On the Information Capacity of Hinge Functions

Gholamreza Alirezaei and Rudolf Mathar
Institute for Theoretical Information Technology
RWTH Aachen University, D-52056 Aachen, Germany
{alirezaei, mathar}@ti.rwth-aachen.de

***Abstract*—So called hinge functions play an important role in many applications, e.g., deep learning, support vector machines, regression, classification and others. A thorough theory, which explains why some of these applications are so successful for their respective purpose, is still missing. This paper aims at filling a knowledge gap by answering the question of how much information can be conveyed across a neural node, e.g., which uses the hinge loss function on weighted agglomerated information from precedent nodes. We hope that insight from artificial neural networks may also have implications for understanding biological information processing. As key results in this paper, an elegant representation of mutual information is derived and, furthermore, some structural properties of a channel, that consists of a certain input signal which is overlaid by additive noise and is filtered by a hinge function, are investigated. Determining the capacity of this channel in an explicit form, although an important fundamental problem, seems to be extremely difficult and unsolved as of today. Thus, necessary and sufficient conditions for an optimal input signal along with upper bounds on the capacity are deduced. Furthermore, we conjecture that exponentially distributed input signals are asymptotically capacity-achieving in the high SNR regime.**

I. Introduction

The nonnegative valued function $Q(z) = \max\{z - c, 0\} = (z - c)^+,\ z \in \mathbb{R},\ c$ some constant, see Fig. 1, is often called *rectifier* or *hinge function*. An early article on the mathematical structure in higher dimensions, convergence properties, and effectiveness for neural networks is [1] which extends the work [2], that investigates approximation bounds for superpositions of sigmoidal functions.

The application of hinge functions in deep learning neural networks is very successful, and is even superior to smooth loss functions like logistic sigmoid, hyperbolic and inverse tangent function [3], cf. [4]. The main advantage is achieved in unsupervised learning. One of the reasons seems to be that rectifying neurons induce sparsity by producing true zeros in an obvious way. This seems to be suitable for modeling corresponding sparse representations in biological neural networks. A thorough theory of understanding the efficiency of rectifier neural networks is however missing.

The present paper aims at contributing to this problem by answering the question of how much information can be conveyed across a neural node when using the hinge loss function on weighted agglomerated information from precedent nodes. We hope that from this insight the functionality of deep learning networks can be better understood, and that such networks can be conceived as a reasonable model for biological information processing.

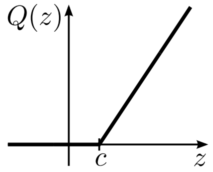

Fig. 1: The hinge function.

The contributions of this paper are as follows. We first derive a compact representation of the mutual information between a certain input variable X and the output Y, where X is subject to additive noise and is then sent through the hinge function. We first investigate the mathematical structure of mutual information. Determining the capacity of this channel requires maximizing mutual information over all input distributions, which is an extremely and, as of today, unsolved problem. Instead, necessary and sufficient conditions for an optimal input signal along with upper bounds on the unknown capacity are derived in the present paper.

II. Channel Model

We assume a time-discrete memoryless channel. Some real input X with cumulative distribution function (CDF) $F(x)$ is subject to additive noise W with density function $f_W(w)$ and corresponding CDF $F_W(w)$, not necessarily Gaussian. X and W are assumed to be stochastically independent. The noisy signal $Z = X + W$ is then filtered by the hinge function Q to generate output Y. The system model is depicted in Fig. 2 and reads as

$$Y = Q(X + W) = \max\{X + W, 0\} = (X + W)^+. \quad (1)$$

In order to understand how much information can be conveyed from the input variable X to the output Y, we will investigate the mathematical structure of the mutual information $I_{X;Y}$. Maximizing $I_{X;Y}$ over all input distributions F yields the capacity $C_{X;Y}$ of this channel. We hope that by training artificial neural networks the input distributions at intermediate nodes come close to the capacity-achieving one. This would help to understand why deep learning networks are this powerful.

III. Entropy of Mixture Distributions

The entropy of a random variable Y with density g with respect to some dominating σ-finite measure μ on the real line is defined as

$$H(Y) = -\int g(y) \log g(y)\, d\mu(y), \quad (2)$$

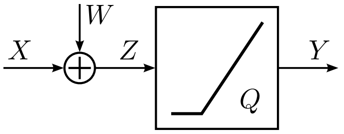

Fig. 2: The system model: some real input X is subject to additive noise W and is then filtered by the hinge function.

see [5]. Channel model (1) leads to a mixture of a one-point distribution at 0 and a continuous distribution, for which the corresponding entropy can be determined by following the work [6].

