An information diffusion Fano inequality

Gábor Braun¹ and Sebastian Pokutta¹

¹ISyE, Georgia Institute of Technology, Atlanta, GA 30332, USA.
Email: {gabor.braun,sebastian.pokutta}@isye.gatech.edu

April 22, 2015

Abstract

In this note, we present an information diffusion inequality derived from an elementary argument, which gives rise to a very general Fano-type inequality. The latter unifies and generalizes the distance-based Fano inequality and the continuous Fano inequality established in [DW13, Corollary 1, Propositions 1 and 2], as well as the generalized Fano inequality in [HV94, Equation following (10)].

1 Introduction

Fano inequality is a crucial tool in information theory with numerous applications. Moreover, it has been heavily used in statistics in the context of minimax theory (see [LC98] and references contained therein) and more recently also in optimization (see e.g., [RR09, ABRW12, BGP13]) to lower bound the rate of convergence of estimators and algorithms. The general setup of Fano inequalities is a Markov chain \(X \rightarrow Y \rightarrow \hat{X}\) and we are interested in the probability of finding a sufficient reconstruction \(\hat{X}\) of the hidden random variable \(X\) by observations \(Y\). Classically the measure of sufficiency has been equality, i.e., we ask for perfect reconstructions \(\hat{X} = X\). This can be relaxed in several ways, by e.g., accepting reconstructions \(\hat{X}\), whenever \(\hat{X}\) is close to \(X\).

In this note we present an elementary information diffusion inequality, which immediately gives rise to a very general Fano inequality, extending and subsuming the versions presented in [DW13]. In particular, we allow for arbitrary relations \(R \subseteq \text{range}(X) \times \text{range}(\hat{X})\) indicating a sufficient reconstruction.

Our notation is standard as to be found in [CT06], and consistent with [DW13]. We denote random variables by capital bold letters such as, e.g., \(X\) and events by scripts letters, such as \(\mathcal{R}\). Let \(\neg \mathcal{R}\) denote the negation of the event \(\mathcal{R}\).

Let \(\log\) be a logarithm with an arbitrary basis \(a > 1\), which also serves as a basis for measuring information, i.e., all information quantities are defined using base \(a\) logarithm \(\log\). Recall that the Rényi divergence of two distributions \(P\) and \(Q\) over the same probability space is defined as

\[
D_\alpha (P \parallel Q) := \log \mathbb{E}_P \left[ \left( \frac{dQ}{dP} \right)^{1-\alpha} \right]
\]

for an order \(0 < \alpha < \infty\) with \(\alpha \neq 1\). By continuity, this extends to orders 0, 1 and \(\infty\). For the order \(\alpha = 1\) one recovers relative entropy, also known as Kullback–Leibler divergence:

\[
D_1 (P \parallel Q) = D (P \parallel Q) := \mathbb{E}_P \left[ \log \left( \frac{dP}{dQ} \right) \right].
\]
When $P$ and $Q$ are Bernoulli distributions with parameters $p$ and $q$ respectively, we obtain the binary versions

$$d_\alpha (p \parallel q) := \frac{\log \left( p^\alpha q^{1-\alpha} + (1-p)^\alpha (1-q)^{1-\alpha} \right)}{\alpha - 1},$$

$$d (p \parallel q) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$  

The binary Rényi entropy and binary entropy is defined as

$$\tilde{H}_\alpha [p] := \frac{\log \left( p^\alpha + (1-p)^\alpha \right)}{1-\alpha},$$

$$\tilde{H} [p] := p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}.$$  

## 2 Information diffusion Fano inequality

In this section we will present a general information diffusion inequality, applicable to a broad range of distributions, including continuous ones. We allow for specification of an arbitrary reconstruction relation $R \subseteq \text{range}(X) \times \text{range}(\hat{X})$, where $X$ is a random variable and $\hat{X}$ its reconstruction. We might want to think of $R$ as specifying the acceptable reconstructions, e.g., those with small $\ell_1$-error.

Our general Fano inequality is inspired by a simple support-based lower bound on relative entropy, see e.g., [vEH14, Theorem 3]: For any two probability distributions $P, Q$ on the same probability space, and denoting in the support $\text{supp} P$ of $P$:

$$D (P \parallel Q) \geq \log \frac{1}{\mathbb{P}_Q[\text{supp} P]}.$$  

The next inequality is an extension of the generalized Fano inequalities in [DW13, Corollary 1 and Proposition 2], where we do not consider the distance between $P_{XY}$ and $P_X \times P_Y$ but rather between two arbitrary distributions $P_{XY}$ and $Q_{XY}$.

