Second-Order Converses via Reverse Hypercontractivity*

Jingbo Liu   Ramon van Handel   Sergio Verdú

Abstract

A strong converse shows that no procedure can beat the asymptotic (as blocklength $n \to \infty$) fundamental limit of a given information-theoretic problem for any fixed error probability. A second-order converse strengthens this conclusion by showing that the asymptotic fundamental limit cannot be exceeded by more than $O(\frac{1}{\sqrt{n}})$. While strong converses are achieved in a broad range of information-theoretic problems by virtue of the "blowing-up method"—a powerful methodology due to Ahlswede, Gács and Körner (1976) based on concentration of measure—this method is fundamentally unable to attain second-order converses and is restricted to finite-alphabet settings. Capitalizing on reverse hypercontractivity of Markov semigroups and functional inequalities, this paper develops the "smoothing-out" method, an alternative to the blowing-up approach that does not rely on finite alphabets and that leads to second-order converses in a variety of information-theoretic problems that were out of reach of previous methods.

Keywords. Strong converse; information-theoretic inequalities; reverse hypercontractivity; blowing-up lemma; concentration of measure.

*This work was supported in part by NSF Grants CCF-1016625, CCF-0939370, DMS-1148711, by ARO Grants W911NF-15-1-0479, W911NF-14-1-0094, and by the Center for Science of Information. The main results of this paper were initially announced in the conference paper [32], and form a part of the first author's Ph.D. thesis at Princeton University [26].

Jingbo Liu, Institute for Data, Systems, and Society, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Bldg. E18-421, Cambridge, MA 02139, USA; e-mail: jingbo@mit.edu

Ramon van Handel, Fine Hall 207, Princeton University, Princeton, NJ 08544, USA; e-mail: rvan@princeton.edu

Sergio Verdú:e-mail: verdu@informationtheory.org

Mathematics Subject Classification (2010): 94A15; 94A24; 68P30; 62B10; 47D07; 60E15

Contents

1 Introduction **2**
1.1 Overview . . . 2
1.2 Weak, strong, and second-order converses . . . 4
1.3 "Blowing up" vs "smoothing out" . . . 7
1.4 Organization . . . 11
1.5 Notation . . . 12

2 Prelude: Binary Hypothesis Testing **12**
2.1 Setup . . . 12
2.2 The blowing-up method . . . 13
2.3 The smoothing-out method . . . 15
2.4 Achievability and optimality . . . 21

3 Second-Order Fano's Inequality **23**
3.1 Bounded probability density case . . . 24
3.2 Gaussian case . . . 27
3.3 Application: output distribution of good channel codes . . . 29
3.4 Application: broadcast channels . . . 31

4 Second-Order Image Size Characterization **34**
4.1 Synopsis . . . 34
4.2 Discrete case . . . 40
4.3 Gaussian case . . . 43
4.4 Application: hypothesis testing with a communication constraint . 46
4.5 Application: source coding with compressed side information . . . 48

A Data processing cannot yield second-order converses **52**

B Brascamp-Lieb divergence: auxiliary results **54**
B.1 Proof of (4.18) in the discrete case . . . 54
B.2 Proof of (4.18) in the Gaussian case . . . 57

1 Introduction

1.1 Overview

What are the fundamental limits of a given data science problem? The investigation of such questions typically follows a two-sided analysis. In the *achievability* part, one shows the existence of a procedure achieving a certain performance, e.g., error probability. In the *converse* part, one shows that no procedure can accomplish better performance than a certain lower bound. This basic structure is common to a wide range of problems that span information theory, statistics, and computer science. While the study of achievability generally depends strongly on the special features of the problem at hand, many converse bounds that arise in different areas rely essentially on information-theoretic methods.

Frequently, the upper and lower bounds on optimal performance provided by the achievability and converse analyses do not coincide. Starting with Shannon (1948) [47], the emphasis in information theory has been on the analysis of the fundamental limits in the *asymptotic* regime of long blocklengths. In that regime, the gap between many existing achievability and converse bounds does vanish and sharp answers can be found to questions such as the maximal transmission rate over noisy channels and the minimal data compression rate subject to a fidelity constraint. The last decade has witnessed a number of information-theoretic results (e.g., [42, 22, 23]), which are applicable in the *non-asymptotic* regime. These results are strongly motivated by the need to understand the relevance of asymptotic limits to practical systems that may be subject to severe delay constraints, or scenarios where the alphabet size or the number of users is large compared to the number of channels or sources [42, 39, 21, 41].

The development of a non-asymptotic information theory has required new and improved methods for investigating fundamental limits. In principle, it is not clear that the quantities that determine the fundamental limits in the asymptotic regime can accurately describe the performance at blocklengths of interest in realistic applications (e.g., 1000 bits). This has led to the development of nonasymptotic bounds that capture more sophisticated distributional information on the relevant information quantities. Unfortunately, however, such bounds are often difficult to compute. A less accurate but more tractable approach to understanding performance at smaller blocklengths is to focus attention on so-called *second-order* analysis, originally pioneered by Wolfowitz [55] and Strassen [50] in the 1960s and significantly refined in recent years. In such bounds, the fundamental limits in the first-order (linear in the blocklength) asymptotics are sharpened by investigating the deviation from this asymptotic behavior to second order (square-root of the blocklength). In particular, in those situations where the first- and second-order asymptotics can be established precisely, the resulting bounds have often proven to be quite accurate except for very short blocklengths.

While precise second-order results are available in various basic information-theoretic problems, more complicated setups, particularly those arising in multiuser information theory, have so far eluded second-order analysis. One of the challenges that emerged from this line of work is the development of second-order converses. Several existing approaches to obtaining second-order converses are briefly reviewed in Section 1.2; however, to date, a variety of information-theoretic problems have remained out of reach of such methods. On the other hand, the powerful general methodology introduced in 1976 by Ahlswede, Gács, and Körner [3] and exploited extensively in the classical book [11] (see also the recent survey [45]) has proven instrumental for proving converses in network information theory. Although this widely used technique yields converse results in a broad range of problems almost as a black box, it is fundamentally unable to yield second-order converses and is restricted to finite-alphabet settings.

Inspired by a result by Margulis [33], the method of Ahlswede, Gács, and Körner is based on a remarkable application of the concentration of measure phenomenon on the Hamming cube [25, 6], which is known in information theory as the "blowing-up lemma". Historically, this is probably the very first application of modern measure concentration to a data science problem. One of the main messages of this paper is that, surprisingly, measure concentration turns out not to be the right approach after all in this original application. Instead, we will revisit the theory of Ahlswede, Gács, and Körner based not on the violent "blowing-up" operation, but on a new and more pacifist "smoothing out" principle that exploits reverse hypercontractivity of Markov semigroups and functional inequalities. With this gentler touch, we are able to eliminate the inefficiencies of the blowing-up method and obtain second-order converses, essentially for free, in many information-theoretic problems that were out of reach of previous methods.

1.2 Weak, strong, and second-order converses

As a concrete basis for discussion, let us consider the basic setup of single-user data transmission through noisy channels, in which there is a three-way trade-off between code size, blocklength, and error probability. Suppose we wish to transmit an equiprobable message $W \in \{1, \ldots, M\}$ through a noisy channel with given blocklength n.[3] We encode each possible message using a codebook $c_1, \ldots, c_M \in \mathcal{X}^n$. What is the largest possible size M of the codebook that can be decoded with error probability (averaged over equiprobable codewords and channel randomness) at most ϵ? For memoryless channels and in various more general situations, the maximum code size $M^*(n, \epsilon)$ satisfies [42]

$$\ln M^*(n, \epsilon) = nC + \mathrm{Q}^{-1}(1-\epsilon)\sqrt{nV} + o_\epsilon(\sqrt{n}), \tag{1.1}$$

where the capacity C and dispersion V determine the precise first- and second-order asymptotics, and $\mathrm{Q}^{-1}(\cdot)$ is the inverse Gaussian tail probability function. For memoryless channels ($P_{Y^n|X^n} = P_{Y|X}^{\otimes n}$), channel capacity and dispersion are

[3] A *channel* is a sequence of random transformations $\{P_{Y^n|X^n}\}$ indexed by blocklength n.

given, respectively, by the quantities [47, 42, 20]

$$C = \max_{P_X} I(X;Y), \tag{1.2}$$

$$V = \mathrm{Var}\left[\imath_{X;Y}(X;Y)\right], \tag{1.3}$$

where $\imath_{X;Y}(a;b) = \log \frac{\mathrm{d}P_{Y|X=a}}{\mathrm{d}P_Y}(b)$, $I(X;Y) = \mathbb{E}[\imath_{X;Y}(X;Y)]$ is the mutual information, and (1.3) is evaluated for a P_X that attains the maximum in (1.2).

To prove a result such as (1.1), we must address two separate questions. The achievability part (that is, the inequality $\geq$) requires us to show existence of a codebook $c_1, \ldots, c_M$ that attains the prescribed error probability. This is usually accomplished using the *probabilistic method* due to Shannon [47] which analyzes the error probability not of a particular code, but rather its average when the codebooks are randomly drawn from an auxiliary distribution. For many problems in information theory with known first-order asymptotics, an achievability bound with $\sim \sqrt{n}$ second-order term can be derived using random coding in conjunction with other techniques [52, 59, 54].

In contrast, the converse part (that is, the inequality $\leq$) claims that *no* code can exceed the size given in (1.1) for the given error probability ϵ. The simplest and most widely used tool in converse analyses is Fano's inequality [15], which yields, in the memoryless case, the following estimate:

$$\ln M^*(n,\epsilon) \leq \frac{n}{1-\epsilon} C + \frac{h(\epsilon)}{1-\epsilon}, \quad \epsilon \in (0,1). \tag{1.4}$$

Such a bound is called a *weak converse*: it yields the correct first-order asymptotics in the limit of vanishing error probability, namely,

$$\lim_{\epsilon \downarrow 0} \lim_{n \to \infty} \frac{1}{n} \ln M^*(n,\epsilon) \leq C. \tag{1.5}$$

However, (1.4) does not rule out that transmission rates exceeding C may be feasible for any given nonzero error probability ϵ. To that end we need a more powerful result known as the *strong converse*, namely,

$$\ln M^*(n,\epsilon) \leq nC + o_\epsilon(n), \quad \epsilon \in (0,1), \tag{1.6}$$

which can be proved by a variety of methods (e.g., [57, 49, 56, 3, 53, 42, 44]). Our interest in this paper is on the even stronger notion of a *second-order converse*

$$\ln M^*(n,\epsilon) \leq nC + O_\epsilon(\sqrt{n}), \quad \epsilon \in (0,1), \tag{1.7}$$

which not only implies a strong converse (1.6) but yields the $\sqrt{n}$ behavior in (1.1). We caution that second-order converses do not always yield the sharpest possible constant in the $\sqrt{n}$-term; however, they aim to capture at least qualitatively various features of the sharp second-order asymptotics illustrated by (1.1).

Of course, the basic data transmission problem that we have discussed here for sake of illustration is particularly simple, as the exact second-order asymptotics (1.1) are known. This is not the case in more complicated information-theoretic problems. In particular, there are many problems in multiuser information theory for which either only weak converses are known, or, at most, strong converses have been obtained using the blowing-up method. It is precisely in such situations that new powerful and robust methods for obtaining second-order converses are needed. Next, we briefly discuss the three main approaches that have been developed in the literature for addressing such problems.

1. The goal of the *single-shot method* (e.g., [48, 53, 42, 22, 23, 44, 51]) is to obtain non-asymptotic achievability and converse bounds without imposing any probabilistic structure on either sources or channels. Therefore, only the essential features of the problem come into play. Those non-asymptotic bounds are expressed not in terms of average quantities such as entropy or mutual information, but in terms of *information spectra*, namely the distribution function of information densities such as $\imath_{X;Y}(X;Y)$. When coupled with the law of large numbers or the ergodic theorem and with central-limit theorem tools, the bounds become second-order tight. The non-asymptotic converse bounds for single-user data transmission boil down to the derivation of lower bounds on the error probability of Bayesian M-ary hypothesis testing followed by anonymization of the actual codebook. Often, those lower bounds are obtained by recourse to the analysis of an associated auxiliary binary hypothesis testing problem. This converse approach has been successfully applied to some problems of multiuser information theory such as Slepian-Wolf coding, multiple access channels [18], and broadcast channels [38]. Its application to other network setups is however a work in progress.

2. *Type class analysis* has been used extensively since [49] and was popularized by [11] mainly in the context of error exponents; however, it applies also to second-order analysis (see, e.g., [51]). The idea behind this method is that to obtain lower bounds, we may consider a situation where the decoder is artificially given access to the *type* (empirical distribution) of the source or channel sequences. Conditioned on each type, the distribution is equiprobable on the type class, so the evaluation of the conditional error probability is reduced to a combinatorial problem (this has been referred to as the "skeleton" or "combinatorial kernel" of the information-theoretic problem [1]). However, this combinatorial problem is not easily solved in side information problems (without additional ideas such as the blowing-up lemma). Moreover, by its nature, the method of types is restricted to finite alphabets and memoryless channels.

3. The method using the *blowing-up lemma (BUL)* of Ahlswede-Gács-Körner [3, 11, 45] uses a completely different idea to attain converse bounds: rather than try to reduce the given converse problem to a simpler one (e.g., to a binary hypothesis testing problem or a codebook that uses a single type), the BUL method is in essence a general technique for bootstrapping a strong converse from a weak converse. Even when the error probability ϵ is fixed, the concentration of measure phenomenon implies that all sequences except those in a set of vanishing probability differ in at most a fraction $o(1)$ of coordinates from a correctly decoded sequence. One can therefore effectively reduce the regime of fixed error probability to one of vanishing error probability, where a weak converse suffices, with negligible cost. The advantage of this method is that it is very broadly applicable. However, as will be discussed below, quantitative bounds obtained from this method are always suboptimal and second-order converses are fundamentally outside its reach. Moreover, the perturbation argument used in this approach is restricted to finite alphabets.

The single-shot and type class analysis methods yield second-order converses, but there are various problems in network information theory that have remained so far outside their reach. In contrast, the BUL method has been successful in establishing strong converses for a wide range of problems, including all settings in [11] with known single-letter rate region; see [11, Ch. 16]. For some problems in network information theory, such as source coding with compressed side information [11], BUL remained hitherto the only method for establishing a strong converse [51, Section 9.2]. However, the generality of the method comes at the cost of an inherent inefficiency, which prevents it from attaining second-order converses and prevents its application beyond the finite alphabet setting.

In this paper, we will show that one can have essentially the best of both worlds: the inefficiency of the blowing-up method can be almost entirely overcome by revisiting the foundation on which it is based. The resulting theory provides a canonical approach for proving second-order asymptotic converses and is applicable to a wide range of information-theoretic problems (including problems with general alphabets) for which no such results were known.

1.3 "Blowing up" vs "smoothing out"

In order to describe the core ingredients of our approach, let us begin by delving into the main elements of the blowing-up method of Ahlswede, Gács and Körner (a detailed treatment in a toy example will be given in Section 2).

The concentration of measure phenomenon is one of the most important ideas in modern probability [25, 6]. It states that for many high-dimensional probability measures, almost all points in the space are within a small distance of any set of fixed probability. This basic principle may be developed in different settings and has numerous important consequences; for example, it implies that Lipschitz functions on high-dimensional spaces are sharply concentrated around their median, a fact that will not be used in the sequel (but is crucial in many other contexts). The following modern incarnation of the concentration property used in the work of Ahlswede-Gács-Körner is due to Marton [35]; see also [45, Lemma 3.6.2].

Lemma 1.1 (Blowing-up lemma). *Denote the r-blowup of $\mathcal{A} \subseteq \mathcal{Y}^n$ by*

$$\mathcal{A}_r := \{v^n \in \mathcal{Y}^n : d_n(v^n, \mathcal{A}) \le r\}, \tag{1.8}$$

where d_n is the Hamming distance on $\mathcal{Y}^n$. Then

$$P^{\otimes n}[\mathcal{A}_r] \ge 1 - e^{-c^2} \quad \textit{for} \quad r = \sqrt{\frac{n}{2}}\left(\sqrt{\ln \frac{1}{P^{\otimes n}[\mathcal{A}]}} + c\right), \tag{1.9}$$

for any $c > 0$ and any probability measure P on $\mathcal{Y}$.

For example, if $n = 5\times 10^9$ and $P^{\otimes n}[\mathcal{A}] = e^{-100}$, we can achieve $P^{\otimes n}[\mathcal{A}_r] \ge 1-e^{-100}$ by letting $r = 10^6 = 0.0002n$. Asymptotically, if $P^{\otimes n}[\mathcal{A}]$ does not vanish, then $P^{\otimes n}[\mathcal{A}_r] \to 1$ as long as $r \gg \sqrt{n}$. In other words, the rather remarkable fact is that we can drastically increase the probability of a set by perturbing only a very small ($\approx n^{-1/2}$) fraction of coordinates of each of its elements.

Ahlswede, Gács and Körner realized how to leverage the BUL to prove strong converses. Suppose one is in a situation where a weak converse, such as (1.4), can be proved through Fano's inequality or any other approach. This can be done in all information-theoretic problems with known first-order asymptotics. However, a weak converse only yields the correct first-order constant when the error probability ϵ is allowed to vanish, while we are interested in the regime of constant ϵ. Let $\mathcal{A}$ be the set of correctly decoded sequences, whose probability is $1-\epsilon$. By the blowing-up lemma, a very slight blow-up $\mathcal{A}_r$ of this set will already have probability $1-o(1)$. We now apply the weak converse argument using $\mathcal{A}_r$ instead of $\mathcal{A}$. On the one hand, this provides the desired first-order term in (1.4), as $1-\epsilon$ is replaced by $1-o(1)$. On the other hand, we must pay a price in the argument for replacing the true decoding set $\mathcal{A}$ by its blowup $\mathcal{A}_r$. If $r = o(n)$, the latter turns out to contribute only to lower order and thus a strong converse is obtained.

The beauty of this approach is that it provides a very general recipe for upgrading a weak converse to a strong converse, and is therefore widely applicable. However, the method has (at least) two significant drawbacks:

- It is designed to yield a strong converse, not the stronger second-order asymptotic converse. Therefore, it is not surprising that it fails to yield second-order behavior that we expect from (1.1): when optimized, the BUL method appears unable to give a bound better than $O(\sqrt{n}\log^{\frac{3}{2}} n)$ (e.g., [45, Thm. 3.6.7]). This is already suggested by Lemma 1.1 itself: to obtain a $\sim \sqrt{n}$ second-order term from (1.4), we would need $\mathcal{A}_r$ to have probability at least $1 - O(n^{-1/2})$. That would require perturbing at least $r \sim \sqrt{n \log n}$ coordinates, which already gives rise to additional logarithmic factors. Thus, the blowing-up operation is too crude to recover the correct second-order behavior. In Appendix A, we will show that this is not an inefficiency in the blowing-up lemma itself, but is in fact an insurmountable problem of any method that is based on set enlargement.

- The argument relies essentially on the finite-alphabet setting. This is not because of the blowing-up lemma, which works for any alphabet $\mathcal{Y}$, but because we must control the price paid for replacing $\mathcal{A}$ by $\mathcal{A}_r$. While Lemma 1.1 gives a lower bound on $P^{\otimes n}[\mathcal{A}_r]$ as a function of $P^{\otimes n}[\mathcal{A}]$, we can also upper bound $P^{\otimes n}[\mathcal{A}_r]$ as a function of $P^{\otimes n}[\mathcal{A}]$ by the following simple argument, which relies crucially on the finiteness of the alphabet.

Lemma 1.2. *Suppose that $|\mathcal{Y}| < \infty$ and that $P(a) > 0$ for all $a \in \mathcal{Y}$. Then*

$$r \ln \frac{r}{neK} \leq \ln \frac{P^{\otimes n}[\mathcal{A}]}{P^{\otimes n}[\mathcal{A}_r]} \leq 0. \tag{1.10}$$

where $K = \frac{|\mathcal{Y}|}{\min_{a \in \mathcal{Y}} P(a)}$. Therefore, if $r = o(n)$, then

$$\lim_{n \to \infty} \frac{1}{n} \ln \frac{P^{\otimes n}[\mathcal{A}]}{P^{\otimes n}[\mathcal{A}_r]} = 0. \tag{1.11}$$

Proof. The right inequality in (1.10) follows from $\mathcal{A} \subset \mathcal{A}_r$. The set $\mathcal{A}_r$ is the overlapping union of spheres centered at the elements of $\mathcal{A}$, each of which contains fewer than $\binom{n}{r}|\mathcal{Y}|^r$ elements. This fact, along with the crude bound

$$\frac{P^{\otimes n}(y^n)}{P^{\otimes n}(z^n)} \geq \left(\min_{a \in \mathcal{Y}} P(a)\right)^{d_n(y^n, z^n)} \tag{1.12}$$

yields

$$P^{\otimes n}[\mathcal{A}_r] \leq \binom{n}{r} K^r P^{\otimes n}[\mathcal{A}]. \tag{1.13}$$

Then, the left inequality in (1.10) follows from $\binom{n}{r} \leq \left(\frac{en}{r}\right)^r$. □

The main contribution of this paper is to show that the shortcomings of the blowing-up method can be essentially eliminated while retaining its wide applicability. This is enabled by two key ideas that play a central role in our theory.

1. *Functional inequalities*. To prove a weak converse such as (1.4), one must relate the relevant information-theoretic quantity (e.g., mutual information) to the error probability (i.e., the probability of the decoding sets). This connection is generally made using a data processing inequality. However, in BUL-type methods, we no longer work directly with the original decoding sets, but rather with a perturbation of these that has better properties. Thus, there is also no reason to restrict attention to *sets*: we can replace the decoding set by an arbitrary *function*, and then control the relevant information functionals through their variational characterization (that is, by convex duality). Unlike the data processing inequality, such variational characterizations are in principle sharp and provide a lot more freedom in how to approximate the decoding set.

2. *"Smoothing out" vs. "blowing up"*. Once one makes the psychological step of working with functional inequalities, it becomes readily apparent that the idea of "blowing up" the decoding sets is much more aggressive than necessary to obtain a strong converse. What turns out to matter is not the overall size of the decoding set, but only the presence of very small values of the function used in the variational principle. Modifying the indicator function of a set to eliminate its small values can be accomplished by a much smaller perturbation than is needed to drastically increase its probability: this basic insight explains the fundamental inefficiency of the classical BUL approach.

