A. Omitted details from Section 2

A.1. Two intervals suffice for $\mathcal{G}$

Here we formally prove (4). In fact, we will prove a slight generalization of this fact which will be useful later on.

We require the following theorem from Hummel and Gidas:

Theorem A.1 ((Hummel & Gidas, 1984)). *Let f be any analytic function with at most n zeros. Then $f \circ \mathcal{N}(0, \sigma^2)$ has at most n zeros.*

This allows us to prove:

Theorem A.2. *Any linear combination $F(x)$ of the probability density functions of k Gaussians with the same variance has at most $k-1$ zeros, provided at least two of the Gaussians have different means. In particular, for any $\mu \neq \nu$, the function $F(x) = D_\mu(x) - D_\nu(x)$ has at most 3 zeroes.*

Proof. If we have more than 1 Gaussian with the same mean, we can replace all Gaussians having that mean with an appropriate factor times a single Gaussian with that mean. Thus, we assume without loss of generality that all Gaussians have distinct means. We may also assume without loss of generality that all Gaussians have a nonzero coefficient in the definition of F.

Suppose the minimum distance between the means of any of the Gaussians is δ. We first prove the statement when δ is sufficiently large compared to everything else. Consider any pair of Gaussians with consecutive means ν, μ. WLOG assume that $\mu > \nu = 0$. Suppose our pair of Gaussians has the same sign in the definition of F. In particular they are both strictly positive. For sufficiently large δ, we can make the contribution of the other Gaussians to F an arbitrarily small fraction of the whichever Gaussian in our pair is largest for all points on $[\nu, \mu]$. Thus, for δ sufficiently large, that there are no zeros on this interval.

Now suppose our pair of Gaussians have different signs in the definition of F. Without loss of generality, assume the sign of the Gaussian with mean ν is positive and the sign of the Gaussian with mean μ is negative. Then the PDF of the first Gaussian is strictly decreasing on $(\nu, \mu]$ and the PDF of the negation of the second Gaussian is decreasing on $[\nu, \mu)$. Thus, their sum is strictly decreasing on this interval. Similarly to before, by making δ sufficiently large, the magnitude of the contributions of the other Gaussians to the derivative in this region can be made an arbitrarily small fraction of the magnitude of whichever Gaussian in our pair contributes the most at each point in the interval. Thus, in this case, there is exactly one zero in the intervale $[\mu, \nu]$.

Also, note that there can be no zeros of F outside of the convex hull of their means. This follows by essentially the same argument as the two positive Gaussians case above.

The general case (without assuming δ sufficiently large) follows by considering sufficiently skinny (nonzero variance) Gaussians with the same means as the Gaussians in the definition of F, rescaling the domain so that they are sufficiently far apart, applying this argument to this new function, unscaling the domain (which doesn't change the number of zeros), then convolving the function with an appropriate (very fat) Gaussian to obtain the real F, and invoking Theorem A.1 to say that the number of zeros does not increase from this convolution. □

A.2. The function L

In this section, we derive the form of L. By definition, we have

$$\sqrt{2\pi}L(\widehat{\mu}, \ell, r) = \sqrt{2\pi}\left(\mathbb{E}_{x\sim G_{\mu^*}}[D(x)] + \mathbb{E}_{x\sim G_\mu}[1 - D(x)]\right)$$
$$= \sqrt{2\pi}\left(\int_I G_{\mu^*}(x) - G_{\widehat{\mu}}(x)dx\right) + \sqrt{2\pi}\,,$$

where $I = [\ell_1, r_1] \cup [\ell_2, r_2]$. We then have

$$\sqrt{2\pi}L(\widehat{\mu}, \ell, r)$$
$$= \sqrt{2\pi}\left(\sum_{i=1,2}\int_{\ell_i}^{r_i} G_{\mu^*}(x) - G_{\widehat{\mu}}(x)dx\right) + \sqrt{2\pi}$$
$$= \sum_{i=1,2}\sum_{j=1,2}\int_{\ell_i}^{r_i} e^{-(x-\mu_j^*)^2/2} - e^{-(x-\widehat{\mu}_j)^2/2}dx + \sqrt{2\pi}\,. \tag{13}$$

It is not to hard to see from the Fundamental theorem of calculus that L is indeed a smooth function of all parameters.

B. Alternative Induced Dynamics

Our focus in this paper is on the dynamics induced by, since it arises naturally from the form of the total variation distance (3) and follows the canonical form of GAN dynamics (1). However, one could consider other equivalent definitions of total variation distance too. And these definitions could, in principle, induce qualitatively different behavior of the first order dynamics.

As mentioned in Section 3, an alternative dynamics could be induced by the definition of total variation distance given in (12). The corresponding loss function would be

$$L'(\mu, \ell, r) = |L(\mu, \ell, r)| = \left|\mathbb{E}_{x\sim G_{\mu^*}}[D(x)] + \mathbb{E}_{x\sim G_\mu}[1 - D(x)]\right|\,, \tag{14}$$

i.e. the same as in (6) but with absolute values on the outside of the expression. Observe that this loss function does not actually fit the form of the general GAN dynamics presented in (1). However, it still constitutes a valid and fairly natural dynamics. Thus one could wonder whether similar behavior

to the one we observe for the dynamics we actually study occurs also in this case.

To answer this question, we first observe that by the chain rule, the (sub)-gradient of L' with respect to μ, ℓ, r are given by

$$\begin{aligned}\nabla_\mu L'(\mu,\ell,r) &= \operatorname{sgn}\left(L(\mu,\ell,r)\right)\nabla_\mu L(\mu,\ell,r)\\ \nabla_\ell L'(\mu,\ell,r) &= \operatorname{sgn}\left(L(\mu,\ell,r)\right)\nabla_\ell L(\mu,\ell,r)\\ \nabla_r L'(\mu,\ell,r) &= \operatorname{sgn}\left(L(\mu,\ell,r)\right)\nabla_r L(\mu,\ell,r)\ ,\end{aligned}$$

that is, they are the same as for L except modulated by the sign of L.

We now show that the *optimal* discriminator dynamics is identical to the one that we analyze in the paper (8), and hence still provably converge. This requires some thought; indeed a priori it is not even clear that the optimal discriminator dynamics are well-defined, since the optimal discriminator is no longer unique. This is because for any μ^*, μ, the sets $A_1 = \{x : G_{\mu^*}(x) \geq G_\mu(x)\}$ and $A_2 = \{x : G_\mu(x) \geq G_{\mu^*}(x)\}$ both achieve the maxima in (12), since

$$\int_{A_1} G_\mu(x) - G_{\mu^*}(x)dx = -\int_{A_2} G_\mu(x) - G_{\mu^*}(x)dx\ . \tag{15}$$

However, we show that the optimal discriminator dynamics are still well-formed. WLOG assume that $\int_{A_1} G_\mu(x) - G_{\mu^*}(x)dx \geq 0$, so that A_1 is also the optimal discriminator for the dynamics we consider in the paper. If we let $\ell^{(i)}, r^{(i)}$ be the left and right endpoints of the intervals in A_i for $i = 1, 2$, we have that the update to μ induced by $(\ell^{(1)}, r^{(1)})$ is given by

$$\nabla_\mu L'(\mu, \ell^{(1)}, r^{(1)}) = \nabla_\mu L(\mu, \ell^{(1)}, r^{(1)})\ ,$$

so the update induced by $(\ell^{(1)}, r^{(1)})$ is the same as the one induced by the optimal discriminator dynamics in the paper. Moreover, the update to μ induced by $(\ell^{(2)}, r^{(2)})$ is given by

$$\begin{aligned}\nabla_\mu L'(\mu, \ell^{(2)}, r^{(2)}) &= \operatorname{sgn}\left(L(\mu, \ell^{(2)}, r^{(2)})\right)\nabla_\mu L(\mu, \ell^{(2)}, r^{(2)})\\ &\overset{(a)}{=} -\nabla_\mu(-L(\mu, \ell^{(1)}, r^{(1)}))\\ &= \nabla_\mu L(\mu, \ell^{(1)}, r^{(1)})\ ,\end{aligned}$$

where (a) follows from the assumption that $\int_{A_1} G_\mu(x) - G_{\mu^*}(x)dx \geq 0$ and from (15), so it is also equal to the the one induced by the optimal discriminator dynamics in the paper. Hence the optimal discriminator dynamics are well-formed and unchanged from the optimal discriminator dynamics described in the paper.

