

# THE AWE AND MYSTERY OF T-NORM

Jiří Navrátil, Ganesh N. Ramaswamy

IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

e-mail: {jiri, ganeshr}@us.ibm.com

## ABSTRACT

A popular score normalization technique termed T-norm is the central focus of this paper. Based on widely confirmed experimental observation regarding T-norm tilting the DET curves of speaker detection systems, we set out to identify the components taking role in this phenomenon. We claim that under certain local assumptions the T-norm performs a gaussianization of the individual true and impostor score populations and further derive conditions for clockwise and counter-clockwise DET rotations caused by this transform.

## 1. INTRODUCTION

In an unrelated presentation which inspired the title of this paper [1], a linear feature transform was referred to by the terms of awe and mystery due to its formal “elegance” and its seemingly unmatched performance across various tasks. By adopting these terms we would like to dedicate this discourse to a particularly popular member of a class of linear score transforms frequently occurring in speaker detection, termed T-norm. Although its complexity is of no match to the scheme cited above, the T-norm is broadly recognized as an extremely powerful normalization technique due to its adaptivity and proven effectivity across tasks and data corpora. However, a certain degree of “mystery” lingers about this subject: the T-norm has an intriguing behavior which is observable in Detection Error Trade-Off (DET) curves (DET is a variate of Receiver Operating Characteristics) in that the performance curves not only shift towards lower error rates but in virtually all cases an apparent rotation (tilt) is introduced in favor of low false alarm rates (i.e. a counter-clockwise tilt). This observation, already noted by Auckenthaler et al. in the work introducing the T-norm [2] (“The cohort approaches seem to rotate the DET plot in favor of lower miss probability at low false alarm rates”) has since been confirmed in numerous publications across different datasets. In this paper, we set out to identify the components that contribute to the above mentioned phenomenon aiming at lifting the veil of mystery attributed to T-norm – the otherwise celebrated technique.

### 1.1. Preliminaries

We will assume reader’s familiarity with the basics of current speaker detection systems, most of which are based on Gaussian Mixture models, as described in [3, 4]. Denote  $X_0$  an acoustic test sample,  $X_{ref}$  a reference sample from the hypothesized speaker,  $M_S := M(X_{ref})$  a model representation of a speaker  $S$  (given that  $X_{ref}$  originates from  $S$ ). A detection system computes a (probabilistic) score  $s(X_0, M_S)$  for the hypothesis that  $X_0$  and  $X_{ref}$  both originate from  $S$ . Statistically, by making a decision, every non-perfect detection system will incur two types of errors according to a False Alarm probability ( $P_{FA}$ ), and a Miss probability ( $P_M$ ) distribution. We refer to a sin-

gle detection test as a “trial,” whereby distinguishing two trial populations, namely a set of true trials  $\mathcal{T}$  and a set of negative (impostor) trials  $\bar{\mathcal{T}}$ . Let  $p(s)$  be an overall score density function composed from individual trial populations

$$p(s) = \pi_{\mathcal{T}} p_{\mathcal{T}}(s) + (1 - \pi_{\mathcal{T}}) p_{\bar{\mathcal{T}}}(s) \quad (1)$$

with a population prior  $\pi_{\mathcal{T}}$ .

A family of linear score normalization functions of the form

$$s' = \frac{s - \mu(\cdot)}{\sigma(\cdot)} \quad (2)$$

has been investigated in the literature with various alternatives of estimating the parameters  $\mu$  and  $\sigma$  (such as Z-, H-, and T-norm)[4, 2]. The T-norm, as introduced in [2], identifies  $\mu, \sigma$  as the mean and the standard deviation of scores over  $\bar{\mathcal{T}}$ , given a particular test sample  $X_0$ . The corresponding estimates are calculated using an impostor score subset  $c \subset \bar{\mathcal{T}}|_{X_0}$ , i.e. utilizing an auxiliary set of false models  $\mathcal{M}_c$  scored against  $X_0$  to generate  $c$ :