**Proposition 2.1** (Information diffusion Fano inequality). Let $P$ and $Q$ be two probability distributions on the same probability space and $R$ an event. Further, choose $0 \leq p_{\min} < 1$ and $0 < p_{\max} \leq 1$ with $p_{\min} + p_{\max} < 1$ to be numbers satisfying

$$p_{\min} \leq \mathbb{P}_Q [R] \leq p_{\max}. \quad (1)$$

Then for any order $0 < \alpha < \infty$ with $\alpha \neq 1$:

$$\mathbb{P}_P [R] \leq \sqrt{\frac{\exp \left( \left( D_\alpha (P \parallel Q) + \tilde{H}_\alpha [\mathbb{P}_P [R]] + \log (1-p_{\min}) \right) \frac{\alpha-1}{\log \epsilon} \right)-1}{\left( \frac{1-p_{\min}}{p_{\max}} \right)^{\alpha-1}-1}}. \quad (2)$$

For the order $\alpha = 1$, the following version holds:

$$\mathbb{P}_P [R] \leq \frac{D (P \parallel Q) + \tilde{H} [\mathbb{P}_P [R]] + \log (1-p_{\min})}{\log \frac{1-p_{\min}}{p_{\max}}}. \quad (3)$$
Proof. The proof is an easy application of the data processing equality. We shall also use the inequality
\[ x^\alpha + y^\alpha \begin{cases} \geq (x + y)^\alpha & \text{if } \alpha \leq 1, \\ \leq (x + y)^\alpha & \text{if } \alpha \geq 1, \end{cases} \] with the choice \( x := P_R \) and \( y := 1 - P_R \):
\[
P_R^\alpha + (1 - P_R)\alpha \begin{cases} \geq 1 & \text{if } \alpha \leq 1, \\ \leq 1 & \text{if } \alpha \geq 1. \end{cases} \quad (4)
\]
One should verify the inequalities below separately for \( \alpha < 1 \) and \( \alpha > 1 \).

\[
\begin{align*}
D_\alpha (P \parallel Q) + \widetilde{H}_\alpha [P_R [R]] & \geq d_\alpha (P_P [P] \parallel P_Q [R]) + \widetilde{H}_\alpha [P_P [R]] & \text{(data processing)} \\
& = \log \left( \frac{P_P [R]^{\alpha} + (1 - P_P [R])^{\alpha}}{P_P [R]^\alpha + (1 - P_P [R])^\alpha} \right) - \frac{\alpha - 1}{\alpha} \log (1 - p_{\min}) - \log(1 - p_{\min}). & \text{(by Eq. (1))}
\end{align*}
\]

The claim follows by rearranging. For the case \( \alpha = 1 \) we provide two proofs: (1) by taking limit when \( \alpha \to 1 \), and (2) via a similar direct argument. To simplify the limit argument, let us introduce some shorthand notation:

\[
A_\alpha := D_\alpha (P \parallel Q) + \widetilde{H}_\alpha [P_P [R]] + \log(1 - p_{\min}),
\]

\[
B := \frac{1 - p_{\min}}{p_{\max}}.
\]

Recall that \( \lim_{\alpha \to 1} D_\alpha (P \parallel Q) = D (P \parallel Q) \), therefore \( A_1 \) is the numerator of (3). The limit of the right-hand side of (2) as \( \alpha \to 1 \)

\[
\lim_{\alpha \to 1} \sqrt[\alpha]{\frac{\exp \left[ \frac{A_\alpha}{\log e} \right] - 1}{B^{\alpha - 1} - 1}} = \lim_{\alpha \to 1} \sqrt[\alpha]{\frac{\exp \left( \frac{A_\alpha}{\log e} \right) - 1}{B^{\alpha - 1} - 1}} \cdot \frac{A_\alpha}{\log e} = \frac{1}{\log B} \cdot \frac{A_1}{\log e} = A_1 \sqrt[\alpha]{\log B},
\]

which is exactly the right-hand side of Eq. (3).
An alternate proof via a direct computation goes as follows, similar to the proof of Eq. (2):