We introduce jointly independent random variables U, V and B. U is assumed to be absolutely-continuous with Lebesgue density $f_c(u)$, and V to be discrete with countably many support points v_i, probabilities p_i and discrete density $f_d(v) = p_i$, whenever $v = v_i$ and $f_d(v) = 0$, otherwise. B is a Bernoulli distributed random variable with $P(B = 1) = \alpha$ and $P(B = 0) = 1 - \alpha$. In this case

$$Y = BU + (1 - B)V$$

has density

$$g(y) = \alpha f_c(y) + (1 - \alpha) f_d(y)$$

with respect to the measure $\mu = \lambda + \chi$, the sum of the Lebesgue measure λ and the counting measure χ on the support points of V.

From (2) it follows that

$$\begin{aligned} H(Y) &= -\alpha \int f_c(y) \log f_c(y)\, dy - \alpha \log \alpha \\ &\quad - (1-\alpha) \sum_i p_i \log p_i - (1-\alpha)\log(1-\alpha) \\ &= H(B) + \alpha H(U) + (1-\alpha) H(V). \end{aligned} \tag{3}$$

The entropy of Y is hence a convex combination of the entropy of U and V with factor α, plus the additional uncertainty introduced by switching the random variable B, cf. [7, Thm. 3].

IV. Mutual Information of the Hinge Channel

Two common abbreviations will be useful in the following. First, self-information is defined as

$$\rho(q) = -q \log q\,, \qquad q \geq 0 \tag{4}$$

and, second, the binary entropy function as

$$\begin{aligned} h(p) &= -p \log p - (1-p)\log(1-p) \\ &= \rho(p) + \rho(1-p)\,, \qquad 0 \leq p \leq 1\,, \end{aligned} \tag{5}$$

where the base of the logarithm is kept general. It is well-known that both $\rho(q)$ and $h(p)$ are strictly concave functions of their arguments q and p, respectively.

The conditional distribution function of $Q(X + W)$ given $X = x$ is easily determined as

$$P\big(Q(X+W) \leq y \mid X = x\big) = \begin{cases} 0\,, & \text{if } y < 0\,, \\ F_W(y - x)\,, & \text{if } y \geq 0\,. \end{cases}$$

Assuming that the noise distribution is absolutely-continuous with density f_W, in which case $X + W$ is also absolutely continuous, the density of $Q(X + W)$ given $X = x$ is a mixture of a discrete single-point distribution at 0 and a continuous one with density $\frac{1}{\alpha} f_W(y - x)$, $y \geq 0$, with $\alpha = \int_0^\infty f_W(w - x)dw = \int_{-x}^\infty f_W(t)dt$.

Now let

$$r(f, x) = \int_{-x}^{\infty} f(t)dt$$

and

$$\ell(f, x) = \int_{-\infty}^{-x} f(t)dt\,.$$

Then by formula (3) we obtain

$$\begin{aligned} &H(Y \mid X = x) \\ &\quad = h\big(r(f_W, x)\big) \\ &\qquad - r(f_W, x) \int_0^\infty \frac{f_W(z-x)}{r(f_W,x)} \log \frac{f_W(z-x)}{r(f_W,x)} dz \\ &\quad = h\big(r(f_W, x)\big) \\ &\qquad - \int_{-x}^\infty f_W(z) \log f_W(z) dz + r(f_W, x) \log r(f_W, x) \\ &\quad = -\ell(f_W, x) \log \ell(f_W, x) - \int_{-x}^\infty f_W(z) \log f_W(z) dz. \end{aligned}$$

The conditional entropy of Y given X is obtained by integrating $H(Y \mid X = x)$ w.r.t. the input distribution F as

$$H(Y \mid X) = \int H(Y \mid X = x) dF(x)\,. \tag{6}$$

In a similar manner, $H(Y)$ is derived as

$$\begin{aligned} H(Y) &= -\ell(f_Z, 0) \log \ell(f_Z, 0) \\ &\qquad - \int_0^\infty f_Z(z) \log f_Z(z) dz \end{aligned} \tag{7}$$

with $f_Z(z) = \int f_W(z - x)\, dF(x)$.

