OPTIMAL STRATEGIES AGAINST GENERATIVE ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative neural models have improved dramatically recently. With this progress comes the risk that such models will be used to attack systems that rely on sensor data for authentication and anomaly detection. Many such learning systems are installed worldwide, protecting critical infrastructure or private data against malfunction and cyber attacks. We formulate the scenario of such an authentication system facing generative impersonation attacks, characterize it from a theoretical perspective and explore its practical implications. In particular, we ask fundamental theoretical questions in learning, statistics and information theory: How hard is it to detect a “fake reality”? How much data does the attacker need to collect before it can reliably generate nominally-looking artificial data? Are there optimal strategies for the attacker or the authenticator? We cast the problem as a maximin game, characterize the optimal strategy for both attacker and authenticator in the general case, and provide the optimal strategies in closed form for the case of Gaussian source distributions. Our analysis reveals the structure of the optimal attack and the relative importance of data collection for both authenticator and attacker. Based on these insights we design practical learning approaches, and show that they result in models that are more robust to various attacks on real-world data. Code will be made publicly available upon publication.

1 INTRODUCTION

Generative models have attracted considerable attention since the introduction of Generative Adversarial Networks (Goodfellow et al., 2014a). Empirically, GANs have been shown to generate novel data instances that resemble those in the true distribution of the data. The success of GANs also comes with the risk that generative models will be used for attacking sensor-based security systems. One example is identity authentication systems, where an individual is identified via her images, and GANs might be able to generate such images to gain access (Thies et al., 2016). Another is anomaly detection systems protecting critical infrastructure. As demonstrated by recent cyber-attacks (notably the Stuxnet attack) sensors of these systems can be hijacked, so that GANs can be used to generate “normal” looking activity while the actual system is being tampered with. The latter is in fact a new form of a man-in-the-middle attack.

Our goal here is to construct a theoretical framework for studying the security risk arising from generative models, and explore its practical implications. We begin with a simple key insight. If the attacker (i.e. the generative model) has unlimited observations of the source it is trying to imitate, it will be able to fool any authenticator. On the other hand, if the attacker has access to fewer sensor observations than the number of fake observations it needs to generate, it seems intuitively clear that it cannot always succeed (as we indeed prove in Sec. 4). Therefore, the optimal defense and attack strategies depend crucially on the amount of information available to the attacker and authenticator.

Motivated by the above insight, we cast the authentication setting as a two player maximin game (authenticator vs. attacker) where all observations are finite. Specifically, there are three key observation sets to consider: those available to the attacker, those that the attacker needs to generate, and those available to the authenticator when designing the system. Our goal is to understand how these three information sources determine the optimal strategies for both players. Under the realistic assumption that cyber attackers are sophisticated enough to play optimal or close to optimal strategies, a characterization of the maximin authentication strategy can be of significant value.

We prove several theoretical results characterizing the optimal strategy for both players. These results highlight the role played by the available observations as well as the functional form for an optimal attacker and authenticator. We refer to the setting above as "GAN in The Middle" (GIM) due to its similarity to "man in the middle" attacks. After describing our theoretical results, we show how to learn both authenticator and attacker policies in practice, where both are based on neural architectures. Our GIM method can be applied to multiple practical problems. The first is building an authentication system that is robust to impersonation attacks. The second is building a data generating mechanism that can generate novel data instances. Finally, we evaluate the method empirically, showing that it outperforms existing methods in terms of resilience to generative attacks, and that it can be used effectively for data-augmentation in the few-shot learning setting.

2 Problem statement

We begin by motivating our problem, and formulating it as a two player zero sum game. As a simple illustrative example, consider a face authentication security system whose goal is to maximize authentication accuracy. The system is initialized by registering k images of an individual θ (the "source"), whose identity is to be authenticated. At test-time, each entity claiming to be θ is required to present to the system n of its images, and the authentication system decides whether the entity is θ or an impersonator. We let m denote the maximum number of "leaked" images any attacker obtained. We observe that if an attacker obtained $m \geqslant n$ images of θ, it can present n of those images. Thus the observations generated by the attacker are indistinguishable from ones generated by θ, leading to failure of the authentication system (see Sec. 4.2 below). Namely, the number of images of θ that the attacker obtains and the size of the authentication sample are of key importance. We now turn to formally stating the problem.

Notation. The set of possible observations is denoted by $\mathcal{X}$. Let $\mathcal{H}$ denote the known set of possible sources θ, where each source $\theta \in \mathcal{H}$ is defined by a probability density f_θ, and an observation of a source $\theta \in \mathcal{H}$ is an $\mathcal{X}$-valued random variable with density f_θ. We assume that subsequent observations of the source are IID, so that n sequential observations have density $f_\theta^{(n)}(x_1, \ldots, x_n) := \prod_{i=1}^{n} f_\theta(x_i)$. We allow θ to be sampled from a known distribution Q on $\mathcal{H}$ and denote the corresponding $\mathcal{H}$-valued random variable by Θ. In what follows we will denote the number of observations leaked to the attacker by m, the number of "registration" observations available to the authenticator by k, and the number of observations required at authentication by n (these may be generated by either the attacker or the true source).

The Authentication Game. The game begins with a random source Θ being drawn from $\mathcal{H}$ according to Q. The authenticator first receives information about the drawn source, and then chooses a decision rule for deciding whether a given test sequence $x \in \mathcal{X}^n$ is an authentic sequence of observations sampled from $f_\theta^{(n)}$, or a fake sequence generated by the attacker. Formally the authenticator learns about the source by seeing k IID "registration" observations $A = A_1, \ldots, A_k \sim f_\theta^{(k)}$. The set of all possible decisions rules is then $\mathcal{D} : \mathcal{X}^k \times \mathcal{X}^n \to \{0, 1\}$ (where a decision of 1 corresponds to the true source, and 0 to an attacker). After the authenticator fixes its strategy, the attacker can seek the best attack strategy. We assume that the attacker has access to m "leaked" IID observations $Y = Y_1, \ldots, Y_m \sim f_\theta^{(m)}$ as information about the source θ. Then it generates an attack sequence $X \in \mathcal{X}^n$ and presents it to the authenticator which uses its decision rule to decide whether X is an authentic sequence of observations sampled from $f_\theta^{(n)}$, or a fake sequence generated by the attacker. Formally, the strategy set of the attacker is all functions $\mathcal{G} : \mathcal{X}^m \to \Delta(\mathcal{X}^n)$, where $\Delta(\mathcal{X}^n)$ is the set of probability distributions over $\mathcal{X}^n$, and $g_{X|Y}$ is the associated conditional probability density. We note that the set $\mathcal{H}$, the parametric family f_θ, and the prior probability Q are known to both players. Also, note that the leaked sample Y revealed to the attacker is *not* available to the authenticator, and the "registration" sample A is not available to the attacker.

The goal of the authenticator is to maximize its expected accuracy, and the goal of the attacker is to minimize it (or equivalently maximize its success probability). We define the utility (payoff) of the game as the expected prediction accuracy of the authenticator. To define expected accuracy we consider the case of equal priors for attack and real samples.[1] Formally, for a pair of strategies

[1] We give equal prior probability to attack and real samples for simplicity and clarity, and since this is common in GAN formulations. All results can be trivially generalized to any prior probability for attack.

$(\mathcal{D}, \mathcal{G})$ and a specific source θ, the expected accuracy of the authenticator is then:

$$V(\theta, \mathcal{D}, \mathcal{G}) = \frac{1}{2}\mathbb{E}_{A\sim f_\theta^{(k)}}\mathbb{E}_{Y\sim f_\theta^{(m)}}\left[\mathbb{E}_{X\sim f_\theta^{(n)}}[\mathcal{D}(A, X)] + \mathbb{E}_{X\sim\mathcal{G}(Y)}[1 - \mathcal{D}(A, X)]\right] \tag{2.1}$$

Since this utility only depends on $\mathcal{G}$ in the second term, minimizing it is equivalent to $\mathcal{G}$ maximizing its success probability. To obtain the overall utility for the authenticator, we take the expected value w.r.t θ and define $V(\mathcal{D}, \mathcal{G}) = \mathbb{E}_{\Theta\sim Q}V(\Theta, \mathcal{D}, \mathcal{G})$. Finally, we arrive at the following maximin game:

$$V_{game} = \max_{\mathcal{D}\in\mathbb{D}}\min_{\mathcal{G}\in\mathbb{G}} V(\mathcal{D}, \mathcal{G}) \tag{2.2}$$

where $\mathbb{D}, \mathbb{G}$ are the set of all possible authenticator and attacker strategies, respectively. In Sec. 4 we show that this game has a Nash equilibrium, we characterize the optimal strategies and game value in general, and find them in closed form for the case of Multivariate Gaussian sources.

3 Related work

Adversarial hypothesis testing (AHT): Hypothesis testing (HT) is a rich field in statistics which studies how one can detect whether a sample was generated by one of two sets of distributions. A variant of HT that is related to our work but distinct from it is AHT (e.g., see Brandão et al., 2014; Barni & Tondi, 2013b;a; 2014; Bao et al., 2011; Zhou et al., 2019; Brückner & Scheffer, 2011; Brückner et al., 2012). These works describe an HT setting where the sample is generated by one of two hypotheses classes, but is then modified by an adversary in some restricted way. e.g., in Barni & Tondi (2013b;a; 2014) the adversary can change the sample of one class up to a fixed distance (e.g., Hamming). Given the quality of current generative models and the rapid pace of progress, when considering an impersonation attack, it seems that the only relevant restriction one can assume on an attacker is on the information it has. This is not captured by prior work since it assumes that the adversary has a restricted strategy set. In contrast, our work considers a novel problem setting where both players are not limited in their strategy set in any way. This leads to a novel analysis that focuses on the dependence on the finite information available to each player (m, n, k).

Adversarial Examples: It has been observed (Goodfellow et al., 2014b) that deep learning models can be very sensitive to small changes in their input. Such "misleading" inputs are known as *adversarial examples* and much recent work has analysed the phenomenon (Ilyas et al., 2019; Shamir et al., 2019; Zhang et al., 2019), addressed the problem of robustness to these (Moosavi-Dezfooli et al., 2016; Papernot et al., 2017; Yuan et al., 2017), and certifying robustness (Raghunathan et al., 2018; Wong et al., 2018). The setting of robust classification in the presence of adversarial examples can also be thought of as a specific case of AHT (see above), where a classifier is required to predict the class of an observation that could have been perturbed by a restricted adversary (Wong et al., 2018; 2019) or generated by an adversary limited to generating examples that will be classified correctly by humans (Song et al., 2018). In contrast, in our setting the attacker is not limited in any way, nor does it have another utility in addition to impersonating the source. Furthermore, in adversarial examples there is no notion of limited information for the adversary, whereas our work focuses on the dependence of the game on the information available to the players (sample sizes n, m, k).

GAN: The GAN model is a game between a generator and discriminator. While our concept of generative attacks is inspired by a GAN it is very different from it: a successful GAN generator is not necessarily a successful attacker in our setting, and vice-versa (e.g., given sufficiently expressive generators and discriminators, GANs can "memorize" the training data, and thus the discriminator will perform at chance level.[2] Such a discriminator will not be useful as a defence against generative attacks). Unlike GANs, in our setting, sample sizes are of key importance. Thus, our attacker will not memorize the data it sees, as this will be detected when generating $n > m$ examples.

Conditional GANs: In conditional GANs (Mirza & Osindero, 2014) the generator uses side information for generating new samples. The attacker in our approach (analogous to GAN generator) has input, but this input is not available to the authenticator (analogous to GAN discriminator). Thus, the objective of the learning process is fundamentally different.

Few shot learning and generation: Our work relates to few shot learning (Snell et al., 2017; Vinyals et al., 2016; Finn et al., 2017; Lake et al., 2011; Koch et al., 2015) and few shot generative models (Rezende et al., 2016; Zakharov et al., 2019; Lake et al., 2015; Edwards & Storkey, 2017; Hewitt et al., 2018) in the sense that both authenticator and attacker need to learn from a limited set

[2] If the generator is not expressive enough, the learned GAN may have small support (Arora et al., 2017).

of observations. However, in our setting the authenticator is required to predict whether a sample came from the true source or an attacker impersonating the source, while taking into consideration the amount of information both players have. These are notions that are not part of the general few shot learning setup. Also, in prior work on few-shot generation, the generator is either measured through human evaluation (Lake et al., 2015) or trained to maximize the likelihood of its generated sample (Rezende et al., 2016; Edwards & Storkey, 2017; Hewitt et al., 2018). In contrast, in our setting the attacker's objective is to maximize the probability that its generated sample will be labeled as real by an authenticator. To this end, we show that the attacker must consider the sample sizes m, n, k, which the generative models in other works do not account for. Furthermore, we show in Sec. F.4, and in Figures 1c,6, that the ML solution is indeed sub-optimal in our setting.

Image to image translation: Several GAN models have been introduced for mapping between two domains (Zhu et al., 2017; Huang et al., 2018; Isola et al., 2017; Wang et al., 2017; Park et al., 2019). This relates to our work since the attacker also needs to learn to map the leaked sample to an attack sample. However, in our setting the mapping is not to a different domain but rather to other images from the same distribution, which results in a different objective.

Data Augmentation: Generative models have also been used for augmenting data in supervised learning, and in particular few-shot learning (Koch et al., 2015; Snell et al., 2017; Vinyals et al., 2016; Lake et al., 2011). One such approach is Data Augmentation GAN (DAGAN) (Antoniou et al., 2018), which takes as input an image and generates a new image. It relates to our framework in the limited case of $m = 1, n = 2, k = 1$, in the sense that the generator objective is to map one image to two. However, in DAGAN the only goal of the discriminator is to improve the generator, and the generator is limited to the functional form of adding a new image to the existing one, which is a sub-optimal attack strategy, as can be seen from the Gaussian case of our problem.

4 Theoretical results

In this section we study the game defined in Eq. 2.2. First, in Sec. 4.1 we show the existence of a Nash equilibrium and characterize the optimal strategies for both players. Specifically we show that the optimal attacker strategy minimizes a certain divergence between its marginal distribution and $f_\Theta^{(n)}$. Next, Sec. 4.2 shows that when there are more leaked samples m than generated samples n, the authenticator will fail. Finally, in Sec. 4.3 we provide a closed form solution for both attacker and authenticator for the case of multivariate Gaussian distributions, and analyze the significance of the observations' dimension, and the sample sizes m, n and k. Proofs are provided in the appendix.

4.1 Characterizing the optimal strategies

We begin by showing that the game defined in Eq. 2.2 admits a Nash equilibrium. Namely, Theorem 4.1 below shows that there exists a pair of strategies $(\mathcal{D}^*, \mathcal{G}^*)$ that satisfy:

$$\max_{\mathcal{D} \in \mathbb{D}} \min_{\mathcal{G} \in \mathbb{G}} V(\mathcal{D}, \mathcal{G}) = \min_{\mathcal{G} \in \mathbb{G}} \max_{\mathcal{D} \in \mathbb{D}} V(\mathcal{D}, \mathcal{G}) = V(\mathcal{D}^*, \mathcal{G}^*) \tag{4.1}$$

Theorem 4.1. *Consider the attacker $\mathcal{G}^*$ defined by:*

$$g^*_{X|Y} \in \operatorname*{argmin}_{g_{X|Y}} \mathbb{E}_{A \sim f_A} \left[\int_{x \in \mathcal{X}^n} \left| f_{X|A}(x|A) - g_{X|A}(x|A) \right| dx \right] \tag{4.2}$$

Where $f_A(a) = \int_{\theta \in \mathcal{H}} Q(\theta) f_\theta^{(k)}(a) d\theta$ is the marginal density of A, $Q_{\Theta|A}$ is the posterior over $\mathcal{H}$ given A, $f_{Y|A}(y|a) = \int_{\theta \in \mathcal{H}} f_\theta^{(m)}(y) Q_{\Theta|A}(\theta|a) d\theta$, and $f_{X|A}(x|a) = \int_{\theta \in \mathcal{H}} f_\theta^{(n)}(x) Q_{\Theta|A}(\theta|a) d\theta$, $g_{X|A}(x|a) = \int_{y \in \mathcal{X}^m} g_{X|Y}(x|y) f_{Y|A}(y|a) dy$ are the conditional densities of X given A generated by the source and the attacker respectively. Consider the authenticator defined by: $\mathcal{D}^*(a, x) = I \left[f_{X|A}(x|a) > g^*_{X|A}(x|a) \right]$, where I is the indicator function. Then $(\mathcal{D}^*, \mathcal{G}^*)$ is a solution of Eq. 2.2 that satisfies Eq. 4.1.

The proof (see Sec. D) follows by first showing that since $\mathcal{D}(a, x) \in \{0, 1\}$, then for any $\mathcal{G}$ its optimal strategy is a MAP test between the two hypotheses (true source or attacker). We then show that given $\mathcal{D}^*$, the game objective for $\mathcal{G}$ becomes the ℓ_1 distance over the space $\mathcal{X}^k \times \mathcal{X}^n$ between the true source distribution and the marginal distribution of the attacker, and thus any optimal $\mathcal{G}$ minimizes it. Therefore it holds that $\min_{\mathcal{G}} V(\mathcal{D}^*, \mathcal{G}) = V(\mathcal{D}^*, \mathcal{G}^*) = \max_{\mathcal{D}} V(\mathcal{D}, \mathcal{G}^*)$ and it follows that $\mathcal{D}^*, \mathcal{G}^*$ satisfy Eq. 4.1.

4.2 Replay Attacks: Authentication failure for $n \leqslant m$

When $n \leqslant m$, the attacker generates a number of observations that is at most the number of observations it has seen. Intuitively, an optimal attack in this case is to simply "replay" a subset of size n from the m observations. This is known as a replay-attack (Syverson, 1994). This subset constitutes an IID sample of length n of the observed source, and is therefore a legitimate "fresh" sample. In this case, it seems like the attack cannot be detected by the authenticator. Indeed it is easy to show using Theorem 4.1 that this attack is optimal and therefore for $n \leqslant m$ we have: $\max_{\mathcal{D}\in\mathbb{D}} \min_{\mathcal{G}\in\mathbb{G}} V(\mathcal{D},\mathcal{G}) = 0.5$ (see Sec. E)

4.3 The Gaussian case

We now turn to the case of multivariate Gaussian distributions where we can find the exact form of the attacker and authenticator, providing insight into the general problem. Specifically, we consider the setting where the source distributions are d-dimensional multivariate Gaussians with an unknown mean and known covariance, and the prior Q over $\mathcal{H}$ is the improper uniform prior.[3] We assume $n > m$ to keep the problem non-trivial. Let the observations be d-dimensional Gaussian vectors with a known covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ and an unknown mean vector $\theta \in \mathbb{R}^d$. The set of possible sources $\mathcal{H}$ becomes $\mathbb{R}^d$, the Gaussian mean vectors. For any sample of n examples $z \in \mathbb{R}^{n\times d}$ we let z_i denote the i'th example, and $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ denote the sample mean. Finally, for any $v \in \mathbb{R}^d, B \in \mathbb{R}^{d\times d}$, we define $\|v\|_B^2 = v^T B v$. The following theorem gives a closed-form solution for both attacker and authenticator for the game defined in Eq. 2.2.

Theorem 4.2. *Define $\delta = m/n \leqslant 1$ and let $\rho = m/k$. Consider the attacker $\mathcal{G}^*$ defined by the following generative process: Given a leaked sample $Y \in \mathbb{R}^{m\times d}$, $\mathcal{G}^*$ generates a sample $X \in \mathbb{R}^{n\times d}$ as follows: it first samples n vectors $W_1,\ldots,W_n \overset{iid}{\sim} \mathcal{N}(0,\Sigma)$ and then sets: $X_i = W_i - \bar{W} + \bar{Y}$. Define the authenticator $\mathcal{D}^*$ by:*

$$\mathcal{D}^*(a,x) = I\left[\|\bar{x}-\bar{a}\|_{\Sigma^{-1}}^2 < \frac{d\,(1+\rho)\left(1+\rho\delta^{-1}\right)}{n(1-\delta)}\log\left(\frac{\rho+1}{\rho+\delta}\right)\right] \tag{4.3}$$

Then $(\mathcal{D}^*,\mathcal{G}^*)$ is a solution of Eq. 2.2 that satisfies Eq. 4.1.

The proof (see Sec. F) starts by showing that $\forall\alpha > 0$, given $\mathcal{D}(a,x) = I[\|\bar{x}-\bar{a}\|_{\Sigma^{-1}}^2 < \alpha]$, the optimal strategy for $\mathcal{G}$ is to set $\bar{x} = \bar{y}$ with probability 1 (as done in $\mathcal{G}^*$). To prove this, we first use the Prekopa-Leindler inequality (Prkopa, 1973) to show that in this case $\mathcal{G}$'s maximization objective is log-concave. We then show that any $\mathcal{G}$ that satisfies $\bar{x} = \bar{y}$ with probability 1 is a local extremum, and since the objective is log-concave it follows that it is the global maximum. We continue by showing that given $\mathcal{G}^*$, $\mathcal{D}^*$ is optimal. To do so, we first find the distribution of $\mathcal{G}^*$'s attack sample using the Woodbury matrix identity, and then show that $\mathcal{D}^*$ is indeed the optimal decision rule. Finally, using the max-min inequality, this implies that $(\mathcal{D}^*,\mathcal{G}^*)$ satisfy Eq. 4.1.

