Data-Driven Estimation of Capacity Upper Bounds

Christian Häger, Member, IEEE, and Erik Agrell, Fellow, IEEE

Abstract — We consider the problem of estimating an upper bound on the capacity of a memoryless channel with unknown channel law and continuous output alphabet. A novel data-driven algorithm is proposed that exploits the dual representation of capacity where the maximization over the input distribution is replaced with a minimization over a reference distribution on the channel output. To efficiently compute the required divergence maximization between the conditional channel and the reference distribution, we use a modified mutual information neural estimator that takes the channel input as an additional parameter. We numerically evaluate our approach on different memoryless channels and show empirically that the estimated upper bounds closely converge to the channel capacity or to best-known lower bounds.

Index Terms — Autoencoders, channel capacity, divergence estimation, duality, dual capacity representation, mutual information neural estimation, neural networks, upper capacity bounds.

I. INTRODUCTION

The capacity of a communication channel is the maximum rate that can be reliably transmitted [1]. Even though capacity is of fundamental importance for both theory and practice, exact analytical expressions are only available in relatively few cases. If the underlying channel law is known, numerical techniques can be used to approximately compute capacity such as the well-known Blahut–Arimoto algorithm [2], [3] and its many variations, see, e.g., [4] and references therein.

Recently, there has been significant interest in developing capacity estimation algorithms based on machine learning [5], [6], [7], [8], [9], [10], [11]. These approaches have their origin in the seminal work [12], where the authors propose to reinterpret the communication problem as a reconstruction task using parameterized transmitters and receivers, similar to autoencoders (AEs) in machine learning. It can be shown that the cross-entropy minimization commonly used to train AEs maximizes a lower bound on mutual information, whereas the transmitter optimization can be regarded as shaping a discrete input distribution. Using this approach, tight lower bounds on the capacity of a nonlinear phase noise (NLPN) channel were for example obtained in [5].

A disadvantage of the AE approach is that it requires a differentiable channel model to compute gradients for the transmitter optimization. To address this problem, [6] proposes to use the sample-based mutual information neural estimation (MINE) technique from [13] and train the AE transmitter based on the (differentiable) MINE. Related approaches were subsequently proposed for more general channels that may include feedback and/or memory in [7] and for memoryless multiple-access channels in [9]. Moreover, a hybrid approach that regularizes the cross-entropy-based AE training using MINE is proposed in [10]. Comparisons of various sample-based mutual information estimators similar to MINE can be found in [8] and [11].

All of the above learning-based approaches target the estimation of lower capacity bounds using the conventional maximization of mutual information (or directed information in [7]) over the input distribution, see (1) below. In this letter, we follow a different path and consider the problem of estimating an upper capacity bound by exploiting the dual representation of channel capacity, which is described in Sec. II. Our work relies on a variation of MINE for estimating the divergence between the conditional channel and a given reference distribution. Compared to the conventional Blahut–Arimoto algorithm, the main advantage of our approach and similar works on neural capacity estimation in [5], [6], [7], [8], [9], [10], [11] stems from the fact that no knowledge about the underlying channel law is required. As such, the resulting estimators can be used in settings where the channel is only accessible via input–output samples (e.g., in experimental setups) and the precise channel law is unknown.

Notation: Random variables are denoted by upper-case letters (e.g., $X$), realizations by lower-case letters (e.g., $x$), and sets by calligraphic letters (e.g., $\mathcal{X}$). The probability distribution of a random variable $X$ is denoted by $f_X$. Expectation is denoted by $\mathbb{E}[\cdot]$, mutual information by $I(\cdot;\cdot)$, and Kullback–Leibler divergence by $D(\cdot||\cdot)$. For an integer $N$, we define the set $[N] = \{1, \ldots, N\}$.

