Strong Data Processing Inequalities for Input Constrained Additive Noise Channels

Flavio P. Calmon  Yury Polyanskiy  Yihong Wu *

December 21, 2015

Abstract

This paper quantifies the intuitive observation that adding noise reduces available information by means of non-linear strong data processing inequalities. Consider the random variables $W \to X \to Y$ forming a Markov chain, where $Y = X + Z$ with $X$ and $Z$ real-valued, independent and $X$ bounded in $L_p$-norm. It is shown that $I(W; Y) \leq F_I(I(W; X))$ with $F_I(t) < t$ whenever $t > 0$, if and only if $Z$ has a density whose support is not disjoint from any translate of itself.

A related question is to characterize for what couplings $(W, X)$ the mutual information $I(W; Y)$ is close to maximum possible. To that end we show that in order to saturate the channel, i.e. for $I(W; Y)$ to approach capacity, it is mandatory that $I(W; X) \to \infty$ (under suitable conditions on the channel). A key ingredient for this result is a deconvolution lemma which shows that post-convolution total variation distance bounds the pre-convolution Kolmogorov-Smirnov distance.

Explicit bounds are provided for the special case of the additive Gaussian noise channel with quadratic cost constraint. These bounds are shown to be order-optimal. For this case simplified proofs are provided leveraging Gaussian-specific tools such as the connection between information and estimation (I-MMSE) and Talagrand's information-transportation inequality.

*F. P. Calmon is with the IBM T.J. Watson Research Center, Yorktown Heights, NY, 10601. E-mail: fdcalmon@us.ibm.com. Y. Polyanskiy is with the Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 02139, USA. E-mail: yp@mit.edu. Y. Wu is with the Department of Electrical Engineering and Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA. E-mail: yihongwu@illinois.edu.

This work is supported in part by the National Science Foundation (NSF) CAREER award under Grant CCF-12-53205, the NSF Grant IIS-1447879 and CCF-1423088 and by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under Grant CCF-09-39370. This paper was presented in part at the 2015 IEEE International Symposium on Information Theory.
1 Introduction

Strong data-processing inequalities (SDPIs) quantify the decrease of mutual information under the action of a noisy channel. Such inequalities have apparently been first discovered by Ahlswede and Gács in a landmark paper [AG76]. Among the work predating [AG76] and extending it we mention [Dob56, Sar62, CIR+93]. Notable connections include topics ranging from existence and uniqueness of Gibbs measures and log-Sobolev inequalities to performance limits of noisy circuits. We refer the reader to the introduction in [PW16] and the recent monographs [Rag14, RS+13] for more detailed discussions of applications and extensions.

For a fixed channel $P_{Y|X}: \mathcal{X} \to \mathcal{Y}$, let $P_{Y|X} \circ P$ be the distribution on $\mathcal{Y}$ induced by the push-forward of the distribution $P$. One approach to strong data processing seeks to find the contraction coefficients

$$
\eta_f \triangleq \sup_{P,Q:P \neq Q} \frac{D_f (P_{Y|X} \circ P \parallel P_{Y|X} \circ Q)}{D_f (P \parallel Q)},
$$

where the $D_f (P \parallel Q) \triangleq \mathbb{E}_Q[f \left( \frac{dP}{dQ} \right)]$ is an $f$-divergence of Csiszár [Csi67]. When the divergence $D_f$ is the KL-divergence and total variation\footnote{The total variation between two distributions $P$ and $Q$ is $d_{TV}(P,Q) \triangleq \sup_{E} |P[E] - Q[E]|$.}, we denote the coefficient $\eta_f$ as $\eta_{KL}$ and $\eta_{TV}$, respectively.

For discrete channels, [AG76] showed equivalence of $\eta_{KL} < 1$, $\eta_{TV} < 1$ and connectedness of the bipartite graph describing the channel. Having $\eta_{KL} < 1$ implies reduction in the usual data-processing inequality for mutual information [CK81, Exercise III.2.12], [AGKN13]:

$$
\forall W \to X \to Y : I(W;Y) \leq \eta_{KL} \cdot I(W;X).
$$
We refer to inequalities of the form (2) linear SDPIs.

When $P_{Y|X}$ is an additive white Gaussian noise channel, i.e. $Y = X + Z$ with $Z \sim \mathcal{N}(0, 1)$, it has been shown [PW16] that restricting the maximization in (1) to distributions with a bounded second moment (or any moment) still leads to no-contraction, giving $\eta_{KL} = \eta_{TV} = 1$ for AWGN. Nevertheless, the contraction does indeed take place, except not multiplicatively. The region

$$\{ (d_{TV}(P, Q), d_{TV}(P * P_Z, Q * P_Z)) : \mathbb{E}(P_{+Q}/2 | X^2) \leq \gamma \},$$

has been explicitly determined in [PW16], where $*$ denotes convolution. The boundary of this region, deemed the Dobrushin curve of the channel, turned out to be strictly bounded away from the diagonal (identity). In other words, except for the trivial case where $d_{TV}(P, Q) = 0$, total variation decreases by a non-trivial amount in Gaussian channels.

Unfortunately, the similar region for KL-divergence turns out to be trivial, so that no improvement in the inequality

$$D(P_X * P_Z \| Q_Z * P_Z) \leq D(P_X \| Q_X)$$

is possible (given the knowledge of the right-hand side and moment constraints on $P_X$ and $Q_X$). In [PW16], in order to study how mutual information dissipates on a chain of Gaussian links, this problem was resolved by a rather lengthy workaround which entails first reducing questions regarding the mutual information to those about the total variation and then converting back.

A more direct approach, in the spirit of the joint-range idea of Harremoës and Vajda [HV11], is to find (or bound) the best possible data-processing function $F_I$ defined as follows.

**Definition 1.** For a fixed channel $P_{Y|X}$ and a convex set $\mathcal{P}$ of distributions on $\mathcal{X}$ we define

$$F_I(t, P_{Y|X}, \mathcal{P}) \triangleq \sup \{ I(W; Y) : I(W; X) \leq t, W \to X \to Y, P_X \in \mathcal{P} \},$$

where the supremum is over all joint distributions $P_{W,X}$ with $P_X \in \mathcal{P}$. When the channel is clear from the context, we abbreviate $F_I(t, P_{Y|X})$ as $F_I(t)$.

For brevity we denote $F_I(t, \gamma)$ the function corresponding to the special case of the AWGN channel and quadratic constraint. Namely, $Y_{\gamma} = \sqrt{\gamma} X + Z$, where $Z \sim \mathcal{N}(0, 1)$ is independent of $X$, we define

$$F_I(t, \gamma) \triangleq \sup \{ I(W; Y_{\gamma}) : I(W; X) \geq t, W \to X \to Y_{\gamma}, \mathbb{E}[X^2] \leq 1 \}.$$  

The significance of the function $F_I$ is that it gives the optimal input-independent strong data processing inequalities. It is instructive to compare definition of $F_I$ with two related quantities considered previously in the literature. Witsenhausen and Wyner [WW75] defined

$$F_T(P_{XY}, h) = \inf H(Y|W),$$

with the infimum taken over all joint distributions satisfying

$$W \to X \to Y, H(X|W) = h, \mathbb{P}[X = x, Y = y] = P_{XY}(x, y).$$

Clearly, by a simple reparametrization $h = H(X) - t$, this function would correspond to $H(Y) - F_I(t)$ if $F_I(t)$ were defined with restriction to a given input distribution $P_X$. The $P_X$-independent version of (5) has also been studied by Witsenhausen [Wit74]:

$$f_T(P_{Y|X}, h) = \inf H(Y|W),$$

with
with the infimum taken over all
\[ W \rightarrow X \rightarrow Y, H(X|W) = h, \mathbb{P}[Y = y|X = x] = P_{Y|X}(y|x). \]

This quantity plays a role in a generalization of Mrs. Gerber’s lemma and satisfies a convenient tensorization property:
\[ f_T((P_{Y|X})^n, nh) = n f_T(P_{Y|X}, h). \]

There is no one-to-one correspondence between \( f_T(P_{Y|X}, h) \) and \( F_I(t) \) and in fact, alas, \( F_I(t) \) does not satisfy any (known to us) tensorization property.

### 1.1 Overview of results

A priori, the only bounds we can state on \( F_I \) are consequences of capacity and the data processing inequality:
\[
F_I(t, P_{Y|X}) \leq \min \{ t, C(P_{Y|X}, \mathcal{P}) \},
\]
where \( C(P_{Y|X}, \mathcal{P}) \triangleq \sup_{P_X \in \mathcal{P}} I(X; Y) \). For the Gaussian-quadratic case, capacity equals
\[
C(\gamma) = \frac{1}{2} \ln(1 + \gamma).
\]

\[
F_I(t, \gamma) \leq \min \{ t, C(\gamma) \},
\]
where \( C(\gamma) = \frac{1}{2} \ln(1 + \gamma) \) is the Gaussian channel capacity.

In this work we show that generally the trivial bound (7) is not tight at any point. Namely, we prove that
\[
F_I(t) \leq t - g_d(t),
\]
\[
F_I(t) \leq C - g_h(t)
\]
and both functions \( g_d \) and \( g_h \) are strictly positive for all \( t > 0 \). We call these two results diagonal and horizontal bounds respectively. See Fig. 1 for an illustration.
For the Gaussian-quadratic case we show explicitly that our estimates are asymptotically sharp. For example, Theorem 1 (Gaussian diagonal bound) shows the lower-bound portion of
\[ gd(t, \gamma) = e^{-\frac{\gamma}{2} \ln \frac{1}{t} + \Theta(\ln \frac{1}{t})}. \]

An application of (10) allows, via a repeated application of (8), to infer that the mutual information between the input \(X_0\) and the output \(Y_n\) of a chain of \(n\) energy-constrained Gaussian relays converges to zero \(I(X_0; Y_n) \to 0\). In fact, (10) recovers the optimal convergence rate of \(\Theta(\frac{\log \log n}{\log n})\) first reported in [PW16, Theorem 1].

We then generalize the diagonal bound to non-Gaussian noise and arbitrary moment constraint (Theorem 2) by an additional quantization argument. It is worth noting that mutual information does not always strictly contract. Consider the following simple example: Let \(Z\) be uniformly distributed over \([0, 1]\) and \(W = X\) is Bernoulli, then \(I(W; X + Z) = I(W; X) = H(X)\) since \(X\) can be decoded perfectly from \(X + Z\). Surprisingly, this turns out to be the only situation for non-contraction of mutual information occur, as the following characterization (Corollary 2) shows: for strict contraction of mutual information it is necessary and sufficient that the noise \(Z\) cannot be perfectly distinguished from a translate of itself (i.e. \(d_{TV}(P_Z, P_{Z+x}) \neq 1\)).

Going to the horizontal bound, we show (for the Gaussian-quadratic case) that \(FI(t, \gamma)\) approaches \(C(\gamma)\) no faster than double-exponentially in \(t\) as \(t \to \infty\). Namely, in Theorem 3 and Remark 4, we prove that \(gh(t)\) satisfies
\[ e^{-c_1(\gamma) e^{4t}} \leq gh(t) \leq e^{-c_2(\gamma) e^{t} + \ln 4(1+\gamma)}, \]
where \(c_1(\gamma)\) and \(c_2(\gamma)\) are strictly positive functions of \(\gamma\).

