Andrew Li-Yang Liu
Nakul Verma
COMS 4774 Unsupervised Learning
Columbia University
November 22, 2022

# UML HW4

## Generalization and equilibrium in GANs:

Background on GANs:
Generative adversarial networks (GANs) are generative models that involve a 2-player mini-max game, whereby a discriminator function tries to maximally differentiate between samples from the true distribution and samples generated by the generator and where the generator tries to minimize the cost by fooling the discriminator. More concretely, let $\{G_u, u \in \mathcal{U}\}$ denote the class of generators $G_u: \mathbb{R}^l \to \mathbb{R}^d$, with $u \in \mathbb{R}^p$ being a vector of $p$ parameters and $\mathcal{U}$ is the space of generator parameters (WLOG, assume $\mathcal{U}$ is a subset of the unit ball). The generator defines a distribution $\mathcal{D}_{G_u}$. To sample from $\mathcal{D}_{G_u}$, simply sample $h$ from a $l$-dimensional spherical Gaussian, then apply $x = G_u(h)$. Similarly, let $\{D_v, v \in \mathcal{V}\}$ be the class of discriminators (also with $p$ parameters) $D_v: \mathbb{R}^d \to [0,1]$, which outputs the probability of $x$ being from $D_{real}$. The objective is:

$$\min_{u \in \mathcal{U}} \max_{v \in \mathcal{V}} \mathbb{E}_{x \sim \mathcal{D}_{real}}[\phi(D_v(x)] + \mathbb{E}_{x \sim \mathcal{D}_{G_u}}[\phi(1 - D_v(x)]$$

Where $\phi$ is a measuring function, which must be concave to guarantee that when $\mathcal{D}_{real} = \mathcal{D}_G$, the optimal strategy of the discriminator is to randomly guess by outputting $\frac{1}{2}$ and the final cost is $2\phi\left(\frac{1}{2}\right)$.

One notes that, in the case of infinite samples and where the discriminator is chosen optimally in a large set of functions containing all neural nets (which amounts to $D(x) = \frac{P_{real}(x)}{P_{real}(x)+P_G(x)}$), performing this optimization is equivalent to having the generator learn the underlying true distribution $D_{true}$. For example, in the case where $\phi(t) = \log(t)$, the optimization minimizes the Jensen-Shannon (JS) divergence $d_{JS}(\mu, \nu) = \frac{1}{2}(KL(\mu \| \frac{\mu+\nu}{2}) + KL(\nu \| \frac{\mu+\nu}{2}))$ between $\mathcal{D}_{real}$ and $\mathcal{D}_G$. Other choices of $\phi$ and discriminator lead to different distance functions being minimized (e.g. Wasserstein distance).

Summary of key contributions:
The paper makes several key contributions to the theory of GANS. First, it demonstrates that using previous notions of distances between distributions (e.g. KL divergence, Wasserstein distance), generalization is not guaranteed (i.e. a generator can win even when $D_G$ and $D_{real}$

are arbitrarily far). The authors then propose a new metric on distributions known as the neural net distance, and show that when the generator wins, generalization occurs with moderate sample complexity.

The authors then explore the existence of equilibria by considering mixtures of generators. Since it is known that every probability density can be approximated by an infinite mixture of Gaussians, it is clear that a mixture of infinitely many generators should win the GAN game (and guarantee an equilibrium). They then show that even a finite mixture of generators well-approximates the performance of the infinite mixture. This allows them to guarantee the existence of a pure approximate equilibrium. Finally, the authors propose a MIX+GAN, which is shown to improve GAN performance on CIFAR dataset.

Key technical results and proofs/proof sketches:
Note: In all results, the authors assume that the generators and discriminators are both $L$-Lipschitz with respect to parameters.

**Definition: (Generalization)**
Given $\widehat{\mathcal{D}}_{real}$ (empirical true distribution with $m$ samples), $\mathcal{D}_G$ generalizes under $d(.,.)$ with error $\epsilon$ if the following holds with high probability:

$$\left|d(\mathcal{D}_{real},\mathcal{D}_G)-d\big(\widehat{\mathcal{D}}_{real},\widehat{\mathcal{D}}_G\big)\right|\leq\epsilon$$

**Lemma: Negative generalization results for JS and Wasserstein divergence**
Let $\mu$ be uniform Gaussian distributions $N\left(0,\frac{1}{d}I\right)$ and $\hat{\mu}$ be its empirical versions with $m$ samples. Then $d_{JS}(\mu,\hat{\mu})=\log 2$ and $d_W(\mu,\hat{\mu})\geq 1.1$.

Consequences:

1. If $\mathcal{D}_{real}=\mathcal{D}_G=\mu$, then $d_W(\mathcal{D}_{real},\mathcal{D}_G)=0$ but $d\big(\widehat{\mathcal{D}}_{real},\widehat{\mathcal{D}}_G\big)>1$ for polynomial number of samples (i.e. not generalizable).
2. If $\mathcal{D}_{real}=\mu$ and $\mathcal{D}_G=\widehat{\mathcal{D}}_{real}=\hat{\mu}$, then since $\mathcal{D}_G$ is discrete with finite support, with enough examples we have $\widehat{\mathcal{D}}_G\cong\mathcal{D}_G$ so $d_W\big(\widehat{\mathcal{D}}_{real},\widehat{\mathcal{D}}_G\big)\cong 0$ but $d_W(\mathcal{D}_{real},\mathcal{D}_G)>1$.

The same argument applies to JS divergence.

