View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by INRIA a CCSD electronic archive server

Capacity Approximation of Continuous Channels by Discrete Inputs

Malcolm Egan, Samir Perlaza

To cite this version:

Malcolm Egan, Samir Perlaza. Capacity Approximation of Continuous Channels by Discrete Inputs. CISS 2018 - 52nd Annual Conference on Information Sciences and Systems, Mar 2018, Princeton, United States. pp.1-6, 10.1109/CISS.2018.8362269 . hal-01686036

HAL Id: hal-01686036

https://hal.archives-ouvertes.fr/hal-01686036

Submitted on 16 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Capacity Approximation of Continuous Channels by Discrete Inputs

Malcolm Egan and Samir M. Perlaza

***Abstract*—In this paper, *discrete approximations* of the capacity are introduced where the input distribution is constrained to be discrete in addition to any other constraints on the input. For point-to-point memoryless additive noise channels, rates of convergence to the capacity of the original channel are established for a wide range of channels for which the capacity is finite. These results are obtained by viewing discrete approximations as a capacity sensitivity problem, where capacity losses are studied when there are perturbations in any of the parameters describing the channel. In particular, it is shown that the discrete approximation converges arbitrarily close to the channel capacity at rate $O(\Delta)$, where Δ is the discretization level of the approximation. Examples of channels where this rate of convergence holds are also given, including additive Cauchy and inverse Gaussian noise channels.**

I. Introduction

The channel capacity is a fundamental limit of a wide range of communication systems, which informs communication strategies both in terms of channel coding and also resource allocation. As such, it is important to obtain computable representations of the capacity. In the case of discrete memoryless (DM) channels, such a representation always exists [1]. However, for continuous memoryless channels it remains challenging to effectively characterize the channel capacity.

Despite the challenges for general channels, there has been progress in special cases. For example, the Gaussian noise channel subject to a transmit power constraint has been studied in [1].

An alternative approach is to establish that the optimal input distribution is discrete. This approach was first developed by Smith [2] for Gaussian channels subject to power and amplitude constraints, and has since been investigated in a range of contexts. In particular, a general characterization of noise distributions and constraints that admit a unique discrete optimal input has recently been obtained by Fahs and Abou-Faycal [3]. Nevertheless, there remain channels arising in emerging communication systems that do not fall into the characterization in [3], including the additive inverse Gaussian noise channel with an absolute moment constraint in molecular communication [4][1].

It is highly desirable that the optimal input distribution is discrete as it is then often possible to obtain arbitrarily accurate approximations of the capacity with error guarantees using convex optimization techniques [2]. This observation motivates the study of capacity approximations obtained by introducing a further constraint that the input distribution is discrete. Although discretization arguments have been used to establish the achievability part of the noisy channel coding theorem (NCT) for additive Gaussian noise channels [5] to exploit the NCT for DM channels, there has not been a systematic study of discrete input approximations of the capacity for general classes of channels.

In this paper, we study the convergence of discrete input approximations to the capacity of continuous memoryless additive noise channels. That is, we consider the capacity of the channel subject to the additional constraint that the input is discrete with mass points separated by a given Euclidean distance, which is often referred to as the discretization level denoted by Δ. In contrast to existing work, which has focused on particular channels (e.g., the additive Gaussian noise channel in [6]), this work considers general classes of channels and input constraints only under assumptions that guarantee the channel capacity exists and is finite.

There are two fundamental questions that are answered in this work: *under what conditions is the channel capacity continuous in the discretization level* and *what is the rate of convergence to the capacity of the original channel?*

In order to answer these questions, we study the discrete input approximation within the framework of the capacity sensitivity [7]. Here, instead of directly characterizing the capacity of a channel, the focus is on how the capacity changes whenever any of the parameters describing the channel are varied. In particular, the behavior of the discrete input approximation can be viewed more generally in terms of how the capacity changes when the constraints on the input distribution are varied. In [7], the effect of varying constraint parameters such as the maximum power level were considered. In this paper, this analysis is extended to the case where the input distribution is discrete.

The key result is that for a wide range of additive noise channels for which the capacity optimization problem is well-posed—i.e., an optimal input distribution exists—the convergence of the capacity subject to a discrete input constraint converges arbitrarily close to the channel capacity according to $O(\Delta)$. This result is very general and in fact can be readily extended to point-to-point channels that are non-linear or even non-additive.

The organization of the paper is as follows. Section II formalizes the problem and Section III establishes the connection

Malcolm Egan and Samir M. Perlaza are with the Laboratoire CITI (a joint laboratory between the Université de Lyon, INRIA, and INSA de Lyon), 6 Avenue des Arts, F-69621, Villeurbanne, France ({malcom.egan, samir.perlaza}@inria.fr). S. Perlaza is also with the Department of Electrical Engineering at Princeton University, Princeton, NJ 08544 USA.

