Abstract—One-shot analogues for various information theory results known in the asymptotic case are proven using smooth min and max Rényi divergences. In particular, we prove that smooth min Rényi divergence can be used to prove one-shot analogue of the Stein's lemma. Using smooth min Rényi divergence we prove a special case of packing lemma in the one-shot setting. Furthermore, we prove a one-shot analogue of covering lemma using smooth max Rényi divergence. We also propose one-shot achievable rate for source coding under maximum distortion criterion. This achievable rate is quantified in terms of smooth max Rényi divergence.

I. INTRODUCTION

Almost all of the results in information theory are derived under the assumption that a random experiment is repeated independently and identically for large enough time. These assumptions have helped to give operational meaning to information theoretic quantities such as the Shannon entropy [1] and mutual information [2]. However, both of these assumptions are immoderate and difficult to justify in most of the practical scenarios. To overcome the limitations posed by such assumptions there has been a considerable interest during the past few years in one-shot or non-asymptotic information theory. One-shot or non-asymptotic information theory relies on the fact that a random experiment is performed only once or finite number of times where these finite repetitions of the experiment can be arbitrarily distributed. Thus removing both the above mentioned assumptions together.

In the literature there has been two kind of approaches to tackle the problems of non-asymptotic information theory. The first approach is due to Renner et al. [3] and the second approach is due to Polyanskiy et al. [4]. The first approach uses the notions of smooth Rényi entropies and divergences to give one-shot bounds for various information theoretic problems which are asymptotically optimal both in the i.i.d. and non-i.i.d. settings, see e.g., Refs. [3], [5], [6], [7], [8].


The second approach uses the notion of channel and rate dispersion to give non-asymptotic bounds for various information theory problems. Ref. [4] carried forward the approach of [11] and defined the notion of channel dispersion to give non-asymptotic bounds for the channel coding problem. This approach was further carried forward to give non-asymptotic bounds for various other information theoretic problems see e.g., Refs. [12], [13], [14], [15], [16], [17].

In this work we carry forward the idea of using smooth Rényi divergences to give one-shot bounds for some information theoretic problems. We first define the smooth min Rényi divergence in a slightly different way than that defined in [6]. We show that this definition of smooth min Rényi divergence helps to prove one-shot equivalent of the Stein’s lemma [2]. This one-shot equivalent of Stein’s lemma gives an operational meaning to smooth min Rényi divergence. Using the smooth min Rényi divergence we prove a special case of the packing lemma [18] in one-shot setting. We also prove a one-shot generalization of covering lemma [18] using smooth max Rényi divergence. The notion of smooth max Rényi divergence was defined in the quantum case in Ref. [19]. We show that smooth max Rényi divergence satisfies the data processing inequality. Furthermore, we define one-shot ε-maximum distortion criterion between two random variables. This definition is motivated by the definition of maximum distortion criterion mentioned in [20]. We then give a one-shot achievable rate under this distortion criterion.

The rest of this paper is organized as follows. In Section II we discuss the notations which we will be using through out this paper. In Section III we first mention the definition of smooth min Rényi divergence as given in Ref. [6]. We then give an equivalent definition for this divergence. This equivalent definition helps us in proving one-shot bounds for some important information theoretic problems. We prove a lemma pertaining to the asymptotic behavior of the smooth min Rényi divergence which is defined according to this equivalent definition. In Section IV we prove one-shot Stein’s lemma. We then give an operational meaning to smooth min Rényi divergence. In Section V one-shot analogue of a special case of the packing lemma is proved. In Section VI we mention the definition of the smooth max Rényi divergence. We show that this smooth divergence satisfies data processing inequality. We
also prove a lemma which quantifies the asymptotic behavior of this divergence. A similar lemma was proved in [8] under certain assumptions. We let go all of this assumptions and prove this lemma in full generality. The result of this lemma is also proved in the quantum case [19]. However, we give a totally different proof. In particular, our proof involves more straightforward arguments. In Section VII we prove one-shot equivalent of the covering lemma known in the asymptotic i.i.d. case. In Section VIII we define \(\varepsilon\)-maximum distortion criterion between two random variables. We then give a one-shot achievable source coding rate under this distortion criterion.

II. NOTATIONS

We will be using the notation \((\Omega, \mathcal{F}, P)\) to represent a probability measure space where \(\Omega\) represents a collection of sets, \(\mathcal{F}\) represents the sigma algebra generated by \(\Omega\) and \(P\) represents the probability measure defined on \(\mathcal{F}\). \(\text{Supp}(P)\) will be used to represent the following set

\[
\bigcup_{i: A_i \in \mathcal{F} \atop P_i(A_i) > 0} A_i.
\]

We will use the notation \(\Omega_n := \Omega^\otimes n\) to represent \(n\)-fold Cartesian product of \(\Omega\) and \(\mathcal{F}_n\) will be used to represent the sigma generated by \(\Omega_n\) with \(P_n\) the corresponding measure on \(\mathcal{F}_n\). Likewise \(\chi^n\) represents the \(n\)-fold Cartesian product of the set \(X\). The notation \(X\) will be used to represent a random variable. \(X = \{X_n\}_{n=1}^\infty\) will be used to represent an arbitrary sequence of random variables where for every \(n \in \mathbb{N}\), \(X^n\) will represent a random sequence of length \(n\). We represent the set \(\{x: 0 < x < \infty\}\) by \(\mathbb{R}^+\). \(X \times Y\) will represent the cartesian product of two sets. We will be using the notation \(\chi(\cdot)\) to represent an indicator function whose definition will be explained according to the context. We will use \(|\cdot|\) to represent the cardinality of a set. The notation \(X^c\) will be used to represent compliment of a set \(X\). Throughout this paper we will assume that log is to the base 2.

III. SMOOTH MIN RÉNYI DIVERGENCE

**Definition 1:** (Wang et al. [6])

Let \((\Omega, \mathcal{F}, P)\) and \((\Omega, \mathcal{F}, Q)\) be two probability measure spaces. For every \(\varepsilon \geq 0\), the smooth Rényi divergence of order zero between the measures \(P\) and \(Q\) is defined as follows

\[
D_\varepsilon^0(P||Q) := \sup_{\phi: \mathcal{F} \to [0,1]} -\log \int_\Omega \phi(\omega)dQ(\omega),
\]

where \(\phi\) is \(\mathcal{F}\) measurable. The supremum in (1) is achieved by choosing \(\phi = 1\) where \(dP/dQ\) is large. For more details on this see [6].

Using the fact that the \(\phi\) which achieves the supremum in (1) is an indicator function we give an equivalent definition for smooth min Rényi divergence of order zero. This equivalent definition will help us in proving one-shot bounds for some important information theoretic problems.

**Definition 2:** (Equivalent definition)

Let \((\Omega, \mathcal{F}, P)\) and \((\Omega, \mathcal{F}, Q)\) be two probability measure spaces. For every \(\varepsilon \geq 0\), the smooth min Rényi divergence between the measures \(P\) and \(Q\) is defined as follows

\[
\hat{D}_\varepsilon^0(P||Q) := \sup_{A \in \mathcal{F}} -\log \int_\Omega \chi_A(\omega)dQ(\omega),
\]

where

\[
\chi_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}
\]

Notice that if \(|\mathcal{F}| < \infty\), then the supremum in (2) can be replaced by maximum.

For \(\varepsilon = 0\), both the above definitions reduce to

\[
D_0(P||Q) := -\log Q[\text{Supp}(P)].
\]

(3) represents Rényi divergence of order zero [21]. From here on we will call smooth min Rényi divergence as the smooth Rényi divergence of order zero. We call Rényi divergence of order zero as min Rényi divergence because Rényi divergence is an increasing function of \(\alpha\) [21].

We will be using the notation \(\hat{D}_\varepsilon^0(P||Q)\) in all of our further discussions to highlight the fact that smooth Rényi divergence of order zero is defined according to Definition 2.

