Quantifying Information Gain in Infinite Space

Teerapat Jenrungrot, Jonathan Cruz
Department of Computer Science
Harvey Mudd College
Claremont, CA 91711
{mjenrungrot,jcruz}@hmc.edu

Abstract

In many machine learning algorithms, a natural way of measuring the reduction of uncertainty is either taking the proportion of the target space to the total space, and taking the negative log base-2 of that ratio, or if the space is under a probability distribution, finding the entropy reduction of moving to the subset versus the initial uncertainty of the full set. However, the problem arises when our space is unbounded and uncountable, such as the real line. We propose an analysis framework by using squashing functions to map infinite space to a finite interval. We demonstrate that it is possible to bijectively map any real number to the interval and show numerical results for computing information gain based on the proposed approach. Finally, we illustrate that the choice of squashing functions indicates a learning algorithm's prior distribution of how parameters lie in the target space.

1 Introduction

Quantifying information gain in infinite space matters because physicist Luke Barnes claims the universe is fine-tuned: "In the set of possible physics, the subset that permit the evolution of life is very small," (4). In other words, if there is some way of assigning measure to infinite spaces to access fine-tuning in physics, then a similar approach can be made to parameters in machine learning. Alexander Pruss, a professor of philosophy at Baylor University, mentions there are issues in an infinite parameter space and how we can no longer place uniform probability mass on the space (3).

Many learning algorithms, such as ID3 algorithm for decision tree (1), are based on computing information gain of some attributes. In information theory, information gain is defined based on the change in information entropy. Since the information entropy is unbounded in general, this problem makes it impossible to quantify the information gain from infinite space to finite space. As a consequence, many learning algorithms require on the assumption that users provide good initialization in finite space, so the algorithms are able to generalize on provided datasets.

In this paper, we propose a simple framework for quantifying information gain from infinite space to finite space by using squashing functions. According to the Schröder-Bernstein theorem, the set of real numbers $\mathbb{R}$ and any non-degenerate closed or open interval in $\mathbb{R}$ (such as the unit interval $[0, 1]$) are equipotent or having the same cardinality (2). Based on this idea, we propose to map a value in infinite space to a finite interval in $\mathbb{R}$ and compute the information gain based on the mapped value.

The rest of this paper outlines as follows. Section 2 describes our proposed frameworks. Section 3 illustrates results of using our framework on numerically evaluating information gain from infinite space to certain finite spaces. Finally, Section 4 concludes our work and suggests some possible future direction.

Preprint. Final Project for HMC CSCI181P, Fall 2018.

2 Proposed Framework

In this section, we describe the proposed framework for quantifying information gain when reducing uncertainty from infinite space to finite space.

Definition 1. A function $f(x)$ is a *squashing function* if and only if the function satisfies the following constraints:

- The function $f(x)$ is monotonically non-decreasing.
- The domain of $f(x)$ is the set of real number $(-\infty, \infty)$.
- The range of $f(x)$ is $(0, 1)$, $[0, 1)$, $(0, 1]$, or $[0, 1]$.

Consider a task of quantifying the reduction of uncertainty from the original space S_o to the target space S_t, where $S_t \subseteq S_o$. In many learning algorithms, the algorithms' parameters are unbounded. In such case, we can safely assume S_o to be a set of real numbers $\mathbb{R}$. Without loss of generality, we can impose a prior distribution regarding how parameters lie in the target space. Suppose X is a random variable indicating the parameters' distribution. From Definition 1, its cumulative distribution function $F_X(x)$ is, therefore, a valid squashing function. In this paper, we demonstrate the framework on one-dimensional space real line. The same technique can be easily adapted to higher dimensional space.

Equation 1 shows the measure of uncertainty reduction from the original space $S_o = \mathbb{R}$ to the target finite space S_t. Suppose that the prior distribution of parameters in the target space is indicated by the random variable X.

$$IG = \log_2 \left(\frac{\lim_{x \to +\infty} F_X(x) - \lim_{x \to -\infty} F_X(x)}{F_X(\sup S_t) - F_X(\inf S_t)} \right) = \log_2 \left(\frac{1}{F_X(\sup S_t) - F_X(\inf S_t)} \right). \quad (1)$$

Instead, if the original space S_o is already a finite space, we can modify Equation 1 to

$$IG = \log_2 \left(\frac{F_X(\sup S_o) - F_X(\inf S_o)}{F_X(\sup S_t) - F_X(\inf S_t)} \right). \quad (2)$$

