Electronic Colloquium on Computational Complexity, Report No. 194 (2016)

The Optimality of Correlated Sampling

Mohammad Bavarian[*] Badih Ghazi[†] Elad Haramaty[‡] Pritish Kamath[§]
Ronald L. Rivest[¶] Madhu Sudan[‖]

December 3, 2016

Abstract

In the *correlated sampling* problem, two players, say Alice and Bob, are given two distributions, say P and Q respectively, over the same universe and access to shared randomness. The two players are required to output two elements, without any interaction, sampled according to their respective distributions, while trying to minimize the probability that their outputs disagree. A well-known protocol due to Holenstein, with close variants (for similar problems) due to Broder, and to Kleinberg and Tardos, solves this task with disagreement probability at most $2\delta/(1+\delta)$, where δ is the total variation distance between P and Q. This protocol has been used in several different contexts including sketching algorithms, approximation algorithms based on rounding linear programming relaxations, the study of parallel repetition and cryptography.

In this note, we give a surprisingly simple proof that this protocol is in fact tight. Specifically, for every $\delta \in (0,1)$, we show that any correlated sampling scheme should have disagreement probability at least $2\delta/(1+\delta)$. This partially answers a recent question of Rivest.

Our proof is based on studying a new problem we call *constrained agreement*. Here, Alice is given a subset $A \subseteq [n]$ and is required to output an element $i \in A$, Bob is given a subset $B \subseteq [n]$ and is required to output an element $j \in B$, and the goal is to minimize the probability that $i \neq j$. We prove tight bounds on this question, which turn out to imply tight bounds for correlated sampling. Though we settle basic questions about the two problems, our formulation also leads to several questions that remain open.

[*] Department of Mathematics and Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge MA 02139. Supported in part by NSF Award CCF-1420692. `bavarian@mit.edu`.

[†] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge MA 02139. Supported in part by NSF CCF-1420956, NSF CCF-1420692 and CCF-1217423. `badih@mit.edu`.

[‡] Harvard John A. Paulson School of Engineering and Applied Sciences. Part of this work supported by NSF Award CCF-1565641. `seladh@gmail.com`.

[§] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge MA 02139. Supported in part by NSF CCF-1420956 and NSF CCF-1420692. `pritish@mit.edu`.

[¶] Institute Professor, MIT. This work supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

[‖] Harvard John A. Paulson School of Engineering and Applied Sciences. Part of this work supported by NSF Award CCF-1565641 and a Simons Investigator Award. `madhu@cs.harvard.edu`.

ISSN 1433-8092

1 Introduction

In this work, we study *correlated sampling*, a very basic task, variants of which have been considered in the context of sketching algorithms [Bro97], approximation algorithms based on rounding linear programming relaxations [KT02, Cha02], the study of parallel repetition [Hol07, Rao11, BHH+08] and very recently cryptography [Riv16].

This problem involves two players, Alice and Bob, attempting to come to agreement non-interactively. Alice and Bob are given distributions P and Q respectively over the same universe Ω. Without any interaction, Alice is required to output an element $i \sim P$ and Bob is required to output an element $j \sim Q$, where the players have access to shared randomness. The goal is to minimize the disagreement probability $\Pr[i \neq j]$ in terms of the total variation distance $d_{\mathrm{TV}}(P, Q)$ (where the probability is over the shared randomness). More formally we define *correlated sampling strategies* as follows.

Definition 1.1 (Correlated Sampling Strategies). *Given a universe*[1] *Ω and a randomness space $\mathcal{R}$, a pair of functions (f, g), where $f : \Delta_\Omega \times \mathcal{R} \to \Omega$ and $g : \Delta_\Omega \times \mathcal{R} \to \Omega$, is said to be a* correlated sampling strategy *with error $\varepsilon : [0, 1] \to [0, 1]$, if for any distribution $P, Q \in \Delta_\Omega$, such that $d_{\mathrm{TV}}(P, Q) = \delta$, it holds that,*

- **[Correctness]** $\{f(P, r)\}_{r\sim\mathcal{R}} = P$ *and* $\{g(Q, r)\}_{r\sim\mathcal{R}} = Q$
- **[Error guarantee]** $\Pr_{r\sim\mathcal{R}} [f(P, r) \neq g(Q, r)] \leq \varepsilon(\delta)$

Here, Δ_Ω is the set of all probability distributions on Ω. Also, we abuse notations slightly to let $\mathcal{R}$ denote a suitable distribution on the set $\mathcal{R}$. Moreover, we will always assume that $\mathcal{R}$ is sufficiently large, and we will often not mention $\mathcal{R}$ explicitly, when talking about correlated sampling strategies. It is also allowed to have a sequence of strategies with increasing size of $\mathcal{R}$, in which case, we want the above constraints to be satisfied in the limit as $|\mathcal{R}| \to \infty$.