**Lemma 1:** Let \(P := \{P_n\}_{n=1}^\infty\) and \(Q := \{Q_n\}_{n=1}^\infty\) be an arbitrary sequences of probability measures defined on a sequence of measurable space \(\{(\Omega_n, \mathcal{F}_n)\}_{n=1}^\infty\), where for every \(n \in \mathbb{N}\), \(\Omega_n := \Omega^\otimes n\) and \(\mathcal{F}_n\) represents the sigma algebra generated by \(\Omega_n\). Then

\[
\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \frac{1}{n} \hat{D}_\varepsilon^0(P_n||Q_n) := I(P; Q),
\]

where

\[
I(P; Q) := \{P_n\}-\liminf_{n \to \infty} \frac{1}{n} \log \frac{dP_n}{dQ_n}.
\]

The notation on the R.H.S. of (4) means \(\lim\inf\) in probability with respect to the sequence of probability measures \(\{P_n\}\), i.e., if \(\{Z_n\}_{n=1}^\infty\) is an arbitrary sequence of random variables then

\[
\{P_n\}-\lim\inf_{n \to \infty} Z_n := \sup \{\alpha : \lim_{n \to \infty} P_n\{Z_n > \alpha\} = 0\}
\]

For more details on this notation see [20]. In particular, if \(P = \{P^n\}_{n=1}^\infty\) and \(Q = \{Q^n\}_{n=1}^\infty\), where \(P^n\) and \(Q^n\) represent the product distribution on the measurable space \((\Omega^\otimes n, \mathcal{F}^\otimes n)\). Then

\[
\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \frac{1}{n} \hat{D}_\varepsilon^0(P_n||Q_n) := D(P||Q),
\]

where \(D(P||Q)\) is the Kullback-Leibler divergence between the probability measures \(P\) and \(Q\) [2].

**Proof:** The essential idea of the proof follows from [6, Lemma 2]. See Appendix for the detailed proof.
IV. BINARY HYPOTHESIS TESTING

We now define the problem of binary hypothesis testing which will help us to give an operational interpretation for smooth Rényi divergence of order zero.

Let $\langle \Omega, \mathcal{F}, P \rangle$ and $\langle \Omega, \mathcal{F}, Q \rangle$ be two probability measure spaces. Let $X : \Omega \rightarrow \mathcal{X}$ be a random variable which is $P$ measurable and let $\bar{X} : \Omega \rightarrow \mathcal{X}$ be another random variable which is $Q$ measurable. We will call $X$ and $\bar{X}$ as two sources from here on. We consider the hypothesis testing problem with the null hypothesis $X$ and the alternative hypothesis $\bar{X}$.

In ordinary hypothesis testing problem we choose a subset $D \subseteq \mathcal{X}$ as an acceptance region. If $x \in \mathcal{X}$, an output from one of the two sources, belongs to $D$, then we choose the null hypothesis $X$ to be true. Otherwise, we choose the alternative hypothesis $\bar{X}$ to be true. Let us define the following two probabilities of error:

$$
\mu := \Pr(X \notin D) \quad (6)
$$

and

$$
\lambda := \Pr(\bar{X} \in D). \quad (7)
$$

Equivalently (6) and (7) can be written as follows

$$
\mu := \int_{\Omega} \chi_{\mathcal{A}^c}(\omega)dP(\omega) \quad (8)
$$

and

$$
\lambda := \int_{\Omega} \chi_{\mathcal{A}}(\omega)dQ(\omega), \quad (9)
$$

where $\mathcal{A} \in \mathcal{F}$ is defined as follows

$$
\mathcal{A} := \{ \omega : X(\omega) \in D \}.
$$

The hypothesis testing is formulated as the problem of choosing an acceptance region $D$ that makes the error $\lambda$ as small as possible subject to the constraint that $\mu$ is upper bounded by some constant. Formally this can be formulated as the following constrained optimization problem

$$
\inf_{\mathcal{A} \in \mathcal{F}} \int_{\Omega} \chi_{\mathcal{A}}(\omega)dP(\omega) < \varepsilon, \quad (10)
$$

Subject to the constraint given in (10) the error probability $\lambda$ can be written as

$$
\lambda \simeq 2^{-\varepsilon R} \quad (R > 0). \quad (11)
$$

For more details on (11) see [20]. It is therefore important to know how large $R$ can be made. The following definitions concisely describe such a situation. These definitions are motivated by the definitions given in [20]. For more details and intuition about these definitions see [20, Definition 4.2.1 and Definition 4.2.2]

**Definition 3:** ($\varepsilon$-achievable rate)

Rate $R$ is $\varepsilon$-achievable $\iff$ There exists $A \in \mathcal{F}$ satisfying $\mu < \varepsilon$ and $-\log \lambda \geq R$.

**Definition 4:** (Supremum $\varepsilon$-achievable error probability exponent)

$$
B^\varepsilon(P||Q) := \sup \{ R \mid R \text{ is } \varepsilon\text{-achievable} \}.
$$

The following theorem is the one-shot generalization of [20, Theorem 4.1.1]

**Theorem 1:** For $\varepsilon \geq 0$, $B^\varepsilon(P||Q) = \hat{D}_0^\varepsilon(P||Q)$.

**Proof:** (Achievability)

Let $R = \hat{D}_0^\varepsilon(P||Q) - \gamma$ where $\gamma > 0$ is arbitrary. From Definition 2 (property of supremum) it follows that for any arbitrary $\gamma > 0$ there exists a set $A \in \mathcal{F}$ such that

$$
\int_{\Omega} \chi_{\mathcal{A}}(\omega)dP(\omega) \geq 1 - \varepsilon \quad (12)
$$

and

$$
-\log \int_{\Omega} \chi_{\mathcal{A}}(\omega)dQ(\omega) \geq \hat{D}_0(\gamma P||Q) - \gamma. \quad (13)
$$

Let $A \in \mathcal{F}$ which satisfies (12) and (13) be the acceptance region for $P$. From (8), (9), (12) and (13) it now easily follows that

$$
\mu := \int_{\Omega} \chi_{\mathcal{A}^c}(\omega)dP(\omega) < \varepsilon \quad (14)
$$

and

$$
\lambda := \int_{\Omega} \chi_{\mathcal{A}}(\omega)dQ(\omega) \leq 2^{-\hat{D}_0(\gamma P||Q) - \gamma}. \quad (15)
$$

Thus taking log on both sides of (15) and rearranging the terms we get

$$
-\log \lambda \geq \hat{D}_0(\gamma P||Q) - \gamma.
$$

We have thus established the fact that $R = \hat{D}_0(\gamma P||Q) - \gamma$ for an arbitrary $\gamma > 0$ is $\varepsilon$-achievable. This further implies that $B^\varepsilon(P||Q) \geq \hat{D}_0(\gamma P||Q)$.

**Converse**

Let $R$ be $\varepsilon$-achievable rate. Then, there exists an $A \in \mathcal{F}$ such that $\mu < \varepsilon$ and $-\log \lambda \geq R$. From (9) and Definition 2 it follows that

$$
\hat{D}_0(\gamma P||Q) \geq -\log \lambda \geq R. \quad (16)
$$

Since (16) is true for every $\varepsilon$-achievable rate $R$, it implies that

$$
B^\varepsilon(P||Q) \leq \hat{D}_0(\gamma P||Q).
$$

This completes the proof.

Thus, $\hat{D}_0(\gamma P||Q)$ is the best error probability exponent possible for binary hypothesis testing problem between the measures $P$ and $Q$ when the probability of making a wrong decision under $P$ is at most $\varepsilon$. Notice that for $\varepsilon = 0$, $B^0(\gamma P||Q) = D_0(\gamma P||Q)$. Thus, the operational interpretation of Rényi divergence of order $0$ is that it is the best error exponent possible for the binary hypothesis testing problem between the measures $P$ and $Q$ when the probability of making a wrong decision under $P$ is zero.