\[
D(P \parallel Q) + \mathbb{H}[P_R] \geq d\left(\frac{P_P[R]}{P_Q[R]} \parallel P_Q[R]\right) + \mathbb{H}[P_P[R]] \\
= P_P[R] \log \frac{P_P[R]}{P_Q[R]} + (1 - P_P[R]) \log \frac{1 - P_P[R]}{1 - P_Q[R]} \\
+ P_P[R] \log \frac{1}{P_Q[R]} + (1 - P_P[R]) \log \frac{1}{1 - P_Q[R]} \\
= P_P[R] \log \frac{1}{p_{\max}} + (1 - P_P[R]) \log \frac{1}{1 - p_{\min}}.
\]

(by (1))

Rearranging finishes the proof.

We obtain a very general version of Fano’s inequality as a consequence. This general form does not require any specific distributional assumptions on \(X\) such as e.g., uniformity. The case \(p_{\min} = 0\) is \([HV94, \text{Equation following (10)}]\).

**Proposition 2.2** (Fano inequality for arbitrary relations). Let \(X \rightarrow Y \rightarrow \hat{X}\) be a Markov chain of random variables and let \(R\) be any set of values \((x, \hat{x})\) with \(x \in \text{range}(X)\) and \(\hat{x} \in \text{range}(\hat{X})\). Further, choose \(0 \leq p_{\min} < 1\) and \(0 < p_{\max} \leq 1\) with \(p_{\min} + p_{\max} < 1\) to be numbers satisfying

\[
p_{\min} \leq \inf_{\hat{x}} \mathbb{P}[(X, \hat{x}) \in R] \quad \text{and} \quad p_{\max} \geq \sup_{\hat{x}} \mathbb{P}[(X, \hat{x}) \in R].
\]

Let \(R\) denote the event \((X, \hat{X}) \in R\). Then

\[
\mathbb{P}[R] \leq \frac{\mathbb{I}[X; \hat{X}] + \mathbb{H}[P[R]] + \log(1 - p_{\min})}{\log \frac{1 - p_{\max}}{p_{\max}}} \leq \frac{\mathbb{I}[X; Y] + \mathbb{H}[P[R]] + \log(1 - p_{\min})}{\log \frac{1 - p_{\min}}{p_{\max}}},
\]

(5)

**Proof.** The second inequality is equivalent to the data processing inequality \(\mathbb{I}[X; \hat{X}] \leq \mathbb{I}[X; Y]\). The first inequality is the following special case of Proposition 2.1. We choose \(P\) to be the joint distribution of \((X, \hat{X})\), which is the distribution used in the statement, i.e., \(\mathbb{P}[R] = P_P[R]\). We choose \(Q\) to be the product of the marginal distributions of \(X\) and \(\hat{X}\), therefore \(D(P \parallel Q) = D(P_{X\hat{X}} \parallel Q_{X\hat{X}})\).

Finally,

\[
\mathbb{P}_Q[R] = \mathbb{P}_Q[(X, \hat{X}) \in R] = \mathbb{E}_{\hat{X} \sim \hat{X}} \mathbb{P}[(X, \hat{x}) \in R] \geq \inf_{\hat{x}} \mathbb{P}[(X, \hat{x}) \in R] \geq p_{\min},
\]

and similarly, \(\mathbb{P}_Q[R] \leq p_{\max}\). Therefore the conditions of Proposition 2.1 are satisfied, and its conclusion provides the first inequality in (5).

We immediately obtain the following corollary by rearranging (5). The condition \(p_{\min} + p_{\max} < 1\) is no longer needed, as it was only used to preserve the direction of inequality while dividing by \(\log[(1 - p_{\min})/p_{\max}]\). This step can be omitted by a direct proof, consisting of repeating the last computation in the proof of Proposition 2.1 and then rearranging.
Corollary 2.3 (Entropy version of Fano inequality). Let $X \to Y \to \hat{X}$ be a Markov chain of random variables and let $R$ be any set of values $(x, \hat{x})$ with $x \in \text{range}(X)$ and $\hat{x} \in \text{range}(\hat{X})$. With notation from Proposition 2.2 we have

$$\mathbb{H} \left[ X \big| \hat{X} \right] \leq \mathbb{H} \left[ X \right] + \log p_{\max} + \mathbb{H} \left[ \mathbb{P} \left[ \neg R \right] \right] + \mathbb{P} \left[ \neg R \right] \log \frac{1 - p_{\min}}{p_{\max}}.$$

Moreover, if $Y = (Y_1, \ldots, Y_n)$ is obtained via independent sampling from a hidden distribution specified by $X$, i.e., the $Y_1, \ldots, Y_n \mid X$ are i.i.d, then we obtain the following corollary, which is sufficient for many applications. The version with the relative entropy is obtained as a direct consequence of the convexity of the relative entropy.

