

We state here a few structural results about (discrete) monotone distributions and their consequences; starting with the very useful fact that they admit a *succinct* approximation, itself monotone, close in total variation distance. Most results here are for discrete distributions, although they admit a continuous counterpart: every non-trivial result stated in this note is from [DDS⁺12, DDS14] (for the discrete case, and the implications for testing and learning).¹

1 Structural result: Birgé’s oblivious decomposition

Definition 1 (Oblivious decomposition). Given a parameter $\varepsilon > 0$, the corresponding *oblivious decomposition* of $[k]$ is the partition \mathcal{I}_ε in ℓ consecutive intervals I_1, \dots, I_ℓ , where $\ell = \Theta\left(\frac{\log(\varepsilon k + 1)}{\varepsilon}\right)$ is the smallest integer such that $\sum_{j=1}^\ell \lfloor (1 + \varepsilon)^j \rfloor \geq k$ and² $|I_j| = \lfloor (1 + \varepsilon)^j \rfloor$ for $1 \leq j < \ell$.

For a distribution p and parameter ε , define $\Phi_\varepsilon(p)$ to be the *flattened distribution* with respect to the oblivious decomposition \mathcal{I}_ε :

$$\forall j \in [\ell], \forall i \in I_j, \quad \Phi_\varepsilon(p)(i) = \frac{p(I_j)}{|I_j|} \quad (1)$$

Note that while $\Phi_\varepsilon(p)$ (obviously) depends on p , the partition \mathcal{I}_ε itself does not; in particular, it can be computed *prior to getting any sample or information about p* .

Theorem 2 ([Bir87, DDS⁺12]). *If p is monotone non-increasing, then $d_{\text{TV}}(p, \Phi_{c\varepsilon}(p)) \leq \varepsilon$ (where $c > 0$ is an absolute constant). Furthermore, $\Phi_{c\varepsilon}(p)$ is also monotone non-increasing.*

Corollary 3 (Robustness). *Suppose p is ε -close to monotone non-increasing. Then $d_{\text{TV}}(p, \Phi_{c\varepsilon}(p)) \leq 2\varepsilon + \alpha$. Furthermore, $\Phi_{c\varepsilon}(p)$ is also ε -close to monotone non-increasing.*

Proof. Let q be a monotone non-increasing distribution such that $d_{\text{TV}}(p, q) \leq \varepsilon$. By the triangle inequality,

$$d_{\text{TV}}(p, \Phi_{c\varepsilon}(p)) \leq d_{\text{TV}}(p, q) + d_{\text{TV}}(q, \Phi_{c\varepsilon}(q)) + d_{\text{TV}}(\Phi_{c\varepsilon}(q), \Phi_{c\varepsilon}(p)) \leq \varepsilon + \alpha + d_{\text{TV}}(\Phi_{c\varepsilon}(q), \Phi_{c\varepsilon}(p))$$

where the last inequality uses the assumption on p and **Theorem 2** applied to it. It only remains to

¹We state all the results for non-increasing probability distributions: a straightforward analogue holds for non-decreasing ones.

²We will often ignore the floors in the definition of the oblivious partition, to avoid more cumbersome analyses and the technicalities that would otherwise arise. However, note that this does not affect the correctness of the proofs: after the first $\approx \frac{1}{\varepsilon} \log \frac{1}{\varepsilon}$ intervals, we do have indeed $|I_{k+1}| \in [1 + \frac{\varepsilon}{2}, 1 + 2\varepsilon] |I_k|$. This multiplicative property, in turn, is the key aspect we shall rely on.

bound the last term: by definition,

$$\begin{aligned}
2d_{\text{TV}}(\Phi_{c\alpha}(q), \Phi_{c\alpha}(p)) &= \sum_{i=1}^k |\Phi_{c\alpha}(p)(i) - \Phi_{c\alpha}(q)(i)| = \sum_{j=1}^{\ell} \sum_{i \in I_j} |\Phi_{c\alpha}(p)(i) - \Phi_{c\alpha}(q)(i)| \\
&= \sum_{j=1}^{\ell} \sum_{i \in I_j} \left| \frac{p(I_j) - q(I_j)}{|I_j|} \right| = \sum_{j=1}^{\ell} |p(I_j) - q(I_j)| = \sum_{j=1}^{\ell} \left| \sum_{i \in I_j} (p(i) - q(i)) \right| \\
&\leq \sum_{j=1}^{\ell} \sum_{i \in I_j} |p(i) - q(i)| = 2d_{\text{TV}}(p, q) \leq 2\varepsilon
\end{aligned}$$