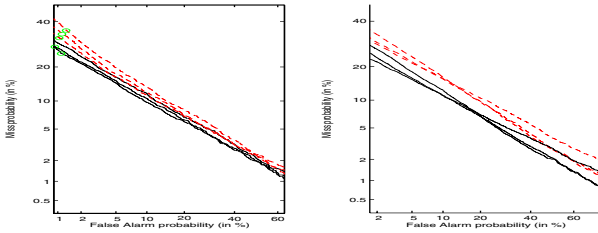
$$\begin{aligned} \hat{\mu}(X_0) &= E[s(X_0, M \in \mathcal{M}_c)|M] \\ \hat{\sigma}^2(X_0) &= var[s(X_0, M \in \mathcal{M}_c)|M] \end{aligned} \quad (3)$$

Unlike other schemes, the T-norm has the advantageous capability of adapting the parameters to the actual test sample  $X_0$ , thus accounting for acoustic environment changes. An acoustic channel may be viewed as a complex non-linear function distorting the input speech signal with propagating effects through the system causing changes in local score distributions. Since most systems use a global threshold on an estimate  $\hat{p}(s)$ , such variations increase the overall error rates. Using (3), the T-norm compensates for such distortions in terms of scale and bias.

There is a noteworthy connection between T-norm and so-called cohort-based systems [5]. By selecting  $\mathcal{M}_c$  so as to focus on models most competing to  $M_S$  and setting  $\sigma(\cdot) \equiv 1$ , the T-norm is identical to a cohort-based system. Furthermore, in Universal-Background-Model (UBM) systems with likelihood-ratio detectors, the  $\mu$ -only norm entirely replaces the UBM contribution in the ratio.

### 1.2. Experimental Motivation

Let us now inspect Figure 1 (left-hand plot) displaying DET curves for multiple configurations of a UBM-based system, obtained using the NIST-2002 evaluation dataset as described in [3]. The configurations differ in GMM size, ranging from small models (256 mixture components) to large (2048 components). Scores from the baseline systems (dashed curves) are normalized by the T-norm (solid curves) resulting in improvements in major parts of the operating region, most distinctively in the low  $P_{FA}$  area. This seems to be due to an obvious counter-clockwise tilt in the curves, observable in all configurations.



**Figure 1.** Comparison of plain (dashed) and T-normed (solid) systems (left-hand plot) and their gaussian approximation via synthetic data (right-hand plot)

It can be shown that for gaussian distributed populations the DET slope is proportional to the ratio of standard deviations of the two score populations  $\frac{\sigma_{\bar{\tau}}}{\sigma_{\tau}}$  ( $\sigma$ -ratio later on) [6, 7]. Therefore, a counter-clockwise line rotation implies a reduced variance in  $p_{\bar{\tau}}(s)$  relative to  $p_{\tau}(s)$ . Although the curves in Figure 1 appear close to linear indicating near gaussianity, a measurement on the data reveals that the actual  $\sigma$ -ratio remains same and in some configurations even increases contradicting the originally seen counter-clockwise tilt. The right-hand part of Figure 1 offers an insight into why this happens: solid and dashed curves again correspond to T-normed and baseline configurations respectively, however all curves were generated using synthetic data drawn randomly from gaussian distributions with same means and variances as the respective real systems. Apparently, if the actual distributions  $p_{\tau}(s)$  and  $p_{\bar{\tau}}(s)$  were gaussian both plots would be identical and the T-norm would in fact cause little or even a slight *clockwise* tilt, contrary to the observation so far (more experimental detail is provided in Section 2.). Obviously, the gaussianity assumption is violated and hence another attribute needs to be measured to explain the counter-clockwise tilt - the degree of gaussianity. For this purpose we employ the negentropy  $J$ , defined as Kullback-Leibler divergence between a density  $f(s)$  and a gaussian  $\mathcal{N}_f$  with identical mean and variance as  $f(s)$

$$J(f(\cdot)) = D_{KL}(f(\cdot) \parallel \mathcal{N}_f) = \int f(s) \log \frac{f(s)}{\mathcal{N}(s, \mu_f, \sigma_f)} ds \quad (4)$$

The negentropy is non-negative and reaches minimum (0) iff  $f(s) \sim \mathcal{N}_f$ . Empirical values of  $J$  obtained via a discretized version of (4) for the individual systems suggest that the T-normed systems experience a “gaussianization” phenomenon, i.e. a process of changing  $p_{\tau \in \mathcal{T}, \bar{\mathcal{T}}}(s)$  in the sense of minimizing (4) (more detail is discussed in Section 2.1.). Reduced negentropy values imply straightened DET curves and consequently account for the visual tilt.

