# Mixture-GAN
# Progress Report

**Jaspal Singh**
Department of Computer Science
Yale University
New Haven, CT 06511
jaspal.singh@yale.edu

**Alexander Tong**
Department of Computer Science
Yale University
New Haven, CT 06511
alexander.tong@yale.edu

## 1 Summary of Current Progress

We originally set out to answer the question of how GANs perform on biased data. We defined biased data in terms of the weighted mixture of two gaussians. Precise definitions can be found in Section 3.1. In effect, we looked at the Li et al. [1] with an optimal discriminator, because it is a simple model that captures the convergence effects of GANs. Specifically, we were interested in how the relative mixture of the two gaussians affects convergence rates under this model, and hopefully by extension, to more realistic models. As a reminder, we were interested in the following generator:

$$\mathcal{G} = \left\{ \alpha\mathcal{N}(\mu_1, 1) + (1-\alpha)\mathcal{N}(\mu_2, 1) \mid \mu_1, \mu_2 \in \mathbb{R} \text{ and } \alpha \in [0, 1] \right\} \tag{1}$$

We consider the following two cases of **Known** $\alpha$ and **Learned** $\alpha$. The known case is assuming apriori knowledge of the value of $\alpha$. In a more realistic setting this is a very weak form of supervision, as it is often easier to know approximately the overall statistics of a dataset than labels for each datapoint. In the learned case we learn the value of $\alpha$ using first order gradient descent. In both of these cases, we consider dynamics under an optimal discriminator and a first order discriminator (potentially with unrolling).

A summary of progress thus far,

1. Implementation of code reproducing the results of Li et al. (2018)
2. Evaluation of the convergence probability for varying alpha values
3. Extension of Li et al. Lemma 4.3 and 4.4 for the fixed $\alpha$ optimal discriminator dynamics

## 2 Experimental Results

**Fixed $\alpha$ Under Optimal Discriminator Dynamics** After implementing the code and replicating the results of Li et al., we first explored known $\alpha$ under the optimal discriminator case. This we consider to be the simplest extension to their results, and the one that is most likely to work. In Figure 1 we can see that even with a large skew in $\alpha$ dynamics under the optimal discriminator always converge to the correct value.

**Rate of convergence** Figure 2 shows the the number of iterations it takes for the means $\widehat{\mu}$ to get within some $\delta$ distance to the optimal $\mu^*$. Each iteration consists of first order updates over $\widehat{\mu}$ and optimal updates over discriminator parameters $l, r$. As can be seen from the plot, though optimal discriminator always converges, its speed of convergence seems to be inversely proportional to $\alpha(1-\alpha)$. We can further question here on how to make the convergence rate independent of $\alpha$ by setting the $\eta$ parameter accordingly.

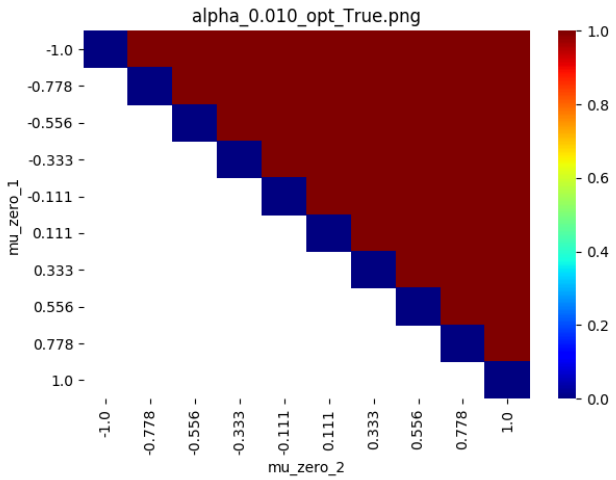



Figure 1: Convergence probabilities over initial values of $\widehat{\mu}$ with $\alpha = 0.01$