A priori it is unclear whether such a protocol can even exist, since the error ε is not allowed to depend on the universe Ω. Somewhat surprisingly, there exists a simple protocol whose disagreement probability can be bounded by roughly twice the total variation distance (and in particular does not degrade with the size of the universe). Variants of this protocol have been rediscovered multiple times in the literature yielding the following theorem.

Theorem 1.2 (Holenstein [Hol07]. See also Broder [Bro97], Kleinberg-Tardos [KT02]). *For any universe Ω, there exists a correlated sampling strategy with error $\varepsilon : [0, 1] \to [0, 1]$ such that,*

$$\forall \delta \in [0, 1] , \qquad \varepsilon(\delta) \leq \frac{2 \cdot \delta}{1 + \delta} \tag{1}$$

Strictly speaking, the work of Broder [Bro97] does not consider the general correlated sampling problem. Rather it gives a strategy (the "MinHash strategy") which happens to solve the correlated sampling problem under the condition that P and Q are *flat* distributions, i.e. they are uniform over some subset of the domain. The above bound applies to the case where these sets have the same size. The technique can also be generalized to other distributions to get the bound above, and this gives a protocol similar to that of Holenstein, though if P and Q are uniform over different sized subsets, the above bound is weaker than that obtained from a direct application of Broder's algorithm! Holenstein [Hol07] appears to be the first to formulate the problem for general distributions and give a solution with the bound claimed above.

For sake of completeness, we give a description of Broder's strategy as well as Holenstein's strategy in Section 3. We point out that variants of the protocol in Theorem 1.2 (sometimes referred to as "consistent sampling" protocols) had been used in several applied works [M+94, GP06, MMT10] before Holenstein's paper.

Given Theorem 1.2, a natural and basic question is whether the bound on the disagreement probability can be improved. Indeed, this question was very recently raised by Rivest [Riv16] in the context of symmetric encryption, and this was one of the motivations behind this work. We give a surprisingly simple proof that the bound in Theorem 1.2 is actually tight (for a coarse parameterization of the problem).

Theorem 1.3 (Main Result). *For every $\delta \in (0, 1)$ and $\gamma > 0$, there exists a family of pairs of distributions (P, Q) satisfying $d_{\mathrm{TV}}(P, Q) \leq \delta$ such that any correlated sampling strategy for this family has error at least $\frac{2\cdot\delta}{1+\delta} - \gamma$.*

[1] we will primarily consider only finite universes.

Our proof of Theorem 1.3 is surprisingly simple and is based on studying the following *constrained agreement* problem that we introduce and which is tightly related to correlated sampling. Alice is given a subset $A \subseteq [n]$ and Bob is given a subset $B \subseteq [n]$, where the pair (A, B) is sampled from some distribution $\mathcal{D}$. Alice is required to output an element $i \in A$ and Bob is required to output an element $j \in B$, such that the disagreement probability $\Pr_{(A,B)\sim D}[i \neq j]$ is minimized.

Definition 1.4 (Constrained Agreement Strategies)**.** *Given a universe $\Omega = [n]$ and a distribution $\mathcal{D}$ over $2^\Omega \times 2^\Omega$ (i.e. pairs of subsets of Ω), a pair of functions (f, g), where $f : 2^\Omega \to \Omega$ and $g : 2^\Omega \to \Omega$, is said to be a* constrained agreement strategy *with error $\mathrm{err}_{\mathcal{D}}(f, g) = \varepsilon \in [0, 1]$, if it holds that,*

- **[Correctness]** $\forall A \subseteq \Omega$, $f(A) \in A$ *and* $\forall B \subseteq \Omega$, $g(B) \in B$
- **[Error guarantee]** $\Pr_{(A,B)\sim\mathcal{D}} [f(A) \neq g(B)] \leq \varepsilon$

We point out that since the constrained agreement problem is defined with respect to an input distribution $\mathcal{D}$ on pairs of sets, we can require, without loss of generality, that the strategies (f, g) be deterministic in Definition 1.4 (this follows from Yao's minimax principle). We arrive at the constrained agreement problem as follows: First we consider the *flat* distribution case of Definition 1.1 and relax the restrictions of $\{f(P, r)\}_{r\sim\mathcal{R}} = P$ and $\{g(Q, r)\}_{r\sim\mathcal{R}} = Q$, although, we still require that $f(P, r) \in \mathrm{supp}(P)$ and $g(Q, r) \in \mathrm{supp}(Q)$ for any $r \in \mathcal{R}$. This makes it a constraint satisfaction problem and we consider a distributional version of the same.