The rest of this paper deals with the two phenomena in more detail: the gaussianization (Section 2.) and the DET rotation in  $\sigma$ -ratio sense (Section 3.) along with additional experimental evidence.

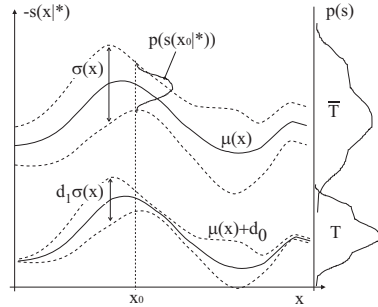
## 2. T-NORM AND GAUSSIANIZATION

Let  $x \in \mathbb{R}^n$  be a continuous random variable describing an acoustic observation and  $s : \mathbb{R}^n \rightarrow \mathbb{R}^1, s(x, M)$  a function of  $x$  and a speaker model  $M$ .

**Definition 1** The function  $s(x, M)$  is said to be locally gaussian-distributed for a model population  $\mathcal{M}_c$  if

$$p(s(x_0, M) | M \in \mathcal{M}_c) \sim \mathcal{N}(s, \mu(x_0), \sigma^2(x_0)), \forall x_0 \in \mathbb{R}^n \quad (5)$$

where  $\mu(x), \sigma(x) > 0$  are differentiable mean and standard deviation functions of  $x$  respectively.



**Figure 2.** A one-dimensional illustration of local score gaussianity

Local gaussianity provides for an assumption that at any acoustic observation point  $x = x_0$ , the scores, e.g. log-likelihood ratios, obtained from all models of a population producing e.g. the impostor trials  $\bar{\mathcal{T}}|_{x_0}$ , will be normally distributed with an arbitrary variance around an arbitrary mean both functions of the acoustic location  $x_0$ .

An illustration of local gaussianity is shown in Figure 2 depicting score distribution parameters for a one-dimensional example of  $x$ , and the global distributions  $p_{\bar{\tau}}(s)$  and  $p_{\tau}(s)$  which are generally non-gaussian.

By noting that for locally gaussian-distributed score functions

$$p_{\bar{\tau}}(s) = \int_{\mathbb{R}^n} \pi(x) \mathcal{N}(s, \mu(x), \sigma^2(x)) dx \quad (6)$$

with  $\pi(x)$  the acoustic prior probability, we state the following proposition on T-norm

**Proposition 1** For score density function  $p_{\tau}(s)$  of a trial population  $\tau = \bar{\mathcal{T}}$  obtained from a locally gaussian-distributed function via (6), the transform:  $s' = c_1 \frac{s - \mu(x)}{\sigma(x)} + c_0$  gaussianizes  $p_{\tau}(s)$ , whereby  $c_0, c_1$  are arbitrary constants.

A proof is given in the Appendix.

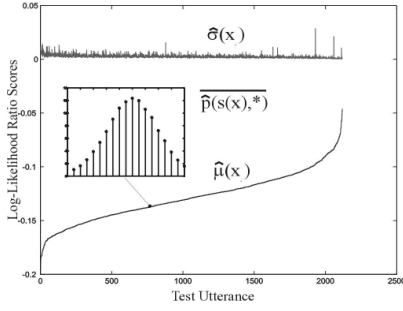
**Remark 1** The non-uniqueness of the T-norm transform within a global bias and scale  $c_0, c_1$  implies that another population, e.g.  $\tau = \mathcal{T}$ , with scores distributed with  $\mu'(x) = \mu(x) + d_0$ , and  $\sigma'(x) = \sigma(x)d_1, d_1 > 0$ , i.e. within a constant scale and bias from the population  $\tau$  is gaussianized simultaneously.

**Remark 2** For the parameter estimates (3) the result holds asymptotically as the expected values converge.

**Remark 3** The result is generally non-unique. For example, a transform  $s' = c_1 u(x) \frac{s - \mu(x)}{\sigma(x)} + c_0$  with  $u(x) \in \{-1, 1\}$  an arbitrary sign-flipping function also results in gaussianity. Furthermore  $p_{\tau}(s)$  being initially gaussian allows for constant transforms.

