In this (short) note, we focus on two techniques used to prove lower bounds for distribution *learning* and *testing*, respectively Assouad's lemma and Le Cam's method. (We do not cover here Fano's lemma, another and somewhat more general result than Assouad's – the interested reader is referred to [Yu97].)

Hereafter, we let $(\Omega, \mathcal{B})$ be a measurable space, and $\Delta(\Omega)$ be the set of all probability distributions on it. Let $\mathrm{d_{TV}}(\cdot,\cdot)$ denote the total variation distance (the theorem would actually apply to any metric d on $\Delta(\Omega)$), and $\mathrm{d_H}(\cdot,\cdot)$ be the *Hellinger distance*, defined as

$$\mathrm{d_H}(D,D') \stackrel{\text{def}}{=} \frac{1}{2}\|\sqrt{D}-\sqrt{D'}\|_2 = \frac{1}{2}\sqrt{\sum_{x\in\Omega}\left(\sqrt{D(x)}-\sqrt{D'(x)}\right)^2} = \sqrt{1-\sum_{x\in\Omega}\sqrt{D(x)D'(x)}}$$

(the last two expressions holding when Ω is countable).

1 Learning Lower Bounds: Assouad's Lemma

Definition 1 (Minimax Risk)**.** Let $\mathcal{C} \subseteq \Delta(\Omega)$ be a family of probability distributions, and $m \geq 1$. The *minimax risk for $\mathcal{C}$ with m samples* (with relation to the total variation distance) is defined as

$$\begin{aligned} R_m(\mathcal{C}) &\stackrel{\text{def}}{=} \inf_{A\in\mathcal{A}_m}\sup_{D\in\mathcal{C}} \mathbb{E}_{s_1,\dots,s_m\sim D}\left[\mathrm{d_{TV}}\left(D,\hat{D}_A\right)\right] \qquad (1)\\ &= \inf_{A\in\mathcal{A}_m}\sup_{D\in\mathcal{C}}\int_{\Omega^m}\mathrm{d_{TV}}(D,A(\mathbf{s}))D^{\otimes m}(d\mathbf{s}) \end{aligned}$$

where $\mathcal{A}_m$ is the set of (deterministic) learning algorithms A which take m samples and output a hypothesis distribution $\hat{D}_A$.

In other terms, $R_m(\mathcal{C})$ is the minimum expected error of any m-sample learning algorithm A when run on the worst possible target distribution (from $\mathcal{C}$) for it. It is immediate from the definition that for any $\mathcal{H} \subseteq \mathcal{C}$, one has $R_m(\mathcal{C}) \geq R_m(\mathcal{H})$.

To prove lower bounds on learning a family $\mathcal{C}$, a very common method is to come up with a (sub)family of distributions in which, as long as a learning algorithm does not take enough samples, there always exist two (far) distributions which still could have yielded indistinguishable "transcripts". In other terms, after running any learning algorithm A on m samples, an adversary can still exhibit two very different distributions (depending on A)[1] that *ought* to be distinguished, yet *could not* possibly have been from only m samples. This is formalized by the following theorem, due to Assouad:

Theorem 2 (Assouad's Lemma [Ass83])**.** *Let $\mathcal{C} \subseteq \Delta(\Omega)$ be a family of probability distributions. Suppose there exists a family of $\mathcal{H} \subseteq \mathcal{C}$ of 2^r distributions and constants $\alpha, \beta > 0$ such that, writing $\mathcal{H} = \{D_z\}_{z\in\{0,1\}^r}$,*

[1] Note that this differs from the standard methodology for proving lower bounds for property testing, where two families of distributions (yes and no-instances) are defined beforehand, and a couple of distributions is "committed to" *before* the algorithm gets to make its move.

(i) for all $x, y \in \{0,1\}^r$, the distance between D_x and D_y is at least proportional to the Hamming distance:

$$\mathrm{d_{TV}}(D_x, D_y) \geq \alpha \|x - y\|_1 \tag{2}$$

(ii) for all $x, y \in \{0,1\}^r$ with $\|x-y\|_1 = 1$, the squared Hellinger distance of D_x, D_y is small:

$$\mathrm{d_H}(D_x, D_y)^2 \leq \beta \tag{3}$$

(or, equivalently, $-\ln(1-\mathrm{h}^2) \leq \ln \frac{1}{1-\beta}$)

Then, for all $m \geq 1$,

$$R_m(\mathcal{H}) \geq \frac{1}{4}\alpha r(1-\beta)^{2m} = \Omega\Big(\alpha r e^{-O(\beta m)}\Big). \tag{4}$$

In particular, to achieve error at most ε, any learning algorithm for $\mathcal{C}$ must have sample complexity $\Omega\Big(\frac{1}{\beta}\log\frac{\alpha r}{\varepsilon}\Big)$.

Remark 3 (High-level idea). Intuitively, every distribution in $\mathcal{H}$ is defined by making r distinct "choices"[2]. With this interpretation, item (i) means that two distributions differing in many choices should be far (so that a learning algorithm has to "figure out" *most* of the choices in order to achieve a small error), while item (ii) requires that two distributions defined by almost the same choices be very close (so that a learning algorithm cannot distinguish them *too easily*).