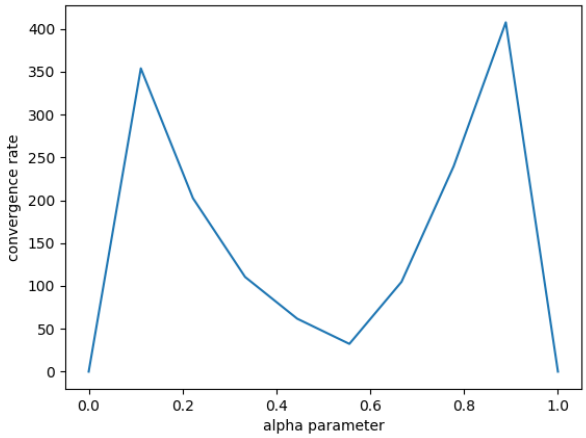
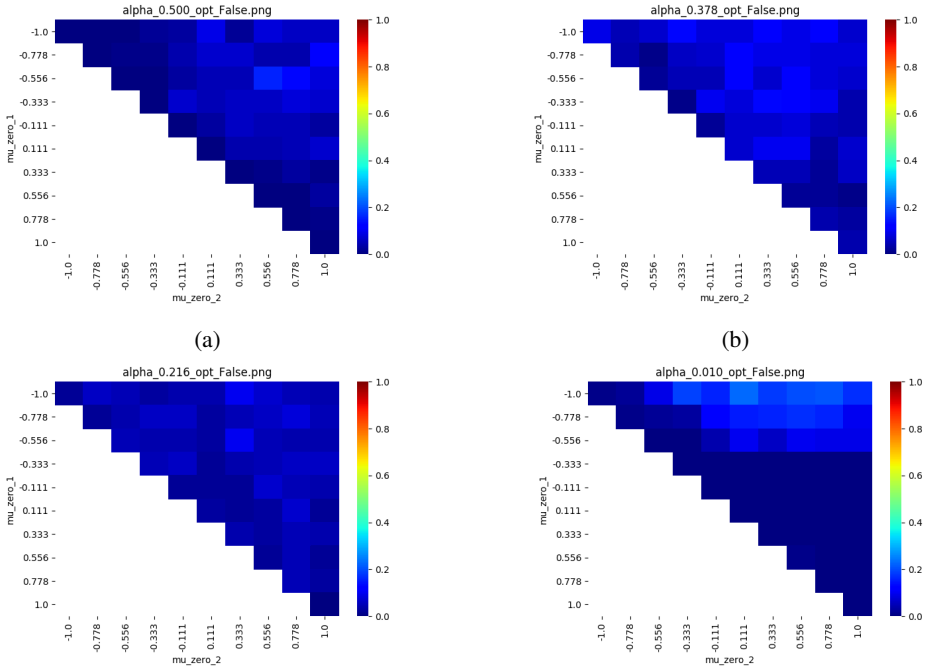



Figure 2: Plot of average rate of convergence for optimal discriminator dynamics vs known $\alpha$ with initialization $\mu^* = (-0.5, 0.5), \alpha \in Linspace(0, 1, 10)$ and the tolerance parameter $\delta = 0.01$

**Fixed $\alpha$ Under First Order Dynamics** Fixed $\alpha$ under first order dynamics can be shown for various levels of $\alpha$ in Figure 3. For $\alpha = 0.5$ we roughly recover the results of Li et al. Figure 2a in terms of convergence probabilities. We take all parameters except for the distribution of initial discriminator intervals which is referred to as "at random". We suspect this leads to small differences in convergence probabilities.

From these plots suspect the following. Slightly uneven weighting gives higher success probability (See Figure 3b) than perfectly even weighting. At extremely low values of $\alpha$ we only succeed on initializations of $\alpha$ where the smaller mean is initialized to below its optimum i.e. $\widehat{\mu_1} < \mu_1^*$ when $\alpha_1 << \alpha_2$. This was unexpected and currently not fully explained.