Proof sketch: For the Wasserstein distance proof, simply invoke standard concentration and union bounds. Given $x_1,\dots,x_m$ and $y\sim N\left(0,\frac{1}{d}I\right)$ we have that $P\big[\|y-x_i\|\geq 1.2\ \ \forall i\in[m]\big]\geq 1-m\exp\big(\Omega(d)\big)\geq 1-o(1)$. Then apply the earth-mover interpretation of Wasserstein distance to obtain $d_W(\mu,\hat{\mu})\geq 1.2P\big[\|y-x_i\|\geq 1.2\ \ \forall i\in[m]\big]\geq 1.1$.

**Definition: ($\mathcal{F}$-distance and neural net distance)**
The $\mathcal{F}$-distance is as follows:

$$d_{\mathcal{F},\phi}(\mu,\nu)=\sup_{D\in\mathcal{F}}\mathbb{E}_{x\sim\mu}\big[\phi\big(D(x)\big)\big]+\mathbb{E}_{x\sim\nu}[1-D(x)]-2\phi\left(\frac{1}{2}\right)$$

This notion of distance is general in the sense that, with specific choices of $\phi$ and $\mathcal{F}$, one recovers the JS and Wasserstein distances. We define the neural net distance as simply the $\mathcal{F}$-distance with $\phi(t) = t$ and $\mathcal{F}$ being the set of neural networks.

**<u>Theorem: (positive generalization results for $\mathcal{F}$-distance)</u>**

$\exists c > 0$ such that when $m \geq \frac{cp\Delta^2 \log\left(\frac{LL_\phi p}{\epsilon}\right)}{\epsilon^2}$, we have that with probability $\geq 1 - \exp(-p)$:

$$\left|d_{\mathcal{F},\phi}(\hat{\mu}, \hat{v}) - d_{\mathcal{F},\phi}(\mu, v)\right| \leq \epsilon$$

i.e. if $d_{\mathcal{F},\phi}(\hat{\mu}, \hat{v})$ is small then $d_{\mathcal{F},\phi}(\mu, v)$ should also be small.

Proof sketch:
It can be shown that for every discriminator $D_v$:

$$\left|\mathbb{E}_{x\sim\mu}[\phi(D_v(x))] - \mathbb{E}_{x\sim\widehat{\mu}}[\phi(D_v(x))]\right| \leq \frac{\epsilon}{2}$$
$$\left|\mathbb{E}_{x\sim v}[\phi(1 - D_v(x))] - \mathbb{E}_{x\sim\widehat{v}}[\phi(1 - D_v(x))]\right| \leq \frac{\epsilon}{2}$$

To show these inequalities, construct a new set $\mathcal{X}$ so that $d(v, \mathcal{X}) \leq \frac{\epsilon}{8} LL_\phi \ \forall v \in \mathcal{V}$ (i.e. an $\frac{\epsilon}{8} LL_\phi$-cover). By Chernoff bound we have $P\left[\left|\mathbb{E}_{x\sim\mu}[\phi(D_v(x))] - \mathbb{E}_{x\sim\tilde{\mu}}[\phi(D_v(x))]\right| \geq \frac{\epsilon}{4}\right] \leq 2\exp\left(-\frac{\epsilon^2 m}{2\Delta^2}\right)$. So for some large $C$, if $m \geq \frac{Cp\Delta^2 \log\left(\frac{LL_\phi p}{\epsilon}\right)}{\epsilon^2}$, then we have that $\forall v \in \mathcal{X}$, $\left|\mathbb{E}_{x\sim\mu}[\phi(D_v(x))] - \mathbb{E}_{x\sim\tilde{\mu}}[\phi(D_v(x))]\right| \geq \epsilon/4$ with probability $\geq 1 - \exp(-p)$ (using union bound). Combining this with the $\frac{\epsilon}{8} LL_\phi$-cover inequality gives the result.

The two inequalities then give rise to $\left|d_{\mathcal{F},\phi}(\hat{\mu}, \hat{v}) - d_{\mathcal{F},\phi}(\mu, v)\right| \leq \epsilon$.

However, it is worth noting that $d_{\mathcal{F}}(\mu, v)$ can be small even if the two distributions are not close, as seen in the following corollary:

**<u>Corollary</u>**: (Low-capacity discriminators cannot detect lack of diversity)

Let $\hat{\mu}$ be the empirical version of $\mu$ with $m$ samples. $\exists c > 0$ such that if $m \geq \frac{cp\Delta^2 \log\left(\frac{LL_\phi p}{\epsilon}\right)}{\epsilon^2}$, then with probability at least $1 - \exp(-p)$ we have that $d_{\mathcal{F},\phi}(\mu, \hat{\mu}) \leq \epsilon$. In other words, a discriminator of size $p$ is unable to distinguish the difference between $\mu$ and a distribution with support $\tilde{O}\left(\frac{p}{\epsilon^2}\right)$.

**<u>On the existence of equilibria:</u>**

Just as zero gradient is necessary to terminate gradient descent, achieving an equilibrium is necessary to terminate the min-max game.

Define the payoff of the game as $F(u,v) = \mathbb{E}_{x\sim\mathcal{D}_{real}}\big[\phi\big(D_v(x)\big)\big] + \mathbb{E}_{x\sim\mathcal{D}_G}\big[\phi\big(1 - D_v(x)\big)\big]$.

