm{P}\left(|\langle \mathbf{m}_1 - \mathbf{m}_2, \mathbf{A}\rangle| \geq \gamma(\sqrt{\mathbf{A}^T\Sigma_1\mathbf{A}} + \sqrt{\mathbf{A}^T\Sigma_2\mathbf{A}}\ )\right)$, *can be lower-bounded by*

$$\mathrm{P}\left(A_1^2 > \beta\frac{(1-\frac{1}{p})}{(1-\frac{\beta}{p})}(1+\tau)\right)(1 - \mathrm{e}^{-\frac{p-1}{2}(\tau - \log(1+\tau))}). \tag{19}$$

Proof 8 *Note that since* $\mathbf{A}^T\Sigma_1\mathbf{A} \geq 0$ *and* $\mathbf{A}^T\Sigma_2\mathbf{A} \geq 0$, *we always have*

$$\sqrt{\mathbf{A}^T\Sigma_1\mathbf{A}} + \sqrt{\mathbf{A}^T\Sigma_2\mathbf{A}} \leq \sqrt{2\mathbf{A}^T(\Sigma_1 + \Sigma_2)\mathbf{A}}.$$

Therefore,

$$\begin{aligned}
&\mathrm{P}\left(|\langle \mathbf{m}_1 - \mathbf{m}_2, \mathbf{A}\rangle| > \gamma(\sqrt{\mathbf{A}^T\Sigma_1\mathbf{A}} + \sqrt{\mathbf{A}^T\Sigma_2\mathbf{A}}\ \ )\right)\\
&\geq \mathrm{P}\left(|\langle \mathbf{m}_1 - \mathbf{m}_2, \mathbf{A}\rangle| > \gamma\sqrt{2\mathbf{A}^T(\Sigma_1 + \Sigma_2)\mathbf{A}}\right)\\
&\geq \mathrm{P}\left(|\langle \mathbf{m}_1 - \mathbf{m}_2, \mathbf{A}\rangle| > \gamma\sqrt{2\lambda_{\max}}\|\mathbf{A}\|\right),
\end{aligned} \tag{20}$$

where the last line follows because, for every $\mathbf{A}$, $\mathbf{A}^T(\Sigma_1 + \Sigma_2)\mathbf{A} \leq \lambda_{\max}\|\mathbf{A}\|^2$. *Therefore, comparing* (20) *with* (2) *reveals that the desired result follows similar to Theorem 1, by replacing* $\sigma_1 + \sigma_2$ *with* $\sqrt{2\lambda_{\max}}$.

Theorem 4 *Consider* $\mathbf{m}_1, \mathbf{m}_2 \in \mathbb{R}^p$ *and semi-positive definite matrices* Σ_1 *and* Σ_2. *Assume that the entries of* $\mathbf{A} = (A_1, \ldots, A_p)$ *are generated i.i.d. according to* $\mathcal{N}(0,1)$. *Let* r *and* $\lambda_{\max}$ *denote the rank and the maximum eigenvalue of* $\Sigma_1 + \Sigma_2$, *respectively. Also, given* $\gamma > 0$, $\tau_1 \in (0,1)$ *and* $\tau_2 > 0$, *let*

$$\beta \triangleq \frac{2(1+\tau_2)\gamma^2\lambda_{\max}r}{(1-\tau_1)\|\mathbf{m}_1 - \mathbf{m}_2\|^2}.$$

Then, for any $\tau > 0$, the probability that the 1-dimensional projected Gaussians using a uniformly random direction are γ-separated, i.e., $\mathrm{P}\left(|\langle \mathbf{m}_1 - \mathbf{m}_2, \mathbf{A}\rangle| \geq \gamma(\sqrt{\mathbf{A}^T\Sigma_1\mathbf{A}} + \sqrt{\mathbf{A}^T\Sigma_2\mathbf{A}}\ )\right)$, *can be lower-bounded by*

$$\begin{aligned}
&2Q\left(\sqrt{\beta\frac{(1-\frac{1}{p})}{(1-\frac{\beta}{p})}(1+\tau)}\right)(1 - \mathrm{e}^{-\frac{p-1}{2}(\tau - \log(1+\tau))})\\
&\quad - \mathrm{e}^{\frac{p}{2}(\tau_1 + \log(1-\tau_1))} - \mathrm{e}^{-\frac{r}{2}(\tau_2 - \log(1+\tau_2))}.
\end{aligned}$$

Proof 9 *Note that since* $\mathbf{A}^T\Sigma_1\mathbf{A} \geq 0$ *and* $\mathbf{A}^T\Sigma_2\mathbf{A} \geq 0$, *we always have*

$$\sqrt{\mathbf{A}^T\Sigma_1\mathbf{A}} + \sqrt{\mathbf{A}^T\Sigma_2\mathbf{A}} \leq \sqrt{2\mathbf{A}^T(\Sigma_1 + \Sigma_2)\mathbf{A}}.$$

Therefore,

$$\begin{aligned}
&\mathrm{P}\left(|\langle \mathbf{m}_1 - \mathbf{m}_2, \mathbf{A}\rangle| > \gamma(\sqrt{\mathbf{A}^T\Sigma_1\mathbf{A}} + \sqrt{\mathbf{A}^T\Sigma_2\mathbf{A}}\ \ )\right)\\
&\geq \mathrm{P}\left(|\langle \mathbf{m}_1 - \mathbf{m}_2, \mathbf{A}\rangle| > \gamma\sqrt{2\mathbf{A}^T(\Sigma_1 + \Sigma_2)\mathbf{A}}\right)
\end{aligned} \tag{21}$$

Since $\Sigma_1 + \Sigma_2$ *is always a semi-positive definite matrix, it can be decomposed as*

$$\Sigma_1 + \Sigma_2 = P^T D P,$$

where $P \in \mathbb{R}^{p\times p}$ *is an orthogonal matrix* ($P^TP = I_p$), *and* $D \in \mathbb{R}^{p\times p}$ *is a diagonal matrix whose diagonal entries are non-negative. Let*

$$D = \mathrm{diag}(_1, \ldots, _p),$$

where $_i \geq 0$, *for all* i. *Using this decomposition,* $\mathbf{A}^T(\Sigma_1 + \Sigma_2)\mathbf{A}$ *can be written as*

$$\mathbf{A}^T(\Sigma_1 + \Sigma_2)\mathbf{A} = (P\mathbf{A})^T D P\mathbf{A}.$$

Let $\mathbf{B} \triangleq P\mathbf{A}$. Since P is an orthogonal matrix, $\mathbf{B}$ is still distributed as $\mathbf{A}$, i.e., $B_1, \dots, B_p$ are i.i.d. $\mathcal{N}(0,1)$. By this change of variable, the probability mentioned in (21) *can be written as*

$$\begin{aligned}
&\mathrm{P}\Big(|\langle \mathbf{m}_1 - \mathbf{m}_2, \mathbf{A}\rangle| \geq \gamma(\sqrt{\mathbf{A}^T\Sigma_1\mathbf{A}} + \sqrt{\mathbf{A}^T\Sigma_2\mathbf{A}}\ )\Big) \\
&\geq \mathrm{P}\left(|\langle \mathbf{m}_1 - \mathbf{m}_2, P^{-1}\mathbf{B}\rangle| > \gamma\sqrt{2\mathbf{B}^T D\mathbf{B}}\right) \\
&= \mathrm{P}\left(|\langle P(\mathbf{m}_1 - \mathbf{m}_2), \mathbf{B}\rangle| > \sqrt{2\gamma^2}\|D^{\frac{1}{2}}\mathbf{B}\|\right). \qquad (22)
\end{aligned}$$