(showing in particular the second part of the claim, as $\Phi_{c\alpha}(q)$ is monotone) and thus

$$d_{\text{TV}}(p, \Phi_{c\alpha}(p)) \leq 2\varepsilon + \alpha$$

as claimed. \square

One can interpret this corollary as saying that the Birgé decomposition provides a tradeoff between becoming *simpler* (and at least as close to monotone) while not staying too far from the original distribution.

Incidentally, the last step of the proof above implies the following easy fact:

Fact 4. For all $\alpha \in (0, 1]$,

$$d_{\text{TV}}(\Phi_{c\alpha}(p), \Phi_{c\alpha}(q)) \leq d_{\text{TV}}(p, q) \quad (2)$$

and in particular, for any property \mathcal{P} left invariant by the Birgé transformation (such as monotonicity)

$$d_{\text{TV}}(\Phi_{c\alpha}(p), \mathcal{P}) \leq d_{\text{TV}}(p, \mathcal{P}) \quad (3)$$

(where, as usual, $d_{\text{TV}}(p, \mathcal{P}) := \inf_{q \in \mathcal{P}} d_{\text{TV}}(p, q)$).

2 Consequences for learning (density estimation)

Write $\Phi_{\text{mon}}(k, \varepsilon, \delta)$ for the sample complexity of learning discrete *non-increasing* distributions over a known domain of size k , to accuracy $\varepsilon > 0$, with error probability $\delta \in (0, 1]$.

Theorem 5. $\Phi_{\text{mon}}(k, \varepsilon, \delta) = \Theta\left(\frac{\log(\varepsilon k + 1)}{\varepsilon^3} + \frac{\log(1/\delta)}{\varepsilon^2}\right)$, where the lower bound hold for $\varepsilon \gtrsim 1/k$.

Note that when $\varepsilon \ll 1/k$, this becomes $\Theta((k + \log(1/\delta))/\varepsilon^2)$ and we get back to the sample complexity of learning *arbitrary* distributions over a known domain of size k . (I.e., for very small ε , the oblivious decomposition becomes trivial and does not buy us anything.)

Proof. The upper bound is a direct consequence of [Theorem 2](#) and the sample complexity of learning (arbitrary) discrete distributions. Specifically, letting $\ell(k, \varepsilon)$ denote the number of intervals in the obvious decomposition with parameters k and $c\frac{\varepsilon}{2}$, we have (i) $d_{\text{TV}}\left(p, \Phi_{c\varepsilon/2}(p)\right) \leq \frac{\varepsilon}{2}$ and (ii) $\Phi_{c\varepsilon/2}(p)$ being a distribution fully determined by the $\ell(k, \varepsilon)$ values $p(I_1), \dots, p(I_{\ell(k, \varepsilon)})$. Let q be the distribution on $[\ell(k, \varepsilon)]$ with $q(i) = p(I_i)$ for all $i \in [\ell(k, \varepsilon)]$.

Given n i.i.d. samples from p , it is immediate to generate n i.i.d. samples from q (simply map each sample from p to the index of the interval I_i it fell into). Thus, with $O\left(\frac{\ell(k, \varepsilon) + \log(1/\delta)}{\varepsilon^2}\right)$ samples from p , one can learn q to total variation distance $\frac{\varepsilon}{2}$ with probability at least $1 - \delta$, obtaining a hypothesis distribution \hat{q} on $[\ell(k, \varepsilon)]$. Now, one can transform this \hat{q} into a piecewise-constant distribution \hat{p} on $[k]$ in the natural way:

$$\forall j \in [\ell], \forall i \in I_j, \quad \hat{p}(i) = \frac{\hat{q}(j)}{|I_j|}. \quad (4)$$

It is immediate to see that, by construction, $d_{\text{TV}}(\Phi_{c\varepsilon/2}(p), \hat{p}) = d_{\text{TV}}(q, \hat{q})$ (as they both are piecewise constant on the same partition). Therefore, we get

$$d_{\text{TV}}(p, \hat{p}) \leq d_{\text{TV}}(p, \Phi_{c\varepsilon/2}(p)) + d_{\text{TV}}(\Phi_{c\varepsilon/2}(p), \hat{p}) \leq 2 \cdot \frac{\varepsilon}{2}$$

with probability at least $1 - \delta$.

