CHENNAI MATHEMATICAL INSTITUTE

MASTERS' THESIS

Capacity Achieving Distribution of Amplitude Constrained AWGN Channel

---

Author:
Ativ Joshi

Supervisor:
Dr. Navin Kashyap

A thesis submitted in fulfillment of the requirements for the degree of MSc in Computer Science

June 9, 2021

CHENNAI MATHEMATICAL INSTITUTE

Abstract

Computer Science

MSc

Capacity Achieving Distribution of Amplitude Constrained AWGN Channel

by Ativ Joshi

The capacity of a communication channel is the maximum "amount of information" that can be obtained about the transmitted (input) signal by observing the received (output) signal. Formally, channel capacity is defined as the maximum (taken over the input) of mutual information between the input and output random variables. We focus on the AWGN(Additive White Gaussian Noise) model of the communication channel with constraint on the maximum amplitude. We survey the nature of the capacity achieving input distribution of the channel, its behavior as the amplitude constraint is relaxed, and certain bounds on its support.

Acknowledgements

Firstly, I would like to thank my supervisor Dr. Navin Kashyap (IISc, Bangalore) for their invaluable advice, support and patience. I am also grateful to Dr. Manjunath Krishnapur for their helpful discussions. I would like to thank Dr. Partha Mukhopadhyay, Dr. Bireswar Das and Dr. Ratnik Gandhi who have helped me countless number of times during my academic journey so far.

Last but not the least, I am grateful to my parents, family and a really long list of friends who have supported me throughout my life.

Contents

to G-13, CMI

Chapter 1

Introduction

1.1 Definitions and Notation

A memoryless communication channel is a channel in which the output signal depends only on the current value of the input signal. One of the most common memoryless channel model is the AWGN model. AWGN (Additive White Gaussian Noise) is characterized by three features:

1. The noise is modeled by adding it to the input signal
2. The power of noise is uniform for all frequencies and
3. The distribution of noise is Gaussian in time domain.

Such a channel can be modeled as

$$Y = X + V$$

where Y is the output of the channel, X is the input of the channel and V is the additive standard Gaussian noise independent of the input. Here, X,Y and V are continuous random variables. We put an extra constraint on the amplitude of the input signal,

$$Pr[|X| > A] = 0.$$

This is called an 'amplitude constraint' or a 'peak-power constraint'.

Entropy (also called Shannon Entropy) of a (discrete) random variable X encodes the average "amount of information" that is inherently present in the variable. Entropy is defined as

$$h(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

where $P(X)$ is the PMF of random variable X. This concept can be extended to the continuous random variables by analogously defining differential entropy as

$$H(X) = -\int_{\mathcal{X}} f(x) \log f(x) \, dx$$

where $f(X)$ is the probability density function of X.

Mutual information of two random variables characterizes the amount of information that can be obtained about one random variable by observing the other. For AWGN, mutual information is defined as

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(V).$$

The capacity of a communication channel is the maximum mutual information between the input and output random variable. Here the maximization is taken over the input variable (since output is also a function of input). For a given amplitude constraint A, the capacity of the channel is given by

$$C(A) = \max_{X} I(X;Y).$$

Let p_V and p_Y be the probability density functions of random variables V and Y (recall that V is Gaussian). Let $F \in \mathcal{F}$ be the cumulative density function of the random variable X, where $\mathcal{F}$ is the class of all CDFs for a given amplitude constraint. p_Y is the convolution of input and noise PDFs

$$p_Y(y;F) = \int_{A}^{A} p_V(y-x)dF(x).$$

Hence the differential entropy of the output is given by

$$H(Y;F) = -\int_{-A}^{A}\int_{-\infty}^{\infty} p_V(y-x)\log\left[p_Y(y;F)\right] dy\, dF(x).$$

We also denote $H(Y;F)$ as $H(A)$ while talking about the output differential entropy for a particular amplitude constraint A. The marginal differential entropy is given by

$$h\left(x;F\right) = -\int_{-\infty}^{\infty} p_V(y-x)\log\left[p_Y\left(y;F\right)\right] dy.$$

The purpose of this survey is to study the nature of the input distribution that maximizes the mutual information, i.e. the capacity achieving input distribution of the AWGN channel.

Note that since $H(V) = \log\sqrt{2\pi e}$, the problem of minimizing the mutual information reduces to that of minimizing the output differential entropy.

1.2 Weak Derivative [Smi71]

Weak derivative is the generalization of the concept of derivative for the functions that are not differentiable. For a convex space Ω and a function $f:\Omega\to\mathbb{R}$, the weak derivative at $x_0\in\Omega$ is defined as

$$f'_{x_0}(x) = \lim_{\theta\downarrow 0}\left\{\frac{f\left[(1-\theta)x_0+\theta x\right]-f\left(x_0\right)}{\theta}\right\},\ \forall x\in\Omega$$

Chapter 2

Finiteness of the Capacity Achieving Distribution

In this chapter, we will study the structure and geometry of the capacity achieving input distribution. We will see that the optimal input distribution is discrete and finite. We will also look at a basic algorithm to compute the optimal distribution.