Note that sine by assumption $\mathrm{rank}(\Sigma_1 + \Sigma_2) = r$*,* $\Sigma_1 + \Sigma_2$ *has only r non-zero eigenvalues. Define $\mathbf{C} \in \mathbb{R}^p$ such that, for $i = 1, \dots, p$,*

$$C_i = B_i \mathbb{1}_{\lambda_i \neq 0}.$$

That is, for every $\lambda_i \neq 0$, C_i is equal to B_i. For every $\lambda_i = 0$, $C_i = 0$. Using this definition, $D^{\frac{1}{2}}\mathbf{B} = D^{\frac{1}{2}}\mathbf{C}$. Note that

$$\|D^{\frac{1}{2}}\mathbf{C}\| \leq \sqrt{\lambda_{\max}}\|\mathbf{C}\|. \qquad (23)$$

Combining (22) *and* (23)*, it follows that*

$$\begin{aligned}
\mathrm{P}\left(|\langle \mathbf{m}_1 - \mathbf{m}_2, \mathbf{A}\rangle| \geq \gamma(\sqrt{\mathbf{A}^T\Sigma_1\mathbf{A}} + \sqrt{\mathbf{A}^T\Sigma_2\mathbf{A}}\ )\right) &\geq \mathrm{P}\left(|\langle P(\mathbf{m}_1 - \mathbf{m}_2), \mathbf{B}\rangle| > \sqrt{2\gamma^2\lambda_{\max}}\|\mathbf{C}\|\right) \\
&= \mathrm{P}\left(|\langle P(\mathbf{m}_1 - \mathbf{m}_2), \frac{\mathbf{B}}{\|\mathbf{B}\|}\rangle| > \sqrt{2\gamma^2\lambda_{\max}}\frac{\|\mathbf{C}\|}{\|\mathbf{B}\|}\right). \qquad (24)
\end{aligned}$$

Given $\tau_1 > 0$ and $\tau_2 > 0$, define events $\mathcal{E}_1$ and $\mathcal{E}_2$ as

$$\mathcal{E}_1 \triangleq \{\|\mathbf{B}\|^2 \geq p(1-\tau_1)\},$$

and

$$\mathcal{E}_2 \triangleq \{\|\mathbf{C}\|^2 \leq r(1+\tau_2)\},$$

respectively. Note that, conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\frac{\|\mathbf{C}\|}{\|\mathbf{B}\|} \leq \sqrt{\frac{r(1+\tau_2)}{p(1-\tau_1)}}.$$

Therefore,

$$\begin{aligned}
\mathrm{P}\left(|\langle P(\mathbf{m}_1 - \mathbf{m}_2), \frac{\mathbf{B}}{\|\mathbf{B}\|}\rangle| \leq \sqrt{2\gamma^2\lambda_{\max}}\frac{\|\mathbf{C}\|}{\|\mathbf{B}\|}\right) =& \mathrm{P}\left(|\langle P(\mathbf{m}_1 - \mathbf{m}_2), \frac{\mathbf{B}}{\|\mathbf{B}\|}\rangle| \leq \sqrt{2\gamma^2\lambda_{\max}}\frac{\|\mathbf{C}\|}{\|\mathbf{B}\|}, \mathcal{E}_1 \cap \mathcal{E}_2\right) \\
&+ \mathrm{P}\left(|\langle P(\mathbf{m}_1 - \mathbf{m}_2), \frac{\mathbf{B}}{\|\mathbf{B}\|}\rangle| \leq \sqrt{2\gamma^2\lambda_{\max}}\frac{\|\mathbf{C}\|}{\|\mathbf{B}\|}, (\mathcal{E}_1 \cap \mathcal{E}_2)^c\right) \\
\leq& \mathrm{P}\left(|\langle P(\mathbf{m}_1 - \mathbf{m}_2), \frac{\mathbf{B}}{\|\mathbf{B}\|}\rangle| \leq \sqrt{\frac{2\gamma^2\lambda_{\max}(1+\tau_2)r}{(1-\tau_1)p}}\ \right) \\
&+ \mathrm{P}(\mathcal{E}_1^c) + \mathrm{P}(\mathcal{E}_2^c). \qquad (25)
\end{aligned}$$

But, from Lemma 2 in [2],

$$\mathrm{P}(\mathcal{E}_1^c) \leq \mathrm{e}^{\frac{p}{2}(\tau_1 + \log(1-\tau_1))}, \qquad (26)$$

and

$$\mathrm{P}(\mathcal{E}_2^c) \leq \mathrm{e}^{-\frac{r}{2}(\tau_2 - \log(1+\tau_2))}.$$

Also, note that since P is an orthogonal matrix, $\|P(\mathbf{m}_1 - \mathbf{m}_2)\| = \|\mathbf{m}_1 - \mathbf{m}_2\|$. Therefore, the desired result follows by comparing $\mathrm{P}(|\langle P(\mathbf{m}_1 - \mathbf{m}_2), \frac{\mathbf{B}}{\|\mathbf{B}\|}\rangle| \leq \sqrt{\frac{2\gamma^2\lambda_{\max}(1+\tau_2)r}{(1-\tau_1)p}})$ *with* (2) *and using the result of Theorem 1.*

Corollary 5 *Consider two c-separated Gaussian distributions in $\mathbb{R}^p$ with means $\mathbf{m}_1, \mathbf{m}_2 \in \mathbb{R}^p$ and covariance matrices Σ_1 and Σ_2. Let $\beta \triangleq \frac{2\gamma^2\lambda_{\max}p}{\|\mathbf{m}_1 - \mathbf{m}_2\|^2}$, where $\lambda_{\max}$ denotes the maximal eigenvalue of the matrix $\Sigma_1 + \Sigma_2$. Then as* $\lim_{p\to\infty} d(\gamma) \leq \frac{1}{2Q(\sqrt{\beta})}$.