The lower bound also follows from the learning lower bound for (arbitrary) discrete distributions, with one small caveat. Namely, it uses the same construction over domain $[\ell(k, c'\varepsilon)]$ analyzed with Assouad's lemma, then “ported over” domain $[k]$ via the Birgé oblivious decomposition. In more detail, letting $\ell := \ell(k, c'\varepsilon)$ (for some constant $c' > 0$) be as before the number of intervals in the corresponding Birgé's decomposition, the lower bound construction is the family of $2^{\ell/2}$ distributions $\{q_z\}_{z \in \{-1, 1\}^{\ell/2}}$ over $[\ell]$ such that

$$\forall i \in [\ell/2], \quad q_z(2i - 1) = \frac{1 + 2z_i\varepsilon}{\ell}, \quad p_z(2i) = \frac{1 - 2z_i\varepsilon}{\ell}, \quad (5)$$

(which we will sometimes refer to as the “Paninski construction,” in view of its use in [Pan08].) Similarly as for the upper bound (cf. (4)), one can map these distributions in a distance-preserving fashion to piecewise-constant distributions $\{p_z\}_{z \in \{-1, 1\}^{\ell/2}}$ over $[k]$, so the lower bound of $\Omega\left(\frac{\ell + \log(1/\delta)}{\varepsilon^2}\right)$ samples obtained by analyzing this family yields a lower bound for our problem. The only subtlety is to make sure the distributions $\{p_z\}_{z \in \{-1, 1\}^{\ell/2}}$ obtained by this mapping are indeed monotone non-increasing. This can be done by choosing a suitable constant c' in the Birgé decomposition, to ensure that $\frac{1 - 2z_i\varepsilon}{|I_j|} \geq \frac{1 + 2z_{i+1}\varepsilon}{|I_{j+1}|}$ for all $1/\varepsilon \lesssim j < \ell$. \square

3 Consequences for testing and distance estimation

Let's now turn to testing, starting with identity and closeness testing (a.k.a. one- and two-sample testing). Actually, let us first look at the most fundamental case, *uniformity* testing: without any monotonicity assumption, this has sample complexity $\Theta\left(\frac{\sqrt{k}}{\varepsilon^2}\right)$ [Pan08, VV17]:

Theorem 6. *Given i.i.d. samples from an arbitrary monotone non-increasing distribution p over $[k]$, testing whether $p = \mathbf{u}$ vs. $d_{\text{TV}}(p, \mathbf{u}) > \varepsilon$ (with success probability at least $2/3$) has sample complexity $\Theta(1/\varepsilon^2)$, independent of k .*

Proof. The idea is actually quite simple: we will show that the only way a monotone distribution can be far from uniform is by “putting significantly too much probability weight on the first half of

the domain.” To do so, assume without loss of generality that k is even.³ Let $H := \{1, \dots, k/2\}$. On the one hand, if $p = \mathbf{u}$, then $p(H) = 1/2$. On the other hand, if $d_{\text{TV}}(p, \mathbf{u}) > \varepsilon$, then by definition of total variation distance there the set $S := \{i \in [k] : p(i) > \mathbf{u}(i)\}$ is such that $p(S) > \mathbf{u}(S) + \varepsilon = |S|/k + \varepsilon$. By monotonicity, this set S is an interval of the form $\{1, \dots, |S|\}$ (since $p(m) > 1/k$ implies $p(m-1) > 1/k$).