First, we state the following basic optimization theorem from [Smi71].

Theorem 1. *Optimization Theorem Let f be a continuous, weakly-differentiable, strictly concave function from a compact, convex topological space Ω to R, and $C = \sup_{x\in\Omega} f(x)$, then:*

1. *$C = max\ f(x) = f(x_0)$ for some unique x_0*
2. *Necessary and sufficient condition for $f(x_0) = C$ is $f'_{x_0}(x) \leq 0, \forall x \in \Omega$*

The output entropy $H(Y; F)$ is a function from $\mathcal{F}$ to $\mathbb{R}$. If we show that $H(Y)$ and $\mathcal{F}$ satisfy the conditions in 1, then we can utilize it to find the optimal distribution.

Definition 1 (Lévy Metric [Wik17])**.** For two cumulative distribution functions, Lévy metric is defined as

$$L(F, G) := \inf\{\varepsilon > 0 | F(x - \varepsilon) - \varepsilon \leq G(x) \leq F(x + \varepsilon) + \varepsilon \forall x \in \mathbb{R}\}.$$

Intuitively, it is the length of the largest square with sides parallel to the axis that can be inscribed between the graphs of F and G.

Theorem 2. *Helly's Theorem [Wik21b] A sequence of monotonically increasing, uniformly bounded functions admits a pointwise convergent subsequence.*

Convexity of $\mathcal{F}$: Using the above definition and theorem, we can say that $\mathcal{F}$ is a metric space in Levy metric. It is easy to see that $\mathcal{F}$ is convex (Convex combination of two functions with bounded support also has bounded support).

Compactness of $\mathcal{F}$: Since $\mathcal{F}$ is a space of CDFs, they are monotonically increasing and uniformly bounded in $[0,1]$. Hence, $\mathcal{F}$ is also sequentially bounded.

Continuity of $H(Y;F)$: To see that $H(Y;F)$ is continuous, we need the following theorem,

Theorem 3. *Helly–Bray theorem [Wik20a] Let F_n be a sequence of cumulative distribution functions converging to F. Let g be a bounded and continuous function on R. Then*

$$\int_{\mathbb{R}} g(x)\, dF_n(x) \quad \xrightarrow[n\to\infty]{} \quad \int_{\mathbb{R}} g(x)\, dF(x).$$

The continuity of $H(Y;F)$ follows directly from Helly-Bray theorem, and the fact that p_Y and $log\ p_Y$ are continuous and bounded since $H(Y;F) = -\int p_Y \log\ p_Y$

Concavity of $H(Y;F)$ [SS10]: Let $F_1, F_2, F_\theta \in \mathcal{F}$ and $F_\theta = (1-\theta)F_1 + \theta F_2$. Then we have $p_Y(y;F_\theta) = (1-\theta)p_Y(y;F_1) + \theta p_Y(y;F_2)$. It follows that

$$\begin{aligned} & H(Y;F_\theta) - (1-\theta)H(Y;F_1) - \theta H(Y;F_2) \\ & = (1-\theta)D\left[p_Y(y;F_1) \| p_Y(y;F_\theta)\right] + \theta D\left[p_Y(y;F_2) \| p_Y(y;F_\theta)\right] \geq 0, \end{aligned}$$

with equality only when θ is 0 or 1. $D(.)$ is KL Divergence (which is always non-negative).

Weak Differentiability: The weak derivative of $H(Y;F)$ is given by

$$H'_{F_1}(F_2) = \int_{-\mathcal{A}}^{\mathcal{A}} h(x;F_1)\, dF_2(x) - H(Y;F_1).$$

Hence, $H(Y;F)$ and $\mathcal{F}$ satisfy the conditions of the theorem 1. Hence, for F^* to be optimal, the necessary and sufficient condition is $H'_{F^*}(F) \leq 0,\ \forall F \in \mathcal{F}$. This gives us the following corollary:

Corollary 4. *[Smi71] F^* is the optimal distribution if and only if the following two conditions are satisfied.*

$$h(x;F^*) \leq H(Y;F^*), \forall x \in [-A, A]$$

$$h(x;F^*) = H(Y;F^*), \forall x \in supp(F^*)$$

Proof. Suppose the first condition is not true. Then there is some x_0 not in support of F^* such that $h(x_0;F^*) > H(Y;F^*)$. Then taking $F = U(x - x_0)$ (unit step function), we get that $H'_{F^*}(F) > 0$ which is a contradiction.

Now if the second condition is not true, then for some subset S of $supp(F^*)$, $h(x,F^*) < H(Y,F^*)$, $\forall x \in S$. This leads to the contradiction $H(Y;F^*) < H(Y;F^*)$ which is clearly a contradiction. □

Before stating the main result of this chapter, we need the following two theorems from analysis.

Theorem 5. *Bolzano–Weierstrass Theorem [Wik21a] Each bounded sequence in R^n has a convergent subsequence.*

Theorem 6. *Identity Thorem [Wik21c] If two analytic function on $\mathbb{C}$ are same on a subdomain $D \subseteq \mathbb{C}$ containing a limit point, then they are same on $\mathbb{C}$.*

Theorem 7. *[Smi71] F^* has a finite support.*

Proof. The extension of $h(y;F^*)$ to the entire complex plane $\mathbb{C}$ is well-defined.

