UNDERSTANDING GAUSSIAN NOISE MISMATCH: A HELLINGER DISTANCE APPROACH

Kexin Huang[1] *Chaohua Shi*[2] *Lu Gan*[3] *Hongqing Liu*[4]

[1] College of Electronic Engineering, National University of Defense Technology, Anhui, China
[2] School of Electronic Engineering, Xidian Unviersity, Xi'an, China
[3] Department of Electrical and Electronics Engineering, Brunel University, London, UK
[4] School of Communication and Information Engineering,
Chongqing University of Posts and Telecommunications, Chongqing, China

ABSTRACT

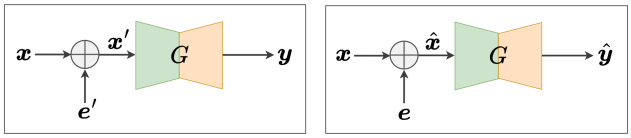
This paper explores noise-mismatched models using the Hellinger distance. In many applications, the design/training stage often assumes an independent and identically distributed (i.i.d.) Gaussian prior noise, but the real world introduces Gaussian noise with arbitrary covariance, creating a mismatch. We analyze the impact on system output and study optimal injected noise intensity for training/design. While theory assumes Gaussian sources, it provides guidance for non-Gaussian settings too. Experiments with CycleGAN for image-to-image translation validate the theory, producing results consistenting with derivations. Overall, this work provides theoretical and empirical insights into designing systems robust to noise uncertainties beyond simplified assumptions.

***Index Terms*—** Noise mismatch, Hellinger distance, f-divergence, Unpaired Image-to-Image translation.

1. INTRODUCTION

Noise mismatch occurs when the actual noise during data collection, processing, or communication deviates from the anticipated model used in system design or analysis. This can degrade system performance, produce faulty estimates, and reduce reliability. Such issues have been explored in various applications such as signal processing [1], communication [2], medical diagnosis [3], image processing [4], information theory [5–7], and machine learning [8, 9]. Prior studies often used Kullback-Leibler (KL) divergence (also known as relative entropy) to quantify the disparity between assumed and actual noise distributions [10–12]. Despite its popularity, KL divergence is not a true distance metric and is unbounded.

This paper explores the noise mismatch problem by leveraging the Hellinger distance for theoretical analysis. The Hellinger distance, bounded between 0 and 1, offers several advantages over KL divergence, such as symmetry, adherence to the triangle inequality, and robustness against outliers and small probability differences, enhancing its overall robustness [13]. Specifically, we examine noise mismatch in Gaussian channels. In particular, we assume that during the analysis/design/training phase, the source signal is corrupted by independent and identically distributed (i.i.d.) Gaussian noise with a probability distribution of $\mathcal{N}(0, \sigma_t^2 \boldsymbol{I}_d)$. However, practical scenarios often involve Gaussian noise with a different covariance matrix, described by the distribution $\mathcal{N}(0, \boldsymbol{\Sigma}_e)$. Consequently, we introduce a noise-mismatched model to investigate the interplay between designed and practical noises. Using the Hellinger distance, we examine the system's behaviour under varying covariance matrices, indirectly highlighting the system's resilience against stochastic disturbances. Furthermore, we determine the optimal solution σ_t^2 for the designed noise to ensure consistent system output under different noisy conditions. The rest of this paper is organized as follows. Section 2 formulates the problem. Section 3 states the main results. Numerical results are presented in Section 4 and conclusions are followed in Section 5.



(a) Analysis stage (b) Practical inference

Fig. 1. Noise mismatched model diagram in a general system: a) System $\boldsymbol{G}$ is designed to receive the perturbed source signal $\boldsymbol{x}'$ to produce $\boldsymbol{y}$; b) $\hat{\boldsymbol{x}}$ (source signal with random noise) outputs $\hat{\boldsymbol{y}}$.

Notations: Hereafter, the use of uppercase letters denotes random variables or vectors, and the corresponding lowercase letters signify their realizations. The notation $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is employed to represent a multidimensional normal (Gaussian) distribution characterized by a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. In the context of a given matrix $\boldsymbol{A}$, the symbols $\mathrm{Tr}(\boldsymbol{A})$ and $|\boldsymbol{A}|$ are utilized to denote the trace and determinant of $\boldsymbol{A}$, respectively. Furthermore, $\lambda_i(\boldsymbol{A})$ signifies the i-the largest eigenvalue of the matrix $\boldsymbol{A}$. $\boldsymbol{0}_d$ and $\boldsymbol{I}_d$ denote $d \times d$ all-zero and identity matrices, respectivley. For symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} > \boldsymbol{B}$ and $\boldsymbol{A} < \boldsymbol{B}$ denote that $\boldsymbol{A} - \boldsymbol{B}$ and $\boldsymbol{B} - \boldsymbol{A}$ are positive definite, respectively.