   To implement this idea, we must identify an efficient method to improve the positivity of an indicator function. To this end, rather than "blowing up" the set by adding all points within Hamming distance r, we will "smooth out" the indicator function by averaging it locally over points at distance $\sim r$. More precisely, this averaging will be performed by means of a suitable Markov semigroup, which enables us to apply the *reverse hypercontractivity* phenomenon [5, 37] to establish strong positivity-improving properties. Such hypercontractivity phenomenon replaces, in our approach, the much better known concentration of measure phenomenon that was exploited in the BUL method. We will show, moreover, that Markov semigroup perturbations are much easier to control than their blowing-up counterparts, so that our method extends readily to general alphabets, Gaussian channels, and channels with memory.

When combined, the above ideas provide a powerful machinery for developing second-order converses. For example, in the basic data transmission problem that we discussed at the beginning of Section 1.2, our method yields

$$\ln M^*(n,\epsilon) \le nC + 2\sqrt{\ln\frac{1}{1-\epsilon}}\sqrt{n(\alpha-1)} + \ln\frac{1}{1-\epsilon} \tag{1.14}$$

(cf. Theorem 3.2), where α is a certain quantity closely related to the dispersion. When compared to the exact second-order asymptotics (1.1), we see that the second-order term has the correct scaling not only in the blocklength $n \to \infty$, but also in the error probability $\epsilon \to 1$ (as $\mathrm{Q}^{-1}(1-\epsilon) \sim \sqrt{2 \ln \frac{1}{1-\epsilon}}$ as $\epsilon \to 1$). However, our method does not recover the exact dispersion in (1.1); nor does it capture the fact that in the small error regime $\epsilon < \frac{1}{2}$, the second-order term in (1.1) is in fact *negative* (the second-order term in (1.14) is always positive). While the latter features are not directly achievable by the methods of this paper, they can be addressed by combining our methods with type class analysis [27]. The details of such a refined analysis are beyond the scope of this paper.

Let us remark that there are various connections between the notions of (reverse) hypercontractivity and concentration of measure; see, e.g., [4] and [24, p. 116] in the continuous case and [36, 37] in the discrete case. However, our present application of reverse hypercontractivity is different in spirit: we are not using it to achieve concentration but only a much weaker effect, which is the key to the efficiency of our method. The sharpness and broad applicability of our method suggests that this may be the "right" incarnation of the pioneering ideas of Ahlswede-Gács-Körner: one might argue that the blowing-up method succeeded in its aims, in essence, because it approximates the natural smoothing-out operation.

1.4 Organization

We have sketched the main ideas behind our approach in broad terms. To describe the method in detail, it is essential to get our hands dirty. To this end, we develop both the blowing-up and smoothing-out approaches in Section 2 in the simplest possible toy example: that of binary hypothesis testing. While this problem is amenable to a (completely classical) direct analysis, we view it as the ideal pedagogical setting in which to understand the main ideas behind our general theory.

The remainder of the paper is devoted to implementing these ideas in increasingly nontrivial situations.

In Section 3, we use our approach to strengthen Fano's inequality with an optimal $O(\sqrt{n})$ second-order term. We develop both the discrete and the Gaussian cases, and illustrate their utility in applications to broadcast channels and the output distribution of good channel codes.

In Section 4, we use our approach to strengthen a basic image size characterization bound of [3], obtaining a $O(\sqrt{n})$ second-order term. The theory is developed once again both in the discrete and the Gaussian cases. We illustrate the utility of our results in applications to hypothesis testing under communication constraints, and to source coding with compressed side information.

The paper concludes with two appendices. In Appendix A, we show that no method based on set enlargement can achieve second-order converses. Thus the functional viewpoint of this paper is essential. Finally, Appendix B contains proofs of some technical results used in section 4.

1.5 Notation

We end this section by collecting common notation that will be used throughout the paper. First, we record two conventions that will always be in force:

- All information-theoretic quantities (such as entropy, relative entropy, mutual information, etc.) will be defined in base e.
- In all variational formulas (such as (2.12), (4.8), (4.19), (4.22), etc.) it is implicit in the notation that we optimize only over functions or measures for which each term of the expression inside the supremum are finite.

We use standard information-theoretic notations for relative entropy $D(P\|Q)$, conditional relative entropy $D(P_{X|U}\|Q_{X|U}|P_U) = D(P_{X|U}P_U\|Q_{X|U}P_U)$, mutual information $I(X;Y)$, entropy $H(X)$, and differential entropy $h(X)$.

We denote by $\mathcal{H}_+(\mathcal{Y})$ the set of nonnegative Borel measurable functions on $\mathcal{Y}$, and by $\mathcal{H}_{[0,1]}(\mathcal{Y})$ the subset of $\mathcal{H}_+(\mathcal{Y})$ with range in $[0,1]$. For a measure ν and $f \in \mathcal{H}_+(\mathcal{Y})$, we write $\nu(f) := \int f\, d\nu$ and $\|f\|_p^p = \|f\|_{L^p(\nu)}^p = \int |f|^p\, d\nu$. We will frequently use $\|f\|_{L^0(\nu)} := \lim_{q\downarrow 0} \|f\|_{L^q(\nu)} = \exp(\nu(\ln f))$ for a probability measure ν. The measure of a set is denoted as $\nu[\mathcal{A}]$, and the restriction of a measure to a set is denoted $\mu|_{\mathcal{C}}[\mathcal{A}] := \mu[\mathcal{A}\cap\mathcal{C}]$. A random transformation $Q_{Y|X}$, mapping measures on $\mathcal{X}$ to measures on $\mathcal{Y}$, is viewed as an operator mapping $\mathcal{H}_+(\mathcal{Y})$ to $\mathcal{H}_+(\mathcal{X})$ according to $Q_{Y|X}(f) := \mathbb{E}[f(Y)|X = \cdot]$ where $(X,Y) \sim Q_{XY}$. The notation $U - X - Y$ denotes that the random variables U, X, Y form a Markov chain. The cardinality of $\mathcal{Y}$ is denoted $|\mathcal{Y}|$, and $\|x^n\|$ denotes the Euclidean norm of a vector $x^n \in \mathbb{R}^n$. Finally, y_i denotes the ith element of a sequence or vector, while y^i denotes the components up to the ith one $(y_j)_{j\le i}$.

2 Prelude: Binary Hypothesis Testing

2.1 Setup

The most elementary setting in which the ideas of this paper can be developed is the classical problem of binary hypothesis testing. It should be emphasized that our theory does not prove anything new in this setting: due to the Neyman-Pearson lemma, the exact form of the optimal tests is known and thus the analysis is amenable to explicit computation (we will revisit this point in Section 2.4). Nonetheless, the simplicity of this setting makes it the ideal toy example in which to introduce and discuss the main ideas of this paper.

In the binary hypothesis testing problem, we consider two competing hypotheses: data is drawn from a probability distribution on $\mathcal{Y}$, which we know is either P or Q. Our aim is to test, on the basis of a data sample, whether it was drawn from P or Q. More precisely, a (possibly randomized) *test* is defined by a function $f \in \mathcal{H}_{[0,1]}(\mathcal{Y})$: when a data sample $y \in \mathcal{Y}$ is observed, we decide hypothesis P with probability $f(y)$, and decide hypothesis Q otherwise. Thus, we must consider two error probabilities:

$$\begin{aligned} \pi_{P|Q} &:= Q(f) = \text{ the probability that } P \text{ is decided when } Q \text{ is true} \\ \pi_{Q|P} &:= 1 - P(f) = \text{ the probability that } Q \text{ is decided when } P \text{ is true} \end{aligned}$$

We aim to investigate the fundamental tradeoff between $\pi_{P|Q}$ and $\pi_{Q|P}$ in the case of product measures $P \leftarrow P^{\otimes n}$, $Q \leftarrow Q^{\otimes n}$; in other words, what is the smallest error probability $\pi_{P^{\otimes n}|Q^{\otimes n}}$ that may be achieved by a test that satisfies $\pi_{Q^{\otimes n}|P^{\otimes n}} \leq \epsilon \in (0,1)$? In this setting, the exact first-order and second-order asymptotics are due to Chernoff [9] and Strassen [50], respectively, resulting in

$$\ln \frac{1}{\pi_{P^{\otimes n}|Q^{\otimes n}}} = nD(P\|Q) + \mathrm{Q}^{-1}(1-\epsilon)\sqrt{nV(P\|Q)} + o_\epsilon(\sqrt{n}), \tag{2.1}$$

where $V(P\|Q) = \mathrm{Var}_P(\ln \frac{dP}{dQ})$.

The achievability ($\leq$) part of (2.1) is straightforward (see Section 2.4), so the main interest in the proof is to obtain the converse ($\geq$). In this section, we will illustrate both the blowing-up method of Ahlswede, Gács and Körner and the new approach of this paper in the context of this simple problem, and compare the resulting bounds to the exact second-order asymptotics (2.1).

2.2 The blowing-up method

For simplicity, to illustrate the blowing-up method in the context of the binary hypothesis testing problem, we restrict attention to *deterministic* tests $f = 1_{\mathcal{A}}$ for some $\mathcal{A} \subseteq \mathcal{Y}$ (that is, we decide hypothesis P if $y \in \mathcal{A}$ and hypothesis Q otherwise). This is not essential, but it simplifies the analysis.

As we mentioned, the blowing-up method is a general technique for upgrading weak converses to strong converses. In the present setting, a weak converse (for any (P, Q), not necessarily product measures) follows in a completely elementary manner from the data processing property of relative entropy.

Lemma 2.1 (Weak converse bound for binary hypothesis testing). *Let P, Q be probability measures on $\mathcal{Y}$ and $\mathcal{A} \subseteq \mathcal{Y}$ define the set of observations for which the deterministic test decides P. If $\pi_{Q|P} = P[\mathcal{A}^c] \leq \epsilon$, then $\pi_{P|Q} = Q[\mathcal{A}]$ satisfies*

$$\ln \frac{1}{\pi_{P|Q}} \leq \frac{D(P\|Q) + \ln 2}{1-\epsilon}. \tag{2.2}$$

Proof. By the data processing property of relative entropy, we have

$$D(P\|Q) \geq P[\mathcal{A}] \ln \frac{P[\mathcal{A}]}{Q[\mathcal{A}]} + P[\mathcal{A}^c] \ln \frac{P[\mathcal{A}^c]}{Q[\mathcal{A}^c]} \tag{2.3}$$

$$\geq (1-\epsilon) \ln \frac{1}{Q[\mathcal{A}]} - h(P[\mathcal{A}^c]), \tag{2.4}$$

where $h(\epsilon) := -\epsilon \ln \epsilon - (1-\epsilon) \ln(1-\epsilon) \leq \ln 2$ is the binary entropy function. □

Specializing to product measures $P \leftarrow P^{\otimes n}$, $Q \leftarrow Q^{\otimes n}$, Lemma 2.1 yields

$$\ln \frac{1}{\pi_{P^{\otimes n}|Q^{\otimes n}}} \leq \frac{nD(P\|Q) + \ln 2}{1-\epsilon} \tag{2.5}$$

for any test that satisfies $\pi_{Q^{\otimes n}|P^{\otimes n}} \leq \epsilon$. While this is sufficient to conclude the weak converse [if $\pi_{Q^{\otimes n}|P^{\otimes n}} \to 0$, then $\pi_{P^{\otimes n}|Q^{\otimes n}}$ cannot vanish faster than $\exp(-nD(P\|Q))$] it falls short of recovering the correct first-order asymptotics in the regime of fixed ϵ, as we saw in (2.1).

The remarkable idea of Ahlswede, Gács and and Körner is that the argument of Lemma 2.1 can be significantly improved by applying the data processing argument (2.4) not to the test set $\mathcal{A}$ satisfying

$$\pi_{Q^{\otimes n}|P^{\otimes n}} = P^{\otimes n}[\mathcal{A}^c] \leq \epsilon, \tag{2.6}$$

but to its blow-up $\mathcal{A}_r$. Then

$$\ln \frac{1}{Q^{\otimes n}[\mathcal{A}_r]} \leq \frac{nD(P\|Q) + \ln 2}{P^{\otimes n}[\mathcal{A}_r]}. \tag{2.7}$$

We must now control both the gain and the loss caused by the blowup. On the one hand, by the blowing-up Lemma 1.1, $P^{\otimes n}[\mathcal{A}_r] = 1 - o(1)$ as long as $r \gg \sqrt{n}$, which eliminates the $1-\epsilon$ factor in the weak converse (2.5). On the other hand, assuming finite alphabets we can invoke Lemma 1.2 with $r \ll n$ and $P \leftarrow Q$ (we may assume without loss of generality that Q is positive on $\mathcal{Y}$) to obtain the strong converse

$$\ln \frac{1}{\pi_{P^{\otimes n}|Q^{\otimes n}}} \leq nD(P\|Q) + o_\epsilon(n). \tag{2.8}$$

With a little more effort, we can optimize the argument over r and quantify the magnitude of the lower-order term.

Proposition 2.2. *Assume $|\mathcal{Y}| < \infty$. Any deterministic test between $P^{\otimes n}$ and $Q^{\otimes n}$ on $\mathcal{Y}^n$ such that $\pi_{Q^{\otimes n}|P^{\otimes n}} \leq \epsilon \in (0,1)$ satisfies*

$$\ln \frac{1}{\pi_{P^{\otimes n}|Q^{\otimes n}}} \leq nD(P\|Q) + O(\sqrt{n} \log^{\frac{3}{2}} n). \tag{2.9}$$

Proof. By the blowing-up Lemma 1.1, we have (since $3 > \sqrt{2} - \frac{1}{\sqrt{n}}$)

$$P^{\otimes n}[\mathcal{A}_r] \geq 1 - e^{-r^2/n} \quad \text{for all } r \geq 3\sqrt{n \ln \frac{1}{1-\epsilon}}. \tag{2.10}$$

Assembling (1.10) (with $P \leftarrow Q$), (2.7) and (2.10) we obtain

$$\ln \frac{1}{\pi_{P^{\otimes n}|Q^{\otimes n}}} \leq \frac{nD(P\|Q) + \ln 2}{1 - e^{-r^2/n}} + r \ln \frac{Ke\,n}{r}. \tag{2.11}$$

Choosing $r \asymp \sqrt{n \log n}$ results in (2.9). □

Re-examining the proof of Proposition 2.2, we can easily verify that no other choice for the growth of r with n may accelerate the decay of the slack term in (2.9). While the blowing-up method almost effortlessly turns a weak converse into a strong one, it evidently fails to result in a second-order converse. The rather crude bounds provided by Lemmas 1.1 and 1.2 may be expected to be the obvious culprits. It will shortly become evident, however, that the inefficiency of the method lies much deeper than expected: the major loss occurs already in the very first step (2.3) where we apply the data processing inequality. We will in fact show in Appendix A that any method based on the data processing inequality necessarily yields a slack term at least of order $\sim \sqrt{n \log n}$. To surmount this obstacle, we have no choice but to go back to the drawing board.

2.3 The smoothing-out method

The aim of this section is to introduce the key ingredients of the new method proposed in this paper. As will be illustrated throughout this paper, this method yields second-order converses while retaining the broad range of applicability of the blowing-up method. In the following, there will be no reason to restrict attention to deterministic tests $f = 1_{\mathcal{A}}$ as in the previous section, so we will consider arbitrary randomized tests $f \in \mathcal{H}_{[0,1]}$ from now on.

2.3.1 Functional inequalities

To prove a converse, we must relate the relevant information-theoretic quantity $D(P\|Q)$ to the properties of any given test f. This was accomplished above by means of the data processing inequality (2.3). However, as was indicated at the end of the previous section, this already precludes us from obtaining sharp quantitative bounds. The first idea behind our approach is to replace the data processing argument by a different lower bound: we will use throughout this paper *functional inequalities* associated to information-theoretic quantities by convex duality. In the present setting, the relevant inequality follows from the Donsker-Varadhan variational principle for relative entropy [12] (see, e.g., [45, (3.4.67)])

$$D(P\|Q) = \sup_{g \in \mathcal{H}_+} \{P(\ln g) - \ln Q(g)\}. \tag{2.12}$$

Unlike the data processing inequality, which can only attain equality in trivial situations, the variational principle (2.12) always attains its supremum by choosing $g \leftarrow \frac{dP}{dQ}$. Therefore, unlike the data processing inequality, in principle an application of (2.12) need not entail any loss.

What we must now show is how to choose the function g in (2.12) to capture the properties of a given test f. Tempting as it is, the choice $g \leftarrow f$ is dismal: for example, in the case of deterministic tests $f = 1_{\mathcal{A}}$, generally $P(\ln 1_{\mathcal{A}}) = -\infty$ and we do not even obtain a weak converse. Instead, inspired by the blowing-up method, we may apply (2.12) to a suitably chosen perturbation of f. Let us first develop the argument abstractly so that we may gain insight into the requisite properties. Suppose we can design a mapping $T : \mathcal{H}_{[0,1]} \to \mathcal{H}_{[0,1]}$ (which plays the role of the blowing-up operation in the present setting) that satisfies:

1. For any test $f \in \mathcal{H}_{[0,1]}$ on $\mathcal{Y}^n$ with $P^{\otimes n}(f) \geq 1 - \epsilon$, we have

$$P^{\otimes n}(\ln Tf) \geq -o_\epsilon(n). \tag{2.13}$$

2. For any test $f \in \mathcal{H}_{[0,1]}$ on $\mathcal{Y}^n$, we have

$$\ln Q^{\otimes n}(Tf) \leq \ln Q^{\otimes n}(f) + o(n). \tag{2.14}$$

Setting $g \leftarrow Tf$ in (2.12) and using (2.13) and (2.14), we immediately deduce a strong converse: for any test $f \in \mathcal{H}_{[0,1]}$ such that $\pi_{Q^{\otimes n}|P^{\otimes n}} := P^{\otimes n}(1 - f) \leq \epsilon$, the error probability $\pi_{P^{\otimes n}|Q^{\otimes n}} := Q^{\otimes n}(f)$ satisfies (2.8). Besides replacing the data processing inequality by the variational principle, the above logic parallels the blowing-up method: (2.13) plays the role of the blowing-up Lemma 1.1, while (2.14) plays the role of the counting estimate (1.11).

Nonetheless, this apparently minor change of perspective lies at the heart of our theory. To explain why it provides a crucial improvement, let us pinpoint the origin of the inefficiency of the blowing-up method. The purpose of the blowing-up operation is to increase the probability of a test: given $P^{\otimes n}(f) \geq 1 - \epsilon$, one designs a blow-up $f \mapsto \tilde{f}$ so that $P^{\otimes n}(\tilde{f}) = 1 - o(1)$. However, when we use the the sharp functional inequality (2.12) rather than the data processing inequality, we do not need to control $P^{\otimes n}(\tilde{f})$, but rather $P^{\otimes n}(\ln \tilde{f})$. The latter is dominated by the *small values* of f, not by its overall magnitude. Therefore, an efficient perturbation of f should not seek to blow it up but only to boost its small values, which may be accomplished at a much smaller cost than the blowing-up operation. It is precisely this insight that will allow us to eliminate the inefficiency of the blowing-up method and attain sharp second-order bounds.

In order to take full advantage of this insight, we must understand how to design efficient perturbations $f \mapsto Tf$. The second key ingredient of our method is its main workhorse: a general mechanism to implement (2.13) and (2.14) so that their speed of decay will be such that the slack term in (2.8) is in fact $O\left(n^{-1/2}\right)$.

2.3.2 Simple semigroups

The essential intuition that arises from the above discussion is that in order to obtain efficient bounds in (2.13) and (2.14), we must design an operation $f \mapsto Tf$ that is *positivity-improving*: it boosts the small values of f sufficiently to ensure that $P^{\otimes n}(\ln Tf)$ is not too small. In this subsection we design a suitable transformation T, and in Section 2.3.3 we show that it achieves the desired goal.

Let $\mathcal{Y}$ be an arbitrary alphabet and let P be any probability measure thereon. We say $(T_t)_{t\geq 0}$ is a *simple semigroup*[4] with stationary measure P if

$$T_t\colon \mathcal{H}_+(\mathcal{Y}) \to \mathcal{H}_+(\mathcal{Y}), \qquad f \mapsto e^{-t}f + (1-e^{-t})P(f). \tag{2.15}$$

In the i.i.d. case $\mathcal{Y} \leftarrow \mathcal{Y}^n$, $P \leftarrow P^{\otimes n}$ we consider their tensor product

$$T_t := [e^{-t} + (1-e^{-t})P]^{\otimes n}. \tag{2.16}$$

We will use $T = T_t$, for a suitable choice of t, as a positivity-improving operation.

It is instructive to examine the effect of the operator in (2.16) on indicator functions. For that purpose, we introduce the following ad-hoc notation: if $(v^n, w^n) \in \mathcal{Y}^n \times \mathcal{Y}^n$ and $I \subset \{1, \ldots, n\}$, then $v^n I w^n \in \mathcal{Y}^n$ is defined by

$$(v^n I w^n)_i = \begin{cases} v_i, & i \in I \\ w_i, & i \in I^c. \end{cases} \tag{2.17}$$

Then, using

$$(aQ + (1-a)P)^{\otimes n} = \sum_{I \subset \{1,\ldots,n\}} a^{n-|I|}(1-a)^{|I|} P^{\otimes I} Q^{\otimes I^c}, \tag{2.18}$$

the application of the operator in (2.16) to the indicator function becomes

$$T_t 1_{\mathcal{A}}(y^n) = \mathbb{E}[1_{\mathcal{A}}(Z^n \mathbf{I} y^n)], \tag{2.19}$$

[4] Readers who are unfamiliar with semigroups may ignore this terminology; while the semigroup property plays an important role in the proof of Theorem 2.3, it is not used directly in this paper.

where I is a random subset of $\{1,\dots,n\}$ obtained by including each element independently with probability $1-e^{-t}$ (in particular, $|\mathbf{I}|\sim \mathrm{Binom}(n,1-e^{-t})$), and $Z^n\sim P^{\otimes n}$ is independent of $\mathbf{I}$. In contrast, the blowing-up operation may be expressed in terms of indicator functions as

$$1_{\mathcal{A}_r}(y^n)=\max_{|I|\le r}\max_{z^n\in\mathcal{Y}^n} 1_{\mathcal{A}}(z^n I y^n). \tag{2.20}$$

From this perspective, we see that the semigroup operation is a *smoothing out* counterpart of the blowing-up operation: while the blowing-up operation *maximizes* the function over a local neighborhood of size r, the semigroup operation *averages* the function over a random neighborhood of size $r\approx n(1-e^{-t})$. What we will gain from smoothing is that it increases the small values of f (it is positivity-improving) without increasing the total mass $P^{\otimes n}(T_tf)=P^{\otimes n}(f)$, so that the mass under $Q^{\otimes n}$ cannot grow too much. In contrast, blowing-up is designed to increase the mass $P^{\otimes n}(f)$; but then the mass under $Q^{\otimes n}$ becomes large as well, which yields the suboptimal rate achieved by the blowing-up method.