Thus the question is whether the first order approximation of this dynamics and/or the unrolled first order dynamics exhibit the same qualitative behavior too. To evaluate the effectiveness, we performed for these dynamics experiments analogous to the ones summarized in Figure 2 in the case of the dynamics we actually analyzed. The results of these experiments are presented in Figure 4. Although the probability of success for these dynamics is higher, they still often do not converge. We can thus see that a similar dichotomy occurs here as in the context of the dynamics we actually study. In particular, we still observe the discriminator collapse phenomena in these first order dynamics.

B.1. Why does discriminator collapse still happen?

It might be somewhat surprising that even with absolute values discriminator collapse occurs. Originally the discriminator collapse occurred because if an interval was stuck in a negative region, it always subtracts from the value of the loss function, and so the discriminator is incentivized to make it disappear. Now, since the value of the loss is always nonnegative, it is not so clear that this still happens.

Despite this, we still observe discriminator collapse with these dynamics. Here we describe one simple scenario in which discriminator collapse still occurs. Suppose the discriminator intervals have left and right endpoints ℓ, r and $L(\mu, \ell, r) > 0$. More if it is the case that $\int_{\ell_i}^{r_i} G_{\mu^*}(x) - G_\mu(x)dx < 0$ for some $i = 1, 2$. that is, on one of the discriminator intervals the value of the loss is negative, then the discriminator is still incentivized locally to reduce this interval to zero, as doing so increases both $L(\mu, \ell, r)$ and hence $L'(\mu, \ell, r)$. Symmetrically if $L(\mu, \ell, r) < 0$ and there is a discriminator interval on which the loss is positive, the discriminator is incentivized locally to reduce this interval to zero, since that increases $L'(\mu, \ell, r)$. This causes the discriminator collapse and subsequently causes the training to fail to converge.

C. Omitted Proofs from Section 4.1

This appendix is dedicated to a proof of Theorem 4.1. We start with some remarks on the proof techniques for these main lemmas. At a high level, Lemmas 4.3, 4.5, 4.6 all follow from involved case analyses. Specifically, we are able to deduce structure about the possible discriminator intervals by reasoning about the structure of the current mean estimate $\widehat{\mu}$ and the true means. From there we are able to derive bounds on how these discriminator intervals affect the derivatives and hence the update functions.

To prove Lemma 4.4, we carefully study the evolution of the optimal discriminator as we make small changes to the generator. The key idea is to show that when the generator means are far from the true means, then the zero crossings of $F(\widehat{\mu}, x)$ cannot evolve too unpredictably as we change $\widehat{\mu}$. We do so by showing that locally, in this setting F can

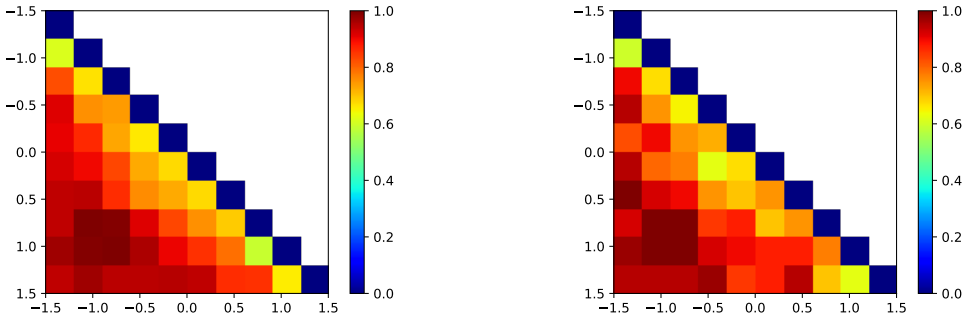

Figure 4. Heatmap of success probability for random discriminator initialization for regular GAN training, unrolled GAN training with dynamics induced by 14

be approximated by a low degree polynomial with large coefficients, via bounding the condition number of a certain Hermite Vandermonde matrix. This gives us sufficient control over the local behavior of zeros to deduce the desired claim. By being sufficiently careful with the bounds, we are then able to go from this to the full generality of the lemma. We defer further details to Appendix C.

C.1. Setup

By inspection on the form of (13), we see that the gradient of the function $f_{\mu^*}(\widehat{\mu})$ if it is defined must be given by

$$\frac{\partial f_{\mu^*}}{\partial \widehat{\mu}_i} = \frac{1}{\sqrt{2\pi}} \sum_{i=1,2} \left(e^{-(\widehat{\mu}_i - r_i)^2/2} - e^{-(\widehat{\mu}_i - \ell_i)^2/2} \right) .$$

Here $I_i = [\ell_i, r_i]$ are the intervals which achieve the supremum in (4). While these intervals may not be unique, it is not hard to show that this value is well-defined, as long as $\widehat{\mu} \neq \mu^*$, that is, when the optimal discriminator intervals are unique as sets.

Recall $F_{\mu^*}(\widehat{\mu}, x) = G_{\mu^*}(x) - G_{\widehat{\mu}}(x)$.

C.2. Basic Math Facts

Before we begin, we require the following facts.

We first need that the Gaussian, and any fixed number of derivatives of the Gaussian, are Lipschitz functions.

Fact C.1. *For any constant i, there exists a constant B such that for all $x, \mu \in \mathbb{R}$, $\frac{\mathrm{d}^i}{\mathrm{d}x^i} \mathcal{N}(x, \mu, \sigma^2 = 1) \leq B$.*

Proof. Note that every derivative of the Gaussian PDF (including the $0th$) is a bounded function. Furthermore, all these derivatives eventually tend to 0 whenever the input goes towards ∞ or $-\infty$. Thus, any particular derivative is bounded by a constant for all $\mathbb{R}$. Furthermore, shifting the mean of the Gaussian does not change the set of values the derivatives of its derivative takes (only their locations). □

We also need the following bound on the TV distance between two Gaussians, which is folklore, and is easily proven via Pinsker's inequality.

Fact C.2 (folklore)**.** *If two univariate Gaussians with unit variance have means within distance at most Δ then their TV distance is at most $O(1) \cdot \Delta$.*

This immediately implies the following, which states that f_{μ^*} is Lipschitz.