Von Neumann's min-max theorem states that if both players can be mixed strategies then the game has an equilibrium. Note that a mixed strategy for the generator is some distribution $S_u$ on $\mathcal{U}$, and for the discriminator it is a distribution $S_v$ on $\mathcal{V}$. More concretely, there is some $V$ and some pair of mixed strategies $(S_u, S_v)$ such that

$$\mathbb{E}_{u\sim S_u}[F(u,v)] \leq V \;\; \forall v$$
$$\mathbb{E}_{v\sim S_v}[F(u,v)] \geq V \;\; \forall u$$

However, a distribution over the class of generators is an infinite mixture of them, which is infeasible in practice. The question then is if we can achieve a similar min-max solution using finitely many generators and discriminators. To this end, we define an $\epsilon$-approximate equilibrium.

**Definition: $\epsilon$-approximate equilibrium**

Mixed strategies $(S_u, S_v)$ is an $\epsilon$-approximate equilibrium if for some value $V$ we have:

$$\mathbb{E}_{u\sim S_u}[F(u,v)] \leq V + \epsilon \;\; \forall v$$
$$\mathbb{E}_{v\sim S_v}[F(u,v)] \geq V - \epsilon \;\; \forall u$$

If $S_u, S_v$ are pure strategies, then we call this equilibrium an $\epsilon$-approximate pure equilibrium.

**Theorem: ($\epsilon$-approximate equilibriums for finite mixture of Lipschitz generators/discriminators)**

Suppose $\phi$ is $L_\phi$-Lipschitz, bounded on $[-\Delta, \Delta]$, generators and discriminators are $L$-Lipschitz with respect to parameters and $L'$-Lipschitz with respect to inputs. If the generator can approximate any point mass (i.e. $\forall x$ and $\forall \epsilon > 0$, there is a generator such that $\mathbb{E}_{h\sim D_h}[\|G(h) - x\|] \leq \epsilon$), then $\exists C > 0$ such that $\forall \epsilon > 0$, exists $T = \frac{C\Delta^2 p \log\left(\frac{LL'L_\phi p}{\epsilon}\right)}{\epsilon^2}$ generators $G_{u_1}, \dots, G_{u_T}$. Let $S_u$ be uniform distribution on $u_i$ and $D$ a discriminator outputting only 1/2, then $(S_u, D)$ is an $\epsilon$-approximate equilibrium.

Proof:

First prove that $V = 2\phi\left(\frac{1}{2}\right)$. A possible strategy for the discriminator is to only output 1/2, in which case $F(u,v) = 2\phi\left(\frac{1}{2}\right)$, so $V \geq 2\phi\left(\frac{1}{2}\right)$. For the other bound, note that we assumed that $\forall x$ and $\forall \epsilon > 0$, there is a generator such that $\mathbb{E}_{h\sim D_h}\big[\big\|G_{x,\epsilon}(h) - x\big\|\big] \leq \epsilon$. For any $\xi > 0$, consider the mixture of generators: Sample $x \sim D_{real}$, then use the generator $G_{x,\xi}$. Denote $\mathcal{D}_\xi$ as the distribution generated by this mixture of generators. Through $d_W\big(\mathcal{D}_\xi, \mathcal{D}_{real}\big) \leq \xi$ and Lipschitz of discriminator $D_v$, we can obtain that:

$$\max_{v\in\mathcal{V}} \mathbb{E}_{x\sim\mathcal{D}_{real}}\big[\phi\big(D_v(x)\big)\big] + \mathbb{E}_{x\sim\mathcal{D}_\xi}\big[\phi\big(1 - D_v(x)\big)\big] \leq 2\phi\left(\frac{1}{2}\right) + O\big(L_\phi L'\xi\big)$$

Taking $\xi \to 0$ gives $V = 2\phi\left(\frac{1}{2}\right)$. This means that in this equilibrium the discriminator can only randomly guess, i.e. the discriminator can only output ½ and nothing else.

How to construct the generator? We use a probabilistic proof. Let $(S'_u, S'_v)$ be optimal von Neumann mixed strategies and $V$ the optimal value. Construct $\frac{\epsilon}{4}LL'L_\phi$-covers $V$ over discriminator parameters $\mathcal{V}$. If we draw $u_1, \dots, u_T$ from $S'_u$, then applying Chernoff bound eventually gives us that for any $v' \in \mathcal{V}$, we have $\mathbb{E}_{i\in[T]}[F(u_i, v')] \le 2\phi\left(\frac{1}{2}\right) + \epsilon$, i.e. the mixture of generators beats any discriminator. So by probabilistic argument, such a mixture of generator exists.

**Theorem: (combining mixture of generators into a single neural network)**

If the generator and discriminator in the mixture are both $k$-layer neural networks ($k \ge 2$) with $p$ parameters, and the last layer uses ReLU activation, then there exists $k+1$-layer neural networks of generators $G$ and discriminator $D$ with $O\left(\frac{\Delta^2 p^2 \log\left(\frac{LL'L_\phi p}{\epsilon}\right)}{\epsilon^2}\right)$ parameters, such that there exists an $\epsilon$-approximate pure equilibrium with value $2\phi\left(\frac{1}{2}\right)$.

Proof outline:

One needs to combine the mixture of generators into a single generator. The paper proposes using a multi-way selector on $G_{u_1}, \dots, G_{u_T}$ that randomly selects $i \in [T]$ and disables all but the $i$th generator.