From $I_{X;Y} = H(Y) - H(Y \mid X)$ with (4), (6) and (7), the following compact representation of the mutual information is achieved:

$$\begin{aligned} I_{X;Y} &= \rho\Big(\int \int_{-\infty}^0 f_W(u-x) du\, dF(x)\Big) \\ &\quad - \int \rho\Big(\int_{-\infty}^0 f_W(u-x) du\Big) dF(x) \\ &\quad + \int_0^\infty \rho\Big(\int f_W(u-x) dF(x)\Big) du \\ &\quad - \int_0^\infty \int \rho\big(f_W(u-x)\big) dF(x)\, du\,. \end{aligned} \tag{8}$$

As is expected, equation (8) is in line with the corresponding result for censored channels, see [8].

The channel capacity $C_{X;Y}$ of the hinge channel under constraints is obtained by maximizing the corresponding mutual information (8) over certain classes of input distributions F. We emphasize the dependence on F by the notation $I_{X;Y}(F)$ hereinafter.

V. Capacity of the Hinge Channel

Determining the channel capacity of the hinge channel subject to moment constraints can be described by the optimization problem

$$\begin{aligned} C_{X;Y} &= \max_F I_{X;Y}(F) \\ \text{s.t.} \int x^i dF(x) &= m_i\,,\ \text{for certain } i \in \mathbb{N}_0\,, \\ \int x^j dF(x) &\leq m_j\,,\ \text{for certain } j \in \mathbb{N}\,, \end{aligned} \tag{9}$$

for given values m_i and m_j. For example, $m_0 = 1$ ensures that a solution $F^\star$ satisfies $\lim_{x\to\infty} F^\star(x) = 1$ and is hence a proper distribution function. Choosing $|m_1| < \infty$ and $0 < m_2 < \infty$ ensures that input X has prescribed mean and power, respectively. Without constraints of the above type the channel capacity may be infinite, depending on the underlying noise distribution. The reason behind this fact is the ramp on the right side of the hinge function by which error-free communication for any data-rate is possible by increasing the power of the input signal to arbitrary high values. This behavior is similar to the common AWGN channel.

The maximization problem (9) is fortunately a convex optimization program, since the self-information ρ is strictly concave. Hence, we use the Karush-Kuhn-Tucker (KKT) conditions to characterize the global optimum. Its Lagrangian reads as

$$\begin{aligned} L(F) = -I_{X;Y}(F) &+ \sum_i \lambda_i \Big(-m_i + \int x^i dF(x) \Big) \\ &+ \sum_j \mu_j \Big(-m_j + \int x^j dF(x) \Big)\,, \end{aligned} \tag{10}$$

where $\lambda_i \in \mathbb{R}$ and $\mu_j \geq 0$ are the Lagrangian multipliers. The functional derivative of the Lagrangian (10) vanishes at any stationary solution, which yields a necessary condition for an optimum solution $F^\star$. Since the above optimization problem belongs to the class of convex programs, the necessary condition is also a sufficient condition. For all points $\tilde{x}$ of increase of $F^\star$, this condition reads as

$$\begin{aligned} 0 = \log\Big(\mathrm{e}\!\int\!\int_{-\infty}^0 f_W(u-x)du\, dF^\star(x)\Big) &\int_{-\infty}^0 f_W(u-\tilde{x})du \\ + \rho\Big(\int_{-\infty}^0 f_W(u-\tilde{x})du\Big) &+ \int_0^\infty \rho\big(f_W(u-\tilde{x})\big)\, du \\ + \int_0^\infty \log\Big(\mathrm{e}\!\int f_W(u-x)dF^\star(x)\Big) &f_W(u-\tilde{x})du \\ + \sum_i \lambda_i^\star\, \tilde{x}^i + \sum_j \mu_j^\star\, \tilde{x}^j\,, & \end{aligned} \tag{11}$$

where the constant e denotes Euler's number. The complicated structure of the condition (11) obviously prevents determining an explicit solution $F^\star$ as a proper candidate for the optimal solution. However, equality (11) is important for numerical methods to solve (9).

VI. Bounds on the Capacity

As equation (11) demonstrates, it is extremely hard to determine the capacity of the hinge channel by maximizing (8) over certain distribution functions F. It is the purpose of this section to determine upper bounds on the unknown capacity. This is done by first constructing a lower bound on (6) and then an upper bound on (7).