In order to prove Theorem 1.3, we show that in fact the correlated sampling strategy (a suitable derandomization thereof) as in Theorem 1.2 is optimal for the constrained agreement problem whenever $\mathcal{D}$ is the distribution $\mathcal{D}_p$ where every coordinate $i \in [n]$ is independently included in each of A and B with probability p.

Lemma 1.5. *For every $p \in [0, 1]$ and the distribution $\mathcal{D}_p$ on $2^{[n]} \times 2^{[n]}$, any constrained agreement strategy (f, g) makes error $\mathrm{err}_{\mathcal{D}_p}(f, g) \geq \frac{2(1-p)}{2-p}$.*

Organization of the paper. In Section 1.1, we discuss some special cases of correlated sampling problem. In Section 1.2, we give some open problems regarding these special cases. In Section 2, we prove Lemma 1.5 and use it to prove Theorem 1.3. In Section 3, we describe the correlated sampling protocols of Broder and Holenstein, thereby proving Theorem 1.2.

1.1 Special Cases

Let $A, B \subseteq [n]$ be such that $|A| = |B|$, and consider the problem of correlated sampling with the uniform distributions $P = \mathcal{U}(A)$ and $Q = \mathcal{U}(B)$. Then, the total variation distance between P and Q is given by

$$d_{\mathrm{TV}}(P, Q) = \frac{1}{2} \cdot \|P - Q\|_1 = \frac{1}{2} \cdot \|\mathcal{U}(A) - \mathcal{U}(B)\|_1 = 1 - \frac{|A \cap B|}{|A|}.$$

Thus, the error probability of the correlated sampling strategy (in Theorem 1.2) is given by

$$\frac{2 \cdot d_{\mathrm{TV}}(P, Q)}{1 + d_{\mathrm{TV}}(P, Q)} = 1 - \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

Rather surprisingly, in the particular case where $|A \cap B| = 1$ and $A \cup B = [n]$, Rivest [Riv16] recently gave a protocol with smaller error probability than the one guaranteed by the correlated sampling protocol of Theorem 1.2.

Theorem 1.6 ([Riv16])**.** *In Definition 1.1, if $\Omega = [n]$, and the distributions P and Q are promised to be of the following form, that there exist $A, B \subseteq [n]$ such that $|A| = |B|$, $|A \cap B| = 1$, $A \cup B = [n]$, and $P = \mathcal{U}(A)$ and $Q = \mathcal{U}(B)$. Then, there is a correlated sampling strategy with error probability at most $1 - 1/|A|$.*

For completeness, we describe this strategy in Section 3.1. Note that for this setting of parameters, we have that

$$1 - \frac{1}{|A|} < 1 - \frac{|A \cap B|}{|A \cup B|} = 1 - \frac{1}{n},$$

and hence Theorem 1.6 improves on the performance (eq. (2)) of the correlated sampling strategy of Theorem 1.2. This naturally leads to the question: Can one similarly improve on the well-known correlated sampling protocol for larger intersection sizes, for example, when $|A \cap B|$ is a constant fraction of $|A|$? The proof of our main result (Theorem 1.3) answers this question negatively. Namely, it implies that the strategy in Theorem 1.2 is tight when $|A \cap B| = \varepsilon \cdot |A|$ with $\varepsilon \in (0, 1)$ being an absolute constant.

Note that in the extreme case where ε is very close to 0, Rivest's protocol (Theorem 1.6) implies that Theorem 1.2 is not tight. What about the other extreme where ε is very close to 1? We show that in this case Theorem 1.2 is in fact tight.

Theorem 1.7. *Let $A, B \subseteq [n]$ be such that $A \cup B = [n]$ and $|A| = |B| = |A \cap B| + 1$, and let $P = \mathcal{U}(A)$ and $Q = \mathcal{U}(B)$. Then, the error probability of any correlated sampling strategy is at least $1 - |A \cap B|/|A \cup B|$.*

We prove Theorem 1.7 in Section 2.1.

1.2 Open questions and Future work

Our work started with a conjecture due to Rivest [Riv16] which informally asserts that Broder's MinHash strategy is optimal except in the case considered in Theorem 1.6. More formally,

Conjecture 1.8 (Rivest [Riv16]). *For every collection of positive integers n, a, b, ℓ with $\ell \geq 2$ and $n \geq a + b - \ell$, and for every pair of probabilistic strategies (f, g) that satisfy correctness as in Definition 1.1, there exist $A, B \subseteq [n]$ with $|A| = a$, $|B| = b$ and $|A \cap B| = \ell$ such that*

$$\Pr[f(A) \neq g(B)] \geq 1 - \frac{\ell}{a + b - \ell} = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Our work does not resolve this conjecture for the general setting of n, a, b and ℓ. It suggests this answer may be asymptotically right as $n \to \infty$ when $a = \alpha n$, $b = \beta n$ and $\ell = \alpha\beta n$, but does not even resolve this setting. (Our set sizes are only approximately αn etc.)