II. DUAL REPRESENTATION OF CHANNEL CAPACITY

We consider a memoryless channel with input $X \in \mathcal{X}$ and output $Y \in \mathcal{Y}$. The channel law conditioned on a particular input $x$ is denoted by $f_{Y|X=x}(y)$. In general, the input is assumed to be constrained by a cost function $c : \mathcal{X} \to \mathbb{R}_{\geq 0}$. The capacity of such a channel is

$$C = \max_{f_X : \mathbb{E}[f_X I(X;Y)] \leq P} I(X;Y),$$

3See also the earlier work in [14] based on reinforcement learning, which, however, requires knowledge about the channel law.

2To keep the notation simple, we focus on scalar channels. However, our approach generalizes to (block-wise) memoryless channels where the input and outputs are (possibly complex-valued) random vectors, see Sec. VI-C.
where \( P \) denotes the maximum average cost. In the following, we work with the dual representation [15, p. 142]

\[
C = \min_{\gamma \geq 0} \left[ F(\gamma) + \gamma P \right],
\]

where

\[
F(\gamma) = \min_{q_Y} \max_{x \in \mathcal{X}} \left[ D(f_Y|X=x||q_Y) - \gamma c(x) \right].
\]

The minimization in (3) is over all distributions \( q_Y \) on \( \mathcal{Y} \). Note that any fixed choice for the reference distribution \( q_Y \) leads to an upper bound in (3) and hence on the capacity (2).

According to [16], the above dual approach first originated in [17] and was further developed in [15], [18], and [19]. Recent work exploiting this approach has mostly focused on choosing \( q_Y \) to obtain a tractable analytical expression for the resulting upper bound, see, e.g., [16], [19]. In this letter, we will instead use (3) as a blueprint for an iterative optimization procedure that alternates between training a divergence estimator (see Sec. III) and the reference distribution \( q_Y \) (see Sec. IV) based on the obtained estimator. The resulting algorithm is described in Sec. V.

### III. DATA-DRIVEN DIVERGENCE ESTIMATION

We use the approach proposed in [13] to estimate the divergence term in (3). This approach is based on the Donsker–Varadhan (DV) representation [13, Th. 1]

\[
D(f_Y|X=x||q_Y) = \sup_{T \in \mathcal{T}} \mathbb{E}_{f_Y|X=x} \left[ T(Y) \right] - \log \left( \mathbb{E}_{q_Y} [e^{T(Y)}] \right),
\]

where the supremum is over all functions \( T: \mathcal{Y} \to \mathbb{R} \) such that the expectations in (4) are finite. The idea in [13] is to approximate the class of functions \( \mathcal{T} \) using a neural network (NN) \( T_\theta: \mathcal{Y} \to \mathbb{R} \), where \( \theta \) are the NN parameters. \( T_\theta \) is also referred to as the statistics network. For a fixed set of parameters \( \theta \), the resulting estimator is

\[
\hat{D}_\theta = \frac{1}{B} \sum_{i=1}^{B} T_\theta(y^{(i)}) - \log \left( \frac{1}{B} \sum_{i=1}^{B} e^{T_\theta(\tilde{y}^{(i)})} \right),
\]

where \( B \) is the batch size, \( y^{(1)}, \ldots, y^{(B)} \sim f_Y|X=x \), and \( \tilde{y}^{(i)}, \ldots, \tilde{y}^{(B)} \sim q_Y \). This estimator can be iteratively trained by running gradient ascent on (5), see [13, Alg. 1] for details.

**Remark:** We use the above estimator mainly for its simplicity and the fact that it empirically tends to work well (see, e.g., [11]). However, one should be aware that using the lower-bound estimator (5) does not guarantee to result in a true upper bound on capacity, which would in principle require a full optimization over \( \mathcal{T} \). Moreover, the gradients resulting from (5) are biased [13] and the estimator has high variance, especially if the true underlying divergence is large [20]. We comment on potential alternative approaches in Sec. VII.

