Information Theoretic Methods in Statistics and Computer Science
Instructor: Yury Polyanskiy
Lecture Date: 04.12.2019

Scribe: Reka Inovan

1 Strong Data Processing Inequality and Applications II

1.1 Input Dependent Contraction Coefficient

Previously we have defined contraction coefficient $\eta_f(P_{Y|X})$, as the maximum contraction of an f-divergences over all channel input distributions. We now define an analogous concept for a specific input distribution P_X.

Definition 1 (Input Dependent Contraction Coefficient)**.** For any input distribution P_X, Markov kernel $P_{Y|X}$ and convex function f, we define

$$\eta_f(P_X, P_{Y|X}) \triangleq \sup_{Q_X: Q_X \neq P_X} \frac{D_f(Q_Y||P_Y)}{D_f(Q_X||P_X)}$$

where $Q_Y = P_{Y|X}Q_X$.

We refer to $\eta_f(P_X, P_{Y|X})$ as the input dependent contraction coefficient, to contrast it with the input independent contraction coefficient $\eta_f(P_{Y|X})$.
Remarks:

- As for $\eta_{KL}(P_{Y|X})$, we also have a corresponding mutual information characterization of $\eta_{KL}(P_X, P_{Y|X})$ as
$$\eta_{KL}(P_X, P_{Y|X}) = \sup_{P_{U|X}: U \to X \to Y} \frac{I(U;Y)}{I(U;X)}.$$
- From the definition, the following inequality holds
$$\eta_f(P_X, P_{Y|X}) \leq \eta_f(P_{Y|X}).$$
- Although we have the equality $\eta_{KL}(P_{Y|X}) = \eta_{\chi^2}(P_{Y|X})$ when $P_{Y|X}$ is a BMS channel, we do not have the same equality for $\eta_{KL}(P_X, P_{Y|X})$.

Example 1. ($\eta_{KL}(P_X, P_{Y|X})$ for Erasure Channel) We define EC_τ as the following channel,

$$Y = \begin{cases} X & \text{w.p. } 1-\tau \\ ? & \text{w.p. } \tau. \end{cases}$$

Let us define an auxiliary random variable $B = \mathbb{1}\{Y =?\}$. Thus we have the following equality,

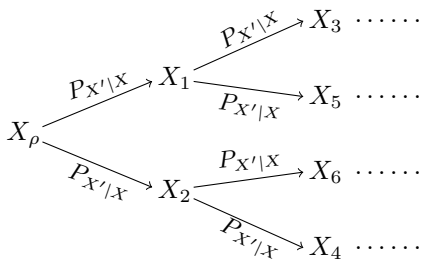
$$I(U;Y) = I(U;Y,B) = \underbrace{I(U;B)}_{0, B \perp\!\!\!\perp U} + I(U;Y|B) = (1-\tau)I(U;X).$$

where the last equality is due to the fact that $I(U;Y|B=1) = 0$ and $I(U;Y|B=0) = I(U;X)$. By the mutual information characterization of $\eta_{KL}(P_X, P_{Y|X})$, we have $\eta_{KL}(P_X, EC_\tau) = 1-\tau$.

The $\eta_{KL}(P_X, EC_\tau)$ is an important quantity. Because in SDPI arguments, we frequently reduces a model into erasure channels by channel degradation argument. As we will see in the subsequent sections.

1.2 Broadcasting on Trees

Consider an infinite b-ary tree $G = (\mathcal{V}, \mathcal{E})$. We assign a random variable X_v for each $v \in \mathcal{V}$. These random variables X_v's are defined on the same alphabet $\mathcal{X}$. In this model, the joint distribution is induced by the distribution on the root vertex π, i.e., $X_\rho \sim \pi$, and the edge kernel $P_{X'|X}$, i.e. $\forall (p, c) \in \mathcal{E}, P_{X_c|X_p} = P_{X'|X}$.



To simplify our discussion, we will assume that π is a reversible measure on kernel $P_{X'|X}$, i.e.,

$$P_{X'|X}(a|b)\pi(b) = P_{X'|X}(b|a)\pi(a).$$

By standard result on Markov chain, this also implies that π is a stationary distribution of the reversed Markov kernel $P_{X|X'}$.