There are several interesting observations about the above optimal strategies. Perhaps the most intuitive strategy for the attacker would have been to sample n elements from a Gaussian with mean $\bar{Y}$ and the known covariance Σ. In expectation this sample would have the correct mean. However, this turns out to be sub-optimal, as can be seen in Figures 1c, 6 (we refer to this as an ML attack, see Sec. F.4 in the appendix for the derivation and visualizations). Instead, the optimal attacker begins by drawing an IID sample, W, from a Gaussian distribution with mean 0, and then "forces" the sample mean to be exactly $\bar{Y}$ by shifting the sample points by $\bar{Y} - \bar{W}$. This optimal attacker strategy can be viewed as matching the sufficient statistics of the leaked sample Y in the generated sample X. The optimal authenticator is a MAP test for the optimal attacker, as in Theorem 4.1.

As a corollary to Theorem 4.2 we obtain the value of the game (i.e., the accuracy of $\mathcal{D}^*$).

Corollary 4.3. *Define δ and ρ as in Theorem 4.2. Then the game value for the Gaussian case is:*

$$\frac{1}{2} + \frac{1}{2\Gamma(\frac{d}{2})}\left[\gamma\left(\frac{d}{2}, \frac{d(1+\rho)}{2(1-\delta)}\log\frac{1+\rho}{\delta+\rho}\right) - \gamma\left(\frac{d}{2}, \frac{d(\delta+\rho)}{2(1-\delta)}\log\frac{1+\rho}{\delta+\rho}\right)\right] \tag{4.4}$$

Where γ is the lower incomplete gamma function, and Γ is the Gamma function.

[3] Similar results can be derived for the conjugate prior case, namely, proper Gaussian priors.

The proof (see Sec. F) follows by showing that the test statistic used by $\mathcal{D}^*$ is Gamma distributed. Fig. 1 demonstrates several interesting aspects of the above results. First, Fig. 1a shows that the authenticator accuracy improves as n the size of the test sample grows. Furthermore, accuracy also improves as the dimension d grows, meaning that for a specified authentication accuracy, the required ratio n/m becomes smaller with the dimension d. This is a very encouraging result, since although this dimensional dependency is proved only for Gaussian sources, it suggests that for real world high-dimensional sources (e.g. faces, video, voice, etc.) the authenticator can achieve high accuracy even when requiring a small (and practical) authentication sample.

Intuitively it may seem like authentication is impossible when the authenticator has less data than the attacker (i.e., $m > k$). Surprisingly, this is not the case. As can be seen in Fig. 1b, even when $m > k$, the authenticator can achieve non trivial accuracy.[4] An intuitive explanation to this phenomenon is that the test statistic used by the authenticator is $\left\|\bar{X} - \bar{A}\right\|$, which, due to the variance in the attacker estimation, has higher variance when X is generated by an attacker than when X is generated by the true source. This in turn allows the authenticator to discriminate between the hypotheses.

A closed-form solution for the general case remains an open problem. We believe that solving for Gaussians is an important step forward, since it exposes interesting structural properties of the solution, which we use in practice. Furthermore, if $\mathcal{G}$ has an encoder-decoder structure, it is not unreasonable that the source distribution in latent space can be approximately Gaussian (as in VAE).

5 GAN IN THE MIDDLE NETWORKS

So far we explored the general formalism of authentication games. Here we consider specific architectures for $\mathcal{D}$ and $\mathcal{G}$. As in GAN based models (Mirza & Osindero, 2014; Mescheder et al., 2018; Karras et al., 2018a;b), we use neural nets to model these, while using insights from our theoretical analysis. Below we provide implementation details for the GIM model (see Sec. H and code for more details). In our analysis we considered the non-differentiable zero-one loss since it is the real accuracy measure. In practice we will use cross-entropy as used in most GAN approaches.

Authenticator Architecture: The authenticator is implemented as a neural network $\mathcal{D}(a, x)$ that maps from a source information sample $a \in \mathcal{X}^k$ and a test sample $x \in \mathcal{X}^n$ to a probability that the test sample came from the true source. Our framework does not restrict the authenticator to any specific function type, but in practice one must of course implement it using some model. We recall that our theoretical results do suggest a certain functional form. The Gaussian results in Sec. 4.3 show that the optimal authenticator is a test on the sufficient statistic of the source parameters. Motivated by this result, and in the spirit of Siamese networks (Koch et al., 2015; Chopra et al., 2005), we consider the following form for the authenticator. We define a function $T_{\mathcal{D}}$ that maps a sample to a fixed sized vector, analogous to the sufficient statistic in the theorem. We apply $T_{\mathcal{D}}$ to both a and x. Then, these two outputs are used as input to a comparison function σ which outputs a scalar reflecting their similarity. Thus the authenticator can be expressed as: $\mathcal{D}(a, x) = \sigma(T_{\mathcal{D}}(a), T_{\mathcal{D}}(x))$.

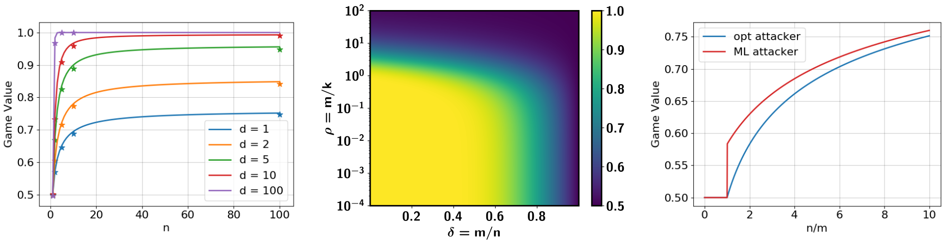
Attacker Architecture: The attacker is implemented as a stochastic neural network $\mathcal{G}(y)$ that maps a leaked sample $y \in \mathcal{X}^m$ to an attack sample $x \in \mathcal{X}^n$. Our theoretical results suggest a certain functional form for this network. The Gaussian analysis in Sec. 4.3 shows that the optimal attacker generates a sample whose sufficient statistic matches that of the leaked sample. Motivated by this result, we consider the following functional form for the attacker. First it applies a function $T_{\mathcal{G}}$ that maps the leaked sample Y to a fixed sized vector $T_{\mathcal{G}}(Y)$, analogous to the sufficient statistic in the theorem. It then draws n random latent vectors $W_1, \ldots, W_n$ and matches their mean to the leaked sufficient statistic to obtain the latent vectors W_i'. I.e it sets $W_i' = W_i - \bar{W} + T_{\mathcal{G}}(Y)$ as done in Theorem 4.2. Finally, it uses a decoder function φ that maps each latent vector W_i' to the domain $\mathcal{X}$. Thus, the attacker can be expressed as: $\mathcal{G}(Y)_i = \varphi(W_i - \bar{W} + T_{\mathcal{G}}(Y)) \quad \forall i \in [n]$.

Optimization Details: Each iteration begins when a source θ is chosen randomly from the set of sources in the training dataset (e.g., a person to be authenticated). Samples A, Y, X_θ are drawn from the set of examples available for θ, where X_θ represents a test sample from θ. Then, given a leaked sample Y, $\mathcal{G}$ generates a fake sample $X_{\mathcal{G}}$, passes it to $\mathcal{D}$ and suffers the appropriate loss. Finally, $\mathcal{D}$ receives as input the source information sample A, outputs a prediction for each of the test samples $X_\theta, X_{\mathcal{G}}$, and suffers the appropriate loss. Optimization is done via gradient ascent on authenticator parameters and descent on attacker parameters, as is typical for GAN problems.

[4] See appendix for additional figures showing that as the dimension grows, the expected accuracy goes to 1.

6 Experiments

We next evaluate our method empirically. In all experiments we use the model described in Sec. 5. We optimize the model with adversarial training, using the loss suggested by Mescheder et al. (2018) and Adam (Kingma & Ba, 2015). See Sec. H for further Implementation details.
Gaussian sources: For the case of Gaussian sources we arrived at a complete characterization of the solution in Sec. 4.3. Thus, we can learn the models using our GIM algorithm and check whether it finds the correct solution. This is important, as the GIM objective is clearly non-convex and GANs are generally hard to train in practice and lack convergence guarantees (Mescheder et al., 2018). We ran all experiments using a multivariate Gaussian with $\Sigma = I_d$, and in each game the source mean was drawn from the prior distribution $Q = \mathcal{N}(0, 10I_d)$. This approximates the improper uniform prior since the prior has much larger variance than the sources. Fig. 1a shows the empirical game value compared to the theoretical one as a function of the test sample size n, for fixed $m = 1, k = 10$, and different values of d. It can be seen that there is an excellent fit between theory and experiment.



(a) (b) (c)

Figure 1: Game value (expected authentication accuracy) for the Gaussian case. (a) A comparison between empirical and theoretical game value for different d values ($m = 1, k = 10$). Solid lines describe the theoretical game values whereas the $*$ markers describe the empirical accuracy when learning with the GIM model. (b) Theoretical game value as a function of δ, ρ (see Corollary 4.3) for $d = 100$. (c) Empirical accuracy of an optimal authenticator against two attacks: the theoretically optimal attack G^* from Theorem 4.2 and a maximum likelihood (ML) attack (See Sec. F.4) for the Gaussian case. It can be seen that the ML attack is inferior in that it results in better accuracy for the authenticator, as predicted by our theoretical results.

Authentication on Faces and Characters: We next evaluate GIM in an authentication setting on two datasets: the VoxCeleb2 faces dataset (Nagrani et al., 2017; Chung & Zisserman, 2018), and the Omniglot handwritten character dataset (Lake et al., 2015). Additional information about the datasets, splits, and modeling details is provided in Sec. G, H. Our goal is to check whether the GIM authenticator is more robust to generative attacks than a state of the art authentication system. To evaluate this, we consider several attackers: 1) A "random source" attacker (RS): a naive attacker that ignores the leaked sample Y. It simply draws a random source from the dataset, and samples n real images of that source. From the authenticator's perspective it's equivalent to a sample version of the verification task (Koch et al., 2015; Schroff et al., 2015; Deng et al., 2018), in which an agent is presented with a pair of real images and needs to decide whether they are from the same source or not. 2) Replay attacker (Replay): an attacker which upon seeing a leaked sample Y, draws n random images of the leaked sample (with replacement). 3) A GIM attack, which is the "worst case" attacker $\mathcal{G}$, learned by our GIM model.

For VoxCeleb we compare the GIM authenticator to the ArcFace method (Deng et al., 2018), which is currently state of the art in face verification. As a baseline for Omniglot we use the Siamese network suggested by Koch et al. (2015), which achieves state of the art in the verification task on Omniglot. Results are shown in Table 1. It can be seen that on average across attacks, GIM outperforms the baselines. The only attack for which GIM is inferior is RS. This is not surprising as this is the objective that both baselines are trained for.

Qualitative evaluation of attacker: In Fig. 2 we provide images generated by the GIM attacker for the test set of both Omniglot and Voxceleb. The images demonstrate qualitatively the strategy learned by the attacker. In the Voxceleb2 dataset, face images are drawn from a video of the person talking. Note that as in real samples from the data, the attack sample varies in pose and expression and not in the background or clothing.

Table 1: Accuracy of GIM and baselines against attacks. Avg acc denotes average over all attacks.

Authenticator	Dataset	m	n	k	RS	Replay	GIM	**Avg acc**
GIM	VoxCeleb2	1	5	5	0.897	0.837	0.822	**0.852**
ArcFace	VoxCeleb2	1	5	5	0.998	0.598	0.526	0.707
GIM	Omniglot	1	5	5	0.912	0.942	0.868	**0.907**
Siamese	Omniglot	1	5	5	0.994	0.509	0.785	0.763

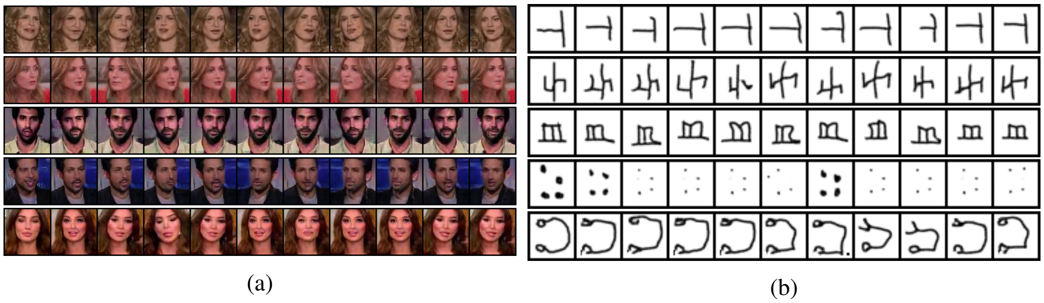


Figure 2: Images generated by the GIM attacker based on one leaked image. In each row the leftmost image is the real leaked image, and the rest of the images are an attack sample generated by the GIM attacker. (a) Voxceleb2 dataset. (b) Omniglot dataset.

Data augmentation: Finally, we use GIM for data augmentation in one-shot classification on Omniglot. This is done by using the GIM attacker to generate more data for a given class. We first train GIM on the training set with parameters $m = 1, n = 5, k = 5$. Then, during both training and testing of one shot classification, given an example, we use the attacker to augment the single example available for each of the classes, by adding to it the $n = 5$ examples our attacker generated from it. We use Prototypical Nets (Snell et al., 2017) as the baseline model. We find that without using our augmentation method, Prototypical Nets achieve 95.9% accuracy on the test split, and with our method they achieve 96.5%, which is similar to the improvement achieved by Antoniou et al. (2018) with Matching networks (Vinyals et al., 2016) as the few shot classification algorithm.

7 Conclusions

We defined the notion of authentication in the face of generative attacks, in which a generative model attempts to produce a fake reality based on observations of reality. These attacks raise numerous interesting theoretical questions and are very important and timely from a practical standpoint. We proposed to study generative attacks as a two-person zero-sum game between attacker and authenticator. In our most general setup both attacker and authenticator have access to a finite set of observations of the source. We show that this game has a Nash equilibrium, and we characterize the optimal strategies. In the Gaussian version of the game a closed form of the optimal strategies is available. A nice outcome of the analysis is that the game value depends on m, n, k only through their ratios $\delta = m/n$ (i.e., the ”expansion ratio” between attack and leaked sample sizes) and $\rho = m/k$ (i.e., the “information ratio” between the number of source observations available to attacker and authenticator). As we show in Fig. 1b, there is a large range of values for which high accuracy authentication is possible, and as d grows we observe that the high authentication accuracy region in the (δ, ρ) plane grows sharply. We introduce the GIM model, which is a practical approach to learning both authenticator and attacker, and whose structure is inspired by our analysis. GIM achieves accuracy that is very close to the theoretical rates in the Gaussian case, and is also more robust to attacks, when compared to state of the art authenticators on real data. Many theoretical and practical questions remain. For example finding closed form optimal strategies for other distributions, and going beyond IID generation. The non IID setting is of particular importance for the problem of fake video (Thies et al., 2016) and audio (Arik et al., 2018) generation, which we intend to study in the future.

REFERENCES

Antreas Antoniou, Amos J. Storkey, and Harrison Edwards. Augmenting image classifiers using data augmentation generative adversarial networks. In *Artificial Neural Networks and Machine Learning*, pp. 594–603, 2018.

Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural voice cloning with a few samples. In *Advances in Neural Information Processing Systems*, pp. 10040–10050, 2018.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 224–232. JMLR. org, 2017.

Ning Bao, O. Patrick Kreidl, and John Musacchio. A network security classification game. In *Game Theory for Networks - 2nd International ICST Conference, GAMENETS 2011, Shanghai, China, April 16-18, 2011, Revised Selected Papers*, pp. 265–280, 2011. doi: 10.1007/978-3-642-30373-9_19. URL https://doi.org/10.1007/978-3-642-30373-9_19.

Mauro Barni and Benedetta Tondi. Multiple-observation hypothesis testing under adversarial conditions. In *2013 IEEE International Workshop on Information Forensics and Security, WIFS 2013, Guangzhou, China, November 18-21, 2013*, pp. 91–96, 2013a. doi: 10.1109/WIFS.2013.6707800. URL https://doi.org/10.1109/WIFS.2013.6707800.

Mauro Barni and Benedetta Tondi. The source identification game: An information-theoretic perspective. *IEEE Trans. Information Forensics and Security*, 8(3):450–463, 2013b. doi: 10.1109/TIFS.2012.2237397. URL https://doi.org/10.1109/TIFS.2012.2237397.

Mauro Barni and Benedetta Tondi. Binary hypothesis testing game with training data. *IEEE Trans. Information Theory*, 60(8):4848–4866, 2014. doi: 10.1109/TIT.2014.2325571. URL https://doi.org/10.1109/TIT.2014.2325571.

Fernando G. S. L. Brandão, Aram Wettroth Harrow, James R. Lee, and Yuval Peres. Adversarial hypothesis testing and a quantum stein's lemma for restricted measurements. In *Innovations in Theoretical Computer Science, ITCS'14, Princeton, NJ, USA, January 12-14, 2014*, pp. 183–194, 2014. doi: 10.1145/2554797.2554816. URL https://doi.org/10.1145/2554797.2554816.

Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, pp. 547–555, 2011. doi: 10.1145/2020408.2020495. URL https://doi.org/10.1145/2020408.2020495.

Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *J. Mach. Learn. Res.*, 13:2617–2654, 2012. URL http://dl.acm.org/citation.cfm?id=2503326.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 539–546, 2005.

Arsha Nagrani Chung, Joon Son and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech*, 2018.

Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018.

Harrison Edwards and Amos Storkey. Towards a neural statistician. In *International Conference on Learning Representations*, 2017.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *CoRR*, abs/1703.03400, 2017. URL http://arxiv.org/abs/1703.03400.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014a.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2014b.

Luke B Hewitt, Maxwell I Nye, Andreea Gane, Tommi Jaakkola, and Joshua B Tenenbaum. The variational homoencoder: Learning to learn high capacity generative models from few examples. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, pp. 988–997, 2018.

Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision*, pp. 1510–1519, 2017.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision*, pp. 172–189, 2018.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *ArXiv*, abs/1905.02175, 2019.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5967–5976, 2017.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018a.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018b.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.

Brenden M. Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua B. Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the 33th Annual Meeting of the Cognitive Science Society*, 2011.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *Proceedings of the 35th International Conference on Machine Learning*, pp. 3478–3487, 2018.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *Interspeech*, 2017.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. *CoRR*, abs/1903.07291, 2019. URL `http://arxiv.org/abs/1903.07291`.

Andrs Prkopa. On logarithmic concave measures and functions. *Acta Scientiarium Mathematicarum*, 34:335–343, 1973.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018.

Danilo Jimenez Rezende, Shakir Mohamed, Ivo Danihelka, Karol Gregor, and Daan Wierstra. One-shot generalization in deep generative models. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pp. 1521–1529, 2016. URL `http://proceedings.mlr.press/v48/rezende16.html`.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

Adi Shamir, Itay Safran, Eyal Ronen, and Orr Dunkelman. A simple explanation for the existence of adversarial examples with small hamming distance. *CoRR*, abs/1901.10861, 2019. URL `http://arxiv.org/abs/1901.10861`.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4080–4090, 2017.

Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 8312–8323. Curran Associates, Inc., 2018. URL `http://papers.nips.cc/paper/8052-constructing-unrestricted-adversarial-examples-with-generative-models.pdf`.

Paul Syverson. A taxonomy of replay attacks [cryptographic protocols]. pp. 187 – 191, 07 1994. ISBN 0-8186-6230-1. doi: 10.1109/CSFW.1994.315935.

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2387–2395, 2016.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pp. 3630–3638, 2016.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. *CoRR*, abs/1711.11585, 2017. URL `http://arxiv.org/abs/1711.11585`.

Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J. Zico Kolter. Scaling provable adversarial defenses. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, pp. 8400–8409. Curran Associates, Inc., 2018.

Eric Wong, Frank R. Schmidt, and J. Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 6808–6817, 2019. URL `http://proceedings.mlr.press/v97/wong19a.html`.

Xiaoyong Yuan, Pan He, Qile Zhu, Rajendra Rana Bhat, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *CoRR*, abs/1712.07107, 2017. URL http://arxiv.org/abs/1712.07107.

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor S. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. *CoRR*, abs/1905.08233, 2019. URL http://arxiv.org/abs/1905.08233.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 7472–7482, 2019. URL http://proceedings.mlr.press/v97/zhang19p.html.

Yan Zhou, Murat Kantarcioglu, and Bowei Xi. A survey of game theoretic approach for adversarial machine learning. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 9(3), 2019. doi: 10.1002/widm.1259. URL https://doi.org/10.1002/widm.1259.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2017.