If $supp(F^*) \subseteq [-A, A]$ is infinite, then $supp(F^*)$ has a limit point according to Bolzano-Weierstrass theorem.

By Identity theorem, since $h(y; F^*)$ has same value on $supp(F^*)$, it has the same value on entire complex plane, and specifically on [-A,A].

This leads to $\int_{-\infty}^{\infty} p_V(y-x) \{-\log[p_Y(y; F^*)] - h(Y; F^*)\} dy = 0$ since $h(y; F^*) = H(Y; F^*)$ for all y.

This implies that $p_Y(y; F^*) = e^{-h(Y;F^*)}$, but output PDF cannot be a uniform distribution.

Thus, the support of F^* is finite. □

The problem of finding capacity achieving input distribution is now reduced to an optimization problem over finite number of variables.

The input distribution can be represented by the tuple $\mathbf{Z} = (\mathbf{x}, \mathbf{q}, n)$ where $\mathbf{x} = [x_1, \ldots, x_n]$ denotes the location of the mass points and $\mathbf{q} = [q_1, \ldots, q_n]$ denotes the corresponding weight.

The output density will be a Gaussian mixture $p_Y(y; \mathbf{Z}) = \sum_{\imath=1}^{n} q_i p_N(y - x_i)$ where $p_N(.)$ is Gaussian with 0 mean and unit variance.

But there is one major hurdle that needs to be overcome to find the optimal input distribution. We do not how many mass points will be there in the optimal distribution. But in practice, there is a nice workaround that can be used to find the optimal n and construct the optimal distribution.

If A is sufficiently small (around 0.1), then $h(x, F^*)$ is concave up and by the necessary and sufficient conditions in Corollary 1, the optimal distribution is an equal pair of mass points at the extremes of the interval $\pm A$, i.e. $\mathbf{x} = [-A, +A]$. Note that here the number of support points $n = 2$.

We take a major assumption that the number of mass points increase monotonically as the amplitude constraint is relaxed.

Optimal F for an arbitrary A can be computed using a simple iterative procedure:

1. Start with a small value of A, $\mathbf{x} = [-A, +A]$ and $n = 2$. Increment A by a small value(say δ) and check if the $\mathbf{x}$ is still optimal.

2. If $\mathbf{x}$ is not optimal, increment n by 1 and compute new $\mathbf{x}$.If $\mathbf{x}$ is optimal, then increase A by δ and continue.

Lastly, we look at one more interesting property of F^* which will be useful later.

Theorem 8. *[SS10] F^* is symmetric around origin.*

Proof. Let f be a function that takes an array as input and flips it, i.e. $f([x_1, \ldots, x_n]) = [x_n, \ldots, x_1]$. Consider the distributions F^* be the optimal input distribution characterized by $(\mathbf{x}, \mathbf{q}, n)$. Let G^* be another distribution characterized by $(-f(\mathbf{x}), f(\mathbf{q}), n)$. F^* and G^* give the same output differential entropy, i.e. $H(y; F^*) = H(y, G^*)$.

We know from Theorem 1 that $H(Y; F)$ is a strictly concave function and hence has a unique maxima. Since we have two optimal functions, it must be the case that they are same. $F^* = G^*$ which means f is symmetric around the origin. □

In the next chapter, we will see how the new mass points appear in F^* as we increase the amplitude constraint A. We will also look at the necessary and sufficient condition for this transition to occur.

Chapter 3

Transition Points of Optimal Distribution

In this chapter, we will state the necessary and sufficient conditions for a new mass point to appear as we increase the amplitude constraint. The constraint A at which the new mass point appears is called a **transition point**.

Let $F^*_{n^*,A}$ be the optimal distribution and let $F^*_{n,A}$ be the optimal n-point distribution for a given n, at constraint A. First we will need to show that $F^*_{n^*,A}$ is a continuous function of A. We need the following maximization theorem by Berge,

Theorem 9. *Berge's Maximization theorem [Wik20b] Let X and Θ be topological spaces, $f : X \times \Theta \to \mathbb{R}$ be a continuous function and $C : \Theta \to X$ be a compact-valued correspondence such that $C(\theta) \neq \phi$, $\forall \theta \in \Theta$. Let f^* be the value function*

$$f^*(\theta) = \sup\{f(x,\theta) : x \in C(\theta)\}$$

and C^* be the maximizers

$$C^*(\theta) = \arg\sup\{f(x,\theta) : x \in C(\theta)\} = \{x \in C(\theta) : f(x,\theta) = f^*(\theta)\}$$

Then, if C is continuous and f is strictly concave, then f^* and C^* are both continuous.

For our case, $\mathcal{F}$ is compact and $H(Y;F)$ is strictly concave in F, thus conditions in the theorem above is satisfied and $F^*_{n^*,A}$ is a continuous function of A, i.e. $\mathbf{q}^*$ and $\mathbf{x}^*$ are continuous as a function of A. But note that we claim the same for n^* because due to continuity of $\mathbf{q}^*$ and $\mathbf{x}^*$, mass points with probability close to zero are also admissible.