We now show that Theorem 1 is the one-shot generalization of [20, Theorem 4.1.1]. In particular, we will show that Stein’s lemma is a special case of this generalization. Let $\{(\Omega_n, \mathcal{F}_n, P_n)\}_{n=1}^\infty$ and $\{(\Omega_n, \mathcal{F}_n, Q_n)\}_{n=1}^\infty$ be sequences of two probability measure spaces. Also, let $X$ and $\bar{X}$ be two arbitrary sequence of random variables where for every $n \in \mathbb{N}$, $X^n : \Omega_n \rightarrow \mathcal{X}^n$ is a random sequence distributed according to $P_n$ and $\bar{X}^n : \Omega_n \rightarrow \mathcal{X}^n$ is a random sequence which is...
distributed according to $Q_n$. Define $\mu_n$ and $\lambda_n$ in a similar way as defined in (8) and (9). Furthermore, let $P := \{P_n\}_n$ and $Q := \{Q_n\}_n$. We now give a definition which will help us in proving our claim.

**Definition 5:** (Asymptotically achievable rate)
Rate $R$ is asymptotically achievable $\iff$ For every $n \in \mathbb{N}$, there exists a set $A_n \in F_n$ satisfying $\mu_n < \varepsilon$ and $\lim_{n \to \infty} \inf \mu_n \geq R$. 

**Definition 6:** (Asymptotically achievable supremum error probability exponent)
$B(P||Q) := \sup \{R : R \text{ is asymptotically achievable} \}$, where $P := \{P_n\}_n$ and $Q := \{Q_n\}_n$ and for every $n \in \mathbb{N}$, $P_n$ and $Q_n$ denote two probability measures on the measurable space $(\Omega \otimes n, F \otimes n)$.

A direct consequence of Theorem 1, Definition 5, Definition 6 and Lemma 1 is the following corollary.

**Corollary 1:**
$$B(P||Q) := I(P; Q).$$ \hspace{1cm} (17)

Notice that Corollary 1 is same as [20, Theorem 4.1.1]. In particular, it is equivalent to Stein’s lemma when $X \sim P$ and $X \sim Q$ where $P := \{P_n\}_n$ and $Q := \{Q_n\}_n$.

Notice that the above corollary is Stein’s lemma in the case when $P = \{P^\otimes n\}_n$ and $Q = \{Q^\otimes n\}_n$, where $P^\otimes n$ and $Q^\otimes n$ represent the product measure on the measurable space $(\Omega \otimes n, F \otimes n)$. For details on Stein’s lemma see [2, Theorem 12.8.1]. Thus Theorem 1 can also be considered as the one-shot generalization of "Stein’s lemma".

**V. ONE-SHOT PACKING LEMMA**

**Lemma 2:** Let $(X, Y) \sim P_{XY}$ be a pair of jointly distributed random variables taking values over the set $\mathcal{X} \times \mathcal{Y}$, where $|\mathcal{X} \times \mathcal{Y}| \leq \infty$. Let $X(i), i \in M$, where $|M| \leq k$, be random variables distributed according to $P_X$. Further assume that $X(i), i \in M$ are independent of $Y$, where $Y \sim P_Y$, but is arbitrarily dependent on other $X(i)$ random variables. Let $\varepsilon > 0$ and consider $A \subseteq \mathcal{X} \times \mathcal{Y}$ such that
$$\int_{\mathcal{X} \times \mathcal{Y}} \chi_A(x, y) dP_{XY}(x, y) \geq 1 - \varepsilon$$
and
$$\hat{D}_0(P_{XY}||P_X \times P_Y) = - \log \int_{\mathcal{X} \times \mathcal{Y}} \chi_A(x, y) d(P_X \times P_Y)(x, y).$$ \hspace{1cm} (18)

Then
$$\Pr \left\{ (X(i), Y) \in A \text{ for some } i \in M \right\} \leq \varepsilon,$$ \hspace{1cm} (19)
if
$$\log k \leq \hat{D}_0(P_{XY}||P_X \times P_Y) + \log \varepsilon.$$ \hspace{1cm} (20)

Furthermore, if the R.H.S. of (20) is negative then there does not exist a $k > 0$ such that (19) is satisfied.

**Proof:**
\begin{align*}
\Pr & \left\{ (X(i), \hat{Y}) \in A \text{ for some } i \in [1 : k] \right\} \\
& \leq \sum_{i=1}^{k} \int_{\mathcal{X} \times \mathcal{Y}} \chi_A(x, y) d(P_X \times P_Y)(x, y) \\
& \leq k^2 - \hat{D}_0(P_{XY}||P_X \times P_Y),
\end{align*}
where $a$ follows from the union bound and the fact that $X(i)$ is independent of $Y$ for every $i \in [1 : k]$ and $b$ follows from (18). Now let
$$k^2 - \hat{D}_0(P_{XY}||P_X \times P_Y) \leq \varepsilon.$$ \hspace{1cm} (21)
Taking log on both sides of (21) and rearranging the terms we get
$$\log k \leq \hat{D}_0(P_{XY}||P_X \times P_Y) + \log \varepsilon.$$ \hspace{1cm} (22)
If the R.H.S. of (20) is negative then in it would imply that it is not possible to satisfy the bound in (19) for any $k > 0$. To see this more precisely we give a proof by contradiction. Suppose there exists $k > 0$, in particular, suppose $k \geq 1$ is such that (19) is satisfied. This assumption implies that
$$\Pr \left\{ (X(1), Y) \in A \right\} \leq \varepsilon,$$ \hspace{1cm} (23)
follows because
$$\Pr \left\{ (X(1), \hat{Y}) \in \mathcal{A} \right\} \leq \Pr \left\{ (X(i), \hat{Y}) \in \mathcal{A} \text{ for some } i \in \mathcal{M} \right\}.$$ \hspace{1cm} (24)
Now since it is given to us that
$$\hat{D}_0(P_{XY}||P_X \times P_Y) + \log \varepsilon < 0,$$ \hspace{1cm} (24)
this would further imply the following
$$\int_{\mathcal{X} \times \mathcal{Y}} \chi_A(x, y) d(P_X \times P_Y)(x, y) > \varepsilon,$$ \hspace{1cm} (25)
where the above inequality follows from (18) and (24). Notice that (25) is in contradiction with (23) because $\hat{Y}$ is independent of $X(1)$. This completes the proof. $\blacksquare$

It is worth mentioning here that Lemma 4 is one-shot analogue of a special case of the famous covering lemma known in the i.i.d. asymptotic case [18]. In a fully generalized packing lemma $\hat{Y}$ need not be distributed with $P_Y$. To see the relation between these two lemmas let
$$(X^n, Y^n) := [(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)],$$
where $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are i.i.d. pairs of random variable distributed according to $P_{XY}$. Furthermore, let
$$\hat{Y}^n := [\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n]$$
be n i.i.d. random variables distributed according to $P_Y$. Also let us assume that $\hat{Y}^n$ is independent of $X^n$. Now notice that for large enough $n$ and $\varepsilon$ arbitrary close to zero the set $\mathcal{A} \subseteq (\mathcal{X} \times \mathcal{Y})^n$ which satisfies
$$\int_{(\mathcal{X} \times \mathcal{Y})^n} \chi_A(x^n, y^n) dP_{X^nY^n}(x^n, y^n) \geq 1 - \varepsilon$$
and
\[ D_0^\phi(P_{X,Y}||P_XP_Y) = -\log \int_{(X\times Y)^n} \chi_{A}(x^n, y^n)d(P_{X^n} \times P_{Y^n})(x^n, y^n) \]
will be equal to the typical set \( T^\varepsilon \) defined with respect to \( P_{XY} \). For more details on \( T^\varepsilon \), see e.g. [2]. It now easily follows from Lemma 3 that the rate derived for this asymptotic case using Lemma 4 (normalized properly) is equal to \( I(X;Y) \) which is the packing rate required in the asymptotic i.i.d. case.