A Introduction to the appendix

Here we provide additional information and proofs for the main paper. In Sec. B we provide a visualization of the game setting. In Sec. C we plot the game value in the Gaussian case for different parameter values. In Sec. D we provide a proof for Theorem 4.1, in Sec. E we formally state the theorem discussed in Sec. 4.2 and provide proof, and in Sec. F we prove Theorem 4.2 and also derive the game value for the sub-optimal "ML attacker". Finally in Sec. G,H we provide additional details about the experiments and the implementation of GIM.

B Game formulation visualization

Our problem setup is illustrated in Figure 3 for a face authentication scenario. A source θ (in this case an individual) generates IID observations (images). k images are used by the authenticator to study the source, which it aims to authenticate. m images are "leaked" and obtained by an attacker who wishes to impersonate the source and pass the authentication. At test time the authenticator is presented with n images that were generated either by the true source θ, or by the attacker, and decides whether the entity that generated the images was the source θ or an attacker.

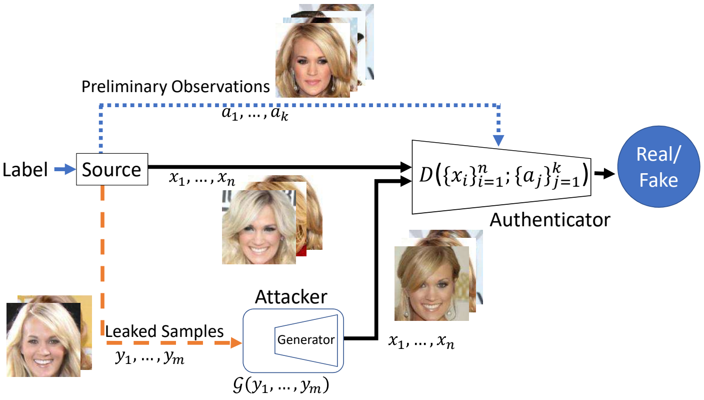


Figure 3: An illustration of the game described in the main text. An authenticator receives a sample of n images and needs to decide whether these were generated by a known source, or by an adversary that had access to leaked examples. In order to decide, the authenticator is supplied with k samples from the real source.

C Game value visualizations

In corollary 4.3 we provide the game value (the expected authenticator accuracy) for multivariate Gaussian sources. In this section we present additional visualizations of the game value as a function of the different parameters of the game. Figure 4 visualizes the game value for different values of d (the observation dimension), as a function of $\delta = \frac{m}{n}$, the "expansion ratio" between the source information available to the attacker through the leaked sample and the size of the attacker's attack sample, and $\rho = \frac{m}{k}$, the information ratio between number of source observations available to attacker and authenticator. One can clearly see that as d, the observation dimension, grows large, so does the accuracy of the authenticator. Even for δ values higher then 0.5, and ρ values which intuitively would give the attacker an advantage (e.g., $\frac{m}{k} = 10$).

Figure 5 visualizes the game value for different values of d (the observation dimension), as a function of $\delta = \frac{m}{n}$ and $\epsilon = \frac{n}{k}$, the"information expansion" of the attacker with respect to the authenticator source information. Again, one can clearly see that as d, the observation dimension, grows large, so does the accuracy of the authenticator.

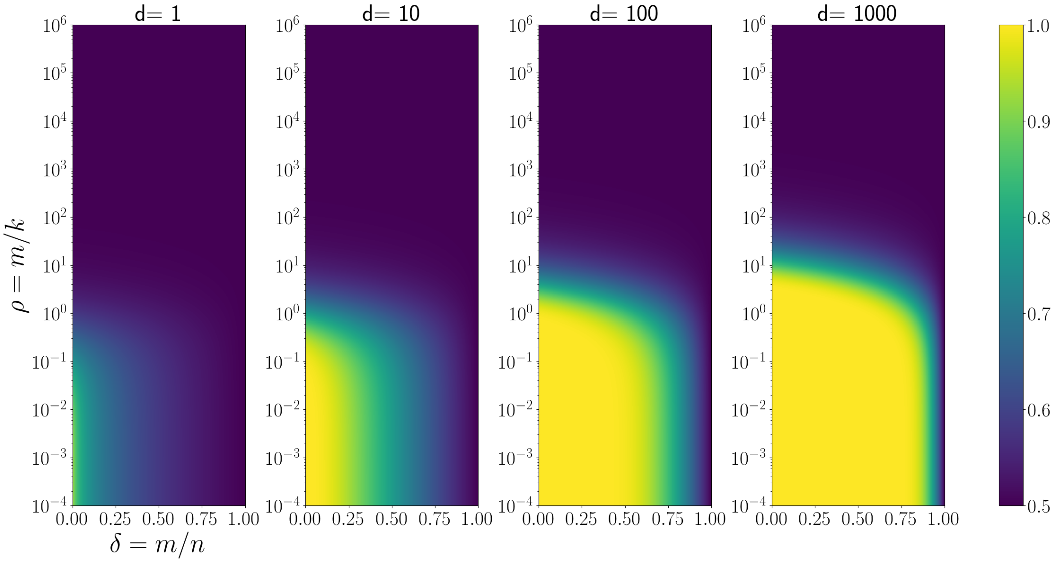


Figure 4: Game value as a function of δ, ρ for different dimensions d.

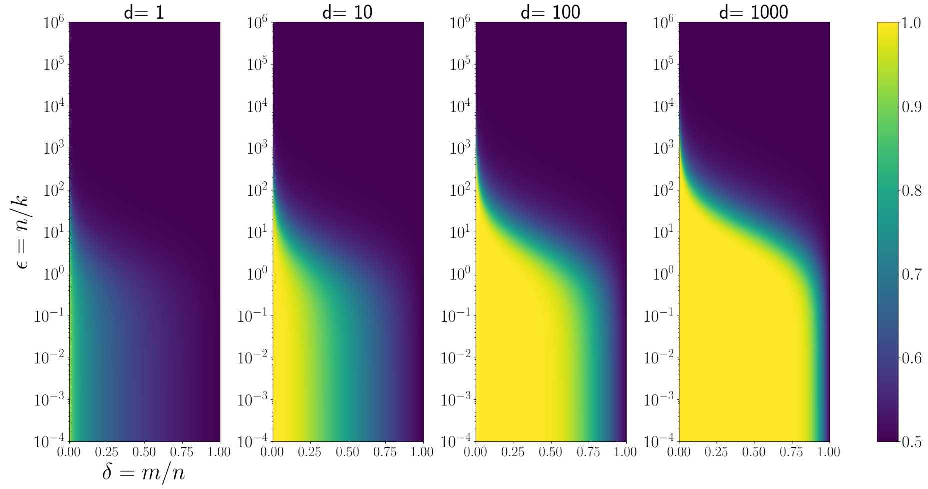


Figure 5: Game value as a function of δ, ϵ for different dimensions d.

D Additional theorems and proofs for Sec. 4.1

We begin with some additional notation. Let

$$f_A(a) = \int_{\theta \in \mathcal{H}} Q(\theta) f_\theta^{(k)}(a) d\theta$$

denote the marginal density of A, let $Q_{\Theta|A}$ denote the posterior probability over $\mathcal{H}$ given A. That is:

$$Q_{\Theta|A}(\theta|a) = \frac{Q(\theta) f_{\theta}^{(k)}(a)}{\int_{\nu \in \mathcal{H}} Q(\nu) f_{\nu}^{(k)}(a) d\nu} \equiv \frac{Q(\theta) f_{\theta}^{(k)}(a)}{f_A(a)}$$

Also let $f_{Y|A}, f_{X|A}, g_{X|A}$, denote the conditional densities defined by:

$$\begin{aligned} f_{Y|A}(y|a) &= \int_{\theta \in \mathcal{H}} f_{\theta}^{(m)}(y) Q_{\Theta|A}(\theta|a) d\theta \\ f_{X|A}(x|a) &= \int_{\theta \in \mathcal{H}} f_{\theta}^{(n)}(x) Q_{\Theta|A}(\theta|a) d\theta \\ g_{X|A}(x|a) &= \int_{y \in \mathcal{X}^m} g_{X|Y}(x|y) f_{Y|A}(y|a) dy \end{aligned}$$

Lemma D.1. *Let $\mathcal{G}$ be an attacker defined by the conditional probability distribution $g_{X|Y}$. Then $\forall a \in \mathcal{X}^k, x \in \mathcal{X}^n$ a best response strategy for $\mathcal{D}$ is:*

$$\mathcal{D}(a, x) = I[f_{X|A}(x|a) > g_{X|A}(x|a)] \tag{D.1}$$

Proof. Given an attacker strategy $g_{X|Y}$, the objective for $\mathcal{D}$ is given by:

$$\begin{aligned} &\operatorname*{argmax}_{\mathcal{D}} \mathbb{E}_{\Theta \sim Q} V(\Theta, \mathcal{D}, \mathcal{G}) \\ &= \operatorname*{argmax}_{\mathcal{D}} \mathbb{E}_{\Theta \sim Q} \mathbb{E}_{A \sim f_{\Theta}^{(k)}} [\mathbb{E}_{X \sim f_{\Theta}^{(n)}} [\mathcal{D}(A, X)] + \mathbb{E}_{Y \sim f_{\Theta}^{(m)}} \mathbb{E}_{X \sim g_{X|Y}(\cdot|y)} [1 - \mathcal{D}(A, X)]] \\ &= \operatorname*{argmax}_{\mathcal{D}} \mathbb{E}_{A \sim f_A} [\mathbb{E}_{X \sim f_{X|A}(\cdot|A)} [\mathcal{D}(A, X)] + \mathbb{E}_{Y \sim f_{Y|A}(\cdot|A)} \mathbb{E}_{X \sim g_{X|Y}(\cdot|Y)} [1 - \mathcal{D}(A, X)]] \\ &= \operatorname*{argmax}_{\mathcal{D}} \int_{a \in \mathcal{X}^k} f_A(a) \int_{x \in \mathcal{X}^n} f_{X|A}(x|a) \mathcal{D}(a, x) + g_{X|A}(x|a)[1 - \mathcal{D}(a, x)] dx] \end{aligned}$$

Note that $\mathcal{D}$ can be optimized independently for each pair $a, x \in \mathcal{X}^k \times \mathcal{X}^n$. Hence, $\forall a, x \in \mathcal{X}^k \times \mathcal{X}^n$ the objective is:

$$\operatorname*{argmax}_{\mathcal{D}(a,x) \in \{0,1\}} \{f_{X|A}(x|a) \mathcal{D}(a, x) + g_{X|A}(x|a)[1 - \mathcal{D}(a, x)]\}$$

And thus the optimal decision rule for $\mathcal{D}$ is:

$$\mathcal{D}(a, x) = I[f_{X|A}(x|a) > g_{X|A}(x|a)]$$

As required. □

Lemma D.2. *$\forall a, x \in \mathcal{X}^k \times \mathcal{X}^n$ let the strategy for $\mathcal{D}$ be defined by:*

$$\mathcal{D}(a, x) = I[f_{X|A}(x|a) > g_{X|A}(x|a)]$$

Then $\mathcal{G}^*$ is a best response strategy for $\mathcal{G}$, iff it minimizes the ℓ_1 distance between the distributions $f_{X|A}, g_{X|A}$ over the space $\mathcal{X}^k \times \mathcal{X}^n$. Namely:

$$g^*_{X|Y} \in \operatorname*{argmin}_{g_{X|Y}} \mathbb{E}_{A \sim f_A} \int_{x \in \mathcal{X}^n} \left| f_{X|A}(x|A) - g_{X|A}(x|A) \right| dx \tag{D.2}$$

Proof. Let $\mathcal{D}$ be defined as in Eq. D.1, the objective for $\mathcal{G}$ is:

$$
\begin{aligned}
&\operatorname*{argmin}_{\mathcal{G}} \frac{1}{2}\mathbb{E}_{\theta\sim Q} V(\Theta,\mathcal{D},\mathcal{G}) \\
&= \operatorname*{argmin}_{g_{X|Y}} \mathbb{E}_{\theta\sim Q}\mathbb{E}_{A\sim f_{\Theta}^{(k)}}[\mathbb{E}_{X\sim f_{\Theta}^{(n)}}[\mathcal{D}(A,X)] + \mathbb{E}_{Y\sim f_{\Theta}^{(m)}}\mathbb{E}_{X\sim g_{X|Y}(\cdot|Y)}[1-\mathcal{D}(A,X)]] \\
&= \operatorname*{argmin}_{g_{X|Y}} \mathbb{E}_{A\sim f_A}[\mathbb{E}_{X\sim f_{X|A}(\cdot|A)}[\mathcal{D}(A,X)] + \mathbb{E}_{X\sim g_{X|A}(\cdot|A)}[1-\mathcal{D}(A,X)]] \\
&= \operatorname*{argmin}_{g_{X|Y}} \int_{a\in\mathcal{X}^k} da f_A(a) \int_{x\in\mathcal{X}^n} dx\{f_{X|A}(x|a)\mathcal{D}(a,x) + g_{X|A}(x|a)[1-\mathcal{D}(a,x)]\} \\
&= \operatorname*{argmin}_{g_{X|Y}} \int_{a\in\mathcal{X}^k} da f_A(a) \int_{x\in\mathcal{X}^n} dx \max\{f_{X|A}(x|a), g_{X|A}(x|a)\} \qquad \text{(D.3)} \\
&= \operatorname*{argmin}_{g_{X|Y}} \frac{1}{2}\int_{a\in\mathcal{X}^k} da f_A(a) \int_{x\in\mathcal{X}^n} dx[f_{X|A}(x|a) + g_{X|A}(x|a) + \left|f_{X|A}(x|a) - g_{X|A}(x|a)\right|] \\
&= \operatorname*{argmin}_{g_{X|Y}} 1 + \frac{1}{2}\int_{a\in\mathcal{X}^k} da f_A(a) \int_{x\in\mathcal{X}^n} dx \left|f_{X|A}(x|a) - g_{X|A}(x|a)\right| \\
&= \operatorname*{argmin}_{g_{X|Y}} \frac{1}{2}\int_{a\in\mathcal{X}^k} da f_A(a) \int_{x\in\mathcal{X}^n} dx \left|f_{X|A}(x|a) - g_{X|A}(x|a)\right| \\
&= \operatorname*{argmin}_{g_{X|Y}} \int_{a\in\mathcal{X}^k} da f_A(a) \int_{x\in\mathcal{X}^n} dx \left|f_{X|A}(x|a) - g_{X|A}(x|a)\right|
\end{aligned}
\qquad \text{(D.4)}
$$

As required. Where in Eq. D.3 we used the definition of $\mathcal{D}$. □

Theorem 4.1. *Consider the attacker defined by:*

$$
g^*_{X|Y} \in \operatorname*{argmin}_{g_{X|Y}} \mathbb{E}_{A\sim f_A}\left[\int_{x\in\mathcal{X}^n} \left|f_{X|A}(x|A) - g_{X|A}(x|A)\right| dx\right] \qquad \text{(D.5)}
$$

and let $\mathcal{G}^*$ be the corresponding map from $\mathcal{X}^m$ to the set of probability distributions over $\mathcal{X}^n$. Consider the authenticator defined by:

$$
\mathcal{D}^*(a,x) = I\left[f_{X|A}(x|a) > g^*_{X|A}(x|a)\right] \qquad \text{(D.6)}
$$

where I is the indicator function. Then $(\mathcal{D}^*,\mathcal{G}^*)$ is a solution of Eq. 2.2 that satisfies Eq. 4.1.

Proof. From Lemmas D.1, D.2, we have that $\max_{\mathcal{D}} V(\mathcal{D},\mathcal{G}^*) = V(\mathcal{D}^*,\mathcal{G}^*) = \min_{\mathcal{G}} V(\mathcal{D}^*,\mathcal{G})$ form which it follows that Eq. 4.1 is satisfied and thus $(\mathcal{D}^*,\mathcal{G}^*)$ is a solution of Eq. 2.2. □

E Theorem and proof for Sec. 4.2

Theorem E.1. *For all $n \leqslant m$ it holds that:*

$$
\max_{\mathcal{D}\in\mathbb{D}} \min_{\mathcal{G}\in\mathbb{G}} V(\mathcal{D},\mathcal{G}) = 0.5
$$

Proof. Consider the attacker $\mathcal{G}_{\text{replay}}$ defined by the following generative process: Given a leaked sample $Y \in \mathbb{R}^{m\times d}$, $\mathcal{G}_{\text{replay}}$ generates a sample $X \in \mathbb{R}^{n\times d}$ such that $X_i = Y_i \quad \forall i \in [n]$. (This is possible since we assumed $n \leqslant m$). Namely, we have:

$$
g_{X|Y}(x|y) = \prod_{i=1}^{n} \delta(x_i - y_i)
$$

Where δ is the Dirac delta. Thus $\forall a, x \in \mathcal{X}^k \times \mathcal{X}^n$:

$$
\begin{aligned}
g_{X|A}(x|a) =& \int_{y \in \mathcal{X}^m} dy g_{X|Y}(x|y) f_{Y|A}(y|a) \\
=& \int_{y \in \mathcal{X}^m} dy g_{X|Y}(x|y) \int_{\theta \in \mathcal{H}} d\theta f_\theta^{(m)}(y) Q_{\Theta|A}(\theta|a) \\
=& \int_{\theta \in \mathcal{H}} d\theta \int_{y \in \mathcal{X}^m} dy \prod_{i=1}^{n} \delta(x_i - y_i) f_\theta^{(m)}(y) Q_{\Theta|A}(\theta|a) \\
=& \int_{\theta \in \mathcal{H}} d\theta Q_{\Theta|A}(\theta|a) \int_{y \in \mathcal{X}^n} dy \prod_{i=1}^{n} \delta(x_i - y_i) f_\theta^{(n)}(y) \int_{y' \in \mathcal{X}^{m-n}} dy' f_\theta^{(m-n)}(y') \\
=& \int_{\theta \in \mathcal{H}} d\theta Q_{\Theta|A}(\theta|a) \int_{y \in \mathcal{X}^n} dy \prod_{i=1}^{n} \delta(x_i - y_i) f_\theta^{(n)}(y) \\
=& \int_{\theta \in \mathcal{H}} d\theta Q_{\Theta|A}(\theta|a) f_\theta^{(n)}(x) \\
=& f_{X|A}(x|a)
\end{aligned}
$$

□

Define:

$$\mathcal{D}_0(a, x) = 0 \quad \forall a, x \in \mathcal{X}^k \times \mathcal{X}^n$$

Then according to Theorem 4.1 $(\mathcal{D}_0, \mathcal{G}_{\text{replay}})$ is a solution of Eq. 2.2 that satisfies Eq. 4.1, and therefore:

$$
\begin{aligned}
\max_{\mathcal{D} \in \mathbb{D}} \min_{\mathcal{G} \in \mathbb{G}} V(\mathcal{D}, \mathcal{G}) =& V(\mathcal{D}_0, \mathcal{G}_{\text{replay}}) \\
=& \frac{1}{2} \mathbb{E}_{\Theta \sim Q} \mathbb{E}_{A \sim f_\theta^{(k)}} \mathbb{E}_{Y \sim f_\theta^{(m)}} \left[\mathbb{E}_{X \sim f_\theta^{(n)}} [\mathcal{D}(A, X)] + \mathbb{E}_{X \sim \mathcal{G}(Y)} [1 - \mathcal{D}(A, X)] \right] \\
=& \frac{1}{2} \mathbb{E}_{\Theta \sim Q} \mathbb{E}_{A \sim f_\theta^{(k)}} \mathbb{E}_{Y \sim f_\theta^{(m)}} \left[\mathbb{E}_{X \sim f_\theta^{(n)}} [0] + \mathbb{E}_{X \sim \mathcal{G}(Y)} [1] \right] \\
=& \frac{1}{2}
\end{aligned}
$$

As required.

F Additional theorems and proofs for Sec. 4.3

F.1 Notation and definitions

In this section we consider the case where the sources are d-dimensional Gaussian vectors with a known covariance matrix $\Sigma = CC^T \in \mathbb{R}^{d \times d}$ and unknown mean vector $\theta \in \mathbb{R}^d$. That is the set of possible sources is $\mathcal{H} = \mathbb{R}^d$, and given $\theta \in \mathcal{H}$ the associated probability density over the domain $\mathcal{X} = \mathbb{R}^d$ is $f_\theta(x) = \frac{1}{\sqrt{(2\pi)^d |\det(\Sigma)|}} \exp\left(-\frac{1}{2}(x - \theta)^T \Sigma^{-1} (x - \theta)\right)$

A sample of n examples $x \in \mathcal{X}^n$, is considered a matrix $x \in \mathbb{R}^{n \times d}$, where the first dimension is the sample and the second is the dimension in the observation space, $\mathbb{R}^d$. i.e. $x_{ij} \in \mathbb{R}$ is the j'th element of the i'th example in the sample $x \in \mathbb{R}^{n \times d}$.