2. PROBLEM FORMULATION

Consider a clean source signal $\boldsymbol{x} \in \mathbb{R}^d$. During the design, analysis, or training phase, a system $\boldsymbol{G}$ processes a noisy signal $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{e}'$, as illustrated in Figure 1a. The term $\boldsymbol{G}$ can have varied interpretations across disciplines: it might be a "generative model" in machine learning, a "channel" in communication, or an "imaging system" in signal processing. The intentional noise $\boldsymbol{e}'$ added during design is i.i.d. Gaussian distributed, given by $\boldsymbol{e}' \sim \mathcal{N}(\boldsymbol{0}, \sigma_t^2 \boldsymbol{I}_d)$, and is independent of the signal. In contrast, the inference stage introduces a noise $\boldsymbol{e}$ to yield a signal $\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{e}$, producing an output $\hat{\boldsymbol{y}} = \boldsymbol{G}(\hat{\boldsymbol{x}})$, as depicted in Figure 1b. Our objective is to understand the system's response when the noise $\boldsymbol{e}$ deviates from the design noise $\boldsymbol{e}'$.

The squared Hellinger distance is a special case of f-divergence, a statistical measure often used to measure the differences between

two probability distributions. Given two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ on a measurable space $\mathcal{X}$ and let p and q represent the Radon–Nikodym derivatives of $\mathbb{P}$ and $\mathbb{Q}$, respectively. The squared Hellinger distance is defined as [14]

$$D_H(\mathbb{P}\|\mathbb{Q}) = \frac{1}{2}\int_{\mathcal{X}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2. \tag{1}$$

In case when $\mathbb{P}$ and $\mathbb{Q}$ are two multivariate normal distributions with $p \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $q \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, $D_H(\mathbb{P}\|\mathbb{Q})$ is given by [14]

$$\begin{aligned} D_{H^2}(\mathbb{P}\|\mathbb{Q}) =& 1 - \frac{|\boldsymbol{\Sigma}_1|^{1/4}|\boldsymbol{\Sigma}_2|^{1/4}}{\left|\frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}\right|^{1/2}} \cdot \exp\Bigg\{-\frac{1}{8}(\boldsymbol{\mu}_1 - \\ & \boldsymbol{\mu}_2)^\top \left(\frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}\right)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\Bigg\}, \end{aligned} \tag{2}$$

which can help us analyze the important mathematical properties of the noise-mismatched model.

In what follows, we provide an in-depth analysis of the noise-mismatched model, addressing key theoretical foundations:

- Under what circumstances does a designed system, introducing perturbed inputs via Gaussian noise, outperform a system processing only the clean signal given mismatched channel noise estimations?
- How does the system behave when the actual noise $\boldsymbol{e}$ has a different distribution from $\boldsymbol{e}'$?
- How can we optimize the value of σ_t^2 during design to enhance resilience against stochastic uncertainties?

3. SQUARED HELLINGER DISTANCE ANALYSIS FOR NOISE MISMATCHED MODEL

Consider $\boldsymbol{X}'$ as the random variable corresponding to the Gaussian noise corrupted source signal $\boldsymbol{x}'$ and $\boldsymbol{Y} = \boldsymbol{G}(\boldsymbol{X}')$ as the corresponding output signal. Likewise, let $\hat{\boldsymbol{X}}$ represent the actual noisy input signal and $\hat{\boldsymbol{Y}} = \boldsymbol{G}(\hat{\boldsymbol{X}})$ its output. Based on the data processing inequality [14], we have

$$D_H(\mathbb{P}_{\hat{\boldsymbol{X}},\hat{\boldsymbol{Y}}}\|\mathbb{P}_{\boldsymbol{X}',\boldsymbol{Y}}) = D_H(\mathbb{P}_{\hat{\boldsymbol{X}}}\|\mathbb{P}_{\boldsymbol{X}'}), \tag{3}$$

$$D_H(\mathbb{P}_{\hat{\boldsymbol{Y}}}\|\mathbb{P}_{\boldsymbol{Y}}) \leq D_H(\mathbb{P}_{\hat{\boldsymbol{X}}}\|\mathbb{P}_{\boldsymbol{X}'}), \tag{4}$$

with equality in (4) if system $\boldsymbol{G}$ is invertible. This implies that the Hellinger divergence of the altered input data distribution, $D_H\left(\mathbb{P}_{\hat{\boldsymbol{X}}}\|\mathbb{P}_{\boldsymbol{X}'}\right)$, remains consistent in the joint distribution of shifted input and output data. Besides, $D_H\left(\mathbb{P}_{\hat{\boldsymbol{X}}}\|\mathbb{P}_{\boldsymbol{X}'}\right)$ bounds the f-divergence $D_H\left(\mathbb{P}_{\hat{\boldsymbol{Y}}}\|\mathbb{P}_{\boldsymbol{Y}}\right)$ of the marginal target distributions. Next, we study $D_H(\mathbb{P}_{\hat{\boldsymbol{X}}}\|\mathbb{P}_{\boldsymbol{X}'})$ to characterize the system's robustness to noise mismatch for Gaussian signals, as described in the following theorem:

Theorem 1. *Let $\boldsymbol{x}$ denote the clean source signal, $\boldsymbol{e}'$ the noise used in training/analysis/design, and $\boldsymbol{e}$ the actual input noise. Assume that each is modeled as d-dimensional Gaussian vectors with probability distributions given by $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$, $\boldsymbol{e}' \sim \mathcal{N}(\boldsymbol{0}, \sigma_t^2\boldsymbol{I}_d)$ and $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_e)$, respectively. Define the squared Hellinger distance function F as*

$$F(\sigma_t^2, \boldsymbol{\Sigma}_e) = D_H\left(\mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s + \sigma_t^2\boldsymbol{I}_d)\|\mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_e)\right).$$

Then, $F(\sigma_t^2, \boldsymbol{\Sigma}_e)$ possesses the following properties:

1. *$F(0, \boldsymbol{\Sigma}_e) > F(\sigma_t^2, \boldsymbol{\Sigma}_e)$ whenever $\boldsymbol{\Sigma}_e > \frac{1}{2}\sigma_t^2\boldsymbol{I}_d$.*
2. *Regarding the 1D case with fixed σ_t^2 and σ_s^2, as σ_e^2 varies:*
   - *$F(\sigma_t^2, \sigma_e^2)$ decreases when $\sigma_e^2 < \sigma_t^2$ and increases when $\sigma_e^2 > \sigma_t^2$.*
   - *The curvature of $F(\sigma_t^2, \sigma_e^2)$ exhibits strict convexity within the interval $0 \leq \sigma_e^2 \leq \epsilon$ and becomes concave when $\sigma_e^2 > \epsilon$, in which $\epsilon = \sigma_t^2 + \frac{2\sqrt{10}}{5}(\sigma_t^2 + \sigma_s^2)$.*
3. *For multi-dimensional cases where $d > 1$ and given that $\boldsymbol{\Sigma}_e = \sigma_e^2\boldsymbol{I}_d$, for a fixed σ_t^2 and as σ_e^2 changes*
   - *$F(\sigma_t^2, \sigma_e^2\boldsymbol{I}_d)$ decreases for $\sigma_e^2 < \sigma_t^2$ and increases when $\sigma_e^2 > \sigma_t^2$.*
   - *The function $F(\sigma_t^2, \sigma_e^2\boldsymbol{I}_d)$ is strictly convex in σ_e^2 if*

$$\sigma_L^2 < \sigma_e^2 < \sigma_t^2 + \sqrt{\frac{8}{d+4}}(\lambda_d(\Sigma_s) + \sigma_t^2),$$

   in which $\sigma_L^2 = \max(0, \sigma_t^2 - \sqrt{\frac{8}{d+4}}(\lambda_d(\Sigma_s) + \sigma_t^2)$ and it is strictly concave if $\sigma_e^2 > \sigma_t^2 + \frac{2\sqrt{10}}{5}(\sigma_t^2 + \lambda_1(\boldsymbol{\Sigma}_s))$, in which $\lambda_d(\boldsymbol{\Sigma}_s)$ and $\lambda_1(\boldsymbol{\Sigma}_s)$ represent the smallest and largest eigenvalues of $\boldsymbol{\Sigma}_s$, respectively.
4. *For a fixed variance σ_t^2 and two distinct covariance matrices $\boldsymbol{\Sigma}_{e1}$ and $\boldsymbol{\Sigma}_{e2}$, the following inequality is satisfied if either $\boldsymbol{\Sigma}_{e1} < \boldsymbol{\Sigma}_{e2} < \sigma_t^2\boldsymbol{I}_d$ or $\sigma_t^2\boldsymbol{I}_d < \boldsymbol{\Sigma}_{e2} < \boldsymbol{\Sigma}_{e1}$,*

$$F(\sigma_t^2, \boldsymbol{\Sigma}_{e1}) > F(\sigma_t^2, \boldsymbol{\Sigma}_{e2}). \tag{5}$$

The outline of the proof can be found in the Appendix.

Property 1 indicates that when every eigenvalue of $\boldsymbol{\Sigma}_e$ is at least $\sigma_t^2/2$, the system's Hellinger distance is smaller with Gaussian injection than without it. This suggests the importance of noise injection in training/design. ***Properties 2-4*** characterize the behaviour of F as the true noise conditions diverge from the design's assumptions. For 1D signals (***Property 2***), F manifests a clear monotonicity and undergoes convex/concave shifts depending on the actual noise variance. In higher-dimensional contexts where the noise is isotropic (***Property 3***), F mirrors these patterns. Notably, the identified convexity zone is merely a sufficient condition — the genuine convex region might extend beyond this. This convex nature reveals that minor noise alterations in real-world scenarios yield pronounced system changes within certain noise intensities. ***Property 4*** illustrates that F diminishes as the actual noise covariance matrix, $\boldsymbol{\Sigma}_e$, edges closer to the design's anticipated noise level, $\sigma_t^2\boldsymbol{I}_d$. Collectively, these insights offer a theoretical foundation for comprehending how models might behave in real-world situations with mismatched noise.