Corollary 6 *Consider the same setup as in Corollary 5. If γ is such that $\sqrt{\beta} = (\ln \ln p)^{\frac{1-\eta}{2}}$, where $\eta > 0$ is a free parameter, then $d(\gamma) = o(\ln p)$.*

Theorem 5 (Theorem 3.10 in [1]) *Consider a mixture of two Gaussian distribution $w\mathcal{N}(\mu_1, \sigma_1) + (1-w)\mathcal{N}(\mu_2, \sigma_2)$. Let $\sigma^2 = w(1-w)(\mu_1-\mu_2)^2 + w\sigma_1^2 + (1-w)\sigma_2^2$ denote the variance of this distribution. Then, given $n = O(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$ samples, Algorithm 3.3, with probability $1-\delta$, returns estimates of the parameters as $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_1, \hat{w})$, which under the right permutation of the indices, satisfy the following guarantees, for $i = 1, 2$,*

- *If $n \geq \left(\frac{\sigma^2}{|\mu_1-\mu_2|^2}\right)^6$, then $|\mu_i - \hat{\mu}_i| \leq \epsilon|\mu_1 - \mu_2|$, $|\sigma_i^2 - \hat{\sigma}_i^2| \leq \epsilon|\mu_1-\mu_2|^2$, and $|w - \hat{w}| \leq \epsilon$.*
- *If $n \geq \left(\frac{\sigma^2}{|\sigma_1^2-\sigma_2^2|}\right)^6$, then $|\sigma_i^2 - \hat{\sigma}_i^2| \leq \epsilon|\sigma_1^2 - \sigma_2^2| + |\mu_1-\mu_2|^2$, and $|w - \hat{w}| \leq \epsilon + \frac{|\mu_1-\mu_2|^2}{|\sigma_1^2-\sigma_2^2|}$.*
- *For any $n \geq 1$, the algorithm performs as well as assuming the mixture is a single Gaussian, and $|\mu_i - \hat{\mu}_i| \leq |\mu_1 - \mu_2| + \epsilon\sigma$, and $|\sigma_i^2 - \hat{\sigma}_i^2| \leq |\mu_1-\mu_2|^2 + |\sigma_1^2 - \sigma_2^2| + \epsilon\sigma^2$.*

Corollary 7 *Let $(X_1, \ldots, X_n)$ denote n i.i.d. samples of a mixture of two c-separated Gaussians $w\mathcal{N}(\mu_1, \sigma_1) + (1-w)\mathcal{N}(\mu_2, \sigma_2)$, where $\mu_1 < \mu_2$ and $\sigma_1 = \sigma_2$. Further assume that the separation $c = |\mu_1 - \mu_2|/(\sigma_1+\sigma_2)$ in 1-dimension is larger than $\gamma_{\min}$. Let $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, \hat{w})$ denote the estimates of $(\mu_1, \mu_2, \sigma_1, \sigma_2, w)$ returned by Algorithm 3.3 of [1]. Then, if $n = O(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$ and $n \geq \frac{1}{(2\gamma_{\min})^{12}}$, then $|\mu_i - \hat{\mu}_i| \leq \epsilon|\mu_1-\mu_2|$, $|\sigma_i^2 - \hat{\sigma}_i^2| \leq \epsilon|\mu_1-\mu_2|^2$, and $|w-\hat{w}| \leq \epsilon$.*

Theorem 6 *Consider $(X_1, \ldots, X_n)$ that are generated i.i.d. according to a mixture of two γ-separation Gaussians $w\mathcal{N}(\mu_1, \sigma_1) + (1-w)\mathcal{N}(\mu_2, \sigma_2)$, where $\sigma_1 = \sigma_2$, $w \in [w_{\min}, 0.5]$, $\mu_1 < \mu_2$ and $\gamma \in [\gamma_{\min}, \gamma_{\max}]$. Let $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, \hat{w})$ denote the estimate of the unknown parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2, w)$. Let $e_{\rm opt}$ and $\hat{e}$ denote the minimum achievable classification error and the achieved clustering error based on the estimated parameters, respectively. Then, if $|\mu_i - \hat{\mu}_i| \leq \epsilon|\mu_1-\mu_2|$, $|\sigma_i^2 - \hat{\sigma}_i^2| \leq \epsilon|\mu_1-\mu_2|^2$, $|w - \hat{w}| \leq \epsilon$, and*

$$(16\gamma_{\max}^2 + 8\gamma_{\max}\ln\frac{1-w_{\min}}{w_{\min}} + 2\gamma_{\max}\epsilon)\epsilon < \frac{1}{2},$$

we have

$$|\hat{e} - e_{\rm opt}| \leq \left(2\gamma + \frac{1}{w_{\min}\gamma} + \left(\frac{1}{\gamma} + 2\gamma\right)\ln\frac{1-w_{\min}}{w_{\min}} + \frac{8\gamma_{\max}^2}{\gamma} + 2\gamma\left(4\gamma + 2\ln\frac{1-w_{\min}}{w_{\min}}\right)^2\right)\epsilon + Q\left(\frac{1}{4\gamma\epsilon} + \epsilon_1\right) + \epsilon_2,$$

where $\epsilon_1 = o(1/\epsilon)$ and $\epsilon_2 = o(\epsilon)$.

Proof 10 *As shown in the proof of Lemma 3, the optimal Bayesian classifier breaks the real line at $t_{\rm opt} = \frac{\mu_1+\mu_2}{2} - \frac{\sigma^2}{(\mu_1-\mu_2)}\ln\frac{w}{1-w}$, and achieves a classification error equal to*

$$\begin{aligned} e_{\rm opt} &= wQ\left(\frac{t_{\rm opt}-\mu_1}{\sigma_1}\right) + (1-w)Q\left(\frac{\mu_2 - t_{\rm opt}}{\sigma_2}\right) \\ &= wQ\left(\gamma + \frac{1}{2\gamma}\ln\frac{w}{1-w}\right) + (1-w)Q\left(\gamma - \frac{1}{2\gamma}\ln\frac{w}{1-w}\right). \end{aligned}$$

On the other hand, without having access to the exact parameters, a clustering algorithm that operates based on the estimated values $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_1, \hat{w})$ finds $\hat{t}_1$ and $\hat{t}_2$, which are the solutions of $\frac{\hat{w}}{\sqrt{2\pi\hat{\sigma}_1^2}}{\rm e}^{\frac{-(t-\hat{\mu}_1)^2}{2\hat{\sigma}_1^2}} = \frac{1-\hat{w}}{\sqrt{2\pi\hat{\sigma}_2^2}}{\rm e}^{\frac{-(t-\hat{\mu}_2)^2}{2\hat{\sigma}_2^2}}$, and puts the decision boundary points at these two points. For $i = 1, 2$, let

$$\tilde{t}_i \triangleq \hat{t}_i - \hat{\mu}_1.$$

and

$$(s_i, \hat{s}_i) \triangleq \left(\frac{1}{\sigma_i^2}, \frac{1}{\hat{\sigma}_i^2}\right).$$