**Learned $\alpha$ Under First Order Dynamics** The next case to consider is that where $\alpha$ is learned along with $\mu, l,$ and $r$. We implemented this case using the same learning rate for $\alpha$ as $\mu, l,$ and $r$. This leads to diverging behavior in the value of $\alpha$ as seen in Figure 4. This implies that simply adding $\alpha$ to the list of learned parameters and treating it in the same way does not work in this model. This draws into question the usefulness of this model for studying behavior of generative models over varying mixture parameters. The next steps in this direction are (1) to figure out how to constrain $\alpha$




(a) (b)

(c) (d)

Figure 3: Shows the convergence probabilities for various settings of $\alpha$ under first order dynamics. Note that the value of $\alpha$ refers to the weight of the $\mu_1^*$, so plots (b-d) are asymmetric. We average each experiment over 100 choices of initial discriminator intervals sampled from $U(-2, 2)$ then sorted. Note that these values of $\alpha$ are selected from $GeomSpace(0.01, 0.5, 10)$.

to known values (perhaps as using a $\log$ transform, and (2) tune the learning rate such that there are fewer wild oscillations of $\widehat{\alpha}$ over training time with the optimal discriminator.

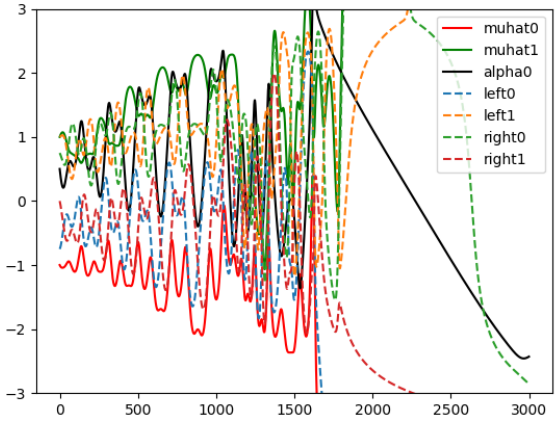



Figure 4: Diverging $\alpha_0$ for first order learned $\alpha$ with initialization $\widehat{\alpha} = (0.5, 0.5)$, $\alpha^* = (0.5, 0.5)$, $\widehat{\mu}_0 = (-1, 1)$, $\mu^* = (-0.5, 0.5)$

**Other Notes** Since we are considering the one dimensional case, when we look at the known $\alpha$ condition we can consider knowing the order of $\alpha$, i.e. whether the gaussian with larger $\alpha^*$ has larger or smaller $\mu^*$ value. In our experiments we fixed the case that our initialization has the correct

ordering, i.e. we do not have to flip the order of the gaussians (as we are not learning $\alpha$). This simplifies the problem but does not generalize to higher dimensions or number of gaussians. We tried the optimal discriminator with the incorrect initialization order and this failed 100% of the time.

# 3 Theoretical Analysis: Optimal Discriminator Dynamics for Fixed Alpha

Given that our experiments show fixed $\alpha$ under optimal discriminator dynamics always converges independent of the value $\alpha$, we wanted to see if we can extend the convergence proof of Li et al. [1] for the fixed $\alpha$ setting as well. We have been successful in extending two of Li et al.'s theorem for this scenario, which are mentioned below. We are still working on extending 2 other theorems, which should complete the convergence proof.