We now explore the optimal choice of σ_t^2, to enhance the system's robustness against stochastic uncertainties. Corollary 1 focuses on the case when $\boldsymbol{e}$ is iid:

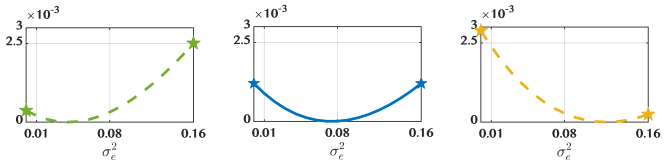
Corollary 1. *Given an iid input noise $\boldsymbol{e}$ with $\boldsymbol{\Sigma}_e = \sigma_e^2\boldsymbol{I}_d$ and a bounded variance $0 \leq \sigma_e^2 \leq M$, define $\sigma_{t,o}^2$ as the optimal noise level that minimizes the worst-case squared Hellinger distance $F(\sigma_t^2, \sigma_e^2\boldsymbol{I}_d)$:*

$$\sigma_{t,o}^2 = \arg\min_{\sigma_t^2}\left\{\max_{0\leq\sigma_e^2\leq M} F(\sigma_t^2, \sigma_e^2\boldsymbol{I}_d)\right\}.$$

For this optimal level, it satisfies $F(\sigma_{t,o}^2, \boldsymbol{0}_d) = F(\sigma_{t,o}^2, M\boldsymbol{I}_d)$.

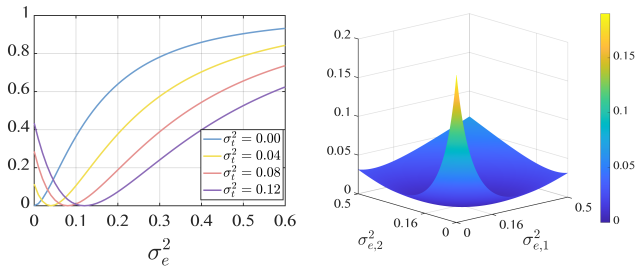
Moreover, when the source $\boldsymbol{x}$ is also iid with $\boldsymbol{\Sigma}_s = \sigma_s^2\boldsymbol{I}_d$, $\sigma_{t,o}^2$ simplifies to:

$$\sigma_{t,o}^2 = \sigma_s^2\left(\sqrt{1 + \frac{M}{\sigma_s^2}} - 1\right). \tag{6}$$



(a) $\sigma_t^2 < \sigma_{t,o}^2$ (b) $\sigma_t^2 = \sigma_{t,o}^2$ (c) $\sigma_t^2 > \sigma_{t,o}^2$

Fig. 2. Geometric visualization of the max-min optimization for σ_t^2 under the constraints of actual input iid Gaussian noise variance, $0 \leq \sigma_e^2 \leq 0.16$. a), b), and c) depict the variations of $F(\sigma_t^2, \mathbf{\Sigma}_e)$ with different values of σ_t^2.



(a) iid noise (b) Non-uniform independent noise

Fig. 3. Visualization of Squared Hellinger distance for AR(1) source signal model with $\mathbf{\Sigma}_s(k,l) = \sigma_s^2 \rho^{|k-l|}$ $(0 \leq k, l \leq d-1)$. (a) $\rho = 0.9$ and $d = 16$. $\boldsymbol{e}$ is isotrophic Gaussian noise with variance σ_e^2. (b) $\rho = 0.95$ and $d = 2$. $\boldsymbol{e}$ is nonuniform independant Gaussian noise with covariance matrix $\mathbf{\Sigma}_e = \text{diag}(\sigma_{e,1}^2, \sigma_{e,2}^2)$.

Proof. For brevity, we outline the main steps: Drawing from Property 3 in Theorem 1, we deduce that for a fixed σ_e^2, the function $F(\sigma_t^2, \sigma_e^2 \boldsymbol{I}_d)$ is decreasing when $\sigma_t^2 < \sigma_e^2$ and increasing for $\sigma_t^2 > \sigma_e^2$. Since $\sigma_{t,o}^2$ satisfies $F(\sigma_{t,o}^2, \mathbf{0}_d) = F(\sigma_{t,o}^2, M\boldsymbol{I}_d)$, for the range $0 < \sigma_{t,o}^2 < \sigma_t^2$, $F(\sigma_t^2, \mathbf{0}_d) > F(\sigma_{t,o}^2, \mathbf{0}_d)$. Likewise, for $\sigma_t^2 < \sigma_{t,o}^2 < M$, $F(\sigma_t^2, M\boldsymbol{I}_d) > F(\sigma_{t,o}^2, M\boldsymbol{I}_d)$. A visual representation for the above argument is provided in Fig. 2. For iid source, (6) is derived directly by setting $F(\sigma_{t,o}^2, \mathbf{0}_d) = F(\sigma_{t,o}^2, M\boldsymbol{I}_d)$. □