### 2.1. Experimental Observations

Empirical negentropy values for GMM systems of sizes ranging from 256 to 2048 components for both populations obtained on the NIST-2002 set are shown in Table 1. It can be seen that the T-norm indeed consistently reduces the divergence and transforms the impostor population distributions closer to gaussianity in all cases, and the target populations in most cases (Remark 2). The fact that both populations become more gaussian simultaneously confirms that the true-trial population  $\mathcal{T}$  tends to be within a global shift and scale from the mean and variance parameters of  $\bar{\mathcal{T}}$  according to Remark 1. Empirical values of  $\hat{\mu}(x), \hat{\sigma}(x)$  for about 2200 female-speaker test utterances of the dataset are shown in Figure 3, whereby the  $\mu$



**Figure 3. Empirical  $\hat{\mu}(x)$ ,  $\hat{\sigma}(x)$  and averaged local score distribution (histogram)**

GMM Size	$J \times 10^{-2}$		$R_\sigma$	
	Plain ( $\mathcal{T}/\bar{\mathcal{T}}$ )	T-norm	Plain	T-norm
256	1.2/1.8	1.6/1.4	.76	.79
512	1.4/2.2	1.3/1.2	.70	.70
1024	2.3/2.4	1.4/1.2	.64	.63
2048	3.6/3.3	2.5/1.1	.58	.55

**Table 1. Empirical negentropy and  $\sigma$ -ratio values obtained on the NIST-2002 dataset using various GMM-system sizes**

values are sorted in ascending order for better viewability. Due to a lack of true trials for the same utterances (only 1 true trial per utterance available) no such plot for  $\mathcal{T}$  is possible preventing a direct validity assessment of Remark 1. Furthermore, an averaged histogram of local scores on a per-utterance basis is embedded in Figure 3. The averaging is done to smooth individual histograms suffering from lacking sample size (each utterance has about 100 imposter scores). This plot should be directly compared to the assumption of local gaussianity in Definition 1 and appears to be not far from gaussian in terms of histogram shape. This suggests a good validity of our assumption made in Proposition 1.

### 3. T-NORM AND DET ROTATION

As described in Section 1.2., besides the gaussianization effect the T-norm transform also changes the  $\sigma$ -ratio [6] accounting for a slope change in the gaussian approximation of the DET curve (see Figure 1). Assuming global gaussianity for each score population, it is interesting to inspect when the T-norm induces a clockwise(+) and counter-clockwise(-) rotation. To simplify the analysis we assume the parameter  $\sigma(x) = \text{constant}$  thus  $s' = s - \hat{\mu}(x)$ . Furthermore, let  $\sigma_{1,2x}$  and  $\mu_{1,2x}$  denote the standard deviations and mean functions of  $x$  whereby the index 1 refers to imposter trials  $\bar{\mathcal{T}}$  and index 2 to true trial population  $\mathcal{T}$ . Using the mean-value notation  $\bar{f}_x = \int \pi(x)f(x)dx$ , the overall population variance can be written as

$$\begin{aligned} \sigma_{1,2}^2 &= \overline{\sigma_{1,2x}^2} + \overline{(\mu_{1,2x} - \mu_{1,2})^2} \\ &= \overline{\sigma_{1,2x}^2} + \sigma_{1,2\mu}^2 \end{aligned} \quad (7)$$

We first assume that the T-norm parameter function is identical to the imposter mean function  $\mu(x) \equiv \mu_{1x}$  (as in Proposition 1) and so in general  $\mu(x) \not\equiv \mu_{2x}$ . Then, the  $\mu$ -only transform modifies the squared  $\sigma$ -ratio as

$$\frac{\sigma_1^2}{\sigma_2^2} = \frac{\overline{\sigma_{1x}^2} + \sigma_{1\mu}^2}{\overline{\sigma_{2x}^2} + \sigma_{2\mu}^2} \rightarrow \frac{\overline{\sigma_{1x}^2}}{\overline{\sigma_{2x}^2} + \sigma_{2\mu}^2 + \sigma_{1\mu}^2 - 2\text{Cov}(\mu_{1x}, \mu_{2x})} \quad (8)$$