Remark 4 (Technical detail). The quantity $1 - \mathrm{d_H}(p,q)^2$ is known as the *Hellinger affinity*; as the Hellinger distance satisfies

$$1 - \sqrt{1 - \mathrm{d_{TV}}(p,q)^2} \leq \mathrm{d_H}(p,q)^2 \leq \mathrm{d_{TV}}(p,q) \tag{5}$$

it is sufficient for (3) to show that the (sometimes easier) condition holds:

$$\mathrm{d_{TV}}(D_x, D_y) \leq \beta.$$

Note that, with (2) this imposes that $\alpha \leq \beta$; while working with the Hellinger distance only requires $\alpha^2 \leq 2\beta - \beta^2$ (from (5) and (2)).

An example of application. To prove a lower bound of $\Omega\Big(\frac{\log n}{\varepsilon^3}\Big)$ for learning monotone distributions over $[n]$, Birgé [Bir87] invokes Assouad's Lemma, defining a family $\mathcal{H}$ achieving parameters $r = \Theta\Big(\frac{\log n}{\varepsilon}\Big)$, $\alpha = \Theta(\varepsilon/r)$ and $\beta = \Theta(\varepsilon^2/r)$. This example shows a very neat feature of Assouad's Lemma – *it enables us to get a dependence on ε in the lower bound.*

2 Testing Lower Bounds: Le Cam's Method

We now turn to another lower bound technique, better suited for proving lower bounds on property testing or parameter estimation – i.e., where the quantity of interest is a functional of the unknown distribution, instead of the distribution itself. We begin with some terminology that will be useful in stating the main result of this section.

[2] E.g., by choosing, for each of r intervals partitioning the support, whether the distribution (a) is uniform on the interval or (b) puts all its weight on the first half of the interval.

Definition 5. Let $\mathcal{C} \subseteq \Delta(\Omega)$ be a family of probability distributions over Ω, and $m \geq 1$. The *convex hull of m-product distributions from $\mathcal{C}$*, denoted $\operatorname{conv}_m(\mathcal{C})$, it the set of probability distributions over Ω^q defined as

$$\operatorname{conv}_m(\mathcal{C}) \stackrel{\text{def}}{=} \left\{ \sum_{k=1}^{\ell} \alpha_k D_k^{\otimes m} \,:\, \ell \geq 1, D_1, \dots, D_\ell \in \mathcal{C}, \alpha_1, \dots, \alpha_\ell \geq 0, \sum_{k=1}^{\ell} \alpha_k = 1 \right\}.$$

That is, $\operatorname{conv}_m(\mathcal{C})$ is the set of mixtures of m-wise product distributions from $\mathcal{C}$. (Note that distributions in $\operatorname{conv}_m(\mathcal{C})$ are not in general product distributions themselves.)

Definition 6 (Estimator)**.** Let $\mathcal{C} \subseteq \Delta(\Omega)$ be a family of probability distributions over Ω, and $m \geq 1$. For any real-valued functional $\varphi\colon \mathcal{C} \to [0,1]$ ("scalar property"), we denote by $\mathcal{E}_m$ the set of *estimators* for φ: that is, the set of (deterministic) algorithms E taking $m \geq 1$ independent samples from a distribution $D \in \mathcal{C}$ and outputting an estimate $\hat{\varphi}_E$ of $\varphi(D)$.

We state the following lemma for estimators taking value in $[0,1]$ endowed with the distance $|\cdot|$, but it holds for more general metric spaces, and in particular for $([0,1], \|\cdot\|_2)$.

Theorem 7 (Le Cam's Method [LC73, LC86, Yu97])**.** *Let $\mathcal{C} \subseteq \Delta(\Omega)$ be a family of probability distributions over Ω, and let $\varphi\colon \mathcal{C} \to [0,1]$ be a scalar property. Suppose there exists $\gamma \in [0,1]$, subsets $A_1, A_2 \subseteq [0,1]$, and families $\mathcal{D}_1, \mathcal{D}_2 \subseteq \mathcal{C}$ such that the following holds.*

(i) A_1 and A_2 are γ-separated*: $|\alpha_1 - \alpha_2| \geq \gamma$ for all $\alpha_1 \in A_1, \alpha_2 \in A_2$;*

(ii) $\varphi(\mathcal{D}_1) \subseteq A_1$ and $\varphi(\mathcal{D}_2) \subseteq A_2$.

Then, for all $m \geq 1$,

$$\inf_{E \in \mathcal{E}_m} \sup_{D \in \mathcal{C}} \mathbb{E}_{s_1,\dots,s_m \sim D}[|\hat{\varphi}_E - \varphi(D)|] \geq \frac{\gamma}{2}\Big(1 - \inf_{\substack{p_1 \in \operatorname{conv}_m(\mathcal{D}_1)\\ p_2 \in \operatorname{conv}_m(\mathcal{D}_2)}} \mathrm{d}_{\mathrm{TV}}(p_1, p_2)\Big). \tag{6}$$

One particular interest of this result is that the infimum is taken over the *convex hull* of the m-fold product distributions from the families $\mathcal{D}_1$ and $\mathcal{D}_2$, and not over the m-fold distributions themselves. While this makes the computations much less straightforward (as a mixture of product distributions is not in general itself a product distribution, one can no longer rely on using the Hellinger distance as a proxy for total variation and leverage its nice properties with regard to product distributions), it also usually yields much tighter bounds – as the infimum over the convex hull is often significantly smaller.