VI. SMOOTH MAX RÉNYI DIVERGENCE

**Definition 7:** Consider a measurable space \((\Omega, \mathcal{F})\). Let \( P \) and \( Q \) be two probability measures on \( \mathcal{F} \) such that \( \text{Supp}(P) \subseteq \text{Supp}(Q) \). For every \( \varepsilon \geq 0 \), the smooth max Rényi divergence between the measures \( P \) and \( Q \) is defined as follows
\[
D_\infty^\phi(P||Q) := \inf_{\phi \in B_\varepsilon(P)} \log \sup_{A \in \mathcal{F}} \frac{\phi(A)}{Q(A)},
\]
where \( B_\varepsilon(P) \) is the set of all measures on \( \mathcal{F} \) for which the following holds
\[
\phi(A) \leq P(A) \forall A \in \mathcal{F} \text{ and } \int \phi(\omega) d\omega \geq 1 - \varepsilon.
\]

Notice that \( B_\varepsilon(P) \) in (26) also contains measures which are not probability measures. Therefore smooth max Rényi divergence can be negative. Furthermore, from the Definition 7 it follows that \( D_\infty^\phi(P||Q) \) is a non-increasing function of \( \varepsilon \). Also for \( \varepsilon = 0 \), the above definition reduces to
\[
D_\infty(P||Q) := \log \sup_{A \in \mathcal{F}} \frac{P(A)}{Q(A)},
\]
which is the Rényi divergence of order infinity [21]. From here on we will call smooth max Rényi divergence as smooth Rényi divergence of order infinity.

**Theorem 2:** (Data Processing inequality)
Let \((\Omega, \mathcal{F}, P)\) and \((\Omega, \mathcal{F}, Q)\) be two probability measure spaces. Consider a stochastic kernel \( W \) from the measurable space \((\Omega, \mathcal{F})\) to \((\Omega', \mathcal{F}')\). For every \( \varepsilon > 0 \),
\[
D_\infty^W(P||Q) \geq D_\infty^{W'}(P'||Q'),
\]
where \( P' := W \circ P \) and \( Q' := W \circ Q \) are probability measures on measurable space \((\Omega', \mathcal{F}')\).

**Proof:** Consider any measure \( \phi \) defined on \( \mathcal{F} \) such that \( \phi \in B_\varepsilon(P) \), \( B_\varepsilon(P) \) is defined in Definition 26. Let us now define a measure \( \phi' \) on \( \mathcal{F}' \) as follows. For every \( A' \in \mathcal{F}' \)
\[
\phi'(A') := \int_{\Omega'} \int \chi_{A'}(\omega')d\phi(\omega)W(d\omega', \omega).
\]
(27)
It is easy to see from (27) that
\[
\phi'(A') \leq P'(A') \forall A' \in \mathcal{F}',
\]
(28)
By changing the order of integration (Fubini’s theorem) in (27) we get
\[
\int_{\Omega'} \phi'(\omega') \geq 1 - \varepsilon.
\]
(29)
Thus, \( \phi' \in B_\varepsilon(W \circ P) \). Let \( \int_{\Omega} d\phi(\omega) = \lambda \). From [21, Theorem 6.11] it follows that
\[
D_\infty \left( \frac{\phi}{\lambda} || Q \right) \geq D_\infty \left(W \circ \frac{\phi}{\lambda} || W \circ Q \right)
\]
\[
= D_\infty \left( \frac{\phi'}{\lambda} || Q' \right),
\]
(30)
where (30) follows from (27). It now easily follows from (30) that
\[
D_\infty(\phi || Q) \geq D_\infty(\phi' || Q') \geq D_\infty(W \circ P || W \circ Q),
\]
(31)
where (31) follows from (28), (29) and Definition 7. Since (31) is true for every \( \phi \in B_\varepsilon(P) \) the result follows.

**Lemma 3:** Let \( P := \{ P_n \}_{n=1}^\infty \) and \( Q := \{ Q_n \}_{n=1}^\infty \) be an arbitrary sequence of probability measures defined on a sequence of measurable space \((\Omega_n, \mathcal{F}_n)\) such that for every \( n \in \mathbb{N} \), \( \Omega_n := \Omega_n^\otimes n \) and \( \mathcal{F}_n \) represents the sigma algebra generated by \( \Omega_n \). Then
\[
\lim_{\varepsilon \to 0} \inf_{n \to \infty} \frac{1}{n} D_\infty^\phi(P_n||Q_n) := I(P;Q),
\]
where
\[
I(P;Q) := \inf \left\{ \alpha \mid \lim_{n \to \infty} \frac{1}{n} \log \frac{dP_n}{dQ_n} = \alpha \right\}.
\]
(32)
The notation on the R.H.S. of (32) means \( \lim sup \) in probability with respect to the sequence of probability measures \( \{ P_n \} \), i.e., if \( \{ Z_n \}_{n=1}^\infty \) is an arbitrary sequence of random variables then
\[
\{ P_n \} \cdot \lim sup_{n \to \infty} Z_n := \inf \left\{ \alpha \mid \lim_{n \to \infty} P_n \{ Z_n < \alpha \} = 1 \right\}.
\]
For more details on this notation see [20]. In particular, if \( P = \{ P^{\otimes n} \}_{n=1}^\infty \) and \( Q = \{ Q^{\otimes n} \}_{n=1}^\infty \), where \( P^{\otimes n} \) and \( Q^{\otimes n} \) represent the product distributions of \( P \) and \( Q \) on the measurable space \((\Omega^{\otimes n}, \mathcal{F}^{\otimes n})\). Then
\[
\lim_{\varepsilon \to 0} \inf_{n \to \infty} \frac{1}{n} D_\infty^\phi(P_n||Q_n) := D(P||Q).
\]
(33)
where \( D(P||Q) \) is the Kullback-Leibler distance between the probability measures \( P \) and \( Q \) [2].

**Proof:** The above lemma was proved in [8] under the assumption that the sequences \( P \) and \( Q \) are sequences of discrete probability mass functions such that for every \( n \in \mathbb{N} \), \( P_n \) and \( Q_n \) are defined on a set which has a finite cardinality. Here we let go of such assumptions. See Appendix for the detailed proof.

VII. ONE SHOT COVERING LEMMA

In all the discussions in this section we will assume that the random variables are discrete and have a finite range.

**Lemma 4:** Let \((X,Y) \sim P_{XY} \) be a pair of jointly distributed discrete random variables taking values over the set \( \mathcal{X} \times \mathcal{Y} \) and let \( 0 < \varepsilon_1 \) be such that \( 2\varepsilon_1 < \varepsilon \). Let \( U^* \) be such that
\[
U^* \in \arg \min_{U \subseteq \mathcal{X} \times \mathcal{Y}} \frac{1}{P_{XY}(U)} > 1 - \varepsilon_1.
\]
(34)
Let \( \hat{Y}(i) \sim P_Y, i \in [1 : k] \), be independent and identically
distributed random variables which are also independent of \( X \).
Then,

\[
\Pr \left\{ (X, \hat{Y}(i)) \not\in \mathcal{U}^*, \forall i \in [1 : k] \right\} \leq \varepsilon, \tag{35}
\]

if

\[
\log k \geq D_{\infty}^{\varepsilon}(P_{XY} \| P_X \times P_Y) + \log[-\ln(\varepsilon - 2\varepsilon_1)]. \tag{36}
\]

Furthermore, if \( D_{\infty}^{\varepsilon}(P_{XY} \| P_X \times P_Y) \) is negative then it would
imply that \( k = 1 \) is sufficient enough for the bound in (35) to
hold.