In a further simplification step, we consider  $\mu_{1x}, \mu_{2x}$  cross-correlated with coefficient  $\rho \in [-1, 1]$  and both functions being of identical variance  $\sigma_\mu^2 := \sigma_{1\mu}^2 = \sigma_{2\mu}^2$ . Then, the squared  $\sigma$ -ratio increases due to the  $\mu$ -only T-norm if the following inequality holds

$$\begin{aligned} \frac{\overline{\sigma_{1x}^2} + \sigma_\mu^2}{\overline{\sigma_{2x}^2} + \sigma_\mu^2} &< \frac{\overline{\sigma_{1x}^2}}{\overline{\sigma_{2x}^2} + 2\sigma_\mu^2(1 - \rho)} \\ \Rightarrow \rho &> \frac{\overline{\sigma_{2x}^2} + \overline{\sigma_{1x}^2} + 2\sigma_\mu^2}{2\overline{\sigma_{1x}^2} + 2\sigma_\mu^2} \end{aligned} \quad (9)$$

Obviously, (9) has no solution for  $\rho \in [-1, 1]$  when  $\overline{\sigma_{2x}^2} \geq \overline{\sigma_{1x}^2}$ , which in turn can be experimentally verified as true in all known cases. Moreover, if  $\sigma_\mu^2 \gg \overline{\sigma_{x1}^2}$  any solution range for  $\rho$  will tend to lie close to +1. Empirical values observed on the evaluation dataset indicate that the  $\mu$ -only T-norm actually increases the  $\sigma$ -ratio contradicting (9) thus indicating that the assumption  $\mu(x) \equiv \mu_{1x}$  may not hold due to estimation errors introduced in  $\hat{\mu}(x)$ . To account for imperfect normalization of the imposter population the estimate  $\hat{\mu}(x)$  is modeled as a partially correlated function of the true  $\mu(x)$  with  $\rho_1 \in [-1, 1]$  and identical variance  $\sigma_\mu$ . The true trial function  $\mu_{2x}$  also correlates with  $\mu(x)$  via  $\rho_2 \in [-1, 1]$ . Equation (9) is then re-stated

$$\frac{\overline{\sigma_{1x}^2} + \sigma_\mu^2}{\overline{\sigma_{2x}^2} + \sigma_\mu^2} < \frac{\overline{\sigma_{1x}^2} + 2\sigma_\mu^2(1 - \rho_1)}{\overline{\sigma_{2x}^2} + 2\sigma_\mu^2(1 - \rho_2)} \quad (10)$$

By inserting  $\rho_2 = \nu\rho_1, \nu \in [-1, 1]$  we can derive the **Clockwise Rotation Condition** on  $\rho_1$ :

$$\rho_1 \cdot 2[\sigma_\mu^2(\nu - 1) + (\overline{\sigma_{1x}^2}\nu - \overline{\sigma_{2x}^2})] > \overline{\sigma_{1x}^2} - \overline{\sigma_{2x}^2} \quad (11)$$

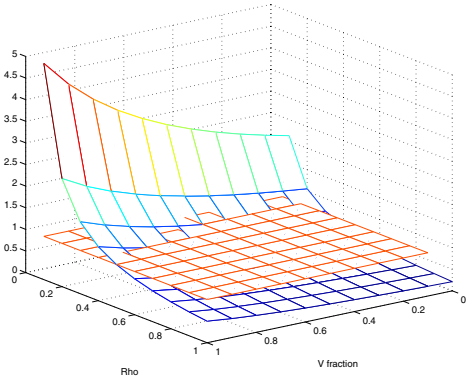
In the particular case of  $\nu = 1$  the condition (11) reduces to

$$\begin{aligned} \rho_1 &> \frac{1}{2} \text{ for } \overline{\sigma_{1x}^2} > \overline{\sigma_{2x}^2} \\ \rho_1 &< \frac{1}{2} \text{ for } \overline{\sigma_{1x}^2} < \overline{\sigma_{2x}^2} \end{aligned} \quad (12)$$