Corollary C.3. *There is some absolute constant C so that for any μ, ν, we have $|f_{\mu^*}(\mu) - f_{\mu^*}(\nu)| \leq C\|\mu - \nu\|_2$.*

We also need the following basic analysis fact:

Fact C.4 (folklore)**.** *Suppose $g : \mathbb{R}^d \to \mathbb{R}$ is B-Lipschitz for some B. Then g is differentiable almost everywhere.*

This implies that f_{μ^*} is indeed differentiable except on a set of measure zero. As mentioned previously, we will always assume that we never land within this set during our analysis.

C.3. Proof of Theorem 4.1 given Lemmata

Before we prove the various lemmata described in the body, we show how Theorem 4.1 follows from them.

Proof of Theorem 4.1. Set δ' be a sufficiently small constant multiple of δ. Provided we make the nonzero constant factor on the step size sufficiently small (compared to δ'/δ), and the exponent on δ in the magnitude step size at least one, the magnitude of our step size will be at most δ'. Thus, in any step where $\widehat{\mu} \in \mathrm{Opt}(\delta')$, we end the step outside of this set but still in $\mathrm{Opt}(2\delta')$. By Lemma C.2, for a sufficiently small choice of constant in the definition of δ', the TV-distance at the end of such a step will be at most δ.

Contrapositively, in any step where the TV-distance at the start is more than δ, we will have at the start that

$\widehat{\mu} \notin \mathrm{Opt}(\delta')$. Then, it suffices to prove that the step decreases the total variation distance additively by at least $1/\operatorname{poly}(C, e^{C^2}, 1/\delta)$ in this case. For appropriate choices of constants in expression for the step size (sufficiently small multiplicative and sufficiently large in the exponent), this is immediately implied by Lemma 4.4 and Lemma 4.2 provided that $\mu^*, \widehat{\mu}, \widehat{\mu}' \in B(2C)$ and $|\widehat{\mu}_1 - \widehat{\mu}_2| \geq \delta$ at the beginning of each step. The condition that we always are within $B(2C)$ at the start of each step is proven in Lemma 4.6 and the condition that the means are separated (ie., that we don't have mode collapse) is proven in Lemma 4.5. □

It is interesting that a critical component of the above proof involves proving explicitly that mode collapse does not occur. This suggests the possibility that understanding mode collapse may be helpful in understanding convergence of Generative Adversarial Models and Networks.

C.4. Proof of Lemma 4.3

In this section we prove Lemma 4.3. We first require the following fact:

Fact C.5 ((Markov, 1892)). *Let $p(x) = \sum_{i=0}^{d} c_j x^j$ be a degree d polynomial so that $|p(x)| \leq 1$ for all $x \in [-1, 1]$. Then $\max_{0 \leq j \leq d} |c_j| \leq (\sqrt{2}+1)^d$. More generally, if $|p(x)| \leq \alpha$ for all $x \in [-\rho, \rho]$, then $\max_{0 \leq j \leq d} |c_j \rho^j| \leq O(\alpha)$.*

We also have the following, elementary lemma:

Lemma C.6. *Suppose $\widehat{\mu}_2 > \mu_i^*$ for all i. Then there is some $x > \widehat{\mu}_2$ so that $F_{\mu^*}(\widehat{\mu}, x) < 0$.*

We are now ready to prove Lemma 4.3

Proof of Lemma 4.3. We proceed by case analysis on the arrangement of the $\widehat{\mu}$ and μ^*.

Case 1: $\mu_1^* < \widehat{\mu}_1$ and $\mu_2^* < \widehat{\mu}_2$ In this case we have $F_{\mu^*}(\widehat{\mu}, x) \leq 0$ for all $x \geq \widehat{\mu}_2$. Hence the optimal discriminators are both to the left of $\widehat{\mu}_2$. Moreover, by a symmetric logic we have $F_{\mu^*}(\widehat{\mu}, x) \geq 0$ for all $x \leq \mu_1^*$, so the optimal discriminator has an interval of the form $I_1 = [-\infty, r_1]$ and possibly $I_2 = [l_2, r_2]$ where $r_1 < l_2 < r_2 < \widehat{\mu}_2$. Then it is easy to see that $\frac{\partial f}{\partial \widehat{\mu}_2}(\widehat{\mu}_2) \geq \frac{1}{\sqrt{2\pi}} e^{-(\widehat{\mu}_2 - r_2)^2/2} \geq \frac{1}{\sqrt{2\pi}} e^{-2C^2}$.

Case 2: $\widehat{\mu}_1 < \mu_1^*$ and $\widehat{\mu}_2 < \mu_2^*$ This case is symmetric to Case (1).

Case 3: $\widehat{\mu}_1 < \mu_1^* < \mu_2^* < \widehat{\mu}_2$ By Lemma C.6, we know that $F_{\mu^*}(\widehat{\mu}, x) < 0$ for some $x \geq \widehat{\mu}_2$, and similarly $F_{\mu^*}(\widehat{\mu}, x) < 0$ for some $x \leq \widehat{\mu}_1$. Since clearly $F(\mu^*)(\widehat{\mu}, x) > 0$ for $x \in [\mu_1^*, \mu_2^*]$, by Theorem A.2 and continuity, the optimal discriminator has one interval. Denote it by $I = [\ell, r]$, so that we have $\ell \leq \mu_1^*$ and $r \geq \mu_2^*$. Suppose $\ell \geq \widehat{\mu}_1$. Then

$$\begin{aligned}
\frac{\partial f}{\partial \widehat{\mu}_1}(\widehat{\mu}_1) &= \frac{1}{\sqrt{2\pi}} \left(e^{-(\widehat{\mu}_1 - \ell)^2/2} - e^{-(\widehat{\mu}_1 - r)^2/2} \right) \\
&= \frac{1}{\sqrt{2\pi}} e^{-(\widehat{\mu}_1 - \ell)^2/2} \left(1 - e^{-(\widehat{\mu}_1 - \ell)(r-\ell)} e^{-(r-\ell)^2/2} \right) \\
&\geq \frac{1}{\sqrt{2\pi}} e^{-2C^2} \left(1 - e^{-\delta^2/2} \right) .
\end{aligned}$$

We get the symmetric bound on $\frac{\partial f_{\mu^*}}{\partial \widehat{\mu}_2}(\widehat{\mu}_2)$ if $r \leq \widehat{\mu}_2$. The final case is if $\ell < \widehat{\mu}_1 < \widehat{\mu}_2 < r$. Consider the auxiliary function

$$H(\mu) = e^{-(\ell - \mu)^2/2} - e^{-(r-\mu)^2/2} .$$

On the domain $[\ell, r]$, this function is monotone decreasing. Moreover, for any $\mu \in [\ell, r]$, we have

$$\begin{aligned}
H'(\mu) &= (\ell - \mu) e^{-(\ell-\mu)^2/2} - (r - \mu) e^{-(r-\mu)^2/2} \\
&\leq -\frac{r - \ell}{2} e^{-(r-\ell)^2/8} \\
&\leq -\frac{\gamma}{2} e^{-\gamma^2/8} .
\end{aligned}$$

In particular, this implies that $H(\widehat{\mu}_1) < H(\widehat{\mu}_2) - \gamma^2 e^{-\gamma^2/8}/2$, so at least one of $H(\widehat{\mu}_2)$ or $H(\widehat{\mu}_1)$ must be $\gamma^2 e^{-\gamma^2/8}/4$ in absolute value. Since $\frac{\partial f_{\mu^*}}{\partial \widehat{\mu}_i}(\widehat{\mu}_i) = H(\widehat{\mu}_i)$, this completes the proof in this case.