**Proof:** Let \( \phi \in \mathcal{B}^\varepsilon(P) \) be such that

\[
D_{\infty}^{\varepsilon}(P_{XY} \| P_X \times P_Y) = \log \max_{(x,y) \in X \times Y} \frac{\phi(x,y)}{P_X(x)P_Y(y)}, \tag{37}
\]

where

\[
\phi(Y = y | X = x) := \begin{cases} \phi(x,y) / P_X(x) & \text{if } P_X(x) > 0 \\ 0 & \text{o.w.} \end{cases}
\]

We assume here that (37) is positive. We will deal with the
case when (37) is negative separately towards the end of the
proof. The existence of a \( \phi \) which satisfies (37) follows from
[8, Lemma 2]. Let us define the following indicator function
which will help us in further calculations

\[
\chi(x,y) = \begin{cases} 1 & \text{if } (x,y) \in \mathcal{U}^* \\ 0 & \text{o.w.} \end{cases}
\]

\[
\Pr \left\{ (X, \hat{Y}(i)) \not\in \mathcal{U}^*, \forall i \in [1 : k] \right\} = \sum_{x \in X} P_X(x) \Pr \left\{ (x, \hat{Y}(i)) \not\in \mathcal{U}^*, \forall i \in [1 : k] \right\}
\]

\[
= \sum_{x \in X} P_X(x) \prod_{i=1}^{k} \Pr \left\{ (x, \hat{Y}(i)) \not\in \mathcal{U}^* \right\} = \sum_{x \in X} P_X(x) \left( 1 - \sum_{y \in Y} P_Y(y) \chi(x,y) \right)^k \leq \sum_{x \in X} P_X(x) \left( 1 - 2^{-D_{\infty}^{\varepsilon}(P_{XY} \| P_X \times P_Y)} \sum_{y \in Y} \phi(Y = y | X = x) \chi(x,y) \right)^k
\]

\[
e \leq \sum_{x \in X} P_X(x) e^{-(k2^{-D_{\infty}^{\varepsilon}(P_{XY} \| P_X \times P_Y)} \sum_{y \in Y} \phi(Y = y | X = x) \chi(x,y))} \leq \sum_{x \in X} P_X(x) \left( 1 - \sum_{y \in Y} \phi(Y = y | X = x) \chi(x,y) \right) + e^{-k2^{-D_{\infty}^{\varepsilon}(P_{XY} \| P_X \times P_Y)}} \leq 2\varepsilon_1 + e^{-k2^{-D_{\infty}^{\varepsilon}(P_{XY} \| P_X \times P_Y)}}
\]

where \( a \) follows because for every \( i \neq j \) and \( i,j \in [1 : k] \),
the random variables \( \hat{X}(i) \) and \( \hat{X}(j) \) are independent of each
other and also \( \hat{X}(i) \) is independent of \( Y \) for every \( i \in [1 : k] \);
\( b \) follows from the Definition 7 and (38); \( c \) follows from the
inequality \((1 - x)^y \leq e^{-xy} \) \((0 \leq x \leq 1, y \geq 0)\); \( d \) follows
from the inequality \( e^{-xy} \leq 1 - y + x \) \((x \geq 0, 0 \leq y \leq 1)\)
and \( e \) follows from the following set of inequalities

\[
1 - \varepsilon_1 \leq \sum_{(x,y) \in X \times Y} \phi(x,y)
\]

\[
= \sum_{(x,y) \in \mathcal{U}^*} \phi(x,y) + \sum_{(x,y) \in \mathcal{U}^*} \phi(x,y) \leq P_{XY}(\mathcal{U}^*) + \sum_{(x,y) \in \mathcal{U}^*} \phi(x,y) \leq \varepsilon_1 + \sum_{(x,y) \in \mathcal{U}^*} \phi(x,y)
\]

where \( a \) and \( b \) both follow from the fact that, \( \phi(x,y) \in \mathcal{B}^\varepsilon(P_{XY}) \) and \( c \) follows from (34). Now let

\[
e^{-k2^{-D_{\infty}^{\varepsilon}(P_{XY} \| P_X \times P_Y)}} + 2\varepsilon_1 \leq \varepsilon.
\]

It is now easy to see that

\[
\log[-\ln(\varepsilon - 2\varepsilon_1)] + D_{\infty}^{\varepsilon}(P_{XY} \| P_X \times P_Y) \leq \log k.
\]
If $D_{\infty}^2(P_{XY}||P_X \times P_Y)$ is negative then it would imply that $k = 1$ is sufficient enough to satisfy the bound in (35). To see this notice the following arguments. Since it is given to us that

$$D_{\infty}^2(P_{XY}||P_X \times P_Y) < 0. \quad (41)$$

it then further implies the following

$$\frac{\phi(x, y)}{P_X(x)P_Y(y)} \leq 1 \ \forall (x, y) \in \mathcal{X} \times \mathcal{Y}. \quad (42)$$

We will now prove that under the assumption of (41) the following holds

$$\Pr \left\{ (X, \hat{Y}(1)) \notin \mathcal{U}^* \right\} \leq \varepsilon. \quad (43)$$

Let us now calculate the following

$$\Pr \left\{ (X, \hat{Y}(1)) \in \mathcal{U}^* \right\} = \sum_{(x, y) \in \mathcal{U}^*} P_X(x)P_Y(y) \geq \sum_{(x, y) \in \mathcal{U}^*} \phi(x, y) \geq 1 - 2\varepsilon_1 \geq 1 - \varepsilon, \quad (44)$$

where $a$ follows because $\hat{Y}(1) \sim P_Y$ is generated independent of $X$; $b$ follows from (55), $c$ follows from (40) and $d$ follows from the fact that $2\varepsilon_1 < \varepsilon$. (43) now trivially follows from (44). This completes the proof. \hfill \blacksquare

It is worth mentioning here that Lemma 4 is one-shot analogue of the famous covering lemma in the asymptotic case [18]. To see the relation between these two lemmas let

$$(X^n, Y^n) := [(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)],$$

where $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are $n$ i.i.d. pairs of random variable distributed according to $P_{XY}$. Furthermore, let

$$\hat{Y}^n := [\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n]$$

be $n$ i.i.d. random variables distributed according to $P_Y$. Also let us assume that $\hat{Y}^n$ is independent of $X^n$. Now notice that for large enough $n$ and $\varepsilon$ arbitrarily close to zero

$$\mathcal{U}^* := \operatorname*{arg min}_{\mathcal{U} \subseteq (\mathcal{X} \times \mathcal{Y})^n; \mu_X(\mathcal{U}) \geq 1 - \varepsilon_1} |\mathcal{U}| = \mathcal{T}^n_{\varepsilon_1},$$

where $\mathcal{T}^n_{\varepsilon_1}$ is the typical set defined with respect to the distribution $P_{XY}$ and $2\varepsilon_1 < \varepsilon$. For more details on $\mathcal{T}^n_{\varepsilon_1}$, see e.g. [2]. It now easily follows from Lemma 3 that the rate derived using Lemma 4 (normalized properly) for this asymptotic case is equal to $I(X; Y)$ which is the coding rate required in the asymptotic i.i.d. case.

VIII. ONE-SHOT SOURCE ENCODING UNDER MAXIMUM DISTORTION CRITERION

In this section we will assume that all the random variables are discrete and take value over a finite set. The definition given below is motivated from maximum distortion criterion mentioned in [20] for the non-i.i.d. asymptotic case.

**Definition 8:** (One-shot $\varepsilon$-maximum distortion criterion) Let $(X, Y) \sim P_{XY}$ with range $\mathcal{X} \times \mathcal{Y}$. For every $\varepsilon > 0$, the $\varepsilon$-maximum distortion between these two random variables for some distortion measure $d$ is defined as

$$\bar{\lambda}_\varepsilon(X, Y) := \inf \left\{ \lambda : \Pr \{d(X, Y) \leq \lambda \} > 1 - \varepsilon \right\},$$

where $d : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$. Throughout this section we will assume that the distortion measure is a bounded function.