Generalization of the horizontal bound to arbitrary noise distribution (Theorem 5) proceeds along a similar route. In the process, we derive a deconvolution estimate that bounds the Kolmogorov-Smirnov distance (\(L_\infty\) norm between CDFs) in terms of the total variation between convolutions with noise. Namely, Corollary 3 shows that for a noise \(Z\) with bounded density and non-vanishing characteristic function we have
\[ d_{KS}(P, Q) \leq f(d_{TV}(P * P_Z, Q * P_Z)) \]
for some continuous increasing function \(f(\cdot)\) with \(f(0) = 0\).

The final result (Theorem 6) addresses the question of bounding \(FI\)-curve for non-scalar channel \(Y = X + Z\). Somewhat surprisingly, we show that for the infinite-dimensional Gaussian case the trivial bound (7) on the \(FI\)-curve is exact.

1.2 Organization and notation
The rest of the paper is organized as follows. Section 2 introduces properties of the \(FI\)-curve, together with a few examples for discrete channels.

Sections 3 and 4 present a (diagonal) lower bound for \(gd(t)\) in the Gaussian and general setting respectively. Section 5 shows that any \(X\) for which close-to-optimal (in MMSE sense) linear estimator of \(Y = X + Z\) exists, must necessarily be close to Gaussian in the sense of Kolmogorov-Smirnov distance. These results are then used in Section 6 to prove a (Gaussian horizontal) lower bound on \(gh(t)\).

Section 7 introduces a deconvolution result that connects KS-distance with TV-divergence. This result is then applied in Section 8 to derive a general horizontal bound for \(FI\) curve for a wide range of additive noise channels.
Finally, in Section 9 we consider the infinite-dimensional discrete Gaussian channel, and show that in this case there exists no non-trivial strong data processing inequality for mutual information. In the appendix, we present a shorter proof of the key step in the Gaussian horizontal bound (namely, Lemma 5) employing Talagrand’s inequality [Tal96].

**Notations** For any distribution $P$ on $\mathbb{R}$, let $F_P(x) = P((−\infty, x])$ denote its cumulative distribution function (CDF). For any random variable $X$, denote its distribution and CDF by $P_X$ and $F_X$, respectively. For any sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write $a_n \gtrless b_n$ or $b_n \lesssim a_n$ when $a_n \geq cb_n$ for some absolute constant $c > 0$.

## 2 Examples and properties of the $F_I$-curves

In this section we discuss properties of the $F_I$-curve, and present a few examples for discrete channels.

**Proposition 1** (Properties of the $F_I$-curve).

1. $F_I$ is an increasing function such that $0 \leq F_I(t) \leq t$ with $F_I(0) = 0$.
2. $t \mapsto \frac{F_I(t)}{t}$ is decreasing. Consequently, $F_I$ is subadditive and $F_I'(0) = \sup_{t > 0} \frac{F_I(t)}{t}$.
3. Value of $F_I(t)$ is unchanged if $W$ is restricted to an alphabet of size $|X| + 1$. Upper concave envelope of $F_I(t)$ equals upper concave envelope of a set of pairs $(I(W; X), I(W; Y))$ achieved by restricting $W$ to alphabet $X$.

**Proof.** The first part follows directly from the definition, the non-negativity and the data processing inequality of mutual information. For the second part, fix $P_{Y|X}$ and let $P_{W|X}$ achieve the pair $(I(W; X), I(W; Y))$. Then by choosing $P'_{W|X} = \lambda P_{W|X} + (1-\lambda)P_WP_X$, the pair $(\lambda I(W; X), \lambda I(W; Y))$ is also achievable. It follows directly that $t \mapsto F_I(t)/t$ is decreasing.

Claim 3 follows by noticing that for a fixed distribution $P_X$, any pair $(H(X|W), H(Y|W))$ can be attained by $W$ with a given restriction on the alphabet, see [WW75, Theorem 2.3]. Similarly, concave envelope of $F_I(t)$ can be found by taking convex closure of extremal points $(H(X) - H(X|W), H(Y) - H(Y|W))$, which can be attained by $W$ with alphabet $|X|$, see paragraph after [WW75, Theorem 2.3].

We present next a few examples of the $F_I(t)$-curve for discrete channels:

1. **Erasure channel** is defined as $P_{Y|X} : X \rightarrow X \cup \{?\}$ with $y = x$ or $?$ with probabilities $1 - \alpha$ and $\alpha$, respectively. In this case we have for any $W - X - Y$ a convenient identity, cf. [VW08]:

$$I(W; Y) = (1 - \alpha)I(W; X),$$

and consequently, the $F_I$-curve is

$$F_I(t) = (1 - \alpha)t \wedge \log |X|$$ (12)

and is achieved by taking $W = X$.

2. **Binary symmetric channel** BSC($\delta$) is defined as $P_{Y|X} : \{0, 1\} \rightarrow \{0, 1\}$ with $Y = X + Z$, $Z \sim \text{Ber}(\delta)$. Here the optimal coupling is $X = W + Z'$ with $Z' \perp \perp W \sim \text{Ber}(1/2)$ and varying bias of $Z'$. This is formally proved in the next Proposition.
Proposition 2. The $F_t$-curve of the $BSC(\delta)$ is given by
\[
F_t(t) = \log 2 - h_b(\delta * h_b^{-1}(|\log 2 - t|^{+})),
\]
where $p * q = p(1-q) + q(1-p)$, $h_b(y) \triangleq -y \log y - (1-y) \log(1-y)$ is the binary entropy function and $h_b^{-1} : [0, \log 2] \to [0, \frac{1}{2}]$ is its functional inverse.

Proof. First, it is clear that
\[
F_t(t) = \max_{p \in [h_b^{-1}(t) - \frac{1}{2}, h_b^{-1}(t) + \frac{1}{2}]} f_t(t, p),
\]
where
\[
f_t(x, p) \triangleq \max \{I(W; Y) : I(W; X) \leq x, X \sim Ber(p)\}
= h_b(p * \delta) - h_b(\delta * h_b^{-1}(h_b(p) - x))
\]
that is $f_t(t, p)$ is an $F_t$-curve for a fixed marginal $P_X$.

It is sufficient to prove that $p = \frac{1}{2}$ is a maximizer in (14) regardless of $t$. To that end, recall Mrs. Gerber’s Lemma [WZ73] states that
\[
x \mapsto h_b(\delta * h_b^{-1}(x))
\]
is convex on $[0, \log 2]$. Consequently for any $0 \leq t \leq u \leq \log 2$, $f_t(t, h_b^{-1}(u)) = h_b(\delta * h_b^{-1}(u) - h_b(\delta * h_b^{-1}(u - t)) \leq h_b(\delta * h_b^{-1}(\log 2) - h_b(\delta * h_b^{-1}(\log 2 - t)) = f_t(t, 1/2).$ \hfill \Box

3 Diagonal bound for Gaussian channels

We now study properties of the $F_t$-curve in the Gaussian case, i.e. $P_Z = N(0, 1)$. In this section, we show that $F_t(t, \gamma)$ is bounded away from $t$ for all $t > 0$ (Theorem 1) and investigate the behavior of $F_t(t, \gamma)$ for small $t$ (Corollary 1). The proofs of the non-linear SDPIs presented in both the current and the next section hinge on the existence of a linear SDPI when the input $X$ is amplitude-constrained. We define
\[
\eta(A) \triangleq \sup_{P, Q \text{ on } [-A, A]} \frac{D(P \ast P_Z \| Q \ast P_Z)}{D(P \| Q)}.
\]
Similarly, define the Dobrushin’s coefficient $\eta_{TV}(A)$ with $D$ replaced by $d_{TV}$ in (15), that is,
\[
\eta_{TV}(A) = \sup_{z, z' \in [-A, A]} d_{TV}(P_{Z+z}, P_{Z+z'}) = \sup_{|\delta| \leq 2A} \theta(\delta),
\]
where
\[
\theta(\delta) \triangleq d_{TV}(P_Z, P_{Z+\delta}).
\]
Observe that for any $W \rightarrow X \rightarrow Y$, where $Y = X + Z$ and $X \in [-A, A]$ almost surely, we have $I(W; Y) \leq \eta(A)I(W; X)$. In the Gaussian case considered in this section, $\eta(A)$ can be upper-bounded as [PW16]
\[
\eta(A) \leq \eta_{TV}(A) = \theta(A) = 1 - 2Q(A),
\]
where $Q(x) \triangleq \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$ is the Gaussian complimentary CDF. This leads to the following general lemma, which also holds for general $P_Z$. 

\[
\]

7
Lemma 1. Let $W \rightarrow X \rightarrow Y$, where $Y = X + Z$. For any $A > 0$, let $\epsilon \triangleq \mathbb{P}[|X| > A]$. Then
\begin{equation}
I(W; Y) \leq I(W; X) - \tilde{\eta}(A) (I(W; X) - \bar{h}_b(\epsilon) - \epsilon I(W; Y|E = 1)),
\end{equation}
where $h_b(x) \triangleq x \ln \frac{x}{x} + (1 - x) \ln \frac{1}{1-x}$ and $\tilde{\eta}(A) \triangleq 1 - \eta(A)$.

Proof. Let $E \triangleq 1_{|X| \geq A}$ and $\bar{\epsilon} \triangleq 1 - \epsilon$. Then
\begin{align*}
I(W; Y) &\leq I(W; Y|E) \\
&= I(W; E) + \epsilon I(W; Y|E = 1) + \epsilon I(W; Y|E = 0) \\
&\leq I(W; E) + \epsilon I(W; Y|E = 1) + \epsilon \eta(A) I(W; X|E = 0),
\end{align*}
where the last inequality follows from the definition of $\eta(t)$ in (15). Observing that
\begin{equation*}
\epsilon I(W; X|E = 0) = I(W; X) - \epsilon I(W; X|E = 1) - I(W; E),
\end{equation*}
and denoting $\tilde{\eta}(A) \triangleq 1 - \eta(A)$, we can further bound (20) by
\begin{align*}
I(W; Y) &\leq \tilde{\eta}(A) (I(W; E) + \epsilon I(W; Y|E = 1)) + \eta(A) I(W; X) + \epsilon \eta(A) (I(W; Y|E = 1) - I(W; X|E = 1)) \\
&\leq \tilde{\eta}(A) (I(W; E) + \epsilon I(W; Y|E = 1)) + \eta(A) I(W; X) \\
&= I(W; X) - \tilde{\eta}(A) (I(W; X) - I(W; E) - \epsilon I(W; Y|E = 1)),
\end{align*}
where (21) follows from $I(W; Y|E = 1) \leq I(W; X|E = 1)$. The result follows by noting that $I(W; E) \leq h_b(\epsilon)$.

We now present explicit bounds for the value of $g_d(t, \gamma)$ when $\mathbb{E}[|X|^2] \leq \gamma$ and $P_Z = \mathcal{N}(0, 1)$.