Even in the setting where the set sizes are allowed to vary slightly, our knowledge is somewhat incomplete. Lemma 1.5 shows optimality of the MinHash strategy when $(A, B) \sim \mathcal{D}_p$. In this case, A and B are independent and p-biased each, so $|A| \approx p \cdot n$, $|B| \approx p \cdot n$ and $|A \cap B| \approx p^2 \cdot n$. We point out that a simple reduction to Lemma 1.5 also implies the optimality of the well-known protocol in the case where A and B are "positively-correlated". Specifically, consider the following distribution $\mathcal{D}_{p,\delta}$ on pairs (A, B) of subsets of $[n]$, where we first sample $S \subseteq [n]$ which independently includes each element of $[n]$ with probability $p/(1 - \delta)$, and then independently including every $i \in S$ in each of A and B with probability $1 - \delta$. In this case, $|A| \approx p \cdot n$, $|B| \approx p \cdot n$ and $|A \cap B| \approx (1 - \delta) \cdot p \cdot n$. Even if we reveal S to both Alice and Bob, Lemma 1.5 implies a lower bound of $2 \cdot \delta/(1 + \delta)$ on the error probability, which is achieved by the MinHash strategy. It is not clear how to use a similar reduction to show optimality in the case where A and B are "negatively-correlated", i.e., when $|A| \approx p \cdot n$, $|B| \approx p \cdot n$ and $|A \cap B| \ll p^2 \cdot n$.

Finally the fact that Holenstein's strategy for correlated sampling can be improved upon in the case where P and Q are uniform distributions on different-sized subsets of the universe clearly shows that strategy as in Theorem 1.2 is not "always optimal". To study questions like this, one could restrict the class of pairs (P, Q) and then give an optimal strategy for every P and every Q. It would be interesting to study what would be the right measure that captures the minimal error probability given the adjacency relationship (P, Q).

2 Lower Bounds on Correlated Sampling

We start by proving lower bounds on error probability in the constrained agreement problem.

Proof of Lemma 1.5. Let $p \in [0, 1]$ and consider the distribution $\mathcal{D}_p$ on pairs (A, B) of subsets $A, B \subseteq [n]$ where for each $i \in [n]$, we independently inlcude i in each of A and B with probability p. Let f be Alice's strategy which satisfies the property that $f(A) \in A$ for every $A \subseteq [n]$. Similarly, let g be Bob's strategy which satisfies the property that $g(B) \in B$ for every $B \subseteq [n]$.

We will construct functions f^* and g^* such that

$$\mathrm{err}_{\mathcal{D}_p}(f, g) \geq \mathrm{err}_{\mathcal{D}_p}(f^*, g) \geq \mathrm{err}_{\mathcal{D}_p}(f^*, g^*) \geq \frac{2(1 - p)}{2 - p}$$

For every $i \in [n]$, we define $\beta_i \triangleq \Pr_B[g(B) = i]$. Since under the distribution $\mathcal{D}_p$, the subsets A and B are independent, we have that when Bob's strategy is fixed to g, the strategy of Alice that results in the largest agreement probability is given by

$$\forall A \subseteq [n], \quad f^*(A) = \operatorname*{argmax}_{i \in A} \beta_i$$

Thus, for a permutation σ of $[n]$ such that $\beta_{\sigma^{-1}(1)} \geq \beta_{\sigma^{-1}(2)} \geq \cdots \geq \beta_{\sigma^{-1}(n)}$, we have,

$$\forall A \subseteq [n], \quad f^*(A) = \operatorname*{argmin}_{i \in A} \sigma(i).$$

Now, for every $i \in [n]$, we define $\alpha_i \triangleq \Pr_A[f^*(A) = i]$. When Alice's strategy is fixed to f^*, the strategy of Bob that results in the largest agreement probability is given by

$$\forall B \subseteq [n], \quad g^*(B) = \operatorname*{argmax}_{i \in B} \alpha_i$$

We now claim that $\alpha_{\sigma^{-1}(1)} \geq \alpha_{\sigma^{-1}(2)} \geq \cdots \geq \alpha_{\sigma^{-1}(n)}$, and hence,

$$\forall B \subseteq [n], \quad g^*(B) = \operatorname*{argmin}_{i \in B} \sigma(i)$$

This follows easily because for each $i \in [n]$, we have that,

$$\alpha_i \;=\; \Pr_A\left[\left(\operatorname*{argmin}_{\ell \in A} \sigma(\ell)\right) = i\right] \;=\; (1-p)^{i-1} \cdot p$$

