

# A Note On the Statistical Difference of Small Direct Products

Leonid Reyzin  
Boston University  
Computer Science  
Boston, MA 02215 USA

Technical Report BUCS-TR-2004-032

September 21, 2004

## Abstract

We demonstrate that if two probability distributions  $D$  and  $E$  of sufficiently small min-entropy have statistical difference  $\varepsilon$ , then the direct-product distributions  $D^l$  and  $E^l$  have statistical difference at least roughly  $\varepsilon\sqrt{l}$ , provided that  $l$  is sufficiently small, smaller than roughly  $1/\varepsilon^{4/3}$ . Previously known bounds did not work for few repetitions  $l$ , requiring  $l > 1/\varepsilon^2$ .

## 1 Introduction

Given a set  $\Sigma$  and two probability distributions  $E$  and  $D$  on  $\Sigma$ , the statistical difference (or statistical distance) is defined as

$$\|D - E\| \stackrel{\text{def}}{=} \max_{S \subseteq \Sigma} |\Pr_E[S] - \Pr_D[S]|.$$

(It is not hard to see that  $\|D - E\| = \frac{1}{2} \sum_{s \in \Sigma} |\Pr_E[s] - \Pr_D[s]|$ .) One is often interested in seeing how statistical difference behaves with repeated sampling, i.e., with the product distributions. As shown by Sahai and Vadhan [SV99], if  $\|D^l - E^l\| = \varepsilon$ , then  $l\varepsilon \geq \|D^l - E^l\| \geq 1 - 2\exp(-l\varepsilon^2/2)$ .

Note that the lowerbound of  $1 - 2\exp(-l\varepsilon^2/2)$  (which is based on Hoeffding's inequality [Hoe63]) says nothing when the number of repetitions  $l$  is small compared to the inverse square  $1/\varepsilon^2$  of the bias: if  $l\varepsilon^2 < 1$ , then  $\exp(-l\varepsilon^2/2) > 1/2$ , and so  $1 - 2\exp(-l\varepsilon^2/2) < 0$ . We are interested in a lowerbound that covers the case of small  $l$ .

However, as pointed out by Sudan and reported in [SV99], there are distributions for which  $\|D^2 - E^2\| = \|D - E\|$ . Thus, we will not be able to improve the bound for all distributions and all  $l$ . Nevertheless, we demonstrate the following.

**Theorem 1.** *Let  $E$  and  $D$  be two probability distributions on a space  $\Sigma$ , with statistical distance  $\|D - E\| = \varepsilon$ . Let  $m_E$  be the maximum probability in  $E$ , and  $m_D$  be the maximum probability in  $D$ . If  $m = \min(m_E, m_D) \leq \varepsilon/2$ , then the statistical difference between  $E^l$  and  $D^l$  is at least*

$$\|D^l - E^l\| > \left( \frac{\sqrt{l}}{\sqrt{2\pi}} - \frac{1}{\sqrt{8\pi l}} \right) \varepsilon - \frac{3}{8}l\varepsilon^2 - \frac{3}{32}l^2\varepsilon^3.$$

*Asymptotically, if  $\varepsilon = o(1/l^{3/4})$ , the statistical difference is at least*

$$\|D^l - E^l\| = \Omega(\varepsilon\sqrt{l}).$$

The above theorem has one restriction: that maximum probability  $m$  in  $D$  or  $E$  must be smaller than half the statistical distance  $\varepsilon$ . The theorem can still be useful when this is not the case, because maximum probability shrinks exponentially. Thus, if  $m \geq \varepsilon/2$ , one may first replace  $D$  and  $E$  with  $D^f$  and  $E^f$ , where  $f = \lceil \log_m(\varepsilon/2) \rceil$ , and then apply the theorem to  $D^f$  and  $E^f$  (it is easy to see that the statistical difference between  $D^f$  and  $E^f$  is at least as big as the statistical difference between  $D$  and  $E$ ). Alternatively, Lemma 1 below can be applied directly to some distributions and does not require any bound on the maximum probability.

