A Lower Bound on the Complexity of Testing Grained Distributions*

Oded Goldreich† Dana Ron‡

September 6, 2021

Abstract

A distribution is called m-grained if each element appears with probability that is an integer multiple of $1/m$. We prove that, for any constant $c < 1$, testing whether a distribution over $[\Theta(m)]$ is m-grained requires $\Omega(m^c)$ samples.

1 Introduction

A distribution $P : \Omega \to [0, 1]$ is called m-grained if $P(x)$ is a multiple of $1/m$ for every x in Ω; that is, for each $x \in \Omega$, there exists an integer m_x, such that $P(x) = m_x/m$ (see [3, Def. 11.7]). Grained distributions have appeared implicitly in several prior works (most conspicuously in [4]), and were defined and studied explicitly in [2]. In particular, the challenge of determining the sample complexity of testing the set of grained distributions (i.e., the property of being grained) was raised explicitly in [2, Sec. 4]. For sake of completeness, we reproduce the standard definition of testing properties of distributions, where distances (like in "ϵ-far") refer to the total variation distance.

Definition 1 (testing properties of distributions): *Let $\mathcal{D} = \{\mathcal{D}_n\}_{n\in\mathbb{N}}$ be a property of distributions such that $\mathcal{D}_n$ is a set of distributions over $[n]$, and $s : \mathbb{N} \times (0, 1] \to \mathbb{N}$. A* tester*, denoted T, of* sample complexity s for the property $\mathcal{D}$ *is a probabilistic machine that, on input parameters n and ϵ, and a sequence of $s(n, \epsilon)$ samples drawn from an unknown distribution P over $[n]$, satisfies the following two conditions.*

1. The tester accepts distributions that belong to $\mathcal{D}$: *If P is in $\mathcal{D}_n$, then*

$$\Pr_{i_1,\ldots,i_s\sim P}[T(n, \epsilon; i_1, \ldots, i_s)=1] \geq 2/3,$$

where $s = s(n, \epsilon)$ and $i_1, \ldots, i_s$ are drawn independently from the distribution P.

2. The tester rejects distributions that are far from $\mathcal{D}$: *If P is ϵ-far from any distribution in $\mathcal{D}_n$* (i.e., P is ϵ-far from $\mathcal{D}$) *with respect to the variation distance, then*

$$\Pr_{i_1,\ldots,i_s\sim P}[T(n, \epsilon; i_1, \ldots, i_s)=0] \geq 2/3,$$

where $s = s(n, \epsilon)$ and $i_1, \ldots, i_s$ are as in the previous item.

*Partially supported by the Israel Science Foundation (grant No. 1041/18).

†Department of Computer Science, Weizmann Institute of Science, Rehovot, Israel. E-mail: oded.goldreich@weizmann.ac.il. Additional funding received from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819702).

‡School of Electrical Engineering, Tel Aviv University, Tel Aviv, Israel. danaron@tau.ac.il

We say that testing $\mathcal{D}$ requires $s'(n)$ samples, if for some constant $\epsilon > 0$ any tester of $\mathcal{D}$ has sample complexity $s(n,\epsilon) \geq s'(n)$.

It is quite easy to prove that testing the set of n-grained distributions requires $\Omega(\sqrt{n})$ samples. In particular, $\Omega(\sqrt{n})$ samples are required in order to distinguish the uniform distribution on $[n]$ from a generic distribution that assigns probability $1/2n$ to each of $n/2$ elements and probability $3n/2n$ to each of the remaining elements. To the best of our knowledge, this was the best lower bound known till this work.[1] In this work we obtain a lower bound of $\Omega(n^c)$, for any constant $c < 1$.

Theorem 2 (main result): *For every constant $c < 1$, the sample complexity of testing whether a distribution over $[n]$ is m-grained, where $m = \Theta(n)$, is $\Omega(n^c)$,*

We mention that the sample complexity of testing the foregoing property of distributions is $O(\epsilon^{-2}n/\log n)$; this follows as a special case from the fact that any label-invariant property of distributions can be tested within this complexity [6] (see also [3, Cor. 11.28]). Recall that a property of distributions over $[n]$ is called label-invariant if for every bijection $\pi : [n] \to [n]$ and every distribution P, it holds that P is in the property if and only if $\pi(P)$ is in the property, where $Q = \pi(P)$ is such that $Q(y) = P(\pi^{-1}(y))$. We conjecture that the aforementioned upper bound is tight; that is:

Conjecture 3 *The sample complexity of testing $\Theta(n)$-grained distributions over $[n]$ is $\Omega(n/\log n)$.*

We mention that the techniques used in our proof of Theorem 2 seem inadequate for proving a lower bound of the form $\Omega(n^{1-o(1)})$. In particular, our proof holds also when guaranteed that the tested distribution assigns probability $O(1/n)$ to each element in its support. However, under this promise, one can even learn the distribution (up to relabeling) using $O(n^{1-\Omega(1)})$ samples.[2]

2 Proof of Theorem 2

Our proof relies on two standard simplifying assumptions:

1. When considering the task of testing a label-invariant property, one may assume, without loss of generality, that the tester is label-invariant [1] (see also [3, Thm. 11.12]); that is, for every bijection π on the potential support, the tester's verdict on the samples $i_1, \ldots, i_s$ is identical to its verdict on the samples $\pi(i_1), \ldots, \pi(i_s)$.

2. To prove a lower bound of L on the sample complexity of testing, it suffices to describe two distributions P and Q that no algorithm of sample complexity $L-1$ can distinguish (with

[1] We mention that a lower bound of $\Omega(n/\log n)$ was known for the tolerant version [3, Thm. 11.31] in which, for some positive constants $\delta < \epsilon$, one is required to distinguish distributions that are δ-close to being n-grained from distributions that are ϵ-far from being n-grained.

[2] Indeed, suppose that a distribution $P : [n] \to [0,1]$ is guaranteed to satisfy $P(i) \leq t/n$ for every $i \in [n]$. For simplicity suppose that P is also $(t \cdot n)$-grained. Then, the histogram $(h_0, ..., h_{t^2})$ such that $h_j = |\{i \in [n] : P(i) = j/(t \cdot n)\}|$ is determined by the probabilities of k-way collisions for $k \in \{2, ..., t^2+2\}$, whereas the probability of k-way collisions can be approximated using $O(n^{(k-1)/k})$ samples of P. The argument can be extended to the case that P is not $O(1/n)$-grained by clustering the elements according to their approximate probability.

gap $\Omega(1))^3$ such that P has the property and Q is $\Omega(1)$-far from having the property (cf. [3, Thm. 7.2]).

Combining these two observations, we focus on presenting distributions that cannot be distinguished by label-invariant algorithms of low complexity such that one distribution is m-grained while the other is $\Omega(1)$-far from being m-grained.

Both distributions that we present are specified by their histograms, which specify how many elements are assigned each value of the probability weight. For $t = O(1/(1-c))$, in both distributions, each element in $[n]$ is assigned weight $\frac{i}{2m}$ such that $i \in [t]$. In particular:

1. In distribution P, n_i^{P} elements are assigned the weight $\frac{i}{2m}$, and $n_i^{\mathsf{P}} = 0$ for every odd $i \in [t]$.
2. In distribution Q, n_i^{Q} elements are assigned the weight $\frac{i}{2m}$, and $n_i^{\mathsf{Q}} = 0$ for every even $i \in [t]$.

Note that $\sum_{i\in[t]} n_i^{\mathsf{P}} \cdot \frac{i}{2m} = 1 = \sum_{i\in[t]} n_i^{\mathsf{Q}} \cdot \frac{i}{2m}$ and $\sum_{i\in[t]} n_i^{\mathsf{P}} = n = \sum_{i\in[t]} n_i^{\mathsf{Q}}$, whereas $2m \in \{n, \ldots, tn\}$. Furthermore, P is m-grained, whereas Q is $\frac{1}{3t}$-far from being m-grained (since the weight on each element has to be modified by at least $\frac{1}{2m}$ units whereas $\frac{n}{2m} \geq \frac{1}{t}$).