We continue with a few more notations that simplify the proofs. Given a matrix $x \in \mathbb{R}^{n \times d}$, we let $x_c = \begin{bmatrix} x_1^T, & \dots, & x_n^T \end{bmatrix}^T \in \mathbb{R}^{nd}$ be the concatenation vector representing x. Given a vector $\theta \in \mathbb{R}^d$ we let $\theta_{c,n} = \begin{bmatrix} \theta^T, & \dots, & \theta^T \end{bmatrix}^T \in \mathbb{R}^{nd}$ be the concatenation of n copies of θ. Given a matrix $x \in \mathbb{R}^{n \times d}$ we let $\bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i \in \mathbb{R}^d$ denote its mean along the sample dimension. For any matrix $B \in \mathbb{R}^{d \times d}$, we denote:

$$diag(B, k) = \begin{bmatrix} B & & 0 \\ & \ddots & \\ 0 & & B \end{bmatrix} \in \mathbb{R}^{kd \times kd}$$

and

$$rep(B,k) = \begin{bmatrix} B & \cdots & B \\ \vdots & \ddots & \vdots \\ B & \cdots & B \end{bmatrix} \in \mathbb{R}^{kd \times kd}$$

Finally we define strategies for both attacker and authenticator which we prove in what follows to be the optimal strategies for the game.

Definition F.1. *Let $\mathcal{G}^*$ denote an attacker defined by its associated conditional probability $g^*_{X|Y}$ as follows: Given a leaked sample $Y \in \mathbb{R}^{m \times d}$, $\mathcal{G}^*$ generates an attack sample $X \in \mathbb{R}^{n \times d}$ as follows. It first samples n vectors $W_1, \ldots, W_n \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$ and then sets:*

$$X_i = W_i - \bar{W} + \bar{Y} \tag{F.1}$$

Definition F.2. *For any $\alpha \in \mathbb{R}_+$, let $\mathcal{D}_\alpha$ denote an authenticator defined as:*

$$\mathcal{D}_\alpha(a, x) = I[\|\bar{x} - \bar{a}\|^2_{\Sigma^{-1}} < \alpha]$$

Where I is the indicator function

F.2 Technical lemmas

Lemma F.3. *Let $X_1, \ldots, X_k \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$ Where $\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}$. Then*

$$\bar{X} = \frac{1}{k} \sum_{j=1}^{k} X_j \sim \mathcal{N}(\mu, \frac{1}{k}\Sigma)$$

Proof. We begin by observing that: $X_c \sim \mathcal{N}(\mu_{c,k}, diag(\Sigma, k))$. Let $B = \frac{1}{k}\begin{bmatrix} I_d & \cdots & I_d \end{bmatrix} \in \mathbb{R}^{d \times kd}$ and observe that $\bar{X} = BX_c$. Therefore, since this is an affine transformation of a Gaussian vector we have:

$$\bar{X} \sim \mathcal{N}(B\mu_{c,k}, Bdiag(\Sigma, k)B^T) = \mathcal{N}(\mu, \frac{1}{k}\Sigma)$$

As required. □

Lemma F.4. *Let $X \in \mathbb{R}^d$ be a Gaussian vector s.t $X \sim \mathcal{N}(\mu, \Sigma)$, and let $X_{c,n} \in \mathbb{R}^{nd}$ be the concatenation of n copies of X. Then:*

$$X_{c,n} \sim \mathcal{N}(\mu_{c,n}, rep(\Sigma, n))$$

Proof. Let

$$B = \begin{bmatrix} I_d \\ \vdots \\ I_d \end{bmatrix} \in \mathbb{R}^{nd \times d}$$

And observe that $X_{c,n} = BX$. Therefore, since this is an affine transformation of a gaussuian vector we have:

$$X_{c,n} \sim \mathcal{N}(B\mu, B\Sigma B^T) = \mathcal{N}(\mu_{c,n}, rep(\Sigma, n))$$

As required. □

Lemma F.5. *Let $\theta \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}$ represent the mean and covariance of a Gaussian distribution. Let $X \in \mathbb{R}^{n \times d}$ be a random sample generated by the attacker defined in Def. F.1. Then:*

$$X_c \sim \mathcal{N}(\theta_{c,n}, diag(\Sigma, n) + rep((\frac{n-m}{mn})\Sigma, n))$$

Proof. Observe that $W_c \sim \mathcal{N}(0, diag(\Sigma, n))$. Using Lemma F.3 we get $\bar{Y} \sim \mathcal{N}(\theta, \frac{1}{m}\Sigma)$ and observe that: $\bar{W}_{c,n} = rep(\frac{1}{n}I_d, n)W_c$. and using Lemma F.4 we get: $\bar{Y}_{c,n} \sim \mathcal{N}(\theta_{c,n}, \frac{1}{m}rep(\Sigma, n))$. We define the following block matrices

$$Z = \begin{bmatrix} W_c \\ \bar{Y}_{c,n} \end{bmatrix}, \quad B = \begin{bmatrix} I_{nd} - \frac{1}{n}rep(I_d, n), & I_{nd} \end{bmatrix}$$

and observe that:

$$Z \sim \mathcal{N}(\begin{bmatrix} 0_{nd} \\ \theta_{c,n} \end{bmatrix}, \begin{bmatrix} diag(\Sigma, n) & 0 \\ 0 & \frac{1}{m}rep(\Sigma, n) \end{bmatrix})$$

Note that $X_c = W_c - \bar{W}_{c,n} + \bar{Y}_{c,n} = BZ$, and therefore we get:

$$\begin{aligned} X_c &\sim \mathcal{N}(B\begin{bmatrix} 0_{nd} \\ \theta_{c,n} \end{bmatrix}, B\begin{bmatrix} diag(\Sigma, n) & 0 \\ 0 & \frac{1}{m}rep(\Sigma, n) \end{bmatrix} B^T) \\ &= \mathcal{N}(\theta_{c,n}, diag(\Sigma, n) + rep((\frac{n-m}{mn})\Sigma, n, n)) \end{aligned}$$

as required. □

Lemma F.6. *Let $\Sigma = CC^T \in \mathbb{R}^{d \times d}$ represent the covariance of a Gaussian distribution, and consider the following covariance matrix:*

$$\Psi = diag(\Sigma, n) + rep((\frac{n-m}{mn})\Sigma, n)$$

Then its inverse is:

$$\Psi^{-1} = diag(\Sigma^{-1}, n) - \frac{n-m}{n^2}rep(\Sigma^{-1}, n)$$

And the determinant is:

$$\det(\Psi) = (\frac{n}{m})^d \det(\Sigma)^n$$

Proof. We begin by defining the following block matrices:

$$U = \begin{bmatrix} \Sigma \\ \vdots \\ \Sigma \end{bmatrix} \in \mathbb{R}^{nd \times d} \tag{F.2}$$

$$V = (\frac{n-m}{mn}) \begin{bmatrix} I_d \\ \vdots \\ I_d \end{bmatrix} \in \mathbb{R}^{nd \times d} \tag{F.3}$$

And note that $rep((\frac{n-m}{mn})\Sigma, n) = UV^T$. Then:

$$\begin{aligned}
\Psi^{-1} &= (diag(\Sigma, n) + rep((\frac{n-m}{mn})\Sigma, n))^{-1} \\
&= (diag(\Sigma, n) + UV^T)^{-1} \\
&\overset{(*)}{=} diag(\Sigma, n)^{-1} - diag(\Sigma, n)^{-1}U(I_d + V^T diag(\Sigma, n)^{-1}U)^{-1}V^T diag(\Sigma, n)^{-1} \\
&\overset{(**)}{=} diag(\Sigma^{-1}, n) - diag(\Sigma^{-1}, n)U(I_d + V^T diag(\Sigma^{-1}, n)U)^{-1}V^T diag(\Sigma^{-1}, n) \\
&= diag(\Sigma^{-1}, n) - \begin{bmatrix} I_d \\ \vdots \\ I_d \end{bmatrix} (I_d + \frac{n-m}{nm} \begin{bmatrix} I_d & \cdots & I_d \end{bmatrix} \begin{bmatrix} I_d \\ \vdots \\ I_d \end{bmatrix})^{-1} \frac{n-m}{nm} \begin{bmatrix} \Sigma^{-1} & \cdots & \Sigma^{-1} \end{bmatrix} \\
&= diag(\Sigma^{-1}, n) - \begin{bmatrix} I_d \\ \vdots \\ I_d \end{bmatrix} \frac{m}{n} I_d \frac{n-m}{nm} \begin{bmatrix} \Sigma^{-1} & \cdots & \Sigma^{-1} \end{bmatrix} \\
&= diag(\Sigma^{-1}, n) - \frac{n-m}{n^2} \begin{bmatrix} I_d \\ \vdots \\ I_d \end{bmatrix} \begin{bmatrix} \Sigma^{-1} & \cdots & \Sigma^{-1} \end{bmatrix} \\
&= diag(\Sigma^{-1}, n) - \frac{n-m}{n^2} rep(\Sigma^{-1}, n)
\end{aligned}$$

As required. Where in $(*)$ we used the Woodbury matrix identity, and in $(**)$ we used the inverse of a diagonal block matrix.

Next, we turn to finding the determinant of Ψ:

$$\begin{aligned}
det(\Psi) &= det(diag(\Sigma, n) + rep((\frac{n-m}{mn})\Sigma, n)) \\
&= det(diag(\Sigma, n) + UV^T) \\
&\overset{(*)}{=} det(diag(\Sigma, n))det(I_d + V^T diag(\Sigma, n)^{-1}U) \\
&\overset{(**)}{=} det(\Sigma)^n det(I_d + V^T diag(\Sigma^{-1}, n)U) \\
&= det(\Sigma)^n det(I_d + \frac{n-m}{nm} \begin{bmatrix} I_d & \cdots & I_d \end{bmatrix} diag(\Sigma^{-1}, n) \begin{bmatrix} \Sigma \\ \vdots \\ \Sigma \end{bmatrix}) \\
&= det(\Sigma)^n det(I_d + \frac{n-m}{m} I_d) \\
&= det(\Sigma)^n (\frac{n}{m})^d
\end{aligned}$$

Where in $(*)$ we used the matrix determinant lemma, and in $(**)$ we used the determinant of a diagonal block matrix. □

Lemma F.7. *Let*

$$h(x, \mu) = \exp\{-\frac{1}{2\sigma^2}(x-\mu)^T\Sigma^{-1}(x-\mu)\} I[\{x \in \mathbb{R}^d : x^T\Sigma^{-1}x < \alpha\}] \quad \forall x, \mu \in \mathbb{R}^d$$

And define the function:

$$\psi(\mu) = \int_{\{x \in \mathbb{R}^d : x^T\Sigma^{-1}x < \alpha\}} dx\, h(x, \mu)$$

Then $\psi(\mu)$ is log-concave over the space $\mathbb{R}^d$.

Proof. We begin by noting that the function:

$$(x-\mu)^T\Sigma^{-1}(x-\mu)$$

is convex w.r.t μ, hence its negative is concave, and by definition the function:

$$\exp\{-\frac{1}{2\sigma^2}(x-\mu)^T\Sigma^{-1}(x-\mu)\} \tag{F.4}$$

is log-concave w.r.t μ. We now show that $h(x,\mu)$ is log concave w.r.t μ. Let $\beta \in [0,1]$, then we have that:

$$\begin{aligned}
&h(\beta\mu_1 + (1-\beta)\mu_2)\\
&= \exp\{-\frac{1}{2\sigma^2}(x-\beta\mu_1-(1-\beta)\mu_2)^T\Sigma^{-1}(x-\beta\mu_1-(1-\beta)\mu_2)\}I[\{x \in \mathbb{R}^d : x^T\Sigma^{-1}x < \alpha\}]\\
&\overset{(*)}{\geqslant} \exp\{-\frac{\beta}{2\sigma^2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\}\exp\{-\frac{(1-\beta)}{2\sigma^2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\}I[\{x \in \mathbb{R}^d : x^T\Sigma^{-1}x < \alpha\}]\\
&\overset{(**)}{=} \exp\{-\frac{\beta}{2\sigma^2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\}(I[\{x \in \mathbb{R}^d : x^T\Sigma^{-1}x < \alpha\}])^{\beta}(I[\{x \in \mathbb{R}^d : x^T\Sigma^{-1}x < \alpha\}])^{(1-\beta)}\\
&= h(\mu_1)^{\beta}h(\mu_2)^{(1-\beta)}
\end{aligned}$$

Therefore $h(\mu)$ is log-concave. Where in $(*)$ we used Eq. F.4, and in $(**)$ we used the fact that $I[\{x \in \mathbb{R}^d : x^T\Sigma^{-1}x < \alpha\}] \in \{0,1\}$.

Finally, by using Prekopa-Leindler inequality (Prkopa, 1973) we have that $\psi(\mu)$ is log concave, as required. □

F.3 Proof of Theorem 4.2

Lemma F.8. *Consider the attacker $\mathcal{G}^*$, defined in Def. F.1 The best response strategy for the authenticator against this attacker is:*

$$\mathcal{D}^*(a,x) = I[\|\bar{x}-\bar{a}\|^2_{\Sigma^{-1}} < \alpha^*]$$

Where

$$\alpha^* = \frac{d(m+k)(n+k)}{k^2(n-m)}\log\frac{n(m+k)}{m(n+k)}$$

Proof. The best response authenticator satisfies:

$$\begin{aligned}
&\mathcal{D}^* \in \underset{\mathcal{D}\in\mathbb{D}}{\operatorname{argmax}}\, V(\mathcal{D},\mathcal{G})\\
&\quad = \underset{\mathcal{D}\in\mathbb{D}}{\operatorname{argmax}}\, \frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_\Theta^{(k)}}\mathbb{E}_{Y\sim f_\Theta^{(m)}}[\mathbb{E}_{X\sim f_\Theta^{(n)}}[\mathcal{D}(A,X)] + \mathbb{E}_{X\sim g_{X|Y}(\cdot|Y)}[1-\mathcal{D}(A,X)]]\\
&\quad = \underset{\mathcal{D}\in\mathbb{D}}{\operatorname{argmax}}\, \mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_\Theta^{(k)}}[\mathbb{E}_{X\sim f_\Theta^{(n)}}[\mathcal{D}(A,X)] + \mathbb{E}_{X\sim g_{X|\Theta}(\cdot|\Theta)}[1-\mathcal{D}(A,X)]]\\
&\quad \overset{F.5}{=} \underset{\mathcal{D}\in\mathbb{D}}{\operatorname{argmax}}\, \mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim\mathcal{N}(\Theta_{c,k},diag(\Sigma,k))}[\mathbb{E}_{X\sim\mathcal{N}(\Theta_{c,n},diag(\Sigma,n))}[\mathcal{D}(A,X)] + \mathbb{E}_{X\sim\mathcal{N}(\theta_{c,n},\Psi)}[1-\mathcal{D}(A,X)]]\\
&\quad \overset{F.6}{=} \underset{\mathcal{D}\in\mathbb{D}}{\operatorname{argmax}}\, \mathbb{E}_{\Theta\sim Q}\int_{a\in\mathbb{R}^{kd}} da \int_{x\in\mathbb{R}^{nd}} dx \exp\{-\frac{1}{2}(a-\Theta_{c,k})^T diag(\Sigma,k)^{-1}(a-\Theta_{c,k})\}[\\
&\qquad \exp\{-\frac{1}{2}(x-\Theta_{c,n})^T diag(\Sigma,n)^{-1}(x-\Theta_{c,n})\}\mathcal{D}(a,x)+\\
&\qquad \sqrt{\frac{1}{(\frac{n}{m})^d}}\exp\{-\frac{1}{2}(x-\Theta_{c,n})^T\Psi^{-1}(x-\Theta_{c,n})\}[1-\mathcal{D}(a,x)]]
\end{aligned}$$

$\mathcal{D}$ can be chosen independently for each pair $(a,x) \in \mathcal{X}^k \times \mathcal{X}^n$ therefore for any $(a,x) \in \mathcal{X}^k \times \mathcal{X}^n$ the decision rule for $\mathcal{D}(a,x) = 1$ is:

$$\begin{aligned}
&\sqrt{(\frac{n}{m})^d}\int_{\theta\in\mathbb{R}^d} Q(\theta)\exp\{-\frac{1}{2}(x-\theta_{c,n})^T diag(\Sigma,n)^{-1}(x-\theta_{c,n})\}\exp\{-\frac{1}{2}(a-\theta_{c,k})^T diag(\Sigma,k)^{-1}(a-\theta_{c,k})\}\\
&> \int_{\theta\in\mathbb{R}^d} Q(\theta)\exp\{-\frac{1}{2}(x-\theta_{c,n})^T\Psi^{-1}(x-\theta_{c,n})\}\exp\{-\frac{1}{2}(a-\theta_{c,k})^T diag(\Sigma,k)^{-1}(a-\theta_{c,k})\}
\end{aligned}$$

Observing the LHS and using the improper uniform prior assumption we have:

$$
\begin{aligned}
&\sqrt{(\frac{n}{m})^d}\int_{\theta\in\mathbb{R}^d} Q(\theta)\exp\{-\frac{1}{2}(x-\theta_{c,n})^T diag(\Sigma,n)^{-1}(x-\theta_{c,n})\}\exp\{-\frac{1}{2}(a-\theta_{c,k})^T diag(\Sigma,k)^{-1}(a-\theta_{c,k})\}\\
=&\sqrt{(\frac{n}{m})^d}\int_{\theta\in\mathbb{R}^d} \exp\{-\frac{1}{2}[\sum_{i=1}^{n}(x_i-\theta)^T\Sigma^{-1}(x-\theta)+\sum_{j=1}^{k}(a_j-\theta)^T\Sigma^{-1}(a_j-\theta)]\}\\
\equiv&\sqrt{(\frac{n}{m})^d}\int_{\theta\in\mathbb{R}^d} \exp\{-\frac{1}{2}[\sum_{i=1}^{n+k}(z_i-\theta)^T\Sigma^{-1}(z_i-\theta)]\}\\
=&\sqrt{(\frac{n}{m})^d}\int_{\theta\in\mathbb{R}^d} \exp\{-\frac{1}{2}[\sum_{i=1}^{n+k}\sum_{r=1}^{d}\sum_{s=1}^{d}(z_{ir}-\theta_r)\Sigma_{rs}^{-1}(z_{is}-\theta_s)]\}\\
=&\sqrt{(\frac{n}{m})^d}\int_{\theta\in\mathbb{R}^d} \exp\{-\frac{1}{2}[\sum_{r=1}^{d}\sum_{s=1}^{d}\Sigma_{rs}^{-1}\sum_{i=1}^{n+k}(z_{ir}z_{is}-z_{ir}\theta_s-\theta_r z_{is}+\theta_r\theta_s)]\}\\
=&\sqrt{(\frac{n}{m})^d}\int_{\theta\in\mathbb{R}^d} \exp\{-\frac{1}{2}(n+k)[\theta^T\Sigma^{-1}\theta-2\theta^T\Sigma^{-1}\bar{z}+\frac{1}{n+k}\sum_{r=1}^{d}\sum_{s=1}^{d}\Sigma_{rs}^{-1}\sum_{i=1}^{n+k}z_{ir}z_{is}]\}\\
=&\sqrt{(\frac{n}{m})^d}\exp\{-\frac{1}{2}[(\sum_{r=1}^{d}\sum_{s=1}^{d}\Sigma_{rs}^{-1}\sum_{i=1}^{n+k}z_{ir}z_{is})-(n+k)\bar{z}^T\Sigma^{-1}\bar{z}]\}\int_{\theta\in\mathbb{R}^d}\exp\{-\frac{1}{2}(n+k)[(\theta-\bar{z})^T\Sigma^{-1}(\theta-\bar{z})]\}\\
=&\sqrt{(\frac{n}{m})^d}\exp\{-\frac{1}{2}[(\sum_{r=1}^{d}\sum_{s=1}^{d}\Sigma_{rs}^{-1}\sum_{i=1}^{n+k}z_{ir}z_{is})-(n+k)\bar{z}^T\Sigma^{-1}\bar{z}]\}\sqrt{(2\pi)^d\det(\frac{\Sigma}{n+k})}
\end{aligned}
$$