We note that our theorem, unlike the bound of [SV99], cannot be used to find  $l$  as a function of  $\varepsilon$  for which  $\|D^l - E^l\|$  is constant. It is likely to be useful in different contexts than the bound of [SV99].

## 2 Proof

At the heart of the proof is Lemma 1 below. Before we can apply it, we need to find a set  $T \subset \Sigma$  such that  $\Pr_D[T] - \Pr_E[T] \geq \varepsilon/2$  and  $\Pr_D[T] \geq 1/2$ , while  $\Pr_E[T] \leq 1/2$ . This is not hard, but relies on the fact that  $\min(m_E, m_D) \leq \varepsilon/2$ .

Let  $A \subset \Sigma$  be such that  $\Pr_D[A] - \Pr_E[A] = \varepsilon$  (it exists by definition of statistical difference). We can assume without loss of generality that  $\Pr_E[A] \leq 1/2$  (if not, we can use  $\Sigma - A$  instead, and swap  $E$  and  $D$ , because  $\Pr_D[\Sigma - A] = 1 - \Pr_D[A] = 1 - \Pr_E[A] - \varepsilon < 1/2$  and  $\Pr_E[\Sigma - A] - \Pr_D[\Sigma - A] = (1 - \Pr_E[A]) - (1 - \Pr_D[A]) = \varepsilon$ ). If  $\Pr_D[A] \geq 1/2$ , then we set  $T = A$  and apply Lemma 1. If  $\Pr_D[A] < 1/2$ , we create  $T$  as follows. Below, we will use the fact that for all  $s \in \Sigma - A$ , we have  $\Pr_E[s] \geq \Pr_D[s]$ , which follows from the definition of  $A$ .

First consider the case when  $\min(m_E, m_D) = m_E$ . The idea is to add elements to  $A$  until we get the weight in  $E$  to be close to  $1/2$ , so that the weight in  $D$  becomes above  $1/2$ . When adding elements, we have to take care to not destroy the statistical difference; we will, reduce it, but not below  $\varepsilon/2$ .

Let  $T_1 \subset \Sigma - A$  be a subset whose weight in  $E$  is maximal among all elements of  $\{R | R \subset \Sigma - A, \Pr_E[R] < 1/2 - \Pr_E[A]\}$ . Note that  $\Pr_E[T_1] \geq 1/2 - \Pr_E[A] - m_E$  (otherwise, we could add an element to  $T_1$ ). Similarly, let  $T_2 \subset \Sigma - A - T_1$  be a subset whose weight in  $E$  is maximal among all elements of  $\{R | R \subset \Sigma - A - T_1, \Pr_E[R] < 1/2 - \Pr_E[A]\}$ . Again,  $\Pr_E[T_2] \geq 1/2 - \Pr_E[A] - m_E$ . By definition of statistical difference,  $0 \leq \Pr_E[T_1 \cup T_2] - \Pr_D[T_1 \cup T_2] \leq \varepsilon$ . Hence,  $\exists i \in \{1, 2\}$  such that  $0 \leq \Pr_E[T_i] - \Pr_D[T_i] \leq \varepsilon/2$ . Now set  $T = A \cup T_i$ . Then  $\Pr_E[T] \leq \Pr_E[A] + (1/2 - \Pr_E[A]) = 1/2$ , and  $\Pr_D[T] = \Pr_D[A] + \Pr_D[T_i] \geq \Pr_E[A] + \varepsilon + \Pr_E[T_i] - \varepsilon/2 = \Pr_E[T] + \varepsilon/2 \geq 1/2 - m_E + \varepsilon/2 \geq 1/2$ . We can now apply Lemma 1 to  $T$ .