[1]In fact, the optimal input distribution for the molecular timing channel with an absolute moment constraint is conjectured to be of mixed-type.

to the capacity sensitivity. Section IV and Section V present the proofs of the main results on the continuity and rate of convergence for the discrete input approximation. Section VI discusses extensions and relationships with other capacity sensitivity problems. Section VII concludes this work.

II. Problem Formulation

This section focuses on real-valued point-to-point channels. For concreteness—generalizations are discussed later in Section VI—consider the linear additive noise channel with real output Y of the form

$$Y = X + N \tag{1}$$

where the input X has an alphabet $\mathcal{X} \subseteq \mathbb{R}$ and the noise has a distribution function on $\mathbb{R}$, denoted by F_N. We assume that F_N has a probability density function, denoted by p_N. Given that the channel is linear and additive, the channel Markov kernel can be written as

$$p_{Y|X}(y|x) = p_N(y - x). \tag{2}$$

As a consequence of the noisy channel coding theorem [Han], when the capacity of (1) exists it is obtained by optimizing the mutual information subject to any constraints on the input X. Let $\mathcal{B}(\mathbb{R})$ be the Borel σ-algebra on $\mathbb{R}$ and let $\mathcal{P}$ denote the collection of Borel probability measures on the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ equipped with the topology of weak convergence.

An important property of the topology of weak convergence is that it is metrized by the Lévy-Prohorov metric [8]. In more detail, recall that a sequence $\mu_1, \mu_2, \ldots$ of probability measures converges weakly[2] to a probability measure μ_0 if for all bounded and continuous functions f,

$$\int f d\mu_n \to \int f d\mu_0, \text{ as } n \to \infty. \tag{3}$$

The Lévy-Prohorov metric is related to weak convergence as follows. A topology on a set (in this case, the set of probability measures) is defined by closed subsets. In turn, a closed set U has the property that all convergent sequences in U have a limit point in U. As such, weak convergence of probability measures induces a topology and a natural question is whether there exists a metric on the space of probability measures that induces the same topology. As already noted, this is indeed the case and one such metric is the Lévy-Prohorov metric.

Denote $I(\mu, p_N)$ as the mutual information for the channel in (1), given by

$$I(\mu, p_N) = \int_{\mathbb{R}} \int_{\mathbb{R}} p_N(y - x) \log \frac{p_N(y - x)}{p_Y(y)} d\mu(x) dy, \tag{4}$$

where μ is the input probability measure. The capacity of (1) is then the solution to the optimization problem

$$\begin{aligned} C(\Lambda) = \sup_{\mu \in \mathcal{P}} \quad & I(\mu, p_N) \\ \text{subject to} \quad & \mu \in \Lambda, \end{aligned} \tag{5}$$

[2]Weak convergence of probability measures should not be confused with weak convergence in a topological vector space.

where Λ is a weakly compact subset of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Key examples of the set Λ are the p-th order constraints $(p > 0)$, defined by

$$\Lambda^p = \{\mu : \mathbb{E}_\mu[|X|^p] \leq b\}, \tag{6}$$

where $b > 0$. The compactness of this constraint set can be readily shown by the application of Prokhorov's Theorem [9].

Now let $\Delta > 0$ and $\mathcal{P}(\Delta\mathbb{Z})$ be the set of probability measures with support on $\Delta\mathbb{Z} = \{\Delta z : z \in \mathbb{Z}\}$ with $\mathbb{Z}$ the set of integers. The question we focus on in this paper is how the problem in (5) be approximated by a problem where the input is required to be discrete. That is, we consider the *discrete input approximation*

$$\begin{aligned} C(\Lambda_\Delta) = \sup_{\mu \in \mathcal{P}} \quad & I(\mu, p_N) \\ \text{subject to} \quad & \mu \in \Lambda_\Delta, \end{aligned} \tag{7}$$

where $\Lambda_\Delta = \cup_{\overline{\Delta} > \Delta} \mathcal{P}(\overline{\Delta}\mathbb{Z}) \cap \Lambda$. As such, (7) corresponds to the capacity of the channel in (1), where the input is constrained to be discrete.

The question addressed in this paper is how well the capacity in (5) can be approximated by (7) as the discretization level Δ tends to zero. Unlike other discrete input approximations such as those based on the Ozarow-Wyner technique [10], [11] with sub-optimal PAM constellations, it is necessary to study the behavior of optimized discrete inputs. As such, the question can be viewed within the framework of capacity sensitivity as detailed in the following section.