## 3.1 Definitions and Notations

- $\mu = (\mu_1, \mu_2) \in \mathbb{R}^2$ and $\alpha \in [0, 1]$
- $G_\mu(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2}$
- $G'_{\mu,\alpha}(x) = \alpha G_\mu(x)$
- $L_{\mu^*}(\widehat{\mu}, l, r) = \left( \int_I G_{\mu^*}(x) - G_{\widehat{\mu}}(x) dx \right)$ where $I = [l_1, r_1] \cup [l_2, r_2]$
- $L'_{\mu^*,\alpha}(\widehat{\mu}, l, r) = \left( \int_I G'_{\mu^*,\alpha}(x) - G'_{\widehat{\mu},\alpha}(x) dx \right)$ where $I = [l_1, r_1] \cup [l_2, r_2]$
- $f_{\mu^*}(\mu) = \max_{l,r} L_{\mu^*}(\mu, l, r)$
- $f'_{\mu^*,\alpha}(\mu) = \max_{l,r} L'_{\mu^*,\alpha}(\mu, l, r)$
- Rect($\delta$) = $\{\mu : |\mu - \mu_i^*| < \delta$ for some $i, j\}$
- $\nabla_{\mu_j} f_{\mu^*}(\mu) = \frac{1}{\sqrt{2\pi}} \sum_{i=1,2} (e^{-(r_i - \widehat{\mu}_j)^2/2} - e^{-(l_i - \widehat{\mu}_j)^2/2})$
- $\nabla_{\mu_j} f'_{\mu^*,\alpha}(\mu) = \alpha_j \nabla_{\mu_j} f_{\mu^*}(\mu)$
- $\nabla_\mu f_{\mu^*}(\mu) = \begin{bmatrix} \nabla_{\mu_1} f_{\mu^*}(\mu) \\ \nabla_{\mu_2} f_{\mu^*}(\mu) \end{bmatrix}$
  we can define $\nabla_\mu f_{\mu^*,\alpha}(\mu)$ similarly.
- $B(C)$ represents the box of side length $C$ around origin.
- $Sep(\gamma) = \{(v_1, v_2) \in \mathbb{R}^2 : |v_1 - v_2| > \gamma\}$
- Let $m = \min(\alpha, 1 - \alpha)$ and $M = \max(\alpha, 1 - \alpha)$
- $||\ ||$ represents the $L2$ norm.

**Lemma 1.** $m||\nabla f_{\mu^*}(\mu)|| \leq ||\nabla f'_{\mu^*,\alpha}(\mu)|| \leq M||\nabla f_{\mu^*}(\mu)||$

*Proof.*

$$\begin{aligned} ||\nabla f'_{\mu^*,\alpha}(\mu)|| &= || \begin{bmatrix} \nabla_{\mu_1} f'_{\mu^*,\alpha}(\mu) \\ \nabla_{\mu_2} f'_{\mu^*,\alpha}(\mu) \end{bmatrix} || \\ &= || \begin{bmatrix} \alpha \nabla_{\mu_1} f_{\mu^*}(\mu) \\ (1-\alpha) \nabla_{\mu_2} f_{\mu^*}(\mu) \end{bmatrix} || \end{aligned}$$

The inequality follows. □

**Lemma 2.** *Let $C \geq 1 \geq \gamma \geq \delta > 0$, $\mu \notin$ Rect$(0)$, $\mu, \mu^* \in B(C)$ and $\mu^* \in Sep(\gamma)$. Then there exist $K = \Omega(1)(\delta e^{-C^2}/C)^{O(1)}$ so that $||\nabla f'_{\mu^*,\alpha}|| \geq mK$*

*Proof.* This follows directly from Lemma 4.3 in the paper [1] and Lemma 1. □

**Lemma 3.** *Let $C \geq 1$ and $\gamma \geq \delta > 0$ so that $\delta$ is sufficiently small. Let $\mu^*, \mu, \mu'$ be such that $L(\mu, \mu') \cap opt(\delta) = \phi$, $\mu^* \in Sep(\gamma)$, $\mu, \mu' \in Sep(\delta)$ and $\mu^*, \mu, \mu' \in B(C)$. Let $K = \Omega(1)(\delta \exp^{-C^2/C})^{O(1)}$ for which Lemma 2 holds and $||\mu' - \mu|| \leq \Omega(1)(\delta \exp^{-C^2}/C)^{O(1)}$, then we get*

$$||\nabla f'_{\mu^*,\alpha}(\mu) - \nabla f'_{\mu^*,\alpha}(\mu')|| \leq (M/2)K \leq (M/2m)||f'_{\mu^*,\alpha}(\mu)||$$