Note that $\sigma_1 = \sigma_2$ by assumptions. Therefore $s_1 = s_2$. Let

$$s \triangleq s_1 = s_2,$$

and

$$\hat{\delta}_\mu \triangleq \hat{\mu}_2 - \hat{\mu}_1.$$

Using the mentioned change of variable, $(\tilde{t}_1, \tilde{t}_2)$ *are the solutions of the following second order equation*

$$(\hat{s}_1 - \hat{s}_2)x^2 + 2\hat{\delta}_\mu \hat{s}_2 x - \hat{\delta}_\mu^2 \hat{s}_2 + 2\ln\frac{\hat{s}_2}{\hat{s}_1} - 2\ln\frac{\hat{w}}{1-\hat{w}} = 0, \tag{27}$$

Assume that $\tilde{t}_1$ *denotes the point that approximates* $t_{\rm opt} - \mu_1$*. A clustering algorithm that decides based on these estimated boundary points estimates its achieved error as* $\hat{e}_{\rm opt}$*, where, if* $\hat{\sigma}_1 \leq \hat{\sigma}_2$*,*

$$\hat{e}_{\rm opt} = \hat{w}\left(Q\left(\frac{\tilde{t}_1}{\hat{\sigma}_1}\right) + Q\left(-\frac{\tilde{t}_2}{\hat{\sigma}_1}\right)\right) + (1-\hat{w})\left(Q\left(\frac{\hat{\delta}_\mu - \tilde{t}_1}{\hat{\sigma}_2}\right) - Q\left(\frac{\hat{\delta}_\mu - \tilde{t}_2}{\hat{\sigma}_2}\right)\right),$$

and if $\hat{\sigma}_1 > \hat{\sigma}_2$*,*

$$\hat{e}_{\rm opt} = \hat{w}\left(Q\left(\frac{\tilde{t}_1}{\hat{\sigma}_1}\right) - Q\left(\frac{\tilde{t}_2}{\hat{\sigma}_1}\right)\right) + (1-\hat{w})\left(Q\left(\frac{\hat{\delta}_\mu - \tilde{t}_1}{\hat{\sigma}_2}\right) + Q\left(\frac{\tilde{t}_2 - \hat{\delta}_\mu}{\hat{\sigma}_2}\right)\right).$$

Since for all x *and* x'*,* $|Q(x) - Q(x')| \leq |x - x'|$*, if* $\hat{\sigma}_1 \leq \hat{\sigma}_2$*,*

$$|e_{\rm opt} - \hat{e}_{\rm opt}| \leq |w - \hat{w}| + \left|\frac{\tilde{t}_1}{\hat{\sigma}_1} - \frac{t_{\rm opt} - \mu_1}{\sigma_1}\right| + \left|\frac{\hat{\delta}_\mu - \tilde{t}_1}{\hat{\sigma}_2} - \frac{\mu_2 - t_{\rm opt}}{\sigma_2}\right| + Q\left(-\frac{\tilde{t}_2}{\hat{\sigma}_1}\right),$$

and if $\hat{\sigma}_1 > \hat{\sigma}_2$*,*

$$|e_{\rm opt} - \hat{e}_{\rm opt}| \leq |w - \hat{w}| + \left|\frac{\tilde{t}_1}{\hat{\sigma}_1} - \frac{t_{\rm opt} - \mu_1}{\sigma_1}\right| + \left|\frac{\hat{\delta}_\mu - \tilde{t}_1}{\hat{\sigma}_2} - \frac{\mu_2 - t_{\rm opt}}{\sigma_2}\right| + Q\left(\frac{\tilde{t}_2 - \hat{\delta}_\mu}{\hat{\sigma}_2}\right).$$

Note that, by the triangle inequality,

$$\left|\frac{\tilde{t}_1}{\hat{\sigma}_1} - \frac{t_{\rm opt} - \mu_1}{\sigma_1}\right| \leq \frac{|\tilde{t}_1 - t_{\rm opt} + \mu_1|}{\hat{\sigma}_1} + |t_{\rm opt} - \mu_1|\left|\frac{1}{\hat{\sigma}_1} - \frac{1}{\sigma_1}\right|. \tag{28}$$

Similarly,

$$\left|\frac{\hat{\delta}_\mu - \tilde{t}_1}{\hat{\sigma}_2} - \frac{\mu_2 - t_{\rm opt}}{\sigma_2}\right| \leq \frac{|\hat{\delta}_\mu - \tilde{t}_1 - \mu_2 + t_{\rm opt}|}{\hat{\sigma}_2} + |t_{\rm opt} - \mu_2|\left|\frac{1}{\hat{\sigma}_2} - \frac{1}{\sigma_2}\right|. \tag{29}$$

But, by assumption, $|\sigma_i^2 - \hat{\sigma}_i^2| \leq \epsilon\delta_\mu^2$*. Therefore,* $\hat{\sigma}_i \leq \sigma_i\sqrt{1 + \epsilon\delta_\mu^2/\sigma_i^2} = \sigma_i\sqrt{1 + 4\gamma^2\epsilon} \leq \sigma_i(1 + 2\gamma^2\epsilon)$*. Similarly,* $\hat{\sigma}_i \geq \sigma_i\sqrt{1 - 4\gamma^2\epsilon} \geq \sigma_i(1 - 4\gamma^2\epsilon)$*. Hence,* $|\sigma_i - \hat{\sigma}_i| \leq 4c^2\epsilon$ *and*

$$\left|\frac{1}{\hat{\sigma}_i} - \frac{1}{\sigma_i}\right| \leq \frac{4\gamma^2\epsilon}{(1 - 4\gamma^2\epsilon)\sigma_i}.$$

Also, note that since $t_{\rm opt} = \frac{\mu_1 + \mu_2}{2} - \frac{\sigma_1^2}{(\mu_1 - \mu_2)}\ln\frac{w}{1-w}$*, for* $i = 1, 2$*,*

$$\frac{|t_{\rm opt} - \mu_i|}{\sigma_1} \leq \gamma + \frac{1}{2\gamma}\ln\frac{1 - w_{\min}}{w_{\min}}.$$