Note that the estimator (5) assumes a fixed channel input \( x \). A different statistics network \( T_\theta \) would thus be required for each input to evaluate the maximization in (3) over the input alphabet. However, this quickly becomes infeasible if the size of the input alphabet is large or infinite. To circumvent this problem, we propose a modified version of (5) where the input \( x \) is taken as an additional input to the statistics network \( T_\theta \), i.e., \( T_\theta: \mathcal{Y} \times \mathcal{X} \to \mathbb{R} \). The resulting modified modified estimator is denoted by

\[
\hat{D}_\theta(x) = \frac{1}{B} \sum_{i=1}^{B} T_\theta(y^{(i)}, x) - \log \left( \frac{1}{B} \sum_{i=1}^{B} e^{T_\theta(\tilde{y}^{(i)}, x)} \right).
\]

This formulation allows us to train a single statistics network that works well for a range of different channel inputs. In particular, this can be accomplished by jointly considering multiple inputs \( \{x^{(1)}, \ldots, x^{(N_t)}\} = \mathcal{X}_t \subseteq \mathcal{X} \) and running gradient descent on an average loss \( \frac{1}{N_t} \sum_{i=1}^{N_t} \hat{D}_\theta(x^{(i)}, \cdot) \). However, averaging the logarithms in (6) leads to a geometric mean \( \frac{1}{N_t} \sum_{i=1}^{N_t} \log a_{ij} = \log((\prod_{i=1}^{N_t} a_{ij})^{1/N_t}) \). We found that replacing this geometric mean with an arithmetic mean gives a numerically more stable training behavior, resulting in

\[
L_\theta = -\frac{1}{N_t B} \sum_{j=1}^{N_t} \sum_{i=1}^{B} T_\theta(y^{(i,j)}, x^{(j)}) + \log \left( \frac{1}{N_t B} \sum_{j=1}^{N_t} \sum_{i=1}^{B} e^{T_\theta(\tilde{y}^{(i,j)}, x^{(j)})} \right),
\]

where \( y^{(1),j}, \ldots, y^{(B),j} \sim f_Y|X=x^{(j)} \) and \( \tilde{y}^{(1),j}, \ldots, \tilde{y}^{(B),j} \sim q_Y \). Note that (7) has the same functional form as the Monte Carlo approximation of the DV representation of \( -I(X; Y) \), assuming that the input is uniformly distributed over \( \mathcal{X}_t \).

**Example:** Assume that \( f_Y|X=x \) and \( q_Y \) correspond to \( \mathcal{N}(x, 1) \) and \( \mathcal{N}(1, 2) \), respectively. To optimize the parameters \( \theta \), we set \( \mathcal{X}_t = \{-1, -0.8, \ldots, 1\} \) and train the statistics network\(^3\) using the loss (7). Fig. 1 compares the accuracy of the resulting estimator \( \hat{D}_\theta(x) \) to the true divergence as a function of \( x \), where \( \theta^* \) refers to the optimized parameters. Note that \( \hat{D}_\theta(x) \) generalizes well even to values of \( x \) that were not seen during training, as illustrated by the red triangles.

### IV. REPRESENTATION OF THE REFERENCE DISTRIBUTION

To allow for the gradient-based optimization of the reference distribution \( q_Y \), two different approaches are described in the following for generating the samples \( y^{(1),j}, \ldots, y^{(B),j} \sim q_Y \) in (7) using NNs. The corresponding block diagrams are shown in Fig. 2. Similar to [7], we refer to the resulting transformation as the neural distribution transformer (NDT).

In the first approach (Fig. 2, top), the NDT consists of an NN \( f_\tau \) with parameters \( \tau \), which is then followed by a batch-wise normalization procedure and transmission over the channel. Thus, this approach generates channel inputs as an intermediate step. The normalization procedure enforces the

\[^3\] The NN architecture and all other training hyperparameters for this example are the same as for case (i) in Tab. 1 below.
average cost constraint\textsuperscript{4} via
\[
\hat{s}^{(i,j)} = \frac{1}{c} \sum_{i=1}^{b} c(s^{(i,j)}),
\]
where \(s^{(i,j)} = f_{\tau}(z^{(i,j)}), i \in [B], j \in [N_i] \) is the NN output and \(z^{(i,j)} \in \mathbb{R}^l \) is a random vector sampled from a fixed latent probability distribution \(f_z\). Note that (8) implicitly assumes that the cost function distributes over division, i.e., \(c(a/b) = c(a)/c(b)\), which is the case for all cost functions considered in this letter. The above approach ensures that \(q_y\) is a valid output distribution for the channel under consideration, given the cost constraint. However, it should be noted that it requires a differentiable channel model in order to compute gradients with respect to \(\tau\).