2.3.3 Reverse hypercontractivity

It is intuitively clear that T_t is positivity improving: it maps any nonnegative function to a strictly positive function. But the goal of lower bounding $P^{\otimes n}(\ln Tf)$ is more ambitious. This idea already appears in the probability theory literature in a very different context: it was realized long ago by Borell [5] that Markov semigroups possess very strong positivity-improving properties, which are described quantitatively by a reverse form of the classical hypercontractivity phenomenon. While Borell was motivated by applications in quantum field theory, we will show in this paper that reverse hypercontractivity provides a powerful mechanism that appears almost tailor-made for our present purposes.

We will presently describe an important generalization of Borell's ideas to general alphabets due to Mossel et al. [37], and show how it may be combined with the above ideas to obtain sharp non-asymptotic converses.

Theorem 2.3 (**Reverse hypercontractivity).** *[37]. Let $(T_t)_{t\ge0}$ be a simple semigroup (2.15) or an arbitrary tensor product of simple semigroups. Then*

$$\|T_tf\|_{L^q}\ge\|f\|_{L^p} \tag{2.21}$$

for any $0<q<p<1$, $f\in\mathcal{H}_+$, and $t\ge\ln\frac{1-q}{1-p}$. In particular, letting $q\downarrow 0$, we have

$$P(\ln T_tf)\ge\log\|f\|_{L^p(P)}. \tag{2.22}$$

An estimate of the form (2.13) is almost immediate from (2.22). On the other hand, the estimate (2.14) will now follow from a simple change of measure argument. The following result combines these ingredients to derive a second-order converse for binary hypothesis testing.

Theorem 2.4. *Any test between $P^{\otimes n}$ and $Q^{\otimes n}$ such that $\pi_{Q^{\otimes n}|P^{\otimes n}} \leq \epsilon$ satisfies*

$$\ln \frac{1}{\pi_{P^{\otimes n}|Q^{\otimes n}}} \leq nD(P\|Q) + 2\sqrt{\ln \frac{1}{1-\epsilon}}\sqrt{n(\alpha-1)} + \ln \frac{1}{1-\epsilon}, \tag{2.23}$$

where $\alpha = \|\frac{dP}{dQ}\|_\infty \geq 1$.

Proof. We establish (2.13) and (2.14) by choosing $T = T_t$ as defined by (2.16). We fix any $t > 0$ initialy and optimize at the end of the proof.

Fix any test $f \in \mathcal{H}_{[0,1]}(\mathcal{Y}^n)$. To establish (2.13), note that by (2.22)

$$\begin{aligned} P^{\otimes n}(\ln T_t f) &\geq \ln \|f\|_{L^{1-e^{-t}}(P^{\otimes n})} && (2.24)\\ &\geq \frac{1}{1-e^{-t}} \ln P^{\otimes n}(f) && (2.25)\\ &\geq \left(\frac{1}{t}+1\right) \ln P^{\otimes n}(f), && (2.26) \end{aligned}$$

where (2.25) used that $f \in [0,1]$ and (2.26) follows from $e^t \geq 1+t$.

On the other hand, to establish (2.14), we argue as follows:

$$\begin{aligned} Q^{\otimes n}(T_t f) &= Q^{\otimes n}((e^{-t} + (1-e^{-t})P)^{\otimes n} f) && (2.27)\\ &= (e^{-t}Q + (1-e^{-t})P)^{\otimes n} f && (2.28)\\ &\leq (e^{-t} + \alpha(1-e^{-t}))^n Q^{\otimes n}(f) && (2.29)\\ &\leq e^{(\alpha-1)nt} Q^{\otimes n}(f), && (2.30) \end{aligned}$$

where $\alpha = \|\frac{dP}{dQ}\|_\infty \geq 1$, and (2.27) is just the definition of T_t.

Now assume that the test f satisfies $\pi_{Q^{\otimes n}|P^{\otimes n}} := P^{\otimes n}(1-f) \leq \epsilon$. Setting $P \leftarrow P^{\otimes n}$, $Q \leftarrow Q^{\otimes n}$, and $g \leftarrow T_t f$ in (2.12), we obtain for all $t > 0$

$$\ln \frac{1}{\pi_{P^{\otimes n}|Q^{\otimes n}}} \leq nD(P\|Q) + \left(\frac{1}{t}+1\right) \ln \frac{1}{1-\epsilon} + (\alpha-1)nt \tag{2.31}$$

using (2.26) and (2.30). Minimizing (2.31) with respect to t yields (2.23). □

Beside resulting in a second-order converse, the smoothing-out method has an additional major advantage over the blowing-up method: the change-of-measure argument (2.30) is purely measure-theoretic in nature and sidesteps the counting

	Blowing-up method	Smoothing-out method
Connecting information measures and observables	Data processing property (2.3)	Convex duality (2.12)
Lower bound w.r.t. a given measure	Concentration of measure (Lemma 1.1)	Reverse hypercontractivity (Theorem 2.3)
Upper bound w.r.t. reference measure	Counting argument (Lemma 1.2)	Change of measure (2.29) [or scaling argument (3.21)]

Table 1: Main ingredients of the blowing-up and smoothing-out approaches.

argument of Lemma 1.2. No analogue of the latter can hold beyond the finite alphabet case: indeed, in general alphabets even the blowup of a set of measure zero will have positive measure, ruling out any estimate of the form (1.11). In contrast, the result of Theorem 2.4 holds for any alphabet $\mathcal{Y}$ and requires only a bounded density assumption. Even the latter assumption is not an essential restriction and can be eliminated in specific situations by working with semigroups other than (2.16). A particularly important example that will be developed in detail later on in this paper is the case of Gaussian measures (see, for example, Section 3.2).

The proof of Theorem 2.4 illustrates the approach of this paper in its simplest possible setting. However, the basic ideas that we have introduced here form the basis of a general recipe that will be applied repeatedly in the following sections to obtain second-order converses in a broad range of applications. The comparison between the key ingredients of the blowing-up method and the smoothing-out approach proposed in this paper is summarized in Table 1.

2.3.4 Beyond product measures

Throughout this paper we focus for simplicity on stationary memoryless systems, that is, those defined by product measures. However, our approach is by no means limited to this setting. For the benefit of the interested reader, let us briefly sketch in the context of Theorem 2.4 what modifications would be needed to adapt our approach to general dependent measures. For an entirely different application of our approach in a dependent setting, see [27].

Consider the problem of testing between two arbitrary (non-product) hypotheses P_n, Q_n on $\mathcal{Y}^n$. To adapt the proof of Theorem 2.4, we need to introduce a hypercontractive semigroup with stationary measure P_n. A natural candidate in this general setting is the so-called *Gibbs sampler* $T_t f(y^n) = \mathbb{E}_{y^n}[f(Y_t^n)]$, where the Markov process Y_t^n is defined by replacing each coordinate with an independent draw from its conditional distribution $P_{Y_i|Y_{\setminus i}}$ (where $Y_{\setminus i} := (Y_j)_{j\neq i}$) at independent exponentially distributed intervals. It was shown in [37] that any semigroup that satisfies a modified log-Sobolev inequality is reverse hypercontractive. In particular, such inequalities may be established for the Gibbs sampler under rather general weak dependence assumptions [8, 34].

On the other hand, we need to establish an upper bound on $Q_n(T_t f)$. This may be done as follows. The Gibbs sampler satisfies the differential equation [8]

$$\frac{d}{dt}T_t f(y^n) = \sum_{i=1}^{n}\{P_{Y_i|Y_{\setminus i}=y_{\setminus i}}(T_t f) - T_t f(y^n)\}. \tag{2.32}$$

We may therefore estimate for any $f \in \mathcal{H}_+(\mathcal{Y}^n)$

$$\frac{d}{dt}Q_n(T_t f) = \sum_{i=1}^{n}\{Q_n(P_{Y_i|Y_{\setminus i}}(T_t f)) - Q_n(T_t f)\} \leq (\alpha-1)n\, Q_n(T_t f), \tag{2.33}$$

where we used the tower property of conditional expectations and we defined

$$\alpha := \max_i \left\| \frac{P_{Y_i|Y_{\setminus i}}}{Q_{Y_i|Y_{\setminus i}}} \right\|_\infty. \tag{2.34}$$

Solving the differential inequality yields precisely the same estimate $Q_n(T_t f) \leq e^{(\alpha-1)nt}Q_n(f)$ as was obtained in (2.30) in the product case.

Putting together these estimates, we obtain for any test with $\pi_{Q_n|P_n} \leq \epsilon$ that

$$\ln\frac{1}{\pi_{P_n|Q_n}} \leq nD(P_n\|Q_n) + \sqrt{C\ln\frac{1}{1-\epsilon}}\sqrt{n(\alpha-1)} + \ln\frac{1}{1-\epsilon}, \tag{2.35}$$

where C is the modified log-Sobolev constant of P_n. This extends our approach for the memoryless case to any dependent situation where a modified log-Sobolev inequality is available for P_n. A deeper investigation of dependent processes is beyond the scope of this paper.

2.4 Achievability and optimality

Theorem 2.4 gives a non-asymptotic converse bound for binary hypothesis testing. To understand whether it is accurate, we also need an upper bound (achievability). Such a bound was already stated in (2.1). To motivate the following discussion, it is instructive to give a quick proof of the achievability part of (2.1).

Lemma 2.5. *There exist a sequence of binary hypothesis tests with $\pi_{Q^{\otimes n}|P^{\otimes n}} \leq \epsilon$ such that*

$$\ln\frac{1}{\pi_{P^{\otimes n}|Q^{\otimes n}}} \geq nD(P\|Q) + \mathrm{Q}^{-1}(1-\epsilon)\sqrt{nV(P\|Q)} + o_\epsilon(\sqrt{n}), \tag{2.36}$$

provided that $V(P\|Q) := \mathrm{Var}_P(\ln\frac{dP}{dQ}) < \infty$.

Proof. For typographical convenience we will write $D := D(P\|Q)$ and $V := V(P\|Q)$. By the central limit theorem

$$\lim_{n\to\infty} P^{\otimes n}\left[\imath_{P^{\otimes n}\|Q^{\otimes n}} \ge nD + \mathrm{Q}^{-1}(1-\epsilon)\sqrt{nV}\right] = 1-\epsilon, \tag{2.37}$$

where we have defined the *relative information*

$$\imath_{P^{\otimes n}\|Q^{\otimes n}}(y^n) := \ln \frac{dP^{\otimes n}}{dQ^{\otimes n}}(y^n) = \sum_{i=1}^{n} \ln \frac{dP}{dQ}(y_i) \tag{2.38}$$

whose mean and variance with $y^n \sim P^{\otimes n}$ are nD and nV, respectively. We may therefore choose a deterministic sequence $a_n = o_\epsilon(\sqrt{n})$ such that the deterministic test $f_n = 1_{\mathcal{A}_n}$ defined by

$$\mathcal{A} = \{y^n \in \mathcal{Y}^n : \imath_{P^{\otimes n}\|Q^{\otimes n}}(y^n) \ge nD + \mathrm{Q}^{-1}(1-\epsilon)\sqrt{nV} - a_n\} \tag{2.39}$$

satisfies $\pi_{Q^{\otimes n}|P^{\otimes n}} \le \epsilon$ for all n. But a simple Chernoff bound now yields

$$\pi_{P^{\otimes n}|Q^{\otimes n}} \le e^{-nD-\mathrm{Q}^{-1}(1-\epsilon)\sqrt{nV}+a_n} Q^{\otimes n}(e^{\imath_{P^{\otimes n}\|Q^{\otimes n}}}), \tag{2.40}$$

and the proof is completed by noting that $Q^{\otimes n}(e^{\imath_{P^{\otimes n}\|Q^{\otimes n}}}) = 1$. $\square$

Comparing the achievability bound (2.36) with our converse (2.23), we see that the main features of the second-order term are captured faithfully by the smoothing-out method, although it fails to recover the precise constant in the second-order term: $V(P\|Q)$ is replaced by its natural uniform bound $V(P\|Q) \le \left\|\frac{dP}{dQ}\right\|_\infty - 1$ in our converse.[5] Beside the optimal order scaling $\sim \sqrt{n}$, we recall that the bound behaves correctly as a function of ϵ (up to universal constant) for large error probabilities $\epsilon \to 1$; cf. the disucussion following (1.14).

As is illustrated by Lemma 2.5, the achievability analysis is conceptually simple in the binary hypothesis testing case thanks to the Neyman-Pearson lemma which identifies the optimal test. However, in information theory optimal procedures are very seldom known explicitly. Thus, the methodology we have introduced says nothing new about binary hypothesis testing. The point of the present method, however, is that it applies broadly in situations where such a direct analysis is far out of reach. In particular, in the general setting of Section 4 the present approach is currently the only known method to achieve sharp second-order converses in a variety of multiuser information theory problems.

[5]To see this, use $x\ln^2 x \le (x-1)^2$ to show $\mathrm{Var}_P(\ln\frac{dP}{dQ}) \le Q(\frac{dP}{dQ}\ln^2\frac{dP}{dQ}) \le P(\frac{dP}{dQ}-1)$.

3 Second-Order Fano's Inequality

The aim of this section is to develop the smoothing-out methodology for channel coding problems, of which a basic example was discussed in Section 1.2.

Weak converses for channel coding problems can be obtained in great generality (cf. [14]): this is the domain of Fano's inequality [15], one of the most basic results in information theory, which gives an implicit upper bound on the error probability of an M-ary hypothesis testing problem. For discrete memoryless channels, when combined with a list decoding argument, the blowing-up method strengthens Fano's inequality to a strong converse with $\sim \sqrt{n}\log^{\frac{3}{2}} n$ second-order term [3, 35, 43, 45]. In this section, we will show that the smoothing-out method results in a strong form of Fano's inequality that not only attains the optimal $\sim \sqrt{n}$ second-order term, but is applicable to a much broader class of channels. The power of this machinery will be illustrated in two typical applications.

Before we turn to the main results of this section, it is instructive to give a short proof of a basic form of Fano's inequality. Although we state it and prove it for deterministic decoding, it also holds for stochastic decoders.

Lemma 3.1 (Fano's inequality). *Let $W \in \{1, \ldots, M\}$ be an equiprobable message to be transmitted over a noisy channel $P_{Y|X}$. Let $c_1, \ldots, c_M \in \mathcal{X}$ be the codewords corresponding to W, and let $\mathcal{D}_1, \ldots, \mathcal{D}_M \subseteq \mathcal{Y}$ be the disjoint decoding sets. Suppose the average probability of correct decoding satisfies*

$$\frac{1}{M}\sum_{m=1}^{M} P_{Y|X=c_m}[\mathcal{D}_m] \geq 1-\epsilon. \tag{3.1}$$

Then

$$\ln M \leq \frac{I(W;Y)+\ln 2}{1-\epsilon}, \tag{3.2}$$

where Y is the output of the channel $P_{Y|X}$ with input $X = c_W$.

Proof. Let $\hat{W}$ be the decoded message, that is, $\hat{W} = c_m$ when the output $Y \in \mathcal{D}_m$. The bound can be shown by reduction to an auxiliary binary hypothesis testing problem: $P \leftarrow P_{W\hat{W}}$, $Q \leftarrow P_W \otimes P_{\hat{W}}$. Then the conclusion follows by applying Lemma 2.1 with $\mathcal{A} = \{(x, \hat{x}) : x = \hat{x}\}$ since $I(W;\hat{W}) \leq I(W;Y)$. □

The proof of Fano's inequality highlights the connection between the channel coding problems investigated in this section and the simple hypothesis testing problem of Section 2. In particular, it suggests that the weak converse (3.2) may be strengthened to a strong converse of the form $\ln M \leq I(W;Y^n) + O(\sqrt{n})$ in the setting of memoryless channels $P_{Y|X} \leftarrow P_{Y^n|X^n} := P_{Y|X}^{\otimes n}$. Unfortunately, the latter does not follow from the strong converse obtained in Section 2 for binary hypothesis testing. Indeed, the framework of Section 2 does not apply as the measures P, Q that appear in the proof of Lemma 3.1 are not product measures even when the channel is memoryless. To sidestep this hurdle we will apply the smoothing-out operation conditionally on X^n: that is, we will introduce semigroups for which the channel $P_{Y^n|X^n=x^n}$ is the stationary measure. The main additional challenge that arises is that the semigroup depends on the channel input x^n and must therefore be controlled uniformly in x^n.

3.1 Bounded probability density case

In this section, we consider a random transformation $P_{X|Y}$ from $\mathcal{X}$ to $\mathcal{Y}$ and denote by $P_{Y^n|X^n=x^n} := P_{Y|X=x_1} \otimes \cdots \otimes P_{Y|X=x_n}$ the corresponding n-fold memoryless random transformation. The main assumption of this section is that there exists a reference probability measure ν on $\mathcal{Y}$ such that

$$\alpha := \sup_{x \in \mathcal{X}} \left\| \frac{dP_{Y|X=x}}{d\nu} \right\|_\infty \in [1, \infty). \tag{3.3}$$

This assumption is automatically satisfied in the finite alphabet setting $|\mathcal{Y}| < \infty$, in which case we may choose ν to be equiprobable and consequently $\alpha \leq |\mathcal{Y}|$. However, the present setting is much more general: it applies to an arbitrary output alphabet $\mathcal{Y}$ and requires only the existence of bounded densities.

The main result of this section is the following strong form of Fano's inequality (cf. Lemma 3.1) exhibiting a $\sqrt{n}$ second-order term..

Theorem 3.2. *Assume (3.3) holds. Let $W \in \{1, \ldots, M\}$ be an equiprobable message, let $c_1, \ldots, c_M \in \mathcal{X}^n$ be the codewords corresponding to W, and let $\mathcal{D}_1, \ldots, \mathcal{D}_M \subseteq \mathcal{Y}^n$ be disjoint decoding sets. Suppose that*

$$\prod_{m=1}^{M} P^{\frac{1}{M}}_{Y^n|X^n=c_m}[\mathcal{D}_m] \geq 1 - \epsilon. \tag{3.4}$$

Then

$$\ln M \leq I(W; Y^n) + 2\sqrt{\ln \frac{1}{1-\epsilon}} \sqrt{n(\alpha - 1)} + \ln \frac{1}{1-\epsilon}, \tag{3.5}$$

where Y^n is the output of the memoryless channel $P_{Y^n|X^n} = P^{\otimes n}_{Y|X}$ with input $X^n = c_W$.

Remark 3.3. The geometric average criterion (3.4) is stronger than the average error criterion (3.1) in Fano's inequality, but is weaker than the maximal error criterion $\min_m P_{Y^n|X^n=c_m}[\mathcal{D}_m] \geq 1-\epsilon$. Both the average and maximal error criteria are commonly used in information theory, and our assumption (3.4) is intermediate between these two conventional criteria.

It is natural to ask whether the strong Fano inequality (3.5) remains valid even under the average error criterion, like the classical Fano inequality. This is not the case. In [31], a general notion of "α-decodability" is introduced which subsumes the geometric average criterion, the average error criterion, and the maximum error criterion as the special cases $\alpha = 0, 1, -\infty$. It is shown there that $\alpha = 0$ is the critical value for the existence of a strong Fano inequality. In this sense, the assumption (3.4) of Theorem 3.2 is essentially the best possible.

While the strong Fano inequality itself cannot hold under the average error criterion, it is possible in some applications of this inequality to upgrade the subsequent results to hold under the average error criterion by an additional argument known as *codebook expurgation*. The idea behind this method is that if the average error criterion is met, then the maximal error criterion will be satisfied on a large subset of the codebook obtained by throwing out the worst half of the codewords (e.g., [10, Theorem 7.7.1]). By combining this device with Theorem 3.2 we can, for example, recover a strong converse under the average error criterion for the basic channel coding problem of Section 1.2. The expurgation argument cannot be applied directly to Theorem 3.2, however, as $I(W;Y^n)$ may change significantly when we throw out codewords. Whether or not an expurgation argument is feasible depends on the manner in which Theorem 3.2 is used in a given application.

Before we turn to the proof of Theorem 3.2, let us prepare for the smoothing-out argument that will appear therein. For any $x \in \mathcal{X}$, denote by $(T_{x,t})_{t\geq 0}$ the simple Markov semigroup on $\mathcal{Y}$ with stationary measure $P_{Y|X=x}$:

$$T_{x,t}f := e^{-t}f + (1-e^{-t})P_{Y|X=x}(f). \tag{3.6}$$

We denote by

$$T_{x^n,t} := T_{x_1,t} \otimes \cdots \otimes T_{x_n,t} \tag{3.7}$$

the corresponding product semigroup with stationary measure $P_{Y^n|X^n=x^n}$.

Unlike the simpler setting of Section 2, the semigroup $T_{x^n,t}$ depends on x^n. To work around this issue, we introduce a linear operator $\Lambda_t : \mathcal{H}_+(\mathcal{Y}^n) \to \mathcal{H}_+(\mathcal{Y}^n)$ that dominates $T_{x^n,t}$ uniformly over x^n:

$$\Lambda_t := [e^{-t} + \alpha(1-e^{-t})\nu]^{\otimes n}. \tag{3.8}$$

Note that Λ_t is not a Markov semigroup; in particular, its total mass satisfies

$$\Lambda_t 1 = \big(e^{-t} + \alpha(1-e^{-t})\big)^n \leq e^{(\alpha-1)nt}. \tag{3.9}$$

However, Λ_t dominates the semigroups $T_{x^n,t}$ in the following sense.

Lemma 3.4. *Assume* (3.3) *holds. Then for any* $f \in \mathcal{H}_+(\mathcal{Y}^n)$

$$\sup_{x^n \in \mathcal{X}^n} T_{x^n,t} f \leq \Lambda_t f. \tag{3.10}$$

Proof. The result follows immediately from the definition of $T_{x^n,t}$, Λ_t, and α upon noting that $P_{Y|X=x}(g) \leq \alpha\nu(g)$ whenever $g \geq 0$. $\square$

We are now ready for the proof of Theorem 3.2.