*Proof.*

$$\begin{aligned}
||\nabla f'_{\mu^*,\alpha}(\mu) - \nabla f'_{\mu^*,\alpha}(\mu')|| &= || \begin{bmatrix} \alpha(\nabla_{\mu_1} f_{\mu^*}(\mu) - \nabla_{\mu_1} f_{\mu^*}(\mu')) \\ (1-\alpha)(\nabla_{\mu_2} f_{\mu^*}(\mu) - \nabla_{\mu_2} f_{\mu^*}(\mu')) \end{bmatrix} || \\
&\leq M||\nabla f_{\mu^*}(\mu) - \nabla f_{\mu^*}(\mu')|| \qquad \text{(by Lemma 1)} \\
&\leq (Mm/2m)K \qquad \text{(by Lemma 4.4 in [1])}
\end{aligned}$$

□

## 4 Some Doubts and Points for Discussion

1. Unable to prove Li et al. Lemma 4.2
2. Is Li et al. Lemma C.6 correct?
3. How can we relate this model to something more realistic?

## 5 Process Notes

For convenience, define $F(\alpha^*, \widehat{\alpha}, \mu^*, \widehat{\mu}, x)$ as follows:

$$F(\alpha^*, \widehat{\alpha}, \mu^*, \widehat{\mu}, x) = \frac{1}{\sqrt{2\pi}} \sum_{j=1,2} \alpha_j^* e^{-(x-\mu_j^*)^2/2} - \widehat{\alpha}_j e^{-(x-\widehat{\mu}_j)^2/2} \tag{2}$$

Looking at the problem more precisely, we consider the loss function $L(\mu, l, r, \alpha)$. Note that we can consider the case where we know $\alpha$, or we would like to learn $\alpha$.

$$\begin{aligned}
L(\mu, l, r, \alpha) &= \mathbb{E}_{x \sim G_{\mu^*}}[D(x)] + \mathbb{E}_{x \sim G_\mu}[1 - D(x)] &(3)\\
&= \left( \int_I G_{\mu^*}(x) - G_{\widehat{\mu}}(x) dx \right) + 1, &(4)\\
&\text{Where } I = [l_1, r_1] \cup [l_2, r_2]. \text{ We then have} \\
&= \left( \sum_{i=1,2} \int_{l_i}^{r_i} G_{\mu^*}(x) - G_{\widehat{\mu}}(x) dx \right) + 1, &(5)\\
&= \frac{1}{\sqrt{2\pi}} \sum_{i=1,2} \sum_{j=1,2} \int_{l_i}^{r_i} \alpha_j^* e^{-(x-\mu_j^*)^2/2} - \widehat{\alpha}_j e^{-(x-\widehat{\mu}_j)^2/2} dx + 1, &(6)\\
&= \frac{1}{\sqrt{2\pi}} \sum_{i=1,2} \sum_{j=1,2} \int_{l_i}^{r_i} F(\alpha^*, \widehat{\alpha}, \mu^*, \widehat{\mu}, x) dx + 1, &(7)
\end{aligned}$$

We next examine the partial derivatives of $L$ in order to understand the first order dynamics. We start with $\frac{\partial}{\partial l_i} L$:

$$\frac{\partial}{\partial l_i} L = \frac{\partial}{\partial l_i} \left( \frac{1}{\sqrt{2\pi}} \sum_{i=1,2} \sum_{j=1,2} \int_{l_i}^{r_i} \alpha_j^* e^{-(x-\mu_j^*)^2/2} - \widehat{\alpha}_j e^{-(x-\widehat{\mu}_j)^2/2} dx + 1 \right) \tag{8}$$