Thus, we conclude that

$$\begin{aligned}
\Pr_{(A,B)\sim\mathcal{D}_p}[f(A) = g(B)] &\leq \Pr_{(A,B)\sim\mathcal{D}_p}[f^*(A) = g(B)] \\
&\leq \Pr_{(A,B)\sim\mathcal{D}_p}[f^*(A) = g^*(B)] \\
&= \sum_{i=1}^{n} \Pr_{(A,B)\sim\mathcal{D}_p}[f^*(A) = g^*(B) = i] \\
&= \sum_{i=1}^{n} \Pr_A[f^*(A) = i] \cdot \Pr_B[g^*(B) = i] \\
&= \sum_{i=1}^{n} (1-p)^{2\cdot(i-1)} \cdot p^2 \\
&\leq \frac{p}{2-p},
\end{aligned}$$

where the second equality uses the fact that under $\mathcal{D}_p$, the subsets A and B are independent. Thus, we obtain that,

$$\mathrm{err}_{D_p}(f, g) \;\geq\; 1 - \frac{p}{2-p} \;=\; \frac{2(1-p)}{2-p}$$

□

We are now ready to prove our main result which is a lower bound on the error in correlated sampling.

Proof of Theorem 1.3. Let $\delta \in (0,1)$ and $\gamma > 0$. Assume for the sake of contradiction that there is a correlated sampling strategy (f^*, g^*) that, when run on distributions at total variation distance up to δ, has error probability at most $\frac{2\cdot\delta}{1+\delta} - \gamma$. Fix $\delta' \in (0,1)$ such that

$$\frac{2\cdot\delta}{1+\delta} - \gamma < \frac{2\cdot\delta'}{1+\delta'} < \frac{2\cdot\delta}{1+\delta}. \tag{3}$$

Note that Equation (3) implies that $\delta' < \delta$. Consider the distribution $\mathcal{D}_p$ over pairs (A, B) of subsets $A, B \subseteq [n]$ where each $i \in [n]$ is independently included in each of A and B with probability $p \triangleq 1 - \delta'$.

We then have that $\mathbb{E}[|A|] = \mathbb{E}[|B|] = p \cdot n$, and $\mathbb{E}[|A \cap B|] = p^2 \cdot n$. Moreover, by the Chernoff bound, we have that

$$\Pr_A[||A| - p \cdot n| > p \cdot n^{0.99}] \leq e^{-p \cdot n^{0.98}/2},$$

$$\Pr_B[||B| - p \cdot n| > p \cdot n^{0.99}] \leq e^{-p \cdot n^{0.98}/2},$$

and

$$\Pr_{A,B}[||A \cap B| - p^2 \cdot n| > p^2 \cdot n^{0.99}] \leq e^{-p^2 \cdot n^{0.98}/2}.$$

Hence, by the union bound and since $p \leq 1$, we get that with probability at least $1 - 3 \cdot e^{-p^2 \cdot n^{0.98}/2}$, we have that $||A| - p \cdot n| \leq pn^{0.99}$, $||B| - p \cdot n| \leq pn^{0.99}$ and $||A \cap B| - p^2 \cdot n| \leq p^2 n^{0.99}$. Consider now the distributions $P = \mathcal{U}(A)$ (on Alice's side) and $Q = \mathcal{U}(B)$ (on Bob's side). Then, with probability at least $1 - 3 \cdot e^{-p^2 \cdot n^{0.98}/2}$, it holds that,

$$\begin{aligned} d_{\mathrm{TV}}(P, Q) &= 1 - \frac{|A \cap B|}{\max\{|A|, |B|\}} \\ &\leq 1 - p + o_n(1) \\ &= \delta' + o_n(1) \\ &< \delta \quad \text{for sufficiently large } n. \end{aligned}$$

Note that, Yao's minimax principle implies that any correlated sampling strategy for (P, Q) pairs with $P = \mathcal{U}(A)$ and $Q = \mathcal{U}(B)$ yields a constrained agreement strategy (f, g) for the corresponding pairs (A, B) of subsets. Hence, Lemma 1.5 implies that

$$\forall\, f, g, \quad \Pr_{(A,B)\sim\mathcal{D}}[f(A) \neq g(B)] \geq \frac{2(1-p)}{2-p} = \frac{2 \cdot \delta'}{1 + \delta'} \tag{4}$$

where (f, g) is any correlated sampling strategies. On the other hand, the property of the assumed strategy (f^*, g^*) implies that

$$\exists\, f, g, \quad \Pr_{(A,B)\sim\mathcal{D}}[f(A) \neq g(B)] \leq \frac{2 \cdot \delta}{1 + \delta} - \gamma + o_n(1). \tag{5}$$

Putting Equations (4) and (5) together contradicts Equation (3) for sufficiently large n. □

2.1 Lower Bound in a Special Case

In this section, we describe the lower bound in Theorem 1.7, which is incomparable to Theorem 1.3.