Case 4: $\mu_1^* < \widehat{\mu}_1 < \widehat{\mu}_2 < \mu_2^*$ By a symmetric argument to Case 3, we know that the optimal discriminator intervals are of the form $(-\infty, r]$ and $[\ell, \infty)$ for some $r < \widehat{\mu}_1 < \widehat{\mu}_2 < \ell$. The form of the derivative is then exactly the same as in the last sub-case in Case 3 with signs reversed, so the same bound holds here.

□

C.5. Proof of Lemma 4.4

We now seek to prove Lemma 4.4. Before we do so, we need to get lower bounds on derivatives of finite sums of Gaussians with the same variance. In particular, we first show:

Lemma C.7. *Fix $\gamma \geq \delta > 0$ and $C \geq 1$. Suppose we have $\mu^*, \widehat{\mu} \in B(C)$, $\mu^*, \widehat{\mu} \in \mathrm{Sep}(\gamma)$, with $\widehat{\mu} \notin \mathrm{Rect}(\delta)$, where all these vectors have constant length k. Then, for any $x \in [-C, C]$, we have that $|\frac{d^i}{dx^i} F_{\mu^*}(\widehat{\mu}, x)| \geq \Omega(1) \cdot (\delta/C)^{O(1)} e^{-C^2/2}$ for some $i = 0, \ldots, 2k-1$.*

Proof. Observe that the value of the ith derivative of $F_{\mu^*}(\widehat{\mu}, x)$ for any x is given by

$$\frac{d^i}{dx^i} F_{\mu^*}(\widehat{\mu}, x) = \frac{1}{\sqrt{2\pi}} \sum_{j=1}^{2k} w_j (-1)^i H_i(z_j) e^{-z_j^2/2} ,$$

where $w_j \in \{-1/k, 1/k\}$, the z_j is either $x - \mu_j^*$ or $x - \widehat{\mu}_j$, and $H_i(z)$ is the ith (probabilist's) Hermite polynomial. Note that the $(-1)^i H_i$ are orthogonal with respect to the Gaussian measure over $\mathbb{R}$, and are orthonormal after some finite scaling that depends only on i and is therefore constant. Hence, if we form the matrix $M_{ij} = (-1)^i H_i(x - z_j)$, if we define $u_i = \frac{d^i}{dx^i} F_{\mu^*}(\widehat{\mu}, x)$ for $i = 0, \ldots, 2k-1$, we have that $Mv = u$, where $v_j = \frac{1}{\sqrt{2\pi}} w_j e^{-(x-z_j)^2/2}$. By assumption, we have $\|v\|_2 \geq \Omega(\sqrt{k} \cdot e^{-C^2/2}) = \Omega(e^{-C^2/2})$. Thus, to show that some u_i cannot be too small, it suffices to show a lower bound on the smallest singular value of M. To do so, we leverage the following fact, implicit in the arguments of (Gautschi, 1990):

Theorem C.8 ((Gautschi, 1990)). *Let $p_r(z)$ be family of orthonormal polynomials with respect to a positive measure $d\sigma$ on the real line for $r = 1, 2, \ldots, t$ and let $z_1, \ldots, z_t$ be arbitrary real numbers with $z_i \neq z_j$ for $i \neq j$. Define the matrix V given by $V_{ij} = p_i(z_j)$. Then, the smallest singular value of V, denoted $\sigma_t(V)$, is at least*

$$\sigma_{min}(V) \geq \left(\int_{\mathbb{R}} \sum_{r=1}^{t} \ell_r(y)^2 d\sigma(y) \right)^{-1/2},$$

where $\ell_r(y) = \prod_{s \neq r} \frac{y - z_s}{z_r - z_s}$ is the Langrange interpolating polynomial for the z_r.

Set $p_r = H_{r-1}$ $t = 2k$, and σ as the Gaussian measure; then apply the theorem. Observe that for any i, j, we have $|z_i - z_j| \geq \min(\delta, \gamma) \geq \delta$ and $|z_i| \leq C$. Hence the largest coefficient of any Lagrange interpolating polynomial through the z_i is at most $(\frac{C}{\delta})^{2k-1}$ with degree $2k-1$. So, the square of any such polynomial has degree at most $2(2k-1)$ and max coefficient at most $2k(\frac{C}{\delta})^{2(2k-1)}$ This implies that

$$\begin{aligned}
&\int_{\mathbb{R}} \sum_{r=1}^{2k} \ell_r(y)^2 \, \mathrm{d}\sigma(y) = \sum_{r=1}^{2k} \int_{\mathbb{R}} \ell_r(y)^2 \, \mathrm{d}\sigma(y) \\
&\leq \sum_{r=1}^{2k} 2(2k-1) \cdot 2k \left(\frac{C}{\delta}\right)^{2(2k-1)} \max_{s \in [2(2k-1)]} \int_{\mathbb{R}} y^s \mathrm{d}\sigma(y) \\
&\leq O(1) \cdot \left(\frac{C}{\delta}\right)^{4k} \max_{s \in [4k]} \int_{-\infty}^{\infty} y^s e^{-y^2/2} \, \mathrm{d}y \\
&\leq O(1) \cdot \left(\frac{C}{\delta}\right)^{O(1)}.
\end{aligned}$$

Hence by Theorem C.8 we have that $\sigma_{\min}(V) \geq \Omega(1) \cdot \left(\frac{\delta}{C}\right)^{O(1)}$. Therefore, we have that $\|u\|_2 \geq \Omega(1) \cdot (\delta/C)^{O(1)} e^{-C^2/2}$, which immediately implies the desired statement. □

We next show that the above Lemma can be slightly generalized, so that we can replace the condition $\widehat{\mu} \notin \mathrm{Rect}(\delta)$ with $\widehat{\mu} \notin \mathrm{Opt}(\delta)$.

Lemma C.9. *Fix $C \geq 1 \geq \gamma \geq \delta \geq \Xi > 0$. Suppose we have $\mu^*, \widehat{\mu} \in B(C)$, $\mu^*, \widehat{\mu} \in \mathrm{Sep}(\gamma)$, with $\widehat{\mu} \notin \mathrm{Opt}(\delta)$. Then for any $x \in [-C, C]$, we have that $|\frac{d^i}{dx^i} F_{\mu^*}(\widehat{\mu}, x)| \geq \Omega(1) \cdot (\delta e^{-C^2}/C)^{O(1)}$ for some $i = 0, \ldots, 3$.*

Proof. Let Ξ be of the form $\Omega(1) \cdot (\delta e^{-C^2}/C)^{O(1)}$, where we will pick its precise value later. Lemma C.7 with δ in that Lemma set to Ξ and $k = 2$ proves the special case when $\widehat{\mu} \notin \mathrm{Rect}(\Xi)$. Thus, the only remaining case is when $\widehat{\mu}_i$ is close to μ_i^* for some i and far away for the other i. Without loss of generality, we assume the first entries are the close pair. Then we have $|\widehat{\mu}_1 - \mu_1^*| \leq \Xi$ and $|\widehat{\mu}_2 - \mu_2^*| \geq \delta$.