Note that the equation $\sum_{i\in[t]} n_i^{\mathsf{P}} = \sum_{i\in[t]} n_i^{\mathsf{Q}}$ asserts that both distributions have the same support size, whereas $\sum_{i\in[t]} n_i^{\mathsf{P}} \cdot i = \sum_{i\in[t]} n_i^{\mathsf{Q}} \cdot i$ asserts that they are assigned the same total probability mass (in terms of units of $\frac{1}{2m}$). Intuitively, a sample complexity lower bound of $\Omega\left(n^{\frac{t-2}{t-1}}\right)$ is related to requiring that, for every $k \in \{2, \ldots, t-2\}$, the probability of a k-way collision is the same in both distributions. Thus, we require that $\sum_{i\in[t]} n_i^{\mathsf{P}} \cdot \left(\frac{i}{2m}\right)^k = \sum_{i\in[t]} n_i^{\mathsf{Q}} \cdot \left(\frac{i}{2m}\right)^k$ for every $k \in \{2, \ldots, t-2\}$, which raises the question of whether such a setting of n_i^{P}'s and n_i^{Q}'s is possible. Before addressing the latter question (as well as the question of why this yields the desired lower bound), we reformulate the foregoing $t-1$ equations in a uniform manner; that is, for every $k \in [[t-2]] \stackrel{\text{def}}{=} \{0, 1, \ldots, t-2\}$, we require

$$\sum_{i\in[t]} n_i^{\mathsf{P}} \cdot i^k = \sum_{i\in[t]} n_i^{\mathsf{Q}} \cdot i^k. \tag{1}$$

Recalling the t initial equalities (i.e., $n_i^{\mathsf{P}} = 0$ for odd $i \in [t]$ and $n_i^{\mathsf{Q}} = 0$ for even $i \in [t]$), we write the foregoing linear system in a matrix form as $Ax = 0$, where $x = (n_1^{\mathsf{P}}, \ldots, n_t^{\mathsf{P}}, n_1^{\mathsf{Q}}, \ldots, n_t^{\mathsf{Q}})^\top$. For $i \in [t]$, the i^{th} row of A is $(0^{i-1}10^{2t-i})$ if i is odd, and $(0^{t+i-1}10^{t-i})$ if i is even, whereas (for $k \in \{0, 1, \ldots, t-2\}$) row $(t+k+1)$ of A is $(1^k, 2^k, \ldots, t^k, -1^k, -2^k, \ldots, -t^k)$. Figure 1 depicts A in case of $t = 5$.

We seek a solution x that is *positive*, which means that each of the entries of x is non-negative, and at least one of the entries is positive. It turns out that such a solution exists if and only if for every $v \in \mathbb{R}^{2t}$ it holds that vA is *not* strongly positive [5, Thm. 15.1(2)], where u is strongly positive if all its entries are positive.

Hence, for every $v \in \mathbb{R}^{2t}$, we show that it is impossible that all entries of vA are positive. Actually, it will suffice to show that it not possible that the entries that correspond to even i's in $[t]$ and to $t+i$'s for odd i's (in $[t]$) are all positive. To verify this, observe that the first t rows in

[3] We say that A distinguishes s samples of P from s samples of P with gap γ if

$$|\Pr_{i_1,\ldots,i_s\sim P}[A(i_1,\ldots,i_s)=1] - \Pr_{i_1,\ldots,i_s\sim P}[A(i_1,\ldots,i_s)=1]| \geq \gamma.$$

0	0	0	0	0	0	0	0	0	0
0	**1**	0	0	0	0	**1**	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	**1**	0	0	0	0	**1**	0
0	0	0	0	0	0	0	0	0	**0**
1	1	1	1	1	-1	-1	-1	-1	-1
1	2	3	4	5	-1	-2	-3	-4	-5
1^2	2^2	3^2	4^2	5^2	-1^2	-2^2	-3^2	-4^2	-5^2
1^3	2^3	3^3	4^3	5^3	-1^3	-2^3	-3^3	-4^3	-5^3

Figure 1: The matrix A and the submatrix considered in the analysis.

the corresponding columns are all-zero. Hence, for even $i \in [t]$ the value of the i^{th} entry (in vA) is $\sum_{k\in[[t-2]]} v_{t+k+1} i^k$, whereas for odd $i \in [t]$ the value of the $(t+i)^{\text{th}}$ entry is $-\sum_{k\in[[t-2]]} v_{t+k+1} i^k$. It follows that $\sum_{k\in[[t-2]]} v_{t+k+1} i^k$ should be positive if $i \in [t]$ is even, and negative otherwise. But this is impossible since the degree of this polynomial (in i) is $t-2$ (and so its sign cannot alternate $t-1$ times).