III. Capacity Sensitivity and Discrete Input Approximations

The capacity of a memoryless additive noise channel can be viewed as a map from the input alphabet $\mathcal{X}$, the output alphabet $\mathcal{Y}$, the noise distribution F_N, and the constraint set Λ to $\mathbb{R}_+$. That is, $(\mathcal{X}, \mathcal{Y}, F_N, \Lambda) \mapsto C$, where C is the optimal value function of the optimization problem in (5) or (7).

In order to study approximations of one channel by another, it is natural to introduce the capacity sensitivity [7], [12]. In particular, the capacity sensitivity is the capacity gap between two channels, and is defined formally as follows.

Definition 1. *Let $\mathcal{K} = (\mathcal{X}, \mathcal{Y}, F_N, \Lambda)$ and $\hat{\mathcal{K}} = (\hat{\mathcal{X}}, \hat{\mathcal{Y}}, \hat{F}_N, \hat{\Lambda})$ be two tuples of channel parameters. The capacity sensitivity due to a perturbation from channel $\mathcal{K}$ to the channel $\hat{\mathcal{K}}$ is defined as*

$$C_{\mathcal{K} \to \hat{\mathcal{K}}} \triangleq |C(\mathcal{K}) - C(\hat{\mathcal{K}})|. \tag{8}$$

The capacity sensitivity problem can be viewed as a special case of analyzing the sensitivity of nonlinear optimization problems, where we identify the capacity as the *optimal value function*. Clearly, the problem of computing the capacity sensitivity is trivial when the capacity is available in closed-form (such as the case of additive Gaussian noise with a power constraint). However, the problem is significantly more challenging in the usual situation in which the only explicit characterization of the capacity is (5) under general perturbations from one channel to another.

In this paper, we are concerned with a class of *constraint perturbations*. In particular, the study of discrete input approximations of the capacity involves analyzing the effect of varying the constraint set Λ. The capacity sensitivity in this case therefore corresponds to

$$C_{\Lambda\to\Lambda_\Delta} = |C(\Lambda) - C(\Lambda_\Delta)|, \tag{9}$$

for $\Delta > 0$.

A first study of general constraint perturbations for the capacity optimization problem is given in [7]; however, the analysis was limited to the case that the constraint set is defined by a finite number of inequalities. This is not the case for the set Λ_Δ and it is therefore necessary to apply further techniques from nonsmooth analysis [13], which are used in the sequel to establish rates of convergence of the discrete input approximation.

IV. Convergence of the Discrete Input Approximation

Before characterizing rates of convergence of the discrete input approximation, it is necessary to first establish that the approximation indeed converges. To this end, in this section the convergence of the discrete input approximation is studied. In particular, we provide sufficient conditions for $C(\Lambda_\Delta) \to C(\Lambda)$ as $\Delta \to 0$, which are detailed in the following theorem.

Theorem 1. *Let Λ be a non-empty compact subset of $\mathcal{P}$. If the mutual information $I(\cdot, p_N)$ is weakly continuous on Λ, then $C(\Lambda_\Delta) \to C(\Lambda)$ as $\Delta \to 0$.*

Observe that the conditions do not depend heavily on the channel statistics nor on the particular structure of the constraint set Λ (other than compactness). As such, Theorem 1 holds for a wide range of channels, with examples in Section IV-C. Another interesting feature of Theorem 1 is that an assumption implying the uniqueness of the optimal input distribution is not required, which is typical in the study of the optimal input distribution.

The proof of Theorem 1 relies on Berge's maximum theorem and the theory of point-to-set maps. As such, we review the concept of point-to-set maps in Section IV-A and prove Theorem 1 in Section IV-B. Examples of channels satisfying the conditions in Theorem 1 are then provided, along with an extension to discrete and compactly supported approximations of the capacity.

A. Preliminaries

In order to establish our convergence result, we first recall useful definitions from the theory of point-to-set maps [13] and a theorem from the sensitivity analysis of optimization problems.

Let (Θ, d) and (S, d_S) be metric spaces. A point-to-set map $\Gamma : \Theta \rightrightarrows S$, also known as a correspondence, is a map from a point in Θ to a subset in S such that for each point $\theta \in \Theta$ the set $\Gamma(\theta)$ is compact. Let $s \in S$ and $\mathcal{S} \subseteq S$ and define $d_S(s, \mathcal{S}) = \inf_{\hat{s}\in\mathcal{S}} d_S(s, \hat{s})$. Furthermore, for any $\epsilon > 0$ define the ϵ-ball centered at $s \in S$ by $B_\epsilon(s) = \{\hat{s} \in S : d_S(s, \hat{s}) < \epsilon\}$.

Definition 2. *Let $\theta \in \Theta$ and $\epsilon > 0$. The ϵ-neighborhood of the set $\Gamma(\theta)$ is defined by*

$$\eta_\epsilon(\Gamma(\theta)) = \{s \in S : d_S(s, \Gamma(\theta)) < \epsilon\} = \bigcup_{\overline{s}\in\Gamma(\theta)} B_\epsilon(\overline{s}) \tag{10}$$

There are two notions of continuity for point-to-set maps, which are detailed as follows.