There are four terms in the expression for $\frac{d^i}{dx^i} F_{\mu^*}(\widehat{\mu}, x)$ corresponding to each of $\widehat{\mu}_1, \widehat{\mu}_2, \mu_1^*, \mu_2^*$. Lemma C.7 with $\delta = \Xi$ and $k = 1$ implies that the contribution of the $\widehat{\mu}_2$ and μ_2^* terms to at least one of the 0th through 3rd derivatives has magnitude at least $\Omega(1) \cdot (\delta e^{-C^2}/C)^{O(1)}$. Fact C.2 and Lemma C.10 (below) imply that the contribution of the $\widehat{\mu}_1$ and μ_1^* terms to these derivatives has magnitude at most $O(1) \cdot \Xi^4$. Thus, there exists a $\Xi = \Omega(1) \cdot (\delta/C)^{O(1)} e^{-C^2/2}$ such that the magnitude of the contribution of these second two terms is less than half that of the first two, which gives a lower bound on the magnitude of the sum of all the terms of $\Omega(1) \cdot (\delta/C)^{O(1)} e^{-C^2/2}$. □

We now show that any function which always has at least one large enough derivative—including its 0th derivaive—on some large enough interval must have a nontrivial amount of mass on the interval.

Lemma C.10. *Let $0 < \xi < 1$ and $t \in \mathbb{N}$. Let $F(x) : \mathbb{R} \to \mathbb{R}$ be a $(t+1)$-times differentiable function such that at every point x on some interval I of length $|I| \geq \xi$, $F(x) \geq 0$ and there exists an $i = i(x) \in 0, \ldots, t$ such that $|\frac{\mathrm{d}^i}{\mathrm{d}x^i} F(x)| \geq B'$ for some B'. Also suppose $|\frac{\mathrm{d}^{t+1}}{\mathrm{d}x^{t+1}} F(x)| \leq B$ for some B. Then,*

$$\int_z^y F(x) dx \geq \left(\frac{B' \cdot (\Omega(1) \cdot \xi)^{t+1} \cdot \min[(B'/B)^{t+2}, 1]}{(t+1)! \cdot (t+1)} \right).$$

Proof. Let $0 < a < 1$ be a non-constant whose value we will choose later. If I has length more than $a\xi$, truncate it to have this length. Let z denote the midpoint of I. By assumption, we know that there is some $i \in 0, \ldots, t$ such that $|\frac{d^i}{dx^i} F(x)| > \xi$. Thus, by Taylor's theorem, we have that $F(\widehat{\mu}, x) \geq p(x - z) - (B/(t+1)!) \cdot |x - z|^{t+1}$ for some degree t polynomial p that has some coefficient of magnitude at least $B'/t!$.

Thus, if we let $G(y) = \int_z^y p(x) dx$, then $G(y)$ is a degree $t+1$ polynomial with some coefficient which is at

least $B'/(t! \cdot t)$. By the nonnegativity of F on I, we have that G is nonnegative on $[-a\xi/2, a\xi/2]$. By this and the contrapositive of Fact C.5 (invoked with α set to a sufficiently small nonzero constant multiple of B), we have for some such y and some constant $B'' > 0$ that $G(y) = |G(y)| \geq B''(|I|/2)^{t+1}B'/(t! \cdot t)$. Therefore, at this point, we have

$$\int_z^y F(x)dx \geq G(y) - \int_z^y (B/(t+1)!) \cdot |x-z|^{t+1}\mathrm{d}x$$
$$\geq \frac{B''a^{t+1}(\xi/2)^{t+1}B'}{t! \cdot t} - \frac{B(a\xi/2)^{t+2}}{(t+1)! \cdot (t+1)}$$
$$\geq \left(\frac{a^{t+1}(\xi/2)^{t+1}(B''B' - B\xi a/2)}{(t+1)! \cdot (t+1)}\right)$$
$$\geq \left(\frac{a^{t+1}(\xi/2)^{t+1}(B''B' - Ba/2)}{(t+1)! \cdot (t+1)}\right).$$

If $B'B'' \leq B$, we set $a = B'B''/B \leq 1$ which gives

$$\int_z^y F(x)dx \geq \left(\frac{(B')^{t+2}(\Omega(1) \cdot \xi/B)^{t+1}}{(t+1)! \cdot (t+1)}\right).$$

Otherwise, $B'B'' \geq B$ and we perform this substitution along with $a = 1$ which gives the similar bound

$$\int_z^y F(x)dx \geq \left(\frac{B'(\Omega(1) \cdot \xi)^{t+1}}{(t+1)! \cdot (t+1)}\right).$$

Together, these bounds imply that we always have

$$\int_z^y F(x)dx \geq \left(\frac{B' \cdot (\Omega(1) \cdot \xi)^{t+1} \cdot \min[(B'/B)^{t+2}, 1]}{(t+1)! \cdot (t+1)}\right).$$

□

This allows us to prove the following lemma, which lower bounds how much mass F can put on any interval which is moderately large. Formally:

Lemma C.11. *Fix $C \geq 1 \geq \gamma \geq \delta > 0$. Let $K = \Omega(1) \cdot (\delta e^{-C^2}/C)^{O(1)}$ be the K for which Lemma 4.3 is always true with those parameters. Let $\mu^*, \widehat{\mu}$ be so that $\widehat{\mu} \notin \mathrm{Opt}(\delta)$, $\widehat{\mu}, \mu^* \in \mathrm{Sep}(\gamma)$, and $\mu^*, \widehat{\mu} \in B(C)$. Then, there is a $\xi = \Omega(1) \cdot (\delta/C)^{O(1)}e^{-C^2})^{O(1)}$ such that for any interval I of length $|I| \geq \xi$ which satisfies $I \cap [-C - 2\sqrt{\log(100/K)}, C + 2\sqrt{\log(100/K)}] \neq \varnothing$ and on which $F(\widehat{\mu}, x)$ is nonnegative, we have*

$$\int_I |F(\widehat{\mu}, x)|dx \geq \Omega(1) \cdot (\delta e^{-C^2}/C)^{O(1)}\xi^{O(1)}.$$

Proof. By Lemma C.9 with C in that lemma set to $C + 2\sqrt{\log(100/K)}$, we get a lower bound of $\Omega(1) \cdot (\delta e^{-C^2}/C)^{O(1)}$ on the magnitude of at least one of the 0th through $3rd$ derivatives of $F(\widehat{\mu}, x)$ with respect to x. Set ξ equal to a sufficiently small (nonzero) constant times this value.

By Fact C.1 there exists a constant B such that the magnitude of the fifth derivative of $F(\widehat{\mu}, x)$ with respect to x—which is a linear combination of four fifth derivatives of Gaussians with constant coefficients—is at most B.

By Lemma C.10 applied to $F(\widehat{\mu}, x)$ as a function of x, we have $\int_I F(\widehat{\mu}, x)dx \geq \Omega(1) \cdot \xi^6$. □

Now we can prove Lemma 4.4.

Proof of Lemma 4.4. Let $A = [C - 2\sqrt{\log(100/K)}, C + 2\sqrt{\log(100/K)}]$ where $K = \Omega(1) \cdot (\delta e^{-C^2}/C)^{O(1)}$ is the K for which Lemma 4.3 is always true with those parameters.