Corollary 2.4 (Fano inequality for independent samples). Let $X \to Y \to \hat{X}$ be a Markov chain of random variables with $Y = (Y_1, \ldots, Y_n)$, so that $Y_1, \ldots, Y_n \mid X$ are i.i.d. Further, let $R$ be any set of values $(x, \hat{x})$ with $x \in \text{range}(X)$ and $\hat{x} \in \text{range}(\hat{X})$. With notation from Proposition 2.2 we have

$$\mathbb{P} \left[ \neg R \right] \leq \frac{n \cdot \mathbb{I} [X; Y_1] + \mathbb{H} \left[ \mathbb{P} \left[ R \right] \right] + \log(1 - p_{\min})}{\log \frac{1 - p_{\min}}{p_{\max}}} \leq \frac{n \cdot \beta + \mathbb{H} \left[ \mathbb{P} \left[ R \right] \right] + \log(1 - p_{\min})}{\log \frac{1 - p_{\min}}{p_{\max}}},$$

(6)

where $\beta = \max_{x, x' \in \text{range}(X)} D \left( Y_1 \mid X = x || Y_1 \mid X = x' \right)$.

2.1 Special cases

We will now show how to obtain [DW13] Corollary 1, Propositions 1 and 2] as special cases of the general Fano inequality from above by choosing the relation $R$ accordingly.

Distance-based Fano inequality

For the distance-based case, let $\rho : \text{range}(X) \times \text{range}(X) \to \mathbb{R}$ be a symmetric function—typically a metric. Let $X$ be a discrete random variable with $2 \leq |\text{range}(X)| \leq \infty$. Furthermore, let $\hat{X}$ denote the reconstruction and assume $\text{range}(\hat{X}) = \text{range}(X)$. For a given radius $t$ denote $P_t := \mathbb{P} \left[ \rho(X, \hat{X}) > t \right]$. We then obtain as corollary, in the case where $X$ is uniform:

Corollary 2.5. (Distance-based Fano inequality [DW13] Proposition 1]) Let $X \to Y \to \hat{X}$ be a Markov chain of random variables with $X$ uniform. For a given radius $t \geq 0$ define

$$N_{t}^{\max} := \max_{x} \left| \{ \hat{x} \mid \rho(x, \hat{x}) \leq t \} \right| \quad \text{and} \quad N_{t}^{\min} := \min_{x} \left| \{ \hat{x} \mid \rho(x, \hat{x}) \leq t \} \right|,$$

then

$$\mathbb{H} [P_t] + P_t \log \frac{|\text{range}(X)| - N_{t}^{\min}}{N_{t}^{\max}} + \log N_{t}^{\max} \geq \mathbb{H} \left[ X \big| \hat{X} \right].$$

Proof. We pick $R := \{(x, \hat{x}) \in \text{range}(X) \times \text{range}(X) \mid \rho(x, \hat{x}) \leq t\}$, so that $\mathbb{P} \left[ \neg R \right] = P_t$, and choose $p_{\min} := \frac{N_{t}^{\min}}{|\text{range}(X)|}$ and $p_{\max} := \frac{N_{t}^{\max}}{|\text{range}(X)|}$. By Corollary 2.3 using $\mathbb{H} \left[ X \right] \leq \log |\text{range}(X)|$

$$\mathbb{H} \left[ X \big| \hat{X} \right] \leq \mathbb{H} \left[ X \right] + \log \frac{N_{t}^{\max}}{|\text{range}(X)|} + \mathbb{H} [P_t] + P_t \log \frac{1 - \frac{N_{t}^{\min}}{|\text{range}(X)|}}{N_{t}^{\max}}$$

$$\leq \log |\text{range}(X)| + \log \frac{N_{t}^{\max}}{|\text{range}(X)|} + \mathbb{H} [P_t] + P_t \log \frac{|\text{range}(X)| - N_{t}^{\min}}{N_{t}^{\max}}$$

$$= \log N_{t}^{\max} + \mathbb{H} [P_t] + P_t \log \frac{|\text{range}(X)| - N_{t}^{\min}}{N_{t}^{\max}},$$
as claimed.