Definition 3. *A point-to-set map $\Gamma : \Theta \rightrightarrows S$ is upper hemicontinuous at $\theta \in \Theta$ if for all $\epsilon > 0$ there exists a $\delta > 0$ such that $d(\overline{\theta}, \theta) < \delta$ implies that $\Gamma(\overline{\theta}) \subseteq \eta_\epsilon(\Gamma(\theta))$.*

Definition 4. *A point-to-set map $\Gamma : \Theta \rightrightarrows S$ is lower hemicontinuous at $\theta \in \Theta$ if for all $\epsilon > 0$ there exists a $\delta > 0$ such that $d(\overline{\theta}, \theta) < \delta$ implies that $\Gamma(\theta) \subseteq \eta_\epsilon(\Gamma(\overline{\theta}))$.*

A point-to-set map $\Gamma : \Theta \rightrightarrows S$ is both upper and lower hemicontinuous at θ, then it is said to be *continuous*[3]. Intuitively, upper hemicontinuity can be viewed as constraining the size of expansions of the set $\Gamma(\theta)$, in the presence of small changes to θ. Conversely, lower hemicontinuity can be viewed as constraining the size of the contractions.

Although point-to-set maps are widely studied in the case the set S is $\mathbb{R}^n$, the definitions also apply to other metric spaces and can even be extended to more general topological spaces [14]. For the purposes of this paper, the set S corresponds to the set of probability measures with the Lévy-Prohorov metric.

In order to establish convergence of the discrete approximation, we will require the following theorem [15] that provides conditions ensuring continuity of the optimal value function in terms of the upper and lower hemicontinuity of the constraint map.

Theorem 2 (Berge's Maximum Theorem)**.** *Let Θ and S be two metric spaces, $\Gamma : \Theta \rightrightarrows S$ a compact-valued correspondence, and $\varphi : S \times \Theta \to \mathbb{R}$ be a continuous function on $S \times \Theta$. Define*

$$\begin{aligned} \sigma(\theta) &= \arg\max\{\varphi(s, \theta) : s \in \Gamma(\theta)\}, \ \forall\theta \in \Theta \\ \varphi^*(\theta) &= \max\{\varphi(s, \theta) : s \in \Gamma(\theta)\}, \ \forall\theta \in \Theta \end{aligned} \tag{11}$$

and assume that Γ is continuous at $\theta \in \Theta$. Then,

(i) *$\sigma : \Theta \rightrightarrows S$ is a compact-valued and upper hemicontinuous at θ.*
(ii) *$\varphi^* : \Theta \to \mathbb{R}$ is continuous at θ.*

Intuitively, Theorem 2 shows that if the constraint set varies continuously and the objective function is also continuous, then the value function is also continuous.

B. Proof of Theorem 1

$\Theta = \mathbb{R}^+$ with the Euclidean metric is a metric space. By [8], $\mathcal{P}$ is a metric space with the Lévy-Prokhorov metric,

[3] In some literature the term hemicontinuity is called semicontinuity. In this case, it is important not to confuse upper and lower hemicontinuity of point-to-set maps with upper and lower semicontinuity of functions.

denoted by ρ. By hypothesis, the mutual information $I(\mu, p_N)$ is weakly continuous on $\mathcal{P}$. We now show that $C_0 = C(\Lambda)$. First, recall the following result from [16, Theorem 6.3].

Theorem 3. *Let X be a separable metric space and $E \subseteq X$ dense in X. Then, the set of all measures whose supports are finite subsets of E is dense in the space of probability measures on X.*

Noting that $\mathbb{R}$ is separable, it then follows from Theorem 3 that $\cup_{\Delta>0}\mathcal{P}(\Delta\mathbb{Z})$ is dense in the topology of weak convergence. Hence, $C_0 = C(\Lambda)$ by the continuity of $I(\mu, p_N)$. Therefore by Theorem 2, the result is established if the point-to-set map $\Lambda_\Delta = \cup_{\overline{\Delta}>\Delta}\mathcal{P}(\overline{\Delta}\mathbb{Z}) \cap \Lambda$ is continuous at $\Delta = 0$.