Output PDF can also be written as $p_Y(y) = \sum_{j=1}^{n} q_j \varphi\left(y, x_j\right)$ where $x_i \geq 0$

$$\varphi(y,x) = \frac{1}{2\sqrt{2\pi}}\left[e^{-(y-x)^2/2} + e^{-(y+x)^2/2}\right]$$

(we have encoded the symmetry of input distribution directly in the definition).

Necessary Conditions: We do not need to consider the symmetry of F^* for this argument to work. Suppose a new mass point appears at x with weight q at the constraint A, then it must have $h(x;F^*_{n,A}) = H(F^*_{n,A})$ and $q = 0$. Further, the derivative of marginal entropy is also zero at x

$$\left.\frac{dh\left(w;F^*_{n,A}\right)}{dw}\right|_{w=x} = h'\left(x;F^*_{n,A}\right) = .0$$

If the mass point x_i splits into two, then at $A+\Delta A$, $h'(x_i \pm \Delta x_i; F^*_{n,A+\Delta A}) = 0$. Hence, taking the limit $\Delta A \to 0$ and $\Delta x \to 0$, we get

$$\left.\frac{d^2h\left(w;F^*_{n,A}\right)}{dw}\right|_{w=x} = h''\left(x;F^*_{n,A}\right) = 0.$$

because

$$h''\left(x;F^*_{n,A}\right) = \lim_{\Delta A,\Delta x \to 0} \frac{h'(x_i \pm \Delta x_i; F^*_{n,A+\Delta A}) - h'\left(x_i;F^*_{n,A}\right)}{\Delta x_i} = 0$$

Sufficient Conditions: Since n increases at most by 1, as we increase A, a new mass point will either appear at the origin (when the number of mass points go from even to odd), or an existing mass point will split into two (when the number of mass points go from odd to even). We handle both the cases separately.

A New Point Appears at Origin: Let A be the amplitude constraint at which the number of mass points in the optimal distribution goes from $n = 2k$ to $n+1$. Hence $h(0; F^*_{n,A}) = H(Y; F^*_{n,A})$. At $A + \Delta A$, the optimal distribution will have n+1 points. We denote *some* (not necessarily optimal) distribution having (n+1) mass points $F_{n+1,A+\Delta A}$ in terms of $F^*_{n,A+\Delta A}$ scaled down by $(1-\varepsilon)$ and a mass point at origin with weight ε.

Hence, the output distribution will be

$$p_Y\left(y; F_{n+1,A+\Delta A}\right) = p_Y\left(y; F^*_{n,A+\Delta A}\right) + \varepsilon\left[\varphi(y,0) - p_Y\left(y; F^*_{n,A+\Delta A}\right)\right]$$

Multiplying the LHS and RHS with $\log(p_Y\left(y; F_{n+1,A+\Delta A}\right))$ equation with and taking the integral over y we get (use $\log(a+b) = \log(a) + \log\left(1+\frac{b}{a}\right)$)

$$\begin{aligned} H\left(Y; F_{n+1,A+\Delta A}\right) = {} & H\left(Y; F^*_{n,A+\Delta A}\right) \\ & + \varepsilon\left[h\left(0; F^*_{n,A+\Delta A}\right) - H\left(Y; F^*_{n,A+\Delta A}\right)\right] + O\left(\varepsilon^2\right) \end{aligned}$$

If for some small $\varepsilon, \Delta A > 0$,

$$h(0; F^*_{n,A+\Delta A}) > h(x^*_i + \Delta x^*_i; F^*_{n,A+\Delta A}),\ \forall i,$$

where $x^*_i + \Delta x^*_i$ are mass points for $F^*_{n,A+\Delta A}$, then the second term in the equation above will be positive.

Hence, for sufficiently small $\varepsilon > 0$ we have

$$H\left(Y; F_{n+1,A+\Delta A}\right) > H\left(Y; F^*_{n,A+\Delta A}\right).$$

This means that for an even n and some $A + \Delta A$, when origin has higher marginal entropy than all other mass points, there exists an (n+1)-point distribution that beats the optimal n-point distribution.

Mass Point at Origin Splits into Two: Let A be the amplitude at which the mass point at origin splits into two. The new mass points at $A + \Delta A$ will be located at $\pm\Delta x$. Proceeding just as before, we obtain $F_{n+1,A+\Delta A}$ in terms of $F^*_{n,A+\Delta A}$ scaled down by $(1-\varepsilon)$ and a mass point at Δx with weight ε. Note that $\varphi(y, \Delta x)$ encodes both $\pm\Delta x$. The resulting PDF will be

$$p_Y\left(y; F_{n+1,A+\Delta A}\right) = p_Y\left(y; F^*_{n,A+\Delta A}\right) + \varepsilon[\varphi(y, \Delta x) - \varphi(y, 0)].$$

Taking the Taylor expansion of $\varphi(y, \Delta x)$ at $\Delta x = 0$, recalling that $h'(\pm\Delta x_i; F^*_{n,A+\Delta A}) = 0$ from the necessary conditions and neglecting the integrals of odd functions (since the integral is over $-\infty$ to ∞), we get

$$H\left(Y; F_{n+1,A+\Delta A}\right) = H\left(Y; F^*_{n,A+\Delta A}\right) + \varepsilon\Delta x^2 h''\left(0; F^*_{n,A+\Delta A}\right) + O\left(\varepsilon^2\right)$$

For some $\varepsilon > 0$, if for some $\Delta A > 0$, $h''\left(0; F^*_{n,A+\Delta A}\right) > 0$ then the (n+1)-point distribution beats the optimal n-point distribution.