Let $Z^{\pm}$ denote the set of all $x \in A$ for which $F(\widehat{\mu}', x)$ and $F(\widehat{\mu}, x)$ have different nonzero signs. Let Z^+ denote the subset of $Z^{\pm}$ where $F(\widehat{\mu}', x) > 0 > F(\widehat{\mu}, x)$ and Z^- denote the subset where $F(\widehat{\mu}', x) < 0 < \widehat{\mu}, x)$. Then $Z^{\pm} = Z^+ \cup Z^-$ and Z^+, Z^- are disjoint and Lebesgue-measurable. If $\mathrm{vol}(Z^+) \leq \mathrm{vol}(Z^-)$, switch $\widehat{\mu}$ and $\widehat{\mu}'$ so that $\mathrm{vol}(Z^+) \geq \mathrm{vol}(Z^-)$.

Note that Z^+ can be obtained by making cuts in the real line at the zeros of $F(\widehat{\mu}', x)$, $F(\widehat{\mu}, x)$, and $F(\widehat{\mu}', x) - F(\widehat{\mu}, x)$, then taking the union of some subset of the open intervals induced by these cuts. By Theorem A.2, the total number of such intervals is $O(1)$. Thus, Z^+ is the union of a constant number of open intervals. By similar arguments, Z^- is also the union of a constant number of open intervals.

We now prove that $\mathrm{vol}(Z^+), Z^{-1} \leq O(1) \cdot \|\widehat{\mu}' - \widehat{\mu}\|_1^{\Theta(1)} \cdot (\delta e^{-C^2}/C)^{-O(1)}$. Since $\mathrm{vol}(Z^+) \geq \mathrm{vol}(Z^-)$, it suffices to prove $\mathrm{vol}(Z^+) \leq O(1) \cdot \|\widehat{\mu}' - \widehat{\mu}\|_1^{\Theta(1)} \cdot (\delta e^{-C^2}/C)^{-O(1)}$. Note also that by Lemma C.2, each of these intervals has mass under $F(\widehat{\mu}', x)$ at most $\int_{\mathbb{R}} |F(\widehat{\mu}', x) - F(\widehat{\mu}, x)|\mathrm{d}x \leq O(1) \cdot \|\widehat{\mu}' - \widehat{\mu}\|_1$. By Lemma C.11 and Lemma 4.3, each of these intervals has length at most $O(1) \cdot \|\widehat{\mu}' - \widehat{\mu}\|_1^{\Theta(1)} \cdot (\delta e^{-C^2}/C)^{-O(1)}$. Since there are at most a constant number of such intervals, this is also a bound on $\mathrm{vol}(Z^+)$ (and $\mathrm{vol}(Z^-)$).

Let Y denote the set of $x \in A$ on which both $F(\widehat{\mu}, x)$ and $F(\widehat{\mu}', x)$ are nonnegative. Let X, X' denote the $x \notin A$ for which $F(\widehat{\mu}, x)$ and $F(\widehat{\mu}', x)$ are respectively positive. Let W, W' denote, respectively, the sets of endpoints of the union of the optimal discriminators for $\widehat{\mu}, \widehat{\mu}'$. Then the union of the optimal discriminators for $\widehat{\mu}, \widehat{\mu}'$ are respectively $Y \cup Z^- \cup X \cup W$ and $Y \cup Z^+ \cup X' \cup W'$. Furthermore, each of these two unions is given by some constant number of closed intervals and more specifically, that X, X' each contain at most two intervals by Lemma A.2. Thus, we have

for any i that

$$\begin{aligned}
&\left|\frac{\partial}{\partial\widehat{\mu}_i}\mathrm{TV}(\mu^*,\widehat{\mu})\right|_{\widehat{\mu}}^{\widehat{\mu}'}\\
&=\left|\int_{Y\cup Z^+\cup W'\cup X'}\frac{\mathrm{d}}{\mathrm{d}x}e^{(x-\widehat{\mu}_i')^2/2}\mathrm{d}x-\int_{Y\cup Z^-\cup W\cup X}\frac{\mathrm{d}}{\mathrm{d}x}e^{(x-\widehat{\mu}_i)^2/2}\mathrm{d}x\right|\\
&\le\left|\int_{Y\cup Z^+\cup W'\cup X'}\frac{\mathrm{d}}{\mathrm{d}x}e^{(x-\widehat{\mu}_i)^2/2}\mathrm{d}x-\int_{Y\cup Z^-\cup W\cup X}\frac{\mathrm{d}}{\mathrm{d}x}e^{(x-\widehat{\mu}_i)^2/2}\mathrm{d}x\right|\\
&\quad+O(1)\cdot|\widehat{\mu}_i'-\widehat{\mu}_i|\,,
\end{aligned}$$

by Lipschitzness, and so

$$\begin{aligned}
&\left|\frac{\partial}{\partial\widehat{\mu}_i}\mathrm{TV}(\mu^*,\widehat{\mu})\right|_{\widehat{\mu}}^{\widehat{\mu}'}\\
&=\left|\int_{Z^+\cup X'}\frac{\mathrm{d}}{\mathrm{d}x}e^{(x-\widehat{\mu}_i)^2/2}\mathrm{d}x-\int_{Z^-\cup X}\frac{\mathrm{d}}{\mathrm{d}x}e^{(x-\widehat{\mu}_i)^2/2}\mathrm{d}x\right|\\
&\quad+O(1)\cdot|\widehat{\mu}_i'-\widehat{\mu}_i|\\
&\le\left|\int_{Z^+}\frac{\mathrm{d}}{\mathrm{d}x}e^{(x-\widehat{\mu}_i)^2/2}\mathrm{d}x\pm\frac{4}{100}\cdot K\right.\\
&\quad\left.-\int_{Z^-}\frac{\mathrm{d}}{\mathrm{d}x}e^{(x-\widehat{\mu}_i)^2/2}\mathrm{d}x\pm\frac{4}{100}\cdot K\right|\\
&\quad+O(1)\cdot|\widehat{\mu}_i'-\widehat{\mu}_i|\\
&\le\left|\int_{Z^+}\frac{\mathrm{d}}{\mathrm{d}x}e^{(x-\widehat{\mu}_i)^2/2}\mathrm{d}x\right|+\left|\int_{Z^-}\frac{\mathrm{d}}{\mathrm{d}x}e^{(x-\widehat{\mu}_i)^2/2}\mathrm{d}x\right|\\
&\quad+\frac{8}{100}\cdot K+O(1)\cdot|\widehat{\mu}_i'-\widehat{\mu}_i|\\
&\le 2\mathrm{vol}(Z^+)\left|\sup_{x\in\mathbb{R}}\frac{\mathrm{d}}{\mathrm{d}x}e^{(x-\widehat{\mu}_i)^2/2}\right|+\frac{8}{100}\cdot K+O(1)\cdot|\widehat{\mu}_i'-\widehat{\mu}_i|\\
&\le O(1)\cdot\|\widehat{\mu}'-\widehat{\mu}\|_2^{\Theta(1)}\cdot(\delta e^{-C^2}/C)^{-O(1)}+\frac{8}{100}\cdot K\,.
\end{aligned}$$

This bound also upper bounds $\|\nabla f_{\mu^*}(\widehat{\mu}')-\nabla f_{\mu^*}(\widehat{\mu})\|_2$ up to a constant factor. Thus, if we choose our step to have magnitude $\|\widehat{\mu}'-\widehat{\mu}\|_2\le\Omega(1)\cdot(\delta e^{-C^2}/C)^{O(1)}$ with appropriate choices of constants, we get

$$\|\nabla f_{\mu^*}(\widehat{\mu}')-\nabla f_{\mu^*}(\widehat{\mu})\|_2\le K/2\le\|\nabla f_{\mu^*}(\widehat{\mu})\|_2/2\,,$$

as claimed □

C.6. Proof of Lemma 4.5

We now prove Lemma 3.4, which forbids mode collapse.