By the following chain of inequalities we derive a lower bound on $H(Y \mid X)$.

$$\begin{aligned} &H(Y \mid X) \\ &\quad= \int \Big[\rho\Big(\int_{-\infty}^{-x} f_W(u)du\Big) + \int_{-x}^{\infty} \rho\big(f_W(u)\big)\, du\Big] dF(x) \\ &\quad\geq \int \int_{-x}^{\infty} \rho\big(f_W(u)\big)\, du\, dF(x) \\ &\quad= \int \int_0^{\infty} \rho\big(f_W(v-x)\big)\, dv\, dF(x) \\ &\quad= -\int \int_0^{\infty} f_W(v-x) \log\big(f_W(v-x)\big)\, dv\, dF(x) \\ &\quad\geq -\int \int_0^{\infty} f_W(u-x)du \log\left[\frac{\int_0^\infty f_W^2(v-x)dv}{\int_0^\infty f_W(v-x)dv}\right] dF(x) \\ &\quad\geq -\rho\big(\mathrm{e}^{-1}\big) \int \int_0^{\infty} f_W^2(v-x)\, dv\, dF(x) \\ &\quad\geq -\rho\big(\mathrm{e}^{-1}\big) \int \int_{-\infty}^{\infty} f_W^2(v-x)\, dv\, dF(x) \\ &\quad= -\rho\big(\mathrm{e}^{-1}\big) \int_{-\infty}^{\infty} f_W^2(v)\, dv\,. \end{aligned} \tag{12}$$

The first inequality is due to the positivity of the first term, the second inequality follows from Jensen's inequality, see e.g. [9]. The third inequality is due to the fact that $p \log \frac{q}{p} \leq q\rho(\mathrm{e}^{-1})$ for any positive numbers p and q. Equality in the above chain is never attained.

Finally, an upper bound on $H(Y)$ follows from

$$\begin{aligned} H(Y) &= \rho\Big(\int_{-\infty}^0 f_Z(z)dz\Big) + \int_0^\infty \rho\big(f_Z(z)\big)dz \\ &= \rho\Big(\int_{-\infty}^0 f_Z(z)dz\Big) + \rho\Big(\int_0^\infty f_Z(z)dz\Big) \\ &\quad + \int_0^\infty f_Z(\tilde{z})d\tilde{z} \int_0^\infty \rho\Big(\frac{f_Z(z)}{\int_0^\infty f_Z(\tilde{z})d\tilde{z}}\Big)dz \\ &= \rho\Big(\int_{-\infty}^0 f_Z(z)dz\Big) + \rho\Big(\int_0^\infty f_Z(z)dz\Big) \\ &\quad + \int_0^\infty f_Z(\tilde{z})d\tilde{z}\ H(\tilde{Z}) \\ &\leq \rho\Big(\int_{-\infty}^0 f_Z(z)dz\Big) + \rho\Big(\int_0^\infty f_Z(z)dz\Big) \\ &\quad + \int_0^\infty f_Z(\tilde{z})d\tilde{z}\ H(S) \\ &= \rho(\alpha_1) + \rho(\alpha_2) + \alpha_2\, H(S), \end{aligned} \tag{13}$$

where $\alpha_1, \alpha_2 \geq 0$ and $\alpha_1 + \alpha_2 = 1$.

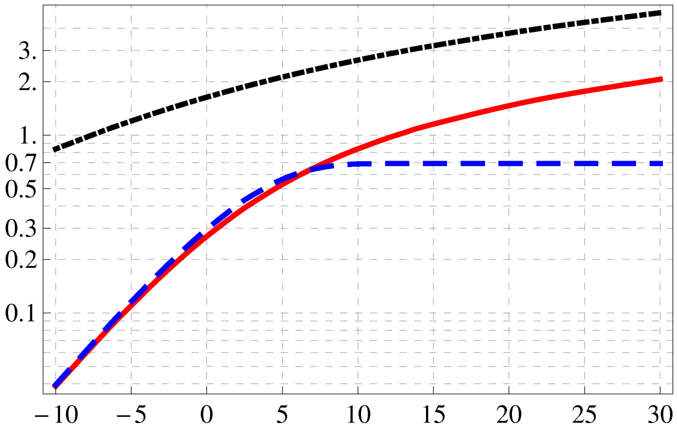

Fig. 3: Mutual information (8) and the upper bound (16) in [nats] over the signal-to-noise ratio in [dB] for a hinge channel subject to additive standard Gaussian noise are visualized. For mutual information the input signal is either Gaussian, depicted by the solid red line, or binary distributed, shown by the dashed blue line. The dotted black line represents the upper bound on the capacity for any input signal with limited variance.

The entropy $H(\tilde{Z})$ corresponds to a random variable $\tilde{Z}$ with density equal to $\frac{f_Z(z)}{\int_0^\infty f_Z(\tilde{z})d\tilde{z}}$, $z \geq 0$, while the entropy $H(S)$ corresponds to a random variable S, that achieves the maximum entropy subject to the given constraints.