In summary, if $\hat{\sigma}_1 \leq \hat{\sigma}_2$*,*

$$|e_{\rm opt} - \hat{e}_{\rm opt}| \leq \left(1 + 4\gamma^3 + 2\gamma\ln\frac{1 - w_{\min}}{w_{\min}}\right)\epsilon + \frac{1}{\sigma_1}|\tilde{t}_1 - t_{\rm opt} + \mu_1| + Q\left(-\frac{\tilde{t}_2}{\hat{\sigma}_1}\right) + o(\epsilon), \tag{30}$$

and if $\hat{\sigma}_1 > \hat{\sigma}_2$*,*

$$|e_{\rm opt} - \hat{e}_{\rm opt}| \leq \left(1 + 4\gamma^3 + 2\gamma\ln\frac{1 - w_{\min}}{w_{\min}}\right)\epsilon + \frac{1}{\sigma_1}|\tilde{t}_1 - t_{\rm opt} + \mu_1| + Q\left(\frac{\tilde{t}_2 - \hat{\delta}_\mu}{\hat{\sigma}_2}\right) + o(\epsilon). \tag{31}$$

In the rest of the proof, we mainly focus on bounding $|\tilde{t}_1 - t_{\rm opt} + \mu_1|$. Since $\tilde{t}_1$ and $\tilde{t}_2$ are the solutions of (27), they can be computed as

$$\tilde{t}_1, \tilde{t}_2 = \frac{-\hat{\delta}_\mu \hat{s}_2 \pm \sqrt{\Delta}}{(\hat{s}_1 - \hat{s}_2)},$$

where

$$\Delta = (\hat{\delta}_\mu \hat{s}_2)^2 - (\hat{s}_1 - \hat{s}_2)\left(-\hat{\delta}_\mu^2 \hat{s}_2 + 2\ln\frac{\hat{s}_2}{\hat{s}_1} - 2\ln\frac{\hat{w}}{1-\hat{w}}\right).$$

Define υ as

$$\upsilon \triangleq \hat{\mu}_2 - \hat{\mu}_1 - (\mu_2 - \mu_1). \tag{32}$$

Note that since by assumption $|\hat{\mu}_i - \mu_i| \leq \epsilon\delta_\mu$, where

$$\delta_\mu \triangleq |\mu_2 - \mu_1|,$$

we have

$$|\upsilon| \leq 2\delta_\mu \epsilon,$$

Define τ_1 and τ_2 as

$$\tau_i \triangleq \hat{s}_i - s.$$

Note that

$$|\tau_i| \leq \frac{\delta_\mu^2 \epsilon}{\hat{\sigma}_i^2 \sigma_i^2} \leq \frac{\delta_\mu^2 \epsilon}{\sigma_i^2(\sigma_i^2 - \epsilon\delta_\mu^2)} = \frac{4\gamma^2 s\epsilon}{1 - 4\epsilon\gamma^2} \leq \frac{4\gamma^2 s\epsilon}{1 - 4\epsilon\gamma_{\max}^2} \leq 8\gamma^2 s\epsilon, \tag{33}$$

where the last inequality holds as long as $4\gamma_{\max}^2 \epsilon \leq \frac{1}{2}$.

Define ε as

$$\varepsilon \triangleq \frac{\hat{s}_2 - \hat{s}_1}{(\hat{\delta}_\mu \hat{s}_2)^2}\left(-\hat{\delta}_\mu^2 \hat{s}_2 + 2\ln\frac{\hat{s}_2}{\hat{s}_1} - 2\ln\frac{\hat{w}}{1-\hat{w}}\right). \tag{34}$$

Then, using this definition, it follows from (27) that

$$\tilde{t}_1 = \frac{\hat{\delta}_\mu \hat{s}_2}{\hat{s}_2 - \hat{s}_1}\left(1 - \sqrt{1+\varepsilon}\,\right), \tag{35}$$

and

$$\tilde{t}_2 = \frac{\hat{\delta}_\mu \hat{s}_2}{\hat{s}_2 - \hat{s}_1}\left(1 + \sqrt{1+\varepsilon}\,\right). \tag{36}$$

Define function f as $f(x) = \sqrt{1+x}$. Then, using the Taylor expansion of function f around zero,

$$f(\varepsilon) = 1 + \frac{1}{2}\varepsilon + \frac{f''(r)}{2}\varepsilon^2, \tag{37}$$

where $|r| \leq |\varepsilon|$. Note that

$$\begin{aligned}\frac{\hat{\delta}_\mu \hat{s}_2}{\hat{s}_2 - \hat{s}_1}\varepsilon &= \frac{1}{\hat{\delta}_\mu \hat{s}_2}\left(-\hat{\delta}_\mu^2 \hat{s}_2 + 2\ln\frac{\hat{s}_2}{\hat{s}_1} - 2\ln\frac{\hat{w}}{1-\hat{w}}\right). \\ &= -\hat{\delta}_\mu + \frac{2}{\hat{\delta}_\mu \hat{s}_2}\left(\ln\frac{\hat{s}_2}{\hat{s}_1} - \ln\frac{\hat{w}}{1-\hat{w}}\right).\end{aligned} \tag{38}$$

Therefore, we have

$$\tilde{t}_1 = \frac{\hat{\delta}_\mu}{2} + \frac{1}{\hat{\delta}_\mu \hat{s}_2}\left(\ln\frac{\hat{w}}{1-\hat{w}} - \ln\frac{\hat{s}_2}{\hat{s}_1}\right) - \frac{f''(r)}{2}\left(\frac{\hat{\delta}_\mu \hat{s}_2}{\hat{s}_2 - \hat{s}_1}\right)\varepsilon^2. \tag{39}$$

and

$$\tilde{t}_2 = \frac{2\hat{\delta}_\mu \hat{s}_2}{\hat{s}_2 - \hat{s}_1} - \tilde{t}_1. \tag{40}$$

As a reminder $t_{\text{opt}} = \frac{\mu_1+\mu_2}{2} - \frac{1}{(\mu_1-\mu_2)s}\ln\frac{w}{1-w}$. *Therefore, from* (39), *we have*

$$\begin{aligned}\left|\tilde{t}_1 - t_{\text{opt}} + \mu_1\right| &= \left|\frac{\hat{\delta}_\mu}{2} + \frac{1}{\hat{\delta}_\mu \hat{s}_2}\left(\ln\frac{\hat{w}}{1-\hat{w}} - \ln\frac{\hat{s}_2}{\hat{s}_1}\right) - \frac{f''(r)}{2}\left(\frac{\hat{\delta}_\mu \hat{s}_2}{\hat{s}_2-\hat{s}_1}\right)\varepsilon^2 - t_{\text{opt}} + \mu_1\right| \\ &\le \delta_\mu\epsilon + \left|\frac{1}{\hat{\delta}_\mu\hat{s}_2}\ln\frac{\hat{w}}{1-\hat{w}} - \frac{1}{s(\mu_2-\mu_1)}\ln\frac{w}{1-w}\right| + \left|\frac{1}{\hat{\delta}_\mu\hat{s}_2}\ln\frac{\hat{s}_2}{\hat{s}_1}\right| + \left|\frac{f''(r)}{2}\left(\frac{\hat{\delta}_\mu\hat{s}_2}{\hat{s}_2-\hat{s}_1}\right)\varepsilon^2\right|. \end{aligned}\tag{41}$$