Proof of Lemma 4.5. Since $\eta\le\delta$, if $|\widehat{\mu}_1-\widehat{\mu}_2|>2\delta$ then clearly $\widehat{\mu}'\in\mathrm{Sep}(\delta)$, since the gradient is at most a constant since the function is Lipschitz. Thus assume WLOG that $|\widehat{\mu}_1-\widehat{\mu}_2|\le 2\delta\le\gamma/50$. There are now six cases:

Case 1: $\widehat{\mu}_1\le\mu_1^*\le\mu_2^*\le\widehat{\mu}_2$ This case cannot happen since we assume $|\widehat{\mu}_1-\widehat{\mu}_2|\le 2\delta\le\gamma/50$.

Case 2: $\mu_1^*\le\widehat{\mu}_1\le\widehat{\mu}_2\le\mu_2^*$ In this case, by Lemma C.6, we know F is negative at $-\infty$ and at $+\infty$. Since clearly $F\ge 0$ when $x\in[\mu_1^*,\mu_2^*]$, by Theorem A.2 and continuity, the discriminator intervals must be of the form $(-\infty,r],[\ell,\infty)$ for some $r\le\widehat{\mu}_1\le\widehat{\mu}_2\le\ell$. Thus, the update to $\widehat{\mu}_i$ is (up to a constant factor of $\sqrt{2\pi}$) given by $e^{-(\ell-\widehat{\mu}_i)^2/2}-e^{-(r-\widehat{\mu}_i)^2/2}$. The function $Q(x)=e^{-(\ell-x)^2/2}-e^{-(r-x)^2/2}$ is monotone on $x\in[r,\ell]$, and thus $\widehat{\mu}_i$ must actually move away from each other in this scenario.

Case 3: $\mu_1^*\le\widehat{\mu}_1\le\mu_2^*\le\widehat{\mu}_2$ In this case we must have $|\mu_2^*-\widehat{\mu}_1|\le 2\delta$ and similarly $|\mu_2^*-\widehat{\mu}_2|\le 2\delta$. We claim that in this case, the discriminator must be an infinitely long interval $(-\infty,m]$ for some $m\le\widehat{\mu}_1$. This is equivalent to showing that the function $F(\widehat{\mu},x)$ has only one zero, and this zero occurs at some $m\le\widehat{\mu}_1$. This implies the lemma in this case since then the update to $\widehat{\mu}_1$ and $\widehat{\mu}_2$ are then in the same direction, and moreover, the magnitude of the update to $\widehat{\mu}_1$ is larger, by inspection.

We first claim that there are no zeros in the interval $[\widehat{\mu}_1,\widehat{\mu}_2]$. Indeed, in this interval, we have that

$$\begin{aligned}
\sqrt{2\pi}D_{\widehat{\mu}}(x)&\ge 2e^{-(\gamma/50)^2/2}\\
&=2e^{-\gamma^2/5000}\\
&\ge 2\left(1-\frac{\gamma^2}{5000}+O(\gamma^4)\right)\\
&\ge 2\left(1-\frac{\gamma^2}{10}\right)\,,
\end{aligned}$$

but

$$\begin{aligned}
\sqrt{2\pi}D_{\mu^*}(x)&\le 1+e^{-(\gamma-2\delta)^2/2}\\
&=1+e^{-(49\gamma/50)^2/2}\\
&\le 2-\frac{\gamma^2}{2}\,.
\end{aligned}$$

Hence $G_{\widehat{\mu}}(x)\ge G_{\mu^*}(x)$ for all $x\in[\widehat{\mu}_1,\widehat{\mu}_2]$, and so there are no zeros in this interval. Clearly there are no zeros of F when $x\ge\widehat{\mu}_2$, because in that regime $e^{-(x-\widehat{\mu}_i)^2/2}\ge e^{-(x-\mu_i^*)^2/2}$ for $i=1,2$. Similarly there are no zeros of F when $x\le\mu_1^*$. Thus all zeros of F must occur in the interval $[-\mu_1^*,\widehat{\mu}_1]$.

We now claim that there are no zeroes of F on the interval $[\alpha+10\delta,\widehat{\mu}_1]$, where $\alpha=(\mu_1^*+\widehat{\mu}_1)/2$. Indeed,

on this interval, we have

$$\begin{aligned}&\sqrt{2\pi}F(\widehat{\mu},x)\\&= e^{-(x-\mu_1^*)^2/2} - e^{-(x-\widehat{\mu}_2)^2/2} + e^{-(x-\mu_1^*)^2/2} - e^{-(x-\widehat{\mu}_1)^2/2}\\&\le e^{-(x-\mu_1^*)^2/2} - e^{-(x-\widehat{\mu}_2)^2/2} < 0\ ,\end{aligned}$$

where the first line follows since moving μ_2^* to $\widehat{\mu}_1$ only increases the value of the function on this interval, and the final line is negative as long as $x > (\mu_1^* + \widehat{\mu}_2)/2$, which is clearly satisfied by our choice of parameters. By a similar logic (moving $\widehat{\mu}_2$ to μ_2^*), we get that on the interval $[\mu_1^*, \alpha - 10\delta]$, the function is strictly positive. Thus all zeros of F must occur in the interval $[\alpha - 10\delta, \alpha + 10\delta]$.

We now claim that in this interval, the function F is strictly decreasing, and thus has exactly one zero (it has at least one zero because the function changes sign). The derivative of F with respect to x is given by

$$\begin{aligned}&\sqrt{2\pi}\frac{\partial F}{\partial x}(\widehat{\mu},x)\\&= (\mu_1^* - x)e^{-(x-\mu_1^*)^2/2} - (\widehat{\mu}_2 - x)e^{-(x-\widehat{\mu}_2)^2/2}\\&\quad + (\mu_2^* - x)e^{-(x-\mu_2^*)^2/2} - (\widehat{\mu}_2 - x)e^{-(x-\widehat{\mu}_1)^2/2}\ .\end{aligned}$$

By Taylor's theorem, we have

$$\begin{aligned}&(\mu_1^* - x)e^{-(x-\mu_1^*)^2/2} - (\widehat{\mu}_2 - x)e^{-(x-\widehat{\mu}_2)^2/2}\\&= -2re^{-\alpha^2/2} + O\left(H_2(\delta)e^{-(r-10\delta)^2/2}\delta^2\right)\ ,\end{aligned}$$

where H_2 is the second (probabilist's) Hermite polynomial, and $r = |\mu_1^* - \alpha|$. On the other hand, by another application of Taylor's theorem, we also have

$$\begin{aligned}&(\mu_2^* - x)e^{-(x-\mu_2^*)^2/2} - (\widehat{\mu}_2 - x)e^{-(x-\widehat{\mu}_1)^2/2}\\&= O\left(\delta H_2(\delta)e^{-(r-10\delta)^2/2}\right)\ .\end{aligned}$$

Thus, altogether we have

$$\begin{aligned}&\sqrt{2\pi}\frac{\partial F}{\partial x}(\widehat{\mu},x)\\&\le -2re^{-\alpha^2/2}\\&\quad + O\left(\delta H_2(\delta)e^{-(r-10\delta)^2/2}\right)\\&< 0\end{aligned}$$

by our choice of δ, and since $r = \gamma/2 > \delta/25$.