The foregoing discussion establishes the existence of n_i^{P}'s and n_i^{Q}'s that satisfy Eq. (1) for every $k \in [[t-2]]$ as well as $n_i^{\mathsf{P}} = 0$ for odd $i \in [t]$ and $n_i^{\mathsf{Q}} = 0$ for even $i \in [t]$. These n_i^{P}'s and n_i^{Q}'s may be assumed to be rational, but they do not necessarily sum-up to n nor are integers. In fact, these n_i^{P}'s and n_i^{Q}'s are independent of n, and so by multiplying them with an adequate number (e.g., the least common multiplier of their denominators) we obtain integers. Hence, we can fit any n that is an integer multiple of the sum of the resulting n_i^{P}'s (and, we can handle other n's by "padding").

We have thus established that distributions P and Q as postulated above do exist; that is, P and Q are $2m$-grained, and it holds that $n_i^{\mathsf{P}} = \left|\left\{j \in [n] : P(j) = \frac{i}{2m}\right\}\right|$ and $n_i^{\mathsf{Q}} = \left|\left\{j \in [n] : Q(j) = \frac{i}{2m}\right\}\right|$ satisfy Eq. (1) for every $k \in [[t-2]]$ as well as $n_i^{\mathsf{P}} = 0$ for odd $i \in [t]$ and $n_i^{\mathsf{Q}} = 0$ for even $i \in [t]$. In order to proceed, we restate the features of the n_i^{P}'s and n_i^{Q}'s in terms of the (probability) histograms of P and Q (or rather their "normalized" forms). Specifically, consider the following random variable: $X = i$ with probability $\frac{n_i^{\mathsf{P}}}{n}$ (resp., $Y = i$ with probability $\frac{n_i^{\mathsf{Q}}}{n}$), representing the fact that there are n_i^{P} (resp., n_i^{Q}) elements in the support of P (resp., Q) that are assigned probability $\frac{i}{2m}$. Observe that $\mathrm{E}[X^k] = \sum_{i\in[t]} \frac{n_i^{\mathsf{P}}}{n} \cdot i^k$ (resp., $\mathrm{E}[Y^k] = \sum_{i\in[t]} \frac{n_i^{\mathsf{Q}}}{n} \cdot i^k$). Hence, we have established the following:

Lemma 4 (main lemma): *For every constant $t \in \mathbb{N}$ and $m, n \in \mathbb{N}$ such that $m \in \{0.5n, \ldots, 0.5tn\}$, there exist $2m$-grained distributions P and Q over $[n]$ such that the following conditions hold.*

1. *P is m-grained, whereas Q is $\frac{1}{3t}$-far from being m-grained.*
2. *For every $k \in [t-2]$, it holds that $\mathrm{E}[X^k] = \mathrm{E}[Y^k]$, where X and Y are the histograms of P and Q* (respectively, as defined above).

At this point we can apply a result of [4], which we slightly modify and rephrase as follows.[4]

[4] Putting aside the many notational modifications, the actual modification is that Lemma 5 refers to the first $t-2$

Lemma 5 (a variant of [4, Thm. 5.6]): *Let P and Q be $2m$-grained distributions over $[n]$ such that their support equals $[n]$, and $a_1, \ldots, a_t \in \mathbb{N}$ such that for every $j \in [n]$ it holds that $P(j) \in \left\{\frac{a_i}{2m} : i \in [t]\right\}$ and $Q(j) \in \left\{\frac{a_i}{2m} : i \in [t]\right\}$. Define a random variable X* (resp., Y) *over $[t]$ such that $X = i$* (resp., $Y = i$) *with probability that represents the fraction of elements in $[n]$ that are assigned probability $\frac{a_i}{2m}$ by P* (resp., Q). *If, for every $k \in [t-2]$, it holds that $\mathrm{E}[X^k] = \mathrm{E}[Y^k]$, then the distinguishing gap of any label-invariant algorithm between $s \leq m/a$ samples of P and s samples of Q is upper-bounded by*

$$O\left(\frac{t^2 \cdot s}{m/a} + \frac{s^{t-1}}{(m/a)^{t-2}}\right) + \exp(-\Omega(s)), \tag{2}$$

where $a = \max_{i \in [t]}\{a_i\}$.

Note that for non-constant $s = o(m/(t^2 a))$, Eq. (2) yields $o(1)$; that is, for any label-invariant algorithm, the distinguishing gap between s samples of P and s samples of Q is $o(1)$. Hence, combining Lemmas 4 and 5, while setting $a_i = i$ and $s = \Omega(m/a)^{(t-2)/(t-1)}$, we obtain the desired bound; Theorem 2 follows by setting $t = \lceil 1/(1-c) \rceil + 1$.