Remarks:

- We can think of this model as a broadcasting scenario, where the root broadcasts its message X_ρ to the leaves through noisy channels $P_{X'|X}$.
- This model arises frequently in community detection, sparse codes and statistical physics.
- The joint distribution of this tree can be written as a Gibbs distribution

$$P_{X_{all}} = \frac{1}{Z} \exp\left(\sum_{(p,c)\in\mathcal{E}} f(X_p, X_c) + \sum_{v\in\mathcal{V}} g(X_v) \right)$$

for a certain f, g, and Z. When $\mathcal{X} = \{0, 1\}$, this model is equivalent to the Ising model.

We can define a corresponding inference problem, where we want to reconstruct the root variable X_ρ given the observations $X_{L_d} = \{X_v : v \in L_d\}$, with $L_d = \{v : v \in \mathcal{V}, \text{depth}(v) = d\}$. A natural question is to upper bound the performance of any inference algorithm on this problem. The following theorem shows that there exists a phase transition depending on the branching factor b and the contraction coefficient of the kernel $P_{X'|X}$.

Theorem 1. *Consider the broadcasting problem on infinite b-ary tree ($b > 1$), with root distribution π and edge kernel $P_{X'|X}$. If π is a reversible measure of $P_{X'|X}$ such that*

$$\eta_{KL}(\pi, P_{X'|X})b < 1,$$

then $I(X_\rho; X_{L_d}) \to 0$ as $d \to 0$.

Proof. For every $v \in L_1$, we define the set $L_{d,v} = \{u : u \in L_d, v \in \text{ancestor}(u)\}$. We can upper bound the mutual information between the root vertex and leaves at depth d

$$I(X_\rho; X_{L_d}) \le \sum_{v\in L_1} I(X_\rho; X_{L_{d,v}}).$$

For each term in the summation, we consider the Markov chain

$$X_{L_{d,v}} \to X_v \to X_\rho.$$

Due to our assumption on π and $P_{X'|X}$, we have $P_{X_\rho|X_v} = P_{X'|X}$ and $P_{X_v} = \pi$. By the definition of the contraction coefficient, we have

$$I(X_{L_{d,v}}; X_\rho) \le \eta_{KL}(\pi, P_{X'|X}) I(X_{L_{d,v}}; X_v).$$

Observe that because $P_{X_v} = \pi$ and all edges have the same kernel, then $I(X_{L_{d,v}}; X_v) = I(X_{L_{d-1}}; X_\rho)$. This gives us the inequality

$$I(X_\rho; X_{L_d}) \le \eta_{KL}(\pi, P_{X'|X}) b I(X_\rho; X_{L_{d-1}}),$$

which implies

$$I(X_\rho; X_{L_d}) \le (\eta_{KL}(\pi, P_{X'|X}) b)^d H(X_\rho).$$

Therefore if $\eta_{KL}(\pi, P_{X'|X}) b < 1$ then $I(X_\rho; X_{L_d}) \to 0$ exponentially fast as $d \to \infty$. □

Remarks: Another version of this theorem for $\eta_{KL}(P_{X'|X}) b \le 1$ is implied by the directed information percolation theorem.

Example 2. (Broadcasting on BSC tree.) Consider a broadcasting problem on b-ary tree with vertex alphabet $\mathcal{X} = \{0, 1\}$, edge kernel $P_{X'|X} = BSC_\delta$, and $\pi = Unif$. Note that uniform distribution is a reversible measure for BSC_δ. In the previous lecture, we calculated $\eta_{KL}(BSC_\delta) = (1 - 2\delta)^2$. Therefore, using theorem 1, we can deduce that if

$$b(1 - 2\delta)^2 < 1$$

then no inference algorithm can recover the root nodes as depth of the tree goes to infinity. This result is originally proved in [BRZ95].

Example 3 (k-coloring on tree)**.** Given a b-ary tree, we assign a k-coloring $X_{v_{all}}$ by sampling uniformly from the ensemble of all valid k-coloring. For this model, we can define a corresponding inference problem, namely given all the colors of the leaves at a certain depth, i.e., X_{L_d}, determine the color of the root node, i.e., X_ρ.