Proof of Theorem 1.7. Let $A, B \subseteq [n]$ be such that $|A| = |B| = |A \cap B| + 1$ and let $P = \mathcal{U}(A)$ and $Q = \mathcal{U}(B)$. Assume for the sake of contradiction that there is a correlated sampling strategy with disagreement probability $< 1 - |A \cap B|/|A \cup B| = 2/n$. Let $\mathcal{D}$ be the uniform distribution over pairs (A, B) of subsets of $[n]$ satisfying $A \cup B = [n]$ and $|A| = |B| = |A \cap B| + 1$. Then, there is a deterministic strategy pair (f, g) solving constrained agreement over $\mathcal{D}$ with error probability

$$\Pr_{(A,B)\sim\mathcal{D}}[f(A) \neq g(B)] < \frac{2}{n}. \tag{6}$$

Let

$$i \overset{\text{def}}{=} \operatorname*{argmax}_{\ell \in [n]} \left| \left\{ A \in \binom{[n]}{n-1} : f(A) = \ell \right\} \right|$$

be the element that is most frequently output by Alice's strategy f, and denote its number of occurences by

$$k \overset{\text{def}}{=} \left| \left\{ A \in \binom{[n]}{n-1} : f(A) = i \right\} \right|.$$

We consider three different cases depending on the value of k:

```
Algorithm 1: MinHash strategy [Bro97]
  Alice's input: $A \subseteq [n]$
  Bob's input: $B \subseteq [n]$
  Shared randomness: a random permutation $\pi : [n] \to [n]$
  Strategy:
    • $f(A, \pi) = \pi(i_A)$, where $i_A$ is the smallest index such that $\pi(i_A) \in A$.
    • $g(B, \pi) = \pi(i_B)$, where $i_B$ is the smallest index such that $\pi(i_B) \in B$.
```

(i) If $k \leq n-3$, then consider any $B \subseteq [n]$ with $|B| = n-1$. For any value of $f(B) \in B$, the conditional error probability $\Pr[f(A) \neq g(B) \,|\, B]$ is at least $2/(n-1)$. Averaging over all such B, we get a contradiction to Equation (6).

(ii) If $k = n-2$, let $A_1 \neq A_2$ be the two subsets of $[n]$ with $|A_1| = |A_2| = n-1$ such that $f(A_1) \neq i$ and $f(A_2) \neq i$. For any $B \subseteq [n]$ with $|B| = n-1$ such that $B \neq A_1$ and $B \neq A_2$, the conditional error probability $\Pr[f(A) \neq g(B) \,|\, B]$ is at least $2/(n-1)$. Note that there are $n-2$ such B's, and that either A_1 or A_2 is the set $[n] \setminus \{i\}$. If $B = [n] \setminus \{i\}$, then the conditional disagreement probability $\Pr[f(A) \neq g(B) \,|\, B]$ is at least $(n-2)/(n-1)$. Averaging over all B, we get that

$$\begin{aligned}\Pr_{(A,B)\sim\mathcal{D}}[f(A) \neq g(B)] &\geq \left(\frac{2}{n-1}\right) \cdot \left(\frac{n-2}{n}\right) + \left(\frac{n-2}{n-1}\right) \cdot \left(\frac{1}{n}\right) \\ &\geq \frac{2}{n},\end{aligned}$$

where the last inequality holds for any $n \geq 2$. This contradicts Equation (6).

(iii) If $k = n-1$, then the only subset A_1 of $[n]$ with $|A_1| = n-1$ and such that $f(A_1) \neq i$ is $A_1 = [n] \setminus \{i\}$. For any $B \neq A_1$, the conditional error probability $\Pr[f(A) \neq g(B) \,|\, B]$ is at least $1/(n-1)$. On the other hand, if $B = A_1$, then the conditional error probability is equal to 1. Averaging over all B, we get that

$$\begin{aligned}\Pr_{(A,B)\sim\mathcal{D}}[f(A) \neq g(B)] &\geq \left(\frac{1}{n-1}\right) \cdot \left(\frac{n-1}{n}\right) + 1 \cdot \left(\frac{1}{n}\right) \\ &= \frac{2}{n},\end{aligned}$$

which contradicts Equation (6).