In the second approach (Fig. 2, bottom), the NDT directly generates samples from the channel output alphabet and simply consists of an NN \(f_{\tau}\) as before but without any additional post-processing, i.e., \(\tilde{y}^{(i,j)} = f_{\tau}(z^{(i,j)})\). While this approach does not necessarily ensure that \(q_y\) is a valid output distribution for cost-constrained channels (which is not required to obtain an upper bound in (3)), it is more universal and can be used even if the channel is only accessible as a black box through input–output samples (e.g., in an experimental setting). On the other hand, we found that this representation typically requires more training steps when optimizing the NN parameters \(\tau\).

\textbf{V. PROPOSED ALGORITHM}

The proposed capacity estimation algorithm is detailed in Algorithm 1. It alternates between training the statistics network \(T_{\theta}\) (lines 3–5) and the NDT network \(f_{\tau}\) (lines 6–8) for a total of \(M\) iterations. The latter optimizes the reference distribution \(q_y\) based on the loss function (cf. (3))

\[
\hat{F}_{\tau}(\gamma) = \max_{j \in [N_d]} \left[ D_{\text{KL}}(x^{(j)} \| q_{Y_{\text{ref}}}) + \gamma c(x^{(j)}) \right],
\]

\[
= \max_{j \in [N_d]} \left[ \frac{1}{B} \sum_{i=1}^{B} T_{\theta}(y_{\text{ref}}^{(i,j)}, x^{(j)}) - \log \left( \frac{1}{B} \sum_{i=1}^{B} e^{T_{\theta}(\tilde{y}^{(i,j)}, x^{(j)})} \right) - \gamma c(x^{(j)}) \right],
\]

where the dependence of \(\hat{F}_{\tau}(\gamma)\) on the parameters \(\tau\) is implicit through the samples \(\tilde{y}^{(i,j)}\) generated by the NDT. Compared to (3), the outer minimization over \(q_y\) is encapsulated in the NN parameters \(\tau\), which are optimized via gradient descent in Algorithm 1.

\textbf{VI. NUMERICAL RESULTS}

In this section, we numerically evaluate the proposed approach.\textsuperscript{6} Note that for cost-constrained channels, Algorithm 1 estimates the capacity in (3) as a function of the Lagrange multiplier \(\gamma \geq 0\). In this case, we use a golden-section search to solve the one-dimensional minimization over \(\gamma\) in (2).

For all considered cases, we use fully-connected NNs with rectified linear unit activation functions in the hidden layers to represent both the NDT network \(f_{\tau}\) and the statistics network \(T_{\theta}\). The NN parameters are summarized in Tab. I. The number of input neurons for \(f_{\tau}\) depends on the latent distribution, which we assume to be a multivariate Gaussian distribution \(\mathcal{N}(0, I_l)\) with \(l = 50\). We found that the results are relatively insensitive to the choice of \(l\), which mainly affects the initial distribution for \(q_y\) before training. Moreover, while we have verified both NDT approaches discussed in Sec. IV, the following numerical results use the first approach (Fig. 2, top) since all channel models below are differentiable.

For the gradient-based optimization steps in Algorithm 1 (lines 5 and 8), we employ the Adam optimizer [21] with learning rate \(\beta = 10^{-3}\) and batch size \(B = 20000\). Lastly, we always pretrain the statistics network \(T_{\theta}\) by running 200 initial iterations of lines 3–5 in Algorithms 1, which we found to improve training convergence.

\textsuperscript{4}Note that this procedure enforces the average cost constraint with equality, even if the inequality constraint was satisfied before the normalization.

\textsuperscript{5}In this case, the data generated in line 3 can be reused in line 6.