This problem can be modeled as a broadcasting problem on tree where the root distribution π is given by the uniform distribution on k colors, and the edge kernel $P_{X'|X}$ is defined as

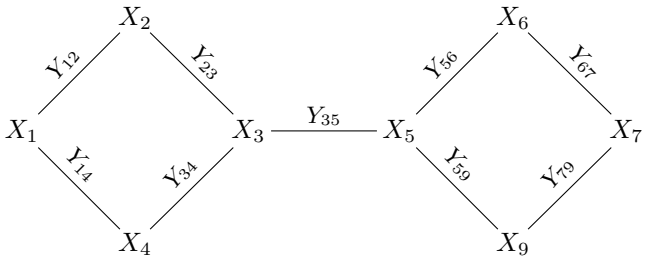
$$P_{X'|X}(a|b) = \begin{cases} \frac{1}{k-1} & a \neq b \\ 0, & a = b. \end{cases}$$

It has been shown in [GP19], that $\eta_{KL}(Unif, P_{X'|X}) = \frac{1}{k \log k(1+o(1))}$. By theorem 1, this implies that if $b < k \log k(1 + o(1))$ then reliable reconstruction of the root node is not possible. This result is originally proved in [Sly09] and [Bha+11]

The other direction $b > k \log k(1 + o(1))$ can be shown by observing that if $b > k \log k(1 + o(1))$ then the probability of the children of a node taking all available colors is close to 1. Hence, an inference algorithm can always determine the color of a node by finding a color that is not assigned to any of its children. In this regime, this reconstruction algorithm will succeed with high probability.

1.3 Undirected Information Percolation

In this section we will study the problem of inference on undirected graph. Consider an undirected graph $G = (\mathcal{V}, \mathcal{E})$. We assign a random variable X_v on the alphabet $\mathcal{X}$ to each vertex v. For each $e = (u, v) \in \mathcal{E}$, we assign Y_e sampled according to the kernel $P_{Y_e|X_e}$ with $X_e = (X_u, X_v)$. The goal of this inference model is to determine the value of X_v's given the value of Y_e's.



Example 4 (Community Detection)**.** In this model, we consider a complete graph with n vertices, i.e. K_n, and the random variables X_v representing the membership of each vertex to one of the m communities. We assume that X_v is sampled uniformly from $[m]$ and independent of the other vertices. The observation $Y_{u,v}$ is defined as

$$Y_{uv} \sim \begin{cases} \text{Ber}(a/n) & X_u = X_v \\ \text{Ber}(b/n) & X_u \neq X_v. \end{cases}$$

Example 5 ($\mathbb{Z}_2$ Synchronization)**.** For any graph G, we sample X_v uniformly from $\{-1, +1\}$ and $Y_e = \text{BSC}_\delta(X_u X_v)$.

Example 6 (Spiked Wigner Model)**.** We consider the inference problem of determining the value of vector $(X_i)_{i\in[n]}$ given the observation $(Y_{ij})_{i,j\in[n],i\leq j}$. The X_i's and Y_{ij}'s are related by a linear model

$$Y_{ij} = \sqrt{\frac{\lambda}{n}} X_i X_j + W_{ij},$$

where the value of X_i is sampled uniformly from $\{-1, +1\}$ and $W_{ij} \sim N(0, 1)$. This model can also be written in matrix form as

$$\mathbf{Y} = \mathbf{X}\mathbf{X}^T \sqrt{\frac{\lambda}{n}} + \mathbf{W}$$

where $\mathbf{W}$ is the Wigner matrix, hence the name of the model.

This problem can also be treated as a problem of inference on undirected graph. In this case, the underlying graph is a complete graph, and we assign X_i to each vertex. Under this model, the edge observations is given by $Y_{ij} = BIAWGN_{\lambda/n}(X_i X_j)$.

Although seemingly different, these problems share similar characteristics, namely:

- X_i's are uniformly distributed,
- If we define an auxiliary random variable $B = \mathbb{1}\{X_u \neq X_v\}$ for any edges $e = (u, v)$, then the following Markov chain holds
$$(X_u, X_v) \to B \to Y_e.$$
In other words, the observation on each edge only depends on whether the random variables on its endpoints are similar.