**MIX+GANs experiments:**

Motivated by the fact that finite mixtures of generators guarantee existence of approximate equilibrium, the authors propose MIX+GANs, whereby one trains a mixture of $T$ generators and $T$ discriminators with same network architecture but different parameters. The optimization objective takes into account the learnable weights $w_{u_i}, w_{v_j}$ associated with each generator/discriminator:

$$\min_{\{u_i\},\{\alpha_{u_i}\}} \max_{\{v_j\},\{\alpha_{v_j}\}} \sum_{i,j\in[T]} w_{u_i} w_{v_j} F(u_i, v_j)$$

Where:

$$w_{u_i} = \frac{e^{\alpha_{u_i}}}{\sum_{k=1}^{T} e^{\alpha_{u_k}}}$$

There is also an entropy regularization term $R_{ent}\left(\{w_{u_i}, w_{v_i}\}\right) = -\frac{1}{T}\sum_{i=1}^{T}\left(\log(w_{u_i}) + \log(w_{v_i})\right)$ to promote more uniform weights.

The results show that MIX+GAN improves the inception score and decreases the neural network divergence of GANs on CIFAR-10 and CelebA datasets.

# Do GANs Learn The Distribution? Some Theory and Empirics

Summary:
GANs are known to suffer from mode collapse – the phenomenon where even if training achieves an optimal GAN cost, the generator probability distribution assigns near-zero probability to perfectly valid samples and assigns most of the probability on a few select modes. In other words, it is when the distribution has a small effective support compared to the support of the true distribution. In the case of facial image generation, for example, this leads to the generator repeatedly generating slight variations (near-duplicates) of the same faces while neglecting other samples.

This paper proposes a test for approximating the support size of discrete probability distributions given finite samples, presents evidence that well-known GANs learn low-support distributions, and proves the inability of encoder-decoder GAN architectures to mitigate mode collapse.

Birthday paradox support test:
One notes that the birthday paradox says that, given a discrete distribution of support $N$, a batch of $\sqrt{N}$ samples is likely to have a duplicate. This motivates the Birthday Paradox Test for approximating support size:

1. Sample a batch of s images from the generator
2. Use a measure of similarity to identify the 20 most similar pairs in the batch.
3. Visually check the pairs for duplicates.
4. Repeat

If the test shows that batches of size $s$ have duplicate images with high probability, one can suspect that the distribution has a support size of $\sim s^2$.

**Theorem 1:**
Given a discrete probability distribution $P$ on $\Omega$, if there exists $S \subset \Omega$ with $|S| = N$ such that $\sum_{s\in S} P(s) \geq \rho$, then the probability of encountering at least one collision among $M$ iid samples from $P$ is $\geq 1 - \exp\left(-\frac{(M^2-M)}{2N}\rho\right)$

Proof:

$$P[exists\ collision\ among\ M\ samples] \geq 1 - P[no\ collision\ among\ M\ samples\ in\ set\ S]$$
$$\geq 1 - 1 * \left(1 - \frac{\rho}{N}\right)\left(1 - \frac{2\rho}{N}\right) * \ldots * \left(1 - \frac{M-1}{N}\rho\right) \geq 1 - \exp\left(-\frac{(M^2-M)\rho}{2N}\right)$$

**Theorem 2:**

Given a discrete probability distribution $P$ on $\Omega$, if the probability of getting at least one collision in $M$ iid samples is $\gamma$, then for $\rho = 1 - o(1)$, $\exists S \subset \Omega$ such that $\sum_{s \in S} P(s) \geq \rho$ with $|S| \leq \frac{2M\rho^2}{\left(-3+\sqrt{9+\frac{24}{M}\ln\left(\frac{1}{1-\gamma}\right)}\right)-2M(1-\rho)^2}$.

**In other words, if one consistently sees collisions, then the distribution has a major component that is limited in support size but nonetheless has very high probability mass and is thus indistinguishable from the full distribution when sampling small number of samples.**

Proof:
Say $X_1, X_2, \ldots$ are iid sampled from discrete distribution $P$. Define $T = \inf\{t \geq 2, X_t \in \{X_1, X_2, \ldots, X_{t-1}\}\}$ to be collision time and use $\beta = \frac{1}{\Pr[T=2]} = \frac{1}{\sum_{X \in \Omega} P(X)^2}$ as a measure of uniformity of $P$. Prior work shows that $\Pr[T \geq M]$ can be bounded with $\beta$. When $\beta > 1000$ and $M \leq 2\sqrt{\beta}\ln\beta$, we have that $\Pr[T \geq M] \geq \exp\left(-\frac{M^2}{2\beta} - \frac{M^3}{6\beta^2}\right)$. This gives:

$$\beta \leq \frac{2M}{-3+\sqrt{9+\frac{24}{M}\ln\frac{1}{1-\gamma}}} = \beta^*$$

i.e. an upper bound for uniformity of $P$. Let $S \subset \Omega$ be the smallest set with $P(\Omega) \geq \rho$. Say $|S| = N$. Then we let $\frac{1}{\left(\frac{\rho}{N}\right)^2 N+(1-\rho)^2} \leq \beta^*$, in which case we get $N \leq \frac{2M\rho^2}{\left(-3+\sqrt{9+\frac{24}{M}\ln\frac{1}{1-\gamma}}\right)-2M(1-\rho)^2}$.

In the case of GANs, even though the distribution is continuous (and thus impossible to obtain exact duplicates), the birthday test still works if one looks for near-duplicates. The authors use the test in several experiments including the CelebA Dataset and CIFAR-10.

## Limitations of encoder-decoder GAN architectures

The authors theoretically demonstrate that even for encoder-decoder GAN architectures – which were meant to mitigate mode collapse – are unable to fully address the issue.

Recall the basic architecture of an encoder-decoder GAN (e.g. BiGAN). The generative player consists of a generator $G$ and an encoder $E$. The generator takes in a latent variable $z$ and outputs $G(z)$ whereas the encoder takes in a data sample $x$ and guesses the latent variable $E(x)$, providing us with two joint distributions over pairs of latent variables and data samples $(z, G(z))$ and $(E(x), x)$. The generative player must convince the discriminator that these two joint distributions are the same.