We now describe the task of one-shot source coding under one-shot $\varepsilon$-maximum distortion criterion. Let $X \sim P_X$, with range $\mathcal{X}$. Consider another set $\mathcal{Y}$ called the estimation set. The encoder describes the source symbol $X$ by an index $e(X) \in \mathcal{M}$, where $\mathcal{M} = \{1, \ldots, \ell^\varepsilon(\gamma|X)\}$ and is called the code. We call

$$\log \ell^\varepsilon(\gamma|X) = \log |\mathcal{M}| \quad (45)$$

as the coding rate of the encoder. The decoder $f$ represents $X$ by an estimate $Y \in \mathcal{Y}$, i.e., $Y = f(e(X))$. We call the pair $(e, f)$ of encoder and decoder with the code of size $\ell^\varepsilon(\gamma|X)$ and $\bar{\lambda}_\varepsilon(X, f(e(X))) \leq \gamma$ a $(1, \ell^\varepsilon(\gamma|X), \gamma)$ one-shot code. In this coding system we wish to minimize $\ell^\varepsilon(\gamma|X)$ such that

$$\bar{\lambda}_\varepsilon(X, f(e(X))) \leq \gamma,$$

where $\gamma > 0$ and $0 < \varepsilon \leq 1$ are predefined parameters which quantify the one-shot $\varepsilon$-maximum distortion between the source symbol and its estimate.

**Definition 9:** (One-shot $\varepsilon$-achievable rate distortion pair) A one-shot rate distortion pair $(R, \gamma)$ for a given distortion measure $d$ is called $\varepsilon$-achievable if and only if there exists a $(1, \ell^\varepsilon(\gamma|X), \gamma)$ one-shot code such that $\bar{\lambda}_\varepsilon(X, f(e(X))) \leq \gamma$ and $\log \ell^\varepsilon(\gamma|X) \leq R$.

**Theorem 3:** Let $X \sim P_X$ and $\varepsilon, \varepsilon_1 \in \mathbb{R}^+$ with $2\varepsilon_1 < \varepsilon$.

**Achievability:** For a given distortion measure $d$ and $\varepsilon$-maximum distortion $\gamma$, the following one-shot rate under the $\varepsilon$-maximum distortion criterion is achievable

$$\log \ell^\varepsilon(\gamma|X) \geq D_{\infty}^2(P_{XY}||P_X \times P_Y) + \log (-\ln(\varepsilon - 2\varepsilon_1)),$$

for some conditional $P_{Y|X}$ such that $\bar{\lambda}_{\varepsilon_1}(X, Y) \leq \gamma$. Furthermore, if $D_{\infty}^2(P_{XY}||P_X \times P_Y) < 0$, then $\ell^\varepsilon(\gamma|X) = 1$ is sufficient to satisfy the $\varepsilon$-maximum distortion criterion.

**Proof:**

1) **Random Codebook generation:** Fix a conditional $P_{Y|X}$ such that

$$\bar{\lambda}_{\varepsilon_1}(X, Y) \leq \gamma. \quad (46)$$

Let $P_Y$ be the marginal distribution of $P_{XY}$. Now generate $Y_1, \ldots, Y_{\ell^\varepsilon(\gamma|X)} \in \mathcal{Y}$ independently with probability
distribution $P_Y$. These $\ell^c(\gamma|X)$ independent realizations of $Y$ forms a random codebook, i.e.,

$$C = \{Y_1, \ldots, Y_{\ell^c(\gamma|X)}\}.$$  

2) **Decoding:** For every $x \in X$ we define the encoder $e$ by $e(x) = i$, where $i$ is determined from

$$d(x, Y_i) = \min_{1 \leq j \leq \ell^c(\gamma|X)} d(x, Y_j).$$

3) **Decoding:** If the decoder receives $i$, then the decoder $f$ outputs $Y_i$, i.e.,

$$f(i) = Y_i.$$  

Let us now calculate $\Pr\{d(X, f(e(X))) > \tilde{\lambda}_e(X, Y) + \alpha\}$ where $\alpha > 0$ is arbitrary. We will show that this probability is less than $\varepsilon$ if

$$\log \ell^c(\gamma|X) \geq D^{\varepsilon^c}(P_{XY}||P_X \times P_Y) + \log[-\ln\varepsilon - 2\varepsilon_1)].$$  

This would further imply that for every $\alpha > 0$, under the condition mentioned in (47) $\tilde{\lambda}_e(X, f(e(X))) \leq \tilde{\lambda}_e(X, Y) + \alpha$ (follows from the Definition 8). This further implies that

$$\tilde{\lambda}_e(X, f(e(X))) \leq \tilde{\lambda}_e(X, Y).$$

We now give few definitions which will help us to prove the desired result. For any $\alpha > 0$, define

$$\Gamma^{(\varepsilon_1)} := \{(x, y) : d(x, y) \leq \tilde{\lambda}_e(X, Y) + \alpha\}. \quad (48)$$

Let $\phi \in \mathcal{B}^{\varepsilon^c}(P)$ be such that

$$D^{\varepsilon^c}(P_{XY}||P_X \times P_Y) = \log \max_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \frac{\phi(x,y)}{P_X(x)P_Y(y)}.$$  

where

$$\phi(Y = y|X = x) := \begin{cases} \frac{\phi(x,y)}{P_X(x)} & \text{if } P_X(x) > 0 \\ 0 & \text{o.w.} \end{cases} \quad (50)$$

The existence of a $\phi$ which satisfies (49) follows from [8, Lemma 2]. We will assume here that (49) is positive. We will take the case when (49) is negative towards the end of the proof. Let us define the following indicator function which will help us in the calculations below

$$\chi(x,y) = \begin{cases} 1 & \text{if } (x, y) \in \Gamma^{(\varepsilon_1)}, \\ 0 & \text{otherwise.} \end{cases} \quad (51)$$

$$\Pr\{d(X, f(e(X))) > \tilde{\lambda}_e(X, Y) + \alpha\} = \Pr\{d(X, Y_i) > \tilde{\lambda}_e(X, Y) + \alpha, \forall i \in [1 : \ell^c(\gamma|X)]\}$$

$$= \sum_{x \in \mathcal{X}} P_X(x) \Pr\{d(x, Y_i) > \tilde{\lambda}_e(X, Y) + \alpha, \forall i \in [1 : \ell^c(\gamma|X)]\}$$

$$\leq \sum_{x \in \mathcal{X}} P_X(x) \left(1 - \sum_{y \in \mathcal{Y}} P_Y(y) \chi(x,y)\right)^{\ell^c(\gamma|X)}$$

$$\leq \sum_{x \in \mathcal{X}} P_X(x) \left(1 - \sum_{y \in \mathcal{Y}} \phi(Y = y|X = x) \chi(x,y)\right)^{\ell^c(\gamma|X)}$$

$$\leq \sum_{x \in \mathcal{X}} P_X(x) \left(1 - \sum_{(x,y) \in \Gamma^{(\varepsilon_1)}} \phi(x,y) \right)^{\ell^c(\gamma|X)}$$

$$= 1 - \sum_{(x,y) \in \Gamma^{(\varepsilon_1)}} \phi(x,y) + e^{-\ell^c(\gamma|X)2^{-D^{\varepsilon^c}(P_{XY}||P_X \times P_Y)}}$$

$$\leq 2\varepsilon_1 + e^{-\ell^c(\gamma|X)2^{-D^{\varepsilon^c}(P_{XY}||P_X \times P_Y)}} +$$

where $a$ follows because $Y_1, \ldots, Y_{\ell^c(\gamma|X)}$ are independent and identically distributed according to $P_Y$; $b$ follows from (49) and (50); $c$ follows from the inequality $(1 - x)^y \leq e^{-xy}$ $(0 \leq x \leq 1, y \geq 0)$; $d$ follows from the inequality $e^{-xy} \leq 1 - y + x$ $(x \geq 0, 0 \leq y \leq 1)$ and $e$ follows because of the following set of inequalities

$$1 - \varepsilon_1 \leq \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \phi(x,y)$$

$$= \sum_{(x,y) \in \Gamma^{(\varepsilon_1)}} \phi(x,y) + \sum_{(x,y) \in \Gamma^{(\varepsilon_1)}} \phi(x,y)$$

$$\leq \Pr\{\Gamma^{(\varepsilon_1)}\} + \sum_{(x,y) \in \Gamma^{(\varepsilon_1)}} \phi(x,y)$$

$$1 - \varepsilon_1 \leq \varepsilon_1 + \sum_{(x,y) \in \Gamma^{(\varepsilon_1)}} \phi(x,y),$$  

where $a$ and $b$ both follow from the fact that, $\phi(x,y) \in \mathcal{B}^{\varepsilon^c}(P_{XY})$ and $c$ follows from Definition 8, (46) and (48). Now let

$$2\varepsilon_1 + e^{-\ell^c(\gamma|X)2^{-D^{\varepsilon^c}(P_{XY}||P_X \times P_Y)}} \leq \varepsilon.$$  