The case when  $\min(m_E, m_D) = m_D$  is handled quite similarly. We choose  $T_1$  to be a subset whose weight in  $D$  is minimal among all elements of  $\{R | R \subset \Sigma - A, \Pr_D[R] \geq 1/2 - \Pr_D[A]\}$ . Note that  $\Pr_D[T_1] \leq 1/2 - \Pr_D[A] + m_D$  (else we could throw out an element of  $T_1$ ). We then choose  $T_2 \subset \Sigma - A - T_1$  be a subset whose weight in  $D$  is minimal among all elements of  $\{R | R \subset \Sigma - A - T_1, \Pr_D[R] \geq 1/2 - \Pr_D[A]\}$  (here we need that  $\Pr_D[\Sigma - A - T_1] \geq 1/2 - \Pr_D[A]$ , which is true because the left-hand side is equal to  $1 - \Pr_D[A] - \Pr_D[T_1] \geq 1/2 - m_D$ , while  $\Pr_D[A] \geq \varepsilon > m_D$ ). We choose  $T_i$  the same way as above, and set  $T = A \cup T_i$ . Similar calculations show that Lemma 1 applies again.

All that is left now is to state and prove the lemma.

**Lemma 1.** *Let  $E$  and  $D$  be two probability distributions on a space  $\Sigma$ , and let  $T \subset \Sigma$  be an event whose probability is at most  $p_1 \leq 1/2$  in  $E$  and  $p_2 \geq 1/2$  in  $D$ . Let  $d = p_2 - p_1$ . Then the statistical*

difference between  $E^l$  and  $D^l$  is at least

$$\left( \frac{\sqrt{2l}}{\sqrt{\pi}} - \frac{1}{\sqrt{2\pi l}} \right) d - \frac{3}{2}ld^2 - \frac{3}{4}l^2d^3.$$

*Proof.* Before proving the lemma, we prove the following bound on the tail of the binomial distribution.

**Proposition 1.** *Let  $X_1, X_2, \dots, X_l$  be independent identically distributed random variables,  $X_i \in \{0, 1\}$ ,  $\Pr[X_i = 1] = 1/2 + x$  for  $x \geq 0$  (Bernoulli trials with  $1/2 + x$  probability of success). Let  $S = X_1 + X_2 + \dots + X_l$ . Then*

$$\Pr[S < l/2] < \begin{cases} \frac{1}{2} - \left( \frac{\sqrt{2l}}{\sqrt{\pi}} - \frac{1}{\sqrt{2\pi l}} \right) x + lx^2 + \frac{l^2}{2}x^3 & \text{if } l \text{ is odd} \\ \frac{1}{2} - 2^{-l-1} \binom{l}{l/2} - \left( \frac{\sqrt{2l}}{\sqrt{\pi}} - \frac{1}{\sqrt{8\pi l}} \right) x + 1.5lx^2 + .75l^2x^3 & \text{if } l \text{ is even} \end{cases}$$

*Proof.* Assume  $l > 1$  (for  $l = 1$  the formula is trivially satisfied).

Consider instead  $\Pr[S \geq l/2] = 1 - \Pr[S < l/2]$ .

$$\begin{aligned} \Pr[S \geq l/2] &= \sum_{i=0}^{\lfloor l/2 \rfloor} \binom{l}{i} (1/2 - x)^i (1/2 + x)^{l-i} \\ &= 2^{-l} \sum_{i=0}^{\lfloor l/2 \rfloor} \binom{l}{i} (1 - y)^i (1 + y)^{l-i} \end{aligned}$$

where  $y = 2x$ . The above is equal to

$$\begin{aligned} &2^{-l} \sum_{i=0}^{\lfloor l/2 \rfloor} \binom{l}{i} (1 - y^2)^i (1 + y)^{l-2i} \\ &\geq 2^{-l} \sum_{i=0}^{\lfloor l/2 \rfloor} \binom{l}{i} (1 - iy^2)(1 + (l - 2i)y) \\ &\geq 2^{-l} \sum_{i=0}^{\lfloor l/2 \rfloor} \binom{l}{i} \left( 1 + (l - 2i)y - \frac{l}{2}y^2 - \frac{l^2}{8}y^3 \right) \end{aligned}$$