- If $|S| \leq k/2$, then by monotonicity $p(H^c) \leq \frac{k}{2} \cdot \frac{1-p(S)}{k-|S|} < \frac{1}{2} \cdot \frac{1-\frac{|S|}{k}-\varepsilon}{1-\frac{|S|}{k}} = \frac{1}{2} \left(1 - \frac{\varepsilon}{1-\frac{|S|}{k}}\right) \leq \frac{1}{2} - \frac{\varepsilon}{2}$
- If $|S| > k/2$, then by monotonicity $p(H) \geq \frac{k}{2} \cdot \frac{p(S)}{|S|} > \frac{1}{2} + \varepsilon \cdot \frac{k}{2|S|} \geq \frac{1}{2} + \frac{\varepsilon}{2}$.

In both cases, we have $p(H) > \frac{1}{2} + \frac{\varepsilon}{2}$. This shows that, mapping each sample from p to the half of the domain it belongs to, the problem boils down to distinguishing a fair coin ($p(H) = 1/2$) from an $\frac{\varepsilon}{2}$ -biased one ($p(H) > \frac{1}{2} + \frac{\varepsilon}{2}$). This can be performed with $O(1/\varepsilon^2)$ samples, giving the upper bound. (The lower bound also follows by a straightforward reduction to this very same biased coin question.) \square

This looks quite surprising: the shape assumption (monotonicity) allowed us to entirely get rid of the dependence on the domain size when testing for uniformity! Since uniformity testing is a special case of the more general *identity* testing, and in the arbitrary (non-monotone) it is known that the two are equivalent [DK16, Gol16, ACT19], one can wonder what happens next... As it turns out, maybe not exactly what one would expect.

Theorem 7. *Given a known reference (monotone non-increasing) distribution q on $[k]$ and i.i.d. samples from an arbitrary monotone non-increasing distribution p over $[k]$, testing whether $p = q$ vs. $d_{\text{TV}}(p, q) > \varepsilon$ (with success probability at least $2/3$) has sample complexity $\Theta(\sqrt{\log(\varepsilon k + 1)}/\varepsilon^{5/2})$, where the lower bound hold for $\varepsilon \gtrsim 1/k$.*

For the proof of this theorem (and the next results), it will be convenient to introduce one more piece of notation. Recall that the flattened distribution of a distribution p (for parameters k, ε) was a piecewise-constant distribution $\Phi_\varepsilon(p)$ on $[k]$ with $\ell := |\mathcal{I}_\varepsilon|$ pieces (cf. (1)). We now also define the *reduction* $\Psi_\varepsilon(p)$ of a distribution p with regard to the Birgé decomposition as the corresponding distribution on $[\ell]$:

$$\forall j \in [\ell], \quad \Psi_\varepsilon(p)(i) = p(I_j). \quad (6)$$

It is immediate to check (and we have done so before, implicitly) that $d_{\text{TV}}(\Psi_\varepsilon(p), \Psi_\varepsilon(q)) = d_{\text{TV}}(\Phi_\varepsilon(p), \Phi_\varepsilon(q))$ for any two distributions p, q on $[k]$. With this in hand, we can proceed with the proof.

Proof of Theorem 9. We will invoke the Birgé decomposition and Theorem 2: with the right parameter $\alpha \asymp \varepsilon$, we have that $p' := \Phi_\alpha(p)$ and $q' := \Phi_\alpha(q)$ are such that: (i) $d_{\text{TV}}(p, p') \leq \frac{\varepsilon}{3}$, (ii) $d_{\text{TV}}(q, q') \leq \frac{\varepsilon}{3}$, (iii) p', q' are both monotone, (iv) q' is explicitly known, and (v) i.i.d. samples from p can be mapped to i.i.d. samples from p' (here, relying crucially on the fact that the decomposition is oblivious and explicit). Further, if $p = q$ then clearly $p' = q'$, while by the triangle inequality if $d_{\text{TV}}(p, q) > \varepsilon$ then $d_{\text{TV}}(p', q') > \varepsilon - 2 \cdot \frac{\varepsilon}{3} = \frac{\varepsilon}{3}$. It is therefore sufficient to test identity of p' and q' with parameter $\frac{\varepsilon}{3}$. But by the above discussion, this is equivalent to testing identity

³Otherwise, one can, e.g., treat separately the last element of the domain and test that $p(k) \ll \varepsilon$, which is a little annoying but does not affect the sample complexity.

of $p'' := \Psi_\alpha(p)$ and $q'' := \Psi_\alpha(q)$ with parameter $\frac{\varepsilon}{3}$: since these two distributions have domain $\ell := \ell(k, \varepsilon/3)$, this can be done with $O(\sqrt{\ell}/\varepsilon^2) = O(\sqrt{\log(\varepsilon k + 1)}/\varepsilon^{5/2})$ samples, as claimed.