According to our formulation, $F_X(x)$ can be replaced with any squashing function, $f_{\text{squash}}(x)$. Due to Definition 1, the range of the squashing functions must be from 0 to 1, so it is possible to determine the prior distribution of how parameters lie in the target space for any squashing functions. Equation 3 shows the probability density function for an arbitrary squashing function $f_{\text{squash}}(x)$:

$$f_X(x) = \frac{d}{dx} \int_{-\infty}^{x} f_{\text{squash}}(t)\, dt. \quad (3)$$

3 Results

In this section, we demonstrate a numerical example of quantifying information gain using our proposed method with different squashing functions. Note that in Section 2 we have shown that if the distribution is known, we can construct a squashing function. We tested the proposed method on two most commonly used distributions: uniform distribution and normal distribution and three other squashing functions mostly used in artificial neural network problems: Sigmoid, TanH, and ArcTan. Specifically, we consider the task of quantifying uncertainty reduction from infinite space $S_o = \mathbb{R}$ to a target finite space S_t.

Uniform distribution If the prior distribution of target values is the uniform distribution $X \sim \mathcal{U}[a, b]$, its cumulative distribution function and its squashing function are given by

$$F_X(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x < b \\ 1 & \text{otherwise} \end{cases} \quad (4)$$

Normal distribution If the prior distribution of target values is the normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$, its cumulative distribution function and its squashing function are given by

$$F_X(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \tag{5}$$

where $\text{erf}(x) = \frac{1}{\sqrt{\pi}}\int_{-x}^{x} e^{-t^2}\, dt$ is an error function.

Other squashing functions In many neural networks tasks, the most common activation functions are Sigmoid, TanH, and ArcTan. Though there are many other novel activation functions such as Rectified Linear Unit (ReLU), we purposely choose only activation functions that have finite range. In addition, we re-scaled all the activation functions so that the range is in the interval $(0, 1)$. Table 1 shows the equation of squashing functions and its probability density function of the prior distribution.

Activation Name	Equation	Prior Distribution
Sigmoid	$\frac{1}{1+e^{-x}}$	$\frac{e^{-x}}{(1+e^{-x})^2}$
TanH	$\frac{1}{2}\tanh(x) + \frac{1}{2}$	$\frac{1}{2}\left(1 - \tanh^2(x)\right)$
ArcTan	$\frac{1}{\pi}\arctan(x) + \frac{1}{2}$	$\frac{1}{\pi(1+x^2)}$

Table 1: Table of common activation functions and their corresponding prior distributions