We will refer to the problem which have this characteristics as the Special Case (S.C.). Due to S.C.,the reconstructed X_v's is symmetric up to any permutation on $\mathcal{X}$. In the case of alphabet $\mathcal{X} = \{-1, +1\}$, this implies that for any realization σ then $P_{X_{\text{all}}|Y_{\text{all}}}(\sigma|b) = P_{X_{\text{all}}|Y_{\text{all}}}(-\sigma|b)$. Consequently, our reconstruction metric also needs to accommodate this symmetry. For $\mathcal{X} = \{-1, +1\}$, this leads to the use of $\frac{1}{n}|\sum_{i=1}^n X_i \hat{X}_i|$ as our reconstruction metric.

Our main theorem for undirected inference problem can be seen as the analogue of the information percolation theorem for DAG. However, instead of controlling the contraction coefficient, the percolation probability is used to directly control the conditional mutual information between any subsets of vertices in the graph.

Before stating our main theorem, we will need to define the corresponding percolation model for inference on undirected graph. For any undirected graph $G = (\mathcal{V}, \mathcal{E})$ we define a percolation model on this graph as follows :

- Every edge $e \in \mathcal{E}$ is open with the probability $\eta_{KL}(P_{Y_e|X_e})$, independent of the other edges,
- For any $v \in \mathcal{V}$ and $S \subset \mathcal{V}$, we define the $v \leftrightarrow S$ as the event that there exists an open path from v to any vertex in S,
- For any $S_1, S_2 \subset \mathcal{V}$, we define the function $\text{perc}_u(S_1, S_2)$ as

$$\text{perc}_u(S_1, S_2) \triangleq \sum_{v \in S_1} P(v \leftrightarrow S_2).$$

  Notice that this function is different from the percolation function for information percolation in DAG. Most importantly, this function is not equivalent to the exact percolation probability. Instead, it is an upper bound on the percolation probability by union bounding with respect to S_1. Hence, it is natural that this function is not symmetric, i.e. $\text{perc}_u(S_1, S_2) \neq \text{perc}_u(S_2, S_1)$.

Theorem 2 (Undirected Information Percolation)**.** *Consider an inference problem on undirected graph $G = (\mathcal{V}, \mathcal{E})$. For any $S_1, S_2 \subset \mathcal{V}$, then*

$$I(X_{S_1}; X_{S_2}|Y) \leq perc_u(S_1, S_2) \log |\mathcal{X}|.$$

The following theorem shows how the undirected information percolation concept allows us to derive a converse result for spiked Wigner model.

Theorem 3. *Consider the spiked Wigner model. If $\lambda \leq 1$, then for any sequence of estimator $\hat{X^n}(Y)$,*

$$\frac{1}{n} E\left[\left|\sum_{i=1}^{n} X_i \hat{X}_i\right|\right] \to 0$$

as $n \to \infty$.

Proof of Theorem 3. First of all, we observe that because spiked Wigner model fulfills the S.C. condition, then there is an inherent symmetry of the solution up to a global flip. Without loss of generality, we take $X_1 = 1$ to break the symmetry.

Due to this choice, the optimal estimator for this problem is equal to

$$\hat{X}_j(y) = \text{argmax}_{\sigma \in \{-1,+1\}} P_{X_j|Y,X_1}(\sigma|y, 1).$$

In our case $I(X_i; X_1, Y) = I(X_i; X_1|Y)$, as $I(X_i; Y) = 0$ due to the symmetry up to a global flip. In other words, it suffices to show that if $I(X_i; X_1|Y) \to 0$ then no reliable reconstruction is possible. Furthermore, by symmetry of the problem, for any $i \neq 1$ then $I(X_i; X_1|Y) = I(X_2; X_1|Y)$.

By using the undirected information percolation theorem, we have

$$I(X_2; X_1|Y) \leq \text{perc}_u(\{1\}, \{2\})$$

in which the percolation model is defined on a complete graph with edge probability λ/n as $\eta_{KL}(BIAWGN_{\lambda/n}) = \frac{\lambda}{n}(1 + o(1))$. This percolation random graph is equivalent to the Erdős-Rényi random graph with n vertices and λ/n edge probability, i.e., $ER(n, \lambda/n)$. Using this observation, the inequality can be rewritten as

$$I(X_2; X_1|Y) \leq P(\text{Vertex 1 and 2 is connected on } ER(n, \lambda/n)).$$

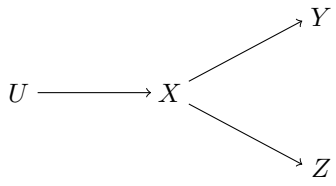
The largest components on $ER(n, \lambda/n)$ contains $O(n^{2/3})$ if $\lambda \leq 1$. This implies that the probability that two specific vertices are connected is $o(1)$, hence $I(X_2; X_1|Y) \to 0$ as $n \to \infty$. □

Remarks: This reduction changes the underlying structure of the graph. Instead of dealing with a complete graph, the percolation problem is defined on an Erdős-Rényi random graph. Moreover, if

η_{KL} is small enough, then the underlying percolation graph tends to have a locally tree-like structure.