Theorem 1. For the AWGN channel with quadratic constraint, see (4), we have $F_1(t, \gamma) = t - g_d(t, \gamma)$ and
\begin{equation}
g_d(t, \gamma) \geq \max_{x \in [0,1/2]} 2Q \left( \sqrt{\frac{x}{2}} \right) \left( t - h(x) - \frac{1}{2} \ln \left( 1 + \frac{\gamma}{x} \right) \right).
\end{equation}

Proof. Let $E = 1_{|X| > A/\sqrt{\gamma}}$ and $\mathbb{E}[E] = \epsilon$. Observe that
\begin{equation}
\mathbb{E}[\gamma X^2|E = 1] \leq \gamma/\epsilon \text{ and } \epsilon \leq \gamma/A^2.
\end{equation}
Therefore, from Lemma 1 and (18),
\begin{equation}
I(W; Y_\gamma) \leq I(W; X) - \tilde{\eta}_{TV}(A) (I(W; X) - I(W; E) - \epsilon I(W; Y_\gamma|E = 1)).
\end{equation}

Now observe that, for $\epsilon = \gamma/A^2 \leq 1/2$,
\begin{equation}
I(W; E) \leq H(E) \leq h_b \left( \gamma/A^2 \right).
\end{equation}
In addition,
\begin{align*}
\epsilon I(W; Y_\gamma|E = 1) &\leq \epsilon I(X; Y_\gamma|E = 1) \\
&\leq \frac{\epsilon}{2} \ln \left( 1 + \frac{\gamma}{p} \right) \\
&\leq \frac{\gamma}{2A^2} \ln (1 + A^2).
\end{align*}
Here (26) follows from the fact that mutual information is maximized when $X$ is Gaussian under the power constraint (23), and (27) follows by noticing that $x \mapsto x \ln(1 + a/x)$ is monotonically increasing for any $a > 0$. Combining (25) and (27), and for $A \geq \sqrt{\gamma}$,

$$I(W; E) + \epsilon I(W; Y | E = 1) \leq h_b \left( \frac{\gamma}{A^2} \right) + \frac{\gamma}{2A^2} \ln \left( A^2 + 1 \right).$$  

(28)

Choosing $A = \sqrt{\gamma/x}$, where $0 \leq x \leq 1/2$, (28) becomes

$$I(W; E) + \epsilon I(W; Y | E = 1) \leq h_b (x) + \frac{x}{2} \ln \left(1 + \frac{\gamma}{x}\right).$$  

(29)

Substituting (29) in (24) yields the desired result. □

**Remark 1.** Note that $f_d(x, \gamma) \triangleq h_b (x) + \frac{\gamma}{2} \ln \left(1 + \frac{\gamma}{x}\right)$ is 0 at $x = 0$; furthermore, $f_d(\cdot, \gamma)$ is continuous and strictly positive on $(0, 1/2)$. Therefore $g_d(t, \gamma)$ is strictly positive for $t > 0$. The next corollary characterizes the behavior of $g_d(t, \gamma)$ for small $t$.

**Corollary 1.** For fixed $\gamma$, $t = 1/u$ and $u$ sufficiently large, there is a constant $c_3(\gamma) > 0$ dependent on $\gamma$ such that

$$g_d(1/u, \gamma) \geq \frac{c_3(\gamma)}{u^\gamma \ln u} e^{-\gamma u \ln u}.$$  

(30)

In particular, $g_d(1/u, \gamma) \geq e^{-\gamma u \ln u + O(\ln u \gamma)}$.

**Proof.** Let $x = \frac{1}{2u \ln u}$ in the expression being maximized in (22). For sufficiently large $t$,

$$Q(\sqrt{2u\gamma \ln u}) = \frac{e^{-\gamma u \ln u}}{2 \sqrt{u \pi u \gamma \ln u}} + O\left(\frac{e^{-\gamma u \ln u}}{(u \gamma \ln u)^{3/2}}\right)$$

and

$$g_d \left( \frac{1}{2u \ln u}, \gamma \right) \geq \frac{3}{4u} + O \left( \frac{\ln u}{u \ln u} \right),$$  

(31)

the result follows. □

**Remark 2.** Fix $\gamma > 0$ and define a binary random variable $X$ with $\mathbb{P}[X = a] = 1/a^2$ and $\mathbb{P}[X = 0] = 1 - 1/a^2$ for $a > 0$. Furthermore, let $\hat{X} \in \{0, a\}$ denote the minimum distance estimate of $X$ based on $Y_\gamma$. Then the probability of error satisfies $P_e = \mathbb{P}[X \neq \hat{X}] \leq Q(\sqrt{\gamma a}/2)$. In addition, $h(Q(\sqrt{\gamma a}/2)) = O(e^{-\gamma a^2/8 \sqrt{\gamma a}})$ and $H(X) = a^{-2} \ln a(2 + o(1))$ as $a \to \infty$. Therefore,

$$h_b(Q(\sqrt{\gamma a}/2)) \leq e^{-\frac{\gamma}{2}(\sqrt{\gamma a})^2} + O(\ln(\gamma/H(X))).$$  

(32)

Using Fano’s inequality, $I(X; Y_\gamma)$ can be bounded as

$$I(X; Y_\gamma) \geq I(X; \hat{X}) \geq H(X) - h_b(P_e) \geq H(X) - h_b(Q(\sqrt{\gamma a}/2)) = H(X) - e^{-\frac{\gamma}{2}(\sqrt{\gamma a})^2} + O(\ln(\gamma/H(X))).$$

Setting $W = X$, this result yields the sharp asymptotics (10).
4 Diagonal bound for general additive noise

In this section, we extend the diagonal bound derived in Theorem 1 to arbitrary noise density and generalizing the power constraint to an $L_p$-norm constraint $\mathbb{E}[|X|^p] \leq \gamma$.

**Theorem 2.** Assume that $W \rightarrow X \rightarrow Y$, where $Y = X + Z$, $X$ and $Z$ are independent, $\mathbb{E}[|X|^p] \leq \gamma$, and $Z$ has an absolute continuous distribution. Then

$$I(W; Y) \leq I(W; X) - g_d(I(W; X), \gamma),$$

(33)

where

$$g_d(t, \gamma) \triangleq \frac{1}{2}(1 - \eta(A^*_2))t,$$

(34)

$$A^*_2 \triangleq \inf \{ A > 0 : 18\gamma A^{-p} \ln(A^p) \leq t, \ A^p \geq \max\{2, 2\gamma, \alpha^* e^3/\gamma\} \},$$

(35)

$$\alpha^* \triangleq \inf \{ \alpha > 0 : \eta \left( \frac{1}{2\alpha} \right) \leq 1/3 \}.$$  

(36)

and the amplitude-constrained contraction coefficient $\eta(\cdot)$ is defined in (15).

**Corollary 2.** For any $p \geq 1$ and any $\gamma > 0$, the following statements are equivalent:

(a) Non-linear SDPI (33) holds with $g_d(t, \gamma) > 0$ whenever $t > 0$.

(b) $S \cap (S + x)$ has non-zero Lebesgue measure for all $x \in \mathbb{R}$, where $S \triangleq \{ z : p_Z(z) > 0 \}$ is the support of the probability density function $p_Z$ of $Z$.

In order to prove these results, we first study the case where $X$ is discrete and a deterministic function of $W$.

**Lemma 2.** Let $W \rightarrow X \rightarrow Y$, $Y = X + Z$, and $W \rightarrow X$ be a deterministic mapping. In addition, assume that $X$ takes values on some $\Delta$-grid for $\Delta > 0$ (i.e. $X/\Delta \in \mathbb{Z}$ almost surely) and $\mathbb{E}[|X|^p] \leq \gamma$, $p \geq 1$. Then

$$I(X; Y) \leq \left( 1 - \frac{\bar{\eta}(A^*_1)}{2} \right) H(X),$$

(37)

where

$$A^*_1 \triangleq \min \left\{ A : A^p \geq \max\{2, 2\gamma, c^3/\gamma\}, A^{-p} \ln A \leq \frac{H(X)}{6\gamma} \right\}.$$  

(38)

**Proof.** Let $E \triangleq 1_{\{|X| \geq A\}}$ and $\epsilon \triangleq P[E = 1]$. Then, from Lemma 1,

$$I(X; Y) \leq H(X) - \bar{\eta}(A)(H(X) - h_b(\epsilon) - \epsilon H(X|E = 1)).$$

(39)

Observe that for $\mathbb{E}[|X|^p] \leq \gamma$,

$$\epsilon = P[|X| \geq A] \leq \gamma/A^p,$$

(40)

and, for $A \geq 1$

$$\mathbb{E}[|X| | E = 1] \leq \mathbb{E}[|X|^p | E = 1] \leq \gamma/\epsilon.$$  

(41)

In addition, for any integer-valued random variable $U$ we have (cf. [CT06, Lemma 13.5.4])

$$H(U) \leq (\mathbb{E}[|U|] + 1) h_b \left( \frac{1}{\mathbb{E}[|U|] + 1} \right) + \ln 2.$$  

(42)
Consequently, for $A^p \geq \max\{2, 2\gamma\}$,
\[
h_b(\epsilon) + \epsilon H(X|E = 1)
\leq h_b(\epsilon) + \left(\frac{\gamma}{\Delta} + \epsilon\right) h_b\left(\frac{\epsilon}{2} + \frac{\epsilon}{\Delta} + \frac{\epsilon}{\Delta} + \epsilon\right) + \epsilon \ln 2
\leq h_b\left(\frac{\gamma}{A^p}\right) + h_b\left(\frac{A^p}{\Delta} + 1\right) h_b\left(\frac{1}{1 + A^p/\Delta}\right) + \frac{\gamma}{A^p} \ln 2
\leq \frac{\gamma}{A^p} \ln A^p + \frac{\gamma}{A^p} \left(1 + \ln \frac{2}{\gamma}\right) + \frac{\gamma}{A^p} \left(\ln \left(\frac{A^p}{\Delta} + 1\right) + A^p \ln \left(1 + \frac{\Delta}{A^p}\right)\right)
\leq \frac{\gamma}{A^p} \ln A^p + \frac{\gamma}{A^p} \left(2 + \frac{2}{\gamma}\right) + \frac{\gamma}{A^p} \ln \left(\frac{A^p}{\Delta} + 1\right)
\leq \frac{2\gamma}{A^p} \ln A^p + \frac{\gamma}{A^p} \left(3 + \ln \frac{2}{\gamma}\Delta\right),
\] (43) (44) (45)

where (43) and (44) follows from the fact that $-(1-x) \ln (1-x) \leq x$ and $\ln (x+1) \leq x$ for $x \in [0, 1]$, respectively, and (45) follows by observing that $\ln (x+1) \leq \ln x + 1$. Assuming $A^p \geq e^3/\gamma\Delta$,
\[
h_b(\epsilon) + \epsilon H(X|E = 1) \leq \frac{3\gamma \ln A^p}{A^p}.
\] (46)

Since the right-hand side of the previous equation is strictly decreasing for $A \geq \exp(1)$, $A$ can be chosen sufficiently large such that $\frac{3\gamma \ln A^p}{A^p} \leq H(X)/2$. Choosing $A = A_1^*$, where $A_1^*$ is given in (38), and combining (46) and (39), we conclude that
\[
I(X; Y) \leq \left(1 - \frac{\eta(A_1^*)}{2}\right) H(X),
\]
proving the lemma.

\[\square\]

**Proof of Theorem 2.** We start by verifying that $\alpha$ defined in (36) is finite and so is $A_2^* \in (35)$. Since $\eta(\alpha) \leq \eta_{TV}(a)$, it suffices to show that $\eta_{TV}(a)$ vanishes as $a \to 0$. Recall $\theta(\delta) = \frac{1}{2} \int |p_Z(z) - p_Z(z + \delta)| \, dz$ as defined in (17). By the denseness of compactly supported continuous functions in $L^1$, $\theta(\delta) \to 0$ as $a \to 0$. Furthermore, the translation invariance and the triangle inequality of total variation imply that $|\theta(a) - \theta(a')| \leq \theta(|a - a'|)$ and hence $\theta$ is uniformly continuous. Therefore,
\[
\eta_{TV}(a) = \max_{|b| \leq 2a} \theta(\delta)
\] (47)
is continuous in $a$ on $\mathbb{R}^+$, which ensures that $a^*$ is finite.