(the first step is true because  $(1 - y)(1 + y) = 1 - y^2$ ; the second step is true because  $(1 + \alpha)^k > 1 + k\alpha$  for  $k > 1$ ; last step is true because  $i \leq l/2$  and  $i(l - 2i) \leq l^2/8$  for any  $i$ ).

We now consider two cases, depending on the parity of  $l$ . Assume that  $l$  is odd. In that case, because  $\sum_{i=0}^{\lfloor l/2 \rfloor} \binom{l}{i} = 2^{l-1}$ , the above becomes

$$\begin{aligned} &\frac{1}{2} + \frac{ly}{2} - \frac{l}{4}y^2 - \frac{l^2}{16}y^3 - 2^{-l} \sum_{i=0}^{\lfloor l/2 \rfloor} 2iy \binom{l}{i} \\ &= \frac{1}{2} + \frac{ly}{2} - \frac{l}{4}y^2 - \frac{l^2}{16}y^3 - 2^{-l+1}y \sum_{i=1}^{\lfloor l/2 \rfloor} l \binom{l-1}{i-1} \end{aligned}$$

(here we use that  $i \binom{l}{i} = l \binom{l-1}{i-1}$ ). Recalling that  $\sum_{i=1}^{\lfloor l/2 \rfloor} \binom{l-1}{i-1} = \sum_{i=0}^{(l-1)/2-1} \binom{l-1}{i} = 2^{l-2} - \frac{1}{2} \binom{l-1}{(l-1)/2}$  leads us to

$$\begin{aligned} & \frac{1}{2} + \frac{ly}{2} - \frac{l}{4}y^2 - \frac{l^2}{16}y^3 - 2^{-l+1}yl \left( 2^{l-2} - \frac{1}{2} \binom{l-1}{(l-1)/2} \right) \\ &= \frac{1}{2} + 2^{-l}ly \binom{l-1}{(l-1)/2} - \frac{l}{4}y^2 - \frac{l^2}{16}y^3. \end{aligned}$$

Finally, using  $\binom{2n}{n} > \frac{4^n}{\sqrt{\pi n}} \left(1 - \frac{1}{8n}\right)$  [GKP89, p. 481], we lowerbound the above by

$$\begin{aligned} & \frac{1}{2} + 2^{-l}ly \frac{4^{(l-1)/2}}{\sqrt{\pi(l-1)/2}} \left(1 - \frac{1}{4(l-1)}\right) - \frac{l}{4}y^2 - \frac{l^2}{16}y^3 \\ &> \frac{1}{2} + \frac{l}{\sqrt{2\pi(l-1)}} \left(1 - \frac{1}{4(l-1)}\right) y - \frac{l}{4}y^2 - \frac{l^2}{16}y^3 \\ &> \frac{1}{2} + \left( \frac{l}{\sqrt{2\pi l}} - \frac{l}{4(l-1)\sqrt{2\pi(l-1)}} \right) y - \frac{l}{4}y^2 - \frac{l^2}{16}y^3 \\ &> \frac{1}{2} + \left( \frac{\sqrt{l}}{\sqrt{2\pi}} - \frac{1}{\sqrt{8\pi l}} \right) y - \frac{l}{4}y^2 - \frac{l^2}{16}y^3 \\ &= \frac{1}{2} + \left( \frac{\sqrt{2l}}{\sqrt{\pi}} - \frac{1}{\sqrt{2\pi l}} \right) x - lx^2 - \frac{l^2}{2}x^3 \end{aligned}$$

(in the second to last step we used  $l/(l-1)^{3/2} < 2/\sqrt{l}$  for  $l \geq 3$ ).