Using the above equation it is easy to see that

$$\log[-\ln(\varepsilon - 2\varepsilon_1)] + D^{\varepsilon^c}(P_{XY}||P_X \times P_Y) \leq \log \ell^c(\gamma|X).$$
Thus we can now conclude that $\lambda_\varepsilon(X, f(e(X))) \leq \gamma$ if

$$\log \ell^\varepsilon(\gamma | X) \geq D_{\infty}^\varepsilon(P_{XY} || P_X \times P_Y) + \log[-\ln(\varepsilon - 2\varepsilon_1)].$$

We will now show that if $D_{\infty}^\varepsilon(P_{XY} || P_X \times P_Y)$ is negative then $\ell^\varepsilon(\gamma | X) = 1$ is sufficient enough to satisfy the $\varepsilon$-maximum distortion bound. An equivalent way of proving this claim is to prove the following

$$\Pr\left\{ (X, Y_1) \notin \Gamma^{(\varepsilon_1)} \right\} \leq \varepsilon,$$  \hspace{1cm} (53)

where $\Gamma^{(\varepsilon_1)}$ is defined in (48). To prove (53) notice the following arguments. Notice that

$$D_{\infty}^\varepsilon(P_{XY} || P_X \times P_Y) < 0,$$  \hspace{1cm} (54)

would further imply that

$$\frac{\phi(x, y)}{P_X(x)P_Y(y)} \leq 1 \ \forall (x, y) \in X \times Y.$$  \hspace{1cm} (55)

Let us now consider the following

$$\Pr\left\{ (X, Y_1) \in \Gamma^{(\varepsilon_1)} \right\} = \sum_{(x, y) \in \Gamma^{(\varepsilon_1)}} P_X(x)P_Y(y)$$

$$a \geq \sum_{(x, y) \in \Gamma^{(\varepsilon_1)}} \phi(x, y)$$

$$b \geq 1 - 2\varepsilon_1$$

$$c \geq 1 - \varepsilon,$$  \hspace{1cm} (56)

where $a$ follows because $Y_1 \sim P_Y$ is generated independently of $X$; $b$ follows from (55); $c$ follows from (52) and $d$ follows from the fact that $2\varepsilon_1 < \varepsilon$. (53) now trivially follows from (56). From (53) it now easily follows that $\ell^\varepsilon(\gamma | X) = 1$ suffices for the bound

$$\lambda_\varepsilon(X, f(e(X))) \leq \gamma,$$

to hold. This completes the proof. \hfill \blacksquare

**IX. APPENDIX**

**A. Proof of Lemma 1**

We will first prove

$$\lim_{\varepsilon \to 0} \limsup_{n \to \infty} \frac{1}{n} D_0^\varepsilon(P_n || Q_n) \leq I(P; Q).$$  \hspace{1cm} (57)

Consider any $0 < \lambda < I(P; Q)$. For every $n \in \mathbb{N}$, let $A_n(\lambda)$ be the union of all the sets in $\mathcal{F}_n$ on which

$$\frac{1}{n} \log \frac{dP_n}{dQ_n} \geq \lambda.$$  \hspace{1cm} (58)

From (4) it easily follows that

$$\lim_{n \to \infty} \int_{\Omega_n} \chi_{A_n(\lambda)} dP_n(\omega) = 1.$$  \hspace{1cm} (59)

We now have the following set of inequalities

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \frac{1}{n} D_0^\varepsilon(P_n || Q_n) \geq \lim_{n \to \infty} \frac{1}{n} \log \int_{\Omega_n} \chi_{A_n(\lambda)} dQ_n(\omega)$$

$$\geq \lim_{n \to \infty} \frac{1}{n} \log \int_{\Omega_n} \chi_{A_n(\lambda)} dP_n(\omega)$$

$$\geq \lambda.$$  \hspace{1cm} (60)

where $a$ follows from Definition 2 and (59), $b$ follows from (58) and $c$ follows from (59).

We now prove the other side, i.e.,

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \frac{1}{n} D_0^\varepsilon(P_n || Q_n) \leq I(P; Q).$$  \hspace{1cm} (61)

Consider any $\gamma > 0$, such that $I(P; Q) < \gamma$. For every $n \in \mathbb{N}$, let $A_n(\gamma)$ be the union of all the sets in $\mathcal{F}_n$ on which

$$\frac{1}{n} \log \frac{dP_n}{dQ_n} \leq \gamma.$$  \hspace{1cm} (62)

From (4) it follows that there exists $\eta \in (0, 1]$, such that

$$\limsup_{n \to \infty} \int_{\Omega_n} \chi_{A_n(\gamma)} dP_n(\omega) = \eta.$$  \hspace{1cm} (63)

Since $\int_{\Omega_n} \chi_{A_n(\gamma)} dP_n(\omega) + \int_{\Omega_n} \chi_{A_n(\gamma)} dP_n(\omega) = 1$, for every $n \in \mathbb{N}$, we have

$$\liminf_{n \to \infty} \int_{\Omega_n} \chi_{A_n(\gamma)} dP_n(\omega) = 1 - \eta.$$  \hspace{1cm} (64)

For every $\varepsilon \in (0, \eta)$, let us define a sequence of sets $\{B_n\}_{n=1}^{\infty}$ such that for every $n \in \mathbb{N}$

$$\int_{\Omega_n} \chi_{B_n} dP_n(\omega) > 1 - \varepsilon.$$  \hspace{1cm} (65)

For every $n \in \mathbb{N}$, notice the following important steps

$$1 - \varepsilon < \int_{\Omega_n} \chi_{B_n \cap A_n} dP_n(\omega) + \int_{\Omega_n} \chi_{B_n \cap A_n} dP_n(\omega)$$

$$\leq \int_{\Omega_n} \chi_{B_n \cap A_n} dP_n(\omega) + \int_{\Omega_n} \chi_{A_n} dP_n(\omega)$$

By rearranging the terms in (65) and taking $\limsup$ on both sides we get

$$\limsup_{n \to \infty} \int_{\Omega_n} \chi_{B_n \cap A_n} dP_n(\omega)$$

$$> 1 - \varepsilon - \liminf_{n \to \infty} \int_{\Omega_n} \chi_{A_n} dP_n(\omega)$$

$$> \eta - \varepsilon.$$  \hspace{1cm} (66)

where (66) follows from (63). Now notice the following inequalities for large enough $n$

$$\int_{\Omega_n} \chi_{B_n} dQ_n(\omega) \geq \int_{\Omega_n} \chi_{B_n \cap A_n} dQ_n(\omega)$$

$$\geq 2^{-n\gamma} \int_{\Omega_n} \chi_{B_n \cap A_n} dP_n(\omega)$$
where $a$ follows from (61). Thus we now have the following
\[
\liminf_{n \to \infty} \left( -\frac{1}{n} \log \int_{\Omega_n} \chi_{B_n}(\omega) dQ_n(\omega) \right) \\
\leq \gamma - \limsup_{n \to \infty} \left( -\frac{1}{n} \log \chi_{B_n \cap A_n}(\omega) dP_n(\omega) \right) \\
\leq \gamma,
\]
where $a$ follows from (66). Thus
\[
\liminf_{n \to \infty} \frac{1}{n} D_\nu^\gamma(P_n||Q_n) \leq \gamma \tag{67}
\]
Since (67) is true for every $\varepsilon \in (0, \eta)$, the result will hold true for $\varepsilon \downarrow 0$. (33) follows from law of large numbers and (4). This completes the proof.