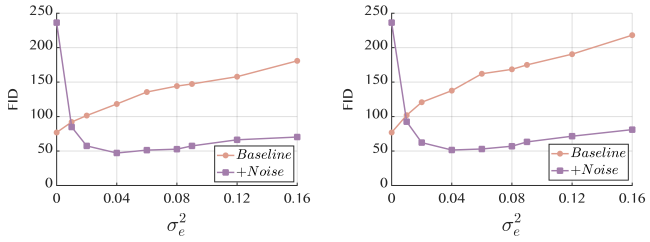
Corollary 1 introduces a min-max optimization framework for choosing the designed noise level σ_t^2, enhancing resilience to discrepancies between the presumed and actual noise distributions. For a small M/σ_s^2, a first-order Taylor series approximation of (6) suggests $\sigma_{t,o}^2 \approx M/2$. This can serve as an initial solution in computational searches for $\sigma_{t,o}^2$ in the case of a non-iid Gaussian source.

4. NUMERICAL RESULTS

4.1. Gaussian signal source

To enable readers have a quick check of Theorem 1, two examples of $F(\sigma_t^2, \mathbf{\Sigma}_e)$ are visualized in Figure 3. Here, we consider an AR(1) signal model with covariance matrix $\mathbf{\Sigma}_s(k,l) = \sigma_s^2 \rho^{|k-l|}$ $(0 \leq k, l \leq d-1)$. Specifically,

Example 1: Figure 3a depicts a signal of length $d = 16$ and $\rho = 0.9$. We plot $F(\sigma_t^2, \sigma_e^2 \boldsymbol{I}_d)$ in the range $0 \leq \sigma_e^2 \leq 0.6$ using $\sigma_t^2 = 0.04j$ for $0 \leq j \leq 3$. Observations from Figure 3a reveal that



(a) Gaussian noise (b) Colored Gaussian noise

Fig. 4. Comparison of FID scores for Horse → Zebra conversion under a) iid Gaussian noise of variance σ_e^2 and b) Colored Gaussian noise, with σ_e^2 denoting the average noise variance.

the behaviours of $F(\sigma_t^2, \sigma_e^2 \boldsymbol{I}_d)$ are consistent with the properties in Theorem 1. Specifically, when $\sigma_t^2 > 0$, $F(\sigma_t^2, \sigma_e^2 \boldsymbol{I}_d) < F(0, \sigma_e^2 \boldsymbol{I}_d)$ holds for $\sigma_e^2 > 0.5\sigma_t^2$, aligning with Property 1. Also, $F(\sigma_t^2, \sigma_e^2 \boldsymbol{I}_d)$ first decreases and then increases in σ_e^2, agreeing with the monotonicity property. Moreover, it exhibits convexity near σ_t^2 and concavity as σ_e^2 becomes large, as suggested by Property 3.

Example 2: In the second example, $d = 2$ and the covariance matrix is $\mathbf{\Sigma}_e = \text{diag}(\sigma_{e,1}^2, \sigma_{e,2}^2)$, where $0 < \sigma_{e,i}^2 \leq 0.5$ $(i = 1, 2)$. The injected noise has $\sigma_t^2 = 0.16$. The graphical representation is shown in Figure 3b. Consistent with Property 4 of Theorem 1, F decreases as both $\sigma_{e,1}^2$ and $\sigma_{e,2}^2$ approach 0.16.

4.2. Non-Gaussian signals

While our theory assumes Gaussian sources, our results provide valuable insights for non-Gaussian signals. To illustrate, we apply Gaussian noise injection to unpaired image-to-image translation, specifically the Horse to Zebra dataset with a Cycle-GAN model [15]. All image pixel values are normalized to $[0, 1]$, and Gaussian noise, $\mathcal{N}(0, 0.04\boldsymbol{I}_d)$, is applied to each training image. The test images are corrupted by both i.i.d. Gaussian and colored Gaussian noise of different intensities. The colored noise is generated by applying a 2D Gaussian filter for iid Gaussian noise, featuring a 0.5 standard deviation and a 7×7 window size. The training follows the original Cycle-GAN setup.

We evaluate translations with the popular FID (Fréchet Inception Distance) score [16], which assesses the distributional gap between translated and target images; lower scores indicate superior quality. Results, alongside baseline counterparts, are in Fig. 4. Fig. 5 shows pictures of noisy horse-to-zebra translations. Unlike the Squared Hellinger distance, FID contrasts the mean and variance between Gaussian feature distributions of real versus generated images. Nevertheless, FID patterns closely resemble Hellinger divergence predictions. One can see that as noise intensity increases, our Gaussian noise-injected model outperforms the baseline, especially when $\sigma_e^2 > \sigma_t^2/2 = 0.02$, reinforcing **Property 1**. Baseline performance drops substantially as noise intensity becomes higher. In alignment with **Properties 3-4** from Theorem 1, our model's FID is optimal when training and test noise variances coincide; but deteriorates with variance disparities, evident in Gaussian (Fig.4a) and colored Gaussian noise contexts (Fig.4b). Notably, the Gaussian noise-injected model faces challenges with clean data due to training noise bias, a trade-off we aim to address in our future work.