i.e. for the typical case  $\overline{\sigma_{1x}^2} < \overline{\sigma_{2x}^2}$  the correlation coefficient of the estimate must to be less than 0.5 otherwise the T-norm operates in the counter-clockwise mode as in the case (9). The correlation coefficient may be reduced by either estimation inaccuracies or by intentional modifications to the model selection when computing  $\hat{\mu}(x)$ , such as in cohort methods. A practical case of  $\sigma_\mu^2 = 2.3 \times 10^{-4}$ ,  $\overline{\sigma_{1x}^2} = .0022$ ,  $\overline{\sigma_{2x}^2} = .004$  as obtained on the NIST-2002 dataset is shown in Figure 4 as function of positive  $(\rho_1, \nu)$ . Here, both sides of (11) were divided by the left side for  $\nu \in [0, 1]$  and  $\rho_1 > 0$  so that the remaining  $f = 1$  plain demarkates the two orientation regions, namely the clockwise-rotating above and the counter-clockwise-rotating below the 1-plain shown in the plot.

### 4. CONCLUSIONS

We have shown that under certain local gaussianity assumptions the T-norm always transforms the global imposter population to a gaussian and that it may simultaneously do so for target score distributions, practically resulting in straightened DET curves. The underlying assumptions were experimentally verified as reasonably valid. Furthermore, it is shown how T-norm induces a rotation on a linear DET curve and derive a simplified condition for its orientation. Both the gaussianization



**Figure 4. Rotation surface based on (11) for empirical values  $\sigma_\mu^2 = 2.3 \times 10^{-4}$ ,  $\sigma_{1x} = .0022$ ,  $\sigma_{2x} = .004$ . In areas above the 1-surface the T-norm operates in +-rotation mode**

and the rotational effect interact in the praxis and result in an apparent counter-clockwise tilt of the DET curve commonly observed in the literature. With a better understanding of the T-norm, more suitable estimation methods may be chosen in order to target specific DET objectives.

## REFERENCES

- [1] G. Saon, “The awe and mystery of FMLLR.” Seminar presentation, IBM Human Language Technologies, Yorktown Heights, NY, November 2001.
- [2] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, “Score normalization for text-independent speaker verification systems,” *Digital Signal Processing*, vol. 10, pp. 42–54, January/April/July 2000.
- [3] G. Ramaswamy, J. Navrátil, U. Chaudhari, and R. Zilca, “The IBM system for the NIST 2002 cellular speaker verification evaluation,” in *Proc. of the ICASSP*, (Hong Kong), April 2003.
- [4] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.
- [5] D. Reynolds, “Comparison of background normalization methods for text-independent speaker verification,” in *Proc. of the EUROSPEECH*, September 1997.
- [6] J. Navrátil and G. Ramaswamy, “DETAC - a discriminative criterion for speaker verification,” in *Proc. of the ICSLP*, (Denver, CO), September 2002.
- [7] K. Fukunaga, *Statistical Pattern Recognition*. Academic Press, 1990. 2nd Edition.
- [8] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley and Sons, 1991.

## 5. APPENDIX

To verify Proposition 1 we consider the scores  $s(x, \cdot)$  locally gaussian-distributed around  $\mu(x)$  with  $\sigma(x)$ , which are, at the same time, the true parameters of the T-norm (2) when  $\tau = \bar{\tau}$ . It follows that the transform

$$s'(x, M) = c_1 \frac{s(x, M) - \mu(x)}{\sigma(x)} + c_0 \quad (13)$$

with  $c_1, c_0$  constant entails

$$p_\tau(s'(x, \cdot)) \sim \mathcal{N}(c_0, c_1^2), \quad \forall x \quad (14)$$

and (6) becomes

$$p_{\bar{\tau}}(s') = \int_{\mathbb{R}^n} \pi(x) \mathcal{N}(s', c_0, c_1^2) dx = \mathcal{N}(s', c_0, c_1^2) \quad (15)$$

For the interested reader we derive an upper bound on  $J(p)$  illustrating how individual  $\mu(x), \sigma(x)$  statistics contribute to the non-gaussianity of  $p_\tau$ . We make use of the Log-Sum Inequality [8], and re-state it for continuous functions:

**Lemma 1** For positive differentiable functions  $a(x)$  and  $b(x)$

$$\int_{\mathbb{R}^n} a(x) \log \frac{a(x)}{b(x)} dx \geq \int_{\mathbb{R}^n} a(x) dx \log \frac{\int_{\mathbb{R}^n} a(x) dx}{\int_{\mathbb{R}^n} b(x) dx} \quad (16)$$

with equality iff  $a(x)/b(x) = \text{constant}$

The distribution  $p_\tau(s)$  is gaussian iff the Negentropy

$$J(p) = \int p_\tau(s) \log \frac{p_\tau(s)}{\mathcal{N}(s, \mu_p, \sigma_p^2)} ds = 0 \quad (17)$$

with  $\mu_p, \sigma_p^2$  the mean and variance of  $\tau$ . For brevity let  $\mathcal{N}_x = \mathcal{N}(s, \mu(x), \sigma^2(x))$  and  $\mathcal{N}_0 = \mathcal{N}(s, \mu_p, \sigma_p^2)$ . Using (6) and  $\int_{\mathbb{R}^n} \pi(x) \mathcal{N}_0 dx = \mathcal{N}_0$ ,  $J(p)$  can be upper-bounded using the Lemma 1 as follows:

$$\begin{aligned} & \int \int_{\mathbb{R}^n} \pi(x) \mathcal{N}_x \log \frac{\int_{\mathbb{R}^n} \pi(x) \mathcal{N}_x dx}{\int_{\mathbb{R}^n} \pi(x) \mathcal{N}_0 dx} ds \\ & \leq \int \int_{\mathbb{R}^n} \pi(x) \mathcal{N}_x \log \frac{\mathcal{N}_x}{\mathcal{N}_0} dx ds + \int_{\mathbb{R}^n} \pi(x) \log \frac{\pi(x)}{\pi(x)} dx \\ & = \int_{\mathbb{R}^n} \pi(x) D_{KL}(\mathcal{N}_x \parallel \mathcal{N}_0) dx + D_{KL}(\pi \parallel \pi) \end{aligned} \quad (18)$$

The KL divergence in the second term vanishes while the first has a closed form solution:

$$D_{KL}(\mathcal{N}_x \parallel \mathcal{N}_0) = \frac{1}{2} \left[ \left( \frac{\sigma_x^2 - \sigma_0^2}{\sigma_x \sigma_0} \right)^2 + \frac{(\sigma_x^2 + \sigma_0^2)(\mu_x - \mu_0)^2}{\sigma_x^2 \sigma_0^2} \right] \quad (19)$$

with  $\sigma_x = \sigma(x)$ ,  $\mu_x = \mu(x)$ , and noting that  $\mu_0 = \int_{\mathbb{R}^n} \pi(x) \mu(x) dx$ , and  $\sigma_0^2 = \int_{\mathbb{R}^n} \pi(x) (\mu(x) - \mu_0)^2 dx + \int_{\mathbb{R}^n} \pi(x) \sigma(x)^2 dx$ . Using the notation  $\bar{f}_x = \int \pi(x) f(x) dx$  we rewrite the upper bound (18):

$$J(p) \leq (\mu_0^2 + \sigma_\mu^2 + \bar{\sigma_x^2}) \frac{1}{\bar{\sigma_x^2}} + \frac{\bar{\mu_x^2}}{\bar{\sigma_x^2}} - 2\mu_0 \frac{\bar{\mu_x}}{\bar{\sigma_x^2}} - 1 \quad (20)$$

whereby  $\sigma_\mu^2 = \int_{\mathbb{R}^n} \pi(x) (\mu(x) - \mu_0)^2 dx$ . The factors in non-gaussianity (in the upper-bound sense) can be seen in the two independent cases as follows:

1. let  $\sigma_x = \text{const} \Rightarrow J(p) \leq 2\sigma_\mu^2$
2. let  $\mu_x = \text{const} \Rightarrow J(p) \leq \frac{\bar{\sigma_x^2}}{\sigma_x^2} \frac{1}{\bar{\sigma_x^2}} - 1$

It can be easily verified that  $\mu_x^* = \text{const}, \sigma_x^* = \text{const}$  are unique minimizers of (20) leading to  $J(p) = 0$  and hence proving Proposition 1.