From Lemma 1, and once more denoting $E \triangleq 1_{\{|X| \geq A\}}$, $\epsilon \triangleq \mathbb{P}[|X| \geq A]$ and $\eta(\bar{A}) = 1 - \eta(A)$, we have
\[
I(W; Y) \leq I(W; X) - \bar{\eta}(A) (I(W; X) - h_b(\epsilon) - \epsilon I(W; Y|E = 1)).
\] (48)

Let $Q_\alpha = [\alpha X]$. Then
\[
I(W; Y) \leq I(Q_\alpha; Y) + I(W; Y|Q_\alpha)
\leq I(Q_\alpha; Y) + \eta\left(\frac{1}{2\alpha}\right) I(W; X|Q_\alpha)
\leq H(Q_\alpha) + \eta\left(\frac{1}{2\alpha}\right) I(W; X).
\]
Thus,
\[ I(W; Y|E = 1) \leq H(Q_\alpha|E = 1) + \eta \left( \frac{1}{2\alpha} \right) I(W; X|E = 1). \] (49)

Since
\[ \epsilon I(W; X|E = 1) \leq I(W; X), \] (50)
combining (48)–(50) gives
\[ I(W; Y) \leq I(W; X) - \bar{\eta}(A) \left( I(W; X) - h_\alpha(\epsilon) - \epsilon H(Q_\alpha|E = 1) - \eta \left( \frac{1}{2\alpha} \right) I(W; X) \right). \] (51)

Since \( E [|Q_\alpha|] \leq \alpha \gamma / A^p \), from (42) and (46) it follows that for \( A^p \geq \alpha e^{3/\gamma} \),
\[ h_\alpha(\epsilon) + \epsilon H(Q_\alpha|E = 1) \leq \frac{3\gamma \ln(A^p)}{A^p}. \] (52)

Thus, choosing \( \alpha \) such that \( \eta(1/2\alpha) \leq 1/3 \), and \( A \) sufficiently large such that \( 3\gamma A^{-p} \ln A^p \leq I(W; X)/6 \), (52) becomes
\[ I(W; Y) \leq I(W; X) \left( 1 - \frac{\bar{\eta}(A)}{2} \right), \] (53)
proving the result upon choosing \( A = A^* \).

Proof of Corollary 2. To show \( (a) \Rightarrow (b) \), suppose that \( S \cap (S + x_0) \) has zero Lebesgue measure for some \( x_0 \). Consider \( W = X = x_0B \), where \( B \sim \text{Bernoulli}(\epsilon) \) with \( E[X^p] = \epsilon|x_0|p \leq \gamma \). Since \( d_{TV}(P_Z, P_{Z+x}) = 0 \), the \( X \) can be perfectly decoded from \( Y = X + Z \) and hence \( I(W; Y) = I(W; X) = H(X) \), which shows that \( F_1(t) = t \) in a neighborhood of zero.

To show \( (b) \Rightarrow (a) \), in view of Theorem 2, it suffices to show that \( \eta(A) < 1 \) for all finite \( A \). Recall that for any channel, \( \eta_{KL} = 1 \) if and only if \( \eta_{TV} = 1 \) ([CKZ98, Proposition II.4.12]). Therefore it is equivalent to show that \( \eta_{TV}(A) < 1 \) for all finite \( A \). Suppose otherwise, i.e., \( \eta_{TV}(A) = 1 \) for some \( A > 0 \). By (47), there exists some \( \delta \in [-A, A] \) such that \( d_{TV}(P_Z, P_{Z+\delta}) = 1 \), which means that \( S \cap (S + \delta) \) has zero Lebesgue, contradicting the assumption \( (b) \) and completing the proof.

5 Minimum mean square error and near-Gaussianity

We now take a step back from strong data-processing inequalities and present an ancillary result of independent interest. We prove that any random variable for which there exists an almost optimal (in terms of the mean-squared error) linear estimator operating on the Gaussian-corrupted measurement must necessarily be almost Gaussian (in terms of the Kolmogorov-Smirnov distance). We will use this result in the next section to bound the horizontal gap \( g_b(t, \gamma) \) for Gaussian noise.

Throughout the rest of the paper we make use of Fourier-analytic tools and, in particular, Esseen’s inequality, stated below for reference.

**Lemma 3 ([Fel66, Eq. (3.13), p. 538]).** Let \( P \) and \( Q \) be two distributions with characteristic functions \( \varphi_P \) and \( \varphi_Q \), respectively. In addition, assume that \( Q \) has a bounded density \( q \). Then
\[ d_{KS}(P, Q) \leq \frac{1}{\pi} \int_{-T}^{T} \left| \frac{\varphi_P(\omega) - \varphi_Q(\omega)}{\omega} \right| \, d\omega + \frac{24 \| q \| \infty}{\pi T}, \] (54)
where \( d_{KS}(P, Q) \triangleq \sup_{x \in \mathbb{R}} |F_P(x) - F_Q(x)| \) is the Kolmogorov-Smirnov distance.
Let $P_\gamma = \mathcal{N}(0, 1)$ and assume that $\mathbb{E} \left[ |X|^2 \right] \leq \gamma$. We show next that if the linear least-square error of estimating $X$ from $Y_\gamma$ is small (i.e., close to the minimum mean-squared error), then $X$ must be almost Gaussian in terms of the KS-distance. With this result in hand, we use the I-MMSE relationship [GSV05] to show that if $I(X; Y_\gamma)$ is close to $C(\gamma)$, then $X$ is also almost Gaussian. This result, in turn, will be applied in the next section to bound $F_1(t, \gamma)$ away from $C(\gamma)$.

Denote the linear least-square error estimator of $X$ given $Y_\gamma$ by $f_L(y) \triangleq \sqrt{\gamma}y/(1 + \gamma)$, whose mean-squared error is

\[ \text{Lmmse}(X|Y_\gamma) \triangleq \mathbb{E} \left[ (X - f_L(Y_\gamma))^2 \right] = \frac{1}{1 + \gamma}. \]

Assume that $\text{Lmmse}(X|Y_\gamma) - \text{mmse}(X|Y_\gamma) \leq \epsilon$. It is well known that $\epsilon = 0$ if and only if $X \sim \mathcal{N}(0, 1)$ (see e.g. [GWSV11]). To develop a finitary version of this result, we ask the following question: If $\epsilon$ is small, how close is $X$ to Gaussian? The next lemma provides a quantitative answer.

**Lemma 4.** If $\text{Lmmse}(X|Y_\gamma) - \text{mmse}(X|Y_\gamma) \leq \epsilon$, then there are absolute constants $a_0$ and $a_1$ such that

\[ d_{\text{KS}}(F_X, \mathcal{N}(0, 1)) \leq a_0 \frac{1}{\gamma \log(1/\epsilon)} + a_1 (1 + \gamma)\epsilon^{1/4} \sqrt{\gamma \log(1/\epsilon)}. \]  

(55)

**Remark 3.** Note that the gap between the linear and nonlinear MMSE can be expressed as the Fisher distance between the convolutions, i.e., $\text{Lmmse}(X|Y_\gamma) - \text{mmse}(X|Y_\gamma) = I(P_\gamma, N(0, 1 + \gamma))$, where $I(P||Q) = \int [(\log \frac{dP}{dQ})^2] dP$ is the Fisher distance, which dominates the KL divergence according to the log-Sobolev inequality. Therefore Lemma 4 can be interpreted as a deconvolution result, where bounds on a stronger (Fisher) distance between the convolutions lead to bounds on the distance between the original distributions under a weaker (KS) metric.

**Proof.** Denote $f_M(y) = \mathbb{E} \left[ X|Y_\gamma = y \right]$. Then

\[ \text{Lmmse}(X|Y_\gamma) - \text{mmse}(X|Y_\gamma) = \mathbb{E} \left[ (X - f_L(Y_\gamma))^2 \right] - \mathbb{E} \left[ (X - f_M(Y_\gamma))^2 \right] \]

\[ = \mathbb{E} \left[ (f_M(Y_\gamma) - f_L(Y_\gamma))^2 \right] \]

\[ \leq \epsilon. \]

Denote $\Delta(y) \triangleq f_M(y) - f_L(y)$. Then $\mathbb{E} \left[ \Delta(Y_\gamma) \right] = 0$ and $\mathbb{E} \left[ \Delta(Y_\gamma)^2 \right] \leq \epsilon$. From the orthogonality principle:

\[ \mathbb{E} \left[ e^{itY_\gamma} (X - f_M(Y_\gamma)) \right] = 0. \]  

(56)

Let $\varphi_X$ denote the characteristic function of $X$. Then

\[ \mathbb{E} \left[ e^{itY_\gamma} (X - f_M(Y_\gamma)) \right] = \mathbb{E} \left[ e^{itY_\gamma} (X - f_L(Y_\gamma) - \Delta(Y_\gamma)) \right] \]

\[ = \frac{1}{1 + \gamma} \left( e^{-t^2/2} \mathbb{E} \left[ e^{\sqrt{\gamma}tX} \right] - \sqrt{\gamma} \varphi_X(\sqrt{\gamma}t) \mathbb{E} \left[ Ze^{itZ} \right] \right) - \mathbb{E} \left[ e^{itY_\gamma} \Delta(Y_\gamma) \right] \]

\[ = \frac{-ie^{-u^2/2\gamma}}{1 + \gamma} \left( \varphi_X(u) + u \varphi_X(u) - \mathbb{E} \left[ e^{itY_\gamma} \Delta(Y_\gamma) \right] \right), \]  

(57)

where the last equality follows by changing variables $u = \sqrt{\gamma}t$. Consequently,

\[ e^{-u^2/2\gamma} \left( \varphi_X(u) + u \varphi_X(u) \right) \leq \mathbb{E} \left[ \Delta(Y_\gamma) \right] \]

\[ \leq \sqrt{\epsilon}. \]  