To show that Λ_Δ is a continuous point-to-set map at $\Delta = 0$, first note that Λ_Δ is monotonically decreasing as Δ increases. This implies that Λ_Δ is lower hemicontinuous at $\Delta = 0$. Moreover, Λ is compact. This means that Λ is also separable and hence $\cup_{\Delta>0}\Lambda_\Delta$ is also dense in Λ. As $\cup_{\Delta>0}\Lambda_\Delta$ is dense, it follows that for any convergent sequence in Λ with limit μ_0, either μ_0 is in Λ_Δ or for all $\epsilon > 0$ there exists $\Delta > 0$ such that $\rho(\mu_0, \Lambda_\Delta) < \epsilon$. This implies that Λ_Δ is upper hemicontinuous. As such, Λ_Δ is both upper and lower hemicontinuous at $\Delta = 0$ and hence continuous at $\Delta = 0$.

Since $I(\mu, p_N)$ is weakly continuous and Λ_Δ is both upper and lower hemicontinuous at $\Delta = 0$, the result then follows by applying Theorem 2.

C. Examples

Theorem 1 provides sufficient conditions for the convergence of the discrete approximation. A key observation is that by the extreme value theorem, these conditions imply that there exists an optimal input distribution [17]. This implies that Theorem 1 holds for a wide range of channels, including the following examples:

(i) *Gaussian model* [1]
- $p_N(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-x^2/(2\sigma^2)\right),\ \sigma > 0.$
- $\Lambda = \{\mu : \mathbb{E}_\mu[X^2] \leq b\},\ b > 0.$

(ii) *Cauchy model* [3]
- $p_N(x) = \frac{1}{\pi\gamma\left(1+\left(\frac{x}{\gamma}\right)^2\right)},\ \gamma > 0.$
- $\Lambda = \{\mu : \mathbb{E}_\mu[|X|^r] \leq b\},\ b > 0.$

(iii) *Inverse Gaussian model* [4], [18]
- $p_N(x) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda(x-\gamma)^2}{2\gamma^2 x}\right),\ x > 0,\ \lambda, \gamma > 0.$
- $\Lambda = \{\mu : \mathbb{E}_\mu[X] \leq b\},\ b > 0.$

Note that these examples consist of models with a continuous optimal input (Gaussian), a discrete optimal input (Cauchy), and a conjectured mixed-type optimal input (inverse Gaussian). Further examples with characterizations of the discreteness of the optimal input can be found in [3].

D. Discrete and Compactly Supported Inputs

A natural question is whether Theorem 1 can be generalized to allow for discrete approximations with *compact support*. To this end, consider the optimization problem

$$\begin{aligned} C_{\Delta,T} = \sup_{\mu\in\mathcal{P}} \quad & I(\mu, p_N) \\ \text{subject to} \quad & \mu \in \Lambda_{\Delta,T}, \end{aligned} \tag{12}$$

where $\Lambda_{\Delta,T} = \cup_{\overline{\Delta}>\Delta}\mathcal{P}(\overline{\Delta}\mathbb{Z} \cap [-T, T]) \cap \Lambda$ with $\mathcal{P}(\overline{\Delta}\mathbb{Z} \cap [-T, T])$ denoting the set of probability measures with support $\overline{\Delta}\mathbb{Z} \cap [-T, T]$. As such, (12) corresponds to the capacity of the channel in (1), where the input is constrained to be discrete and supported on $[-T, T]$. The following result then holds.

Theorem 4. *Let Λ be a non-empty compact subset of $\mathcal{P}$. If the mutual information $I : \mathcal{P} \to \mathbb{R}$ is weakly continuous on Λ, then $C_{\Delta,T} \to C(\Lambda)$ as $T \to \infty, \Delta \to 0$.*

The proof of Theorem 4 follows that of Theorem 1. The key observation is that the set of all discrete measures on all compact subsets of $\mathbb{R}$ is also dense in $\mathcal{P}$, which follows from Theorem 3.

V. Rate of Convergence

In this section, we obtain a rate of convergence result for the discrete input approximation in (7). In particular, we establish that the following theorem holds.

Theorem 5. *Suppose that Λ is a non-empty compact subset of $\mathcal{P}$ and the mutual information $I : \mathcal{P} \to \mathbb{R}$ is weakly continuous on Λ . If $C(\Lambda) = \sup_{\mu\in\Lambda} I(\mu, p_N) < \infty$ in (5), then for all $\epsilon > 0$ there exists a $v \in \mathbb{R}$ such that*

$$C(\Lambda) - C(\Lambda_\Delta) - \epsilon \leq |v|\Delta + o(\Delta), \tag{13}$$

where $C(\Lambda_\Delta)$ is defined in (7).

If, in addition, $C(\Lambda_\Delta)$ is a convex function in Δ, then there exists $g \in \mathbb{R}$ such that

$$C(\Lambda) - C(\Lambda_\Delta) \leq |g|\Delta. \tag{14}$$

We remark that convergence rates for discrete approximations have been studied in special cases, such as Gaussian channels subject to a power constraint [6]. Also, the problem of characterizing channels with optimal discrete inputs [2], [3] can be viewed as a study of discrete approximations that have very good convergence properties.