For example, when neglecting positivity and thus extending the set of admissible distributions, S would be Gaussian if the constraint refers to upper bounding the variance. S would have an exponential distribution if the constraint fixes the mean. Note that $H(S)$ is a function of the values m_i and m_j as well as the moments of the noise density f_W. Thus, the entropy $H(S)$ is independent of the input distribution F. Equality in (13) is attained if S has the same density as $\tilde{Z}$.

A global upper bound for $H(Y)$ is found by maximizing the right hand side over α_1 and α_2. Since the objective function $\rho(\alpha_1) + \rho(\alpha_2) + \alpha_2\, H(S)$ is concave, we use simple convex optimization methods to obtain the optimum $\alpha_1^\star = \left(1 + \mathrm{e}^{H_\mathrm{e}(S)}\right)^{-1}$, where $H_\mathrm{e}(S)$ is the entropy of S based on the natural logarithm. Hence, by aid of the binary entropy (5) we deduce

$$H(Y) \leq h\Big(\frac{1}{1 + \mathrm{e}^{-H_\mathrm{e}(S)}}\Big) + \frac{H(S)}{1 + \mathrm{e}^{-H_\mathrm{e}(S)}}. \qquad (14)$$

Finally, combining (12) and (14) yields the upper bound

$$C_{X;Y} \leq h\big(\tfrac{1}{1+\mathrm{e}^{-H_\mathrm{e}(S)}}\big) + \tfrac{H(S)}{1+\mathrm{e}^{-H_\mathrm{e}(S)}} + \rho\big(\mathrm{e}^{-1}\big)\int_{-\infty}^{\infty} f_W^2(v)\, dv. \qquad (15)$$

The right hand term of the upper bound corresponds to the energy of the noise density. Inequality (15) shows that the capacity of the hinge channel is finite once the maximum entropy $H(S)$ is finite due to certain constraints, and the noise density is square-integrable. This is an interesting result in itself. In order to tighten the upper bound (15), we consider the following technique.

A further upper bound on the channel capacity can be derived by using the *data processing inequality*. Following this approach, the mutual information $I_{X;Y}$ for the Markov chain $X \to Z \to Y$ is bounded by $\min\{I_{X;Z}, I_{Z;Y}\}$. Hence, the channel capacity $C_{X;Y}$ is bounded by $\min\{C_{X;Z}, C_{Z;Y}\}$. For this upper bound the mutual information $I_{Z;Y} = H(Y) - H(Y|Z)$ is needed, which can similarly be deduced as shown for $I_{X;Y}$. It follows that $H(Y|Z) = 0$, since the conditional probability $P(Y|Z = z)$ is a single-point distribution, while $H(Y)$ is equal to (7). With these results, an upper bound on the mutual information of the hinge channel is given by $I_{X;Y} \leq I_{Z;Y} = H(Y)$. Using the bound in (14), we obtain the upper bound

$$C_{X;Y} \leq h\Big(\frac{1}{1 + \mathrm{e}^{-H_\mathrm{e}(S)}}\Big) + \frac{H(S)}{1 + \mathrm{e}^{-H_\mathrm{e}(S)}}\,, \qquad (16)$$

which is obviously tighter than the upper bound in (15).

VII. Numerical Experiments

In this section, we consider two scenarios for numerical experiments. In both scenarios additive standard Gaussian noise W applies. Mutual information (8) and the upper bound (16) are compared in both scenarios for different input signals and constraints.

In Fig. 3 the results for the first scenario are shown. In this scenario we compare two input signals where the first one is Gaussian distributed while the second one has a symmetric binary distribution. Both input signals are zero mean and have the same variance $E\{X^2\}$. As is shown by the solid red line, the Gaussian input signal achieves a better throughput for high signal-to-noise ratios (SNR) than the binary input, depicted by the dashed blue line. The reason behind this effect is that the capacity for binary signaling is achieved at approximately 7dB and a further increase is only possible by including additional signaling points. The upper bound, depicted by the dotted black line, corresponds to a Gaussian random variable S with a variance equal to $E\{X^2\}$. The difference between the upper bound and the mutual information of the Gaussian input signal shows an interesting effect. The difference is not necessarily decreasing for high SNR values, since the left-half of the input power is always censored by the hinge function and hampers the increase of the corresponding mutual information.