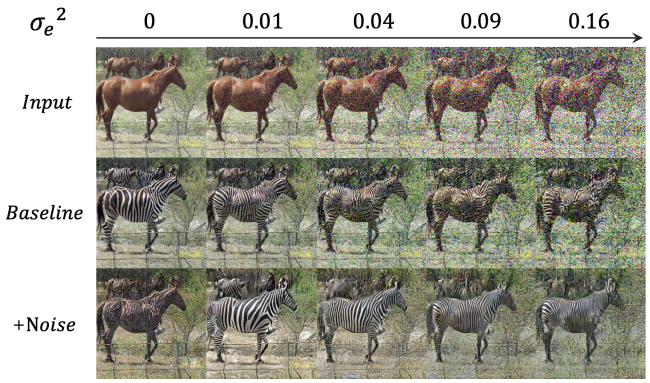


Fig. 5. Horse-to-zebra translation with "Baseline" and noise-injected CycleGAN under colored Gaussian noise.

5. CONCLUSION

This work provides theoretical and empirical insights into designing robust systems under noise mismatch. Our analysis using the Hellinger distance characterizes the impact of mismatch between assumed i.i.d. Gaussian noise and actual arbitrary covariance Gaussian noise, guiding the selection of optimal injected noise levels during training/design to maximize robustness. Experiments with CycleGAN for image translation validate the theory, providing useful guidance even for non-Gaussian settings. Overall, by accounting for noise uncertainties through proper noise injection and optimization, this research enables reliable system desgins that can withstand mismatches between simplified noise assumptions and reality.

A. PROOF OF THEOREM 1

Proof. Due to lack of space, we present proof details of Properties 1 and 2 and part of Property 3. The rest of the proof will be left in the journal version.

Property 1: According to Eq. 2, $D_H(\mathbb{P}_{\boldsymbol{X}'}, \mathbb{P}_{\hat{\boldsymbol{X}}})$, and the squared distence between $D_H(\mathbb{P}_{\boldsymbol{X}}, \mathbb{P}_{\hat{\boldsymbol{X}}})$ are given by

$$F(\sigma_t^2, \boldsymbol{\Sigma}_e) = 1 - \frac{|\boldsymbol{\Sigma}_s + \sigma_t^2 \boldsymbol{I}_d|^{\frac{1}{4}} |\boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_e|^{\frac{1}{4}}}{\left|\boldsymbol{\Sigma}_s + \frac{\boldsymbol{\Sigma}_e + \sigma_t^2 \boldsymbol{I}_d}{2}\right|^{\frac{1}{2}}}, \tag{7}$$

$$F(0, \boldsymbol{\Sigma}_e) = 1 - \frac{|\boldsymbol{\Sigma}_s|^{\frac{1}{4}} |\boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_e|^{\frac{1}{4}}}{\left|\boldsymbol{\Sigma}_s + \frac{\boldsymbol{\Sigma}_e}{2}\right|^{\frac{1}{2}}}. \tag{8}$$

Using Eq. 7 and Eq. 8, the proof is equivalent to showing

$$\begin{aligned}
&\frac{|\boldsymbol{\Sigma}_s|^{1/4} |\boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_e|^{1/4}}{\left|\boldsymbol{\Sigma}_s + \frac{\boldsymbol{\Sigma}_e}{2}\right|^{1/2}} < \frac{\left|\boldsymbol{\Sigma}_s + \sigma_t^2 \boldsymbol{I}_d\right|^{1/4} |\boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_e|^{1/4}}{\left|\boldsymbol{\Sigma}_s + \frac{\boldsymbol{\Sigma}_e + \sigma_t^2 \boldsymbol{I}_d}{2}\right|^{1/2}} \\
\iff &\frac{\left|\boldsymbol{\Sigma}_s + \frac{\boldsymbol{\Sigma}_e + \sigma_t^2 \boldsymbol{I}_d}{2}\right|^2}{\left|\boldsymbol{\Sigma}_s + \frac{\boldsymbol{\Sigma}_e}{2}\right|^2} < \frac{|\boldsymbol{\Sigma}_s + \sigma_t^2 \boldsymbol{I}_d|}{|\boldsymbol{\Sigma}_s|} \\
\iff &\prod_{i=1}^{d} \left(1 + \frac{\frac{1}{2}\sigma_t^2}{\lambda_i\left(\boldsymbol{\Sigma}_s + \frac{\boldsymbol{\Sigma}_e}{2}\right)}\right)^2 < \prod_{i=1}^{d} \left(1 + \frac{\sigma_t^2}{\lambda_i\left(\boldsymbol{\Sigma}_s\right)}\right) \\
\iff &2\sum_{i=1}^{d} \ln\left(1 + \frac{\frac{1}{2}\sigma_t^2}{\lambda_i\left(\boldsymbol{\Sigma}_s + \frac{\boldsymbol{\Sigma}_e}{2}\right)}\right) < \sum_{i=1}^{d} \ln\left(1 + \frac{\sigma_t^2}{\lambda_i\left(\boldsymbol{\Sigma}_s\right)}\right).
\end{aligned} \tag{9}$$