Proof of Theorem 3.2. Let $f_m := 1_{\mathcal{D}_m}$ and $t > 0$ to be optimized later. Note that

$$I(W;Y^n) = \frac{1}{M}\sum_{m=1}^{M} D(P_{Y^n|X^n=c_m} \| P_{Y^n}) \tag{3.11}$$

$$\geq \frac{1}{M}\sum_{m=1}^{M} P_{Y^n|X^n=c_m}(\ln \Lambda_t f_m) - \frac{1}{M}\sum_{m=1}^{M} \ln P_{Y^n}(\Lambda_t f_m) \tag{3.12}$$

where (3.12) is from the variational formula (2.12).

We can lower bound the first term in (3.12) as follows:

$$\frac{1}{M}\sum_{m=1}^{M} \ln \|\Lambda_t f_m\|_{L^0(P_{Y^n|X^n=c_m})}$$

$$\geq \frac{1}{M}\sum_{m=1}^{M} \ln \|T_{c_m,t} f_m\|_{L^0(P_{Y^n|X^n=c_m})} \tag{3.13}$$

$$\geq \frac{1}{1-e^{-t}} \frac{1}{M}\sum_{m=1}^{M} \ln P_{Y^n|X^n=c_m}(f_m) \tag{3.14}$$

$$\geq -\left(\frac{1}{t}+1\right) \ln \frac{1}{1-\epsilon} \tag{3.15}$$

where (3.13) is from (3.10); (3.14) is from Theorem 2.3 and $f \in [0,1]$; and (3.15) is from (3.4) and $e^t \geq 1+t$. For the second term in (3.12), we can estimate

$$-\frac{1}{M}\sum_{m=1}^{M} \ln P_{Y^n}(\Lambda_t f_m) \geq -\ln P_{Y^n}\left(\frac{1}{M}\sum_{m=1}^{M} \Lambda_t f_m\right) \tag{3.16}$$

$$\geq \ln M - (\alpha - 1)nt \tag{3.17}$$

using Jensen's inequality, $\sum_m f_m \leq 1$, and (3.9). Combining (3.12), (3.15), and (3.17) and optimizing over $t > 0$ concludes the proof. $\square$

3.2 Gaussian case

The strong Fano inequality of the previous section applies in principle to arbitrary alphabets $\mathcal{X}, \mathcal{Y}$. However, in certain situations the bounded density assumption (3.3) may be overly restrictive, particularly when the alphabets are unbounded. In this section, we will adapt the argument of the previous section to the prototypical example where this issue arises, namely, Gaussian channels. Beside the intrinsic utility of the result, this will illustrate the broader applicability of the smoothing-out method beyond the simple semigroup setting exploited so far.

Throughout this section, we consider the random transformation $P_{Y|X=x} = \mathcal{N}(x, 1)$ from $\mathbb{R}$ to $\mathbb{R}$, so that the corresponding memoryless channel is defined by $P_{Y^n|X^n=x^n} = \mathcal{N}(x^n, \mathbf{I}_n)$. In this setting, we have the following Gaussian analogue of Theorem 3.2. For later applications to broadcast channels, we consider a slight extension of the setting of Theorem 3.2 to allow stochastic encoders that use $\log_2 L$ random bits.

Theorem 3.5. *Let $P_{Y|X=x} = \mathcal{N}(x, 1)$. Consider any encoding function $\phi : [M] \times [L] \to \mathbb{R}^n$ and disjoint decoding sets $\mathcal{D}_1, \ldots, \mathcal{D}_M \subseteq \mathbb{R}^n$. Suppose that*

$$\prod_{w,v} P_{Y^n|X^n=\phi(w,v)}^{\frac{1}{ML}}[\mathcal{D}_w] \geq 1 - \epsilon. \tag{3.18}$$

Then

$$\ln M \leq I(W; Y^n) + \sqrt{2 \ln \frac{1}{1-\epsilon}} \sqrt{n} + \ln \frac{1}{1-\epsilon}, \tag{3.19}$$

where Y^n is the output of the channel $P_{Y^n|X^n}$ with input $X^n = \phi(W, V)$, and (W, V) is equiprobable on $[M] \times [L]$.

In the Gaussian setting, it is natural to work not with simple semigroups but rather with the *Ornstein-Uhlenbeck semigroup* with stationary measure $\mathcal{N}(x^n, \mathbf{I}_n)$:

$$T_{x^n,t} f(y^n) := \mathbb{E}[f(e^{-t} y^n + (1 - e^{-t}) x^n + \sqrt{1 - e^{-2t}} V^n)], \tag{3.20}$$

where $V^n \sim \mathcal{N}(0^n, \mathbf{I}_n)$. In this setting, Borell [5] showed that reverse hypercontractivity (2.21) holds under the even weaker assumption $t \geq \frac{1}{2} \ln \frac{1-q}{1-p}$.

The proof will proceed a little differently than in Section 3.1. Here, the analogue of Λ_t in (3.8) is simply $T_{0^n,t}$. Instead of Lemma 3.4, we will exploit a simple change-of-variable formula: for any $f \geq 0$, $t > 0$ and $x^n \in \mathbb{R}^n$, we have

$$P_{Y^n|X^n=x^n}(\ln T_{0^n,t} f) = P_{Y^n|X^n=e^{-t}x^n}(\ln T_{e^{-t}x^n,t} f), \tag{3.21}$$

which can be verified directly from (3.20).

Proof of Theorem 3.5. Let $f_w = 1_{\mathcal{D}_w}$ for $w \in [M]$. Define also $\bar{X}^n = e^t X^n$, and denote by $\bar{Y}^n$ the output of the channel $P_{Y^n|X^n}$ with input $X^n \leftarrow \bar{X}^n$. Note that

$$\begin{aligned} & D(P_{\bar{Y}^n|W=w} \| P_{\bar{Y}^n}) \\ & \geq \frac{1}{L}\sum_{v=1}^{L} P_{Y^n|X^n=e^t\phi(w,v)}(\ln T_{0^n,t} f_w) - \ln P_{\bar{Y}^n}(T_{0^n,t} f_w) && (3.22) \\ & = \frac{1}{L}\sum_{v=1}^{L} P_{Y^n|X^n=\phi(w,v)}(\ln T_{\phi(w,v),t} f_w) - \ln P_{\bar{Y}^n}(T_{0^n,t} f_w) && (3.23) \end{aligned}$$

for each w, where the key step (3.23) used (3.21).

The summand in the first term of the right side of (3.23) can be bounded using reverse hypercontractivity for the Ornstein-Uhlenbeck semigroup (3.20) as

$$P_{Y^n|X^n=\phi(w,v)}(\ln T_{\phi(w,v),t} f_w) \geq \frac{1}{1-e^{-2t}} \ln P_{Y^n|X^n=\phi(w,v)}(f_w). \tag{3.24}$$

Therefore, using assumption (3.18), we obtain

$$\frac{1}{ML}\sum_{w,v} P_{Y^n|X^n=\phi(w,v)}(\ln T_{\phi(w,v),t} f_w) \geq -\frac{1}{1-e^{-2t}} \ln \frac{1}{1-\epsilon}. \tag{3.25}$$

On the other hand, by Jensen's inequality,

$$\frac{1}{M}\sum_{w=1}^{M} \ln P_{\bar{Y}^n}(T_{0^n,t} f_w) \leq \ln P_{\bar{Y}^n}\left(\frac{1}{M}\sum_{w=1}^{M} T_{0^n,t} f_w\right) \leq \ln \frac{1}{M}, \tag{3.26}$$

where we used $\sum_w f_w \leq 1$. Thus, averaging over w in (3.23), we find

$$\ln M \leq I(W; \bar{Y}^n) + \frac{1}{1-e^{-2t}} \ln \frac{1}{1-\epsilon}. \tag{3.27}$$

We must now bound $I(W; \bar{Y}^n)$ in terms of $I(W; Y^n)$. To this end, let $G^n \sim \mathcal{N}(0^n, \mathbf{I}_n)$ be independent of X^n and note that

$$\begin{aligned} h(\bar{Y}^n) &= h(e^t X^n + G^n) && (3.28) \\ &= h(X^n + e^{-t} G^n) + nt && (3.29) \\ &\leq h(Y^n) + nt, && (3.30) \end{aligned}$$

where (3.30) can be seen from the entropy power inequality. On the other hand,

$$\begin{aligned} & h(\bar{Y}^n|W=w) - h(Y^n|W=w) \\ & = I(\bar{Y}^n; X^n|W=w) - I(Y^n; X^n|W=w) \geq 0 && (3.31) \end{aligned}$$

for every w, where the equality follows as $h(\bar{Y}^n|X^n, W = w) = h(Y^n|X^n, W = w) = h(G^n)$ and nonnegativity can be seen from [17, Theorem 1]. Combining (3.27), (3.30), and (3.31) and using $e^t \geq 1 + t$ yields

$$\ln M \leq I(W; Y^n) + \left(\frac{1}{2t} + 1\right) \ln \frac{1}{1-\epsilon} + nt, \tag{3.32}$$

and the conclusion follows by optimizing over t. □

Remark 3.6. If the geometric average criterion (3.18) is replaced by the stronger maximal error criterion, a Gaussian Fano inequality with $\sim \sqrt{n}$ second-order term can also be obtained using the information spectrum method and the Gaussian Poincaré inequality; see [43, Theorem 8]. However, beside the stronger assumption, the resulting bound is much less explicit and does not recover the correct dependence of the second-order term on the error probability as $\epsilon \to 1$.

3.3 Application: output distribution of good channel codes

As a very first application, let us note that Theorem 3.2 effortlessly gives a sharp non-asymptotic converse for the basic data transmission example of Section 1.2 for discrete memoryless channels. Since by the data processing inequality

$$I(W; Y^n) \leq n \sup_{P_X} I(X; Y) =: nC,$$

the maximal code size satisfies

$$\ln M^*(n, \epsilon) \leq nC + 2\sqrt{|\mathcal{Y}| \ln \frac{1}{1-\epsilon}}\sqrt{n} + \ln \frac{1}{1-\epsilon}, \tag{3.33}$$

where we chose the reference measure ν in (3.3) to be uniform on $\mathcal{Y}$.

It was observed in [19, 46] that if a code is close to achieving the upper bound (3.33) , this places a strong constraint on what the transmitted message looks like. In particular, it was shown there that if an (n, M, ϵ)-code satisfies

$$\ln M = nC + o(n), \tag{3.34}$$

then it is necessarily the case that

$$\frac{1}{n} D(P_{Y^n} \| P^\star_{Y^n}) = o(\epsilon), \tag{3.35}$$

where P_{Y^n} is the channel output distribution induced by the codebook and $P^\star_{Y^n} := P_Y^{\star \otimes n}$ and $P^\star_Y$ is the unique maximal mutual information output distribution (note that corresponding optimal input distribution need not be unique). That is, the output distribution of a good code must approximate the capacity-achieving output distribution in the small error probability regime. Such a necessary condition for good channel codes sometimes provides guidelines for their design.

Theorem 3.2 enables us to develop a sharp quantitative form of this phenomenon for *any* discrete memoryless channel.

Theorem 3.7. *Consider a discrete memoryless channel $P_{Y|X}$ with capacity C. An (n, M, ϵ) code (under the maximal error criterion) satisfies*

$$D(P_{Y^n}\|P^\star_{Y^n}) \le nC - \ln M + 2\sqrt{|\mathcal{Y}|\ln\frac{1}{1-\epsilon}}\sqrt{n} + \ln\frac{1}{1-\epsilon} \tag{3.36}$$

where $P^\star_{Y^n}$ is the unique capacity achieving output distribution defined above.

Two second-order bounds on the approximation error $D(P_{Y^n}\|P^\star_{Y^n})$ were previously derived in [43]. In [43, Theorem 7], an analogue of Theorem 3.7 is obtained using the blowing-up method; as usual, this results in a suboptimal second-order term. However, in [43, Theorem 6] (see [45, Theorem 3.6.6] for a sharper formulation) a result analogous to Theorem 3.7 is obtained with sharp second-order term under the following assumption (the Burnashev condition [7]):

$$\sup_{x,x'}\left\|\frac{dP_{Y|X=x}}{dP_{Y|X=x'}}\right\|_\infty < \infty. \tag{3.37}$$

Theorem 3.7 shows that the Burnashev condition is unnecessary: the result holds for any discrete memoryless channel. Its proof is completely straightforward; one simply follows the steps in [43, (64)-(66)] using the optimal second-order Fano inequality (Theorem 3.2) in lieu of the blowing-up argument.

Proof of Theorem 3.7. It suffices to note that

$$D(P_{Y^n}\|P^\star_{Y^n}) = D(P_{Y^n|X^n}\|P^\star_{Y^n}|P_{X^n}) - I(X^n;Y^n) \tag{3.38}$$

$$\le nC - I(X^n;Y^n), \tag{3.39}$$

where (3.38) is a well-known identity, and (3.39) follows from the saddle point property of the capacity-achieving distribution [43, eq. (19)]. It remains to bound $I(X^n;Y^n)$ using Theorem 3.2, where we choose ν in (3.3) to be uniform on $\mathcal{Y}$. □

Remark 3.8. We have stated the above form of Theorem 3.7 for simplicity. Without any additional difficulty, the finite alphabet assumption may be weakened to a bounded density assumption (3.3) and the maximal error criterion may be weakened to the geometric average criterion (3.4). A Gaussian counterpart of Theorem 3.7 can also be obtained in a similar manner using Theorem 3.5, strengthening the result of [43, Theorem 8] to the geometric average criterion.

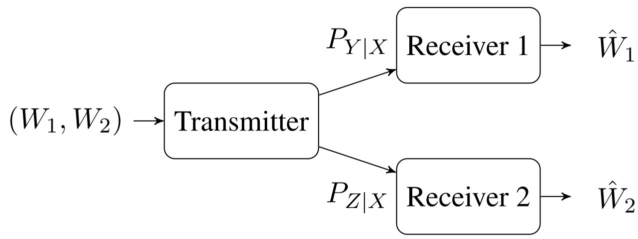

Figure 1: The broadcast channel

Remark 3.9. By codebook expurgation (cf. Remark 3.3), one can slightly modify the strong channel coding converse (3.33) to remain valid even under the average error criterion. In contrast, the proof of Theorem 3.7 relied on the sharper bound on $I(W; Y^n)$ given in Theorem 3.2, and thus the codebook expurgation method does not apply in this setting. It is shown in [31] that the geometric average criterion developed here is (in a particular sense) the weakest criterion under which Theorem 3.7 may hold, so that our approach is optimal also in this sense.

3.4 Application: broadcast channels

Determining the capacity region of general broadcast channels is a long-standing open problem in information theory [13]. The general setting is illustrated in Figure 3.4. A transmitter wishes to send messages (W_1, W_2), equiprobable on $[M_1] \times [M_2]$, to two respective receivers, using n repetitions of the random transformations $P_{Y|X}$ and $P_{Z|X}$. A code consisting of codewords $c_{u,v} \in \mathcal{X}^n$ (with $u \in [M_1]$, $v \in [M_2]$) is said to be an (n, M_1, M_2, ϵ)-*code* under the average error criterion, if the receivers can reconstruct $\hat{W}_1$ and $\hat{W}_2$ from the outputs of their respective channels with error probabilities

$$\mathbb{P}[\hat{W}_1 \neq W_1] \leq \epsilon, \qquad \mathbb{P}[\hat{W}_2 \neq W_2] \leq \epsilon. \tag{3.40}$$

The problem of understanding precisely what codebook sizes M_1, M_2 can be transmitted with a given error probability remains, in general, an open problem. However, we will consider the case of *degraded broadcast channels* which is much better understood. The additional structural assumption in this case is that $P_{Z|X}$ equals the concatenation of $P_{Y|X}$ and a certain random transformation $P_{Z|Y}$ (or the other way around). In particular, in the setting of Gaussian broadcast channels, this additional assumption is always satisfied.

The aim of this section is to show how to obtain second-order converses for both Gaussian and discrete degraded broadcast channels using our methods, improving the best previously known bounds. To avoid digressions unrelated to the topic of this paper, we will only sketch how Theorems 3.2 and 3.5 enter the proofs.

The subsequent manipulations of the first-order terms (using the chain rule and the entropy power inequality) are standard; we refer to [13] for the omitted steps and for matching achievability results. The maximal error criterion for the broadcast channel reads as

$$\max_{w_1,w_2} \mathbb{P}[\hat{W}_1 \neq W_1 | (W_1, W_2) = (w_1, w_2)] \leq \epsilon, \tag{3.41}$$

$$\max_{w_1,w_2} \mathbb{P}[\hat{W}_2 \neq W_2 | (W_1, W_2) = (w_1, w_2)] \leq \epsilon. \tag{3.42}$$

The reduction to the average error criterion for the broadcast channel can also be done by a codebook expurgation argument, albeit more sophisticated than the single-user setting; see [13, Problem 8.11].

We now introduce the following geometric average decodability criterion for the broadcast channel, which interpolates the average and the maximal error criteria above (see Remark 3.3).

$$\prod_{w_1,w_2} \mathbb{P}^{\frac{1}{M_1 M_2}}[\hat{W}_1 = W_1 | (W_1, W_2) = (w_1, w_2)] \geq 1 - \epsilon, \tag{3.43}$$

$$\prod_{w_1,w_2} \mathbb{P}^{\frac{1}{M_1 M_2}}[\hat{W}_2 = W_2 | (W_1, W_2) = (w_1, w_2)] \geq 1 - \epsilon. \tag{3.44}$$

The geometric average criterion integrates seamlessly with our proof, which does not involve an expurgation argument.

We begin by stating the Gaussian result. In the stationary memoryless Gaussian broadcast channel, it is assumed that $P_{Y^n|X^n} = P_{Y|X}^{\otimes n}$ and $P_{Z^n|X^n} = P_{Z|X}^{\otimes n}$ with $P_{Y|X=x} = \mathcal{N}(x, \sigma_1^2)$ and $P_{Z|X=x} = \mathcal{N}(x, \sigma_2^2)$. We assume moreover that the codewords $c \in \mathbb{R}^n$ must satisfy the power constraint $\|c\|^2 \leq nP$. The signal-to-noise ratios (SNR) of the two channels are then defined by $S_i := P/\sigma_i^2$.

Theorem 3.10. *Consider a stationary memoryless Gaussian broadcast channel with SNRs $S_1, S_2 \in (0, \infty)$. Suppose there exists an (n, M_1, M_2, ϵ)-code (under the geometric average criterion). Then*

$$\ln M_1 \leq n\mathsf{C}(\alpha S_1) + \sqrt{2n \ln \frac{1}{1-\epsilon}} + \ln \frac{1}{1-\epsilon}, \tag{3.45}$$

$$\ln M_2 \leq n\mathsf{C}\left(\frac{(1-\alpha)S_2}{\alpha S_2 + 1}\right) + \sqrt{2n \ln \frac{1}{1-\epsilon}} + \ln \frac{1}{1-\epsilon} \tag{3.46}$$

for some $\alpha \in [0, 1]$, where $\mathsf{C}(t) := \frac{1}{2}\ln(1+t)$.

Proof. We first use Theorem 3.5 to obtain[6]

$$\ln M_1 \leq I(W_1; Y^n|W_2) + \sqrt{2n \ln \frac{1}{1-\epsilon}} + \ln \frac{1}{1-\epsilon}, \tag{3.47}$$

$$\ln M_2 \leq I(W_2; Z^n) + \sqrt{2n \ln \frac{1}{1-\epsilon}} + \ln \frac{1}{1-\epsilon}. \tag{3.48}$$

To see (3.47), we first apply Theorem 3.5 to obtain:

$$\ln M_1 \leq I(W_1; Y^n|W_2 = w_2) + \sqrt{2n \ln \frac{1}{1-\epsilon_{w_2}}} + \ln \frac{1}{1-\epsilon_{w_2}}, \tag{3.49}$$

where we defined

$$\epsilon_{w_2} = 1 - \prod_{w_1} \mathbb{P}^{\frac{1}{M_1}}[\hat{W}_1 = W_1|(W_1, W_2) = (w_1, w_2)].$$

Then we obtain (3.47) by averaging (3.49) over w_2 and applying Jensen's inequality:

$$\frac{1}{M_2} \sum_{w_2} \sqrt{2n \ln \frac{1}{1-\epsilon_{w_2}}} \leq \sqrt{2n \frac{1}{M_2} \cdot \sum_{w_2} \ln \frac{1}{1-\epsilon_{w_2}}} \tag{3.50}$$

$$= \sqrt{2n \ln \frac{1}{\prod_{w_2} (1-\epsilon_{w_2})^{\frac{1}{M_2}}}} \tag{3.51}$$

$$\leq \sqrt{2n \ln \frac{1}{1-\epsilon}}, \tag{3.52}$$

where we used the fact that

$$\prod_{w_2} (1-\epsilon_{w_2})^{\frac{1}{M_2}} = \prod_{w_1, w_2} \mathbb{P}^{\frac{1}{M_1 M_2}}[\hat{W}_1 = W_1|(W_1, W_2) = (w_1, w_2)] \geq 1 - \epsilon.$$

Now it remains to bound the mutual informations $I(W_1; Y^n|W_2)$ and $I(W_2; Z^n)$ by the respective capacities; the proof of this part of the argument is identical that of the weak converse [13, Theorem 5.3], and we omit the details. □

The analogous result in the discrete case is as follows.

[6]Note that as no power constraint is assumed in Theorem 3.5, its conclusion extends verbatim to channels with arbitrary positive variance by scaling.

Theorem 3.11. *Consider a degraded discrete memoryless broadcast channel* $(P_{Y|X}, P_{Z|X})$ *and* (n, M_1, M_2, ϵ)*-code (under the geometric average criterion). Then*

$$\ln M_1 \le nI(X;Y|U) + 2\sqrt{|\mathcal{Y}| n \ln \frac{1}{1-\epsilon}} + \ln \frac{1}{1-\epsilon}, \tag{3.53}$$

$$\ln M_2 \le nI(U;Z) + 2\sqrt{|\mathcal{Z}| n \ln \frac{1}{1-\epsilon}} + \ln \frac{1}{1-\epsilon} \tag{3.54}$$

for some distribution P_{UX}*.*

Proof. The steps are identical to those of Theorem 3.10, modulo the (trivial) generalization of Theorem 3.2 to allow stochastic encoders as we did in Theorem 3.5. The omitted manipulations of the mutual information terms follow identically the proof of the weak converse in this setting [13, Theorem 5.2]. □

Theorem 3.11 appears to be the first second-order converse for the discrete degraded broadcast channel, improving the suboptimal converse obtained by the blowing-up method [3] (cf. [45, Theorem 3.6.4]). Our approach further extends beyond finite alphabets to the bounded density setting (3.3). For the Gaussian broadcast channel, a strong converse with $\sim \sqrt{n}$ second-order term was previously obtained in [16] via the information spectrum method. The present bound (3.46) is sharper; in particular, it has the correct asymptotic dependence not only on $n \to \infty$ but also on the error probability $\epsilon \to 1$.