Note that these conditions also improve the previously stated algorithm since we do not need to check all $x \in [-A, A]$. Instead, it is enough to monitor the origin as the amplitude constraint is relaxed.

Chapter 4

Bound on the Size of Optimal Support

So far we have established that the support of optimal input is discrete for an amplitude constrained AWGN channel. We have also stated the necessary and sufficient conditions for a new mass point to appear assuming the conjecture that the number of mass point increase monotonically and at most by 1 as the amplitude constraint is relaxed. In this chapter we bound the size of the support of the optimal input distribution.

Let $\mathrm{N}(\mathcal{D}, f)$, the number of zeroes of the function f in domain $\mathcal{D}$. Let $H^*(A)$ be the optimal entropy. Let $P_{X^\star}$ be optimal input distribution and $f_{Y^\star}(y)$ be optimal output PDF.

From Corollary 1, we already know that all the support points satisfy the condition that marginal entropy at that point is equal to the optimal entropy, i.e.

$$\operatorname{supp}\left(P_{X^\star}\right) \subseteq \left\{x : h\left(x; P_{X^\star}\right) - H^*(\,\mathrm{A}) = 0\right\}.$$

Hence we can upper bound the cardinality of support in terms of number of roots of some continuous function

$$\begin{aligned}
\left|\operatorname{supp}\left(P_{X^\star}\right)\right| &\leq \mathrm{N}\left([-\mathrm{A}, \mathrm{A}], \Xi_{\mathrm{A}}\left(\cdot; P_{X^\star}\right)\right) \\
&\leq \mathrm{N}\left(\mathbb{R}, \Xi_{\mathrm{A}}\left(\cdot; P_{X^\star}\right)\right)
\end{aligned}$$

where $\Xi_{\mathrm{A}} = h\left(x; P_{X^\star}\right) - H^*(\mathrm{~A})$.

We first state the following oscillation theorem, which will play a central role in determining the upper. It will help us to bound the support of the optimal input distribution in terms of the number of zeros of the shifted optimal output PDF f_{Y^*}

Theorem 10 (Oscillation Theorem [Dyt+19]). *Given open intervals $\mathbb{I}_1$ and $\mathbb{I}_2$, let $p : \mathbb{I}_1 \times \mathbb{I}_2 \rightarrow \mathbb{R}$ be a strictly totally positive kernel. For an arbitrary y, suppose $p(\cdot, y) : \mathbb{I}_1 \rightarrow \mathbb{R}$ is an n -times differentiable function. Assume that μ is a measure on $\mathbb{I}_2$, and let $\xi : \mathbb{I}_2 \rightarrow \mathbb{R}$ be a function with $S(\xi) = n$. For $x \in \mathbb{I}_1$, define*

$$\Xi(x) = \int \xi(y) p(x, y) \mathrm{d}\mu(y)$$

If $\Xi : \mathbb{I}_1 \rightarrow \mathbb{R}$ is an n -times differentiable function, then either $\mathrm{N}\left(\mathbb{I}_1, \Xi\right) \leq n$, or $\Xi \equiv 0$, where $S(f)$ defines the number of sign changes in the function f.

Upper Bound:

Since the Gaussian kernel is totally strictly positive, we can directly use the above theorem to bound $|\operatorname{supp}\left(P_{X^\star}\right)|$.

Recall that

$$h\left(x; P_{X^\star}\right) = \int_{\mathbb{R}} \frac{\mathrm{e}^{-\frac{(y-x)^2}{2}}}{\sqrt{2\pi}} \log \frac{1}{f_{Y^\star}(y)} \mathrm{d}y.$$

Hence, as per the theorem we have

$$\Xi_{\mathrm{A}}\left(x; P_{X^\star}\right) = \int_{\mathbb{R}} \frac{\xi_{\mathrm{A}}(y)}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(y-x)^2}{2}} \mathrm{~d}y$$

where

$$\xi_{\mathrm{A}}(y) = \log \frac{1}{f_{Y^\star}(y)} - H^*(\mathrm{~A})$$

From the observation above, we can write

$$|\operatorname{supp}\left(P_{X^\star}\right)| \leq \mathrm{N}\left(\mathbb{R}, \Xi_{\mathrm{A}}\left(\cdot; P_{X^\star}\right)\right) \leq S\left(\xi_{\mathrm{A}}\right) \leq \mathrm{N}\left(\mathbb{R}, \xi_{\mathrm{A}}\right)$$

since number of sign changes of a continuous function is bounded above by number of zeros of the function.