\textsuperscript{6}The source code to reproduce all numerical results in this letter is available at \url{https://github.com/chaeger/upper_capacity_bounds}. 

---

\textbf{Algorithm 1} Estimation of Capacity Upper Bounds.

\begin{algorithm}
\begin{algorithmic}
\State \textbf{Inputs:} number of iterations \(M\), batch size \(B\), learning rate \(\beta\), Lagrange multiplier \(\gamma\) (for cost-constrained channels), input sets \(\mathcal{X}_l\) (see Sec. III) and \(\mathcal{X}_d\) (see Sec. V)
\For {\(l = 1, 2, \ldots, M\)}
\State \(\forall x^{(j)} \in \mathcal{X}_l, i \in [B]: \) generate \(y_{\text{ref}}^{(i,j)} \sim f_{\tau}[x = x^{(j)}, y_{\text{ref}}^{(i,j)}] \sim q_{Y}\)
\State \(L_{\theta} \leftarrow \text{compute average loss according to (7)}\)
\State \(\theta \leftarrow \theta - \beta \nabla_{\theta} L_{\theta} \quad /\!\!/ \quad \text{update statistics network }\)
\State \(\forall x^{(j)} \in \mathcal{X}_d, i \in [B]: \) generate \(y^{(i,j)} \sim f_{\tau}[x = x^{(j)}, y^{(i,j)}] \sim q_{Y}\)
\State \(\hat{F}_{\tau}(\gamma) \leftarrow \text{estimate upper bound according to (9)}\)
\State \(\tau \leftarrow \tau - \beta \nabla_{\tau} \hat{F}_{\tau}(\gamma) \quad /\!\!/ \quad \text{update NDT network}\)
\EndFor
\State \textbf{return} \(\hat{F}_{\tau}(\gamma)\)
\end{algorithmic}
\end{algorithm}

---

Note that this procedure enforces the average cost constraint with equality, even if the inequality constraint was satisfied before the normalization.
ensure that the channel input satisfies the amplitude constraint. For the numerical estimation, we set $\alpha = 0.4$ (cf. [24, Fig. 2]). Results are shown in Fig. 4 after $M = 500$ iterations, where we compare to the upper and lower capacity bounds developed in [24], see in particular [24, Fig. 2]. The gap of the estimated upper bound with respect to the lower bound is due to the fact that the latter is not tight. Indeed, to verify that the channel capacity is close to our estimated upper bound, we used the Blahut–Arimoto algorithm for cost-constrained channels [15, p. 140] (black circles). We also note that tighter analytical upper bounds can potentially be obtained by extending the methodology in [23] to this channel model.

### B. Nonlinear Phase Noise Channel

Lastly, we consider an NLPN channel for coherent optical communication which is based on a split-step solution of the nonlinear Schrödinger equation without dispersive effects. The resulting complex-valued channel is defined by the recursion

$$X_{k+1} = X_k e^{\gamma L |X_k|^2/K} + N_{k+1}, \quad 0 \leq k \leq K,$$

(10) where $X = X_0 \in \mathbb{C}$ is the input, $Y = X_K \in \mathbb{C}$ is the output, $N_{k+1} \sim \mathcal{CN}(0, \sigma^2/K)$, $\sigma^2$ is the total noise power, $\gamma$ is a nonlinearity parameter, $L$ is the transmission distance, and $c(x) = |x|^2$. This channel has a long history in terms of capacity analysis, see [25], [26], [5], [27], [28] and references therein. Here, we use the same parameters as in [5], [27], i.e., $K = 50$, $\sigma^2 = -21.3$ dBm, $\gamma = 1.27$ rad/km/W, and $L = 5000$ km.