4 Second-Order Image Size Characterization

The aim of this section is to develop the smoothing-out methodology for coding problems in network information theory that are proved by the "image size characterization" method. This powerful machinery was introduced in the original work of Ahlswede, Gács and Körner [3] for proving a strong converse for source coding with side information (cf. Section 4.5), and was further developed systematically in the classic monograph of Csiszár and Körner [11, chapters 15–16] to enable the analysis of a wide variety of source and channel networks.

Unfortunately, the general formulation of the image size method may appear rather technical at first sight, particularly if the reader is unfamiliar with the classical theory developed in [11]. To make the theory more accessible, we will first provide some motivating discussion and background in Section 4.1. The main results of this section, which give second-order forms of the image-size characterization method in discrete and Gaussian cases, are given in Sections 4.2 and 4.3, respectively. Finally, we will illustrate the application of our framework in two representative applications in Sections 4.4 and 4.5.

It should be emphasized that in order to keep this paper to a reasonable length, we treat here only the more basic setting of the image size problem that appears in the original paper [3]. However, our methods extend in full generality to the theory developed in [11], which yields the strong converse property of all source-channel networks with known first-order rate region. In the terminology of [11], the problem considered here only contains a "forward channel" while the general theory allows for the addition of a "reverse channel"; in fact, besides yielding second-order converses and extending to general alphabets, our methods have the further advantage of simplifying certain technical aspects in dealing with the reverse channel. A detailed development of our framework for the most general form of the image size characterization problem may be found in [26, Chapter 5].

4.1 Synopsis

4.1.1 A motivating example

Before we introduce the general setting for image size characterization, it is instructive to motivate its form through a simple application. To this end, let us begin by describing a variant of the binary hypothesis testing problem, due to Ahlswede and Csiszár [2], in which the image size problem arises naturally. After we have developed the general results, we will return to this example and provide a second-order converse using our methods in Section 4.4.

Suppose

we are given a joint probability measure Q_{XY}, and we obtain n independent observations drawn from either Q_{XY} or from $Q_X \otimes Q_Y$. We would like to construct an optimal hypothesis test between these two alternatives:

$$\begin{aligned} &\mathrm{H}_0 \colon Q_{XY}^{\otimes n}, \\ &\mathrm{H}_1 \colon Q_X^{\otimes n} Q_Y^{\otimes n}. \end{aligned} \tag{4.1}$$

When stated in this form, this is just a special case of the general binary hypothesis testing problem discussed in Section 2. In particular, for any test $\mathcal{B} \subseteq \mathcal{X}^n \times \mathcal{Y}^n$ with error probability $\pi_{1|0} := Q_{XY}^{\otimes n}[\mathcal{B}^c] \le \epsilon$, we would like to understand how small an error probability $\pi_{0|1} := Q_X^{\otimes n} Q_Y^{\otimes n}[\mathcal{B}]$ can be achieved (we will focus here on deterministic tests for simplicity). In other words, the aim of the binary hypothesis testing problem is to understand the quantity

$$\min_{Q_{X^n Y^n}[\mathcal{B}] \ge 1-\epsilon} Q_{X^n} Q_{Y^n}[\mathcal{B}]. \tag{4.2}$$

Both a second-order converse (lower bound) and the matching achievability argument (upper bound) follow immediately from the results in Section 2.

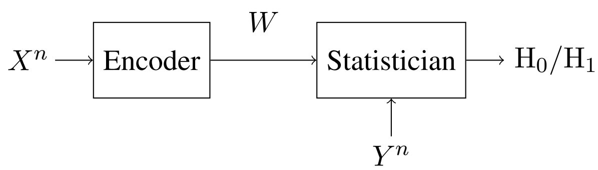

Figure 2: Hypothesis testing with a communication constraint.

We now, however, add a new ingredient. Suppose only Y^n is directly revealed to the statistician; the data X^n is observed elsewhere, and must be communicated to the statistician through a (noiseless) communication channel with limited capacity. That is, an encoder must encode the data X^n into a message $W \in [M]$ that is transmitted noiselessly to the statistician; the capacity of the channel limits the size of the codebook to $M \in \mathbb{N}$. This situation is illustrated in Figure 2. The communication constraint may be described by restricting the original binary hypothesis testing problem to a special class of tests of the form

$$\mathcal{B} = \bigcup_{i=1}^{M} (\mathcal{A}_i \times \mathcal{B}_i) \subseteq \mathcal{X}^n \times \mathcal{Y}^n, \tag{4.3}$$

where $\mathcal{A}_i = \{W = i\}$ is the set of strings X^n for which the message i is transmitted to the statistician, and $\mathcal{B}_i$ is the set of strings Y^n for which the statistician declares H_0 given that the message i was received. In this setting, we aim to understand the interplay between the channel capacity and the attainable error probabilities of the test. That is, the aim of the hypothesis testing problem with communication constraint is to understand the quantity

$$\min_{Q_{X^nY^n}[\mathcal{B}] \geq 1-\epsilon,\ \mathcal{B} \text{ as in } (4.3)} Q_{X^n} Q_{Y^n}[\mathcal{B}]. \tag{4.4}$$

In accordance with the theme of this paper, we will be concerned here with the converse direction only; the achievability argument may be found in [2].

In Section 4.4, we will give a second-order converse for the problem in (4.4). For the purposes of the present motivating discussion, it will be convenient for simplicity to consider the stronger maximum error criterion variant[7] of this problem: how small can $Q_{X^n} Q_{Y^n}[\mathcal{B}|X^n]$ be made for communication-constrained tests (4.3), given that $Q_{X^nY^n}[\mathcal{B}|X^n] \geq 1 - \epsilon$? By applying (4.3), this question can be reformulated as follows: if we have

$$\max_{x^n \in \mathcal{A}_i} Q_{Y^n|X^n=x^n}[\mathcal{B}_i^c] \leq \epsilon, \tag{4.5}$$

[7] In Section 4.4 we reduce the original problem to such a variant by expurgation (cf. Remark 3.3); however, for reasons explained in Section 4.1.3, this must be done carefully to attain the correct rate.

how small can $Q_{Y^n}[\mathcal{B}_i]$ be made? In this formulation, it is not entirely clear how the codebook size M enters the picture. However, note that as there are only M codewords, at least one of the codewords must have large probability $Q_{X^n}[\mathcal{A}_i] \geq 1/M$. Therefore, a lower bound for the maximal error variant of (4.4) naturally reduces to the problem of understanding the following quantity:

$$\min_{Q_{X^n}[Q_{Y^n|X^n}[\mathcal{B}'] \geq 1-\epsilon] \geq 1/M} Q_{Y^n}[\mathcal{B}']. \tag{4.6}$$

The latter problem, which was originally considered in [3], is the most basic form of the *image size characterization problem* that is studied in this section.[8]

The motivation we have given here for the image size problem may appear rather specific to binary hypothesis testing with a communication constraint. However, just as ideas from binary hypothesis testing form the basis for a wide variety of coding problems, the image size problem turns out to arise in a broad range of problems in network information theory. After developing the general theory, we will discuss in some detail two applications: the present hypothesis testing problem with communication constraint (Section 4.4) and source coding with compressed side information (Section 4.5). Further applications are discussed in [11, Chapter 16]. Another interesting application to the problem of common randomness generation is developed in detail in [26, Section 4.4].

4.1.2 The image size problem

We now turn to the formulation of the basic objects and results that will be investigated in this section.

In the following, let $Q_{Y|X}$ be a given random transformation, and let ν and μ_n be positive measures on $\mathcal{Y}$ and $\mathcal{X}^n$, respectively. The general problem that will be investigated in this section is how to lower bound the $\nu^{\otimes n}$-measure of a set $\mathcal{A} \subseteq Y^n$ in terms of its "$(1-\epsilon)$-preimage" under $Q_{Y^n|X^n} := Q_{Y|X}^{\otimes n}$. Instead of the formulation in (4.6), we find it convenient to introduce a Lagrange multiplier c and investigate the following unconstrained version of the problem: given any $c > 0$ and $\epsilon \in (0,1)$, find an upper bound on the quantity

$$\sup_{\mathcal{A} \subseteq \mathcal{Y}^n} \{\ln \mu_n[Q_{Y^n|X^n}[\mathcal{A}] > 1-\epsilon] - c \ln \nu^{\otimes n}[\mathcal{A}]\}. \tag{4.7}$$

To understand the behavior of this quantity, we begin by defining the information measure that controls it to first order.

---

[8] The name arises from the fact that we think of $\mathcal{B}_i$ as an "$(1-\epsilon)$-image" of $\mathcal{A}_i$ under the random transformation $Q_{Y^n|X^n}$ [11].

Definition 4.1. Given positive measures μ on $\mathcal{X}$ and ν on $\mathcal{Y}$, a random transformation $Q_{Y|X}$, and constant $c > 0$, we define

$$\mathrm{d}(\mu, Q_{Y|X}, \nu, c) := \sup_{P_X} \{cD(P_Y \| \nu) - D(P_X \| \mu)\} \tag{4.8}$$

where the probability measures P_X, P_Y satisfy $P_X \to Q_{Y|X} \to P_Y$.

Remark 4.2. The quantity (4.8) was called "Brascamp-Lieb divergence" in [29]. Let us also note that the largest $c > 0$ for which $d(Q_X, Q_{Y|X}, Q_Y, c) = 0$ is the reciprocal of the *strong data processing constant* of $Q_{Y|X}$, cf. [28].

To establish the natural connection between (4.7) and (4.8), let us begin by proving a weak converse. Evidently the method of proof has much in common with that of Lemma 2.1 and is completely elementary (cf. [3, (19)-(21)]).

Lemma 4.3 (**weak converse**)**.** *If ν is a probability measure, then*

$$\begin{aligned} &\ln \mu_n[Q_{Y^n|X^n}[\mathcal{A}] > 1 - \epsilon] - c(1 - \epsilon) \ln \nu^{\otimes n}[\mathcal{A}] \\ &\quad \le \mathrm{d}(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) + c \ln 2 \end{aligned} \tag{4.9}$$

for every set $\mathcal{A} \subseteq \mathcal{Y}^n$ and $c > 0$, $\epsilon \in (0, 1)$, $Q_{Y|X}$, and positive measure μ_n.

Proof. Define the event $\mathcal{B}$ and probability measure Q_{X^n} by

$$\mathcal{B} = \{x^n : Q_{Y^n|X^n=x^n}[\mathcal{A}] > 1 - \epsilon\}, \qquad Q_{X^n}[\mathcal{C}] = \frac{\mu_n[\mathcal{B} \cap \mathcal{C}]}{\mu_n[\mathcal{B}]}, \tag{4.10}$$

and let Q_{Y^n} be defined by $Q_{X^n} \to Q_{Y^n|X^n} \to Q_{Y^n}$. Then

$$D(Q_{X^n} \| \mu_n) = \ln \frac{1}{\mu_n[\mathcal{B}]}, \tag{4.11}$$

while by the data processing inequality

$$D(Q_{Y^n} \| \nu^{\otimes n}) \ge Q_{Y^n}[\mathcal{A}] \ln \frac{Q_{Y^n}[\mathcal{A}]}{\nu^{\otimes n}[\mathcal{A}]} + Q_{Y^n}[\mathcal{A}^c] \ln \frac{Q_{Y^n}[\mathcal{A}^c]}{\nu^{\otimes n}[\mathcal{A}^c]} \tag{4.12}$$

$$\ge (1 - \epsilon) \ln \frac{1}{\nu^{\otimes n}[\mathcal{A}]} - h(Q_{Y^n}[\mathcal{A}^c]). \tag{4.13}$$

Thus,

$$\mathrm{d}(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}) \ge cD(Q_{Y^n} \| \nu^{\otimes n}) - D(Q_{X^n} \| \mu_n) \tag{4.14}$$

$$\ge c(1 - \epsilon) \ln \frac{1}{\nu^{\otimes n}[\mathcal{A}]} - \ln \frac{1}{\mu_n[\mathcal{B}]} - c \ln 2, \tag{4.15}$$

which is (4.9). $\square$

Lemma 4.3 is a weak converse due to the extraneous factor $1-\epsilon$ in front of $\ln \nu^{\otimes n}[\mathcal{A}]$ in (4.9), which prevents us from attaining the optimal rate in the regime of nonvanishing error probability. Instead, we aim to prove a strong converse

$$\begin{aligned} \ln \mu_n[Q_{Y^n|X^n}[\mathcal{A}] > 1-\epsilon] - c \ln \nu^{\otimes n}[\mathcal{A}] \\ \leq \mathrm{d}(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) + o_\epsilon(n). \end{aligned} \tag{4.16}$$

Ahlswede, Gács and Körner [3] used the blowing-up method to establish (4.16) for finite alphabets. As usual, such an argument can only attain a suboptimal second-order term. In the follows subsections, we will use the smoothing-out method to obtain new sharp forms of (4.16) that attain both the $O(\sqrt{n})$ second-order term, and that may be extended beyond the finite-alphabet setting. In particular, we develop the relevant theory for discrete channels in Section 4.2, and for Gaussian channels in Section 4.3.

4.1.3 The first-order term

Before we proceed to the main results of this section, we must discuss an independent issue that plays an important role in applications of image size characterizations. Consider the motivating example of hypothesis testing with a communication constraint discussed above. A lower bound in (4.6) is readily obtained by applying Lemma 4.3 or (4.16) with $\mu_n = Q_X^{\otimes n}$. When all measures in $\mathrm{d}(\cdot)$ are product measures, it is readily evaluated by means of the *tensorization property*

$$\mathrm{d}(Q_X^{\otimes n}, Q_{Y|X}^{\otimes n}, Q_Y^{\otimes n}, c) = n\,\mathrm{d}(Q_X, Q_{Y|X}, Q_Y, c) \tag{4.17}$$

(a simple proof may be found, for example, in [28]). Thus, to first order, the quantity defined in (4.6) grows linearly in n with rate $\mathrm{d}(Q_X, Q_{Y|X}, Q_Y, c)$ (with the optimal choice of Lagrange multiplier c). This is in fact the correct growth rate of (4.6), as is demonstrated in [3, Theorem 1]. Unfortunately, however, this quantity does *not* give the correct first-order rate for the hypothesis testing problem with communication constraint: the best achievability result for the latter has a strictly smaller rate than is suggested by the above arguments.

It turns out that the origin of this inefficiency does not lie in (4.16) itself, but rather in the initial argument that reduced the information-theoretic problem to the study of the quantity (4.6). When the argument is developed in detail, we will see that it is not necessary to consider the probability of codewords under the full measure Q_{X^n}, but only under the restriction of this measure to any set $\mathcal{C}$ of sufficiently high probability. While the information-theoretic problem is not affected by throwing out a small probability event, doing so has a significant effect on the resulting bounds. In particular, as was noted in [3] in the finite alphabet setting, the best possible improvement to the first-order rate is obtained by choosing $\mu_n = Q_{X^n}|_{\mathcal{C}_n}$ to be the restriction to a set $\mathcal{C}_n$ of *typical sequences*. As we will show in Appendices B.1 and B.2 for the discrete and Gaussian cases, respectively,[9] such a choice of μ_n in fact gives rise to the estimate

$$\mathrm{d}(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) \leq n\,\mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) + O(\sqrt{n}) \tag{4.18}$$

where the modified rate $\mathrm{d}^\star(\cdot)$ is defined as follows.

Definition 4.4. Given a probability measure Q_X on $\mathcal{X}$, a positive measure ν on $\mathcal{Y}$, a random transformation $Q_{Y|X}$, and constant $c > 0$, we define

$$\begin{aligned}&\mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) := \\ &\qquad \sup_{P_{UX}\colon P_X = Q_X} \left\{ cD(P_{Y|U}\|\nu|P_U) - D(P_{X|U}\|Q_X|P_U) \right\}\end{aligned} \tag{4.19}$$

where the joint distribution of U, X, Y is given by $P_{UXY} := P_{UX}Q_{Y|X}$.

We observe that it follows immediately from Definitions 4.1 and 4.4 that $\mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) \leq \mathrm{d}(Q_X, Q_{Y|X}, \nu, c)$. Therefore, restricting to a typical set always results in an improved bound. It will turn out that the rate (4.19) captures precisely the correct first-order behavior of the information-theoretic applications in which image size characterizations will be applied. Moreover, as we are able to establish in (4.18) a second-order term of order $O(\sqrt{n})$, the combination of such an estimate with the smoothing-out method enables us to attain second-order converses for applications of the image size problem. As the methods needed to establish (4.18) are unrelated to the main topic of this paper, we focus presently on the smoothing-out method and relegate the proof of (4.18) to Appendix B.

4.2 Discrete case

The aim of this section is to established a non-asymptotic form of the strong converse for image size characterization (4.16) in the finite alphabet setting. We will, in fact, initially work in the more general bounded density setting as in Section 3.1. We specialize subsequently to the finite alphabet setting only in order to achieve the appropriate characterization (4.18) of the first-order term.

[9] The claim in (4.18) is the main theme of the conference paper [29] and the journal paper [30]. The proof of (4.18) in the present paper is based on the concentration of the empirical distribution of X^n, which is similar in spirit to [29]. The proof in [30] adopts a different convex analysis argument which establishes (4.18) for general distributions and also determines the exact prefactor in the $O(\sqrt{n})$ term. However, for conceptual simplicity we do not adopt the approach of [30] in the present paper.

For the time being, let $\mathcal{X}, \mathcal{Y}$ be general alphabets, let $Q_{Y|X}$ be a random transformation from $\mathcal{X}$ to $\mathcal{Y}$, and let ν be a probability measure on $\mathcal{Y}$. We define, as usual, $Q_{Y^n|X^n} := Q_{Y|X}^{\otimes n}$, and introduce the bounded density assumption

$$\alpha := \sup_{x \in \mathcal{X}} \left\| \frac{dQ_{Y|X=x}}{d\nu} \right\|_\infty \in [1, \infty). \tag{4.20}$$

The basic result of this section is the following quantitative form of (4.16).

Theorem 4.5. *For any positive measure μ_n on $\mathcal{X}^n$, $f \in \mathcal{H}_{[0,1]}(\mathcal{Y}^n)$, $\eta \in (0,1)$, and $c > 0$, we have*

$$\begin{aligned} &\ln \mu_n[Q_{Y^n|X^n}(f) \geq \eta] - c \ln \nu^{\otimes n}(f) \\ &\quad \leq \mathrm{d}(\mu_n, Q_{Y^n|X^n}, \nu^{\otimes n}, c) + 2c\sqrt{\ln \frac{1}{\eta}}\sqrt{n(\alpha - 1)} + c \ln \frac{1}{\eta}. \end{aligned} \tag{4.21}$$

The proof of Theorem 4.5 relies on similar ideas to the proof of Theorem 3.2. As a first step, we must develop a variational characterization for the quantity $\mathrm{d}(\cdot)$, generalizing the variational formula (2.12) for relative entropy.

Proposition 4.6. *In the setting of Definition 4.1,*

$$\mathrm{d}(\mu, Q_{Y|X}, \nu, c) = \sup_{f \in \mathcal{H}_+(\mathcal{Y})} \left\{ \ln \mu(e^{cQ_{Y|X}(\ln f)}) - c \ln \nu(f) \right\}. \tag{4.22}$$

Proof. There are several proofs in the literature (cf. [28]); here we give a short proof using (2.12). First, for any f, define a probability measure P_X by

$$\frac{dP_X}{d\mu} = \frac{e^{cQ_{Y|X}(\ln f)}}{\mu(e^{cQ_{Y|X}(\ln f)})}, \tag{4.23}$$

and define P_Y by $P_X \to Q_{Y|X} \to P_Y$. Then by (2.12), we have

$$\begin{aligned} &\ln \mu(e^{cQ_{Y|X}(\ln f)}) - c \ln \nu(f) \\ &\leq \ln \mu(e^{cQ_{Y|X}(\ln f)}) + cD(P_Y \| \nu) - cP_Y(\ln f) \qquad (4.24) \\ &= cD(P_Y \| \nu) - D(P_X \| \mu). \qquad (4.25) \end{aligned}$$

This proves the $\geq$ part of (4.22).

Conversely, for any P_X, we have using (2.12)

$$\begin{aligned} &cD(P_Y \| \nu) - D(P_X \| \mu) \\ &\leq cD(P_Y \| \nu) - P_X(cQ_{Y|X}(\ln f)) + \ln \mu(e^{cQ_{Y|X}(\ln f)}) \qquad (4.26) \\ &= \ln \mu(e^{cQ_{Y|X}(\ln f)}) - c \ln \nu(f) \qquad (4.27) \end{aligned}$$

where we chose $f = \frac{dP_Y}{d\nu}$. This proves the $\leq$ part of (4.22). $\square$

We can now complete the proof of Theorem 4.5. In the following, we will use the same semigroup device as in Section 3.1: we define the simple semigroup

$$T_{x,t}f := e^{-t}f + (1-e^{-t})Q_{Y|X=x}(f) \tag{4.28}$$

and its product $T_{x^n,t} := T_{x_1,t}\otimes\cdots\otimes T_{x_n,t}$, and we define the corresponding dominating operator Λ_t according to (3.8). We recall, in particular, the key properties (3.9) and (3.10) that will also be used in the present proof.

Proof of Theorem 4.5. We begin by noting that, by the variational principle given in Proposition 4.6, we can estimate

$$\int \|g\|_{L^0(Q_{Y^n|X^n=x^n})} d\mu_n(x^n) = \int e^{Q_{Y^n|X^n}(\ln g)} d\mu_n \tag{4.29}$$

$$\leq e^{\mathrm{d}(\mu_n, Q_{Y^n|X^n}, \nu^{\otimes n}, c)} \|g\|_{L^{1/c}(\nu^{\otimes n})} \tag{4.30}$$

for any function $g \in \mathcal{H}_+(\mathcal{Y}^n)$ for which the integrals are finite.