Case 4: $\widehat{\mu}_1 \le \mu_1^* \le \widehat{\mu}_2 \le \mu_2^*$ This case is symmetric to Case 3, and so we omit it.

Case 5: $\mu_1^* \le \mu_2^* \le \widehat{\mu}_1 \le \widehat{\mu}_2$ In this case, we proceed as in the proof of Theorem A.2. If the Gaussians were sufficiently skinny, then by the same logic as in the proof of Theorem A.2, there is exactly one zero crossing. The lemma then follows in this case by Theorem A.1.

Case 6: $\widehat{\mu}_1 \le \widehat{\mu}_2 \le \mu_1^* \le \mu_2^*$ This case is symmetric to Case 5.

This completes the proof. □

C.7. Proof of Lemma 4.6

We also show that no terribly divergent behavior can occur. Formally, we show that if the true means are within some bounded box, then the generators will never leave a slightly larger box.

Proof of Lemma 4.6. If $\widehat{\mu} \in B(C)$, then since f is Lipschitz and η is sufficiently small, clearly $\widehat{\mu}' \in B(C)$. Thus, assume that there is an $i = 1, 2$ so that $|\widehat{\mu}_i| > C$, and let $\widehat{\mu}_1$ be the largest such i in magnitude. WLOG take $\widehat{\mu}_2 > 0$. In particular, this implies that $\widehat{\mu}_2 > \mu_i^*$ for all $i = 1, 2$. There are now 3 cases, depending on the position of $\widehat{\mu}_1$.

Case 1: $\widehat{\mu}_1 \ge \mu_2^*$**:** Here, as in Case 2 in Lemma 4.5, the optimal discriminator is of the form $(-\infty, r]$ for some $r \le \widehat{\mu}_1, \widehat{\mu}_2$. In particular, the update step will be

$$\widehat{\mu}_i' = \widehat{\mu}_i - \eta e^{-(r-\widehat{\mu}_i)^2/2} < \widehat{\mu}_i\ .$$

Thus, in this case our update moves us in the negative direction. By our choice of η, this implies that $0 \le \widehat{\mu}_2' < \widehat{\mu}_2'$. Moreover, since $\widehat{\mu}_1 \ge \mu_2^*$, this implies that $|\widehat{\mu}_1| \le C$, and thus $|\widehat{\mu}_1'| \le 2C$. Therefore in this case we stay within the box.

Case 2: $\mu_1^* \le \widehat{\mu}_1 \le \mu_2^*$**:** As in Case 1, we know that $\widehat{\mu}_1$ cannot leave the box after a single update, as $|\widehat{\mu}_1| \le C$. Thus it suffices to show that $\widehat{\mu}_2$ moves in the negative direction. By Lemma C.6, we know there is a discriminator interval at $-\infty$, and there is no discriminator interval at ∞. Moreover, in this case, we know that $F(\widehat{\mu}, x) \ge 0$ for all $x \ge \widehat{\mu}_2$. Thus, all discriminators must be to the left of $\widehat{\mu}_2$. Therefore, the update to $\widehat{\mu}_2$ is given by

$$\begin{aligned}&\widehat{\mu}_2' = \widehat{\mu}_2\\&- \eta\left(e^{-(r_2-\widehat{\mu}_2)^2/2} - e^{-(\ell_2-\widehat{\mu}_2)^2/2} + e^{-(r_1-\widehat{\mu}_2)^2/2}\right)\ ,\end{aligned}$$

for some $r_1 \le \ell_2 \le r_2 \le \widehat{\mu}_2'$. Clearly this update has the property that $0 \le \widehat{\mu}_2' < \widehat{\mu}_2$, and so the new iterate stays within the box.

Case 3: $\widehat{\mu}_1 \le \mu_2^*$ In this case we must prove that neither $\widehat{\mu}_1$ nor $\widehat{\mu}_2$ leave the box. The two arguments are symmetric, so we will focus on $\widehat{\mu}_2$. Since η is small, we may assume that $\widehat{\mu}_2 > 3C/2$, as otherwise $\widehat{\mu}_2'$ cannot leave the box. As in Case 1, it suffices to show that the endpoints of the discriminator intervals are all less than $\widehat{\mu}_2$. But in this case, we have that for all

$x \geq \widehat{\mu}_2'$, the value of the true distribution at x is at most $2e^{-(x-C)^2/2}$, and the value of the discriminator is at $e^{-(x-\widehat{\mu}_2')^2/2} \geq e^{-(x-1.5C)^2/2}$. By direct calculation, this is satisfied for any choice of C satisfying $2e^{5C^2/8} < e^{3C^2/4}$, which is satisfied for $C \geq 3$.

□

D. Single Gaussian

Although our proof of the two Gaussian case implies the single Gaussian case, it is possible to prove the single Gaussian case in a somewhat simpler fashion, while still illustrating several of the high-level components of the overall proof structure. Therefore, we sketch how to do so, in hopes that it provides additional intuition for the proof for a mixture of two Gaussians.

In order to prove convergence, we can use the following.

1. The fact that the gradient is only discontinuous on a measure 0 set of points.
2. An absolute lower bound on the magnitude of the gradient from below over all points that are not close to the optimal solution that we might encounter over the course of the algorithm
3. An upper bound on how much the gradient can change if we move a certain distance.

Then, as long as we take steps that are small enough to guarantee that the gradient never changes by more than half the absolute lower bound, we will get by Lemma 4.2 that we always make progress towards the optimum solution in function value unless we are already close to the optimal solution.

The proof of these facts is substantially simplified in the single Gaussian case. Suppose we have a true univariate Gaussian distribution with unit variance and mean μ^*, along with a generator distribution with unit variance and mean $\widehat{\mu}$. Then the optimal discriminator for this pair of distributions starts at the midpoint between their means and goes in the direction of the true distribution off to ∞ or $-\infty$. Therefore, unless the generator mean is within one step length of the true mean, it cannot move away from the true mean. One can also argue that the gradient of $\widehat{\mu}$ with respect to the optimal discriminator (ie., the gradient of total variation distance) is only discontinuous when $\widehat{\mu} = \mu^*$, and has magnitude roughly $e^{(\widehat{\mu}-(\widehat{\mu}+\mu^*)/2)^2/2}$ for $\widehat{\mu} \neq \mu^*$. This implies the first two items. For the last item, note that the midpoint $e^{z^2/2}$, which implies the gradient is Lipschitz as long as we are not at the optimal solution, which gives bounds on how much the gradient can change if we move a certain distance.

The preceding discussion implies convergence for an appropriately chosen step size, and all this can be made fully quantitative if one works out the quantitative versions of the statements in the preceding argument.

This analysis is simpler the the two Gaussians analysis in several respects. In particular, the proofs of the second two items are substantially more involved and require many separate steps. For example, in the two Gaussian case, the gradient can be 0 if mode collapse happens, so we have to directly prove both that mode collapse does not happen and that the gradient is large if mode collapse doesn't happen and we aren't too close to the optimal solution, which is a substantially more involved condition to prove. Additionally, the gradient in the two Guassian case does not seem to be Lipschitz away from the optimum like it is in the single Gaussian case. Instead, we will have to use a weaker condition which is considerably more difficult to reason about. This is further complicated by the fact that the optimal discriminators can move in a discontinuous fashion when we vary the generator means.