Instead of proving theorem 2 in its full generality, we will prove the theorem under S.C. condition. The main step of the proof utilizes the fact we can upper bound the mutual information of any channel by its degraded channel. To this end, we will define the less noisy partial ordering on the channels.

Definition 2 (Less Noisy Ordering)**.** We define $P_{Y|X} \leq_{LN} P_{Z|X}$ iff for every $P_{U,X}$ on the following Markov chain



we have $I(U;Y) \leq I(U;Z)$.

Remarks: We also have the equivalent definition in terms of the divergence, namely $P_{Y|X} \leq_{LN} P_{Z|X}$ if and only if for all P_X, Q_X we have $D(Q_Y||P_Y) \leq D(Q_Z||P_Z)$.

Proposition 1. *$\eta_{KL}(P_{Y|X}) \leq 1-\tau$ if and only if $P_{Y|X} \leq_{LN} EC_\tau$.*

Proof. For EC_τ we always have

$$I(U;Z) = (1-\tau)I(U;X).$$

By the mutual information characterization of η_{KL} we have,

$$I(U;Y) \leq (1-\tau)I(U;X).$$

Combining these two inequalities gives us

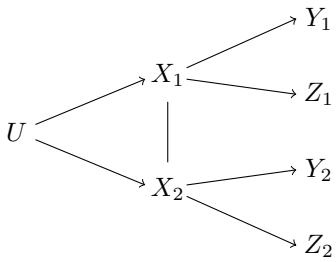
$$I(U;Y) \leq I(U;Z).$$

□

Remarks: This proposition gives us an intuitive interpretation of contraction coefficient as the worst erasure channel that still dominates the channel.

Proposition 2. *(Tensorization of Less Noisy Ordering) If for all $i \in [n], P_{Y_i|X_i} \leq_{LN} P_{Z_i|X_i}$, then $P_{Y_1|X_1} \otimes P_{Y_2|X_2} \leq_{LN} P_{Z_1|X_1} \otimes P_{Z_2|X_2}$. Note that $P \otimes Q$ refers to the product channel of P and Q.*

Proof. Consider the following Markov chain.



It can be seen from the Markov chain that $I(U;Y_1,Y_2) \leq I(U;Y_1,Z_2)$ implies $I(U;Y_1,Y_2) \leq I(U;Z_1,Z_2)$. Consider the following inequalities,

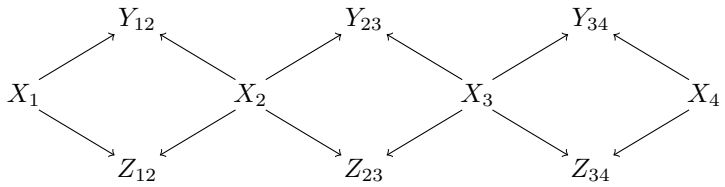
$$\begin{aligned} I(U;Y_1,Y_2) &= I(U;Y_1) + I(U;Y_2|Y_1) \\ &\leq I(U;Y_1) + I(U;Z_2|Y_1) \\ &= I(U;Y_1,Z_2). \end{aligned}$$

Hence $I(U;Y_1,Y_2) \leq I(U;Y_1,Z_2)$ for any $P_{X_1,X_2,U}$. □

Theorem 4. *Consider the problem of inference on undirected graph $G = (\mathcal{V},\mathcal{E})$ with $X_1,...,X_n$ are not necessarily independent. If $P_{Y_e|X_e} \leq_{LN} P_{Z_e|X_e}$, then for any $S_1,S_2 \subset \mathcal{V}$ and $E \subset \mathcal{E}$*

$$I(X_{S_1};Y_E|X_{S_2}) \leq I(X_{S_1};Z_E|X_{S_2})$$

Proof. Consider the following Markov chain.