Now take $g = (\Lambda_t f)^c$ and observe that by (3.9),

$$\|g\|_{L^{1/c}(\nu^{\otimes n})} \leq e^{c(\alpha-1)nt}[\nu^{\otimes n}(f)]^c. \tag{4.31}$$

On the other hand,

$$\int \|g\|_{L^0(Q_{Y^n|X^n=x^n})} d\mu_n(x^n)$$

$$= \int \|\Lambda_t f\|^c_{L^0(Q_{Y^n|X^n=x^n})} d\mu_n(x^n) \tag{4.32}$$

$$\geq \int \|f\|^c_{L^{1-e^{-t}}(Q_{Y^n|X^n=x^n})} d\mu_n(x^n) \tag{4.33}$$

$$\geq \int Q_{Y^n|X^n}(f)^{\frac{c}{1-e^{-t}}} d\mu_n \tag{4.34}$$

$$\geq \eta^{\frac{c}{1-e^{-t}}} \mu_n[Q_{Y^n|X^n}(f) \geq \eta], \tag{4.35}$$

where (4.33) used (3.10) and reverse hypercontractivity (Theorem 2.3); (4.34) used that $f \in [0,1]$; and (4.35) follows from Markov's inequality. Putting together the estimates (4.30), (4.31), and (4.35), we have shown that for every $t > 0$

$$\ln \mu_n[Q_{Y^n|X^n}(f) \geq \eta] - c \ln \nu^{\otimes n}(f)$$

$$\leq \mathrm{d}(\mu_n, Q_{Y^n|X^n}, \nu^{\otimes n}, c) + \frac{c}{1-e^{-t}} \ln \frac{1}{\eta} + c(\alpha-1)nt. \tag{4.36}$$

The desired inequality (4.21) follows using $\frac{1}{1-e^{-t}} \leq \frac{1}{t} + 1$ and optimizing over t. $\square$

As was explained in Section 4.1.2, it is essential for applications of Theorem 4.5 to make a judicious choice of measure μ_n in order to attain the correct first-order rate of the information-theoretic problems of interest. To this end, we now specialize to the case of finite alphabets and state a ready-to-use result along these lines. The precise construction of μ_n may be found in Appendix B.1.

Corollary 4.7. *Let $|\mathcal{X}| < \infty$, Q_X be a probability measure on $\mathcal{X}$, ν be a probability measure on $\mathcal{Y}$, and $Q_{Y|X}$ be a random transformation. Define*

$$\beta_X := \frac{1}{\min_x Q_X(x)} \in [1, \infty), \tag{4.37}$$

and define α as in (4.20). Then for any $\delta \in (0,1)$ and $n > 3\beta_X \ln \frac{|\mathcal{X}|}{\delta}$, we may choose a set $\mathcal{C}_n \subseteq \mathcal{X}^n$ with $Q_X^{\otimes n}[\mathcal{C}_n] \geq 1 - \delta$ such that

$$\begin{aligned} &\ln \mu_n[Q_{Y^n|X^n}(f) \geq \eta] - c \ln \nu^{\otimes n}(f) \\ &\quad \leq n\, \mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) + A\sqrt{n} + c \ln \frac{1}{\eta} \end{aligned} \tag{4.38}$$

for all $f \in \mathcal{H}_{[0,1]}(\mathcal{Y}^n)$, $c > 0$, $\eta \in (0,1)$, where we defined $\mu_n := Q_X^{\otimes n}|_{\mathcal{C}_n}$ and

$$A := \ln(\alpha^c \beta_X^{c+1}) \sqrt{3\beta_X \ln \frac{|\mathcal{X}|}{\delta}} + 2c\sqrt{(\alpha - 1) \ln \frac{1}{\eta}}. \tag{4.39}$$

Proof. This follows immediately by combining Theorem 4.5 and Theorem B.1 in Appendix B.1 (note that α_Y defined in the Appendix satisfies $\alpha_Y \leq \alpha$). □

While the formulation of Theorem 4.5 and Corollary 4.7 is closer in spirit to the formulation of the image-size characterization problem in [3, 11], it is worth noting that our approach naturally gives rise to a somewhat sharper variant of these results that can sometimes be used to obtain cleaner bounds in converse proofs. In particular, it is not really necessary to apply the Markov inequality (4.35); we may work directly in applications with the following smoother version of the problem (stated, for future reference, in the setting of Corollary 4.7).

Corollary 4.8. *Let β_X, α, δ, n and μ_n be as in Corollary 4.7. Then*

$$\begin{aligned} &\ln \int Q_{Y^n|X^n}^{c\left(1+\frac{1}{t}\right)}(f)\, d\mu_n - c \ln \nu^{\otimes n}(f) \\ &\leq n\, \mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) + c(\alpha - 1)nt + \ln(\alpha^c \beta_X^{c+1}) \sqrt{3n\beta_X \ln \frac{|\mathcal{X}|}{\delta}} \end{aligned} \tag{4.40}$$

for every $c, t > 0$ and $f \in \mathcal{H}_{[0,1]}(\mathcal{Y}^n)$.

Proof. Simply omit the use of Markov's inequality (4.35) and the subsequent optimization over t in the proof of Theorem 4.5. □

In the sequel, we will illustrate both Corollaries 4.7 and 4.8 in applications.

4.3 Gaussian case

Beside giving rise to $O(\sqrt{n})$ second-order terms, another key advantage of the smoothing-out method is that it is applicable beyond the finite alphabet setting. We will presently further illustrate this feature by developing a Gaussian analogue of the image size theory of the previous section, opening the door to systematic extension of many applications of this methodology to the Gaussian setting. The basic result of this section is the following Gaussian version of Theorem 4.5.

Theorem 4.9. *Let $Q_{Y|X=x} = \mathcal{N}(x,1)$ and ν be Lebesgue on $\mathbb{R}$. Then we have*

$$\begin{aligned}\ln \mu_n[Q_{Y^n|X^n}(f) > \eta] - c \ln \nu^{\otimes n}(f) \\ \le \mathrm{d}(\mu_n, Q_{Y^n|X^n}, \nu^{\otimes n}, c) + c\sqrt{2n \ln \frac{1}{\eta}} + c \ln \frac{1}{\eta}\end{aligned} \tag{4.41}$$

for any positive measure μ_n on $\mathbb{R}^n$, $f \in \mathcal{H}_{[0,1]}(\mathbb{R}^n)$, $\eta \in (0,1)$, and $c > 0$.

The proof uses some ideas that were introduced in Section 3.2. In particular, in this section $T_{x^n,t}$ will denote the Ornstein-Uhlenbeck semigroup (3.20), and we will again exploit heavily the change-of-variables formula (3.21).

Proof of Theorem 4.9. Denote by $\bar{\mu}_n$ the dilation of μ_n by the factor e^t, that is,

$$\bar{\mu}_n[e^t \mathcal{C}] := \mu_n[\mathcal{C}] \tag{4.42}$$

for any set $\mathcal{C}$. Applying Proposition 4.6, we can estimate

$$\int e^{Q_{Y^n|X^n}(\ln g)} d\bar{\mu}_n \le e^{\mathrm{d}(\bar{\mu}_n, Q_{Y^n|X^n}, \nu^{\otimes n}, c)} \|g\|_{L^{1/c}(\mathbb{R}^n)} \tag{4.43}$$

for any function $g \in \mathcal{H}_+(\mathbb{R}^n)$ for which the integrals are finite.

Now take $g = (T_{0^n,t} f)^c$ and observe that

$$\begin{aligned}\|g\|_{L^{1/c}(\mathbb{R}^n)} &= \|T_{0^n,t} f\|^c_{L^1(\mathbb{R}^n)} &(4.44)\\ &= e^{cnt} \|f\|^c_{L^1(\mathbb{R}^n)} &(4.45)\\ &= e^{cnt} [\nu^{\otimes n}(f)]^c, &(4.46)\end{aligned}$$

where (4.45) can be verified using Fubini's theorem and (3.20). On the other hand,

$$\int e^{Q_{Y^n|X^n}(\ln g)} d\bar{\mu}_n = \int e^{cQ_{Y^n|X^n=x^n}(\ln T_{0^n,t}f)} d\bar{\mu}_n(x^n) \tag{4.47}$$

$$= \int e^{cQ_{Y^n|X^n=e^{-t}x^n}(\ln T_{e^{-t}x^n,t}f)} d\bar{\mu}_n(x^n) \tag{4.48}$$

$$= \int \|T_{e^{-t}x^n,t}f\|^c_{L^0(Q_{Y^n|X^n=e^{-t}x^n})} d\bar{\mu}_n(x^n) \tag{4.49}$$

$$\geq \int \|f\|^c_{L^{1-e^{-2t}}(Q_{Y^n|X^n=e^{-t}x^n})} d\bar{\mu}_n(x^n) \tag{4.50}$$

$$\geq \int \|f\|^{\frac{c}{1-e^{-2t}}}_{L^1(Q_{Y^n|X^n=e^{-t}x^n})} d\bar{\mu}_n(x^n) \tag{4.51}$$

$$\geq \eta^{\frac{c}{1-e^{-2t}}} \bar{\mu}_n[x^n \colon Q_{Y^n|X^n=e^{-t}x^n}(f) > \eta] \tag{4.52}$$

$$= \eta^{\frac{c}{1-e^{-2t}}} \mu_n[Q_{Y^n|X^n}(f) > \eta], \tag{4.53}$$

where (4.48) follows from the change of variables formula (3.21); (4.50) is from (2.21) (with the sharp constant for Gaussian hypercontractivity, see secion 3.2); (4.51) used that $f \in [0,1]$; and (4.53) follows from the definition (4.42) of $\bar{\mu}_n$.

Combining (4.43), (4.46), and (4.53), we obtain

$$\ln \mu_n[Q_{Y^n|X^n}(f) > \eta] - c\ln \nu^{\otimes n}(f)$$

$$\leq \mathrm{d}(\bar{\mu}_n, Q_{Y^n|X^n}, \nu^{\otimes n}, c) + \inf_{t>0}\left\{\frac{c}{1-e^{-2t}}\ln\frac{1}{\eta} + cnt\right\} \tag{4.54}$$

$$\leq \mathrm{d}(\bar{\mu}_n, Q_{Y^n|X^n}, \nu^{\otimes n}, c) + c\sqrt{2n\ln\frac{1}{\eta}} + c\ln\frac{1}{\eta}, \tag{4.55}$$

where (4.55) used $\frac{1}{1-e^{-2t}} \leq \frac{1}{2t} + 1$.

To conclude the proof, it remains only to show that

$$\mathrm{d}(\bar{\mu}_n, Q_{Y^n|X^n}, \nu^{\otimes n}, c) \leq \mathrm{d}(\mu_n, Q_{Y^n|X^n}, \nu^{\otimes n}, c), \tag{4.56}$$

for which we use the following modification of the argument that led to (3.31). When P_{X^n} and μ_n are both scaled by e^t, the relative entropy $D(P_{X^n}\|\mu_n)$ remains unchanged. On the other hand, as $D(P_{Y^n}\|\nu^{\otimes n}) = -h(P_{Y^n})$, the same argument as was used in (3.31) yields $D(P_{\bar{Y}^n}\|\nu^{\otimes n}) \leq D(P_{Y^n}\|\nu^{\otimes n})$. Substituting these observations into Definition 4.1 concludes the proof of (4.56). □

We finally state a ready-to-use Gaussian analogue of Corollary 4.7. The precise construction of μ_n may be found in Appendix B.2.

Corollary 4.10. *Let Q_{XY} be any nondegenerate Gaussian measure, and let ν be the Lebesgue measure on $\mathbb{R}$. Then for any $\delta \in (0,1)$ and $n \geq 20 \ln \frac{2}{\delta}$, we may choose a set $\mathcal{C}_n \subseteq \mathbb{R}^n$ with $Q_X^{\otimes n}[\mathcal{C}_n] \geq 1-\delta$ such that*

$$\ln \mu_n[Q_{Y^n|X^n}(f) > \eta] - c \ln \nu^{\otimes n}(f)$$
$$\leq n\, \mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) + \sqrt{6n \ln \frac{2}{\delta}} + c\sqrt{2n \ln \frac{1}{\eta}} + c \ln \frac{1}{\eta} \tag{4.57}$$

for all $f \in \mathcal{H}_{[0,1]}(\mathbb{R}^n)$, $c > 0$, $\eta \in (0,1)$, where we defined $\mu_n := Q_X^{\otimes n}|_{\mathcal{C}_n}$.

Proof. For $Q_X = \mathcal{N}(0,\sigma^2)$ and $Q_{Y|X=x} = \mathcal{N}(x,1)$, this follows immediately by combining Theorem 4.9 and Theorem B.4 in Appendix B.2.

For a general Gaussian measure Q_{XY}, we argue as follows. Suppose first that $Q_{Y|X=x} = \mathcal{N}(ax, a^2)$ for some $a \neq 0$. We can reduce to the case already proved by applying (4.57) with $\tilde{Y} := \frac{Y}{a}$, $\tilde{f}(y^n) = f(ay^n)$; as $\nu^{\otimes n}(f) = |a|^n \nu^{\otimes n}(\tilde{f})$ and

$$\mathrm{d}^\star(Q_X, Q_{\tilde{Y}|X}, \nu, c) = \mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) + c \ln |a| \tag{4.58}$$

(the latter follows directly from Definition 4.4), it is readily verified that the conclusion of Corollary 4.10 remains valid for any $a \neq 0$. Now suppose that we further scale X, Y simultaneously by some factor $b \neq 0$; then clearly both sides of (4.57) remain unchanged. As any nondegenerate Gaussian measure may be obtained by scaling X and Y in this manner, the proof is complete. $\square$

4.4 Application: hypothesis testing with a communication constraint

We now revisit the hypothesis testing problem with communication constraint introduced in Section 4.1.1, and show how the general framework of Section 4.2 enables us to achieve a second-order converse in this setting. Let us note that the first-order term in the following result gives the correct rate for this problem: the matching achievability argument can be found in [2].

Theorem 4.11. *Let $|\mathcal{X}| < \infty$, and consider the hypothesis testing problem with communication constraint defined in Section 4.1.1. Let $\mathcal{B}$ be any test of the form (4.3) that uses at most M codewords and satisfies $\pi_{1|0} := Q_{XY}^{\otimes n}[\mathcal{B}^c] \leq \epsilon$. Then the error probability $\pi_{0|1} := Q_X^{\otimes n} Q_Y^{\otimes n}[\mathcal{B}]$ satisfies*

$$\ln \pi_{0|1} \geq -n \sup_{U:\, U-X-Y} \left\{ I(U;Y) \colon I(U;X) \leq \frac{1}{n} \ln M \right\} \tag{4.59}$$
$$- \left(2\ln(\alpha\beta)\sqrt{3\beta \ln \frac{4|\mathcal{X}|}{1-\epsilon}} + 2\sqrt{\alpha \ln \frac{4}{1-\epsilon}} \right)\sqrt{n} - 2 \ln \frac{4}{1-\epsilon}$$

for $n > 3\beta \ln \frac{4|\mathcal{X}|}{1-\epsilon}$, where $\beta := \max_x \frac{1}{Q_X(x)}$ and $\alpha := \max_x \left\| \frac{dQ_{Y|X=x}}{dQ_Y} \right\|_\infty$.

Proof. We proceed in three steps.

Step 1. We begin with an expurgation argument. Suppose the test $\mathcal{B}$ of the form (4.3) satisfies $\pi_{1|0} \le \epsilon$ and $\pi_{0|1} = \rho$. We claim that for any $\epsilon' \in (0, 1-\epsilon)$, we can modify the coding scheme such that the error under H_0 is below $\epsilon + \epsilon'$, and the error under H_1 is below $\frac{\rho}{\epsilon'}$ *conditioned on each message*.

Indeed, assume without loss of generality that the messages $i = 1, \ldots, M$ are ordered such that $Q_{X^n}Q_{Y^n}[\mathcal{B}|W = i]$ is increasing in i. Let

$$i^\dagger := \min\{i : Q_{X^n}[W > i] \le \epsilon'\}, \tag{4.60}$$

and define a new test $\mathcal{B}'$ that always declares H_1 upon receiving $i > i^\dagger$, and coincides with $\mathcal{B}$ otherwise. Then $\mathcal{B}'$ satisfies

$$Q_{X^nY^n}[\mathcal{B}'^c] = Q_{X^nY^n}[\mathcal{B}^c \cap \{W \le i^\dagger\}] + Q_{X^n}[W > i^\dagger] \le \epsilon + \epsilon'. \tag{4.61}$$

On the other hand, as $Q_{X^n}Q_{Y^n}[\mathcal{B}'|W = i] = Q_{X^n}Q_{Y^n}[\mathcal{B}|W = i]$ for $i \le i^\dagger$ and as we assumed this quantity is increasing in i, we can estimate

$$\epsilon' Q_{X^n}Q_{Y^n}[\mathcal{B}'|W = i] \le Q_{X^n}Q_{Y^n}[\mathcal{B} \cap \{W \ge i^\dagger\}] \le \rho \tag{4.62}$$

for all $i = 1, \ldots, M$ (this is trivial for $i > i^\dagger$ as then $Q_{X^n}Q_{Y^n}[\mathcal{B}'|W = i] = 0$). Thus, the claimed properties are satisfied for the modified test $\mathcal{B}'$.

Step 2. Our aim is to apply Corollary 4.8 to the modified test $\mathcal{B}'$. Fix for the time being any $\delta \in (0, 1-\epsilon-\epsilon')$, and define μ_n as in Corollary 4.7. Then

$$1 - \epsilon - \epsilon' \le \int Q_{X^nY^n}[\mathcal{B}'|X^n]\, dQ_{X^n} \le \int Q_{X^nY^n}[\mathcal{B}'|X^n]\, d\mu_n + \delta \tag{4.63}$$

by (4.61). Thus, we may bound, for any $c \ge 1$,

$$(1 - \epsilon - \epsilon' - \delta)^{c(1+\frac{1}{t})} \le \int (Q_{X^nY^n}[\mathcal{B}'|X^n])^{c(1+\frac{1}{t})}\, d\mu_n, \tag{4.64}$$

where we used Jensen's inequality and the fact that μ_n is a sub-probability measure.

Now denote by $\mathcal{B}'_i \subseteq \mathcal{Y}^n$ the set of sequences y^n for which the test $\mathcal{B}'$ declares H_0 upon receiving message i (cf. (4.3)). Then we may crudely estimate

$$\begin{aligned}
&(1 - \epsilon - \epsilon' - \delta)^{c(1+\frac{1}{t})} \\
&\le \sum_{i=1}^{M} \int (Q_{Y^n|X^n}[\mathcal{B}'_i])^{c(1+\frac{1}{t})}\, d\mu_n && (4.65) \\
&\le e^{n\mathrm{d}^\star(Q_X, Q_{Y|X}, Q_Y, c) + c(\alpha-1)nt + \ln(\alpha^c\beta^{c+1})\sqrt{3n\beta\ln\frac{|\mathcal{X}|}{\delta}}} \sum_{i=1}^{M} (Q_{Y^n}[\mathcal{B}'_i])^c && (4.66)
\end{aligned}$$

for $n > 3\beta \ln \frac{|\mathcal{X}|}{\delta}$ and $t > 0$, where we used Corollary 4.8 in (4.66). But as $Q_{Y^n}[\mathcal{B}'_i] \leq \frac{\rho}{\epsilon'}$ by (4.62), we may rearrange the above estimate to obtain

$$\begin{aligned}\ln \rho \geq &- \frac{n}{c}\,\mathrm{d}^\star(Q_X, Q_{Y|X}, Q_Y, c) - \frac{\ln M}{c}\\ &- (\alpha - 1)nt - \left(1 + \frac{1}{t}\right) \ln \frac{1}{1 - \epsilon - \epsilon' - \delta}\\ &- \frac{\ln(\alpha^c \beta^{c+1})}{c} \sqrt{3n\beta \ln \frac{|\mathcal{X}|}{\delta}} - \ln \frac{1}{\epsilon'}. \end{aligned} \tag{4.67}$$

Finally, making the convenient choices

$$\epsilon' = \frac{1 - \epsilon}{2}, \qquad \delta = \frac{1 - \epsilon}{4}, \tag{4.68}$$

recalling that $\rho = \pi_{0|1}$, and optimizing over $t > 0$ yields

$$\begin{aligned}\ln \pi_{0|1} \geq &- \frac{n}{c}\,\mathrm{d}^\star(Q_X, Q_{Y|X}, Q_Y, c) - \frac{\ln M}{c} \qquad (4.69)\\ &- \left(2 \ln(\alpha\beta) \sqrt{3\beta \ln \frac{4|\mathcal{X}|}{1 - \epsilon}} + 2\sqrt{(\alpha - 1) \ln \frac{4}{1 - \epsilon}}\right) \sqrt{n} - 2 \ln \frac{4}{1 - \epsilon}\end{aligned}$$

for $n > 3\beta \ln \frac{4|\mathcal{X}|}{1-\epsilon}$ and $c \geq 1$.

Step 3. To achieve the correct first-order rate it remains to optimize (4.69) over $c \geq 1$. To this end, let us denote the rate parameter that appears in (4.59) by

$$\theta(R) := \sup_{U\colon U - X - Y} \{I(U;Y) \colon I(U;X) \leq R\}. \tag{4.70}$$

When we choose $\nu = Q_Y$ in Definition 4.4, the latter may be written as

$$\begin{aligned}\frac{1}{c}\mathrm{d}^\star(Q_X, Q_{Y|X}, Q_Y, c) &= \sup_{U\colon U - X - Y} \left\{ I(U;Y) - \frac{1}{c} I(U;X) \right\} && (4.71)\\ &= \sup_{0 \leq R \leq \ln|\mathcal{X}|} \left\{ \theta(R) - \frac{R}{c} \right\}, && (4.72)\end{aligned}$$

where we used that $I(U;X) \leq \ln|\mathcal{X}|$. But is not hard to show that $\theta(R)$ is a concave function of R, cf. [2, Lemma 1]. Thus, we obtain for any $R' \geq 0$

$$\begin{aligned}&\inf_{c \geq 1} \left\{ \frac{1}{c}\mathrm{d}^\star(Q_X, Q_{Y|X}, Q_Y, c) + \frac{R'}{c} \right\} = && (4.73)\\ &\inf_{c > 0} \left\{ \frac{1}{c}\mathrm{d}^\star(Q_X, Q_{Y|X}, Q_Y, c) + \frac{R'}{c} \right\} = \theta(R'), && (4.74)\end{aligned}$$

where we used in the first line that $\mathrm{d}^\star(Q_X, Q_{Y|X}, Q_Y, c) = 0$ for $c \leq 1$ by the data processing inequality $I(U;Y) \leq I(U;X)$; and we used the minimax theorem in (4.74). Setting $R' = \frac{1}{n} \ln M$ completes the proof. □

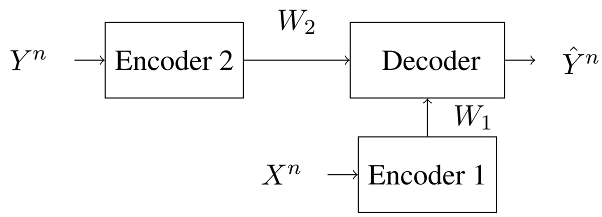

Figure 3: Source coding with compressed side information

4.5 Application: source coding with compressed side information

In this section, we revisit one of the original applications of the blowing-up method: the problem of source coding with compressed side information [58, 3]. Using the methods of this section, we will obtain second-order converses for the discrete and Gaussian variants of this problem. The present setting is a typical example of *side information problems*, whose converses have so far proved to be particularly elusive in non-asymptotic information theory (see, e.g., [51, Section 9.2]). Using the smoothing-out method, second-order converses can now be obtained by a straightforward replacement of the classical blowing-up analysis. For second-order achievability bounds for side information problems, we refer the reader to [54].