References

[1] Tugkan Batu. *Testing properties of distributions.* PhD thesis, Computer Science department, Cornell University, 2001.

[2] Oded Goldreich. The Uniform Distribution is Complete with respect to Testing Identity to a Fixed Distribution. *ECCC*, TR16-015, February 2016.

[3] Oded Goldreich. *Introduction to Property Testing.* Cambridge University Press, 2017.

[4] Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong Lower Bounds for Approximating Distribution Support Size and the Distinct Elements Problem. *SIAM Journal on Computing*, Vol. 39 (3), pages 813–842, 2009. Extended abstract in *48th FOCS*, 2007.

[5] Steven Roman. *Advanced Linear Algebra.* Graduate Texts in Mathematics, Vol. 135, Springer, 2005.

[6] Gregory Valiant and Paul Valiant. Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new CLTs. In *43rd ACM Symposium on the Theory of Computing*, pages 685–694, 2011.

Appendix: Deriving Lemma 5 from the proof of [4, Thm. 5.6]

There are several differences between Lemma 5 and [4, Thm. 5.6].

powers of X and Y, whereas [4, Thm. 5.6] refers to the first $t-1$ powers. In fact, we present a generalization of [4, Thm. 5.6] in which the number of powers is a free parameter. In the appendix we outline how this generalization (and in particular Lemma 5) follows from the proof of [4, Thm. 5.6].

1. Lemma 5 refers to algorithms that obtain samples drawn from ($2m$-grained) distributions whereas [4, Thm. 5.6] refers to algorithms that see the colors of balls drawn uniformly and independently (with replacement) among N balls.

   Note that samples drawn from a $2m$-grained distribution over $[n]$ correspond to the colors of uniformly selected balls, where the number of balls equals $2m$ and the number of colors is n. That is, a $2m$-grained distribution D corresponds to a collection of $2m$ balls such that (for every $\chi \in [n]$) exactly $2m \cdot D(\chi)$ balls are assigned the color χ.

2. Lemma 5 refers to algorithms that obtain s samples, whereas [4, Thm. 5.6] refers to algorithms that obtain $\texttt{Poi}(s)$ balls, where $\texttt{Poi}(s)$ denotes the Poisson distribution with parameter s.

   Recall that $\Pr[\texttt{Poi}(s) < s/2] = \exp(-\Omega(s))$, which means that an algorithm that gets $\texttt{Poi}(s)$ samples can emulate an algorithm that expects $s/2$ samples, with error probability $\exp(-\Omega(s))$. The latter error term is accounted for by the last term in Eq. (2).

3. In Lemma 5 the distribution P and Q play the main role while their histograms X and Y appear as secondary players, whereas in [4, Thm. 5.6] the histograms appear as main players and the corresponding distributions of colors appear in the second role.

4. Most importantly, Lemma 5 presupposes equality between the first $t-2$ powers of X and Y, whereas in [4, Thm. 5.6] the hypothesis refers to the first $t-1$ powers (but merely presupposes that they are at a fixed proportion).

   However, as we observe and is detailed below, the actual proof of [4, Thm. 5.6] supports a generalization in which the number of powers is $d-1$, where d and t are free parameters. Hence, we may use $d = t-1$ (for our application) rather than $d = t$ (as in [4, Thm. 5.6]).

We now turn to the actual presentation of [4], but do so using slightly different notation.[5] It refers to N balls, where each ball has a *color*, and there are n colors. The presentation starts from a histogram that describes the frequencies of colors that appear in a specific number of balls; that is, for natural numbers $a_1 < a_2 < \ldots < a_t$ and non-negative $p_1, \ldots, p_t$ that sum-up to 1, a p_i fraction of the colors each occur in a_i balls (i.e., $|C_i| = p_i \cdot n$ and for each $\chi \in C_i$ there are a_i balls that have color χ).