Note that we require $X$ to be uniform in Corollary 2.5 to easily match the form of [DW13, Proposition 1]. However, the uniformity requirement can be removed. With the same choice for $R$, we also immediately obtain [DW13, Corollary 1], either by following the approach in [DW13] or by directly invoking Proposition 2.2.

**Corollary 2.6.** (Mutual information version of distance-based Fano inequality [DW13, Proposition 2]) With the notation of Corollary 2.5, let $X \rightarrow Y \rightarrow \hat{X}$ be a Markov chain of random variables with $X$ uniform. For any radius $t \geq 0$ we have

$$P_t \geq 1 - \frac{I[X; Y] + \tilde{H}(P_t)}{\log \frac{|\text{range}(X)|}{N_{\max}}}.$$  

**Continuous Fano inequality**

In a next step, we will show how to obtain the continuous Fano inequality of [DW13], avoiding the discretization argument altogether. Our version is slightly more general.

Let $X$ be a continuous random variable so that that $\text{range}(X)$ has finite non-zero Lebesgue measure. Moreover, let $\text{range}(\hat{X}) = \text{range}(X)$ as in the discrete distance-based setup. With the notation from above, we define $B_\rho(t, x) := \{ \hat{x} \in \text{range}(X) \mid \rho(x, \hat{x}) \leq t \}$. We obtain

**Corollary 2.7.** (Continuous Fano inequality [DW13, Proposition 2]) Let $X \rightarrow Y \rightarrow \hat{X}$ be a Markov chain of random variables with $X$ uniform. For a given radius $t \geq 0$ we have

$$P_t \geq 1 - \frac{I[X; Y] + \log 2}{\log \frac{\text{vol}(\text{range}(X))}{\sup_x \text{vol}(B_\rho(t, x) \cap \text{range}(X))}}.$$  

**Proof.** As before, we choose $R := \{ (x, \hat{x}) \in \text{range}(X) \times \text{range}(X) \mid \rho(x, \hat{x}) \leq t \}$, so that $P[R] = 1 - P_t$. We apply Proposition 2.2 with the choice $p_{\min} = 0$ and $p_{\max} = \frac{\sup_x \text{vol}(B_\rho(t, x) \cap \text{range}(X))}{\text{vol}(\text{range}(X))}$ and obtain

$$1 - P_t \leq \frac{I[X; Y] + \tilde{H}(P_t)}{\log \frac{\text{vol}(\text{range}(X))}{\sup_x \text{vol}(B_\rho(t, x) \cap \text{range}(X))}},$$

which is the claim rearranged. □

**Acknowledgements**

Research reported in this paper was partially supported by NSF grant CMMI-1300144 and CCF-1415496.

**References**

[ABRW12] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *Information Theory, IEEE Transactions on*, 58(5):3235–3249, 2012.

[BGP13] G. Braun, C. Guzmán, and S. Pokutta. Unifying lower bounds on the oracle complexity of nonsmooth convex optimization. *submitted*, 2013.
[CT06] T. Cover and J. Thomas. *Elements of information theory*. Wiley-interscience, 2006.

[DW13] J. C. Duchi and M. J. Wainwright. Distance-based and continuum Fano inequalities with applications to statistical estimation. *arXiv:1311.2669v2*, 2013.

[HV94] T. S. Han and S. Verdú. Generalizing the Fano inequality. *IEEE Transactions on Information Theory*, 40(4):1247–1251, July 1994.

[LC98] E. L. Lehmann and G. Casella. *Theory of point estimation*, volume 31. Springer, 1998.

[RR09] M. Raginsky and A. Rakhlin. Information complexity of black-box convex optimization: A new look via feedback information theory. In *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, pages 803–510. IEEE, 2009.

[vEH14] T. van Erven and P. Harremoës. Rényi divergence and Kullback–Leibler divergence. *arXiv:1206.2459v2 [cs.IT]*, April 2014.