For the numerical estimation, we consider a renormalized version of (10), where $X_{k}' = X_k'/\sqrt{P}$. The input space of the renormalized channel is discretized using 9 uniformly spaced grid points in the interval $[-1.75, 1.75]$ for both the real and imaginary part, i.e., $N_d = 81$ total discretization points. Separating the real and imaginary parts of the channel input and output, respectively, the number of NDT output neurons is defined by

$$Y = X + Z,$$
is increased to 2 and the number of input neurons of the statistics network to 4. Fig. 5 shows the obtained results after $M = 2500$ iterations as a function of $P$ (see the top axis for a conversion to SNR $= P/\sigma^2$). It can be seen that the estimated upper bound closely follows the lower bound in [27], which is based on a Gaussian input distribution.

VII. DISCUSSION AND FUTURE WORK

We have proposed a novel data-driven approach for estimating upper bounds on channel capacity. Similar to recent work in [6], [7], the proposed algorithm relies on the DV representation for estimating divergence, with the consequence that the resulting estimates are neither true upper nor lower bounds for a finite sample size [29]. Even assuming an infinite sample size, one cannot guarantee that the resulting estimates are true upper bounds on capacity, which would in principle require a full optimization over the function class $T$ (cf. (4)). It is therefore important to properly choose the NN and training parameters. For example, more training iterations were required for the NLPN channel compared to the other cases to ensure convergence of the NDT and statistics networks.

An overview of potential alternative divergence estimation approaches can be found in [8] and [11]. Moreover, [9] recently proposed an approach to obtain outer bounds on the achievable rate region of memoryless multiple-access channels by exploiting the upper bounds based on $f$-divergence inequalities from [30]. However, the histogram-based approach to numerically evaluate these bounds in [9] does not directly lead to a differentiable loss function. Therefore, additional modifications (e.g., based on ideas similar to [31]) would be required to be able to use such inequalities in conjunction with the NDT optimization in Algorithm 1.

Lastly, another reason why the proposed approach does not necessarily compute true upper bounds for continuous-input channels is that the maximization over $x$ is only done approximately via discretization. To ensure that the capacity of the resulting input-discretized channel is close to that of the original channel, one approach is to successively increase the number of discretization points $N_x$ until convergence. For future work, it may be interesting to develop native estimation approaches that do not require such a discretization. This could be done, for example, by considering an auxiliary distribution over the input space and casting the maximization in (3) as an optimization problem over this auxiliary distribution. Similar to the NDT, the auxiliary distribution could then again be parameterized using an NN.