Observing the RHS and using the improper uniform prior assumption we have:

$$
\begin{aligned}
&\int_{\theta\in\mathbb{R}^d} Q(\theta)\exp\{-\frac{1}{2}(x-\theta_{c,n})^T\Psi^{-1}(x-\theta_{c,n})\}\exp\{-\frac{1}{2}(a-\theta_{c,k})^T diag(\Sigma^{-1},k)(a-\theta_{c,k})\}\\
=&\int_{\theta\in\mathbb{R}^d} \exp\{-\frac{1}{2}[(x-\theta_{c,n})^T\Psi^{-1}(x-\theta_{c,n})+(a-\theta_{c,k})^T diag(\Sigma^{-1},k)(a-\theta_{c,k})]\}\\
\overset{F.6}{=}&\int_{\theta\in\mathbb{R}^d} \exp\{-\frac{1}{2}[(x-\theta_{c,n})^T(diag(\Sigma^{-1},n)-\frac{n-m}{n^2}rep(\Sigma^{-1},n))(x-\theta_{c,n})+(a-\theta_{c,k})^T diag(\Sigma^{-1},k)(a-\theta_{c,k})]\}\\
=&\int_{\theta\in\mathbb{R}^d} \exp\{-\frac{1}{2}[\sum_{i=1}^{n}(x_i-\theta)^T\Sigma^{-1}(x-\theta)+\sum_{j=1}^{k}(a_j-\theta)^T\Sigma^{-1}(a_j-\theta)-\frac{n-m}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-\theta)^T\Sigma^{-1}(x_j-\theta)]\}\\
=&\int_{\theta\in\mathbb{R}^d} \exp\{-\frac{1}{2}[\sum_{i=1}^{n}(x_i-\theta)^T\Sigma^{-1}(x-\theta)+\sum_{j=1}^{k}(a_j-\theta)^T\Sigma^{-1}(a_j-\theta)-(n-m)(\bar{x}-\theta)^T\Sigma^{-1}(\bar{x}-\theta)]\}\\
=&\exp\{-\frac{1}{2}[(\sum_{r=1}^{d}\sum_{s=1}^{d}\Sigma_{rs}^{-1}\sum_{i=1}^{n+k}z_{ir}z_{is})-(n-m)\bar{x}^T\Sigma^{-1}\bar{x}]\}\\
&\int_{\theta\in\mathbb{R}^d} \exp\{-\frac{1}{2}[(m+k)\theta^T\Sigma^{-1}\theta-2(n+k)\theta^T\Sigma^{-1}\bar{z}+2(n-m)\theta^T\Sigma^{-1}\bar{x}]\}\\
=&\exp\{-\frac{1}{2}[(\sum_{r=1}^{d}\sum_{s=1}^{d}\Sigma_{rs}^{-1}\sum_{i=1}^{n+k}z_{ir}z_{is})-(n-m)\bar{x}^T\Sigma^{-1}\bar{x}]\}\\
&\int_{\theta\in\mathbb{R}^d} \exp\{-\frac{1}{2}[(m+k)\theta^T\Sigma^{-1}\theta-2(n+k)\theta^T\Sigma^{-1}(\frac{n}{n+k}\bar{x}+\frac{k}{n+k}\bar{a})+2(n-m)\theta^T\Sigma^{-1}\bar{x}]\}\\
=&\exp\{-\frac{1}{2}[(\sum_{r=1}^{d}\sum_{s=1}^{d}\Sigma_{rs}^{-1}\sum_{i=1}^{n+k}z_{ir}z_{is})-(n-m)\bar{x}^T\Sigma^{-1}\bar{x}]\}\\
&\int_{\theta\in\mathbb{R}^d} \exp\{-\frac{1}{2}[(m+k)\theta^T\Sigma^{-1}\theta-2n\theta^T\Sigma^{-1}\bar{x}-2k\theta^T\Sigma^{-1}\bar{a}+2(n-m)\theta^T\Sigma^{-1}\bar{x}]\}
\end{aligned}
$$

$$
\begin{aligned}
&= \exp\{-\frac{1}{2}[(\sum_{r=1}^{d}\sum_{s=1}^{d}\Sigma_{rs}^{-1}\sum_{i=1}^{n+k} z_{ir}z_{is}) - (n-m)\bar{x}^T\Sigma^{-1}\bar{x}]\} \\
&\quad \int_{\theta\in\mathbb{R}^d} \exp\{-\frac{1}{2}[(m+k)\theta^T\Sigma^{-1}\theta - 2k\theta^T\Sigma^{-1}\bar{a} - 2m\theta^T\Sigma^{-1}\bar{x}]\} \\
&= \exp\{-\frac{1}{2}[(\sum_{r=1}^{d}\sum_{s=1}^{d}\Sigma_{rs}^{-1}\sum_{i=1}^{n+k} z_{ir}z_{is}) - (n-m)\bar{x}^T\Sigma^{-1}\bar{x}]\} \\
&\quad \int_{\theta\in\mathbb{R}^d} \exp\{-\frac{1}{2}[(m+k)\theta^T\Sigma^{-1}\theta - 2\theta^T\Sigma^{-1}(k\bar{a} + m\bar{x})]\} \\
&\equiv \exp\{-\frac{1}{2}[(\sum_{r=1}^{d}\sum_{s=1}^{d}\Sigma_{rs}^{-1}\sum_{i=1}^{n+k} z_{ir}z_{is}) - (n-m)\bar{x}^T\Sigma^{-1}\bar{x}]\} \int_{\theta\in\mathbb{R}^d} \exp\{-\frac{1}{2}[(m+k)\theta^T\Sigma^{-1}\theta - 2\theta^T\Sigma^{-1}(m+k)\bar{v}]\} \\
&= \exp\{-\frac{1}{2}[(\sum_{r=1}^{d}\sum_{s=1}^{d}\Sigma_{rs}^{-1}\sum_{i=1}^{n+k} z_{ir}z_{is}) - (n-m)\bar{x}^T\Sigma^{-1}\bar{x} - (m+k)\bar{v}^T\Sigma^{-1}\bar{v}]\} \\
&\quad \int_{\theta\in\mathbb{R}^d} \exp\{-\frac{1}{2}(m+k)[\theta^T\Sigma^{-1}\theta - 2\theta^T\Sigma^{-1}\bar{v} + \bar{v}^T\Sigma^{-1}\bar{v}]\} \\
&= \exp\{-\frac{1}{2}[(\sum_{r=1}^{d}\sum_{s=1}^{d}\Sigma_{rs}^{-1}\sum_{i=1}^{n+k} z_{ir}z_{is}) - (n-m)\bar{x}^T\Sigma^{-1}\bar{x} - (m+k)\bar{v}^T\Sigma^{-1}\bar{v}]\} \\
&\quad \int_{\theta\in\mathbb{R}^d} \exp\{-\frac{1}{2}(m+k)(\theta-\bar{v})^T\Sigma^{-1}(\theta-\bar{v})\} \\
&= \exp\{-\frac{1}{2}[(\sum_{r=1}^{d}\sum_{s=1}^{d}\Sigma_{rs}^{-1}\sum_{i=1}^{n+k} z_{ir}z_{is}) - (n-m)\bar{x}^T\Sigma^{-1}\bar{x} - (m+k)\bar{v}^T\Sigma^{-1}\bar{v}]\}\sqrt{(2\pi)^d \det(\frac{\Sigma}{m+k})}
\end{aligned}
$$

Therefore the decision rule is

$$
\begin{aligned}
&\sqrt{(\frac{n}{m})^d}\exp\{-\frac{1}{2}[(\sum_{r=1}^{d}\sum_{s=1}^{d}\Sigma_{rs}^{-1}\sum_{i=1}^{n+k} z_{ir}z_{is}) - (n+k)\bar{z}^T\Sigma^{-1}\bar{z}]\}\sqrt{(\frac{2\pi}{n+k})^d \det(\Sigma)} > \\
&\quad \exp\{-\frac{1}{2}[(\sum_{r=1}^{d}\sum_{s=1}^{d}\Sigma_{rs}^{-1}\sum_{i=1}^{n+k} z_{ir}z_{is}) - (n-m)\bar{x}^T\Sigma^{-1}\bar{x} - (m+k)\bar{v}^T\Sigma^{-1}\bar{v}]\}\sqrt{(\frac{2\pi}{m+k})^d \det(\Sigma)} \\
&\Leftrightarrow \sqrt{(\frac{n(m+k)}{m(n+k)})^d}\exp\{-\frac{1}{2}[-(n+k)\bar{z}^T\Sigma^{-1}\bar{z}]\} > \exp\{-\frac{1}{2}[-(n-m)\bar{x}^T\Sigma^{-1}\bar{x} - (m+k)\bar{v}^T\Sigma^{-1}\bar{v}]\} \\
&\Leftrightarrow d\log\frac{n(m+k)}{m(n+k)} > \\
&\quad (n-m)\bar{x}^T\Sigma^{-1}\bar{x} + (m+k)\bar{v}^T\Sigma^{-1}\bar{v} - (n+k)\bar{z}^T\Sigma^{-1}\bar{z} \\
&\Leftrightarrow d\log\frac{n(m+k)}{m(n+k)} > \\
&\quad (n-m)\bar{x}^T\Sigma^{-1}\bar{x} + \frac{1}{m+k}(m\bar{x}+k\bar{a})^T\Sigma^{-1}(m\bar{x}+k\bar{a}) - \frac{1}{(n+k)}(n\bar{x}+k\bar{a})^T\Sigma^{-1}(n\bar{x}+k\bar{a}) \\
&\Leftrightarrow d\log\frac{n(m+k)}{m(n+k)} > \\
&\quad (n-m)\bar{x}^T\Sigma^{-1}\bar{x} + \frac{1}{m+k}[m^2\bar{x}^T\Sigma^{-1}\bar{x} + 2mk\bar{x}^T\Sigma^{-1}\bar{a} + k^2\bar{a}^T\Sigma^{-1}\bar{a}] \\
&\quad - \frac{1}{(n+k)}[n^2\bar{x}^T\Sigma^{-1}\bar{x} + 2nk\bar{x}^T\Sigma^{-1}\bar{a} + k^2\bar{a}^T\Sigma^{-1}\bar{a}] \\
&\Leftrightarrow d\log\frac{n(m+k)}{m(n+k)} > \\
&\quad \bar{x}^T\Sigma^{-1}\bar{x}(n-m+\frac{m^2}{m+k}-\frac{n^2}{n+k}) + \bar{x}^T\Sigma^{-1}\bar{a}(\frac{2mk}{m+k}-\frac{2nk}{(n+k)}) + \bar{a}^T\Sigma^{-1}\bar{a}k^2(\frac{1}{m+k}-\frac{1}{(n+k)})
\end{aligned}
$$

$$
\begin{aligned}
&\Leftrightarrow d(m+k)(n+k)\log\frac{n(m+k)}{m(n+k)} > \\
&\quad \bar{x}^T\Sigma^{-1}\bar{x}[(n-m)(m+k)(n+k)+m^2(n+k)-n^2(m+k)] \\
&\quad +[2mk(n+k)-2nk(m+k)]\bar{x}^T\Sigma^{-1}\bar{a}+(n-m)k^2\bar{a}^T\Sigma^{-1}\bar{a} \\
&\Leftrightarrow \frac{d(m+k)(n+k)}{k^2(n-m)}\log\frac{n(m+k)}{m(n+k)} > \bar{x}^T\Sigma^{-1}\bar{x}-2\bar{x}^T\Sigma^{-1}\bar{a}+\bar{a}^T\Sigma^{-1}\bar{a} \\
&\Leftrightarrow \frac{d(m+k)(n+k)}{k^2(n-m)}\log\frac{n(m+k)}{m(n+k)} > (\bar{x}-\bar{a})^T\Sigma^{-1}(\bar{x}-\bar{a})
\end{aligned}
$$

As required. □

Lemma F.9. *Consider the authenticator $\mathcal{D}_\alpha$ as defined in Def. F.2. Then any attacker $\mathcal{G}$, represented by a conditional probability $g_{X|Y}$, that satisfies the condition: $\bar{x}=\bar{y}$ for any leaked sample $y\in\mathbb{R}^{m\times d}$, and attacker generated sample $x\in\{\mathbb{R}^{n\times d}: g_{X|Y}(x|y)>0\}$ satisfies:*

$$\mathcal{G}\in\underset{\mathcal{G}'\in\mathbb{G}}{\operatorname{argmin}}\, V(\mathcal{D}_\alpha,\mathcal{G}') \quad \forall\alpha\in\mathbb{R}_+$$

Proof. The best response attacker satisfies:

$$
\begin{aligned}
g'_{X|Y} &\in \underset{g_{X|Y}}{\operatorname{argmin}}\,\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_\Theta^{(k)}}\mathbb{E}_{Y\sim f_\Theta^{(m)}}[\mathbb{E}_{X\sim f_\Theta^{(n)}}[\mathcal{D}_\alpha(A,X)]+\mathbb{E}_{X\sim g_{X|Y}(\cdot|Y)}[1-\mathcal{D}_\alpha(A,X)]] \\
&= \underset{g_{X|Y}}{\operatorname{argmin}}\,\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_\Theta^{(k)}}\mathbb{E}_{Y\sim f_\Theta^{(m)}}\mathbb{E}_{X\sim g_{X|Y}(\cdot|Y)}[1-\mathcal{D}_\alpha(A,X)] \\
&= \underset{g_{X|Y}}{\operatorname{argmax}}\,\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_\Theta^{(k)}}\mathbb{E}_{Y\sim f_\Theta^{(m)}}\mathbb{E}_{X\sim g_{X|Y}(\cdot|Y)}[\mathcal{D}_\alpha(A,X)] \\
&= \underset{g_{X|Y}}{\operatorname{argmax}}\,\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\overset{\text{iid}}{\sim}\mathcal{N}(\Theta,\Sigma)}\mathbb{E}_{Y\overset{\text{iid}}{\sim}\mathcal{N}(\Theta,\Sigma)}\mathbb{E}_{X\sim g_{X|Y}(\cdot|Y)}\left[I\left[\|\bar{X}-\bar{A}\|^2_{\Sigma^{-1}}<\alpha\right]\right] \\
&\overset{\text{Lemma F.3}}{=} \underset{g_{X|Y}}{\operatorname{argmax}}\,\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{\bar{A}\sim\mathcal{N}(\Theta,\frac{1}{k}\Sigma)}\mathbb{E}_{Y\overset{\text{iid}}{\sim}\mathcal{N}(\Theta,\Sigma)}\mathbb{E}_{X\sim g_{X|Y}(\cdot|Y)}\left[I\left[\|\bar{X}-\bar{A}\|^2_{\Sigma^{-1}}<\alpha\right]\right] \\
&= \underset{g_{X|Y}}{\operatorname{argmax}}\int_{y\in\mathbb{R}^{m\times d}}dy\int_{x\in\mathbb{R}^{n\times d}}dx g_{X|Y}(x|y)\int_{\bar{a}\in\mathbb{R}^d}d\bar{a}I\left[\|\bar{x}-\bar{a}\|^2_{\Sigma^{-1}}<\alpha\right]\int_{\theta\in\mathbb{R}^d}d\theta Q(\theta) \\
&\quad \exp\{-\frac{k}{2}(\bar{a}-\theta)^T\Sigma^{-1}(\bar{a}-\theta)\}\exp\{-\frac{1}{2}\sum_{j=1}^m(y_j-\theta)^T\Sigma^{-1}(y_j-\theta)\}
\end{aligned}
$$

Note that since $g_{X|Y}(x|y)$ can be chosen independently for each $y\in\mathbb{R}^{m\times d}$, we can optimize it independently for each $y\in\mathbb{R}^{m\times d}$, and we have:

$$
\begin{aligned}
g'_{X|Y}(\cdot|y)\in\underset{g_{X|Y}(\cdot|y)}{\operatorname{argmax}}&\int_{x\in\mathbb{R}^{n\times d}}dx g_{X|Y}(x|y)\int_{\bar{a}\in\mathbb{R}^d}d\bar{a}I\left[\|\bar{x}-\bar{a}\|^2_{\Sigma^{-1}}<\alpha\right]\int_{\theta\in\mathbb{R}^d}d\theta Q(\theta) \\
&\exp\{-\frac{k}{2}(\bar{a}-\theta)^T\Sigma^{-1}(\bar{a}-\theta)\}\exp\{-\frac{1}{2}\sum_{j=1}^m(y_j-\theta)^T\Sigma^{-1}(y_j-\theta)\}
\end{aligned}
$$

Note that for any PDF f over $\mathbb{R}^{n\times d}$ and a function $\varphi:\mathbb{R}^{n\times d}\to\mathbb{R}$, it holds that $\int_{x\in\mathbb{R}^{n\times d}}dx f(x)\varphi(x)\leqslant sup_x\varphi(x)$. Therefore there exists a deterministic distribution $g'_{X|Y}(x|y)=\delta(x-x')$ that achieves the maximum. And thus its sufficient to find a vector $x_\mathcal{G}$ that achieves the maximum:

$$
\begin{aligned}
x_\mathcal{G}\in\underset{x}{\operatorname{argmax}}&\int_{x'\in\mathbb{R}^{n\times d}}dx\delta(x'-x)\int_{\bar{a}\in\mathbb{R}^d}d\bar{a}I\left[\|\bar{x}'-\bar{a}\|^2_{\Sigma^{-1}}<\alpha\right]\int_{\theta\in\mathbb{R}^d}d\theta Q(\theta) \\
&\exp\{-\frac{k}{2}(\bar{a}-\theta)^T\Sigma^{-1}(\bar{a}-\theta)\}\exp\{-\frac{1}{2}\sum_{j=1}^m(y_j-\theta)^T\Sigma^{-1}(y_j-\theta)\}
\end{aligned}
$$

$$
\begin{aligned}
&= \underset{x}{\operatorname{argmax}} \int_{\bar{a}\in\mathbb{R}^d} d\bar{a} I\left[\|\bar{x}-\bar{a}\|^2_{\Sigma^{-1}} < \alpha\right] \int_{\theta\in\mathbb{R}^d} d\theta Q(\theta) \\
&\quad \exp\{-\frac{k}{2}(\bar{a}-\theta)^T\Sigma^{-1}(\bar{a}-\theta)\} \exp\{-\frac{1}{2}\sum_{j=1}^{m}(y_j-\theta)^T\Sigma^{-1}(y_j-\theta)\} \\
&\overset{(*)}{=} \underset{x}{\operatorname{argmax}} \int_{\bar{a}\in\mathbb{R}^d} d\bar{a} I\left[\|\bar{x}-\bar{a}\|^2_{\Sigma^{-1}} < \alpha\right] \int_{\theta\in\mathbb{R}^d} d\theta \\
&\quad \exp\{-\frac{1}{2}[k(\bar{a}-\theta)^T\Sigma^{-1}(\bar{a}-\theta) + \sum_{j=1}^{m}(y_j-\theta)^T\Sigma^{-1}(y_j-\theta)]\} \\
&= \underset{x}{\operatorname{argmax}} \int_{\bar{a}\in\mathbb{R}^d} d\bar{a} I\left[\|\bar{x}-\bar{a}\|^2_{\Sigma^{-1}} < \alpha\right] \int_{\theta\in\mathbb{R}^d} d\theta \\
&\quad \exp\{-\frac{1}{2}[k\bar{a}^T\Sigma^{-1}\bar{a} - 2k\theta^T\Sigma^{-1}\bar{a} + k\theta^T\Sigma^{-1}\theta + \sum_{j=1}^{m} y_j^T\Sigma^{-1}y_j - 2m\theta^T\Sigma^{-1}\bar{y} + m\theta^T\Sigma^{-1}\theta]\} \\
&= \underset{x}{\operatorname{argmax}} \int_{\bar{a}\in\mathbb{R}^d} d\bar{a} \exp\{-\frac{1}{2}[k\bar{a}^T\Sigma^{-1}\bar{a}]\} I\left[\|\bar{x}-\bar{a}\|^2_{\Sigma^{-1}} < \alpha\right] \int_{\theta\in\mathbb{R}^d} d\theta \\
&\quad \exp\{-\frac{1}{2}[(m+k)\theta^T\Sigma^{-1}\theta - 2\theta^T\Sigma^{-1}(m\bar{y}+k\bar{a})]\} \\
&= \underset{x}{\operatorname{argmax}} \int_{\bar{a}\in\mathbb{R}^d} d\bar{a} \exp\{-\frac{1}{2}[k\bar{a}^T\Sigma^{-1}\bar{a} - \frac{1}{m+k}(m\bar{y}+k\bar{a})^T\Sigma^{-1}(m\bar{y}+k\bar{a})]\} I\left[\|\bar{x}-\bar{a}\|^2_{\Sigma^{-1}} < \alpha\right] \\
&\quad \int_{\theta\in\mathbb{R}^d} d\theta \exp\{-\frac{(m+k)}{2}(\theta - \frac{m\bar{y}+k\bar{a}}{m+k})^T\Sigma^{-1}(\theta - \frac{m\bar{y}+k\bar{a}}{m+k})\} \\
&= \underset{x}{\operatorname{argmax}} \int_{\bar{a}\in\mathbb{R}^d} d\bar{a} \exp\{-\frac{1}{2}[k\bar{a}^T\Sigma^{-1}\bar{a} - \frac{1}{m+k}(m\bar{y}+k\bar{a})^T\Sigma^{-1}(m\bar{y}+k\bar{a})]\} I\left[\|\bar{x}-\bar{a}\|^2_{\Sigma^{-1}} < \alpha\right] \\
&= \underset{x}{\operatorname{argmax}} \int_{\bar{a}\in\mathbb{R}^d} d\bar{a} \exp\{-\frac{1}{2}[\frac{mk}{m+k}\bar{a}^T\Sigma^{-1}\bar{a} - \frac{2mk}{m+k}\bar{y}^T\Sigma^{-1}\bar{a} + \frac{mk}{m+k}\bar{y}^T\Sigma^{-1}\bar{y}]\} I\left[\|\bar{x}-\bar{a}\|^2_{\Sigma^{-1}} < \alpha\right] \\
&= \underset{x}{\operatorname{argmax}} \int_{\bar{a}\in\mathbb{R}^d} d\bar{a} \exp\{-\frac{mk}{2(m+k)}[(\bar{a}-\bar{y})^T\Sigma^{-1}(\bar{a}-\bar{y})]\} I\left[\|\bar{x}-\bar{a}\|^2_{\Sigma^{-1}} < \alpha\right]
\end{aligned}
$$