From our assumption and the tensorization property of less noisy ordering, we have $P_{Y_E|X_{S_1},X_{S_2}} \leq_{LN} P_{Z_E|X_{S_1},X_{S_2}}$. This implies that for σ as a valid realization of X_{S_2} we will have

$$I(X_{S_1};Y_E|X_{S_2}=\sigma) = I(X_{S_1},X_{S_2};Y_E|X_{S_2}=\sigma) \leq I(X_{S_1},X_{S_2};Z_E|X_{S_2}=\sigma) = I(X_{S_1};Z_E|X_{S_2}=\sigma).$$

As this inequality holds for all realization of X_{S_2}, then the following inequality also holds

$$I(X_{S_1};Y_E|X_{S_2}) \leq I(X_{S_1};Z_E|X_{S_2}).$$

□

Using these results, we can give a proof for our main theorem under S.C. conditions.

Proof of Theorem 2. Under S.C. conditions, we have the following equalities for any $i \in S_1$

$$I(X_i;X_{S_2}|Y_E) = I(X_i;X_{S_2},Y_E) = I(X_i;Y_E|X_{S_2}) \tag{1.1}$$

where the first inequality is due to the fact $B_E \perp\!\!\!\perp X_i$ under S.C, and the second inequality is due to $X_i \perp\!\!\!\perp X_{S_2}$ under S.C.

Due to our previous result, if $\eta_{KL}(P_{Y_e|X_e}) = 1-\tau$ then $P_{Y_e|X_e} \leq_{LN} P_{Z_e|X_e}$ where $P_{Z_e|X_e} = EC_\tau$. By tensorization property, this ordering also holds for the channel $P_{Y_E|X_E}$, thus we have

$$I(X_i;Y_E|X_{S_2}) \leq I(X_j;Z_E|X_{S_2}).$$

Let us define another auxiliary random variable $D = \mathbb{1}\{i \leftrightarrow S_2\}$, namely it is the indicator that there is an open path from i to S_2. Notice that D is fully determined by Z_E. By the same argument as in (1.1), we have

$$\begin{aligned} I(X_i;Z_E|X_{S_2}) &= I(X_i;X_{S_2}|Z_E) \\ &= I(X_i;X_{S_2}|Z_E,D) \\ &= (1-P(i \leftrightarrow S_2)) \underbrace{I(X_i;X_{S_2}|Z_E,D=0)}_{0} + P(i \leftrightarrow S_2) \underbrace{I(X_i;X_{S_2}|Z_E,D=1)}_{\leq \log|\mathcal{X}|} \\ &\leq P(i \leftrightarrow S_2) \log|\mathcal{X}|. \end{aligned}$$

Summing over all the elements of S_1 gives us

$$I(X_{S_1}; Z_E | X_{S_2}) \leq \sum_{i \in S_1} I(X_i; Z_E | X_{S_2}) \leq \log |\mathcal{X}| \sum_{i \in S_1} P(i \leftrightarrow S_2) = \mathrm{perc}_u(S_1, S_2) \log |\mathcal{X}|.$$

□

References

[Bha+11] Nayantara. Bhatnagar, Juan. Vera, Eric. Vigoda, and Dror. Weitz. "Reconstruction for Colorings on Trees". In: *SIAM Journal on Discrete Mathematics* 25.2 (2011), pp. 809–826. DOI: 10.1137/090755783. eprint: https://doi.org/10.1137/090755783. URL: https://doi.org/10.1137/090755783.

[BRZ95] P. M. Bleher, J. Ruiz, and V. A. Zagrebnov. "On the purity of the limiting gibbs state for the Ising model on the Bethe lattice". In: *Journal of Statistical Physics* 79.1 (1995), pp. 473–482. ISSN: 1572-9613. DOI: 10.1007/BF02179399. URL: https://doi.org/10.1007/BF02179399.

[GP19] Y. Gu and Y. Polyanskiy. *Nonlinear log-Sobolev inequalities for the Potts channel with applications to reconstruction problems*. draft. May 2019.

[Sly09] Allan Sly. "Reconstruction of Random Colourings". In: *Communications in Mathematical Physics* 288.3 (2009), pp. 943–961. ISSN: 1432-0916. DOI: 10.1007/s00220-009-0783-7. URL: https://doi.org/10.1007/s00220-009-0783-7.