We now consider the case of even  $l$ . In that case, because  $\sum_{i=0}^{\lfloor l/2 \rfloor} \binom{l}{i} = 2^{l-1} + \frac{1}{2} \binom{l}{l/2}$ , we get

$$\begin{aligned} & 2^{-l} \sum_{i=0}^{\lfloor l/2 \rfloor} \binom{l}{i} \left( 1 + (l-2i)y - \frac{l}{2}y^2 - \frac{l^2}{8}y^3 \right) \\ &= \frac{1}{2} + \frac{ly}{2} - \frac{l}{4}y^2 - \frac{l^2}{16}y^3 + 2^{-l-1} \binom{l}{l/2} \left( 1 + ly - \frac{l}{2}y^2 - \frac{l^2}{8}y^3 \right) - 2^{-l} \sum_{i=0}^{\lfloor l/2 \rfloor} 2iy \binom{l}{i} \\ &\geq \frac{1}{2} + \frac{ly}{2} - \frac{3l}{8}y^2 - \frac{3l^2}{32}y^3 + 2^{-l-1} \binom{l}{l/2} (1+ly) - 2^{-l+1}y \sum_{i=1}^{\lfloor l/2 \rfloor} l \binom{l-1}{i-1} \end{aligned}$$

(here we use that  $2^{-l} \binom{l}{l/2} \leq 1/2$  for  $l \geq 2$  and  $i \binom{l}{i} = l \binom{l-1}{i-1}$ ). Recalling that  $\sum_{i=1}^{\lfloor l/2 \rfloor} \binom{l-1}{i-1} = \sum_{i=0}^{\lfloor (l-1)/2 \rfloor} \binom{l-1}{i} = 2^{l-2}$  leads us to

$$\begin{aligned} & \frac{1}{2} + \frac{ly}{2} - \frac{3l}{8}y^2 - \frac{3l^2}{32}y^3 + 2^{-l-1} \binom{l}{l/2} (1+ly) - 2^{-l+1}yl2^{l-2} \\ &= \frac{1}{2} - \frac{3l}{8}y^2 - \frac{3l^2}{32}y^3 + 2^{-l-1} \binom{l}{l/2} (1+ly). \end{aligned}$$

Finally, using  $\binom{2n}{n} > \frac{4^n}{\sqrt{\pi n}} \left(1 - \frac{1}{8n}\right)$  [GKP89, p. 481], we get that the above is greater than

$$\frac{1}{2} + 2^{-l-1} \binom{l}{l/2} + 2^{-l-1}ly \frac{4^{l/2}}{\sqrt{\pi l/2}} \left(1 - \frac{1}{4l}\right) - \frac{3l}{8}y^2 - \frac{3l^2}{32}y^3$$

$$\begin{aligned}
&= \frac{1}{2} + 2^{-l-1} \binom{l}{l/2} + \left( \frac{\sqrt{l}}{\sqrt{2\pi}} - \frac{1}{4\sqrt{2\pi l}} \right) y - \frac{3l}{8} y^2 - \frac{3l^2}{32} y^3 \\
&= \frac{1}{2} + 2^{-l-1} \binom{l}{l/2} + \left( \frac{\sqrt{2l}}{\sqrt{\pi}} - \frac{1}{\sqrt{8\pi l}} \right) x - 1.5lx^2 - .75l^2x^3.
\end{aligned}$$

□

The lemma follows from the tail bound as follows. Let  $B \subset \Sigma^l$  be the set of sequences  $s_1s_2 \dots s_l$  for which at least half of the members are in  $T$ . We will lowerbound the probability of  $B$  in  $D^l$  and upperbound it in  $E^l$ .