(58)

\[ \leq \mathbb{E} \left[ \Delta(Y_\gamma) \right] \]

\[ \leq \sqrt{\epsilon}. \]  

(59)
Put $\phi_X(u) = e^{-u^2/2} (1 + z(u))$. Then

$$|\phi'_X(u) + u\phi_X(u)| = e^{-u^2/2} |z'(u)|,$$

and, from \((59)\), $|z'(u)| \leq (1 + \gamma)\sqrt{e e^{-u^2/2}}$. Since $z(0) = 0$,

$$|z(u)| \leq \int_0^u |z'(x)| dx \leq u(1 + \gamma)\sqrt{e e^{-u^2/2}}. \quad (60)$$

Observe that $|\phi_X(u) - e^{-u^2/2}| = e^{-u^2/2} |z(u)|$. Then, from \((60)\),

$$\left|\frac{\phi_X(u) - e^{-u^2/2}}{u}\right| \leq (1 + \gamma)\sqrt{e e^{-u^2/2}}. \quad (61)$$

Thus, Lemma 3 yields

$$d_{KS}(F_X, \mathcal{N}(0,1)) \leq \frac{1}{\pi} \int_{-T}^{T} (1 + \gamma)\sqrt{e e^{-u^2/2}} du + \frac{12\sqrt{2}}{\pi^{3/2}T}$$

$$\leq \frac{2T}{\pi} (1 + \gamma)\sqrt{e e^{-u^2/2}} + \frac{12\sqrt{2}}{\pi^{3/2}T}.$$

Choosing $T = \sqrt{\frac{2}{\gamma} \ln(\frac{1}{\epsilon})}$, we find

$$d_{KS}(F_X, \mathcal{N}(0,1)) \leq a_0 \sqrt{\frac{1}{\gamma \ln(1/\epsilon)}} + a_1 (1 + \gamma)\epsilon^{1/4} \sqrt{\gamma \ln(1/\epsilon)},$$

where $a_0 = \frac{24}{\pi^{3/2}}$ and $a_1 = \sqrt{2}/\pi$. \hfill \Box

Through the I-MMSE relationship \cite{GSV05}, the previous lemma can be extended to bound the KS-distance between the distribution of $X$ and the Gaussian distribution when $I(X; Y_\gamma)$ is close to $C(\gamma)$.

**Lemma 5.** Assume that $C(\gamma) - I(X; Y_\gamma) \leq \epsilon$. Then, for $\gamma > 4\epsilon$,

$$d_{KS}(F_X, \mathcal{N}(0,1)) \leq a_0 \sqrt{\frac{2}{\gamma \ln(\frac{1}{\epsilon})}} + a_1 (1 + \gamma)(\gamma\epsilon)^{1/4} \sqrt{2 \ln \left(\frac{\gamma}{4\epsilon}\right)}. \quad (62)$$

**Proof.** From the I-MMSE relationship \cite{GSV05}:

$$C(P) - I(X; Y_P) = \frac{1}{2} \int_0^P \frac{1}{1 + \gamma} - \text{mmse}(X|Y_\gamma) d\gamma \leq \epsilon. \quad (63)$$

Since $\text{mmse}(X|Y_\gamma) \leq \frac{1}{1 + \gamma}$, for any $\delta \in [0, P)$

$$\frac{1}{\delta} \int_{P-\delta}^P \frac{1}{1 + \gamma} - \text{mmse}(X|Y_\gamma) d\gamma \leq \frac{2\epsilon}{\delta}. \quad (64)$$

The function $\text{mmse}(X|Y_\gamma)$ is continuous in $\gamma$. Then, from the mean-value theorem for integrals, there exists $\gamma^* \in (P - \delta, P)$ such that

$$\frac{1}{1 + \gamma^*} - \text{mmse}(X|Y_{\gamma^*}) \leq \frac{2\epsilon}{\delta}. \quad (65)$$
From Lemma 4, we find
\[
d_{KS}(F_X, \mathcal{N}(0,1)) \leq a_0 \sqrt{\frac{1}{\gamma^* \ln(\delta/2\epsilon)}} + a_1(1 + \gamma^*) \left( \frac{2\epsilon}{\delta} \right)^{1/4} \sqrt{\gamma^* \ln(\delta/2\epsilon)}
\]
\[
\leq a_0 \sqrt{\frac{1}{(P - \delta) \ln(\delta/2\epsilon)}} + a_1(1 + P) \left( \frac{2\epsilon}{\delta} \right)^{1/4} \sqrt{P \ln(\delta/2\epsilon)}.
\]
The desired result is found by choosing \( \delta = P/2 \).

\[\square\]

6 Horizontal bound for Gaussian channels

Using the results from the previous section, we show that, for \( PZ \sim \mathcal{N}(0,1) \), \( FI(t,\gamma) \) is bounded away from the capacity \( C(\gamma) \) for all \( t \).

**Theorem 3.** For the AWGN channel with quadratic constraint, see (4), we have \( FI(t,\gamma) = C(\gamma) - gh(t,\gamma) \) and
\[
gh(t,\gamma) \geq e^{-c_1(\gamma) e^{4t}},
\]
where \( c_1(\gamma) \) is some positive constant depending on \( \gamma \).

We first give an auxiliary lemma.

**Lemma 6.** If \( D(\mathcal{N}(0,1)\| P_X \ast \mathcal{N}(0,1)) \leq 2\epsilon \), then there exists an absolute constant \( a_2 > 0 \) such that
\[
\mathbb{P}[|X| > \epsilon^{1/8}] \leq a_2 \epsilon^{1/8}.
\]

**Proof.** Let \( Z \sim \mathcal{N}(0,1) \perp X \). For any \( \delta \in (0,1) \), Pinsker’s inequality yields
\[
\mathbb{P}[Z \in B(0,\delta)] - \mathbb{P}[Z + X \in B(0,\delta)] \leq d_{TV}(P_Z, P_{Z+X}) \leq \sqrt{\frac{\epsilon}{2}}.
\]
Observe that
\[
\mathbb{P}[Z + X \in B(0,\delta)] = \mathbb{P}[Z \in B(-X,\delta) \mid |X| \leq 3\delta] \mathbb{P}[|X| < 3\delta] + \mathbb{P}[Z \in B(-X,\delta) \mid |X| > 3\delta] \mathbb{P}[|X| > 3\delta]
\]
\[
\leq \mathbb{P}[Z \in B(0,\delta)] \mathbb{P}[|X| \leq 3\delta] + \mathbb{P}[Z \in B(3\delta,\delta)] \mathbb{P}[|X| > 3\delta]
\]
\[
= \mathbb{P}[|X| > 3\delta] (\mathbb{P}[Z \in B(3\delta,\delta) - \mathbb{P}[Z \in B(0,\delta)]) + \mathbb{P}[Z \in B(0,\delta).
\]
Consequently,
\[
\mathbb{P}[|X| > 3\delta] (\mathbb{P}[Z \in B(0,\delta)] - \mathbb{P}[Z \in B(3\delta,\delta)]) \leq \sqrt{\frac{\epsilon}{2}}
\]
Since
\[
\mathbb{P}[Z \in B(0,\delta)] - \mathbb{P}[Z \in B(3\delta,\delta)] \geq 2\delta(\varphi(\delta) - \varphi(2\delta)) \geq \frac{1}{4} \delta^3,
\]
then
\[
\mathbb{P}[|X| > 3\delta] \leq \frac{\delta^{-3}}{4} \sqrt{\frac{\epsilon}{2}}.
\]
The result follows by choosing \( \delta = \epsilon^{1/8} \) with constant \( a_2 = 27/4\sqrt{2} \). \[\square\]
Proof of Theorem 3. We will show an equivalent statement: If \( t > 0 \) is such that \( C(\gamma) - F_t(t, \gamma) \leq \epsilon \) then
\[
t \geq \frac{1}{4} \ln \frac{1}{\epsilon} - \ln c_1(\gamma). \tag{68}
\]
Since \( t \geq 0 \), by choosing \( \ln c_1(\gamma) \geq \frac{1}{4} \ln \frac{1}{\epsilon} \), it suffices to consider \( \epsilon \geq \frac{1}{4} \). Observe that
\[
I(W; Y_\gamma) = C(\gamma) - D(P_{\sqrt{\gamma}X} \ast \mathcal{N}(01)) - I(X; Y_\gamma|W). \tag{69}
\]
Therefore, if \( I(W; Y_\gamma) \) is close to \( C(\gamma) \), then (a) \( P_X \) needs to be Gaussian like, and (b) \( P_{X|W} \) needs to be almost deterministic with high \( P_W \)-probability. Consequently, \( P_{X|W} \) and \( P_X \) are close to being mutually singular and hence \( I(W; X) \) will be large, since
\[
I(W; X) = D(P_{X|W}||P_X|P_W).
\]
Let \( \tilde{X} \triangleq \sqrt{\gamma}X \) and then \( W \rightarrow \tilde{X} \rightarrow Y_\gamma \). Define
\[
d(x, w) \triangleq D(P_{Y_\gamma|\tilde{X}=x}||P_{Y_\gamma|W=w}) = D(\mathcal{N}(x1)||P_{\tilde{X}|W=w} \ast \mathcal{N}(01)). \tag{70}
\]
Then \( (x, w) \rightarrow d(x, w) \) is jointly measurable\(^2\) and \( I(X; Y|W) = \mathbb{E}[d(\tilde{X}, W)] \). Similarly, \( w \rightarrow \tau(w) \triangleq D(P_{X|W=w}||P_X) \) is measurable and \( I(X; W) = \mathbb{E}[\tau(W)] \). Since \( \epsilon \geq I(X; Y|W) \) in view of (69), we have
\[
\epsilon \geq \mathbb{E}[d(\tilde{X}, W)] \geq 2\epsilon \cdot \mathbb{P}[d(\tilde{X}, W) \geq 2\epsilon]. \tag{71}
\]
Therefore
\[
\mathbb{P}[d(\tilde{X}, W) < 2\epsilon] > \frac{1}{2}. \tag{72}
\]
Denote \( B(x, \delta) \triangleq [x - \delta, x + \delta] \). In view of Lemma 6, if \( d(x, w) < 2\epsilon \), then
\[
\mathbb{P}[\tilde{X} \in B(x, \epsilon^{1/8})|W = w] = \mathbb{P} \left[ X \in B \left( \frac{x}{\sqrt{\gamma}}, \frac{\epsilon^{1/8}}{\sqrt{\gamma}} \right)|W = w \right] \geq 1 - a_2\epsilon^{1/8}.
\]
Therefore, with probability at least \( 1/2 \), \( \tilde{X} \) and, consequently, \( X \) is concentrated on a small ball. Furthermore, Lemma 5 implies that there exist absolute constants \( a_3 \) and \( a_4 \) such that
\[
\mathbb{P} \left[ X \in B \left( \frac{x}{\sqrt{\gamma}}, \frac{\epsilon^{1/8}}{\sqrt{\gamma}} \right) \right] \leq \mathbb{P} \left[ Z \in B \left( \frac{x}{\sqrt{\gamma}}, \frac{\epsilon^{1/8}}{\sqrt{\gamma}} \right) \right] + 2d_{KS}(F_X, \mathcal{N}(01)) \leq \frac{\sqrt{2}\epsilon^{1/8}}{\gamma^{1/4}} + a_3 \sqrt{\frac{1}{\gamma \ln \left( \frac{2}{4\epsilon} \right)}} + a_4 (1 + \gamma) (\gamma \epsilon)^{1/4} \sqrt{\ln \left( \frac{\gamma}{4\epsilon} \right)} \leq \kappa(\gamma) \left( \ln \frac{1}{\epsilon} \right)^{-1/2},
\]
\(^2\)By definition of the Markov kernel, both \( x \mapsto P_{Y_\gamma|A|\tilde{X}=x} \) and \( w \mapsto P_{Y_\gamma|A|W=w} \) are measurable for any measurable subset \( A \). Let \( \lceil k \rceil \triangleq \lceil ky \rceil / k \) denote the uniform quantizer. By the data processing inequality and the lower semicontinuity of divergence, we have \( D(P_{Y_\gamma|\tilde{X}=x}||P_{Y_\gamma|W=w}) \rightarrow D(P_{Y_\gamma|\tilde{X}=x}||P_{Y_\gamma|W=w}) \) as \( k \rightarrow \infty \). Therefore the joint measurability of \( (x, w) \mapsto D(P_{Y_\gamma|\tilde{X}=x}||P_{Y_\gamma|W=w}) \) follows from that of \( (x, w) \mapsto D(P_{Y_\gamma|\tilde{X}=x}||P_{Y_\gamma|W=w}) \).
where $\kappa(\gamma)$ is some positive constant depending only on $\gamma$. Therefore, for any $w \in B$ and $\epsilon$ sufficiently small, denoting $E = B(\frac{x}{\sqrt{\gamma}}, \frac{\epsilon^{1/8}}{\sqrt{\gamma}})$, we have by data processing inequality:

$$
\tau(w) = D(P_{X|W=w} \| P_X) \geq P_{X|W=w}(E) \ln \frac{P_{X|W=w}(E)}{P_X(E)} + P_{X|W=w}(E^c) \ln \frac{P_{X|W=w}(E^c)}{P_X(E^c)} \geq \frac{1}{2} \ln \ln \frac{1}{\epsilon} - \ln \kappa(\gamma) - a_5,
$$

where $a_5$ is an absolute positive constant. Combining (74) with (72) and letting $c_1^2(\gamma) \triangleq e^{a_5} \kappa(\gamma)$, we obtain

$$
\mathbb{P}\left[\tau(W) \geq \frac{1}{2} \ln \ln \frac{1}{\epsilon} - 2 \ln c_1(\gamma)\right] \geq \mathbb{P}[d(\tilde{X}, W) < 2\epsilon] \geq \frac{1}{2},
$$

which implies that $I(W; X) = \mathbb{E}[\tau(W)] \geq \frac{1}{4} \ln \ln \frac{1}{\epsilon} - \ln c_1(\gamma)$, proving the desired (68).

**Remark 4.** The double-exponential convergence rate in Theorem 3 is in fact sharp. To see this, note that [WV10, Theorem 8] showed that there exists a sequence of zero-mean and unit-variance random variables $X_m$ with $m$ atoms, such that

$$
C(\gamma) - I(X_m; \sqrt{\gamma} X_m + Z) \leq 4(1 + \gamma) \left(\frac{\gamma}{1 + \gamma}\right)^{2m}.
$$

Consequently,

$$
C(\gamma) - F_I(t, \gamma) \leq C(\gamma) - F_I(\ln[\epsilon^t], \gamma) \leq 4(1 + \gamma) \left(\frac{\gamma}{1 + \gamma}\right)^{2(\epsilon^t-1)} = e^{-2\epsilon^t \ln \frac{1+\gamma}{\gamma} + O(\ln \gamma)},
$$

proving the right-hand side of (11).

7 Deconvolution results for total variation

The proof of the horizontal gap for the scalar AWGN channel in Section 6 consists of four steps:

(a) Notice that if $C(\gamma) - I(W; Y)$ is small, then both $X$ is Gaussian-like and $P_X$ and $P_{X|W}$ are close to being mutually singular;

(b) Use Lemma 5 to show that $P_X$ cannot be concentrated on any ball of small radius if it is Gaussian-like;

(c) Apply Lemma 6 to show that $P_{X|W}$, in turn, is concentrated on a small ball with high $W$-probability;

(d) Use (75) to show that $I(W; X)$ must explode.

In Section 8, we will implement the above program to extend the results in Theorem 3 (i.e. $I(W; Y)$ approaches capacity only as $I(W; X) \to \infty$) for a range of noise distributions. We also generalize the moment constraint on the input distribution, allowing $P_X$ to be restricted to an arbitrary convex set. However, the extension of the AWGN result to a wider class of noise distributions
requires new deconvolution results that are similar in spirit to Lemmas 5 and 6. These results are the focus of the present section.

If \( \mathcal{P} \) is convex and \( C(\mathcal{P}) \triangleq \sup_{P_X \in \mathcal{P}} I(X;Y) < \infty \), then there exists a unique capacity-achieving output distribution \( P_Y^* \). [Kem74]. In addition, by the saddle-point characterization of capacity,

\[
C(\mathcal{P}) = \sup_{P_X \in \mathcal{P}} D(P_{Y|X} \Vert P_Y^* | P_X).
\]

Consequently, for any \( P_X \in \mathcal{P} \), we can decompose

\[
I(W;Y) = I(X;Y) - I(X;Y|W) \leq C(\mathcal{P}) - D(P_Y \Vert P_Y^*) - I(X;Y|W). \tag{77}
\]

If the capacity-achieving input distribution \( P_X^* \) is unique, then the same intuition for the Gaussian case should hold: (i) \( P_X \) must be close to the capacity achieving input distribution \( P_X^* \) and (ii) \( P_{X|W} \) must be concentrated on a small ball with high probability. Therefore, as long as \( P_X^* \) is assumed to have no atoms, then \( P_{X|W} \) and \( P_X \) are close to being mutually singular, which, in view of the fact that

\[
I(W;X) = D(P_{X|W} \Vert P_X | P_W), \tag{78}
\]

implies that \( I(W;X) \) will explode.

In order to make this proof concrete, we require additional results to quantify the distance between \( P_X \) and \( P_X^* \) (analogous to Lemma 5 in the Gaussian case), and to show that \( P_{X|W} \) is concentrated in a small ball (analogous to Lemma 6) for general \( P_Z \). These are precisely the results we present in this section, once again making use of Lemma 3 and Fourier-analytic tools. In particular, we prove a deconvolution result in terms of total variation for a wide range of additive noise distributions \( P_Z \) (e.g., Gaussian, uniform). The main result in this section (Theorem 4 and Corollary 3) states that, under first moment constraints and certain conditions on the characteristic function of \( P_Z \) (e.g., no zeros, cf. Lemma 7), if \( d_{TV}(P * P_Z, Q * P_Z) \) is small and \( Q \) has a bounded density, then \( d_{KS}(P,Q) \) is also small.

Let \( v : \mathbb{R} \to \mathbb{R} \) be the positive, symmetric function

\[
v(x) \triangleq \frac{2(1 - \cos x)}{x^2} \tag{79}
\]

and \( \hat{v} \) its Fourier transform

\[
\hat{v}(\omega) \triangleq \int v(x)e^{i\omega x} \, dx = 2\pi (1 - \vert \omega \vert)^+, \tag{80}
\]

where \((x)^+ \triangleq \max\{x,0\} \).

We have the following deconvolution lemma.

**Lemma 7.** Assume \( P_Z \) has density bounded by \( m_1 \) and that there exists a decreasing function \( g_1 : (0,1] \to \mathbb{R}^+ \) with \( g_1(0+) = \infty \) such that

\[
\text{Leb} \{ \omega : \vert \varphi_Z(\omega) \vert \leq \sqrt{u}, \vert \omega \vert \leq g_1(u) \} \leq \sqrt{g_1(u)} \quad \forall u \in (0,1]. \tag{81}
\]

Then for all distributions \( P,Q \) and all \( x_0 \in \mathbb{R}:

\[
\left| \mathbb{E}_P [v(TX - x_0)] - \mathbb{E}_Q [v(TX - x_0)] \right| \leq \frac{c}{\sqrt{T}}, \quad T = g_1 \left( m_1 d_{TV}(P * P_Z, Q * P_Z) \right), \tag{82}
\]

where \( c \) is an absolute constant.
**Remark 5.** 1. The implication of the previous lemma is that $P$ and $Q$ are almost the same on all balls of size approximately $\frac{1}{T}$.

2. For Gaussian $P_Z$, $g_1(u) = \sqrt{-\ln u}$. For uniform $P_Z$, $g_1(u) = u^{-1/3}$.

3. Without assumptions similar to those of Lemma 7, it is impossible to have any deconvolution inequality. For example, if $\varphi_Z = 0$ outside of a neighborhood of 0 (e.g. $p_Z$ is proportional to (79)), then one may have $P*P_Z = Q*P_Z$, but $P \neq Q$.

**Proof.** Denote the density of $Z$ by $p_Z$. From Plancherel’s theorem, we have

\[
\| (\varphi_P - \varphi_Q) \varphi_Z \|^2 = 2\pi \| P * p_Z - Q * p_Z \|^2 \\
\leq 2\pi \| P * p_Z - Q * p_Z \|_1 \| P * p_Z - Q * p_Z \|_\infty \\
\leq 4\pi m_1 \text{d}_\text{TV}(P * p_Z, Q * p_Z) = 4\pi \delta,
\]

where the first inequality follows from Hölder’s inequality, and the second inequality follows from $\| (P * p_Z - Q * p_Z) \|_\infty \leq \max\{\| P * p_Z \|_\infty, \| Q * p_Z \|_\infty\} \leq \| p_Z \|_\infty$.

Assume there exist positive functions $g$ and $h$ and $T > 0$ such that

\[
|\{\omega : |\varphi_Z(\omega)| \leq g(T), |\omega| \leq T\}| \leq h(T).
\]

Put $D = \{\omega : |\varphi_Z(\omega)| \leq g(T), |\omega| \leq T\}$ and $D^c = [-T, T] \setminus D$. Then

\[
\frac{1}{T} \int_{-T}^{T} |\varphi_P(\omega) - \varphi_Q(\omega)| d\omega = \frac{1}{T} \left( \int_D |\varphi_P(\omega) - \varphi_Q(\omega)| d\omega + \int_{D^c} |\varphi_P(\omega) - \varphi_Q(\omega)| d\omega \right) \\
\overset{(84)}{\leq} \frac{h(T)}{T} + \frac{1}{T} \int_{D^c} |\varphi_P(\omega) - \varphi_Q(\omega)| \left( \frac{|\varphi_Z(\omega)|}{g(T)} \right) d\omega \\
\leq \frac{h(T)}{T} + \frac{1}{T g(T)} \int_{-T}^{T} |\varphi_P(\omega) - \varphi_Q(\omega)||\varphi_Z(\omega)| d\omega \\
\leq \frac{h(T)}{T} + \frac{\sqrt{2}\| (\varphi_P - \varphi_Q) \varphi_Z \|_2}{g(T) \sqrt{T}} \\
\overset{(83)}{\leq} \frac{h(T)}{T} + \frac{\sqrt{8\pi \delta}}{\sqrt{T} g(T)},
\]

where the third inequality follows Cauchy-Schwartz inequality.