Where in $(*)$ we used the fact that $Q(\theta)$ is the improper uniform prior. Note that the expression depends only on the mean $\bar{x}$, therefore its sufficient to find an $\bar{x}$ that maximizes the expression. We substitute the integration variable to $\varphi = \bar{a} - \bar{x}$ and obtain:

$$
\begin{aligned}
\bar{x}_{\mathcal{G}} &\in \underset{\bar{x}}{\operatorname{argmax}} \int_{\{\varphi\in\mathbb{R}^d:\varphi^T\Sigma^{-1}\varphi<\alpha\}} d\varphi \exp\{-\frac{mk}{2(m+k)}[(\varphi+\bar{x}-\bar{y})^T\Sigma^{-1}(\varphi+\bar{x}-\bar{y})]\} \\
&\equiv \underset{\bar{x}}{\operatorname{argmax}}\, \psi(\bar{y}-\bar{x}) \equiv \underset{\bar{x}}{\operatorname{argmax}}\, \psi(\mu)
\end{aligned}
$$

Where ψ is defined as in Lemma F.7 (with $\sigma = \frac{m+k}{mk}$), from which it follows that $\psi(\mu)$ is log-concave, and therefore has at most one local extremum which can be only a maximum. Therefore it is sufficient to show that $\mu = 0$ (i.e $\bar{x} = \bar{y}$) is a local extremum by equating the gradient at the point to 0.

$$
\begin{aligned}
\frac{\partial}{\partial\mu}\psi(\mu) =& \frac{\partial}{\partial\mu}\int_{\{\varphi\in\mathbb{R}^d:\varphi^T\Sigma^{-1}\varphi<\alpha\}} d\varphi \exp\{-\frac{mk}{2(m+k)}[(\varphi-\mu)^T\Sigma^{-1}(\varphi-\mu)]\} \\
=& -\frac{mk}{2(m+k)}\int_{\{\varphi\in\mathbb{R}^d:\varphi^T\Sigma^{-1}\varphi<\alpha\}} d\varphi \exp\{-\frac{mk}{2(m+k)}[(\varphi-\mu)^T\Sigma^{-1}(\varphi-\mu)]\}\frac{\partial}{\partial\mu}(\varphi-\mu)^T\Sigma^{-1}(\varphi-\mu) \\
=& -\frac{mk}{(m+k)}\int_{\{\varphi\in\mathbb{R}^d:\varphi^T\Sigma^{-1}\varphi<\alpha\}} d\varphi \exp\{-\frac{mk}{2(m+k)}[(\varphi-\mu)^T\Sigma^{-1}(\varphi-\mu)]\}\Sigma^{-1}(\mu-\varphi)
\end{aligned}
$$

Therefore:

$$
\frac{\partial}{\partial\mu}\psi(\mu)|_{\mu=0} = \frac{mk}{(m+k)}\int_{\{\varphi\in\mathbb{R}^d:\varphi^T\Sigma^{-1}\varphi<\alpha\}} d\varphi \exp\{-\frac{mk}{2(m+k)}[\varphi^T\Sigma^{-1}\varphi]\}\Sigma^{-1}\varphi
$$

And since the domain of integration is symmetric about the origin and the integrand is odd with respect to the integration variable, the integral is equal to zero. i.e $\frac{\partial}{\partial\mu}\psi(\mu)|_{\mu=0} = 0$. Therefore $\bar{x} = \bar{y}$ ($\mu = 0$) achieves the global maximum, and any attacker that satisfies the condition: $\bar{x} = \bar{y}$ for any leaked sample $y \in \mathbb{R}^{m\times d}$ and attacker generated sample $x \in \{\mathbb{R}^{n\times d} : g_{X|Y}(x|y) > 0\}$ satisfies:

$$\mathcal{G} \in \operatorname*{argmin}_{\mathcal{G}'\in\mathbb{G}} V(\mathcal{D}_\alpha, \mathcal{G}') \quad \forall\alpha \in \mathbb{R}_+$$

As required. □

Corollary F.10. *Consider an authenticator $\mathcal{D}_\alpha$ as defined in Def. F.2. Then the attacker $\mathcal{G}^*$ defined in Def. F.1 is a best response. i.e.:*

$$\mathcal{G}^* \in \operatorname*{argmin}_{\mathcal{G}'\in\mathbb{G}} V(\mathcal{D}_\alpha, \mathcal{G}') \quad \forall\alpha \in \mathbb{R}_+$$

Proof. Directly from Lemma F.9 □

Theorem F.11. *The game value is:*

$$\max_{\mathcal{D}}\min_{\mathcal{G}} V(\mathcal{D},\mathcal{G}) = \min_{\mathcal{G}}\max_{\mathcal{D}} V(\mathcal{D},\mathcal{G}) = V(\mathcal{D}^*,\mathcal{G}^*) =$$
$$\frac{1}{2} + \frac{1}{2\Gamma(\frac{d}{2})}[\gamma(\frac{d}{2}, \frac{dn(m+k)}{2k(n-m)}\log\frac{n(m+k)}{m(n+k)}) - \gamma(\frac{d}{2}, \frac{dm(n+k)}{2k(n-m)}\log\frac{n(m+k)}{m(n+k)})]$$

Where γ is the lower incomplete gamma function.

Proof. From the max-min inequality we have:

$$\max_{\mathcal{D}}\min_{\mathcal{G}} V(\mathcal{D},\mathcal{G}) \leqslant \min_{\mathcal{G}}\max_{\mathcal{D}} V(\mathcal{D},\mathcal{G})$$

On the other hand using Lemma F.8 and Corollary F.10 we have:

$$\max_{\mathcal{D}}\min_{\mathcal{G}} V(\mathcal{D},\mathcal{G}) \geqslant \min_{\mathcal{G}} V(\mathcal{D}^*,\mathcal{G}) \overset{F.10}{=} V(\mathcal{D}^*,\mathcal{G}^*) \overset{F.8}{=} \max_{\mathcal{D}} V(\mathcal{D},\mathcal{G}^*) \geqslant \min_{\mathcal{G}}\max_{\mathcal{D}} V(\mathcal{D},\mathcal{G})$$

And therefore:

$$\max_{\mathcal{D}}\min_{\mathcal{G}} V(\mathcal{D},\mathcal{G}) = \min_{\mathcal{G}}\max_{\mathcal{D}} V(\mathcal{D},\mathcal{G}) = V(\mathcal{D}^*,\mathcal{G}^*)$$

The game value is given by:

$$\begin{aligned}
V(\mathcal{D}^*,\mathcal{G}^*) =& \mathbb{E}_{\Theta\sim Q} V(\Theta,\mathcal{D}^*,\mathcal{G}^*) \\
=& \frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_\Theta^{(k)}}\mathbb{E}_{Y\sim f_\Theta^{(m)}}\left[\mathbb{E}_{X\sim f_\Theta^{(n)}}[\mathcal{D}^*(A,X)] + \mathbb{E}_{X\sim g^*_{X|Y}(\cdot|Y)}[1-\mathcal{D}^*(A,X)]\right] \\
=& \frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_\Theta^{(k)}}\mathbb{E}_{Y\sim f_\Theta^{(m)}} \\
& \left[\mathbb{E}_{X\sim f_\Theta^{(n)}}\left[I[\|\bar{X}-\bar{A}\|^2_{\Sigma^{-1}} < \alpha^*]\right] + \mathbb{E}_{X\sim g^*_{X|Y}(\cdot|Y)}\left[1 - I[\|\bar{X}-\bar{A}\|^2_{\Sigma^{-1}} < \alpha^*]\right]\right] \\
=& \frac{1}{2} + \frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_\Theta^{(k)}}\mathbb{E}_{X\sim f_\Theta^{(n)}}\left[I[\|\bar{x}-\bar{A}\|^2_{\Sigma^{-1}} < \alpha^*]\right] - \\
& \frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_\Theta^{(k)}}\mathbb{E}_{X\sim g^*_{X|\Theta}(\cdot|\Theta)}\left[I[\|\bar{x}-\bar{A}\|^2_{\Sigma^{-1}} < \alpha^*]\right]
\end{aligned}$$

Observing the first term we have:

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_{\Theta}^{(k)}}\mathbb{E}_{X\sim f_{\Theta}^{(n)}}\left[I[\|\bar{X}-\bar{A}\|_{\Sigma^{-1}}^{2}<\alpha^{*}]\right]\\
=&\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_{\Theta}^{(k)}}\mathbb{E}_{X\sim f_{\Theta}^{(n)}}\left[I[(\bar{X}-\bar{A})^{T}\Sigma^{-1}(\bar{X}-\bar{A})<\alpha^{*}]\right]\\
=&\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_{\Theta}^{(k)}}\mathbb{E}_{X\sim f_{\Theta}^{(n)}}\left[I[(\bar{X}-\bar{A})^{T}(CC^{T})^{-1}(\bar{X}-\bar{A})<\alpha^{*}]\right]\\
=&\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_{\Theta}^{(k)}}\mathbb{E}_{X\sim f_{\Theta}^{(n)}}\left[I[(\bar{X}-\bar{A})^{T}C^{-T}C^{-1}(\bar{X}-\bar{A})<\alpha^{*}]\right]\\
=&\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_{\Theta}^{(k)}}\mathbb{E}_{X\sim f_{\Theta}^{(n)}}\left[I[(C^{-1}(\bar{X}-\bar{A}))^{T}(C^{-1}(\bar{X}-\bar{A}))<\alpha^{*}]\right]\\
\equiv&\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_{\Theta}^{(k)}}\mathbb{E}_{X\sim f_{\Theta}^{(n)}}\left[I[Z^{T}Z<\alpha^{*}]\right]\\
=&(*)
\end{aligned}$$

Observe that

$$Z=C^{-1}(\bar{X}-\bar{A})=C^{-1}[I_d,-I_d]\begin{bmatrix}\bar{X}\\ \bar{A}\end{bmatrix}=[C^{-1},-C^{-1}]\begin{bmatrix}\bar{X}\\ \bar{A}\end{bmatrix}$$

And since:

$$\begin{bmatrix}\bar{X}\\ \bar{A}\end{bmatrix}\sim\mathcal{N}(0_{2d},\begin{bmatrix}\frac{1}{n}\Sigma & 0_{d\times d}\\ 0_{d\times d} & \frac{1}{k}\Sigma\end{bmatrix})$$

Then:

$$\begin{aligned}
Z&\sim\mathcal{N}(0_d,[C^{-1},-C^{-1}]\begin{bmatrix}\frac{1}{n}\Sigma & 0_{d\times d}\\ 0_{d\times d} & \frac{1}{k}\Sigma\end{bmatrix}\begin{bmatrix}C^{-T}\\ -C^{-T}\end{bmatrix})\\
&=\mathcal{N}(0_d,(\frac{1}{n}+\frac{1}{k})C^{-1}\Sigma C^{-T})\\
&=\mathcal{N}(0_d,\frac{n+k}{nk}C^{-1}CC^{T}C^{-T})\\
&=\mathcal{N}(0_d,\frac{n+k}{nk}I_d)
\end{aligned}$$

We denote $\tilde{Z}=\sqrt{\frac{nk}{n+k}}Z\sim\mathcal{N}(0_d,I_d)$, and thus $\tilde{Z}_1,\dots,\tilde{Z}_d$ are independent standard normal random variables and $\tilde{Z}^T\tilde{Z}\sim\chi^2(d)$. Therefore $Z^TZ=\frac{n+k}{nk}\tilde{Z}^T\tilde{Z}\sim\Gamma(k=\frac{d}{2},\theta=2\frac{n+k}{nk})$ and:

$$\begin{aligned}
(*)=&\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_{\Theta}^{(k)}}\mathbb{E}_{X\sim f_{\Theta}^{(n)}}\left[I[Z^{T}Z<\alpha^{*}]\right]\\
=&\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{Z^TZ\sim\Gamma(k=\frac{d}{2},\theta=2\frac{n+k}{nk})}\left[I[Z^{T}Z<\alpha^{*}]\right]\\
\overset{(i)}{=}&\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\frac{1}{\Gamma(\frac{d}{2})}\gamma(\frac{d}{2},\frac{nk\alpha^{*}}{2(n+k)})\\
=&\frac{1}{2}\frac{1}{\Gamma(\frac{d}{2})}\gamma(\frac{d}{2},\frac{nk\alpha^{*}}{2(n+k)})
\end{aligned}$$

Where in (i) We used the CDF of the Gamma distribution in which γ is the lower incomplete gamma function.

Similarly, observing the second term we have:

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_{\Theta}^{(k)}}\mathbb{E}_{X\sim g^{*}_{X|\Theta}(\cdot|\Theta)}\left[I[\|\bar{X}-\bar{A}\|_{\Sigma^{-1}}^{2}<\alpha^{*}]\right]\\
\equiv&\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_{\Theta}^{(k)}}\mathbb{E}_{X\sim g^{*}_{X|\Theta}(\cdot|\Theta)}\left[I[V^{T}V<\alpha^{*}]\right]\\
=&(**)
\end{aligned}$$

Where:

$$V = C^{-1}(\bar{X} - \bar{A}) = C^{-1}[I_d, -I_d]\begin{bmatrix}\bar{X}\\ \bar{A}\end{bmatrix} = [C^{-1}, -C^{-1}]\begin{bmatrix}\bar{X}\\ \bar{A}\end{bmatrix}$$

Using the definition of $\mathcal{G}^*$ (Definition F.1) and Lemma F.3 we have:

$$\begin{bmatrix}\bar{X}\\ \bar{A}\end{bmatrix} \sim \mathcal{N}(0_{2d}, \begin{bmatrix}\frac{1}{m}\Sigma & 0_{d\times d}\\ 0_{d\times d} & \frac{1}{k}\Sigma\end{bmatrix})$$

Therefore:

$$V \sim \mathcal{N}(0_d, [C^{-1}, -C^{-1}]\begin{bmatrix}\frac{1}{m}\Sigma & 0_{d\times d}\\ 0_{d\times d} & \frac{1}{k}\Sigma\end{bmatrix}\begin{bmatrix}C^{-T}\\ -C^{-T}\end{bmatrix}) = \mathcal{N}(0_d, \frac{m+k}{mk}I_d)$$

And similarly to the first term we get:

$$V^TV \sim \Gamma(k = \frac{d}{2}, \theta = 2\frac{m+k}{mk})$$

And thus:

$$(**) = \frac{1}{2}\frac{1}{\Gamma(\frac{d}{2})}\gamma(\frac{d}{2}, \frac{mk\alpha^*}{2(m+k)})$$

Therefore the game value is given by:

$$\begin{aligned} V(\mathcal{D}^*, \mathcal{G}^*) &= \frac{1}{2} + \frac{1}{2}\frac{1}{\Gamma(\frac{d}{2})}[\gamma(\frac{d}{2}, \frac{nk\alpha^*}{2(n+k)}) - \gamma(\frac{d}{2}, \frac{mk\alpha^*}{2(m+k)})] \\ &= \frac{1}{2} + \frac{1}{2}\frac{1}{\Gamma(\frac{d}{2})}[\gamma(\frac{d}{2}, \frac{dn(m+k)}{2k(n-m)}\log\frac{n(m+k)}{m(n+k)}) - \gamma(\frac{d}{2}, \frac{dm(n+k)}{2k(n-m)}\log\frac{n(m+k)}{m(n+k)})] \end{aligned}$$

As required. □

Finally we prove Theorem 4.2 and Corollary 4.3.

Theorem 4.2. *Define $\delta = m/n \leqslant 1$ and let $\rho = m/k$. Consider the attacker $\mathcal{G}^*$ defined by the following generative process: Given a leaked sample $Y \in \mathbb{R}^{m\times d}$, $\mathcal{G}^*$ generates a sample $X \in \mathbb{R}^{n\times d}$ as follows: it first samples n vectors $W_1, \dots, W_n \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$ and then sets: $X_i = W_i - \bar{W} + \bar{Y}$. Define the authenticator $\mathcal{D}^*$ by:*

$$\mathcal{D}^*(a, x) = I\left[\|\bar{x} - \bar{a}\|^2_{\Sigma^{-1}} < \frac{d\,(1+\rho)\left(1+\rho\delta^{-1}\right)}{n(1-\delta)}\log\left(\frac{\rho+1}{\rho+\delta}\right)\right] \tag{F.5}$$

Then $(\mathcal{D}^*, \mathcal{G}^*)$ is a solution of Eq. 2.2 that satisfies Eq. 4.1.

Proof. Directly from Lemma F.8, Corollary F.10, and Theorem F.11 by assigning $\delta = \frac{m}{n}, \rho = \frac{m}{k}$. □

Corollary 4.3. *Define δ and ρ as in Theorem 4.2. Then the game value for the Gaussian case is:*

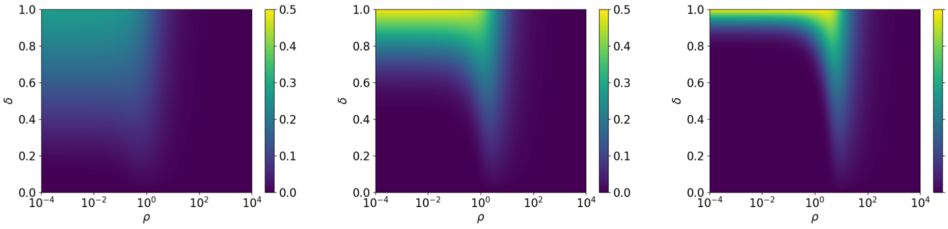
$$\frac{1}{2} + \frac{1}{2\Gamma(\frac{d}{2})}\left[\gamma\left(\frac{d}{2}, \frac{d(1+\rho)}{2(1-\delta)}\log\frac{1+\rho}{\delta+\rho}\right) - \gamma\left(\frac{d}{2}, \frac{d(\delta+\rho)}{2(1-\delta)}\log\frac{1+\rho}{\delta+\rho}\right)\right] \tag{F.6}$$

Where γ is the lower incomplete gamma function, and Γ is the Gamma function.

Proof. Directly from Theorem F.11 by assigning $\delta = \frac{m}{n}, \rho = \frac{m}{k}$. □

F.4 Game value for a maximum likelihood attacker

In this section we show that the intuitive “optimal attacker” is sub optimal as can be seen in Figure 1c in the main paper. We consider an attacker that draws the attack sample from a Gaussian distribution with the maximum likelihood estimate of the mean and the known covariance. We denote this attacker by the name ML attacker. We find the best response autenticator for this attacker and associated the game value. Figure 6 visualizes the difference in theoretical game value between the ML attacker (see Definition F.12) and the optimal attacker (see Definition F.1) for different values of d (the observations dimension), and demonstrates that the ML attacker is indeed sub-optimal.



(a) (b) (c)

Figure 6: The difference in game value (expected authentication accuracy of the optimal authenticator) between the ML attacker and the optimal attacker for different values of the observations dimension d, as a function of the parameters $\rho = \frac{m}{k}, delta = \frac{m}{n}$. Namely: $\max_{\mathcal{D}}\{V(\mathcal{D}, \mathcal{G}_{ML})\} - \max_{\mathcal{D}}\{V(\mathcal{D}, \mathcal{G}^{*})\}$. (a) Difference in game value for $d = 10$. (b) Difference in game value for $d = 100$. (c) Difference in game value for $d = 1000$.

Definition F.12. *Let $\mathcal{G}_{ML}$ denote an attacker defined by the following generative process: Given a leaked sample $Y \in \mathbb{R}^{m \times d}$, $\mathcal{G}_{ML}$ generates an attack sample $X \overset{iid}{\sim} \mathcal{N}(\bar{Y}, \Sigma)$*

Lemma F.13. *Let $\theta \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}$ represent the mean and covariance of a Gaussian distribution. Let $X \in \mathbb{R}^{n \times d}$ be a random sample generated by the attacker defined in Def. F.12. Then:*

$$X_c \sim \mathcal{N}(\theta_{c,n}, diag(\Sigma, n) + rep((\frac{1}{m})\Sigma, n)) \equiv \mathcal{N}(\theta_{c,n}, \Psi_{ML})$$

Proof. Let $W_1, \ldots, W_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma)$, observe that $X_i = \bar{Y} + W_i \quad \forall i \in [n]$, and thus:

$$X_c = \bar{Y}_{c,n} + W_c$$

Where $W_c \sim \mathcal{N}(0 \cdot 1_{dn}, diag(\Sigma, n)$. Using Lemma F.3 we have $\bar{Y} \sim \mathcal{N}(\theta, \frac{1}{m}\Sigma)$, and using Lemma F.4 we have $\bar{Y}_{c,n} \sim \mathcal{N}(\theta_{c,n}, rep(\frac{1}{m}\Sigma, n))$.