**B. Proof of Lemma 3**

We will first prove
\[
\lim \sup_{\varepsilon \to 0} \limsup_{n \to \infty} \frac{1}{n} D_\nu^\gamma(P_n||Q_n) \leq \bar{I}(P; Q).
\]
Consider any $\lambda > \bar{I}(P; Q)$. Let $A_n(\lambda)$ be the union of all $P_n$ measurable sets on which
\[
\frac{1}{n} \log \frac{dP_n}{dQ_n} \leq \lambda \tag{68}
\]
Let $\phi_n : F_n \to [0, 1]$, for every $n \in \mathbb{N}$, be a measure such that
\[
\phi_n(B_n) = \begin{cases} 
P_n(B_n) & \text{if } B_n \subseteq A_n \text{ is } P_n \text{ measurable}, \\
0 & \text{otherwise}. 
\end{cases} \tag{69}
\]
From (32) it easily follows that
\[
\lim_{n \to \infty} \int_{\Omega_n} \chi_{A_n(\lambda)}(\omega) dP_n(\omega) = 1. \tag{70}
\]
Thus from our construction of $\phi_n$, (69), it follows that
\[
\lim_{n \to \infty} \int_{\Omega_n} \phi_n(B_n) = \lim \sup_{n \to \infty} \int_{\Omega_n} \chi_{A_n(\lambda)}(\omega) dP_n(\omega) = 1. \tag{71}
\]
Using (69) and (71) observe that for $n$ large enough
\[
\lim \sup_{\varepsilon \to 0} \limsup_{n \to \infty} \frac{1}{n} D_\nu^\gamma(P_n||Q_n) \leq \lim \sup_{n \to \infty} \frac{1}{n} \log \sup_{B_n \subseteq A_n} \frac{\phi_n(B_n)}{Q_n(B_n)} \leq \lambda,
\]
where $a$ follows from (68) and (69).

We now prove the other side, i.e.,
\[
\lim \sup_{\varepsilon \to 0} \limsup_{n \to \infty} \frac{1}{n} D_\nu^\gamma(P_n||Q_n) \geq \bar{I}(P; Q).
\]
Consider any $\gamma < \bar{I}(P; Q)$. For every $n \in \mathbb{N}$, let $A_n(\gamma)$ be the union of sets in $F_n$ on which
\[
\frac{1}{n} \log \frac{dP_n}{dQ_n} \geq \gamma. \tag{72}
\]
From (32) it follows that there exists $\eta \in (0, 1]$, such that
\[
\limsup_{n \to \infty} \Pr\{A_n(\gamma)\} = \eta. \tag{73}
\]
Since $\int_{\Omega_n} \chi_{A_n(\gamma)}(\omega) dP_n(\omega) + \int_{\Omega_n} \chi_{A_n(\gamma)}(\omega) dP_n(\omega) = 1$, for every $n \in \mathbb{N}$, we have
\[
\liminf_{n \to \infty} \int_{\Omega_n} \chi_{A_n(\gamma)}(\omega) dP_n(\omega) = 1 - \eta. \tag{74}
\]
For every $\varepsilon \in (0, \eta)$, let us define a sequence of measure $\{\phi_n\}_{n=1}^\infty$, such that for every $n \in \mathbb{N}$
\[
\phi_n(B_n) \leq P_n(B_n) \forall B_n \in F_n \text{ and } \int_{\Omega_n} d\phi_n(\omega) \geq 1 - \varepsilon, \tag{75}
\]
and $\phi_n$ for every $n \in \mathbb{N}$ is defined on $F_n$, the sigma algebra generated by $\Omega_n$.

For every $n \in \mathbb{N}$, we have
\[
1 - \varepsilon \leq \int_{\Omega_n} \chi_{A_n(\gamma)}(\omega) d\phi_n(\omega) + \int_{\Omega_n} \chi_{A_n(\gamma)}(\omega) dP_n(\omega) \leq \int_{\Omega_n} \chi_{A_n(\gamma)}(\omega) d\phi_n(\omega),
\]
where $a$ follows from the fact that for every $B_n \in F_n$, $\phi_n(B_n) \leq P_n(B_n)$. By rearranging the terms in the above equation we get
\[
1 - \varepsilon - \int_{\Omega_n} \chi_{A_n(\gamma)}(\omega) dP_n(\omega) \leq \int_{\Omega_n} \chi_{A_n(\gamma)}(\omega) d\phi_n(\omega).
\]
Taking $\lim sup$ on both sides of the above equation, we have
\[
\limsup_{n \to \infty} \int_{\Omega_n} \chi_{A_n(\gamma)}(\omega) d\phi_n(\omega) \geq 1 - \varepsilon - \liminf_{n \to \infty} \int_{\Omega_n} \chi_{A_n(\gamma)}(\omega) dP_n(\omega) \geq \eta - \varepsilon. \tag{76}
\]
(76) follows from (74). Now notice the following set of inequalities for large enough $n$
\[
1 \geq \int_{\Omega_n} \chi_{A_n(\gamma)}(\omega) dP_n(\omega) \geq 2^{n\gamma} \int_{\Omega_n} \chi_{A_n(\gamma)}(\omega) dQ_n(\omega) \geq 2^{n\gamma} \sup_{B_n \subseteq A_n(\gamma)} \frac{\phi_n(B_n)}{Q_n(B_n)} \int_{\Omega_n} \chi_{A_n(\gamma)}(\omega) d\phi_n(\omega) \tag{77}
\]
where $a$ follows from (72); $b$ follows from the fact that for every $B_n \in F_n$, which is a subset of $A_n(\gamma)$,
\[
\frac{\phi_n(B_n)}{Q_n(B_n)} \leq \sup_{B_n \subseteq A_n(\gamma)} \frac{\phi_n(B_n)}{Q_n(B_n)} \leq \sup_{B_n \subseteq A_n(\gamma)} \frac{\phi_n(B_n)}{Q_n(B_n)}.
\]
By taking $\log$ on both sides of (77) and rearranging the terms we get
\[
\sup_{B_n \subseteq A_n(\gamma)} \frac{1}{n} \log \frac{\phi_n(B_n)}{Q_n(B_n)} \geq \gamma + \frac{1}{n} \log \int_{\Omega_n} \chi_{A_n(\gamma)}(\omega) d\phi_n(\omega).
Taking $\limsup$ on both sides of the above equation we have

\[
\limsup_{n \to \infty} \sup_{B_n \in F} \frac{1}{n} \log \frac{\phi_n(B_n)}{Q_n(B_n)} \geq \gamma + \limsup_{n \to \infty} \frac{1}{n} \log \int_{\Omega_n} \chi_{A_n(\gamma)}(\omega) d\phi_n(\omega) \\
\geq \gamma, \tag{78}
\]

where (78) follows from (76). Notice that (78) is true for every $\phi_n$ satisfying (75). Thus

\[
\limsup_{n \to \infty} \frac{1}{n} D^\varepsilon_\infty(P_n||Q_n) \geq \gamma. \tag{79}
\]

Since (79) is true for every $\varepsilon \in (0, \eta)$, the result will hold true for $\varepsilon \downarrow 0$.

(5) easily follows from the law of large numbers and (32). This completes the proof.

REFERENCES