Note that it is sufficient to consider $x_0 = 0$, since otherwise we can simply shift the distributions $P$ and $Q$ without affecting the value of $\delta$. In addition, Plancherel’s theorem and (80) yield

\[
\mathbb{E}_P \left[ v(TX) \right] = \frac{1}{T} \int_{-T}^{T} \varphi_P(\omega) \left( 1 - \frac{|\omega|}{T} \right) d\omega.
\]

Thus, we have

\[
|\mathbb{E}_P [v(TX)] - \mathbb{E}_Q [v(TX)]| \leq \frac{1}{T} \int_{-T}^{T} |\varphi_P(\omega) - \varphi_Q(\omega)| d\omega \\
\leq \frac{h(T)}{T} + \frac{\sqrt{8\pi \delta}}{\sqrt{T} g(T)}.
\]

Finally, choosing $T = g_1(\delta)$, $h(T) = \sqrt{T}$ and $g(T) = \sqrt{\delta}$, the result follows.

\[\square\]
The methods used in the proof of the previous theorem and, in particular, Eq. (85), can be used to bound the KS-distance between $P$ and $Q$, as demonstrated in the next theorem.

**Theorem 4.** Assume $P_Z$ has density bounded by $m_1$ and that there exists functions $g(T)$ and $h(T)$ that satisfy assumption (84). Then for any pair of distributions $P$, $Q$ where $Q$ has a density bounded by $m_2$ we get for all $T > 0$:

$$d_{KS}(P, Q) \leq \frac{Th(T)}{\pi} + \frac{24m_2 + 2(\mathbb{E}_P [|X|] + \mathbb{E}_Q [|X|])}{\pi T} + \frac{(2T)^{3/2}}{\sqrt{\pi} g(T)} \sqrt{m_1 d_{TV}(P * P_Z, Q * Q_Z)}, \quad (87)$$

Proof. By assumption, we can choose $g(T)$ in as (91) and $h(T) = 0$ to fulfill (84). Then (87) leads to

$$d_{KS}(P, Q) \leq \frac{C}{T(d_{TV}(P * P_Z, Q * Q_Z))}, \quad (92)$$

where $C$ is a constant depending only on $m_1$ and $m_2 + \mathbb{E}_P [|X|] + \mathbb{E}_Q [|X|]$.

In particular, for $Z \sim \mathcal{N}(0, 1)$,

$$d_{KS}(P, Q) \leq C' \left( \log \frac{1}{d_{TV}(P * \mathcal{N}(0, 1), Q * \mathcal{N}(0, 1))} \right)^{-1/2}. \quad (93)$$

where $C'$ is a constant depending only on $m_2 + \mathbb{E}_P [|X|] + \mathbb{E}_Q [|X|]$.

Proof. By assumption, we can choose $g(T)$ in as (91) and $h(T) = 0$ to fulfill (84). Then (87) leads to

$$d_{KS}(P, Q) \leq \frac{C}{T} \left( 1 + \frac{\sqrt{d_{TV}(P * P_Z, Q * Q_Z) \cdot T^5}}{g(T)} \right),$$

where $C_0 = (\max\{24m_2 + 2(\mathbb{E}_P [|X|] + \mathbb{E}_Q [|X|]), \sqrt{8m_1 \pi}\})/\pi$. Since $P_Z$ has a density, $g(T) \leq |\psi_Z(T)| \to 0$ by Riemann-Lebesgue lemma. Since $g(T)$ is decreasing and $g(0) = 1$, $\alpha T^5 = g^2(T)$ always has a unique solution $T(\alpha) > 0$. Choosing $T = T(d_{TV}(P * P_Z, Q * Q_Z))$ yields $d_{KS}(P, Q) \leq 2C_0/T$, completing the proof. When $Z \sim \mathcal{N}(0, 1)$, we have $g(T) = e^{-T^2/2}$. Choosing $T = \sqrt{-\log d_{TV}(P * P_Z, Q * P_Z)}$, the result follows.
Remark 6. Consider a Gaussian $Z$. Then $P_n \xrightarrow{w} Q \iff P_n * P_Z \xrightarrow{w} Q * P_Z \iff P_n * P_Z \xrightarrow{TV} P * P_Z$, where the last part follows from pointwise convergence of densities (Scheffe’s lemma, see, e.g., [Pet95, 1.8.34]). Furthermore, when one of the distributions has bounded density the Levy-Prokhorov distance (that metrizes weak convergence) is equivalent to the Kolmogorov-Smirnov distance, cf. [Pet95, 1.8.32]. In this perspective, Theorem 4 can be viewed as a finitary version of the implication $d_{TV}(P_n * P_Z, Q * P_Z) \to 0 \Rightarrow d_{KS}(P_n, Q) \to 0$.

Remark 7. A slightly better bound may be obtained if $\mathbb{E}_{P,Q}[|X + Z|^2] < \infty$. Namely, $T^3$ in the third term in (87) can be reduced to $T$. Indeed if $\delta = d_{TV}(P * P_Z, Q * P_Z)$ then elementary truncation shows

$$W_1(P * P_Z, Q * P_Z) \lesssim \sqrt{\delta}$$

and then following (108) we get

$$|\phi_P(\omega) - \phi_Q(\omega)| |\phi_Z(\omega)| \lesssim \sqrt{\delta} |\omega|.$$ 

Now the left-hand side of (88) can be bounded by $\frac{T}{g(T)}$ for the choice of $g(T)$ as in (91) and a straightforward modification for the general case of (84). This improves the constant in (93).

8 Horizontal bound for general additive noise

With the results introduced in the previous section in hand, we are now ready to extend Theorem 3 to a broader class of additive noise and channel input distributions.

Theorem 5. Let $Y = X + Z$ and let $\mathcal{P}$ be a convex set of distributions. Assume that

(a) $P_Z$ satisfies the assumption of Lemma 7;

(b) The capacity $C(\mathcal{P}) \triangleq \sup_{P_X \in \mathcal{P}} I(X;Y)$ is finite and attained at some $P_{X^*} \in \mathcal{P}$.

Then there exists a constant $c_0$ and a decreasing function $\rho : (0, c_0) \to (0, \infty)$ (depending on $P_Z$ and $\mathcal{P}$), such that any $P_{W,X}$ with $P_X \in \mathcal{P}$ satisfies

$$I(W;X) \geq \rho(C(\mathcal{P}) - I(W;Y)).$$  \hspace{1cm} (94)

Furthermore, if $P_{X^*}$ has no atoms, then $\rho$ satisfies $\rho(0+) = \infty$.

Remark 8. Theorem 5 translates into the following bound on the gap between the $F_I$ curve and the capacity:

$$F_I(t) \leq C(\mathcal{P}) - \rho^{-1}(t).$$

The function $\rho$ can be chosen to be

$$\rho(\epsilon) = -\frac{1}{2} \ln \left( \mathcal{L}(X^*; T^{-3/4}) + \frac{4 + 2c}{\sqrt{T}} \right),$$  \hspace{1cm} (95)

where $T = g_1(m_1 \sqrt{t})$, $c$, $g_1$, $m_1$ are as in Lemma 7, and

$$\mathcal{L}(X^*; \delta) \triangleq \sup_{x \in \mathbb{R}} P[X^* \in B(x, \delta)]$$  \hspace{1cm} (96)

is the Lévy concentration function [Pet95, p. 22] of $X^*$. For the AWGN channel with $P_Z \sim \mathcal{N}(0,1)$ and $\mathcal{P} = \{P_X : \mathbb{E}[X^2] \leq \gamma \}$ this gives

$$\rho(\epsilon) = \frac{1}{8} \ln \ln \frac{1}{\epsilon} + c_0(\gamma)$$

for some constant $c_0(\gamma)$. Compared to the Gaussian-specific bound (68), the general proof loses a factor of two, which is due to the application of Pinsker’s inequality.
Proof. Throughout the proof we assume that
\[ C(P) - I(W; Y) \leq \epsilon, \]  
and, from \((77)\), \(I(X; Y|W) \leq \epsilon\) and \(D(P_X \ast P_Z \| P_X \ast P_Z) \leq \epsilon\), where \(P_X \ast\) is capacity-achieving. Denote
\[ t(x, w) \triangleq d_{TV}(P_{Z+x}, P_{X|W=w} \ast P_Z), \]
which is joint measurable in \((x, w)\) for the same reason that \(d\) defined in \((70)\) is jointly measurable. Pinsker’s inequality yields
\[ \epsilon \geq I(X; Y|W) = \mathbb{E}_{X, W}[D(P_{Z+W} \| P_X \ast P_Z)] \geq 2\mathbb{E}[t(X, W)^2] \geq 2\epsilon \mathbb{P}[t(X, W)^2 \geq \epsilon]. \]  
Define
\[ \mathcal{F} \triangleq \{(x, w) : t(x, w) \leq \sqrt{\epsilon}\} \]
\[ \mathcal{G} \triangleq \{w : \exists x, t(x, w) \leq \sqrt{\epsilon}\}. \]

Then, from \((98)\),
\[ \mathbb{P}[W \in \mathcal{G}] \geq \mathbb{P}[(X, W) \in \mathcal{F}] \geq \frac{1}{2}. \]  
Therefore, for any \(w \in \mathcal{G}\), there exists \(\hat{x}_w \in \mathbb{R}\) such that \(t(x, \hat{x}_w) \leq \sqrt{\epsilon}\). Applying Lemma 7 with \(P = P_{X|W=w}, Q = \delta_{\hat{x}_w}\) and \(x_0 = T\hat{x}_w\), we conclude that
\[ |\mathbb{E}[v(T(X - \hat{x}_w))|W = w] - 1| \leq \frac{c}{\sqrt{T}}, \]  
where \(v\) is defined in \((79)\), \(c\) is the absolute constant in \((82)\) and \(T = g_1(m_1 \sqrt{\epsilon})\).