It is easy to see that $\xi_{\mathrm{A}} = 0$ iff $f_{Y^\star} = \kappa_1$ where $\kappa_1 = e^{-C(A)-H(V)} = e^{-H^*(A)}$. So we have

$$\boxed{|\text{supp}\,(P_{X^\star})| \leq \mathrm{N}\,(\mathbb{R}, f_{Y^\star} - \kappa_1)}$$

Note that since $I(X;Y) \geq 0$, $H(Y) \geq H(V) = log(\sqrt{2\pi e})$ for a given A. Hence $\kappa_1 \in (0, \frac{1}{\sqrt{2\pi}})$.

Lower Bound:

Since we know that optimal input distribution has finite support, let $|\text{supp}\,(P_{X^\star})| = n$, where the support points are $x_1 < x_2 \ldots < x_n$. We can write the output PDF as

$$f_{Y^\star}(y) = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{n} P_{X^\star}(x_i) \exp\left(-\frac{1}{2}(y - x_i)^2\right).$$

The number of zeros of $f_{Y^\star}(y) - \kappa_1$ is same as number of zeros of the right shifted function $f(y) - \kappa_1 = f_{Y^\star}(y - |x_1| - 1)$. $f(y)$ can be written as

$$f(y) = \sum_{i=1}^{n} a_i \exp\left(-\frac{1}{2}(y - u_i)^2\right) - a_0,\ \forall i \in [1, \ldots, n]$$

for

$$u_i = x_i + |x_1| + 1 \text{ for } i = 1, \ldots n,$$
$$a_i = \begin{cases} \kappa_1 & i = 0 \\ \frac{1}{\sqrt{2\pi}} P_{X^\star}(x_i) & i = 1, \ldots, n \end{cases}$$

For some arbitrary $0 < \epsilon_1 < \ldots < \epsilon_n$, consider the perturbed version of $f(y)$

$$\mathrm{e}^{-\frac{1}{2}y^2} \tilde{f}(y, \epsilon_1, \ldots, \epsilon_n) = \sum_{i=0}^{n} b_i \exp\left(-\frac{(2 + \epsilon_i)}{2}(y - v_i)^2\right)$$

where,

$$\epsilon_0 = -1,$$

$$b_i = \begin{cases} -a_0 & i = 0, \\ a_i \exp\left(-\frac{1+\epsilon_i}{2(2+\epsilon_i)} u_i^2\right) & i = 1, \dots, n, \end{cases}$$

$$v_i = \begin{cases} 0 & i = 0, \\ \frac{1+\epsilon_i}{2+\epsilon_i} u_i & i = 1, \dots, n \end{cases}$$

The important thing to note is that $e^{-\frac{1}{2}y^2} \tilde{f}(y, \epsilon_1, \dots, \epsilon_n)$ is a heteroscedastic (having different variance) linear combination of n+1 Gaussians and hence, it can have maximum $n+1$ peaks and $2n$ zeros.

Since $\tilde{f}(y, \epsilon_1, \dots, \epsilon_n) \to f(y)$ as $(\epsilon_1, \dots, \epsilon_n) \to (0, \dots, 0)$. So $2\left|\text{supp}(P_{X^\star})\right| \geq \mathrm{N}(\mathbb{R}, f_{Y^\star} - \kappa_1)$. Hence, finally we have

$$\boxed{\frac{1}{2}\mathrm{N}(\mathbb{R}, f_{Y^\star} - \kappa_1) \leq \left|\text{supp}(P_{X^\star})\right| \leq \mathrm{N}(\mathbb{R}, f_{Y^\star} - \kappa_1)}$$

Since we have no information about the optimal input or output distributions, we have to work with generic f_Y and $\kappa_1 \in (0, \frac{1}{\sqrt{2\pi}})$.

We need a bound on the number of zeros of $f_Y - \kappa_1$ where f_Y is the PDF of a random variable Y=X+V where $|X| \leq A$ and $\mathbb{E}[X] = 0$ and V is standard Gaussian independent of X.

First, we see that all the zeros are contained in a bounded interval. For all $|y| > A$,

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \mathbb{E}\left[e^{-\frac{(y-X)^2}{2}}\right] \leq \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-A)^2}{2}}$$

$f_Y(y) - \kappa_1 < 0$ for all $|y| > A + \log^{\frac{1}{2}}\left(\frac{1}{2\pi\kappa_1^2}\right) = B_{\kappa_1}$. So

$$\boxed{\mathrm{N}(\mathbb{R}, f_Y - \kappa_1) = \mathrm{N}([-B_{\kappa_1}, B_{\kappa_1}], f_Y - \kappa_1)}$$

Note that f_Y is analytic (as convolution with Gaussian preserves analyticity). Following a similar argument as in theorem 1, using Bolzano-Weierstrass theorem and Identity theorem, it follows that $f_Y - \kappa_1$ has finitely many zeros in a bounded interval. This is an alternate proof of Theorem 1.

The κ_1 in the above term is a hurdle, and we can get rid of it using a simple trick. By Rolle's theorem, w.k.t for a function f continuous on $[-R, R]$ and differentiable on $(-R, R)$ with finite number of zeros, then $\mathrm{N}([-R, R], f) \leq \mathrm{N}([-R, R], f') + 1$.