Which by Leibniz integral rule,

$$= -\frac{1}{\sqrt{2\pi}} \sum_{j=1,2} \alpha_j^* e^{-(l_i-\mu_j^*)^2/2} - \widehat{\alpha}_j e^{-(l_i-\widehat{\mu}_j)^2/2} \tag{9}$$

$$= -F(\alpha^*, \widehat{\alpha}, \mu^*, \widehat{\mu}, l_i) \tag{10}$$

Similarly for $\frac{\partial}{\partial r_i} L$:

$$\frac{\partial}{\partial r_i} L = \frac{\partial}{\partial r_i} \left( \frac{1}{\sqrt{2\pi}} \sum_{i=1,2} \sum_{j=1,2} \int_{l_i}^{r_i} \alpha_j^* e^{-(x-\mu_j^*)^2/2} - \widehat{\alpha}_j e^{-(x-\widehat{\mu}_j)^2/2} dx + 1 \right) \tag{11}$$

Which by Leibniz integral rule,

$$= \frac{1}{\sqrt{2\pi}} \sum_{j=1,2} \alpha_j^* e^{-(r_i-\mu_j^*)^2/2} - \widehat{\alpha}_j e^{-(r_i-\widehat{\mu}_j)^2/2} \tag{12}$$

$$= F(\alpha^*, \widehat{\alpha}, \mu^*, \widehat{\mu}, r_i) \tag{13}$$

Next, $\frac{\partial}{\partial \widehat{\mu}_j} L$ is a little more involved:

$$\frac{\partial}{\partial \widehat{\mu}_j} L = \frac{\partial}{\partial \widehat{\mu}_j} \left( \frac{1}{\sqrt{2\pi}} \sum_{i=1,2} \sum_{j=1,2} \int_{l_i}^{r_i} \alpha_j^* e^{-(x-\mu_j^*)^2/2} - \widehat{\alpha}_j e^{-(x-\widehat{\mu}_j)^2/2} dx + 1 \right) \tag{14}$$

Which by Leibniz integral rule,

$$= \widehat{\alpha}_j \frac{1}{\sqrt{2\pi}} \sum_{i=1,2} e^{-(r_i-\widehat{\mu}_j)^2/2} - e^{-(l_i-\widehat{\mu}_j)^2/2} \tag{15}$$

Finally, in the case of learning $\widehat{\alpha}$ by first order methods it is important to calculate the loss gradient with respect to $\alpha$, $\frac{\partial}{\partial \widehat{\alpha}_j} L$.

$$\frac{\partial}{\partial \widehat{\alpha}_j} L = \frac{\partial}{\partial \widehat{\alpha}_j} \left( \frac{1}{\sqrt{2\pi}} \sum_{i=1,2} \sum_{j=1,2} \int_{l_i}^{r_i} \alpha_j^* e^{-(x-\mu_j^*)^2/2} - \widehat{\alpha}_j e^{-(x-\widehat{\mu}_j)^2/2} dx + 1 \right) \tag{16}$$

$$= -\frac{1}{\sqrt{2\pi}} \sum_{i=1,2} \int_{l_i}^{r_i} e^{-(x-\widehat{\mu}_j)^2/2} dx \tag{17}$$

Note that most of these can easily be formulated in terms of the probability distribution function (pdf) or cumulative distribution function (cdf) of a standard normal distribution.

**Implementation Discoveries** While implementing Li et al. we discovered a few sticking points. The most major was how to actually implement an efficient optimal discriminator. Essentially, finding the optimal discriminator values boils down to finding the zeros of the loss function. With two gaussians (even with $\alpha \neq \frac{1}{2}$), the loss function has at most 3 zeros. The problem is how to search and find all three of the zeros not knowing anything else about the loss function. This is far from optimal. Currently, we use a brentq function to find zeros in between any of the 4 $\mu$ values $\widehat{\mu}_1, \widehat{\mu}_2, \mu_1^*, \mu_2^*$, and if there are any zeros within distance 1 of the minimum and maximimum $\mu$ values. Empirically, this seems to work well enough for the cases we have tried (i.e. all convergence probabilities of the optimal discriminator are exactly 1, not near to 1), but we cannot prove that this procedure finds all of the zeros of the function, and in fact does not in some cases.

## References


[1] Jerry Li, Aleksander Madry, John Peebles, and Ludwig Schmidt. On the Limitations of First-Order Approximation in GAN Dynamics. In *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweeden, 2018.