Ideally, we want the two distributions to eventually match, giving $p(z, x)$. Hence, the BiGAN objective is given as:

$$\min_{G,E}\max_{D}\left|\mathbb{E}_{x\sim\hat{\mu}}\phi\left(D\left(x,E(x)\right)\right)-\mathbb{E}_{z\sim\hat{\nu}}\phi\left(D(G(z),z)\right)\right|$$

Where $\hat{\mu}$ is the empirical distribution over data $x$ and $\hat{\nu}$ is a distribution over random seeds for latent variables, and $\phi$ is the concave measuring function. Once again, assume $\phi$ has values between $[-\Delta,\Delta],\Delta\geq 1$, and that it is $L_\phi$-Lipschitz. Assume the discriminators are $L$-Lipschitz with respect to trainable parameters.

Theorem 3:

There exists a generator $G$ of support $\frac{p\Delta^2\log^2\left(\frac{p\Delta L L_\phi}{\epsilon}\right)}{\epsilon^2}$ and encoder $E$ with at most $\tilde{d}$ non-zero weights, such that for all discriminators $D$ that are $L$-Lipschitz and have capacity $\leq p$ we have that:

$$\left|\mathbb{E}_{x\sim\mu}\phi\left(D\left(x,E(x)\right)-\mathbb{E}_{z\sim\nu}\phi\left(D(G(z),z)\right)\right|\leq\epsilon$$

**In other words, even though $E$ has small complexity and $G$ has small support (and thus is very far from the actual data distribution), the BiGAN objective is still arbitrarily small (i.e. training succeeded). Hence, the BiGAN encoder-decoder architecture does not effectively mitigate mode collapse.**

Proof:

The proof of Theorem 3 requires first proving two lemmas:

Lemma D.1: Let $G$ be a fixed generator and $D$ a fixed discriminator. Then:

$$\mathbb{E}_{z\sim\nu}\phi\left(D(G(z),z)\right)=\mathbb{E}_{T\sim\mathcal{T}_{nc}}\mathbb{E}_{z\sim T}\phi\left(D(G(z),z)\right)$$

Lemma D.2 (Concentration of good generators). With probability $1-\exp\left(-\Omega\left(p\log\left(\frac{\Delta}{\epsilon}\right)\right)\right)$ over the choice of $G$, we have:

$$\left|\mathbb{E}_{T\sim\mathcal{T}_{nc}}\mathbb{E}_{z\sim T}\phi\left(D(G(z),z)\right)-\mathbb{E}_{G}\mathbb{E}_{T\sim\mathcal{T}_{nc}}\mathbb{E}_{z\sim T}\phi\left(D(G(z),z)\right)\right|\leq\epsilon$$

For all discriminators $D$ of capacity $\leq p$.

The proof of Lemma D.1. is a simple one-liner, whereas the proof for Lemma D.2. is somewhat involved so I will skip it.

From these two Lemmas we get that with probability $1-\exp\left(-\Omega\left(p\log\left(\frac{\Delta}{\epsilon}\right)\right)\right)$ over choice of $G$ we have that $\left|\mathbb{E}_{z\sim\nu}\phi\left(D(G(z),z)\right)-\mathbb{E}_{G,T\sim\mathcal{T}_G}\phi\left(D(G(z),z)\right)\right|\leq\epsilon$ for all discriminators $D$ of capacity $\leq p$.

# Similarity Search in High Dimensions via Hashing (Locality-Sensitive Hashing)

Summary:
Most nearest neighbor similarity search algorithms suffer from the curse of dimensionality. For example, in high dimensions, searching in k-d trees requires comparisons with a large part of the database and is therefore not much better than linear search. The authors propose locality-sensitive hashing (LSH) – an approximate nearest neighbor method that hashes data points so that those that are close are highly likely to be mapped to the same bucket whereas those that are far are highly unlikely to collide.

We make two assumptions about the data:

1. The distance is measured using L1 norm. This is not restrictive since L1 and L2 tend to have similar performance for similarity search.
2. All coordinates of data points are positive integers. This is also not very restrictive since we can apply a uniform shift of the datapoints (which is an isometry) to make all coordinates positive. We can then scale each coordinate by a large number and round to the nearest integer. Note that the rounding error can be made arbitrarily small.

The LSH algorithm:
Given a dataset $P = (p_1, \dots, p_n)$ of points in $d$-dimensional space, denote $C$ as the largest coordinate value in all points of $P$. One can embed $P$ into a Hamming cube $H^{d'}$ with $d' = Cd$. This is done by mapping each point $p = (x_1, \dots, x_d)$ to:

$$v(p) = Unary_C(x_1) \dots Unary_C(x_d)$$

Note that $Unary_C(x)$ is the unary representation of $x$ (a vector of $x$ 1s followed by $C - x$ 0s).

We note that this mapping is isometric (assuming we also use L1 norm in the embedding space).

We now define hash functions. For some integer $l$, choose $l$ subsets $I_1, \dots, I_l$ of the coordinate indices $\{1, \dots, d'\}$ (e.g. $I_1 = \{1,4,6\}$). Denote $p_{|I}$ as the projection of $p$ onto the coordinate set $I$, e.g. $p_{|I_1} = (x_1, x_4, x_6)$. Denote $g_j(p) = p_{|I_j}$. For each point $p \in P$, we store $p$ into buckets $g_j(p)$ for each $j = 1, \dots, l$ (i.e. each point gets stored into $l$ projection buckets). Now we have the bucket representations of the dataset. Since the number of buckets $l$ may be large, we do a further compression by resorting to another standard hash function. This maps the contents of the buckets into a hash table of size $M$. Denote the maximal bucket size of the hash table as $B$.