The actual presentation of [4] starts with a random variable Φ that ranges over $\{a_1, \ldots, a_t\} \subset \mathbb{N}$, and lets $p_i = \Pr[\Phi = a_i]$. Given Φ and an integer N, it defines the following instance of the *colored balls* problem, denoted $B_{\Phi,N}$: For each $i \in [t]$, there are $\lfloor Np_i/\mathrm{E}[\Phi] \rfloor$ colors of type i such each color of type i occurs in a_i balls. In our case, the p_i's are multiples of $1/n$ and $N = \sum_{i\in[t]} p_i \cdot n \cdot a_i$ is an integer, which implies that

$$\frac{Np_i}{\mathrm{E}[\Phi]} = p_i \cdot \frac{\sum_{j\in[t]} p_j \cdot n \cdot a_j}{\sum_{j\in[t]} p_j \cdot a_j} = p_i \cdot n$$

is an integer (and there is no need additional tweaks as in [4]). That is, there are $n_i = p_i n$ colors of type i, and the total number of balls is $\sum_{i\in[t]} n_i \cdot a_i$, which equals $2m$ in our case. We next state a generalization of [4, Thm. 5.6], in which the hypothesis refers to the first $d-1$ powers of Φ_1 and Φ_2, while noting that in [4, Thm. 5.6] $d = t$ (whereas in our application $d = t-1$).

[5] For example, we replace n by N (as denoting the number of balls), replace k by t, and $(a_1, \ldots, a_t)$ by $(a_0, \ldots, a_{k-1})$. The number of colors is implicit in [4], but is explicit here.

Lemma 6 (a generalization of [4, Thm. 5.6], slightly rephrased):[6] *Let* Φ_1 *and* Φ_2 *be random variables over positive integers* $a_1 < a_2 < \ldots < a_t$ *such that*

$$\frac{\mathrm{E}[\Phi_1]}{\mathrm{E}[\Phi_2]} = \frac{\mathrm{E}[\Phi_1^2]}{\mathrm{E}[\Phi_2^2]} = \ldots = \frac{\mathrm{E}[\Phi_1^{d-1}]}{\mathrm{E}[\Phi_2^{d-1}]}. \tag{3}$$

Then, for $s \leq \frac{N}{2a_t}$, *the distinguishing gap between* $B_{\Phi_1,N}$ *and* $B_{\Phi_2,N}$ *as judged by any label-invariant algorithm that takes* $\mathtt{Poi}(2s)$ *samples is upper-bounded by*

$$O\left(\frac{t \cdot d \cdot 2s}{N/a_t} + \frac{d}{\lfloor d/2 \rfloor! \cdot \lceil d/2 \rceil!} \cdot \frac{(2s)^d}{(N/a_t)^{d-1}}\right). \tag{4}$$

Lemma 5 follows from Lemma 6 by using $\Phi_1 = X$ and $\Phi_2 = Y$, observing that $N = 2m$ and $B_{\Phi_1,N} \equiv P$ (resp., $B_{\Phi_2,N} \equiv Q$), setting $d = t-1$, simplifying the upper bound, and accounting for the error term of $\exp(-\Omega(s))$.

Recall that Lemma 6 generalizes [4, Thm. 5.6] by allowing d and t to be arbitrary natural numbers rather than mandating that $d = t$. However, the proof of [4, Thm. 5.6] does not use $d = t$ in an essential manner, and so going over that proof one merely needs to keep track of when k stands for t and when it stands for d (and observe that in all places a_{k-1} merely stands for the maximal a_i).[7] In particular, denoting $a = \max_{i \in [t]}\{a_i\}$, the upper bound in [4, Lem. 5.9] is $\delta_1 \stackrel{\mathrm{def}}{=} O(\frac{a^{d-1}}{d!} \cdot \frac{(2s)^d}{N^{d-1}})$, the upper bound in [4, Lem. 5.10] is $\delta_2 \stackrel{\mathrm{def}}{=} \frac{2t \cdot a \cdot 2s}{N}$, the upper bound δ_3 in [4, Lem. 5.12] is $\Theta(1/d)$ of the bound in Eq. (4), and the final upper bound is $2 \cdot \delta_1 + 2 \cdot \delta_2 + (d-1) \cdot \delta_3$, which matches Eq. (4).

[6]In the case of $d = t$, our rephrasing is merely notational (e.g., $(a_1, \ldots, a_t)$ replaces $(a_0, \ldots, a_{k-1})$, and N replaces n). In addition, we incorporate Eq. (3) in our formulation of the lemma rather than referring to a notion (i.e. "proportional moments") defined before, and avoid a notation for the gap of an algorithm (i.e., a notation as in Footnote 3 is avoided in Eq. (4)).

[7]Recall that the parameter s in [4] is replaced here by $2s$, and n is replaced by N.