The problem that we investigate in this section is illustrated in Figure 3. Two correlated memoryless sources X^n, Y^n with distribution $Q_{X^nY^n} = Q_{XY}^{\otimes n}$ are encoded into messages $W_1 \in [M_1], W_2 \in [M_2]$, respectively, and transmitted to a decoder. The (noiseless) communication channels are rate-constrained in that they limit the sizes M_1, M_2 of the two codebooks. The decoder aims to reconstruct the source Y^n with error probability at most ϵ (the side information X^n is not reconstructed, but helps decode Y^n due to the correlation between the sources). We aim to understand the fundamental interplay between M_1, M_2 and ϵ.

We begin by developing a second-order converse in the finite alphabet setting.

Theorem 4.12. *Let $|\mathcal{X}|, |\mathcal{Y}| < \infty$, $\epsilon \in (0,1)$, and $n \geq 3\beta_X \ln \frac{4|\mathcal{X}|}{1-\epsilon}$, where β_X is defined in (4.37). Then for any encoders $f\colon \mathcal{X}^n \to [M_1]$, $g\colon \mathcal{Y}^n \to [M_2]$, and decoder $\hat{Y}^n\colon [M_1] \times [M_2] \to \mathcal{Y}^n$ such that $\mathbb{P}[Y^n \neq \hat{Y}^n] \leq \epsilon$, we have*

$$\begin{aligned} \ln M_2 \geq{} & n \inf_{U\colon U-X-Y} \left\{ H(Y|U)\colon I(U;X) \leq \frac{1}{n} \ln M_1 \right\} \\ & - \left(2\ln(|\mathcal{Y}|\beta_X) \sqrt{3\beta_X \ln \frac{4|\mathcal{X}|}{1-\epsilon}} + 2\sqrt{|\mathcal{Y}| \ln \frac{2}{1-\epsilon}} \right) \sqrt{n} - 2\ln \frac{4}{1-\epsilon}. \end{aligned} \tag{4.75}$$

Note that the first term on the right-hand side of (4.75) corresponds precisely to the rate region for the present problem (see, e.g., [13, Theorem 10.2]).

Proof. The proof follows from Corollary 4.7 using similar steps as in [3, Theorem 3]. Define for every $i \in [M_1]$ the set

$$\mathcal{B}_i := \{y^n \in \mathcal{Y}^n \colon y^n = \hat{Y}^n(i, g(y^n))\} \tag{4.76}$$

of correctly decoded sequences y^n, given that the side information message i was received. Then we have by assumption

$$Q_{X^n}[Q_{Y^n|X^n}[\mathcal{B}_{f(X^n)}]] \geq 1 - \epsilon. \tag{4.77}$$

Thus, for any $\epsilon' \in (\epsilon, 1)$, we obtain

$$Q_{X^n}[Q_{Y^n|X^n}[\mathcal{B}_{f(X^n)}] \geq 1 - \epsilon'] \geq 1 - \frac{\epsilon}{\epsilon'} \tag{4.78}$$

by applying Markov's inequality to $Q_{X^n}[1 - Q_{Y^n|X^n}[\mathcal{B}_{f(X^n)}] > \epsilon']$.

Fix for the time being any $\delta \in (0, 1 - \frac{\epsilon}{\epsilon'})$ such that $n > 3\beta_X \ln \frac{|\mathcal{X}|}{\delta}$, and define μ_n as in Corollary 4.7. Then

$$\mu_n[Q_{Y^n|X^n}[\mathcal{B}_{f(X^n)}] \geq 1 - \epsilon'] \geq 1 - \frac{\epsilon}{\epsilon'} - \delta. \tag{4.79}$$

As f takes at most M_1 possible values, there exists i^* such that

$$\mu_n[Q_{Y^n|X^n}[\mathcal{B}_{i^*}] \geq 1 - \epsilon'] \geq \frac{1 - \frac{\epsilon}{\epsilon'} - \delta}{M_1} \tag{4.80}$$

by the union bound. On the other hand, if we let ν be the uniform distribution on $\mathcal{Y}$, then we can estimate by the definition of $\mathcal{B}_i$

$$\nu^{\otimes n}[\mathcal{B}_i] = |\mathcal{Y}|^{-n} |\mathcal{B}_i| \leq M_2 |\mathcal{Y}|^{-n} \tag{4.81}$$

for every i. Applying Corollary 4.7 with $f = 1_{\mathcal{B}_{i^*}}$, $\eta = 1 - \epsilon'$ yields

$$\begin{aligned}
&\ln \frac{1 - \frac{\epsilon}{\epsilon'} - \delta}{M_1} - c \ln M_2 |\mathcal{Y}|^{-n} \\
&\leq \ln \mu_n[Q_{Y^n|X^n}[\mathcal{B}_{i^*}] \geq 1 - \epsilon'] - c \ln \nu^{\otimes n}[\mathcal{B}_{i^*}] && (4.82)\\
&\leq n \, \mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) + A\sqrt{n} + c \ln \frac{1}{1 - \epsilon'}, && (4.83)
\end{aligned}$$

where A is defined in (4.39). Rearranging yields

$$\begin{aligned}
\ln M_1 + c \ln M_2 \geq &- n\{\mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) - c \ln |\mathcal{Y}|\} \\
&- \sqrt{n}\left(\ln(|\mathcal{Y}|^c \beta_X^{c+1}) \sqrt{3\beta_X \ln \frac{4|\mathcal{X}|}{1 - \epsilon}} + 2c\sqrt{|\mathcal{Y}| \ln \frac{2}{1 - \epsilon}} \right) \\
&- c \ln \frac{2}{1 - \epsilon} - \ln \frac{4}{1 - \epsilon},
\end{aligned} \tag{4.84}$$

for $n > 3\beta_X \ln\frac{4|\mathcal{X}|}{1-\epsilon}$, where we made the choices $\epsilon' = \frac{1+\epsilon}{2}$ and $\delta = \frac{1}{2}(1-\frac{\epsilon}{\epsilon'})$.

It remains to manipulate the first-order terms. Note first that

$$\mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) - c\ln|\mathcal{Y}| = \sup_{U\colon U-X-Y}\{-cH(Y|U) - I(U;X)\} \tag{4.85}$$

$$= \mathrm{d}^\star(Q_X, Q_{Y|X}, Q_Y, c) - cH(Q_Y), \tag{4.86}$$

which follows readily from Definition 4.4 and (4.71). We may therefore follow verbatim the arguments in Step 3 of the proof of Theorem 4.11 to optimize (4.84) over $c \geq 1$, which readily completes the proof of (4.75). $\square$

We now give a analogue of Theorem 4.12 for Gaussian sources under quadratic distortion. We note again that the first-order term of (4.87) corresponds to the known rate region for this problem (e.g., let $D_2 \to \infty$ in [13, Theorem 12.3]).

Theorem 4.13. *Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, $Q_{XY} = \mathcal{N}(0, \left(\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right))$, and $D > 0$, $\epsilon \in (0,1)$, $n \geq 20\ln\frac{8}{1-\epsilon}$. Then for any encoders $f\colon \mathbb{R}^n \to [M_1]$, $g\colon \mathbb{R}^n \to [M_2]$, and decoder $\hat{Y}^n\colon [M_1]\times[M_2] \to \mathbb{R}^n$ such that $\mathbb{P}[\|Y^n - \hat{Y}^n\|^2 > nD] \leq \epsilon$, we have*

$$\ln M_2 \geq \frac{n}{2}\ln\frac{1-\rho^2+\rho^2 e^{-\frac{2}{n}\ln M_1}}{D} - 4\sqrt{\ln\frac{8}{1-\epsilon}}\sqrt{n} - 2\ln\frac{4}{1-\epsilon}. \tag{4.87}$$

Proof. Define for every $i \in [M_1]$ the set

$$\mathcal{B}_i := \{y^n \in \mathbb{R}^n\colon \|y^n - \hat{Y}^n(i, g(y^n))\|^2 \leq nD\} \tag{4.88}$$

of correctly decoded sequences y^n, given that the side information message i was received. Let $\epsilon' \in (\epsilon, 1)$, $\delta \in (0, 1-\frac{\epsilon}{\epsilon'})$ such that $n \geq 20\ln\frac{2}{\delta}$, define μ_n as in Corollary 4.10, and let ν be the Lebesgue measure on $\mathbb{R}$. Repeating verbatim the arguments in the proof of Theorem 4.12, we find that there exists i^* such that

$$\mu_n[Q_{Y^n|X^n}[\mathcal{B}_{i^*}] \geq 1-\epsilon'] \geq \frac{1-\frac{\epsilon}{\epsilon'}-\delta}{M_1}. \tag{4.89}$$

On the other hand, in the present setting, we estimate by the union bound

$$\nu^{\otimes n}[\mathcal{B}_i] \leq M_2 \mathrm{Vol}(\mathrm{Ball}(0,\sqrt{nD})) \leq M_2(2\pi e D)^{n/2}. \tag{4.90}$$

Applying Corollary 4.10 with $f = 1_{\mathcal{B}_{i^*}}$, $\eta = 1-\epsilon'$ and rearranging yields

$$\begin{aligned}\ln M_1 + c\ln M_2 \geq &-n\{\mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) + \tfrac{c}{2}\ln(2\pi e D)\}\\ &-\sqrt{n}\left(\sqrt{6\ln\frac{8}{1-\epsilon}} + c\sqrt{2\ln\frac{2}{1-\epsilon}}\right)\\ &-c\ln\frac{2}{1-\epsilon} - \ln\frac{4}{1-\epsilon}\end{aligned} \tag{4.91}$$

for $n \geq 20 \ln \frac{8}{1-\epsilon}$, where we made the choices $\epsilon' = \frac{1+\epsilon}{2}$ and $\delta = \frac{1}{2}(1 - \frac{\epsilon}{\epsilon'})$.

It remains to manipulate the first-order terms. To this end, we first note that using (B.43) in Appendix B.2 and the subsequent discussion, we may compute

$$\begin{aligned} &\mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) + \tfrac{c}{2}\ln(2\pi e D) \\ &= \sup_{\sigma\in[0,1]} \left\{ \frac{1}{2}\ln \sigma^2 - \frac{c}{2}\ln\frac{1-\rho^2+\rho^2\sigma^2}{D} \right\} \quad (4.92) \\ &= \frac{c}{2}\vartheta^*\left(-\frac{1}{c}\right), \quad (4.93) \end{aligned}$$

where ϑ^* is the convex conjugate of the convex function $\vartheta(x) := \ln \frac{1-\rho^2+\rho^2 e^{-x}}{D}$ for $x \geq 0$ and $\vartheta(x) := +\infty$ for $x < 0$. Therefore, by convex duality

$$\begin{aligned} &\inf_{c>0} \frac{1}{c}\left\{ \mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) + \tfrac{c}{2}\ln(2\pi e D) + \frac{1}{n}\ln M_1 \right\} \quad (4.94) \\ &= -\frac{1}{2}\vartheta\left(\frac{2}{n}\ln M_1\right) = -\frac{1}{2}\ln\frac{1-\rho^2+\rho^2 e^{-\frac{2}{n}\ln M_1}}{D}. \quad (4.95) \end{aligned}$$

We claim that the same conclusion follows if we take the infimum over $c \geq 1$ only. Indeed, it is readily verified that the function inside the supremum in (4.92) is increasing in σ^2 for any $c < 1$, so that $\mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) + \frac{c}{2}\ln(2\pi e D) = -\frac{c}{2}\ln\frac{1}{D}$ for $c < 1$. Thus, the infimum in (4.94) cannot be attained for $c < 1$. The proof is now readily completed using (4.95) by optimizing (4.91) over $c \geq 1$. $\square$

Acknowledgements

Suggestions on the presentation of the paper by Yury Polyanskiy are greatly appreciated. We thank Wei Yang for discussions on multiple access channels.

A Data processing cannot yield second-order converses

The aim of this appendix is to explain the claim made in Section 2.2 that second-order converses are fundamentally outside the reach of any method that is based on the data processing inequality: in particular, the inefficiency of the blowing-up method already appears in the very first step in the argument. This claim is little more than the Neyman-Pearson lemma in the following disguise.

Lemma A.1. *Let P, Q be probability measures on $\mathcal{Y}$ such that $\left\| \ln \frac{dP}{dQ} \right\|_\infty < \infty$. Then for every $\delta > 0$, there exists $C_\delta > 0$ independent of n so that any set $\tilde{\mathcal{A}} \subseteq \mathcal{Y}^n$ with $P^{\otimes n}[\tilde{\mathcal{A}}] \geq 1 - n^{-\delta}$ must satisfy $\ln Q^{\otimes n}[\tilde{\mathcal{A}}] \geq -nD(P\|Q) + C_\delta\sqrt{n \ln n}$.*

Let us first explain why this lemma rules out obtaining sharp bounds by data processing. Suppose we start our argument by applying the data processing inequality (2.3) to an arbitrary set $\tilde{\mathcal{A}}$. Then we obtain (cf. (2.7))

$$\ln Q^{\otimes n}[\tilde{\mathcal{A}}] \geq -n\frac{D(P\|Q)}{P^{\otimes n}[\tilde{\mathcal{A}}]} - \ln 2. \tag{A.1}$$

In order to deduce from this a non-asymptotic converse with $\sim \sqrt{n}$ second-order term, we must introduce some operation $\mathcal{A} \mapsto \tilde{\mathcal{A}}$ that associates to *any* deterministic test $\mathcal{A}$ with $P^{\otimes n}[\mathcal{A}] \geq 1 - \epsilon$ a set $\tilde{\mathcal{A}}$ with the following properties:

$$P^{\otimes n}[\tilde{\mathcal{A}}] \geq 1 - O(n^{-1/2}), \qquad Q^{\otimes n}[\tilde{\mathcal{A}}] \leq e^{O(\sqrt{n})}Q^{\otimes n}[\mathcal{A}]. \tag{A.2}$$

But Lemma A.1 shows such an operation cannot exist. Indeed, let $\mathcal{A}$ be a test with $P^{\otimes n}[\mathcal{A}] = 1-\epsilon$ and $\ln Q^{\otimes n}[\mathcal{A}] \leq -nD(P\|Q) + O(\sqrt{n})$ as in Lemma 2.5. Then by Lemma A.1, any set $\tilde{\mathcal{A}}$ such that $P^{\otimes n}[\tilde{\mathcal{A}}] \geq 1 - O(n^{-1/2})$ satisfies

$$Q^{\otimes n}[\tilde{\mathcal{A}}] \geq e^{-nD(P\|Q)+C\sqrt{n\ln n}} \geq e^{C\sqrt{n\ln n}+O(\sqrt{n})}Q^{\otimes n}[\mathcal{A}], \tag{A.3}$$

which contradicts the desired property of the mapping $\mathcal{A} \mapsto \tilde{\mathcal{A}}$. The same argument shows that converses obtained from the data processing inequality can attain at best a second-order term of order no smaller than $\sim \sqrt{n \log n}$.

We conclude this appendix with the proof of Lemma A.1.

Proof of Lemma A.1. By the Neyman-Pearson lemma, among all sets $\tilde{\mathcal{A}}$ such that $P^{\otimes n}[\tilde{\mathcal{A}}] \geq 1 - n^{-\delta}$, the probability $Q^{\otimes n}[\tilde{\mathcal{A}}]$ is minimized by

$$\tilde{\mathcal{A}} = \{y^n \in \mathcal{Y}^n : \imath_{P^{\otimes n}\|Q^{\otimes n}}(y^n) \geq nD(P\|Q) - \gamma\}, \tag{A.4}$$

where γ is chosen so that $P^{\otimes n}[\tilde{\mathcal{A}}] = 1 - n^{-\delta}$. It therefore suffices to prove a suitable lower bound on $Q^{\otimes n}[\tilde{\mathcal{A}}]$ for this particular choice of $\tilde{\mathcal{A}}$.

We will twice use the classical Kolmogorov lower bound for the tail of sums of independent random variables [40, p. 266]. First, this bound implies

$$P^{\otimes n}[\imath_{P^{\otimes n}\|Q^{\otimes n}} \geq nD(P\|Q) - \sqrt{\delta V(P\|Q)}\sqrt{n\ln n}] \leq 1 - n^{-\delta} \tag{A.5}$$

for n sufficiently large (depending on $\delta, V(P\|Q), \|\ln\frac{dP}{dQ}\|_\infty$), where $V(P\|Q) := \mathrm{Var}_P(\ln\frac{dP}{dQ})$. This shows that we must have $\gamma \geq \gamma_n := \sqrt{\delta V(P\|Q)}\sqrt{n\ln n}$.

On the other hand, we can now estimate

$$\begin{aligned} Q^{\otimes n}[\tilde{\mathcal{A}}] &= P^{\otimes n}\big(e^{-\imath_{P^{\otimes n}\|Q^{\otimes n}}}1_{\{\imath_{P^{\otimes n}\|Q^{\otimes n}} \geq nD(P\|Q)-\gamma\}}\big) && \text{(A.6)} \\ &\geq e^{-nD(P\|Q)+\sqrt{\delta V(P\|Q)/4}\sqrt{n\ln n}} \times \\ &\qquad P^{\otimes n}[nD(P\|Q) - \tfrac{1}{2}\gamma_n \geq \imath_{P^{\otimes n}\|Q^{\otimes n}} \geq nD(P\|Q) - \gamma]. && \text{(A.7)} \end{aligned}$$

But applying again (A.5), we find that

$$P^{\otimes n}[nD(P\|Q) - \tfrac{1}{2}\gamma_n \geq \imath_{P^{\otimes n}\|Q^{\otimes n}} \geq nD(P\|Q) - \gamma] \tag{A.8}$$

$$\geq 1 - n^{-\delta} - P^{\otimes n}[\imath_{P^{\otimes n}\|Q^{\otimes n}} \geq nD(P\|Q) - \tfrac{1}{2}\gamma_n] \geq n^{-\delta/4} - n^{-\delta}. \tag{A.9}$$

Consequently $\ln Q^{\otimes n}[\tilde{\mathcal{A}}] \geq -nD(P\|Q) + \sqrt{\delta V(P\|Q)/4}\sqrt{n \ln n} + O(\ln n)$. □

B Brascamp-Lieb divergence: auxiliary results

B.1 Proof of (4.18) in the discrete case

The aim of this section is to prove the following result.

Theorem B.1. *Let $|\mathcal{X}| < \infty$, Q_X be a probability measure on $\mathcal{X}$, ν be a probability measure on $\mathcal{Y}$, and $Q_{Y|X}$ be a random transformation. Define*

$$\beta_X := \frac{1}{\min_x Q_X(x)}, \tag{B.1}$$

$$\alpha_Y := \left\| \frac{dQ_Y}{d\nu} \right\|_\infty. \tag{B.2}$$

Then for every $\delta \in (0,1)$ and $n > 3\beta_X \ln \frac{|\mathcal{X}|}{\delta}$, we may choose a set $\mathcal{C}_n \subseteq \mathcal{X}^n$ with $Q_X^{\otimes n}[\mathcal{C}_n] \geq 1 - \delta$ such that

$$\begin{aligned} &\mathrm{d}(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) \\ &\qquad \leq n\,\mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) + \ln(\alpha_Y^c \beta_X^{c+1})\sqrt{3n\beta_X \ln \frac{|\mathcal{X}|}{\delta}} \end{aligned} \tag{B.3}$$

for every $c > 0$, where we defined $\mu_n := Q_X^{\otimes n}|_{\mathcal{C}_n}$.

Let us fix in the sequel $(Q_X, Q_{Y|X}, \nu)$ and $c > 0$, and define

$$\phi(P_X) := \mathrm{d}^\star(P_X, Q_{Y|X}, \nu, c) \tag{B.4}$$

for any $P_X \ll Q_X$. The idea behind the proof Theorem B.1 is roughly as follows. Using the chain rule of relative entropy, we will show that $\mathrm{d}(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c)$ can be bounded by a quantity of the form $n\phi(P_X)$, where P_X is a mixture of empirical measures of sequences in the support of μ_n. Thus, if we choose $\mathcal{C}_n$ to be the set of sequences with empirical measure close to Q_X, then $Q_X^{\otimes n}[\mathcal{C}_n]$ will be large by the law of large numbers and $\phi(P_X) \approx \phi(Q_X) = \mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c)$, completing the proof.

To make these ideas precise, we require quantitative forms of the continuity of ϕ and of the law of large numbers. The former is provided by the following lemma. Here we adopt the same notations as in Theorem B.1.