□

3 Correlated Sampling Strategies

In this section, we describe the correlated sampling strategy that proves Theorem 1.2. First, let's consider the case of *flat distributions* where, the distributions P and Q are promised to be of the following form, that there exist $A, B \subseteq [n]$ such that $P = \mathcal{U}(A)$ and $Q = \mathcal{U}(B)$ over the universe $[n]$. In this case, it is easy to show that the protocol given in Algorithm 1 achieves an error probability of $1 - \frac{|A \cap B|}{|A \cup B|}$. Since π is a random permutation, it is clear that $f(A, \pi)$ is uniformly distributed over A and $g(B, \pi)$ is uniformly distributed over B. Let i_0 be the smallest index such that $\pi(i_0) \in A \cup B$. The probability that $\pi(i_0) \in A \cap B$ is exactly $\frac{|A \cap B|}{|A \cup B|}$, and this happens precisely when $f(A, \pi) = g(B, \pi)$. Hence, we get the claimed error probability.

The strategy desired in Theorem 1.2 can now be obtained by a reduction to the case of *flat* distributions, and subsequently using the MinHash strategy.

Proof of Theorem 1.2. Given a universe Ω, define a new universe $\Omega' = \Omega \times \Gamma$, where $\Gamma = \{0, \gamma, 2\gamma, \cdots, 1\}$ for a sufficiently small value of $\gamma > 0$. Thus, $|\Omega'| = \frac{1}{\gamma} \cdot |\Omega|$. Suppose we are given distributions P and Q such that $d_{\mathrm{TV}}(P, Q) = \delta$. Define $A = \{(\omega, p) \in \Omega \times \Gamma : p < P(\omega)\}$ and $B = \{(\omega, q) \in \Omega \times \Gamma : q < Q(\omega)\}$.

Algorithm 2: Holenstein's strategy [Hol07]

Alice's input: $P \in \Delta_\Omega$
Bob's input: $Q \in \Delta_\Omega$
Pre-processing: Let $\Omega' = \Omega \times \Gamma$, where $\Gamma = \{0, \gamma, 2\gamma, \cdots, 1\}$ (for suitable $\gamma > 0$)
Shared randomness: $r \sim \mathcal{R}$ as required by the MinHash strategy on Ω'

Strategy:

- Let $A = \{(\omega, p) \in \Omega \times \Gamma : p < P(\omega)\}$ and $B = \{(\omega, q) \in \Omega \times \Gamma : q < Q(\omega)\}$.
- Alice and Bob use MinHash strategy (Algorithm 1) with inputs A, B on universe Ω' to obtain (ω_A, p_A) and (ω_B, p_B) respectively.
- Alice outputs ω_A.
- Bob outputs ω_B.

Holenstein's strategy can now be simply described as follows: Alice and Bob use the MinHash strategy on inputs A and B over the universe Ω', to obtain elements (ω_A, p_A) and (ω_B, p_B) respectively, and they simply output ω_A and ω_B respectively. This strategy is summarized in Algorithm 2.

It can easily seen that, $|A| = \sum_{\omega \in \Omega} \left\lfloor \frac{P(\omega)}{\gamma} \right\rfloor$ and hence,

$$\sum_{\omega \in \Omega} \left(\frac{P(\omega)}{\gamma} - 1 \right) \le |A| \le \sum_{\omega \in \Omega} \frac{P(\omega)}{\gamma}$$

And hence,

$$\frac{1}{\gamma} - |\Omega| \le |A|, |B| \le \frac{1}{\gamma}$$

Similarly, $|A \cap B| = \sum_{\omega \in \Omega} \min \left\{ \left\lfloor \frac{P(\omega)}{\gamma} \right\rfloor, \left\lfloor \frac{Q(\omega)}{\gamma} \right\rfloor \right\}$ and $|A \cup B| = \sum_{\omega \in \Omega} \max \left\{ \left\lfloor \frac{P(\omega)}{\gamma} \right\rfloor, \left\lfloor \frac{Q(\omega)}{\gamma} \right\rfloor \right\}$ and hence,

$$\begin{aligned} \frac{1-\delta}{\gamma} - |\Omega| \ &\le\ |A \cap B| \ \le\ \frac{1-\delta}{\gamma} \\ \frac{1+\delta}{\gamma} - |\Omega| \ &\le\ |A \cup B| \ \le\ \frac{1+\delta}{\gamma} \end{aligned}$$

The probability that Alice outputs ω_A is $\frac{\left\lfloor \frac{P(\omega_A)}{\gamma} \right\rfloor}{|A|}$, which is bounded as,

$$P(\omega_A) - \gamma \le \frac{\left\lfloor \frac{P(\omega_A)}{\gamma} \right\rfloor}{|A|} \le \frac{P(\omega_A)}{1 - \gamma \cdot |\Omega|}$$

Thus, it follows that as $\gamma \to 0$, Alice's output is distributed according to P, and similarly Bob's output is distributed according to Q. Moreover, we have that,

$$\Pr[\omega_A \ne \omega_B] \ = \ 1 - \frac{|A \cap B|}{|A \cup B|} \ \le \ 1 - \frac{1 - \delta - \gamma \cdot |\Omega|}{1 + \delta} \ = \ \frac{2\delta + \gamma \cdot |\Omega|}{1 + \delta} \to \frac{2\delta}{1 + \delta}$$

This gives us the desired error probability. □

3.1 Strategy in a Special Case

In this section, we describe the correlated sampling strategy of [Riv16] that proves Theorem 1.6. To do so, we will need the well-known Hall's Theorem.