As $\boldsymbol{\Sigma}_e > \frac{1}{2}\sigma_t^2 \boldsymbol{I}_d$, it is straightforward that $\boldsymbol{\Sigma}_s + \frac{\boldsymbol{\Sigma}_e}{2} > \boldsymbol{\Sigma}_s + \frac{1}{4}\sigma_t^2 \boldsymbol{I}_d$. Using Weyl's inequality [17] for eigenvalues, we know that

$$\lambda_i(\boldsymbol{\Sigma}_s + \frac{\boldsymbol{\Sigma}_e}{2}) > \lambda_i(\boldsymbol{\Sigma}_s) + \frac{1}{4}\sigma_t^2.$$

Hence, the LHS (Left-Hand Side) of Eq. 9 can be bounded by [18]

$$2\sum_{i=1}^{d} \ln\left(1 + \frac{\frac{1}{2}\sigma_t^2}{\lambda_i\left(\boldsymbol{\Sigma}_s + \frac{\boldsymbol{\Sigma}_e}{2}\right)}\right) < 2\sum_{i=1}^{d} \ln\left(1 + \frac{\frac{1}{2}\sigma_t^2}{\lambda_i\left(\boldsymbol{\Sigma}_s\right) + \frac{1}{4}\sigma_t^2}\right). \tag{10}$$

As $\lambda_i(\boldsymbol{\Sigma}_s) > 0$, it can be easily shown that

$$\left(1 + \frac{\frac{1}{2}\sigma_t^2}{\lambda_i\left(\boldsymbol{\Sigma}_s\right) + \frac{1}{4}\sigma_t^2}\right)^2 < \left(1 + \frac{\sigma_t^2}{\lambda_i\left(\boldsymbol{\Sigma}_s\right)}\right). \tag{11}$$

By combining (10) and (11), we know that (9) holds and then Property 1 is proved.

Property 2: For $1d$ case, one can derive that the first order partial derivative $\frac{\partial F(\sigma_t^2, \sigma_e^2)}{\partial \sigma_e^2}$ with respect to σ_e^2 is

$$\frac{\partial F(\sigma_t^2, \sigma_e^2)}{\partial \sigma_e^2} = \frac{\sqrt{2}(\sigma_s^2 + \sigma_t^2)^{\frac{1}{4}}(\sigma_e^2 - \sigma_t^2)}{4(\sigma_s^2 + \sigma_t^2)^{\frac{3}{4}}(\sigma_e^2 + 2\sigma_s^2 + \sigma_t^2)^{\frac{3}{2}}}.$$

This implies that $F(\sigma_t^2, \sigma_e^2)$ is increasing for $\sigma_e^2 > \sigma_t^2$ and decreasing for $\sigma_e^2 < \sigma_t^2$. The second-order partial derivative is

$$\frac{\partial F^{(2)}(\sigma_t^2, \sigma_e^2)}{\partial^{(2)} \sigma_e^2} = \frac{\sqrt{2}(\sigma_s^2 + \sigma_t^2)^{\frac{1}{4}}\psi(\sigma_t^2, \sigma_e^2)}{16(\sigma_e^2 + \sigma_s^2)^{\frac{7}{4}}(\sigma_e^2 + 2\sigma_s^2 + \sigma_t^2)^{\frac{5}{2}}},$$

in which $\psi(\sigma_t^2, \sigma_e^2) = -5\sigma_e^4 + 10\sigma_e^2\sigma_t^2 + 8\sigma_s^4 + 16\sigma_s^2\sigma_t^2 + 3\sigma_t^4$. It can be shown $\psi(\sigma_t^2, \sigma_e^2) > 0$ when $\sigma_e^2 < \epsilon$ and negative otherwise, with $\epsilon = \sigma_t^2 + \frac{2\sqrt{10}}{5}(\sigma_t^2 + \sigma_s^2)$. This proves the convex/concave properties of $F(\sigma_t^2, \sigma_e^2)$.

Property 3: For $d > 1$ and given $\boldsymbol{\Sigma}_e = \sigma_e^2 \boldsymbol{I}_d$, leveraging the properties of determinants allows us to express the squared Hellinger distance as $F(\sigma_t^2, \sigma_e^2 \boldsymbol{I}_d) = 1 - \prod_{i=1}^{d} \zeta_i(\sigma_{s,i}^2, \sigma_t^2, \sigma_e^2)$ in which

$$\zeta_i(\sigma_{s,i}^2, \sigma_t^2, \sigma_e^2) = \frac{(\sigma_{s,i}^2 + \sigma_t^2)^{\frac{1}{4}}(\sigma_{s,i}^2 + \sigma_e^2)^{\frac{1}{4}}}{(\sigma_{s,i}^2 + 0.5\sigma_t^2 + 0.5\sigma_e^2)^{\frac{1}{2}}},$$

and $\sigma_{s,i}^2$ is the i-th eigenvalue of $\boldsymbol{\Sigma}_s$. From Property 2, it's evident that each ζ_i is increasing when $\sigma_e^2 < \sigma_t^2$ and decreasing behaviour otherwise, which in turn indicates the monotonic nature of F.