REFERENCES

[1] C. E. Shannon, “A mathematical theory of communication,” Bell Syst. Tech. J., vol. 27, no. 3, pp. 379–423, Jul. 1948.
[2] R. E. Blahut, “Computation of channel capacity and rate-distortion functions,” IEEE Trans. Inf. Theory, vol. IT-18, no. 4, pp. 460–473, Jul. 1972.
[3] S. Arimoto, “An algorithm for computing the capacity of arbitrary discrete memoryless channels,” IEEE Trans. Inf. Theory, vol. IT-18, no. 1, pp. 14–20, Jan. 1972.
[4] I. Naiss and H. Permuter, “Extension of the Blahut–Arimoto algorithm for maximizing directed information,” IEEE Trans. Inf. Theory, vol. 59, no. 1, pp. 204–222, Jan. 2013.
[5] S. Li et al., “Achievable information rates for nonlinear fiber communication via end-to-end autoencoder learning,” in Proc. Eur. Conf. Opt. Commun. (ECOC), Rome, Italy, Sep. 2018, pp. 1–3.
[6] R. Fritschek et al., “Deep learning for channel coding via neural mutual information estimation,” in Proc. IEEE 20th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC), Cannes, France, Jul. 2019, pp. 1–5.
[7] Z. Aharoni et al., “Capacity of continuous channels with memory via directed information neural estimator,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Los Angeles, CA, USA, Jun. 2020, pp. 2014–2019.
[8] R. Fritschek et al., “Neural mutual information estimation for channel coding: State-of-the-art estimators, analysis, and performance comparison,” in Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC), Atlanta, GA, USA, May 2020, pp. 1–5.
[9] F. Mirkarimi and N. Farsad, “Neural computation of capacity region of memoryless multiple access channels,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Melbourne, VIC, Australia, Jul. 2021, pp. 2066–2071.
[10] N. A. Letizia and A. M. Tonello, “Capacity-driven autoencoders for communications,” IEEE Open J. Commun. Soc., vol. 2, pp. 1366–1378, 2021.
[11] F. Mirkarimi et al., “Neural capacity estimators: How reliable are they?” Nov. 2021, arXiv:2111.07401.
[12] T. O’Shea and J. Hoydis, “An introduction to deep learning for the physical layer,” IEEE Trans. Cogn. Commun. Netw., vol. 3, no. 4, pp. 563–575, Dec. 2017.
[13] M. I. Belghazi et al., “Mutual information neural estimation,” in Proc. Int. Conf. Mach. Learn. (ICML), Stockholm, Sweden, 2018, pp. 531–540.
[14] Z. Aharoni et al., “Computing the feedback capacity of finite state channels using reinforcement learning,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Jul. 2019, pp. 837–841.
[15] I. Csiszár and J. Körner, Information Theory: Coding Theorems for Discrete Memoryless Systems, New York, NY, USA: Academic, 1981.
[16] A. Thangaraj et al., “Capacity bounds for discrete-time, amplitude-constrained, additive white Gaussian noise channels,” IEEE Trans. Inf. Theory, vol. 63, no. 7, pp. 4172–4182, Jul. 2017.
[17] F. Topsøe, “An information theoretical identity and a problem involving capacity,” Studia Sci. Math. Hungarica, vol. 2, pp. 291–292, Jan. 1967.
[18] J. H. B. Kemperman, “On the Shannon capacity of an arbitrary channel,” Indagationes Math., vol. 77, no. 2, pp. 101–115, Jan. 1974.
[19] A. Lapidoth and S. M. Moser, “Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels,” IEEE Trans. Inf. Theory, vol. 49, no. 10, pp. 2426–2467, Oct. 2003.
[20] D. McAllester and K. Stratos, “Formal limits on the measurement of mutual information,” in Proc. Int. Conf. Artif. Intell. Statist. (AISTATS), 2020, pp. 875–884.
[21] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, 2015, pp. 1–15.
[22] A. McKellips, “Simple tight bounds on capacity for the peak-limited discrete-time channel,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Chicago, IL, USA, Jun./Jul. 2004, p. 348.
[23] B. Rassouli and B. Clerckx, “An upper bound for the capacity of amplitude-constrained scalar AWGN channel,” IEEE Commun. Lett., vol. 20, no. 10, pp. 1924–1926, Oct. 2016.
[24] A. Lapidoth et al., “On the capacity of free-space optical intensity channels,” IEEE Trans. Inf. Theory, vol. 55, no. 10, pp. 4449–4461, Oct. 2009.
[25] K. S. Turitsyn et al., “Information capacity of optical fiber channels with zero average dispersion,” Phys. Rev. Lett., vol. 91, Nov. 2003, Art. no. 203901.
[26] M. I. Yousef and F. R. Kschischang, “On the per-sample capacity of nondispersive optical fibers,” IEEE Trans. Inf. Theory, vol. 57, no. 11, pp. 7522–7541, Nov. 2011.
[27] K. Keykhoosavari et al., “Accuracy assessment of nondispersive optical perturbative models through capacity analysis,” Entropy, vol. 21, no. 8, pp. 1–19, Aug. 2019.
[28] A. V. Reznichenko and I. S. Terekhov, “Path integral approach to nondispersive optical fiber communication channel,” Entropy, vol. 22, no. 1, pp. 1–30, May 2020.
[29] B. Poole et al., “On variational bounds of mutual information,” in Proc. Int. Conf. Mach. Learn. (ICML), Long Beach, CA, USA, 2019, pp. 5171–5180.
[30] I. Sason and S. Verdú, “$f$-Divergence inequalities,” IEEE Trans. Inf. Theory, vol. 62, no. 11, pp. 5973–6006, Nov. 2016.
[31] E. Ustinova and V. Lempitsky, “Learning deep embeddings with histogram loss,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), Barcelona, Spain, 2016, pp. 1–6.