Hence, the problem reduces to bounding the number of zeros of

$$f'_Y(y) = \frac{1}{\sqrt{2\pi}} \mathbb{E}\left[(X - y)\mathrm{e}^{-\frac{(y-X)^2}{2}}\right].$$

We will bound the number of zeros of $f'_Y(y)$ using its maximum value in an interval.

Tijdeman's Number of Zeros Lemma bounds the number of zeros of a complex analytic function in an open disc in terms of its maximum value. For $R, t > 0$ and $s > 1$, the complex valued function $f \neq 0$ which is analytic on $|z| < (st + s + t)R$, its number of zeros $\mathrm{N}(\mathcal{D}_R, f)$ within the disk $\mathcal{D}_R = \{z : |z| \leq R\}$ satisfies

$$\mathrm{N}(\mathcal{D}_R, f) \leq \frac{1}{\log s}\left(\log \max_{|z| \leq (st+s+t)R} |f(z)| - \log \max_{|z| \leq tR} |f(z)|\right)$$

Now, all we need is a lower and upper bound on the maximum value of $\breve{f}'_Y(z)$, which is the complex analytic extension of $f'_Y(y)$.

Upper Bound: Using Jensen's inequality and triangle inequality, it can be proved that

$$\max_{|z| \leq \mathrm{B}} \left|\breve{f}'_Y(z)\right| \leq \frac{1}{\sqrt{2\pi}}(\mathrm{A} + \mathrm{B}) \exp\left(\frac{\mathrm{B}^4}{2}\right)$$

Lower Bound: For $|X| \leq A \leq B$ using the fact that $\mathbb{E}[X] = 0$, we have

$$\max_{|z| \leq \mathrm{B}} \left|\breve{f}'_Y(z)\right| \geq \frac{\mathrm{A}}{\sqrt{2\pi}} \exp\left(-2\,\mathrm{A}^2\right)$$

For $R > A + \log^{\frac{1}{2}}\left(\frac{1}{2\pi\kappa_1^2}\right)$, we have a final bound

$$\begin{aligned}&\mathrm{N}\left([-R,R], f_Y - \kappa_1\right)\\&\leq 1 + \min_{s>1}\left\{\frac{\left(\frac{((\mathrm{A}+R)s+\mathrm{A})^2}{2} + 2\ \mathrm{A}^2 + \log\left(2 + \frac{(\mathrm{A}+R)s}{\mathrm{A}}\right)\right)}{\log s}\right\}\end{aligned}$$

But the bound is in terms of R and κ_1, which needs to be resolved.

An amplitude constraint $|X| \leq \mathrm{A}$ induces a second moment constraint $\mathbb{E}\left[X^2\right] \leq \mathrm{A}^2$, and therefore

$$\mathrm{C}(\mathrm{A}) = \max_{|X|\leq \mathrm{A},\ \mathbb{E}[X^2]\leq \mathrm{A}^2} I(X;Y) \leq \frac{1}{2}\log\left(1+\mathrm{A}^2\right)$$

This gives us a bound on $\kappa_1 = e^{-(\mathrm{C}(\mathrm{A})+h(\mathrm{Z}))} \leq \sqrt{2\pi \mathrm{e}\left(1+\mathrm{A}^2\right)}$ which implies $B_{\kappa_1} \leq 2A+1$.

Hence, letting $R \leftarrow (2A+1)$ we get the final bound $\mathrm{N}\left(\mathbb{R}, f_{Y^*} - \kappa_1\right) = \mathrm{N}\left(\left[-\mathrm{B}_{\kappa_1},\ \mathrm{B}_{\kappa_1}\right], f_{Y^*} - \kappa_1\right) = \mathcal{O}(A^2)$ and we have

$$\boxed{|\mathrm{supp}\left(P_{X^\star}\right)| \leq \mathrm{N}\left(\mathbb{R}, f_{Y^\star} - \kappa_1\right) \leq cA^2}$$

for some constant c.

Chapter 5

A Discrete Setting

The problem of finding the optimal support has a natural discrete analogue [proposed by Dr. Kashyap and Dr. Krishnapur].

Let Z_m be a random variable with binomial distribution $B(m, 0.5)$. So $P(Z_m = k) = \binom{m}{k} 2^{-m}$, $\forall k \in (0, \ldots, m)$. Let X be a random variable taking values on $(x_1, \ldots, x_n)$. We want to maximize $H(X + Z_m)$ where $H(.)$ is the Shannon entropy $H(X) = -\sum p_n \log p_n$ where $p_n = \mathbb{P}\{X = n\}$.

Note that $lim_{m \to \infty} \frac{Z_m - \frac{m}{2}}{\sqrt{\frac{m}{4}}} = \mathcal{N}(0, 1)$. Hence, as $m \to \infty$ and $n \to A\sqrt{m}$, the discrete problem converges to the continuous one.

Let X' be an equal probability distribution with two support points at the extremes, i.e. $[0.5, \ldots, 0.5]$. Experimental results suggest that for a fixed n, as $m \to \infty$, the optimal distribution $H(X^*) \to H(X')$.