To compute the nearest neighbor of a query $q$, we search through $g_1(q), \dots, g_l(q)$ until encountering at least $cl$ (for some $c$) or used all $l$ indices. Say we encounter points $p_1, \dots, p_t$ in this search. Then, for (approximate) $K$-nearest neighbors, output the $K$ points $p_i$ closest to $q$.

The question then is how to choose the coordinate subsets $I_j$ for $j = 1, \dots, l$. The paper proposes constructing them by sampling $k$ coordinates uniformly at random from $\{1, \dots, d'\}$. $k$ and $l$ should be chosen to maximize the probability that points close to each other fall in the same bucket, and minimize the probability that points far apart collide.

**Analysis of the algorithm:**

Recall that the goal of LSH is to achieve locality-sensitiveness of the hashing, i.e. close points have high probability of being hashed to the same bucket and far points have low probability. This is formalized with the following definition of $(r_1, r_2, p_1, p_2)$-sensitivity.

**Definition:**

A family $\mathcal{H}$ of functions from $S$ to $U$ is $(r_1, r_2, p_1, p_2)$-sensitive for $D(.,.)$ if for any $q, p \in S$ we have that:

- If $p \in B(q, r_1)$ then $P_{\mathcal{H}}[h(q) = h(p)] \geq p_1$
- If $p \notin B(q, r_2)$ then $P_{\mathcal{H}}[h(q) = h(p)] \leq p_2$

We assume that $p_1 > p_2$ and $r_1 < r_2$.

With this notion of sensitivity, one can generalize the previous algorithm to any class of locality-sensitive functions $\mathcal{H}$. To do this, for $i = 1, \dots, l$, choose $g_i$ to be:

$$g_i(p) = \left(h_{i_1}(p), \dots, h_{i_k}(p)\right)$$

Where $h_{i_1}, \dots, h_{i_k}$ are sampled randomly with replacement from $\mathcal{H}$.

The authors prove that the LSH algorithm can be used to solve the $(r, \epsilon)$-Neighbor problem: Determine if there is a point $p$ in $B(q, r_1 = r)$, in which case return a point $p'$ in $B(q, (1+\epsilon)r)$, or if all points are at least $r_2 = r(1+\epsilon)$ distance away from $q$. The following theorem shows that LSH solves this problem for appropriate choices of $k$ and $l$.

**Core theorem:**

Let $P' = \{p' \in P | d(q, p') > r_2\}$.

Define the following properties:

P1: If $\exists p^*$ such that $p^* \in B(q, r_1)$, then $g_j(p^*) = g_j(q)$ for $j = 1, \dots, l$

P2: The number of blocks pointed by $q$ and containing only points from $P'$ is less than $cl$

Assume $\mathcal{H}$ is $(r_1, r_2, p_1, p_2)$-sensitive and define $\rho = \frac{\ln \frac{1}{p_1}}{\ln \frac{1}{p_2}}$

**If $k = \log_{1/p_2}\left(\frac{n}{B}\right)$ and $l = \left(\frac{n}{B}\right)^{\rho}$, then P1 and P2 hold with probability $\geq \frac{1}{2} - \frac{1}{e} \geq 0.132$.**

**Proof idea:**

Denote $P_1$ as the probability of P1 holding and $P_2$ as the probability of P2 holding. Setting $k = \log_{1/p_2}\left(\frac{n}{B}\right)$ and assuming $\exists p^* \in B(q, r_1)$, then for $p' \in P \backslash B(q, r_2)$ we have $\Pr[g(p) = g(q)] \leq p_2^k = \frac{B}{n}$. One notes that the expected number of blocks for a particular $g_j$ containing only points from $P'$ is at most 2 and thus the expected number of blocks for all $g_j$ is at most $2l$. Using Markov inequality, the probability that the expected number of blocks is $\geq 4l$ is less than $1/2$. Let $c = 4$, then $P_2 > \frac{1}{2}$.

The probability of $g_j(p^*) = g_j(q)$ similarly has a lower bound of $p_1^k = \left(\frac{n}{B}\right)^{-\rho}$

Setting $l = \left(\frac{n}{B}\right)^{\rho}$, one can give the following bound: $\Pr\left[g_j(p^*) \neq g_j(q)\ \forall j = 1, \dots, l\right] \leq 1/e$. Therefore, $P_1 \geq 1 - \frac{1}{e}$.

With the bounds for $P_1$ and $P_2$, the probability that both hold is $\geq \frac{1}{2} - \frac{1}{e}$.

Remark: Note that one can repeat LSH $O\left(\frac{1}{\delta}\right)$ times to increase the probability that P1 and P2 hold for at least one trial to $1 - \delta$.

The $\epsilon$-Nearest Neighbor problem (efficiently return a point $p \in P$ such that $d(q, p) \leq (1 + \epsilon)d(q, P)$) can be reduced to the above $(r, \epsilon)$-Neighbor problem simply by gradually increasing $r$. However, in previous work it has been shown that the distribution of distances $d(q, p)$ generally does not depend query $q$ and depends instead on intrinsic geometry of the dataset. This means that a single fixed choice of $r$ (and thus $k, l$) generally works well.

Experiments:

In experiments, the authors measure LSH error using $E = \frac{1}{|Q|}\sum_{query\ q \in Q} \frac{d_{LSH}}{d^*}$, where $d_{LSH}$ is the distance between the query $q$ and the point returned by LSH and $d^*$ is the distance between $q$ and the actual nearest neighbor.