On the other hand, \((97)\) implies that \(D(P_X \ast P_Z \| P_{Y\ast}) \leq \epsilon\) and hence \(d_{TV}(P_X \ast P_Z, P_{Y\ast}) \leq \sqrt{\epsilon}\) by Pinsker’s inequality. Applying Lemma 7 with \(P = P_X, Q = P_{X\ast}\) and \(x_0 = T\hat{x}_w\), we have
\[ |\mathbb{E}[v(T(X - \hat{x}_w))] - \mathbb{E}[v(T^\ast - \hat{x}_w)]| \leq \frac{c}{\sqrt{T}}. \]  
For any \(x\), since \(0 \leq v \leq 1\),
\[ \mathbb{E}[v(T(X\ast - x))] = 2\mathbb{E}\left[\frac{1 - \cos(T(X\ast - x))}{T^2(X\ast - x)^2}\right] \leq \mathbb{P}[X\ast \in B(x, T^{-3/4})] + \frac{4}{\sqrt{T}}. \]  
Therefore,
\[ 0 \leq \mathbb{E}[v(T(X\ast - x))] \leq \mathcal{L}(X\ast; T^{-3/4}) + \frac{4}{\sqrt{T}} \]  
(102)

Note that the function \(v\) takes values in \([0, 1]\). Using the fact that
\[ d_{TV}(P, Q) = \sup_{|\phi| \leq 1} \int f dP - \int f dQ, \]
and assembling (100)-(102), we have for any \( w \in \mathcal{G} \)

\[
d_{TV}(P_X, P_{X|W=w}) \geq \mathbb{E} [v(T(X - \hat{x}_w)) | W = w] - \mathbb{E} [v(T(X - \hat{x}_w))]
\geq 1 - \mathcal{L}(X^*; T^{-3/4}) - \frac{4 + 2c}{\sqrt{T}}. \tag{103}
\]

Using (78) and the fact that \( D(P\|Q) \geq -\ln(1 - d_{TV}(P, Q)) \), we have

\[
I(W; X) \geq \mathbb{E} \left[ \ln \frac{1}{1 - d_{TV}(P_X, P_{X|W})} \right]
\geq \mathbb{E} \left[ \ln \frac{1}{1 - d_{TV}(P_X, P_{X|W})} 1_{W \in \mathcal{G}} \right]
\geq \frac{1}{2} \mathbb{E} \left[ \ln \frac{1}{\mathcal{L}(X^*; T^{-3/4}) + \frac{4 + 2c}{\sqrt{T}}} \right],
\]

where the last inequality follows from (99) and (103). Lemma 9 in Appendix B implies that 
\( \mathcal{L}(X^*; 0+) = \max_{x \in \mathbb{R}} \mathbb{P} [X = x] < 1 \). Denote by \( \epsilon_0 \) the supremum of \( \epsilon \) such that \( \mathcal{L}(X^*; T^{-3/4}) + \frac{4 + 2\epsilon}{T^{3/4}} < 1 \) and define \( \rho(\epsilon) \) as in (95). This completes the proof of (94). Finally, by Lemma 9 we have that for diffuse \( P_{X^*} \) it holds that \( \rho(0+) = \infty \).

9 Infinite-dimensional case

It is possible to extend the results and proof techniques to the case when the channel \( X \mapsto Y \) is a \( d \)-dimensional Gaussian channel subject to a total-energy constraint \( \mathbb{E} \left[ \sum_i X_i^2 \right] \leq 1 \). Unfortunately, the resulting bound strongly depends on the dimension; in particular, it does not improve the trivial estimate (7) as \( d \to \infty \). It turns out that this dependence is unavoidable as we show next that (7) holds with equality when \( d = \infty \).

To that end we consider an infinite-dimension discrete-time Gaussian channel. Here the input \( X = (X_1, X_2, \ldots) \) and \( Y = (Y_1, Y_2, \ldots) \) are sequences, where \( Y_i = X_i + Z_i \) and \( Z_i \sim \mathcal{N}(0, 1) \) are i.i.d. Similar to Definition 1, we define

\[
F_{I^\infty}^X(t, \gamma) = \sup \{ I(W; Y) : I(W; X) \leq t, W \to X \to Y \}, \tag{104}
\]

where the supremum is over all \( P_{W,X} \) such that \( \mathbb{E} \left[ \|X\|_2^2 \right] = \mathbb{E} \left[ \sum_i X_i^2 \right] \leq \gamma \). Note that, in this case,

\[
F_{I^\infty}^X(t, \gamma) \leq \min \{t, \gamma/2\}. \tag{105}
\]

The next theorem shows that unlike in the scalar case, there is no improvement over the trivial upper bound (105) in the infinite-dimensional case. This is in stark contrast with the strong data processing behavior of total variation in Gaussian noise which turns out to be dimension-free [PW16, Corollary 6].

Theorem 6. \( F_{I^\infty}^X(t, \gamma) = \min \{t, \gamma/2\} \).

Proof. For any \( \epsilon > 0 \) and all sufficiently large \( \beta > 0 \), there exists \( n \) and a code of size of \( M_\beta \) for the \( n \)-parallel Gaussian channel, where each codeword has energy (squared \( \ell_2 \)-norm) less than \( \beta \), the probability of error is at most \( \epsilon \), and \( M_\beta = e^{\beta/2 + o(\beta)} \) as \( \beta \to \infty \) (see, e.g. Gal68, Thm. 7.5.2]). Choosing \( X \) uniformly at random over the codewords, we have from Fano’s inequality

\[
I(X; Y) \geq (1 - \epsilon) \ln M - h(\epsilon) = \frac{(1 - \epsilon)\beta}{2} + o(\beta) - h(\epsilon).
\]
For any \( \beta > \gamma \), define

\[
X' = \begin{cases} 
x_0 & \text{w.p. } 1 - \frac{\gamma}{\beta} \\
X & \text{w.p. } \frac{\gamma}{\beta}.
\end{cases}
\]

where \( x_0 \) is an arbitrary vector outside the codebook. Then, \( \mathbb{E}[\|X'\|_2^2] \leq \gamma \). Furthermore, as \( \beta \to \infty \),

\[
H(X') = \frac{\gamma}{\beta} \ln M + h \left( \frac{\gamma}{\beta} \right) = \frac{\gamma}{2} + o(1),
\]

and, by the concavity of the mutual information in the input distribution,

\[
I(X'; Y) \geq \frac{\gamma}{\beta} I(X; Y) \geq \frac{(1 - \epsilon)\gamma}{2} + o(1).
\]

Since \( F_{I,2}(\gamma/2, \gamma) \geq \frac{I(X; Y)}{H(X)} \), first sending \( \beta \to \infty \) then \( \epsilon \to 0 \), we have \( F_{I,2}(\gamma/2, \gamma) = \gamma/2 \). The result then follows by noting that \( t \mapsto F_{I,2}(t, \gamma)/t \) is decreasing and \( t \mapsto F_{I,2}(t, \gamma) \) is increasing (Proposition 1).

**Appendix A  Alternative version of Lemma 5**

**Lemma 8.** Assume that \( C(\gamma) - I(X; Y, \gamma) \leq \epsilon < 1 \). Then

\[
d_{KS}(P_X, \mathcal{N}(0, 1)) \leq \frac{24}{\pi^{3/2} \sqrt{\gamma \log(1/\epsilon)}} + \frac{2\sqrt{2(1+\gamma)}\epsilon^{1/4} \sqrt{\log(1/\epsilon)}}{\pi} \tag{106}
\]

**Proof.** Abbreviate \( Y_\gamma = \sqrt{\gamma}X + Z \) by \( Y \). From Talagrand’s inequality [Tal96, Thm 1.1]

\[
W_2(P, \mathcal{N}(0, 1), \mathcal{N}(0, \gamma + 1)) \leq 2\sqrt{1 + \gamma}\epsilon.
\]

Since \( W_1(\mu, \nu) \leq W_2(\mu, \nu) \) for any measures \( \mu, \nu \), there exists a random variable \( G \sim \mathcal{N}(0, \gamma + 1) \) such that

\[
\mathbb{E} [|Y - G|] \leq 2\sqrt{1 + \gamma}\epsilon. \tag{107}
\]

Let \( \varphi_Y(t) \) and \( \varphi_G(t) \) be the characteristic functions of \( Y \) and \( G \), respectively. Then

\[
|\varphi_Y(t) - \varphi_G(t)| = |\mathbb{E} [e^{itY} - e^{itG}]| \leq \mathbb{E} |t(Y - G)| \leq 2|t|\sqrt{1 + \gamma}\epsilon \tag{108}
\]

where the second inequality follows from [Fel66, Lemma 4.1], and the last inequality from (107). Using Esseen’s inequality (Lemma 3) and the fact that the PDF of \( G \) is upper bounded by \( 1/\sqrt{2\pi T} \), for all \( T > 0 \)

\[
\left| P_{\sqrt{\gamma}X}(t) - \mathcal{N}(0, P) \right| \leq \frac{1}{\pi} \int_{-T}^{T} \left| \varphi_X(t) - e^{-\gamma t^2/2} \right| \frac{dt}{t} + \frac{12\sqrt{2}}{\pi^{3/2} T \sqrt{\gamma}}
\]

\[
= \frac{1}{\pi} \int_{-T}^{T} e^{t^2/2} \left| \frac{\varphi_Y(t) - \varphi_G(t)}{t} \right| \frac{dt}{t} + \frac{12\sqrt{2}}{\pi^{3/2} T \sqrt{\gamma}}
\]

\[
\leq \frac{4\sqrt{(1+\gamma)}\epsilon T e^{T^2/2}}{\pi} + \frac{12\sqrt{2}}{\pi^{3/2} T \sqrt{\gamma}}.
\]
Choosing \( T = \sqrt{\frac{1}{2} \log(1/\epsilon)} \) yields

\[
\left| P_{\gamma X}(t) - \mathcal{N}(0, \gamma) \right| \leq \frac{2\sqrt{2(1 + \gamma)} \epsilon^{1/4} \sqrt{-\log(\epsilon)}}{\pi} + \frac{24}{\pi^{3/2} \sqrt{-\gamma \log(\epsilon)}}.
\]

(109)

The proof is complete upon observing that \( d_{KS}(P_{\sqrt{\gamma X}}, \mathcal{N}(0, \gamma)) = d_{KS}(P_X, \mathcal{N}(0, 1)) \).

Appendix B Lévy concentration function near zero

We show that the Lévy concentration function defined in (96) is continuous at zero if and only if the distribution has no atoms.

**Lemma 9.** For any \( X \), \( \lim_{\delta \to 0} \mathcal{L}(X; \delta) = \max_{x \in \mathbb{R}} P \left[ X = x \right] \). Consequently, \( \mathcal{L}(X; 0+) = 0 \) if and only if \( X \) has no atoms.

**Proof.** Let \( a \triangleq \lim_{\delta \to 0} \mathcal{L}(X; \delta) \), which exists since \( \delta \mapsto \mathcal{L}(X; \delta) \) is increasing. Since \( \mathcal{L}(X; \delta) \geq P \left[ X = x \right] \) for any \( \delta > 0 \) and any \( x \), it is sufficient to show that \( a \leq \max_{x \in \mathbb{R}} P \left[ X = x \right] \). Assume that \( a > 0 \) for otherwise there is nothing to prove. By definition, for any \( n \), there exists \( x_n \) so that \( P \left[ X \in B(x_n, 1/n) \right] \geq a - 1/n \). Let \( T > 0 \) so that \( P \left[ |X| > T \right] \leq a/2 \). Then \( |x_n| \leq T \) for all sufficiently large \( n \). By restricting to a subsequence, we can assume that \( x_n \) converges to some \( x \) in \([-T, T]\). By triangle inequality, \( P \left[ X \in B(x, |x_n - x| + 1/n) \right] \geq P \left[ X \in B(x_n, 1/n) \right] \geq a - 1/n \). By bounded convergence theorem, \( P \left[ X = x \right] \geq a \), completing the proof.

References