To establish the concavity of F, we start by observing from Property 2 that each function $\zeta_i(\sigma_{s,i}^2, \sigma_t^2, \sigma_e^2)$ is positive and strictly convex in σ_e^2 for $\sigma_e^2 > \sigma_t^2 + \frac{2\sqrt{10}}{5}(\sigma_t^2 + \sigma_{s,1}^2)$. We then invoke a mathematical principle: for two positive, strictly convex functions p and q defined on an interval $\mathcal{I}$, if the derivatives p' and q' are both negative across this interval, then their product, pq, remains convex on $\mathcal{I}$. This is justified as $(pq)'' = p''q + 2p'q' + q'' > 0$ under these conditions. Employing this principle and inductive reasoning, we can ascertain that the product $\prod_{i=1}^{d} \zeta_i(\sigma_{s,i}^2, \sigma_t^2, \sigma_e^2)$ is strictly convex for $\sigma_e^2 > \sigma_t^2 + \frac{2\sqrt{10}}{5}(\sigma_t^2 + \sigma_{s,1}^2)$. Consequently, $F(\sigma_t^2, \sigma_e^2 \boldsymbol{I}_d)$ exhibits concavity within this domain. The proof of convexity will be shown in the journal version.

□

B. REFERENCES

[1] François Chapeau-Blondeau and David Rousseau, "Noise-enhanced performance for an optimal bayesian estimator," *IEEE Transactions on Signal Processing*, vol. 52, no. 5, pp. 1327–1334, 2004.

[2] Hao Wu, "LMMSE channel estimation in OFDM systems: A vector quantization approach," *IEEE Communications Letters*, vol. 25, no. 6, pp. 1994–1998, 2021.

[3] Ryutaro Tanno, Daniel Worrall, Enrico Kaden, Aurobrata Ghosh, Francesco Grussu, Alberto Bizzi, Stamatios Sotiropoulos, Antonio Criminisi, and Daniel Alexander, "Uncertainty modelling in deep learning for safer neuroimage enhancement: Demonstration in diffusion MRI," *NeuroImage*, vol. 225, pp. 117366, 01 2021.

[4] Darwin T Kuan, Alexander A Sawchuk, Timothy C Strand, and Pierre Chavel, "Adaptive noise smoothing filter for images with signal-dependent noise," *IEEE transactions on pattern analysis and machine intelligence*, , no. 2, pp. 165–177, 1985.

[5] Dongning Guo, Shlomo Shamai, and Sergio Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE transactions on information theory*, vol. 51, no. 4, pp. 1261–1282, 2005.

[6] Miquel Payaro and Daniel P. Palomar, "Hessian and concavity of mutual information, differential entropy, and entropy power in linear vector Gaussian channels," *IEEE Transactions on Information Theory*, vol. 55, no. 8, pp. 3613–3628, 2009.

[7] Dongning Guo, Yihong Wu, Shlomo S. Shitz, and Sergio Verdú, "Estimation in Gaussian noise: Properties of the minimum mean-square error," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2371–2385, 2011.

[8] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard, "Robustness of classifiers: from adversarial to random noise," *Advances in neural information processing systems*, vol. 29, 2016.

[9] Qing Da, Yang Yu, and Zhi-Hua Zhou, "Learning with augmented class by exploiting unlabeled data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2014, vol. 28.

[10] Saloua Chlaily, Chengfang Ren, Pierre-Olivier Amblard, Olivier Michel, Pierre Comon, and Christian Jutten, "Information–estimation relationship in mismatched gaussian channels," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 688–692, 2017.

[11] Sergio Verdú, "Mismatched estimation and relative entropy," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 3712–3720, 2010.

[12] Minhua Chen and John Lafferty, "Mismatched estimation and relative entropy in vector Gaussian channels," in *2013 IEEE International Symposium on Information Theory*, 2013, pp. 2845–2849.

[13] Brett E Bissinger, R Lee Culver, and NK Bose, "Minimum Hellinger distance based classification of underwater acoustic signals," in *2009 43rd Annual Conference on Information Sciences and Systems*. IEEE, 2009, pp. 47–49.

[14] Yury Polyanskiy, "Information theoretic methods in statistics and computer science," 2019, online lecture notes, available at https://people.lids.mit.edu/yp/homepage/data/LN_fdiv.pdf.

[15] 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, pp. 2223–2232.

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[17] Joel N Franklin, *Matrix theory*, Courier Corporation, 2012.

[18] Neven Elezovic and Josip Pečarić, "A note on Schur-convex functions," *Rocky Mountain Journal of Mathematics*, vol. 30, pp. 853–856, 2000.