Let the probabilities for Z_m, X and Y respectively be

$$b_i = \binom{m}{i} 2^{-m}, \;\; p_j = P(X = j), \; and \;\; q_l = \sum_{k=0}^{n} p_k b_{l-k}$$

The marginal entropy is given by $h(x_j, X) = -\sum_l b_{l-j} \; log \; q_l$ and $H(Y) = \sum_k p_k h(x_k, X)$.

Let the optimal input be $X^* = argmax_X H(Y)$ and the optimal output entropy be $max_X H(Y) = H^*$

Similar to Corollary 1, it can be seen that if $x_i \in Supp(X^*)$, then $h(x_i, X^*) = H^*$.

We want a bound on $|supp(X^{*})| \leq n$. Similar to the continuous case, we have

$$\operatorname{supp}\left(X^{\star}\right) \subseteq \left\{x : h\left(x; X^{\star}\right) - H^{*} = \Xi(., X^{*}) = 0\right\}.$$

and hence

$$\left|\operatorname{supp}\left(X^{\star}\right)\right| \leq \mathrm{N}\left(\mathbb{N}, \Xi\left(\cdot; X^{\star}\right)\right)$$

As the binomial distribution is also a strictly positive kernel, assuming that the oscillation theorem holds for the discrete setting, we have the following main issue in extending the results of [Dyt+19] into the discrete case:

Discreteness of ξ is an issue:

Let $\xi_{\mathrm{A}}(y) = \log \frac{1}{q_y} - H^{*}$. By the oscillation theorem,

$$\left|\operatorname{supp}\left(X^{\star}\right)\right| \leq \mathrm{N}\left(\mathbb{N}, \Xi\left(\cdot; X^{\star}\right)\right) \leq S(\xi)$$

where $S(f)$ is the number of sign changes of f.

In [Dyt+19], since ξ is continuous, $S(\xi)$ is bounded above by the number of zeros of ξ. The issue here is that $S(\xi)$ cannot be bounded above by the number of roots of $\xi(y)$ since it is a discrete function.

Using Descartes' rule:

Consider ξ to be the list of coefficients of a polynomial $P(\xi)$ of degree m+n. Descartes' rule of sign can only give a lower bound on $S(\xi)$ in terms of $N(\mathbb{R}, P(\xi))$.

A possible solution would be to interpolate a polynomial passing through all the points $(y, \xi(y))$ and try to bound its number of roots. A common way to count the number of real roots of a polynomial is to count the number of sign changes of the leading terms in the Sturm sequence of the polynomial. The problem with this approach is again, that we do not know $\xi(y)$ explicitly, and working with a generic polynomial will only give us a bound in the degree of the polynomial.

Bibliography

[Dyt+19] Alex Dytso et al. "The capacity achieving distribution for the amplitude constrained additive Gaussian channel: An upper bound on the number of mass points". In: *IEEE Transactions on Information Theory* 66.4 (2019), pp. 2006–2022.

[Smi71] Joel G Smith. "The information capacity of amplitude-and variance-constrained sclar Gaussian channels". In: *Information and control* 18.3 (1971), pp. 203–219.

[SS10] Naresh Sharma and Shlomo Shamai. "Transition points in the capacity-achieving distribution for the peak-power limited AWGN and free-space optical intensity channels". In: *Problems of Information Transmission* 46.4 (2010), pp. 283–299.

[Wik17] Wikipedia contributors. *Lévy metric — Wikipedia, The Free Encyclopedia*. [Online; accessed 6-June-2021]. 2017. URL: https://en.wikipedia.org/w/index.php?title=L%C3%A9vy_metric&oldid=811060730.

[Wik20a] Wikipedia contributors. *Helly–Bray theorem — Wikipedia, The Free Encyclopedia*. [Online; accessed 6-June-2021]. 2020. URL: https://en.wikipedia.org/w/index.php?title=Helly%E2%80%93Bray_theorem&oldid=973392803.

[Wik20b] Wikipedia contributors. *Maximum theorem — Wikipedia, The Free Encyclopedia*. [Online; accessed 7-June-2021]. 2020. URL: https://

en.wikipedia.org/w/index.php?title=Maximum_theorem&oldid=976077619.

[Wik21a] Wikipedia contributors. *Bolzano–Weierstrass theorem — Wikipedia, The Free Encyclopedia*. [Online; accessed 6-June-2021]. 2021. URL: https://en.wikipedia.org/w/index.php?title=Bolzano%E2%80%93Weierstrass_theorem&oldid=1026653272.

[Wik21b] Wikipedia contributors. *Helly's selection theorem — Wikipedia, The Free Encyclopedia*. [Online; accessed 6-June-2021]. 2021. URL: https://en.wikipedia.org/w/index.php?title=Helly%27s_selection_theorem&oldid=1000180793.

[Wik21c] Wikipedia contributors. *Identity theorem — Wikipedia, The Free Encyclopedia*. [Online; accessed 6-June-2021]. 2021. URL: https://en.wikipedia.org/w/index.php?title=Identity_theorem&oldid=1026216096.