The experiments generally show that LSH consistently has fewer disk accesses than SR-tree and has dramatically lower error rates.

# A concentration theorem for projections:

Summary:
For a random variable $X \in \mathbb{R}^D$ with mean zero, finite moments, most linear projections of $X$ onto $\mathbb{R}^d$ ($d < D$) look like a scale-mixture of spherical Gaussians $N(0, \sigma^2 I_d)$, where $\sigma$ follows the same distribution as $\|X\|/\sqrt{D}$. The extent of this Gaussian resemblance depends on $d/D$ and on coefficient of eccentricity of $X$'s distribution.

We denote $X \in \mathbb{R}^D$, $\mathbb{E}[X] = 0$, $\mathbb{E}[X^2] < \infty$. We let $\mu$ be the distribution of $\|X\|/\sqrt{D}$ and $\nu_\sigma$ be the d-dimensional Gaussian distribution $N(0, \sigma^2 I_d)$. Let $\bar{F}$ be the scale-mixture of spherical Gaussians:

$$\bar{F} = \int \nu_\sigma \mu(d\sigma)$$

For any dxD matrix $\Theta$, denote $F_\Theta$ as the distribution of the projection $\frac{1}{\sqrt{D}}\Theta X$. Also, for any ball $B \subset \mathbb{R}^d$, let $F_\Theta(B)$ is the total probability mass $F_\Theta$ assigns to $B$, i.e. $F_\Theta(B) = \mathbb{P}_X\left[\frac{\Theta X}{\sqrt{D}} \in B\right] = \mathbb{E}_X\left[1\left(\frac{\Theta X}{\sqrt{D}} \in B\right)\right]$

We note that concentration theorems exist for means of Gaussians. This can be generalized to any Lipschitz function/statistic $f$.

Theorem:
$\gamma_N$ denotes the distribution $N(0, I_N)$. Suppose $f: \mathbb{R}^N \to \mathbb{R}$ is $C$-Lipschitz. Then:

$$\gamma_N\{z: |f(z) - \mathbb{E}[f]| \geq r\} \leq 2e^{-r^2/2C^2}$$

Lemma 2:
If $\Theta$ is a dxD matrix with iid $N(0,1)$ entries, then $\forall X \in \mathbb{R}^D$, the distribution over the projection $\frac{1}{\sqrt{D}}\Theta X$ is $N\left(0, \frac{\|X\|^2}{D} I_d\right)$

Lemma 3:
$\forall B \subset \mathbb{R}^d$, we have $\mathbb{E}_\Theta[F_\Theta(B)] = \bar{F}(B)$

Proof: We achieve a random sample of $X$ through first sampling $\sigma = \frac{\|X\|}{\sqrt{D}}$ from $\mu$, and then picking $X$ such that $\|X\|^2 = \sigma^2 D$. So:

$$\mathbb{E}_\Theta[F_\Theta(B)] = \mathbb{E}_\Theta\left[\mathbb{E}_X\left[1\left(\frac{1}{\sqrt{D}}\Theta X \in B\right)\right]\right] = \mathbb{E}_X\left[\mathbb{E}_\Theta\left[1\left(\frac{1}{\sqrt{D}}\Theta X \in B\right)\right]\right]$$

$$= \mathbb{E}_\sigma\left[\mathbb{E}_X\left[\mathbb{E}_\Theta\left[1\left(\frac{1}{\sqrt{D}}\Theta X \in B\right)\right] \mid \|X\|^2 = \sigma^2 D\right]\right]$$

$$= \mathbb{E}_\sigma[\mathbb{E}_X[\nu_\sigma(B) \mid \|X\|^2 = \sigma^2 D]] = \mathbb{E}_\sigma[\nu_\sigma(B)] = \bar{F}(B)$$

Note that we cannot apply the concentration bound to $F$ because $F$ might not be Lipschitz. So instead consider a smoothed version of $F_\Theta$:

We note that $F_\Theta(B) = \mathbb{E}_X\left[1\left(\frac{\Theta X}{\sqrt{D}} \in B\right)\right]$. Consider the random variable $1\left(\frac{\Theta X}{\sqrt{D}} \in B\right)$. To smooth out $F$, we can consider smoothing out the indicator function. Define:

$$h_B(z) = \begin{cases} 1 \ \ if\ d(z,B) = 0 \\ 1 - \left(\dfrac{d(z,B)}{\Delta}\right) \ \ if\ 0 < d(z,B) < \Delta \\ 0 \ \ if\ d(z,B) > \Delta \end{cases}$$

We note that $h_B$ is $\frac{1}{\Delta}$-Lipschitz. Define the smoothed function as $\tilde{F}(\Theta, B) = \mathbb{E}_X\left[h_B\left(\frac{\Theta X}{\sqrt{D}}\right)\right]$. It can be shown straightforwardly that $\tilde{F}(.\,, B)$ is $\sqrt{\frac{\lambda_{max}}{D\Delta^2}}$-Lipschitz, where $\lambda_{max}$ is the largest eigenvalue of $\mathbb{E}_X[XX^T]$. Therefore, applying the concentration bound to $\tilde{F}$ gives us:

Claim 5:
Fix $B \subset \mathbb{R}^d$ and any $\epsilon > 0$. Pick $\Theta$ at random. Then:

$$\mathbb{P}_\Theta\left[\left|\tilde{F}(\Theta, B) - \mathbb{E}_\Theta[\tilde{F}(\Theta, B)]\right| \geq \epsilon\right] \leq 2e^{-\epsilon^2\Delta^2 D/2\lambda_{max}}$$

However, we are interested in a concentration bound for $F_\Theta$, not its smoothed version. So note the following relationship:

$$F_\Theta(B) \leq \tilde{F}_\Theta(B) \leq F_\Theta(B_\Delta)$$

Where $B_\Delta = B \cup B(0, \Delta)$ (i.e. grow the radius of B by $\Delta$). Denote $B_{-\Delta}$ to be B with radius smaller by $\Delta$.