Lemma B.2. *If* $P_X \le (1+\epsilon)Q_X$ *for some* $\epsilon \in [0,1)$*, then*

$$\phi(P_X) \le \phi(Q_X) + \epsilon \ln(\beta_X^{c+1}\alpha_Y^c). \tag{B.5}$$

Proof. Let P_X be as in the statement, and let $P_{U|X}$ be a maximizer for (4.19) (we assume its existence for notational simplicity only; if a maximizer does not exist, the argument is readily adapted to work with near-maximizers). We now modify this distribution by allowing U to take an additional value as follows: let $\tilde{\mathcal{U}} = \mathcal{U} \cup \{\star\}$, $\tilde{\mathcal{X}} = \mathcal{X}$, and define $P_{\tilde{U}\tilde{X}}$ by

$$P_{\tilde{U}} := \frac{1}{1+\epsilon}P_U + \frac{\epsilon}{1+\epsilon}\delta_\star; \tag{B.6}$$

$$P_{\tilde{X}|\tilde{U}=u} := P_{X|U=u}, \quad \forall u \in \mathcal{U}; \tag{B.7}$$

$$P_{\tilde{X}|\tilde{U}=\star} := \frac{1+\epsilon}{\epsilon}\left(Q_X - \frac{1}{1+\epsilon}P_X\right) \tag{B.8}$$

where $\delta_\star$ is a point mass on $\star$. Observe that $P_{\tilde{X}} = Q_X$,

$$\begin{aligned} D(P_{\tilde{X}|\tilde{U}} \| Q_X | P_{\tilde{U}}) &= \frac{1}{1+\epsilon} D(P_{X|U}\|Q_X|P_U) \\ &\quad + \frac{\epsilon}{1+\epsilon} D\left(\frac{1+\epsilon}{\epsilon}Q_X - \frac{1}{\epsilon}P_X \middle\| Q_X\right) && \text{(B.9)} \\ &\le D(P_{X|U}\|Q_X|P_U) + \epsilon \ln \beta_X && \text{(B.10)} \end{aligned}$$

where we used $D(P\|Q_X) \le \ln \beta_X$ for every probability measure P, and

$$\begin{aligned} cD(P_{\tilde{Y}|\tilde{U}}\|\nu|P_{\tilde{U}}) &= \frac{c}{1+\epsilon} D(P_{Y|U}\|\nu|P_U) \\ &\quad + \frac{c\epsilon}{1+\epsilon} D\left(\frac{1+\epsilon}{\epsilon}Q_Y - \frac{1}{\epsilon}P_Y \middle\| \nu\right) && \text{(B.11)} \\ &\ge cD(P_{Y|U}\|\nu|P_U) - c\epsilon \ln(\beta_X\alpha_Y), && \text{(B.12)} \end{aligned}$$

where we used $\frac{1}{1+\epsilon} \ge 1-\epsilon$ and

$$\begin{aligned} D(P_{Y|U}\|\nu|P_U) &= D(P_{Y|U}\|Q_Y|P_U) + P_Y(\ln \tfrac{dQ_Y}{d\nu}) && \text{(B.13)} \\ &\le D(P_{X|U}\|Q_X|P_U) + P_Y(\ln \tfrac{dQ_Y}{d\nu}) && \text{(B.14)} \\ &\le \ln(\beta_X\alpha_Y), && \text{(B.15)} \end{aligned}$$

where (B.14) follows from the data processing inequality and (B.15) follows since $D(P_{X|U}\|Q_X|P_U) \le \ln \beta_X$. The proof is concluded by subtracting (B.10) from (B.12). ∎

As a quantitative form of the law of large numbers, we will use the following standard result (see, e.g., [6]).

Lemma B.3 (Chernoff Bound for Bernoulli variables)**.** *Assume that* $X_1, \ldots, X_n$ *are i.i.d.* $\mathrm{Ber}(p)$*. Then for any* $\epsilon \in (0, \infty)$

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i \geq (1+\epsilon)np\right] \leq e^{-\frac{1}{3}\min\{\epsilon^2, \epsilon\}np}. \tag{B.16}$$

We can now conclude the proof of Theorem B.1.

Proof of Theorem B.1. We denote by $\widehat{P}_{X^n}$ the empirical measure of $X^n \sim Q_X^{\otimes n}$. Let $n > 3\beta_X \ln \frac{|\mathcal{X}|}{\delta}$ and define

$$\epsilon_n := \sqrt{\frac{3\beta_X}{n} \ln \frac{|\mathcal{X}|}{\delta}} \in (0,1), \tag{B.17}$$

$$\mathcal{C}_n := \{x^n \colon \hat{P}_{x^n} \leq (1+\epsilon_n) Q_X\}. \tag{B.18}$$

As for each $x \in \mathcal{X}$

$$\mathbb{P}[\widehat{P}_{X^n}(x) > (1+\epsilon_n)Q_X(x)] \leq e^{-\frac{n}{3}Q_X(x)\epsilon_n^2} \leq \frac{\delta}{|\mathcal{X}|} \tag{B.19}$$

by Lemma B.3, it follows by the union bound that $Q_X^{\otimes n}[\mathcal{C}_n] \geq 1-\delta$.

Now consider any probability measure $P_{X^n} \ll \mu_n := Q_X^{\otimes n}|_{\mathcal{C}_n}$. Then we can estimate, following essentially [29, Lemma 9],

$$\begin{aligned}
&cD(P_{Y^n}\|\nu^{\otimes n}) - D(P_{X^n}\|\mu_n) \\
&= cD(P_{Y^n}\|\nu^{\otimes n}) - D(P_{X^n}\|Q_X^{\otimes n}) && \text{(B.20)}\\
&= c\sum_{i=1}^{n} D(P_{Y_i|Y^{i-1}}\|\nu|P_{Y^{i-1}}) - \sum_{i=1}^{n} D(P_{X_i|X^{i-1}}\|Q_X|P_{X^{i-1}}) && \text{(B.21)}\\
&\leq c\sum_{i=1}^{n} D(P_{Y_i|X^{i-1}}\|\nu|P_{X^{i-1}}) - \sum_{i=1}^{n} D(P_{X_i|X^{i-1}}\|Q_X|P_{X^{i-1}}) && \text{(B.22)}\\
&= n[c\, D(P_{Y_I|IX^{I-1}}\|\nu|P_{IX^{I-1}}) - D(P_{X_I|IX^{I-1}}\|Q_X|P_{IX^{I-1}})] && \text{(B.23)}\\
&\leq n\phi(P_{X_I}). && \text{(B.24)}
\end{aligned}$$

(B.20) follows from the fact that P_{X^n} is supported on $\mathcal{C}_n$; (B.21) is the chain rule of relative entropy; (B.22) follows from the convexity of relative entropy since $Y_i - X^{i-1} - Y^{i-1}$ under $P_{X^nY^n}$; in (B.23) we defined I to be a random variable uniformly distributed on $\{1, \ldots, n\}$ and independent of X^n, Y^n; and in (B.24) ϕ

is defined in (B.4). But note that $P_{X_I} = P_{X^n}(\hat{P}_{X^n}) \le (1+\epsilon_n)Q_X$ as P_{X^n} is supported on $\mathcal{C}_n$. Thus, (B.24), Lemma B.2 and (B.4) yield

$$\mathrm{d}(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) = \sup_{P_{X^n} \ll \mu_n} \{cD(P_{Y^n} \| \nu^{\otimes n}) - D(P_{X^n} \| \mu_n)\} \tag{B.25}$$

$$\le n\mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) + \ln(\beta_X^{c+1}\alpha_Y^c)n\epsilon_n, \tag{B.26}$$

and the proof is complete. □

B.2 Proof of (4.18) in the Gaussian case

We now prove the following Gaussian analogue of Theorem B.1.

Theorem B.4. *Let $Q_X = \mathcal{N}(0,\sigma^2)$, let $Q_{Y|X=x} = \mathcal{N}(x,1)$, and let ν be the Lebesgue measure on $\mathbb{R}$. Then for any $\delta \in (0,1)$ and $n \ge 20\ln\frac{2}{\delta}$, we may choose a set $\mathcal{C}_n \subseteq \mathbb{R}^n$ with $Q_X^{\otimes n}[\mathcal{C}_n] \ge 1-\delta$ such that*

$$\mathrm{d}(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) \le n\mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) + \sqrt{6n\ln\frac{2}{\delta}} \tag{B.27}$$

for any $c > 0$, where we defined $\mu_n := Q_X^{\otimes n}|_{\mathcal{C}_n}$.

In the Gaussian case, the choice of typical set $\mathcal{C}_n$ is rather simple: it is a spherical shell of radius $\sim \sqrt{n}$ and width $O(1)$. Controlling the probability of such a spherical shell is a standard exercise (cf. [6, p. 43]).

Lemma B.5 (Chi-square tail bound)**.** *For $X^n \sim \mathcal{N}(0, \mathbf{I}_n)$ and any $t > 0$*

$$\mathbb{P}[|\|X^n\|^2 - n| \ge 2\sqrt{nt} + 2t] \le 2e^{-t}. \tag{B.28}$$

We now turn to the proof of Theorem B.4.

Proof of Theorem B.4. Define

$$A := \sqrt{6\ln\frac{2}{\delta}}, \tag{B.29}$$

$$\mathcal{C}_n := \{x^n : n - A\sqrt{n} \le \tfrac{1}{\sigma^2}\|x^n\|^2 \le n + A\sqrt{n}\}. \tag{B.30}$$

Then we have $Q_X^n[\mathcal{C}_n] \ge 1-\delta$ for $n \ge 20\ln\frac{2}{\delta}$ by Lemma B.5.

We proceed similarly to the proof of Theorem B.1 (following essentially [29, Theorem 13]). Consider any probability measure $P_{X^n} \ll \mu_n := Q_X^{\otimes n}|_{\mathcal{C}_n}$. Then

$$cD(P_{Y^n}\|\nu^{\otimes n}) - D(P_{X^n}\|\mu_n)$$

$$= cD(P_{Y^n}\|\nu^{\otimes n}) - D(P_{X^n}\|Q_X^{\otimes n}) \tag{B.31}$$

$$= cD(P_{Y^n}\|\nu^{\otimes n}) - D(P_{X^n}\|\nu^{\otimes n}) + P_{X^n}(\ln\tfrac{dQ_X^{\otimes n}}{d\nu^{\otimes n}}) \tag{B.32}$$

$$\le h(P_{X^n}) - ch(P_{Y^n}) - nh(Q_X) + \tfrac{1}{2}A\sqrt{n}, \tag{B.33}$$

where in (B.31) we used that P_{X^n} is supported on $\mathcal{C}_n$; (B.32) follows from the definition of relative entropy; and (B.33) follows by substituting the explicit form of the Gaussian density $\frac{dQ_X}{d\nu}$ and using the definition of $\mathcal{C}_n$.

We now proceed to estimate using the chain rule

$$\begin{aligned}
&h(P_{X^n}) - ch(P_{Y^n}) \\
&= \sum_{i=1}^{n} \{h(X_i|X^{i-1}) - ch(Y_i|Y^{i-1})\} && \text{(B.34)} \\
&\leq \sum_{i=1}^{n} \{h(X_i|X^{i-1}) - ch(Y_i|X^{i-1})\} && \text{(B.35)} \\
&= n\{h(X_I|IX^{I-1}) - ch(Y_I|IX^{I-1})\} && \text{(B.36)} \\
&\leq nF((1+\tfrac{A}{\sqrt{n}})\sigma^2), && \text{(B.37)}
\end{aligned}$$

where in (B.35) we used concavity of differential entropy; in (B.36) we defined I to be uniformly distributed on $\{1,\ldots,n\}$ and independent of everything else; and in (B.37) we used that $\mathrm{Var}(X_I) \leq (1+\frac{A}{\sqrt{n}})\sigma^2$ by the definition of $\mathcal{C}_n$, with

$$F(M) := \sup_{P_{UX}\colon \mathrm{Var}(X)\leq M} \{h(X|U) - c\, h(Y|U)\} \tag{B.38}$$

(this quantity plays the role of $\phi(P_X)$ in the present setting). But note that by the scaling property of the differential entropy, we readily obtain

$$nF((1+\tfrac{A}{\sqrt{n}})\sigma^2) = nF(\sigma^2) + \tfrac{1-c}{2} n \ln(1+\tfrac{A}{\sqrt{n}}) \leq nF(\sigma^2) + \tfrac{1}{2}A\sqrt{n}. \tag{B.39}$$

Combining (B.33), (B.37), and (B.39), we have shown

$$\begin{aligned}
\mathrm{d}(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) &= \sup_{P_{X^n} \ll \mu_n} \{cD(P_{Y^n}\|\nu^{\otimes n}) - D(P_{X^n}\|\mu_n)\} && \text{(B.40)} \\
&\leq nF(\sigma^2) - nh(Q_X) + A\sqrt{n}. && \text{(B.41)}
\end{aligned}$$

To conclude the proof, it remains to show that

$$F(\sigma^2) - h(Q_X) = \mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c). \tag{B.42}$$

To this end, we note that arguing as in (B.32), we may write Definition 4.4 as

$$\mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) = \sup_{P_{UX}\colon P_X = Q_X} \{h(X|U) - c\, h(Y|U)\} - h(Q_X). \tag{B.43}$$

Thus, the inequality $\geq$ in (B.42) is immediate from the definition (B.38). For the converse direction we require the fact, proved in [28, Theorem 14], that for Gaussian channels the supremum in (B.38) is achieved for P_{UX} such that $U = 0$ and X is Gaussian. In particular, $F(\sigma^2) = h(X) - ch(Y)$ with $X \sim \mathcal{N}(0, a^2)$ for some $a \leq \sigma$. Thus, choosing P_{UX} in (B.43) so that $X = U + Z$ with $Z \sim \mathcal{N}(0, a^2)$, $U \sim \mathcal{N}(0, \sigma^2 - a^2)$, and U, Z independent yields $\leq$ in (B.42). □

References

[1] R. Ahlswede. Coloring hypergraphs: A new approach to multi-user source coding, part I. *Journal of Combinatorics, Information and System Sciences*, 4(1):76–115, 1979. Part II in pp. 220-268, 1980.

[2] R. Ahlswede and I. Csiszár. Hypothesis testing with communication constraints. *IEEE Trans. Inf. Theory*, 32(4):533–542, July 1986.

[3] R. Ahlswede, P. Gács, and J. Körner. Bounds on conditional probabilities with applications in multi-user communication. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 34(2):157–177, 1976. Correction (1977). ibid, 39(4), 353–354.

[4] S. G. Bobkov, I. Gentil, and M. Ledoux. Hypercontractivity of Hamilton-Jacobi equations. *J. Math. Pures Appl.*, 80(7):669–696, 2001.

[5] C. Borell. Positivity improving operators and hypercontractivity. *Mathematische Zeitschrift*, 180(3):225–234, 1982.

[6] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univesity Press, 2013.

[7] M. V. Burnashev. Data transmission over a discrete channel with feedback: Random transmission time. *Problemy Peredachi Informatsii*, 12, no. 4:10–30, 1976.

[8] P. Caputo, G. Menz, and P. Tetali. Approximate tensorization of entropy at high temperature. *Ann. Fac. Sci. Toulouse Math.*, 24:691–716, 2015.

[9] H. Chernoff. Large-sample theory: Parametric case. *Annals of Mathematical Statistics*, 27:1–22, 1956.

[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2nd edition, 2012.

[11] I. Csiszar and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2nd edition, 2011.

[12] M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, III. *Communications on Pure and Applied Mathematics*, 29, no. 4:369–461, 1977.

[13] A. El Gamal and Y.-H. Kim. *Network Information Theory*. Cambridge University Press, 2011.

[14] A. El Gamal and E. C. van der Meulen. A proof of Marton's coding theorem for the discrete memoryless broadcast channel. *IEEE Trans. Inf. Theory*, 27(1):120–122, Jan. 1981.

[15] R. M. Fano. Transmission of information. *Unpublished course notes, Massachusetts Institute of Technology, Cambridge, MA.*, 1952.

[16] S. L. Fong and V. Y. F. Tan. A proof of the strong converse theorem for Gaussian broadcast channels via the Gaussian Poincaré inequality. *IEEE Trans. Inf. Theory*, 63(12):7737–7746, Dec. 2017.

[17] D. Guo, S. Shamai, and S. Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Trans. Inf. Theory*, 51(4):1261–1282, 2005.

[18] T. S. Han. *Information-Spectrum Methods in Information Theory*. Springer, 2003.

[19] T. S. Han and S. Verdú. Approximation theory of output statistics. *IEEE Trans. Inf. Theory*, 39(3):752–772, May 1993.

[20] M. Hayashi. Information spectrum approach to second-order coding rate in channel coding. *IEEE Trans. Inf. Theory*, 55:4947–4966, Nov 2009.

[21] B. Kelly. *Information Theory With Large Alphabets And Source Coding Error Exponents*. PhD thesis, Cornell University, 2011.

[22] I. Kontoyiannis and S. Verdú. Optimal lossless data compression: Non-asymptotics and asymptotics. *IEEE Trans. Inf. Theory*, 60, no. 2:777–795, Feb. 2014.

[23] V. Kostina and S. Verdú. Fixed-length lossy compression in the finite blocklength regime. *IEEE Trans. Inf. Theory*, 58(6):3309–3338, June 2012.

[24] M. Ledoux. Isoperimetry and Gaussian analysis. In *Lectures on Probability Theory and Statistics*, pages 165–294. Springer, 1996.

[25] M. Ledoux. *The Concentration of Measure Phenomenon*. AMS, 2001.

[26] J. Liu. *Information Theory from a Functional Viewpoint*. PhD thesis, Princeton University, Princeton, NJ, 2017.

[27] J. Liu. Dispersion bound for the Wyner-Ahlswede-Körner network via reverse hypercontractivity on types. In *Proceedings of 2018 IEEE International Symposium on Information Theory*, pages 1854–1858, Vail, Colorado, June 2018.

[28] J. Liu, T. A. Courtade, P. Cuff, and S. Verdú. Information theoretic perspectives on Brascamp-Lieb inequality and its reverse. *arXiv:1702.06260*, 2015.

[29] J. Liu, T. A. Courtade, P. Cuff, and S. Verdú. Smoothing Brascamp-Lieb inequalities and strong converses for common randomness generation. In *Proc. of IEEE International Symposium on Information Theory*, pages 1043–1047, July 2016, Barcelona, Spain.

[30] J. Liu, T. A. Courtade, P. W. Cuff, and S. Verdú. Smoothing Brascamp-Lieb inequalities and strong converses of coding theorems. *IEEE Trans. Inf. Theory (to appear)*, 2019.

[31] J. Liu, P. Cuff, and S. Verdú. On α-decodability and α-likelihood decoder. In *55th Annual Allerton Conference on Communication, Control, and Computing*, 2017.

[32] J. Liu, R. van Handel, and S. Verdú. Beyond the blowing-up lemma: Sharp converses via reverse hypercontractivity. In *Proceedings of 2017 IEEE International Symposium on Information Theory*, Aachen, Germany, June 2017.

[33] G. A. Margulis. Probabilistic characteristics of graphs with large connectivity. *Problems of Information Transmission*, 10(2):101–108, 1974.

[34] K. Marton. Logarithmic Sobolev inequalities in discrete product spaces. *Combin. Probab. Comput.*, 28:919–935, 2019.

[35] K. Marton. A simple proof of the blowing-up lemma. *IEEE Trans. Inf. Theory*, 24(3):445–446, May 1986.

[36] E. Mossel, R. O. Donnell, O. Regev, J. E. Steif, and B. Sudakov. Non-interactive correlation distillation, inhomogeneous Markov chains, and the reverse Bonami-Beckner inequality. *Israel Journal of Mathematics*, 154(1):299–336, 2006.

[37] E. Mossel, K. Oleszkiewicz, and A. Sen. On reverse hypercontractivity. *Geometric and Functional Analysis*, 23(3):1062–1097, 2013.

[38] Y. Oohama. Strong converse exponent for degraded broadcast channels at rates outside the capacity region. In *Proc. of IEEE International Symposium on Information Theory*, pages 939–943, June 2015, Hong Kong, China.

[39] A. Orlitsky, N. Santhanam, and J. Zhang. Relative redundancy for large alphabets. In *Proc. of IEEE International Symposium on Information Theory*, pages 2672–2676, Seattle, WA, July 2006.

[40] D. Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, 2002.

[41] Y. Polyanskiy. A perspective on massive random-access. In *Proc. 2017 IEEE International Symposium on Information Theory*, Aachen, Germany, July 2017.

[42] Y. Polyanskiy, H. V. Poor, and S. Verdú. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory*, 56(5):2307–2359, May 2010.

[43] Y. Polyanskiy and S. Verdú. Empirical distribution of good channel codes with nonvanishing error probability. *IEEE Trans. Inf. Theory*, 60(1):5–21, Jan. 2014.

[44] Y. Polyanskiy and S. Verdú. Arimoto channel coding converse and Rényi divergence. In *Proc. 48th Annual Allerton Conference on Communication, Control, and Computing*, pages 1327–1333, Monticello, Illinois, Oct. 2010.

[45] M. Raginsky and I. Sason. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends in Communications and Information Theory*, 10(1-2):1–250, 2014.

[46] S. Shamai and S. Verdú. The empirical distribution of good codes. *IEEE Transactions on Information Theory*, 43, no. 3:836–846, May 1997.

[47] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27, no. 3, 4:379–423, 623–656, July-Oct. 1948.

[48] C. E. Shannon. Certain results in coding theory for noisy channels. *Information and Control*, 1:6–25, Sep. 1957.

[49] C. E. Shannon, R. G. Gallager, and E. Berlekamp. Lower bounds to error probability for coding on discrete memoryless channels, I. *Information and Control*, 10(1):65–103, 1967.

[50] V. Strassen. Asymptotische abschäzungen in Shannons informationstheorie. In *Transactions of the Third Prague Conference on Information Theory, on Statistics, Decision Functions, Random Processes, (Liblice, 1962)*, pages 689–723. Publishing House of the Czechoslovak Academy of Sciences, Prague, Liblice, Czechoslovaquia, 1964.

[51] V. Y. F. Tan. Asymptotic Estimates in Information Theory with Non-Vanishing Error Probabilities. *Foundations and Trends in Communications and Information Theory*, 11(1-2):1–184, 2014.

[52] S. Verdú. Non-asymptotic achievability bounds in multiuser information theory. In *Proc. 50th Annual Allerton Conference on Communication, Control and Computing*, pages 1–8, Monticello, Illinois, Oct. 2012.

[53] S. Verdú and T. S. Han. A general formula for channel capacity. *IEEE Trans. Inf. Theory*, 40:1147–1157, July 1994.

[54] S. Watanabe, S. Kuzuoka, and V. Tan. Nonasymptotic and second-order achievability bounds for coding with side-information. *IEEE Trans. Inf. Theory*, 61:1574–1605, April 2015.

[55] J. Wolfowitz. *Coding Theorems of Information Theory*. Springer, Berlin, Germany, 1 edition, 1961.

[56] J. Wolfowitz. Notes on a general strong converse. *Information and Control*, 12:1–4, 1968.

[57] J. Wolfowitz. Strong converse of the coding theorem for semicontinuous channels. *Illinois J. Math.*, 3, no. 4:477–489, Dec. 1959.

[58] A. D. Wyner. On source coding with side information at the decoder. *IEEE Trans. Inf. Theory*, 21(3):294–300, May 1975.

[59] M. H. Yassaee, M. R. Aref, and A. Gohari. A technique for deriving one-shot achievability results in network information theory. In *Proceedings of 2013 IEEE International Symposium on Information Theory*, pages 1287–1291, Istanbul, Turkey, July 2013.