The lower bound, as for [Theorem 5](#), follows directly from the corresponding testing lower bound construction and analysis for arbitrary distributions (over support size $[\ell]$), namely, the ‘‘Paninski family’’ defined in [\(5\)](#). Here too, one only needs to set the parameters of the Birgé partition so that when mapping the distributions of this family (which are over $[\ell]$) to the larger domain $[k]$, the resulting probability distributions are monotone non-increasing. \square

In summary, we showed the following:

$$p \stackrel{?}{=} q, \text{ with } k, \varepsilon \longrightarrow \Phi_\alpha(p) \stackrel{?}{=} \Phi_\alpha(q), \text{ with } k, \varepsilon/3 \longleftrightarrow \Psi_\alpha(p) \stackrel{?}{=} \Psi_\alpha(q), \text{ with } \ell, \varepsilon/3$$

which is the template of reductions we will rely on again for the next two results.

Remark 8. Note that, again, when $\varepsilon \ll 1/k$, the sample complexity of [Theorem 9](#) becomes $\Theta(\sqrt{k}/\varepsilon^2)$ and we get back to the sample complexity of identity testing *arbitrary* distributions over a known domain of size k . But, interestingly, under the monotonicity assumption identity testing is no longer equivalent to uniformity testing! Roughly, this is because the aforementioned reductions from [\[DK16, Gol16, ACT19\]](#) do not preserve monotonicity, so one cannot use [Theorem 6](#) at the end.

Now, unto two-sample (i.e., closeness) testing:

Theorem 9. *Given i.i.d. samples from two arbitrary monotone non-increasing distributions p, q over $[k]$, testing whether $p = q$ vs. $d_{\text{TV}}(p, q) > \varepsilon$ (with success probability at least $2/3$) has sample complexity $O\left(\max\left(\log(\varepsilon k + 1)^{2/3}/\varepsilon^2, \sqrt{\log(\varepsilon k + 1)}/\varepsilon^{5/2}\right)\right)$ and, as long as $\varepsilon \gtrsim 1/k$, $\Omega\left(\max\left(\log(\varepsilon k + 1)^{2/3}/(\log \log(\varepsilon k + 1)\varepsilon^{4/3}), \sqrt{\log(\varepsilon k + 1)}/\varepsilon^{5/2}\right)\right)$.*

Proof. The upper bound follows verbatim the same template of reductions, and the argument of [Theorem 9](#) goes through with the difference that now both p, q are unknown: so that the result invoked at the end is the fact that closeness testing for two arbitrary distributions over $[\ell]$ has sample complexity $\Theta\left(\max\left(\ell^{2/3}/\varepsilon^{4/3}, \sqrt{\ell}/\varepsilon^2\right)\right)$ [\[CDVV14\]](#).

The lower bound, however, is not as straightforward: indeed, the lower bound constructions for arbitrary distributions over $[\ell]$ from [\[Val08, CDVV14\]](#), which we would like to use as a near-blackbox similarly to what we did in [Theorem 5](#) and [Theorem 9](#) (there with the Paninski construction), are not immediately amenable to this. However, we can rely on the more involved reduction from [\[DDS⁺12, Proposition 6\]](#), here adapted for the case of monotone distributions: Let $\rho := \frac{\rho_{\max}}{\rho_{\min}}$, $\alpha := c \cdot \frac{\varepsilon}{2}$ (where $c > 0$ is as in [Theorem 2](#)), and $\lambda_{\rho, \varepsilon} := 1 + \log_{1+\alpha} \rho$. Define the geometric distribution $\gamma_{\rho, \varepsilon}$ on $[\lambda_{\rho, \varepsilon}]$ as the one such that $\gamma_{\rho, \varepsilon}(j) \propto (1 + \alpha)^j$ for $j \in [\lambda_{\rho, \varepsilon}]$. Now, given any distribution p on $[k]$, we can define the distribution $p \odot \gamma_{\rho, \varepsilon}$ on $[\lambda_{\rho, \varepsilon} \cdot k]$ by