Let $Z = \begin{bmatrix} W_c \\ \bar{Y}_{c,n} \end{bmatrix}$ and $B = [I_{nd \times nd} \quad , \quad I_{nd \times nd}]$, then $X_c = W_c + \bar{Y}_{c,n} = BZ$. Note that:

$$Z \sim \mathcal{N}(\begin{bmatrix} 0_{nd} \\ \theta_{c,n} \end{bmatrix}, \begin{bmatrix} diag(\Sigma, n) & 0 \\ 0 & rep(\frac{1}{m}\Sigma, n) \end{bmatrix})$$

and therefore we have:

$$\begin{aligned} X_c \sim & \mathcal{N}(B \begin{bmatrix} 0_{nd} \\ \theta_{c,n} \end{bmatrix}, B \begin{bmatrix} diag(\Sigma, n) & 0 \\ 0 & rep(\frac{1}{m}\Sigma, n) \end{bmatrix} B^T) \\ = & \mathcal{N}(\theta_{c,n}, diag(\Sigma, n) + rep(\frac{1}{m}\Sigma, n)) \end{aligned}$$

as required. □

Lemma F.14. *Let $\Sigma = CC^T \in \mathbb{R}^{d \times d}$ represent the covariance of a Gaussian distribution, and consider the following covariance matrix: $\Psi_{ML} = diag(\Sigma, n) + rep(\frac{1}{m}\Sigma, n)$. Then:*

$$\Psi_{ML}^{-1} = diag(\Sigma^{-1}, n) - \frac{1}{n+m} rep(\Sigma^{-1}, n)$$

And the determinant is:

$$\det(\Psi) = \det(\Sigma)^n (\frac{n+m}{m})^d$$

Proof. To find the inverse of Ψ_{ML} we first define

$$U = \begin{bmatrix} \Sigma \\ \vdots \\ \Sigma \end{bmatrix} \in \mathbb{R}^{nd \times d}, \quad V = \frac{1}{m} \begin{bmatrix} I_d \\ \vdots \\ I_d \end{bmatrix} \in \mathbb{R}^{nd \times d} \tag{F.7}$$

then we have:

$$\begin{aligned}
\Psi_{ML}^{-1} =& (diag(\Sigma, n) + rep(\frac{1}{m}\Sigma, n))^{-1} \\
=& (diag(\Sigma, n) + UV^T)^{-1} \\
\overset{(i)}{=}& diag(\Sigma, n)^{-1} - diag(\Sigma, n)^{-1}U(I_d + V^T diag(\Sigma, n)^{-1}U)^{-1}V^T diag(\Sigma, n)^{-1} \\
\overset{(ii)}{=}& diag(\Sigma^{-1}, n) - diag(\Sigma^{-1}, n)U(I_d + V^T diag(\Sigma^{-1}, n)U)^{-1}V^T diag(\Sigma^{-1}, n) \\
=& diag(\Sigma^{-1}, n) - \begin{bmatrix} I_d \\ \vdots \\ I_d \end{bmatrix} (I_d + \frac{1}{m}\begin{bmatrix} I_d & \cdots & I_d \end{bmatrix} \begin{bmatrix} I_d \\ \vdots \\ I_d \end{bmatrix})^{-1} \frac{1}{m} \begin{bmatrix} \Sigma^{-1} & \cdots & \Sigma^{-1} \end{bmatrix} \\
=& diag(\Sigma^{-1}, n) - \begin{bmatrix} I_d \\ \vdots \\ I_d \end{bmatrix} (I_d + \frac{n}{m}I_d)^{-1} \frac{1}{m} \begin{bmatrix} \Sigma^{-1} & \cdots & \Sigma^{-1} \end{bmatrix} \\
=& diag(\Sigma^{-1}, n) - \begin{bmatrix} I_d \\ \vdots \\ I_d \end{bmatrix} ((\frac{n+m}{m})I_d)^{-1} \frac{1}{m} \begin{bmatrix} \Sigma^{-1} & \cdots & \Sigma^{-1} \end{bmatrix} \\
=& diag(\Sigma^{-1}, n) - \begin{bmatrix} I_d \\ \vdots \\ I_d \end{bmatrix} \frac{m}{n+m} I_d \frac{1}{m} \begin{bmatrix} \Sigma^{-1} & \cdots & \Sigma^{-1} \end{bmatrix} \\
=& diag(\Sigma^{-1}, n) - \begin{bmatrix} I_d \\ \vdots \\ I_d \end{bmatrix} \frac{1}{n+m} I_d \begin{bmatrix} \Sigma^{-1} & \cdots & \Sigma^{-1} \end{bmatrix} \\
=& diag(\Sigma^{-1}, n) - \begin{bmatrix} I_d \\ \vdots \\ I_d \end{bmatrix} \frac{1}{n+m} I_d \begin{bmatrix} \Sigma^{-1} & \cdots & \Sigma^{-1} \end{bmatrix} \\
=& diag(\Sigma^{-1}, n) - \frac{1}{n+m} \begin{bmatrix} I_d \\ \vdots \\ I_d \end{bmatrix} \begin{bmatrix} \Sigma^{-1} & \cdots & \Sigma^{-1} \end{bmatrix} \\
=& diag(\Sigma^{-1}, n) - \frac{1}{n+m} rep(\Sigma^{-1}, n)
\end{aligned}$$

As required. Where in (i) we used the Woodbury matrix identity, and in (ii) we used the inverse of a diagonal block matrix.

Next, we turn to finding the determinant of Ψ_{ML}:

$$\begin{aligned}
\det(\Psi) &= \det(diag(\Sigma, n) + rep((\frac{1}{m})\Sigma, n)) \\
&= \det(diag(\Sigma, n) + UV^T) \\
&\overset{(iii)}{=} \det(diag(\Sigma, n)) \det(I_d + V^T diag(\Sigma, n)^{-1}U) \\
&\overset{(iv)}{=} \det(\Sigma)^n \det(I_d + V^T diag(\Sigma^{-1}, n)U)
\end{aligned}$$

$$
\begin{aligned}
&= \det(\Sigma)^n \det(I_d + \frac{1}{m}\begin{bmatrix} I_d & \cdots & I_d \end{bmatrix} diag(\Sigma^{-1}, n) \begin{bmatrix} \Sigma \\ \vdots \\ \Sigma \end{bmatrix}) \\
&= \det(\Sigma)^n \det((\frac{n+m}{m}) I_d) \\
&= \det(\Sigma)^n (\frac{n+m}{m})^d
\end{aligned}
$$

Where in (iii) we used the matrix determinant lemma, and in (iv) we used the determinant of a diagonal block matrix. □

Lemma F.15. *Consider the attacker $\mathcal{G}_{ML}$, defined in F.12. The best response strategy for the authenticator against this attacker is:*

$$\mathcal{D}_{ML}(a, x) = I\left[\|\bar{x} - \bar{a}\|^2_{\Sigma^{-1}} < \alpha_{ML}\right]$$

Where:

$$\alpha_{ML} = \frac{d(n+k)(nm+nk+mk)}{k^2 n^2} \log(\frac{nm+nk+mk}{m(n+k)})$$

Proof. The best response authenticator satisfies:

$$
\begin{aligned}
\mathcal{D}^* &\in \operatorname*{argmax}_{\mathcal{D} \in \mathbb{D}} V(\mathcal{D}, \mathcal{G}_{ML}) \\
&= \operatorname*{argmax}_{\mathcal{D} \in \mathbb{D}} \frac{1}{2} \mathbb{E}_{\Theta \sim Q} \mathbb{E}_{A \sim f_\Theta^{(k)}} \mathbb{E}_{Y \sim f_\Theta^{(m)}} \left[\mathbb{E}_{X \sim f_\Theta^{(n)}} [\mathcal{D}(A, X)] + \mathbb{E}_{X \sim g^{ML}_{X|Y}(\cdot|Y)} [1 - \mathcal{D}(A, X)] \right] \\
&= \operatorname*{argmax}_{\mathcal{D} \in \mathbb{D}} \mathbb{E}_{\Theta \sim Q} \mathbb{E}_{A \sim f_\Theta^{(k)}} \left[\mathbb{E}_{X \sim f_\Theta^{(n)}} [\mathcal{D}(A, X)] + \mathbb{E}_{X \sim g^{ML}_{X|\Theta}(\cdot|\Theta)} [1 - \mathcal{D}(A, X)] \right] \\
&\overset{\text{Lemma F.13}}{=} \operatorname*{argmax}_{\mathcal{D} \in \mathbb{D}} \mathbb{E}_{\Theta \sim Q} \mathbb{E}_{A \sim \mathcal{N}(\Theta_{c,k}, diag(\Sigma, k))} \left[\mathbb{E}_{X \sim \mathcal{N}(\Theta_{c,n}, diag(\Sigma, n))} [\mathcal{D}(A, X)] + \mathbb{E}_{X \sim \mathcal{N}(\theta_{c,n}, \Psi_{ML})} [1 - \mathcal{D}(A, X)] \right] \\
&\overset{\text{Lemma F.14}}{=} \operatorname*{argmax}_{\mathcal{D} \in \mathbb{D}} \mathbb{E}_{\Theta \sim Q} \int_{a \in \mathbb{R}^{kd}} da \int_{x \in \mathbb{R}^{nd}} dx \exp\{-\frac{1}{2}(a - \Theta_{c,k})^T diag(\Sigma, k)^{-1} (a - \Theta_{c,k})\} \\
&\quad [\sqrt{\frac{1}{(2\pi)^{nd} |\det(diag(\Sigma, n))|}} \exp\{-\frac{1}{2}(x - \Theta_{c,n})^T diag(\Sigma, n)^{-1} (x - \Theta_{c,n})\} \mathcal{D}(a, x) + \\
&\quad \sqrt{\frac{1}{(2\pi)^{nd} |\det(\Psi_{ML})|}} \exp\{-\frac{1}{2}(x - \Theta_{c,n})^T \Psi_{ML}^{-1} (x - \Theta_{c,n})\} [1 - \mathcal{D}(a, x)]] \\
&\overset{\text{Lemma F.14}}{=} \operatorname*{argmax}_{\mathcal{D} \in \mathbb{D}} \mathbb{E}_{\Theta \sim Q} \int_{a \in \mathbb{R}^{kd}} da \int_{x \in \mathbb{R}^{nd}} dx \exp\{-\frac{1}{2}(a - \Theta_{c,k})^T diag(\Sigma, k)^{-1} (a - \Theta_{c,k})\} [\\
&\quad \sqrt{(\frac{n+m}{m})^d} \exp\{-\frac{1}{2}(x - \Theta_{c,n})^T diag(\Sigma, n)^{-1} (x - \Theta_{c,n})\} \mathcal{D}(a, x) + \\
&\quad \exp\{-\frac{1}{2}(x - \Theta_{c,n})^T \Psi_{ML}^{-1} (x - \Theta_{c,n})\} [1 - \mathcal{D}(a, x)]]
\end{aligned}
$$

$\mathcal{D}$ can be chosen independently for each pair $(a, x) \in \mathcal{X}^k \times \mathcal{X}^n$ therefore for any $(a, x) \in \mathcal{X}^k \times \mathcal{X}^n$ the decision rule for $\mathcal{D}(a, x) = 1$ is:

$$
\begin{aligned}
&\sqrt{(\frac{n+m}{m})^d} \int_{\theta \in \mathbb{R}^d} Q(\theta) \exp\{-\frac{1}{2}(x - \Theta_{c,n})^T diag(\Sigma, n)^{-1} (x - \Theta_{c,n})\} \exp\{-\frac{1}{2}(a - \Theta_{c,k})^T diag(\Sigma, k)^{-1} (a - \Theta_{c,k})\} \\
&> \int_{\theta \in \mathbb{R}^d} Q(\theta) \exp\{-\frac{1}{2}(x - \Theta_{c,n})^T \Psi_{ML}^{-1} (x - \Theta_{c,n})\} \exp\{-\frac{1}{2}(a - \Theta_{c,k})^T diag(\Sigma, k)^{-1} (a - \Theta_{c,k})\}
\end{aligned}
$$

Observing the LHS and using the improper uniform prior assumption we obtain:

$$\sqrt{(\frac{n+m}{m})^d}\int_{\theta\in\mathbb{R}^d} Q(\theta)\exp\{-\frac{1}{2}(x-\theta_{c,n})^T diag(\Sigma,n)^{-1}(x-\theta_{c,n})\}\exp\{-\frac{1}{2}(a-\theta_{c,k})^T diag(\Sigma,k)^{-1}(a-\theta_{c,k})\}$$

$$=\sqrt{(\frac{n+m}{m})^d}\int_{\theta\in\mathbb{R}^d}\exp\{-\frac{1}{2}[\sum_{i=1}^{n}x_i^T\Sigma^{-1}x_i+\sum_{j=1}^{k}a_j^T\Sigma^{-1}a_j-2\theta^T\Sigma^{-1}(n\bar{x}+k\bar{a})+(n+k)\theta^T\Sigma^{-1}\theta]\}$$

$$=\sqrt{(\frac{n+m}{m})^d}\exp\{-\frac{1}{2}[\sum_{i=1}^{n}x_i^T\Sigma^{-1}x_i+\sum_{j=1}^{k}a_j^T\Sigma^{-1}a_j-(n+k)(\frac{n\bar{x}+k\bar{a}}{n+k})^T\Sigma^{-1}(\frac{n\bar{x}+k\bar{a}}{n+k})]\}$$

$$\int_{\theta\in\mathbb{R}^d}\exp\{-\frac{n+k}{2}[(\theta-\frac{n\bar{x}+k\bar{a}}{n+k})^T\Sigma^{-1}(\theta-\frac{n\bar{x}+k\bar{a}}{n+k})]\}$$

$$=\sqrt{(\frac{n+m}{m})^d}\exp\{-\frac{1}{2}[\sum_{i=1}^{n}x_i^T\Sigma^{-1}x_i+\sum_{j=1}^{k}a_j^T\Sigma^{-1}a_j-(n+k)(\frac{n\bar{x}+k\bar{a}}{n+k})^T\Sigma^{-1}(\frac{n\bar{x}+k\bar{a}}{n+k})]\}\sqrt{(\frac{2\pi}{n+k})^d\det(\Sigma)}$$

Observing the RHS and using the improper uniform prior assumption we have:

$$\int_{\theta\in\mathbb{R}^d} Q(\theta)\exp\{-\frac{1}{2}(x-\theta_{c,n})^T\Psi_{ML}^{-1}(x-\theta_{c,n})\}\exp\{-\frac{1}{2}(a-\theta_{c,k})^T diag(\Sigma^{-1},k)(a-\theta_{c,k})\}$$

$$=\int_{\theta\in\mathbb{R}^d}\exp\{-\frac{1}{2}(x-\theta_{c,n})^T\Psi_{ML}^{-1}(x-\theta_{c,n})\}\exp\{-\frac{1}{2}(a-\theta_{c,k})^T diag(\Sigma^{-1},k)(a-\theta_{c,k})\}$$

$$\overset{F.14}{=}\int_{\theta\in\mathbb{R}^d}\exp\{-\frac{1}{2}(x-\theta_{c,n})^T(diag(\Sigma^{-1},n)-\frac{1}{n+m}rep(\Sigma^{-1},n))(x-\theta_{c,n})\}$$

$$\exp\{-\frac{1}{2}(a-\theta_{c,k})^T diag(\Sigma^{-1},k)(a-\theta_{c,k})\}$$

$$=\int_{\theta\in\mathbb{R}^d}\exp\{-\frac{1}{2}[\sum_{i=1}^{n}(x_i-\theta)^T\Sigma^{-1}(x_i-\theta)+\sum_{j=1}^{k}(a_j-\theta)^T\Sigma^{-1}(a_j-\theta)-\frac{n^2}{n+m}(\bar{x}-\theta)^T\Sigma^{-1}(\bar{x}-\theta)]\}$$

$$\overset{(i)}{=}\exp\{-\frac{1}{2}[\sum_{i=1}^{n}x_i^T\Sigma^{-1}x_i+\sum_{j=1}^{k}a_j^T\Sigma^{-1}a_j-\frac{n^2}{n+m}\bar{x}^T\Sigma^{-1}\bar{x}-\frac{nm+nk+mk}{n+m}v^T\Sigma^{-1}v]\}$$

$$\int_{\theta\in\mathbb{R}^d}\exp\{-\frac{1}{2}\frac{nm+nk+mk}{n+m}[v^T\Sigma^{-1}v]\}$$

$$=\exp\{-\frac{1}{2}[\sum_{i=1}^{n}x_i^T\Sigma^{-1}x_i+\sum_{j=1}^{k}a_j^T\Sigma^{-1}a_j-\frac{n^2}{n+m}\bar{x}^T\Sigma^{-1}\bar{x}-\frac{nm+nk+mk}{n+m}v^T\Sigma^{-1}v]\}$$

$$\sqrt{(\frac{2\pi(n+m)}{nm+nk+mk})^d\det(\Sigma)}$$

Where in (i) we denoted $v=\frac{nm\bar{x}+k(n+m)\bar{a}}{nm+nk+mk}$. Therefore the decision rule is

$$\exp\{\frac{1}{2}[(n+k)(\frac{n\bar{x}+k\bar{a}}{n+k})^T\Sigma^{-1}(\frac{n\bar{x}+k\bar{a}}{n+k})]\}\sqrt{(\frac{1}{m(n+k)})^d}>$$

$$\exp\{\frac{1}{2}[\frac{n^2}{n+m}\bar{x}^T\Sigma^{-1}\bar{x}+\frac{nm+nk+mk}{n+m}v^T\Sigma^{-1}v]\}\sqrt{(\frac{1}{nm+nk+mk})^d}$$

$$\Leftrightarrow\sqrt{(\frac{nm+nk+mk}{m(n+k)})^d}>$$

$$\exp\{\frac{1}{2}[\frac{n^2}{n+m}\bar{x}^T\Sigma^{-1}\bar{x}+\frac{nm+nk+mk}{n+m}v^T\Sigma^{-1}v-(n+k)(\frac{n\bar{x}+k\bar{a}}{n+k})^T\Sigma^{-1}(\frac{n\bar{x}+k\bar{a}}{n+k})]\}$$

$$\Leftrightarrow d\log(\frac{nm+nk+mk}{m(n+k)})>$$

$$\frac{n^2}{n+m}\bar{x}^T\Sigma^{-1}\bar{x}+\frac{nm+nk+mk}{n+m}v^T\Sigma^{-1}v-(n+k)(\frac{n\bar{x}+k\bar{a}}{n+k})^T\Sigma^{-1}(\frac{n\bar{x}+k\bar{a}}{n+k})$$

$$\begin{aligned}
&\Leftrightarrow d\log(\frac{nm+nk+mk}{m(n+k)}) > \frac{k^2n^2}{(n+k)(nm+nk+mk)}(\bar{x}-\bar{a})^T\Sigma^{-1}(\bar{x}-\bar{a})\\
&\Leftrightarrow (\bar{x}-\bar{a})^T\Sigma^{-1}(\bar{x}-\bar{a}) < \frac{d(n+k)(nm+nk+mk)}{k^2n^2}\log(\frac{nm+nk+mk}{m(n+k)})
\end{aligned}$$

As required. □

Theorem F.16. *Fix the attacker to be $\mathcal{G}_{ML}$ as defined in F.12, then the game value is:*

$$\max_{\mathcal{D}} V(\mathcal{D},\mathcal{G}_{ML}) \overset{\textit{Lemma F.15}}{=} V(\mathcal{D}_{ML},\mathcal{G}_{ML}) =$$

$$\frac{1}{2}+\frac{1}{2}\frac{1}{\Gamma(\frac{d}{2})}[\gamma(\frac{d}{2},\frac{d(nm+nk+mk)}{2nk}\log\frac{nm+nk+mk}{m(n+k)})-\gamma(\frac{d}{2},\frac{dm(n+k)}{2nk}\log\frac{nm+nk+mk}{m(n+k)})]=$$

$$\frac{1}{2}+\frac{1}{2}\frac{1}{\Gamma(\frac{d}{2})}[\gamma(\frac{d}{2},\frac{d}{2}(1+\rho+\delta)\log\frac{1+\rho+\delta}{\rho+\delta})-\gamma(\frac{d}{2},\frac{d}{2}(\rho+\delta)\log\frac{1+\rho+\delta}{\rho+\delta})]$$

Where $\rho=\frac{m}{k},\delta=\frac{m}{n}$, and γ is the lower incomplete gamma function

Proof. The game value is given by:

$$\begin{aligned}
V(\mathcal{D}^{ML},\mathcal{G}^{ML}) =&\mathbb{E}_{\Theta\sim Q}V(\Theta,\mathcal{D}_{ML},\mathcal{G}_{ML})\\
=&\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_\Theta^{(k)}}\mathbb{E}_{Y\sim f_\Theta^{(m)}}\left[\mathbb{E}_{X\sim f_\Theta^{(n)}}[\mathcal{D}_{ML}(A,X)]+\mathbb{E}_{X\sim g^{ML}_{X|Y}(\cdot|Y)}[1-\mathcal{D}_{ML}(A,X)]\right]\\
=&\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_\Theta^{(k)}}\mathbb{E}_{Y\sim f_\Theta^{(m)}}\\
&\left[\mathbb{E}_{X\sim f_\Theta^{(n)}}\left[I[\|\bar{X}-\bar{A}\|^2_{\Sigma^{-1}}<\alpha_{ML}]\right]+\mathbb{E}_{X\sim g^{ML}_{X|Y}(\cdot|Y)}\left[1-I[\|\bar{X}-\bar{A}\|^2_{\Sigma^{-1}}<\alpha_{ML}]\right]\right]\\
=&\frac{1}{2}+\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_\Theta^{(k)}}\mathbb{E}_{X\sim f_\Theta^{(n)}}\left[I[\|\bar{x}-\bar{A}\|^2_{\Sigma^{-1}}<\alpha_{ML}]\right]-\\
&\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_\Theta^{(k)}}\mathbb{E}_{X\sim g^{ML}_{X|\Theta}(\cdot|\Theta)}\left[I[\|\bar{x}-\bar{A}\|^2_{\Sigma^{-1}}<\alpha_{ML}]\right]
\end{aligned}$$