We next bound the error terms in (41). *Note that, by the triangle inequality,*

$$\begin{aligned}\left|\frac{1}{\hat{\delta}_\mu\hat{s}_2}\ln\frac{\hat{w}}{1-\hat{w}} - \frac{1}{s(\mu_2-\mu_1)}\ln\frac{w}{1-w}\right| &= \left|\frac{1}{\hat{\delta}_\mu\hat{s}_2}\left(\ln\frac{\hat{w}}{1-\hat{w}} - \ln\frac{w}{1-w} + \ln\frac{w}{1-w}\right) - \frac{1}{s(\mu_2-\mu_1)}\ln\frac{w}{1-w}\right| \\ &\le \frac{1}{|\hat{\delta}_\mu|\hat{s}_2}\left|\ln\frac{\hat{w}}{1-\hat{w}} - \ln\frac{w}{1-w}\right| + \left|\ln\frac{w}{1-w}\right|\left|\frac{1}{\hat{\delta}_\mu\hat{s}_2} - \frac{1}{s(\mu_2-\mu_1)}\right|\end{aligned}\tag{42}$$

Since $|\mu_i - \hat{\mu}_i| \le \epsilon\delta_\mu$, $|\hat{\delta}_\mu| = |\hat{\mu}_1 - \hat{\mu}_2| \ge \delta_\mu(1-2\epsilon)$. *Therefore, we have*

$$\frac{1}{|\hat{\delta}_\mu|} \le \frac{1}{\delta_\mu(1-2\epsilon)}. \tag{43}$$

Let $g(w) = \ln\frac{w}{1-w}$. *Then,* $g'(w) = \frac{1}{w} + \frac{1}{1-w}$. *Therefore, since by assumption,* $|w-\hat{w}| \le \epsilon$, *we have*

$$\left|\ln\frac{\hat{w}}{1-\hat{w}} - \ln\frac{w}{1-w}\right| \le \left(\frac{1}{w_{\min}} + \frac{1}{1-w_{\min}}\right)\epsilon \le \frac{2\epsilon}{w_{\min}}. \tag{44}$$

Note that since $w \in (w_{\min}, 0.5)$, *and since* $\frac{w}{1-w}$ *is an increasing function of* w *in this interval, we have*

$$|\ln\frac{w}{1-w}| \le \ln\frac{1-w_{\min}}{w_{\min}}. \tag{45}$$

Note that

$$\ln\frac{\hat{s}_2}{\hat{s}_1} = \ln\frac{s+\tau_1}{s+\tau_2} = \ln\frac{1+\tau_1/s}{1+\tau_2/s}.$$

Hence,

$$\left|\ln\frac{\hat{s}_2}{\hat{s}_1}\right| \le \ln\frac{1+|\frac{\tau_1}{s}|}{1-|\frac{\tau_2}{s}|} \le \ln\frac{1+\frac{4\epsilon\gamma^2}{1-4\epsilon\gamma^2_{\max}}}{1-\frac{4\epsilon\gamma^2}{1-4\epsilon\gamma^2_{\max}}} = \ln\frac{1}{1-8\epsilon\gamma^2_{\max}} \le \frac{8\gamma^2_{\max}\epsilon}{1-8\epsilon\gamma^2_{\max}} \le 16\gamma^2_{\max}\epsilon, \tag{46}$$

where the last line holds if $8\gamma^2_{\max}\epsilon < \frac{1}{2}$. *Combining* (42), (43), (44), (45) *and* (46) *with* (41), *it follows that*

$$\begin{aligned}\left|\tilde{t}_1 - t_{\text{opt}} + \mu_1\right| \le&\, \delta_\mu\epsilon + \frac{2\epsilon}{(1-2\epsilon)(1-8\gamma^2 s\epsilon)w_{\min}\delta_\mu s} + \frac{1}{\hat{s}_2}\ln\frac{1-w_{\min}}{w_{\min}}\left|\frac{1}{\hat{\delta}_\mu} - \frac{1}{(\mu_2-\mu_1)}\right| \\ &+ \frac{1}{\delta_\mu}\ln\frac{1-w_{\min}}{w_{\min}}\left|\frac{1}{\hat{s}_2} - \frac{1}{s}\right| + \frac{16\gamma^2_{\max}\epsilon}{(1-2\epsilon)(1-8\gamma^2 s\epsilon)\delta_\mu s} + \left|\frac{f''(r)}{2}\left(\frac{\hat{\delta}_\mu\hat{s}_2}{\hat{s}_2-\hat{s}_1}\right)\varepsilon^2\right| \\ \le&\, \delta_\mu\epsilon + \frac{2\epsilon}{(1-2\epsilon)(1-8\gamma^2 s\epsilon)w_{\min}\delta_\mu s} + \left(\frac{2\epsilon}{(1-2\epsilon)(1-8\gamma^2\epsilon)\delta_\mu s} + \epsilon\delta_\mu\right)\ln\frac{1-w_{\min}}{w_{\min}} \\ &+ \frac{16\gamma^2_{\max}\epsilon}{(1-2\epsilon)(1-8\gamma^2 s\epsilon)\delta_\mu s} + \left|\frac{f''(r)}{2}\left(\frac{\hat{\delta}_\mu\hat{s}_2}{\hat{s}_2-\hat{s}_1}\right)\varepsilon^2\right| \\ =&\left(\delta_\mu + \frac{2}{w_{\min}\delta_\mu s} + \left(\frac{2}{\delta_\mu s} + \delta_\mu\right)\ln\frac{1-w_{\min}}{w_{\min}} + \frac{16\gamma^2_{\max}}{\delta_\mu s}\right)\epsilon + \frac{f''(r)}{2}\left|\frac{\hat{\delta}_\mu\hat{s}_2}{\hat{s}_2-\hat{s}_1}\right|\varepsilon^2 + o(\epsilon).\end{aligned}\tag{47}$$

Finally, we need to bound ε, defined in (34). By the triangle inequality and (43), it follows that.