The main feature of Theorem 5 is that it applies to a very wide range of channels, including those that do not yet have a characterization of their optimal input. This includes the additive inverse Gaussian noise channel subject to a first order moment constraint arising in molecular communications [4]. Moreover, as discussed further in Section V-B, the scaling factor v can be characterized as belonging to the (non-empty) set of *regular subgradients* of $C(\Lambda_\Delta)$.

A. Preliminaries

In general, $C(\Lambda_\Delta)$ is neither differentiable nor convex. As such, standard methods to obtain rates of convergence based on directional derivatives or the subgradient are not applicable. In more general non-smooth and non-convex settings, a useful notion is that of the regular subgradient, which is defined as follows.

Definition 5. *Consider a function $f : \mathbb{R}^n \to \mathbb{R}$ and a point $\overline{\mathbf{x}} \in \mathbb{R}^n$ with $f(\overline{\mathbf{x}})$ finite. For a vector, $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{v}$ is a regular subgradient of f at $\overline{\mathbf{x}}$, denoted by $\mathbf{v} \in \hat{\partial} f(\overline{\mathbf{x}})$, if there exists $\delta > 0$ such that for all $\mathbf{x} \in B_\delta(\overline{\mathbf{x}})$*[4]

$$f(\mathbf{x}) \geq f(\overline{\mathbf{x}}) + \mathbf{v}^T(\mathbf{x} - \overline{\mathbf{x}}) + o(|\mathbf{x} - \overline{\mathbf{x}}|). \tag{15}$$

Furthermore, recall the notion of *f-attentive convergence* [13].

Definition 6. *A sequence $(\mathbf{x}^k)_k$ is said to f-converge to $\overline{\mathbf{x}}$ if $(\mathbf{x}^k, f(\mathbf{x}^k)) \to (\overline{\mathbf{x}}, f(\overline{\mathbf{x}}))$ as $k \to \infty$, and is denoted by $\mathbf{x}^k \underset{f}{\to} \overline{\mathbf{x}}$.*

We will also require the following result, which is proven in [13, Result 8.10].

Theorem 6. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is finite and lower semicontinuous at $\overline{\mathbf{x}} \in \mathbb{R}^n$. Then, there exists a sequence $\mathbf{x}^k \underset{f}{\to} \overline{\mathbf{x}}$ with $\hat{\partial} f(\mathbf{x}^k) \neq \emptyset$ for all k.*

The main consequence of Theorem 6 is that it provides conditions for a regular subgradient to exist for a dense subset of $\mathbb{R}$.

B. Proof of Theorem 5

We now prove Theorem 5. Since $C(\Lambda_\Delta)$ is finite and continuous at $\Delta = 0$, it follows that it is lower semicontinuous at $\Delta = 0$ and by Theorem 6 that there exists a sequence $(\Delta^k)_k \to 0$ such that for each k, $\hat{\partial}(C(\Delta^k)) \neq \emptyset$. By the definition of the regular subgradient, this implies that for each k there exists a $v \in \mathbb{R}$ such that

$$C(\Lambda_\Delta) - C_{\Delta^k} \geq v\Delta + o(\Delta). \tag{16}$$

By applying the triangle inequality, it follows that

$$C_{\Delta^k} - C(\Lambda_\Delta) \leq |v|\Delta + o(\Delta). \tag{17}$$

By Theorem 1, $C(\Lambda_\Delta)$ is continuous at $\Delta = 0$. This implies that for all $\epsilon > 0$, there exists $N_\epsilon \in \mathbb{Z}$ such that for all $k > N_\epsilon$, $C(\Lambda) - C_{\Delta^k} \leq \epsilon$. It then follows that for all $k > N_\epsilon$,

$$C - C(\Lambda_\Delta) - \epsilon \leq C_{\Delta^k} - C(\Lambda_\Delta) \leq |v|\Delta + o(\Delta). \tag{18}$$

as desired.

Now suppose, in addition, that $C(\Lambda_\Delta)$ is convex. By Theorem 1, $C(\Lambda_\Delta)$ is continuous at $\Delta = 0$ and hence there exists a subgradient $g \in \mathbb{R}$ such that

$$C(\Lambda_\Delta) - C(\Lambda) \geq g\Delta. \tag{19}$$

Moreover,

$$C(\Lambda) - C(\Lambda_\Delta) \leq |g|\Delta, \tag{20}$$

which completes the proof.

[4] $B_\delta(\mathbf{x})$ is the δ-ball in $\mathbb{R}^n$ centered at $\mathbf{x}$.

VI. Discussion

The discrete approximation provides an alternative means of characterizing the capacity. This is achieved adopting the capacity sensitivity point of view and studying the effect of restricting the input to be discrete. A key observation from our study of the discrete approximation is the result for the rate of convergence in Theorem 5 holds for a very wide range of memoryless additive noise channels.