Corollary 6: Fix $B \subset \mathbb{R}^d, \epsilon > 0$, $\Theta$ picked at random. Then:

$$\mathbb{P}_\Theta[\bar{F}(B_{-\Delta}) - \epsilon \leq F(\Theta, B) \leq \bar{F}(B_\Delta) + \epsilon] \geq 1 - 2e^{-\epsilon^2\Delta^2 D/2\lambda_{max}}$$

To relate $\bar{F}(B)$ to $\bar{F}(B_\Delta)$, we first relate $\nu_\sigma(B)$ and $\nu_\sigma(B_\Delta)$. Note that if $\Delta$ is small, $\nu_\sigma(B_\Delta)$ is not much larger than $\nu_\sigma(B)$. This is formalized by the following statement: Pick $0 < \epsilon < 1$ and $\sigma > 0$, then if $\Delta \leq \frac{\sigma}{2\sqrt{d}} \ln\left(1 + \frac{\epsilon}{8}\right) \frac{1}{1 + \sqrt{\frac{2}{d}\ln\frac{8}{\epsilon}}}$ then $\nu_\sigma(B_\Delta) \leq \nu_\sigma(B) + \epsilon$ for any ball $B$.

This leads to directly to corollary 8, which relates $\bar{F}(B)$ to $\bar{F}(B_\Delta)$:

Corollary 8:
Pick $0 < \epsilon < 1$ and $\sigma_\epsilon > 0$ such that $\mu\{\sigma : \sigma < \sigma_\epsilon\} \leq \epsilon$. If $\Delta \leq \frac{\sigma_\epsilon}{2\sqrt{d}} \ln\left(1 + \frac{\epsilon}{8}\right) \frac{1}{1 + \sqrt{\frac{2}{d}\ln\frac{8}{\epsilon}}}$ then

$\bar{F}(B_\Delta) \leq \bar{F}(B) + 2\epsilon.$

Proof:

$$\bar{F}(B_\Delta) - \bar{F}(B) = \mathbb{E}_\sigma[\nu_\sigma(B_\Delta)] - \mathbb{E}_\sigma[\nu_\sigma(B)] \le \mathbb{E}_\sigma[\nu_\sigma(B_\Delta) - \nu_\sigma(B) |\, \sigma \ge \sigma_\epsilon] + \mathbb{P}_\sigma(\sigma < \sigma_\epsilon) \le 2\epsilon$$

This gives rise to Theorem 9, which states that $F_\Theta(B) \cong \bar{F}(B)$ with high probability on any given ball $B$ (i.e. a concentration theorem for $F_\Theta(B)$! More concretely, pick $0 < \epsilon < 1$ and $\sigma_\epsilon > 0$ such that $\mu\{\sigma: \sigma < \sigma_\epsilon\} \le \epsilon$. Pick $B \subset \mathbb{R}^d$. Then:

$$\mathbb{P}_\Theta[|F_\Theta(B) - \bar{F}(B)| > \epsilon] \le \exp\left(-\widetilde{\Omega}\left(\frac{\epsilon^4 D}{d}\frac{\sigma_\epsilon^2}{\lambda_{max}}\right)\right)$$

Finally we wish to prove this uniform convergence for all balls simultaneously. To do this, construct balls $B_1, \dots, B_M$ through first creating a grid with resolution $2\epsilon_0$ on $\left[-c\sqrt{d}, c\sqrt{d}\right]^d$, then creating a set of balls centered at each point in the grid, with ball radii $\epsilon_0\sqrt{d}, 2\epsilon_0\sqrt{d}, \dots, (2c + 2\epsilon_0)\sqrt{d}$. With this construction, one can show the following lemma holds:

Lemma 10:

Let $c \ge \sqrt{\lambda_{avg}/2\epsilon}$ and $\epsilon_0 \le \frac{\Delta}{4\sqrt{d}}$. Pick any $B \subset \mathbb{R}^d$ centered in $B\left(0, c\sqrt{d}\right)$. Then there exists $B_i, B_j$ such that $B_i \subset B \subset B_j$ and $\bar{F}\left(B_j\right) - \bar{F}(B_i) \le 2\epsilon$.

This leads directly to the final result of the paper, i.e. that the distribution of the projection converges uniformly to the scale-mixture of Gaussians:

Theorem 11:

For any $0 < \epsilon < 1$ and $\sigma_\epsilon > 0$ such that $\mu(\sigma: \sigma < \sigma_\epsilon) \le \epsilon$ we have:

$$\mathbb{P}_\Theta\left[\sup_{balls\ B \subset \mathbb{R}^d} |F_\Theta(B) - \bar{F}(B)| > \epsilon\right] \le \left(\tilde{O}\left(\frac{d^3}{\epsilon^3}\frac{\lambda_{avg}}{\sigma_\epsilon^2}\right)\right)^{\frac{d}{2}} \exp\left\{-\widetilde{\Omega}\left(\frac{\epsilon^4 D}{d}\frac{\sigma_\epsilon^2}{\lambda_{max}}\right)\right\}$$