$$
\begin{aligned}
|\varepsilon| \leq & \frac{|\hat{s}_2-\hat{s}_1|}{\delta_\mu^2(1-2\epsilon)^2(\hat{s}_2)^2}\left(\delta_\mu^2(1+2\epsilon)^2\hat{s}_2+2\left|\ln\frac{\hat{s}_2}{\hat{s}_1}\right|+2\left|\ln\frac{\hat{w}}{1-\hat{w}}\right|\right)\\
= & \frac{|\hat{s}_2-\hat{s}_1|}{\delta_\mu^2(1-2\epsilon)^2(\hat{s}_2)^2}\left(\delta_\mu^2(1+2\epsilon)^2\hat{s}_2+2\left|\ln\frac{\hat{s}_2}{\hat{s}_1}\right|+2\left|\ln\frac{\hat{w}}{1-\hat{w}}-\ln\frac{w}{1-w}+\ln\frac{w}{1-w}\right|\right)\\
\overset{(a)}{\leq} & \frac{|\hat{s}_2-\hat{s}_1|}{\delta_\mu^2(1-2\epsilon)^2(1-8\gamma^2\epsilon)^2 s^2}\left(\delta_\mu^2(1+2\epsilon)^2(1+8\gamma^2\epsilon)s+32\gamma_{\max}^2\epsilon+\frac{4\epsilon}{w_{\min}}+2\ln\frac{1-w_{\min}}{w_{\min}}\right)\\
\overset{(b)}{=} & \frac{|\hat{s}_2-\hat{s}_1|}{s}\left(1+\frac{1}{2\gamma}\ln\frac{1-w_{\min}}{w_{\min}}+O(\epsilon)\right),
\end{aligned}
\tag{48}
$$

where (a) follows from (33), (44), (45) and (46), and (b) holds because $\delta_\mu^2 s=4\gamma$. Also, note that, from (33),

$$
|\varepsilon| \leq 8\gamma\left(2\gamma+\ln\frac{1-w_{\min}}{w_{\min}}+O(\epsilon)\right)\epsilon. \tag{49}
$$

Therefore, if $(16\gamma_{\max}^2+8\gamma_{\max}\ln\frac{1-w_{\min})}{w_{\min}}+2\gamma_{\max}\epsilon)\epsilon<\frac{1}{2}$,

$$
|\varepsilon| \leq \frac{1}{2},
$$

and $|f''(r)|=\frac{1}{4}(1+r)^{-\frac{3}{2}}\leq\frac{1}{\sqrt{2}}<1$. Combining (48) with (47), it follows that

$$
\begin{aligned}
\left|\tilde{t}_1-t_{\text{opt}}+\mu_1\right| \leq & \left(\delta_\mu+\frac{2}{w_{\min}\delta_\mu s}+\left(\frac{2}{\delta_\mu s}+\delta_\mu\right)\ln\frac{1-w_{\min}}{w_{\min}}+\frac{16\gamma_{\max}^2}{\delta_\mu s}\right)\epsilon\\
& +|f''(r)||\hat{\delta}_\mu\hat{s}_2||\hat{s}_2-\hat{s}_1|\left(1+\frac{1}{2\gamma}\ln\frac{1-w_{\min}}{w_{\min}}+O(\epsilon)\right)^2+o(\epsilon)\\
\leq & \left(\delta_\mu+\frac{2}{w_{\min}\delta_\mu s}+\left(\frac{2}{\delta_\mu s}+\delta_\mu\right)\ln\frac{1-w_{\min}}{w_{\min}}+\frac{16\gamma_{\max}^2}{\delta_\mu s}\right)\epsilon\\
& +|f''(r)||\delta_\mu s|(1+\epsilon)(16\gamma^2\epsilon)\left(1+\frac{1}{2\gamma}\ln\frac{1-w_{\min}}{w_{\min}}+O(\epsilon)\right)^2+o(\epsilon)\\
\leq & \left(\delta_\mu+\frac{2}{w_{\min}\delta_\mu s}+\left(\frac{2}{\delta_\mu s}+\delta_\mu\right)\ln\frac{1-w_{\min}}{w_{\min}}+\frac{16\gamma_{\max}^2}{\delta_\mu s}\right.\\
& \left.+|f''(r)||\delta_\mu s|\left(4\gamma+2\ln\frac{1-w_{\min}}{w_{\min}}\right)^2\right)\epsilon+o(\epsilon).
\end{aligned}
\tag{50}
$$

Dividing both sides of (50) by σ_1, and noting that $\delta_\mu/(2\sigma_1)=\gamma$ and $|f''(r)|\leq 1$, we derive

$$
\begin{aligned}
\frac{\left|\tilde{t}_1-t_{\text{opt}}+\mu_1\right|}{\sigma_1} \leq & \left(2\gamma+\frac{1}{w_{\min}\gamma}+\left(\frac{1}{\gamma}+2\gamma\right)\ln\frac{1-w_{\min}}{w_{\min}}+\frac{8\gamma_{\max}^2}{\gamma}\right.\\
& \left.+2\gamma\left(4\gamma+2\ln\frac{1-w_{\min}}{w_{\min}}\right)^2\right)\epsilon+o(\epsilon).
\end{aligned}
\tag{51}
$$

Finally, as a reminder, from (40), $\tilde{t}_2=\frac{2\hat{\delta}_\mu\hat{s}_2}{\hat{s}_2-\hat{s}_1}-\tilde{t}_1$. From (33), $|\hat{s}_2-\hat{s}_1|\leq 16\gamma^2 s\epsilon$. Hence, if $\hat{\sigma}_1\leq\hat{\sigma}_2$,

$$
-\frac{\tilde{t}_2}{\hat{\sigma}_1}=\frac{1}{\hat{\sigma}_1}\left(\frac{2\hat{\delta}_\mu\hat{s}_2}{\hat{s}_1-\hat{s}_2}+\tilde{t}_1\right)\geq\frac{1}{4\gamma\epsilon}+o\left(\frac{1}{\epsilon}\right).
$$

Similarly, if $\hat{\sigma}_2\leq\hat{\sigma}_1$,

$$
\frac{\tilde{t}_2-\hat{\delta}_\mu}{\hat{\sigma}_2}=\frac{1}{\hat{\sigma}_2}\left(\frac{2\hat{\delta}_\mu\hat{s}_2}{\hat{s}_2-\hat{s}_1}-\tilde{t}_1-\hat{\delta}_\mu\right)\geq\frac{1}{4\gamma\epsilon}+o\left(\frac{1}{\epsilon}\right).
$$

Combining (51) and the above equations with (30) and (31) yields the desired result.