In this section, we discuss implications of the discrete approximation point of view for characterizations of the optimal input and extensions to more general point-to-point channels. We also point out how the techniques used to establish Theorem 5 can be applied to obtain scaling laws for other capacity sensitivity problems arising from constraint perturbations.

A. On the Optimal Input of (5)

A key question in the study of the capacity in (5) is whether or not the optimal input is discrete. The standard method in order to establish discreteness of the input is due to Smith [2] and is based on the behavior of the entropy density. A sufficient condition for the optimal input to be discrete has recently been established by generalizing Smith's technique in [3]. Here, we note another sufficient condition obtained by considering the form of the function $C(\Lambda_\Delta) : \mathbb{R}_+ \to \mathbb{R}_+$.

Proposition 1. *Suppose that the optimal input probability measure for the problem in (5) exists and is unique. If $C(\Lambda_\Delta)$ from (7) is convex in Δ, then the optimal input is not discrete.*

Proof. First, observe that $C(\Lambda_\Delta)$ is non-increasing in Δ. Moreover if the optimal input is discrete, then there exists $\Delta_0 > 0$ such that $C(\Lambda_\Delta) = C(\Lambda)$ for all $\Delta \leq \Delta_0$. Observe that the epigraph of $C(\Lambda_\Delta)$ in this case is non-convex, which in turn implies that $C(\Lambda_\Delta)$ is non-convex. □

At this point, we do not have a characterization of channels for which $C(\Lambda_\Delta)$ is in fact convex. Nevertheless, Proposition 1 suggests that the behavior of $C(\Lambda_\Delta)$ may provide further insight into the behavior of the optimal input distribution.

B. General Scalar Point-to-Point Memoryless Channels

A key observation from the conditions in Theorem 1 and Theorem 5 is that they do not depend directly on the assumption that the channel is linear and additive with the noise distribution admitting a density. As such, these theorems can be applied to more general point-to-point memoryless channels under the assumption that it is possible to establish that the mutual information is continuous on the constraint set Λ in these settings. However, this can be challenging as the output Y may no longer admit a density and the mutual information given an input X must be written in the general form developed by Pinsker [19], given by

$$I(X;Y) = \sup \sum_{i,j} P_{X,Y}(E_i \times F_j) \log \frac{P_{X,Y}(E_i \times F_j)}{P_X(E_i)P_Y(F_j)}, \tag{21}$$

where $P_{X,Y}$ is the joint distributions, P_X, P_Y the marginal distributions, and the supremum is taken over all partitions $\{E_i\}$ of $\mathrm{supp}(X)$ and $\{F_j\}$ of $\mathrm{supp}(Y)$.

C. General Constraint Perturbations

In order to develop results on the rate of convergence of discrete approximations, the problem is formulated in terms of the capacity sensitivity in the case of a particular class of constraint perturbation. It is also possible to consider other classes of constraint perturbations by using similar techniques. Of interest is the study of capacity sensitivity in the context of channels subject to constraints of the form

$$\Lambda(b) = \{\mu : \mathbb{E}_\mu[f(|X|)] \leq b\}, \tag{22}$$

where f is positive, non-decreasing and lower semicontinuous. That is, the capacity $C(b)$ is given by

$$\begin{aligned} &\sup_{\mu \in \mathcal{P}} && I(\mu, p_N) \\ &\text{subject to} && \mu \in \Lambda(b). \end{aligned} \tag{23}$$

For $\bar{b}, \tilde{b} \in \mathbb{R}_+$, we seek to establish a bound on the capacity sensitivity $|C(\bar{b}) - C(\tilde{b})|$. From [7], we have the following result.

Theorem 7. *Let $\bar{b} \in \mathbb{R}_+$ and suppose that the following conditions hold:*

(i) *$\Lambda(\bar{b})$ in (22) is non-empty and compact.*

(ii) *$I(\mu, p_N)$ is weakly continuous on $\Lambda(\bar{b})$.*

Then, $C(b)$ in (23) is continuous at $\bar{b}$.