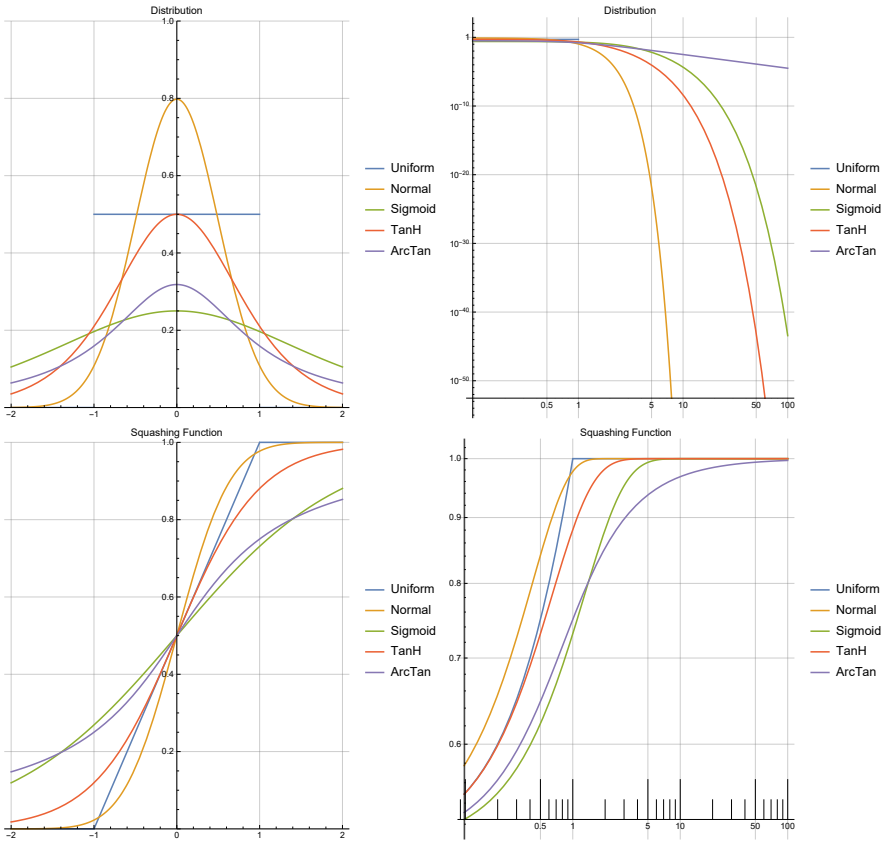

Figure 1: (top-left) the probability distribution function of each squashing function; (bottom-left) the corresponding squashing function from $x = -2$ to $x = 2$; (top-right) the probability distribution function of each squashing function in log-scale from $x = 0$ to $x = 100$; (bottom-right) the corresponding squashing functions in log-scale.

From Figure 1 and Table 2, consider the squashing function for the uniform distribution $\mathcal{U}[-1, 1]$, the information gain from $(-\infty, \infty)$ to $(-10, 10)$ or $(-1, 1)$ is 0 because we impose a prior assumption

Squashing Function	Information Gain from $(-\infty, \infty)$ to						
	$(-10, 10)$	$(-1, 1)$	$(0, 1)$	$(0, 10)$	$(0, 100)$	$(1, 100)$	$(10, 100)$
$\mathcal{U}[-1, 1]$	0.0000	0.0000	1.0000	1.0000	1.0000	∞	∞
$\mathcal{N}(0, 0.5)$	0.0000	0.0672	1.0672	1.0000	1.0000	5.4580	∞
Sigmoid	0.0001	1.1137	2.1137	1.0001	1.0000	1.8946	14.427
TanH	0.0000	0.3929	1.3929	1.0000	1.0000	3.0685	28.854
ArcTan	0.0946	1.0000	2.0000	1.0946	1.0092	2.0185	5.1307

Table 2: Information Gain from infinite space to finite spaces for different squashing functions.

that the target value must be in the interval $(-1, 1)$. On the other hand, the information gain from $(-\infty, \infty)$ to $(1, 100)$ or $(10, 100)$ is infinite because we do not have prior assumption that the target values will be outside the interval $(-1, 1)$.

Since the majority of probability mass of all squashing functions is around $x = 0$, the information gain from $(-\infty, \infty)$ to intervals containing $x = 0$ is relatively small. In contrast, the information gains from $(-\infty, \infty)$ to $(1, 100)$ and $(10, 100)$ are relatively large because the probability masses are small in these intervals.

Comparing the information gains from $(-\infty, \infty)$ to intervals $(0, 10)$ and $(0, 100)$, we do not observe significantly different information gains. Because the probability mass of prior assumption of target values lying in the interval $(10, 100)$ is very small, the difference between the information gain for $(0, 10)$ and $(0, 100)$ is therefore relatively small.

Comparing the squashing function for the normal distribution $\mathcal{N}(0, 0.5)$ with the three common activation functions, the probability mass is very low in the interval $(1, 100)$. Hence, the information gain from infinite space to this interval is bigger than the information gain based on other activation functions.

4 Conclusion

In this paper, we have demonstrated a theoretical framework for quantifying information gain in infinite space using squashing functions. Furthermore, we have shown that the choice of squashing functions can be used as a proxy of how we specify prior assumption regarding the values on the target space. If a distribution is known, the corresponding squashing function is defined based on its cumulative distribution function. We demonstrate the proposed framework for computing information gain from infinite space to different non-degenerate, finite spaces. Information gain based on different squashing functions give different values based on the prior assumptions regarding how the target values lying in the target space.

With many machine learning algorithms developed based on information theory, we believe that the results of this work can lead to more insights on how the machine learning algorithms work when dealing with parameters in infinite space. For instance, a decision tree learning algorithm ID3 constructs a decision tree based on choosing parameters that maximize the information gain. For future works, we believe that modifying the ID3 algorithm to support our proposed framework and then conducting the formal analysis on choice of different squashing functions may give more insights on how different activation functions impact learning algorithms.

References

[1] Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81–106.

[2] Cantor, Georg (1874), "Ueber eine Eigenschaft des Inbegriffes aller reellen algebraischen Zahlen", Journal für die Reine und Angewandte Mathematik, 77: 258–262, doi:10.1515/crll.1874.77.258.

[3] Pruss, Alexander R. “Fine-and Coarse-Tuning, Normalizability, and Probabilistic Reasoning.” (2005).

[4] Barnes, Luke A. “The fine-tuning of the universe for intelligent life.” Publications of the Astronomical Society of Australia 29.4 (2012): 529-564.