Let  $X_i$  be a random variable on  $D^l$  defined as  $X_i(s_1s_2 \dots s_l) = 1$  if  $s_i \in T$ . Let  $x_1 = \Pr_D[s \in T] - 1/2$  and  $S_X = X_1 + X_2 + \dots + X_l$ . Note that  $\Pr_{D^l}[B] = \Pr[S_X \geq l/2]$ . By the tail bound,

$$\Pr_{D^l}[B] = 1 - \Pr[S_X < l/2] > \begin{cases} \frac{1}{2} + \left( \frac{\sqrt{2l}}{\sqrt{\pi}} - \frac{1}{\sqrt{2\pi l}} \right) x_1 - lx_1^2 - \frac{l^2}{2}x_1^3 & \text{if } l \text{ is odd} \\ \frac{1}{2} + 2^{-l-1} \binom{l}{l/2} + \left( \frac{\sqrt{2l}}{\sqrt{\pi}} - \frac{1}{\sqrt{8\pi l}} \right) x_1 - 1.5lx_1^2 - .75l^2x_1^3 & \text{if } l \text{ is even} \end{cases}$$

On the other hand, let  $Y_i$  be a random variable on  $E^l$  defined as  $Y_i(s_1s_2 \dots s_l) = 1$  if  $s_i \notin T$ . Let  $x_2 = \Pr_E[s \notin T] - 1/2$  and  $S_Y = Y_1 + Y_2 + \dots + Y_l$ . Then the tail bound states that

$$\Pr[S_Y < l/2] < \begin{cases} \frac{1}{2} - \left( \frac{\sqrt{2l}}{\sqrt{\pi}} - \frac{1}{\sqrt{2\pi l}} \right) x_2 + lx_2^2 + \frac{l^2}{2}x_2^3 & \text{if } l \text{ is odd} \\ \frac{1}{2} - 2^{-l-1} \binom{l}{l/2} - \left( \frac{\sqrt{2l}}{\sqrt{\pi}} - \frac{1}{\sqrt{8\pi l}} \right) x_2 + 1.5lx_2^2 + .75l^2x_2^3 & \text{if } l \text{ is even} \end{cases}$$

Observe that  $\Pr_{E^l}[B] = \Pr[S_Y \leq l/2] = \Pr[S_Y < l/2] + \Pr[S_Y = l/2]$ . If  $l$  is odd, then  $\Pr[S_Y = l/2] = 0$ . If  $l$  is even, then  $\Pr[S_Y = l/2] = \binom{l}{l/2} (1/2 - x_2)^{l/2} (1/2 + x_2)^{l/2} = \binom{l}{l/2} (1/4 - x_2^2)^{l/2} \leq 2^{-l} \binom{l}{l/2}$ .

Finally, subtracting the two bounds, and recalling that  $d = x_1 + x_2$  (hence  $d^2 \leq x_1^2 + x_2^2$  and  $d^3 \leq x_1^3 + x_2^3$ ), we get that the statistical difference between  $D^l$  and  $E^l$  is at least

$$\Pr_{D^l}[B] - \Pr_{E^l}[B] > \begin{cases} \left( \frac{\sqrt{2l}}{\sqrt{\pi}} - \frac{1}{\sqrt{2\pi l}} \right) d - ld^2 - \frac{l^2}{2}d^3 & \text{if } l \text{ is odd} \\ \left( \frac{\sqrt{2l}}{\sqrt{\pi}} - \frac{1}{\sqrt{8\pi l}} \right) d - 1.5ld^2 - .75l^2d^3 & \text{if } l \text{ is even} \end{cases}$$

□

### 3 Acknowledgments

I thank Murad Taqqu for a suggestion on approaching the tail bound.

### References

- [GKP89] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, 1989.
- [Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

- [SV99] Amit Sahai and Salil Vadhan. Manipulating statistical difference. In Panos Pardalos, Sanguthevar Rajasekaran, and José Rolim, editors, *Randomization Methods in Algorithm Design (DIMACS Workshop, December 1997)*, volume 43 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 251–270. American Mathematical Society, 1999.