With Theorem 7 in hand, it is possible to obtain rates of convergence in an analogous fashion to the proof of Theorem 5. In particular, if $C(\bar{b}) < \infty$, by Theorem 6 it then follows that there exists a sequence $(b^k)_k \to \bar{b}$ such that for each k, $\hat{\partial}(C(b^k)) \neq \emptyset$. By the definition of the regular subgradient, for each k there exists a $v \in \mathbb{R}$ such that

$$C(\tilde{b}) - C(b^k) \geq v(\tilde{b} - b^k) + o(|\tilde{b} - b^k|). \tag{24}$$

If $\bar{b} \geq \tilde{b}$, by the continuity of $C(b)$ at $\bar{b}$, it then follows that for all $\epsilon > 0$ sufficiently small

$$|C(\bar{b}) - C(\tilde{b}) - \epsilon| \leq |v||\bar{b} - \tilde{b}| + o(|\bar{b} - \tilde{b}|), \tag{25}$$

which is consistent with [7, Theorem 2], with the difference that the convergence is to a point *arbitrarily close* to $C(\bar{b})$. Nevertheless, the proof technique is considerably simpler and is of the same form as that of Theorem 5, which suggests that it is useful for the study of general constraint perturbations.

VII. Conclusion

In this paper, discrete approximations of the capacity for continuous channels were studied. In particular, the rate of convergence was analyzed within the capacity sensitivity framework. The key conclusion is that for a very wide range of channels, the approximation converges to arbitrarily close to the capacity of the continuous channel at rate $O(\Delta)$, where Δ is the discretization level of the approximation. This result was obtained by exploiting results from the theory of point-to-set maps and techniques from nonsmooth analysis of optimization problems.

The work in this paper suggests a number of further avenues of investigation. For instance, the conditions on the channel require that the mutual information is continuous on the constraint set. It remains an open question to establish this continuity condition in the general setting of point-to-point channels with mixed-type outputs. Another aspect of this work is the emphasis on the behavior of the capacity of the discrete approximation $C(\Lambda_\Delta)$. An interesting direction is therefore whether new conditions for discreteness of the optimal input distribution can be obtained from properties of $C(\Lambda_\Delta)$. Finally, rates of convergence for general constraint perturbations can be established via the regular subgradient for finite dimensional constraint parameters. An open question is to find rates of convergence in the infinite dimensional case.

References

[1] C. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423,623–656, Jul., Oct. 1948.

[2] J. Smith, "The information capacity of amplitude- and variance-constrained scalar Gaussian channels," *Information and Control*, vol. 18, no. 3, pp. 203–219, Apr. 1971.

[3] J. Fahs and I. Abou-Faycal, "On properties of the support of capacity-achieving distributions for additive noise channel models with input cost constraints," *To appear in IEEE Transactions on Information Theory*.

[4] K. Srinivas, A. Eckford, and R. Adve, "Molecular communication in fluid media: The additive inverse Gaussian noise channel," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4678–4692, Jul. 2012.

[5] R. McEliece, *The Theory of Information and Coding*. Cambridge University Press, 2002.

[6] Y. Wu and S. Verdú, "The impact of constellation cardinality on Gaussian channel capacity," in *Proc. of the 48th Annu. Allerton Conf. Commun., Control, Comput.*, Monticello, IL, Sep. 2010.

[7] M. Egan and S. Perlaza, "Capacity sensitivity in additive non-gaussian noise channels," in *Proc. IEEE International Symposium on Information Theory*, Aachen, Germany, Jun. 2017.

[8] P. Billingsley, *Convergence of Probability Measures*, 2nd ed. New York, NY: John Wiley and Sons, 1999.

[9] J. Fahs and I. Abou-Faycal, "On the finiteness of the capacity of continuous channels," *IEEE Transactions on Communications*, vol. 54, no. 1, pp. 166–173, Jan. 2016.

[10] L. Ozarow and A. Wyner, "On the capacity of the Gaussian channel with a finite number of input levels," *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1426–1428, Nov.

[11] A. Dytso, M. Goldenbaum, H. Poor, and S. Shamai, "A generalized Ozarow-Wyner capacity bound with applications," in *Proc. IEEE International Symposium on Information Theory*, Aachen, Germany, Jun. 2017.

[12] M. Pinsker, V. Prelov, and S. Verdú, "Sensitivity of channel capacity," *IEEE Transactions on Information Theory*, vol. 41, no. 6, pp. 1877–1888, Nov. 1995.

[13] R. Rockafellar and R. Wets, *Variational Analysis*. Berlin Heidelberg: Springer-Verlag, 1997.

[14] C. Berge, *Topological Spaces*. Mineola, NY: Dover, 1963.

[15] E. Ok, *Real Analysis with Economic Applications*. Princeton, N.J.: Princeton University Press, 2007.

[16] K. Parthasarathy, *Probability Measures on Metric Spaces*. New York, NY: Academic Press, 1967.

[17] D. Luenberger, *Optimization By Vector Space Methods*. New York, NY: Wiley, 1969.

[18] H. Li, S. Moser, and D. Guo, "Capacity of the memoryless additive inverse gaussian noise channel," *IEEE Journal on Selected Areas on Communications*, vol. 32, no. 12, pp. 2315–2329, Dec. 2014.

[19] M. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco, C.A.: Holden-Day, 1964.