$$\forall i \in [k], \forall j \in [\lambda_{\rho, \varepsilon}], \quad p \odot \gamma_{\rho, \varepsilon}(\lambda_{\rho, \varepsilon} \cdot (i - 1) + j) = p(i) \gamma_{\rho, \varepsilon}(j).$$

This seemingly contrived construction has the following key property, easy to check by definition of $\gamma_{\rho, \varepsilon}$ and our choice of $\lambda_{\rho, \varepsilon}$: if $\frac{\max_i p(i)}{\min_i p(i)} \leq \rho$, then $p \odot \gamma_{\rho, \varepsilon}(i - 1) \leq (1 + \alpha)p \odot \gamma_{\rho, \varepsilon}(i)$ for all $2 \leq i \leq \lambda_{\rho, \varepsilon} \cdot k$. With that in hand, we can proceed as in the lower bound proof of [Theorem 9](#), but this time leveraging the $\Omega(m^{2/3}/\varepsilon^{4/3})$ lower bound construction of [\[CDVV14\]](#), straightforwardly modified so that $\rho_{\max} \asymp \varepsilon/k^{2/3}$ and $\rho_{\min} = 1/(2k)$ (instead of 0). The end result loses an

$\frac{1}{\varepsilon} \log \log(\varepsilon k + 1)$ factor due to the first domain expansion of size $\lambda_{\rho, \varepsilon} = 1 + \log_{1+\alpha} \varepsilon \ell$. (The other term of the lower bound follows directly from [Theorem 9](#), since closeness testing is at least as hard as identity testing.) \square

As a final remark, note that this has similar implications for *distance estimation* between two distinct monotone distributions p, q , i.e., the task of estimating $d_{TV}(p, q)$ up to an additive ε . The interested (or bored) reader may check by themselves what upper and lower bound one obtains, from the fact that the corresponding task for *arbitrary* distributions over domain size m has sample complexity $\Theta\left(\frac{m}{\varepsilon^2 \log m}\right)$ [[VV11](#), [JHW18](#)].

References

- [ACT19] Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Communication-constrained inference and the role of shared randomness. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 30–39. PMLR, 2019. [3](#), [8](#)
- [Bir87] Lucien Birgé. On the risk of histograms for estimating decreasing densities. *Ann. Statist.*, 15(3):1013–1022, 1987. [2](#)
- [CDVV14] Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1193–1203. ACM, New York, 2014. [3](#)
- [DDS⁺12] Constantinos Daskalakis, Ilias Diakonikolas, Rocco A. Servedio, Gregory Valiant, and Paul Valiant. Testing k -modal distributions: optimal algorithms via reductions. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1833–1852. SIAM, Philadelphia, PA, 2012. [\(document\)](#), [2](#), [3](#)
- [DDS14] Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning k -modal distributions via testing. *Theory Comput.*, 10:535–570, 2014. [\(document\)](#)
- [DK16] Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *FOCS*, pages 685–694. IEEE Computer Society, 2016. [3](#), [8](#)
- [Gol16] Oded Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:15, 2016. [3](#), [8](#)
- [JHW18] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the l_1 distance. *IEEE Trans. Inf. Theory*, 64(10):6672–6706, 2018. [3](#)
- [Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inform. Theory*, 54(10):4750–4755, 2008. [2](#), [3](#)
- [Val08] Paul Valiant. Testing symmetric properties of distributions. In *STOC’08*, pages 383–392. ACM, New York, 2008. [3](#)

- [VV11] 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 *STOC*, pages 685–694. ACM, 2011. [3](#)
- [VV17] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM J. Comput.*, 46(1):429–455, 2017. [3](#)