Observing the first term, we can see that by replacing α^* with α_{ML} in the analogue part of the proof for Theorem F.11 we get:

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_\Theta^{(k)}}\mathbb{E}_{X\sim f_\Theta^{(n)}}\left[I[\|\bar{X}-\bar{A}\|^2_{\Sigma^{-1}}<\alpha_{ML}]\right]\\
=&\frac{1}{2}\frac{1}{\Gamma(\frac{d}{2})}\gamma(\frac{d}{2},\frac{nk\alpha_{ML}}{2(n+k)})
\end{aligned}$$

Again, similarly to analogue part of the proof for Theorem F.11, observing the second term we have:

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_\Theta^{(k)}}\mathbb{E}_{X\sim g^{ML}_{X|\Theta}(\cdot|\Theta)}\left[I[\|\bar{X}-\bar{A}\|^2_{\Sigma^{-1}}<\alpha_{ML}]\right]\\
\equiv&\frac{1}{2}\mathbb{E}_{\Theta\sim Q}\mathbb{E}_{A\sim f_\Theta^{(k)}}\mathbb{E}_{X\sim g^{ML}_{X|\Theta}(\cdot|\Theta)}\left[I[V^TV<\alpha_{ML}]\right]\\
=&(*)
\end{aligned}$$

Where:

$$V=C^{-1}(\bar{X}-\bar{A})=C^{-1}[I_d,-I_d]\begin{bmatrix}\bar{X}\\ \bar{A}\end{bmatrix}=[C^{-1},-C^{-1}]\begin{bmatrix}\bar{X}\\ \bar{A}\end{bmatrix}$$

Using the definition of $\mathcal{G}_{ML}$ (Definition F.12) we have $\bar{X}\sim\mathcal{N}(\theta,\frac{n+m}{nm}\Sigma)$, using Lemma F.3 we have $\bar{A}\sim\mathcal{N}(\theta,\frac{1}{k}\Sigma)$, and thus:

$$\begin{bmatrix}\bar{X}\\ \bar{A}\end{bmatrix}\sim\mathcal{N}(\begin{bmatrix}\theta\\ \theta\end{bmatrix},\begin{bmatrix}\frac{n+m}{nm}\Sigma & 0\\ 0 & \frac{1}{k}\Sigma\end{bmatrix})$$

Therefore:

$$V \sim \mathcal{N}(0, [C^{-1}, -C^{-1}] \begin{bmatrix} \frac{n+m}{nm}\Sigma & 0 \\ 0 & \frac{1}{k}\Sigma \end{bmatrix} \begin{bmatrix} C^{-T} \\ -C^{-T} \end{bmatrix}) = \mathcal{N}(0_d, \frac{nm+nk+mk}{nmk} I_d)$$

We denote $\tilde{V} = \sqrt{\frac{nmk}{nm+nk+mk}} V$ and thus $\tilde{V}_1, \ldots, \tilde{V}_d$ are independent standard normal random variables and $\tilde{V}^T \tilde{V} \sim \chi^2(d)$. Therefore

$$V^T V \sim \Gamma(k = \frac{d}{2}, \theta = 2\frac{nm+nk+mk}{nmk})$$

And thus:

$$(*) = \frac{1}{2}\frac{1}{\Gamma(\frac{d}{2})}\gamma(\frac{d}{2}, \frac{nmk\alpha_{ML}}{2(nm+nk+mk)})$$

Therefore the game value is given by:

$$V(\mathcal{D}^*, \mathcal{G}^*) = \frac{1}{2} + \frac{1}{2}\frac{1}{\Gamma(\frac{d}{2})}[\gamma(\frac{d}{2}, \frac{nk\alpha_{ML}}{2(n+k)}) - \gamma(\frac{d}{2}, \frac{nmk\alpha_{ML}}{2(nm+nk+mk)})] =$$

$$\frac{1}{2} + \frac{1}{2}\frac{1}{\Gamma(\frac{d}{2})}[\gamma(\frac{d}{2}, \frac{d(nm+nk+mk)}{2nk}\log\frac{nm+nk+mk}{m(n+k)}) - \gamma(\frac{d}{2}, \frac{dm(n+k)}{2nk}\log\frac{nm+nk+mk}{m(n+k)}))]$$

As required. □

G EXPERIMENTS - DATASETS

Below we provide details on the datasets used in the authentication on faces and characters experiments. The VoxCeleb2 (Nagrani et al., 2017; Chung & Zisserman, 2018) dataset contains cropped face videos of 6112 identities. We used the original split of 5994 identities for training and 118 for test. For each identity we saved every fifth frame and resized each frame to 64×64 and augmented it using horizontal flip. The Omniglot dataset (Lake et al., 2015) contains handwritten character images from 50 alphabets. There are 1623 different characters, and 20 examples for each character. We use the splits and augmentations suggested by Vinyals et al. (2016) and used by Snell et al. (2017).

H EXPERIMENTS - IMPLEMENTATION DETAILS

In this section we describe our implementation of the GIM model in detail for the different experiments. Recall from Sec. 5, that in general the authenticator is a neural network $\mathcal{D}(a, x)$ that can be expressed as $\mathcal{D}(a, x) = \sigma(T_{\mathcal{D}}(a), T_{\mathcal{D}}(x))$, and the generator is a neural network $\mathcal{G}(y)$ that can be expressed as $\mathcal{G}(Y)_i = \varphi(W_i - \bar{W} + T_{\mathcal{G}}(Y)) \quad \forall i \in [n]$. In what follows we describe our implementation of these models for each of the experiments.

H.1 GAUSSIAN SOURCES

Authenticator Architecture: For the statistic function $T_{\mathcal{D}}$, we use a concatenation of the mean and standard deviation of the sample. For the comparison function σ, we use the element-wise absolute difference between the statistics $T_{\mathcal{D}}(a), T_{\mathcal{D}}(x)$, followed by a linear layer.

Attacker Architecture: For the statistic function $T_{\mathcal{G}}$ we use the sample mean, i.e $T_{\mathcal{G}}(y) = \bar{y}$. The noise vectors W_i are generated as follows: First, n Gaussian noise vectors $Z_1, \ldots, Z_n \sim \mathcal{N}(0, I_d)$ are drawn, then each vector Z_i is passed through a linear layer to obtain W_i. Finally, the decoder φ is the identity function.

Optimization details: The model is trained in an authentication setup as in our theoretical setup, using alternating gradient descent as is common in GAN optimization (Mescheder et al., 2018). Each iteration begins when a source $\theta \in \mathbb{R}^d$ is drawn from the prior distribution $Q = \mathcal{N}(0, 10I_d)$. Samples $A \in \mathbb{R}^{k \times d}, Y \in \mathbb{R}^{m \times d}, X_\theta \in \mathbb{R}^{n \times d}$ are drawn IID from

$f_\theta = \mathcal{N}(\theta, I_d)$, where X_θ represents a real sample from θ. The attacker, given the leaked sample Y, generates a fake sample $X_\mathcal{G} = \mathcal{G}(Y) \in \mathbb{R}^{n \times d}$, passes it to $\mathcal{D}$, and suffers the loss $-\log(\mathrm{sigmoid}(\mathcal{D}(\mathrm{A}, \mathrm{X}_\mathcal{G})))$. The authenticator, $\mathcal{D}$ receives as input the source information sample A, outputs a prediction for each of the test samples $X_\theta, X_\mathcal{G}$, and suffers the binary cross entropy loss. $-0.5\left(\log\left(\mathrm{sigmoid}(\mathcal{D}(\mathrm{A}, \mathrm{X}_\theta))\right) + \log\left(\mathrm{sigmoid}(1 - \mathcal{D}(\mathrm{A}, \mathrm{X}_\mathcal{G}))\right)\right)$. Each experiment is trained for $200K$ iterations with a batch size of 4000 using the Adam optimizer (Kingma & Ba, 2015) with learning rate 10^{-4}.

H.2 Experiments on Voxceleb2 and Omniglot

To describe the models we begin with some notation. We let c denote the number of image channels, h denote the image size (we only consider square images of size $c \times h \times h$), and l denote the latent dimension of the model.

Authenticator Architecture: The authenticator model consists of the following components: Two image encoders $E_{\mathcal{D},i} : [0,1]^{c \times h \times h} \to \mathbb{R}^l$ that map an image to a latent vector, a statistic module ζ that maps a set of latent vectors to a single fixed sized vector of size $5l$, and the comparator module σ that takes in two latent vectors and outputs a scalar. The exact specification of these components can be seen in the published code.

The statistic function $T_\mathcal{D}$ maps a sample of n images $x \in [0,1]^{n \times c \times h \times h}$ to a statistic vector $T_\mathcal{D}(x) \in \mathbb{R}^{5l}$ as follows: Every image X_i in the sample is mapped using $E_{\mathcal{D},1}$ and $E_{\mathcal{D},2}$ to latent vectors $v_i, u_i \in \mathbb{R}^l$. The latent sample u is passed through ζ to obtain a single vector $\tilde{u} \in \mathbb{R}^{5l}$ which gets concatenated to $\bar{v}$, the sample mean of v. Namely,

$$T_\mathcal{D}(x) = \mathrm{concat}\left(\bar{E_{\mathcal{D},1}(x)}, \zeta\left(E_{\mathcal{D},2}(x)\right)\right)$$

The comparison function $\sigma : \mathbb{R}^{5l} \to \mathbb{R}$ receives two latent vectors s_A, s_X representing the statistics of the samples A and X respectively. The vectors are concatenated and then passed through a Multi-Layered Perceptron which outputs a scalar reflecting their similarity. Namely:

$$\sigma(s_A, s_X) = \mathrm{MLP}\left(\mathrm{concat}\left(s_A, s_X\right)\right)$$

The full architecture of the authenticator is depicted in Fig. 7.

Attacker Architecture: Our implementation of the attacker is inspired by the architecture suggested in (Zakharov et al., 2019), which relies on an implicit assumption that an image could be modeled as a mapping of two latent vectors to image space. The first vector represents the source θ and is the same for any image of θ, the second vector represents the environment (e.g pose, lighting, expression) and is different for each image of the source.

The attacker model consists of the following components: An image encoder $\mathrm{E}_{\mathcal{G},src} : [0,1]^{c \times h \times h} \to \mathbb{R}^l$ that maps an image to a latent vector representing the source θ. an image encoder $\mathrm{E}_{\mathcal{G},env} : [0,1]^{c \times h \times h} \to \mathbb{R}^l$ that maps an image to a latent vector representing the environment, a Multi-layered Perceptron $\mathrm{MLP}_\mathcal{G} : \mathbb{R}^l \to \mathbb{R}^l$ that maps gaussian noise to the environment latent space, an environment decoder $\varphi_{env} : \mathbb{R}^l \to \mathbb{R}^{c \times h \times h}$ that maps a latent vector to an environment image, which could represent aspects of the environment such as facial landmarks[5], and finally a generator $\phi : \mathbb{R}^{2c \times h \times h} \times \mathbb{R}^l \to [0,1]^{c \times h \times h}$ that maps an environment image concatenated to the real image to a new image. The generator is based on the image to image model used in (Zakharov et al., 2019; Johnson et al., 2016) and uses the source latent vector as input for Adaptive instance normalization (Huang & Belongie, 2017). The exact specification of these components can be seen in the published code.

The attacker generates a fake sample $X \in [0,1]^{n \times c \times h \times h}$ based on a leaked sample $Y \in [0,1]^{m \times c \times h \times h}$ in the following way: Each image Y_j in the leaked sample is mapped using $E_{\mathcal{G},src}$ and $E_{\mathcal{G},env}$ to latent vectors $src_j, env_j \in \mathbb{R}^l$. A latent environment vector, $\tilde{env}_i$, for each fake image X_i is constructed in the following way: First, n IID Gaussian noise vectors $Z_1, \ldots, Z_n \sim \mathcal{N}(0, I_l)$ are drawn, then each vector Z_i is passed through $\mathrm{MLP}_\mathcal{G}$ to obtain W_i, and finally $\tilde{env}_i$ is obtained by matching the mean of the latent environment vectors to $\bar{env}$. Namely:

$$\tilde{env}_i = W_i - \bar{W} + \bar{env} \quad \forall i \in [n]$$

[5] In (Zakharov et al., 2019), our so called environment image is indeed a facial landmarks image which is used as input to the model. In our work we allow the model to learn which environment image is useful

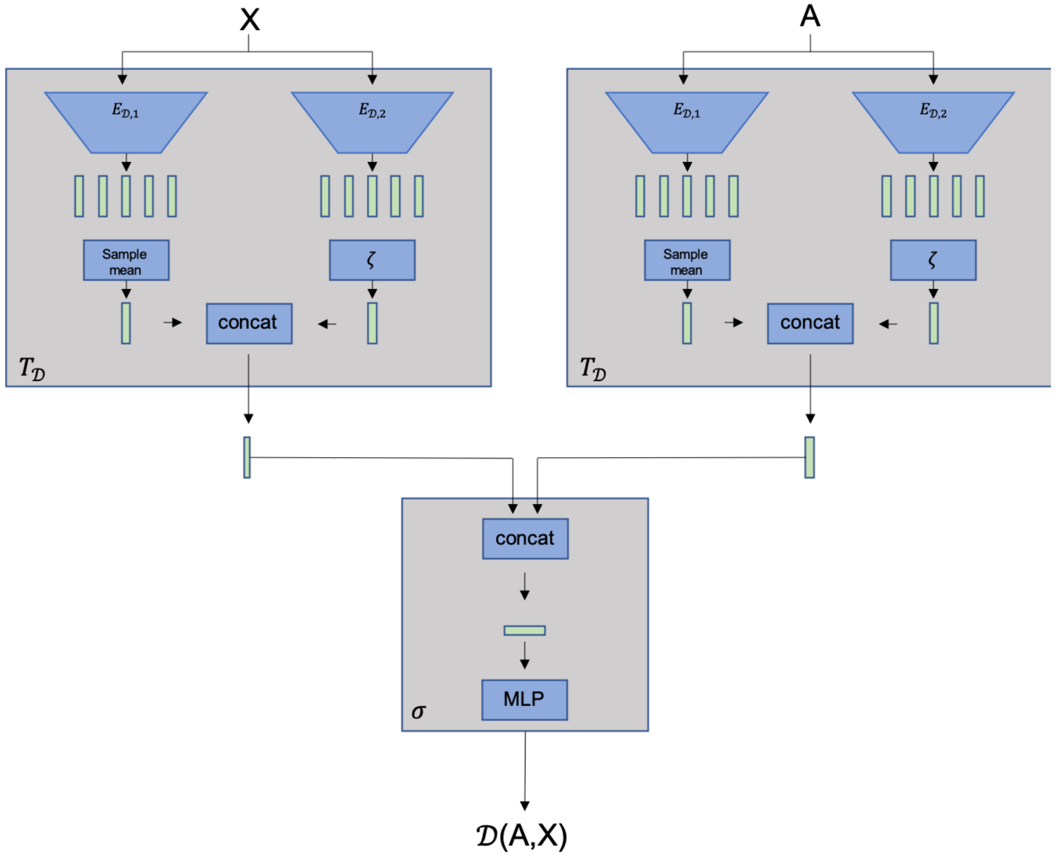


Figure 7: an overview on the implementation of the GIM authenticator architecture for the experiments on the Voxceleb2 and Omniglot datasets.

Each fake image X_i is then generated deterministically from the latent vectors $src, \tilde{env}_i$. $\tilde{env}_i$ is used as input to the decoder φ_{env} which outputs an environment image. This image is concatenated along the channel dimension to a random image from the leaked sample Y, and then passed as input to the generator ϕ, which also receives src as input to its Adaptive instance norm layers. The output of the generator is the fake image X_i for all $i \in [n]$. The full architecture of the attacker is depicted in Fig. 8.

Optimization details: The model is trained in an authentication setup as in our theoretical setup, using alternating gradient descent with the regularization parameter as suggested by Mescheder et al. (2018). Each iteration begins when a source $\theta \in \mathbb{R}^d$ is drawn uniformly from the dataset. Samples $A \in [-1, 1]^{k \times c \times h \times h}, Y \in [-1, 1]^{m \times c \times h \times h}, X_\theta \in [-1, 1]^{n \times c \times h \times h}$ are sampled uniformly from the images available for the source θ. The attacker, given the leaked sample Y, generates a fake sample $X_{\mathcal{G}} = \mathcal{G}(Y)$, passes it to $\mathcal{D}$, and suffers the loss $-\log(\mathrm{sigmoid}(\mathcal{D}(\mathrm{A}, \mathrm{X}_{\mathcal{G}})))$. The authenticator, $\mathcal{D}$ receives as input the source information sample A, outputs a prediction for each of the test samples $X_\theta, X_{\mathcal{G}}$, and suffers the binary cross entropy loss $-0.5\,(\log(\mathrm{sigmoid}(\mathcal{D}(\mathrm{A}, \mathrm{X}_\theta))) + \log(\mathrm{sigmoid}(1 - \mathcal{D}(\mathrm{A}, \mathrm{X}_{\mathcal{G}}))))$.

The experiments on Omniglot were trained for $520k$ iterations with batch size 128 using the Adam optimizer (Kingma & Ba, 2015) with learning rate 10^{-6} for $\mathcal{D}$, 10^5 for $\mathcal{G}$ and 10^7 for $\mathrm{MLP}_{\mathcal{G}}$ (as done in (Karras et al., 2018b)). The regularization parameter was set to 0.

The experiments on Voxceleb2 were trained for $250k$ iterations with batch size 64 using the Adam optimizer (Kingma & Ba, 2015) with learning rate 10^{-4} for both $\mathcal{D}$ and $\mathcal{G}$ and 10^6 for $\mathrm{MLP}_{\mathcal{G}}$. The regularization parameter was set to 10 (as done in (Karras et al., 2018b)) since we noticed that it

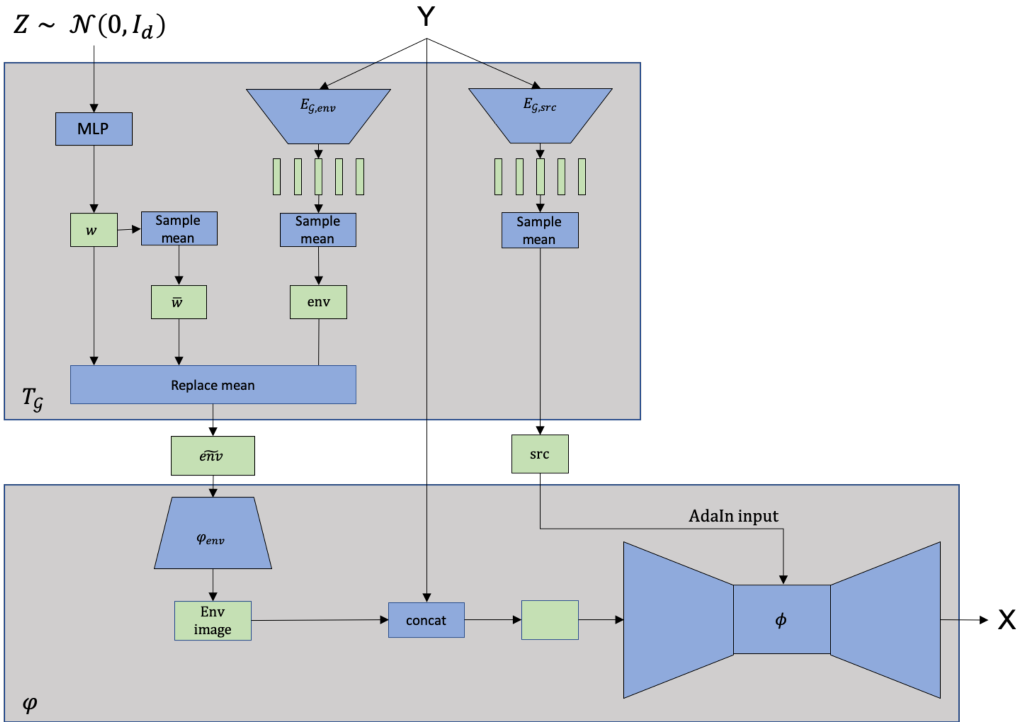


Figure 8: an overview on the implementation of the GIM attacker architecture for the experiments on the Voxceleb2 and Omniglot datasets.

stabilized and sped up the training, and in contrast to Omniglot and the Gaussian experiments did not seem to hurt the results.