In Fig. 4 the results of the second scenario are visualized. Here, an exponentially distributed input signal is compared to a uniformly distributed binary input signal with one signaling point at zero. Both input signals have the same positive mean $E\{X\} > 0$. For the sake of comparability with the results above, the curves are depicted over the ratio between the squared expected value of the input signal and the noise variance. As is shown by the solid red line, the exponentially distributed input signal achieves a better throughput for high SNRs than the binary input, depicted by the dashed blue line. The reason behind this effect is the same as described above.

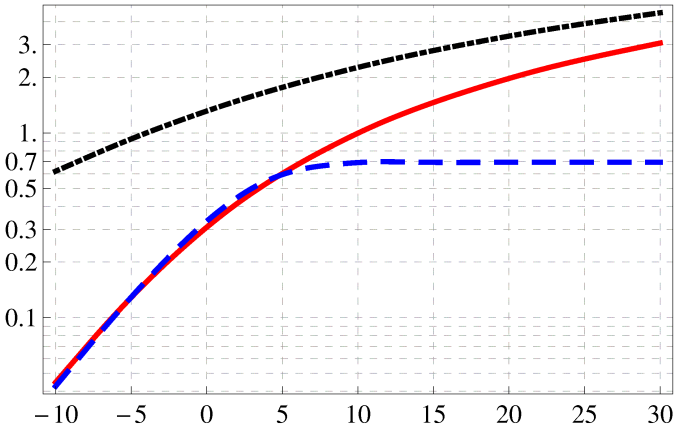

Fig. 4: Mutual information (8) and the upper bound (16) in [nats] over the signal-to-noise ratio in [dB] for a hinge channel subject to additive standard Gaussian noise are visualized. For mutual information the input signal is either exponentially, depicted by the solid red line, or binary distributed, shown by the dashed blue line. The dotted black line represents the upper bound on the capacity for any input signal with limited mean.

But the exponentially distributed input signal achieves also better results for very low SNRs, which can be explained by the bad placement of both signaling points of the binary input signal. The upper bound, depicted by the dotted black line, refers to an exponentially distributed random variable S with expected value equal to $E\{X\}$. The difference between the upper bound and the mutual information of the exponentially distributed signal seems to converge to zero as the SNR tends to infinity.

Comparing the results in both figures reveals that input signals with a positive support achieve higher mutual information than input signals that also have support on the negative real line. Furthermore, the upper bound on the capacity seems to be tighter for input signals with positive support as can be expected from the equality condition in (13).

In summary, we conjecture that an input signal with at most one single mass-point on the negative real line should be preferable to increase the mutual information of the hinge channel. Moreover, we conjecture that exponentially distributed input signals for a wide range of noise distributions are asymptotically capacity-achieving as SNR tends to infinity.

VIII. Conclusion

The hinge channel is an important member of communication channels which has wide applications mainly for artificial neural networks. Information theoretic investigations of this channel are extremely hard, as shown in the present paper. We have achieved a concise description of the mutual information of the hinge channel and provided necessary and sufficient conditions for the capacity-achieving input distribution. Since an explicit form of the information capacity seems to be out of reach, we have developed upper bounds on the unknown capacity. The tightness of the upper bounds under different constraints is demonstrated by numerical investigations. We conjecture that the exponential distribution is asymptotically capacity-achieving as the SNR tends to infinity.

References

[1] L. Breiman, “Hinging hyperplanes for regression, classification, and function approximation,” *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 999–1013, May 1993.

[2] A. R. Barron, “Universal approximation bounds for superpositions of a sigmoidal function,” *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 930–945, May 1993.

[3] G. Alirezaei, “A sharp double inequality for the inverse tangent function,” 2013. [Online]. Available: http://arxiv.org/abs/1307.4983

[4] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *14th International Conference on Artificial Intelligence and Statistics*, Ford Lauderdale, April 2011, pp. 315–323.

[5] M. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, California, USA: Holden-Day, 1964.

[6] D. N. Politis, “Maximum entropy modelling of mixture distributions,” *Kybernetes*, vol. 23, no. 1, pp. 49–54, 1994.

[7] A. Rényi, “On the dimension and entropy of probability distributions,” *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 10, no. 1–2, pp. 193–215, 1959.

[8] G. Alirezaei and R. Mathar, “An upper bound on the capacity of censored channels,” in *The 9th International Conference on Signal Processing and Communication Systems (ICSPCS’15)*, Cairns, Australia, Dec. 2015.

[9] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, ser. Cambridge Mathematical Library. Cambridge University Press, 1952.