Theorem 3.1 (Hall; [vLW01])**.** *Consider a bipartite graph G on vertex sets L and R. Then, there is a matching that entirely covers L if and only if for every subset $S \subseteq L$, we have that $|S| \le |N_G(S)|$, where $N_G(S)$ denotes the set of all neighbors of elements of S in G.*

Algorithm 3: Rivest's strategy [Riv16]

Alice's input: $A \subseteq [n]$
Bob's input: $B \subseteq [n]$
Promise: $|A| = |B| = k$, $|A \cap B| = 1$ and $A \cup B = [n]$
Pre-processing: Let G be the bipartite graph on vertices $\binom{[n]}{k} \times \binom{[n]}{k}$, with an edge between vertices A and B if $|A \cap B| = 1$. Decompose the edges of G into k disjoint matchings $M_1, \cdots, M_k$.
Shared randomness: Index $r \in [k]$

Strategy:

- Let (A, B') and (A', B) be edges present in M_r.
- Alice outputs the unique element in $A \cap B'$.
- Bob outputs the unique element in $A' \cap B$.

Proof of Theorem 1.6. Alice and Bob have subsets $A, B \subseteq [n]$ respectively such that $|A| = |B| = k$, $|A \cap B| = 1$ and $A \cup B = [n]$. This forces $n = 2k - 1$. Consider the bipartite graph G on vertices $\binom{[n]}{k} \times \binom{[n]}{k}$, with an edge between vertices A and B if $|A \cap B| = 1$. It is easy to see that G is k-regular. Iteratively using Hall's theorem (Theorem 3.1), we get that the edges of G can be written as a disjoint union of k matchings. Let's denote these as $M_1, M_2, \cdots, M_k$.

The strategy of Alice and Bob is as follows: Use the shared randomness to sample a random index $r \in [k]$ and consider the matching M_r. If (A, B') is the edge present in M_r, then Alice outputs the unique element in $A \cap B'$. Similarly, if (A', B) is the edge present in M_r, then Bob outputs the unique element in $A' \cap B$. This protocol is summarized in Algorithm 3.

It is easy to see that both Alice and Bob are outputting uniformly random elements in A and B respectively. Moreover, the probability that they output the same element, is exactly $1/k$, which is the probability of choosing the unique matching M_r which contains the edge (A, B) (i.e. enforcing $A = A'$ and $B = B'$). □

References

[BHH⁺08] Boaz Barak, Moritz Hardt, Ishay Haviv, Anup Rao, Oded Regev, and David Steurer. Rounding parallel repetitions of unique games. In *Foundations of Computer Science, 2008. FOCS'08. IEEE 49th Annual IEEE Symposium on*, pages 374–383. IEEE, 2008. 2

[Bro97] Andrei Z Broder. On the resemblance and containment of documents. In *Compression and Complexity of Sequences 1997. Proceedings*, pages 21–29. IEEE, 1997. 2, 7

[Cha02] Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388. ACM, 2002. 2

[GP06] Sreenivas Gollapudi and Rina Panigrahy. A dictionary for approximate string search and longest prefix search. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 768–775. ACM, 2006. 2

[Hol07] Thomas Holenstein. Parallel repetition: simplifications and the no-signaling case. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 411–419. ACM, 2007. 2, 8

[KT02] Jon Kleinberg and Eva Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and markov random fields. *Journal of the ACM (JACM)*, 49(5):616–639, 2002. 2

[M⁺94] Udi Manber et al. Finding similar files in a large file system. In *Usenix Winter*, volume 94, pages 1–10, 1994. 2

[MMT10] Mark Manasse, Frank McSherry, and Kunal Talwar. Consistent weighted sampling. *Unpublished technical report) http://research. microsoft. com/en-us/people/manasse*, 2010. 2

[Rao11] Anup Rao. Parallel repetition in projection games and a concentration bound. *SIAM Journal on Computing*, 40(6):1871–1891, 2011. 2

[Riv16] Ronald L. Rivest. Symmetric encryption via keyrings and ecc. http://arcticcrypt.b.uib.no/files/2016/07/Slides-Rivest.pdf, 2016. 2, 3, 4, 8, 9

[vLW01] Jacobus Hendricus van Lint and Richard Michael Wilson. *A course in combinatorics*. Cambridge university press, 2001. 8