Lemma 3 *Consider i.i.d. points generated as* $w\mathcal{N}(\mu_1,\sigma)+(1-w)\mathcal{N}(\mu_2,\sigma)$. *Without loss of generality, assume that* $\mu_1 \leq \mu_2$ *and* $w < 0.5$. *Let* $\gamma = (\mu_2-\mu_1)/(2\sigma)$. *Also, let* $e_{\rm opt}$ *denote the error probability of an optimal Bayesian classifier. Then, if* $w \leq 0.1$,

$$e_{\rm opt} \geq wQ\left(-\frac{1}{\gamma}+\gamma\right). \tag{52}$$

For $w \in (0.1, 0.5]$,

$$e_{\rm opt} \geq wQ\left(\gamma\right). \tag{53}$$

Proof 11 *The optimal* Bayesian *classifier, which has access to the parameters* (μ_1,μ_2,σ,w), *divides the real line at*

$$t_{\rm opt} = \frac{\mu_1+\mu_2}{2} - \frac{\sigma^2}{(\mu_1-\mu_2)}\ln\frac{w}{1-w}, \tag{54}$$

and achieves a classification error equal to

$$\begin{aligned} e_{\rm opt} &= w\,{\rm P}(\mu_1+\sigma_1 Z \geq t_{\rm opt}) + (1-w)\,{\rm P}(\mu_2+\sigma_2 Z \leq t_{\rm opt}) \\ &= wQ\left(\frac{\mu_2-\mu_1}{2\sigma} - \frac{\sigma}{(\mu_1-\mu_2)}\ln\frac{w}{1-w}\right) + (1-w)Q\left(\frac{\mu_2-\mu_1}{2\sigma} + \frac{\sigma}{(\mu_1-\mu_2)}\ln\frac{w}{1-w}\right) \\ &= wQ\left(\gamma + \frac{1}{2\gamma}\ln\frac{w}{1-w}\right) + (1-w)Q\left(\gamma - \frac{1}{2\gamma}\ln\frac{w}{1-w}\right), \end{aligned} \tag{55}$$

where $Z \sim \mathcal{N}(0,1)$. *Note that since by assumption* $w < 1-w$, $\ln\frac{w}{1-w} \leq 0$. *Therefore,*

$$Q\left(\gamma - \frac{1}{2\gamma}\ln\frac{w}{1-w}\right) \leq Q\left(\gamma + \frac{1}{2\gamma}\ln\frac{w}{1-w}\right).$$

Keeping the larger Q *term, it follows from* (55) *that*

$$e_{\rm opt} \geq wQ\left(-\frac{1}{2\gamma}\ln\frac{1-w}{w}+\gamma\right). \tag{56}$$

For $w \leq 0.1$, $0.5\ln\frac{1-w}{w} \geq 0.5\ln\frac{1-0.1}{0.1} > 1$. *Therefore, since* $Q(\cdot)$ *is a monotonically decreasing function of its argument,* (52) *follows. The result for* $w \in (0.1, 0.5)$ *stated in* (52) *follows by noting that* $-\frac{1}{2\gamma}\ln\frac{1-w}{w}+\gamma \leq \gamma$.

Lemma 4 *Let* $(X_1,\ldots,X_n)$ *denote* n *i.i.d. samples of a mixture of two* γ*-separated Gaussians* $w\mathcal{N}(\mu_1,\sigma_1)+(1-w)\mathcal{N}(\mu_2,\sigma_2)$, *where* $\sigma_1=\sigma_2$, $\gamma=(\mu_2-\mu_1)/(\sigma_1+\sigma_2)<1/2$ *and* $\mu_1<\mu_2$. *Let* $(\hat{\mu}_1,\hat{\mu}_2,\hat{\sigma}_1,\hat{\sigma}_2,\hat{w})$ *denote the estimates of* $(\mu_1,\mu_2,\sigma,\sigma,w)$ *returned by Algorithm 3.3 of [1]. Then, if* $n = O(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$, *with probability larger than* $1-\delta$,

$$\frac{|\hat{\mu}_1-\hat{\mu}_2|}{\hat{\sigma}_1+\hat{\sigma}_2} \leq \frac{3\gamma+\epsilon}{1-2\sqrt{\gamma^2+\epsilon}}.$$

Proof 12 *By Theorem 5, for* $n = O(\epsilon^2\log\frac{1}{\delta})$, *with probability* $1-\delta$, *there exists a permutations of indices, such that* $|\mu_i-\hat{\mu}_i| \leq |\mu_1-\mu_2|+\epsilon\sigma$ *and* $|\sigma_i^2-\hat{\sigma}_i^2| \leq |\mu_1-\mu_2|^2+|\sigma_1^2-\sigma_2^2|+\epsilon\sigma^2 = |\mu_1-\mu_2|^2+\epsilon\sigma^2$. *Therefore, by the triangle inequality,*

$$|\hat{\mu}_1-\hat{\mu}_2| \leq \sum_{i=1}^{2}|\mu_i-\hat{\mu}_i| + |\mu_1-\mu_2| \leq 3|\mu_1-\mu_2|+\epsilon\sigma.$$

Hence, since $\sigma^2 = w(1-w)(\mu_1-\mu_2)^2 + \sigma_1^2$,

$$\begin{aligned}
\frac{|\hat{\mu}_1 - \hat{\mu}_2|}{\hat{\sigma}_1 + \hat{\sigma}_2} &\leq \frac{3|\mu_1-\mu_2| + \epsilon\sigma}{2\sigma_1 - 2\sqrt{|\mu_1-\mu_2|^2 + \epsilon\sigma^2}} \\
&= \frac{3|\mu_1-\mu_2| + \epsilon\sqrt{w(1-w)(\mu_1-\mu_2)^2 + \sigma_1^2}}{2\sigma_1 - 2\sqrt{|\mu_1-\mu_2|^2 + \epsilon w(1-w)(\mu_1-\mu_2)^2 + \epsilon\sigma_1^2}} \\
&\overset{(a)}{=} \frac{3\gamma + \epsilon\sqrt{w(1-w)\gamma^2 + 0.25}}{1 - \sqrt{4\gamma^2 + 4\epsilon w(1-w)\gamma^2 + \epsilon}} \\
&\overset{(b)}{\leq} \frac{3\gamma + 0.5\epsilon\sqrt{\gamma^2+1}}{1 - \sqrt{4\gamma^2 + \epsilon(1+\gamma^2)}} \\
&\overset{(c)}{\leq} \frac{3\gamma + \epsilon}{1 - 2\sqrt{\gamma^2 + \epsilon}},
\end{aligned} \tag{57}$$

where (a) *follows by dividing the nominator and denominator by* $2\sigma_1$ *and* (b) *holds because* $w(1-w) \leq 0.25$. *Finally* (c) *holds, since by assumption* $\gamma^2 < 1$.

References

M. Hardt and E. Price. Tight bounds for learning a mixture of two gaussians. In *Proc. of the 47th Ann. ACM Sym. on Theory of Comp.*, pages 753–760. ACM, 2015.

S. Jalali and A. Maleki. Minimum complexity pursuit. In *Proc. 49th Annual Proc. Allerton Conf. Comm., Cont., and Comp.*, pages 1764–1770, Sep. 2011.