We now state an immediate corollary in terms of property testing, where a testing algorithm is said to *fail* if it outputs ACCEPT on a no-instance or REJECT on a yes-instance. Note as usual that if the samples originate from a distribution which is neither a yes nor no-instance, then the any output is valid and the tester cannot fail.

Corollary 8. *Fix $\varepsilon \in (0,1)$, and a property $\mathcal{P} \subseteq \Delta(\Omega)$. Let $\mathcal{D}_1, \mathcal{D}_2 \subseteq \Delta(\Omega)$ be families of respectively* yes- *and* no-*instances, i.e. such that $\mathcal{D}_1 \subseteq \mathcal{P}$, while any $D \in \mathcal{D}_2$ has $\mathrm{d}_{\mathrm{TV}}(D, \mathcal{P}) > \varepsilon$. Then, for all $m \geq 1$,*

$$\inf_{T \in \mathcal{T}_m} \sup_{D \in \Delta(\Omega)} \Pr_{s_1,\dots,s_m \sim D}[\, T(s_1,\dots,s_m) \text{ fails}\,] \geq \frac{1}{2}\Big(1 - \inf_{\substack{p_1 \in \operatorname{conv}_m(\mathcal{D}_1)\\ p_2 \in \operatorname{conv}_m(\mathcal{D}_2)}} \mathrm{d}_{\mathrm{TV}}(p_1, p_2)\Big). \tag{7}$$

where $\mathcal{T}_m$ is the set of (deterministic) testing algorithms T with sample complexity m.

As any (possibly randomized) *bona fide* testing algorithm can only fail with probability $1/3$, the above combined with Yao's Principle implies a lower bound of $\Omega(m)$ as soon as m and $\mathcal{D}_1, \mathcal{D}_2$ satisfy $\inf_{p_1,p_2} \mathrm{d}_{\mathrm{TV}}(p_1, p_2) < 1/3$ in (7).

Proof of Corollary 8. We apply Theorem 7 with the following parameters: $A_1 = \{0\}$, $A_2 = \{1\}$, $\gamma = 1$, and $\varphi\colon D \in \mathcal{C} \mapsto \mathbb{1}_{\mathcal{P}}(D) \in \{0, 1\}$, where $\mathcal{C} = \mathcal{P} \cup \{\, D \in \Delta(\Omega) \;:\; \mathrm{d}_{\mathrm{TV}}(D, \mathcal{P}) > \varepsilon \,\}$ is the set of valid instances. $\square$

An example of application. To prove a lower bound of $\Omega(\sqrt{n}/\varepsilon^2)$ for testing uniformity over $[n]$, Paninski [Pan08] defines the families $\mathcal{D}_1 = \mathcal{P} = \{\mathcal{U}_n\}$ and $\mathcal{D}_2$ as the set of distributions D obtained by perturbing each disjoint pair of consecutive elements $(2i-1, 2i)$ by either $(\frac{\varepsilon}{n}, -\frac{\varepsilon}{n})$ or $(-\frac{\varepsilon}{n}, \frac{\varepsilon}{n})$ (for a total of $2^{\frac{n}{2}}$ distinct distributions). He then analyzes the total variation distance between $\mathcal{U}_n^{\otimes m}$ and the uniform mixture

$$p \stackrel{\text{def}}{=} \frac{1}{2^{\frac{n}{2}}} \sum_{D \in \mathcal{D}_2} D^{\otimes m}.$$

By an approach similar as that of [Pol03, Section 14.4], Paninski shows that $\inf_{p_2 \in \mathrm{conv}_m(\mathcal{D}_2)} \mathrm{d}_{\mathrm{TV}}(\mathcal{U}_n^{\otimes m}, p_2) \leq \mathrm{d}_{\mathrm{TV}}(\mathcal{U}_n^{\otimes m}, p) \leq \frac{1}{2}\sqrt{e^{m^2\varepsilon^4/n} - 1}$, which for $m \leq \frac{c\sqrt{n}}{\varepsilon^2}$ is less than $1/3$ – establishing the lower bound.

References

[Ass83] Patrice Assouad. Deux remarques sur l'estimation. *Comptes Rendus des Séances de l'Académie des Sciences. Série I. Mathématique*, 296(23):1021–1024, 1983. 2

[Bir87] Lucien Birgé. Estimating a Density under Order Restrictions: Nonasymptotic Minimax Risk. *The Annals of Statistics*, 15(3):995–1012, 09 1987. 1

[LC73] Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1:38–53, 1973. 7

[LC86] Lucien Le Cam. *Asymptotic methods in statistical decision theory.* Springer Series in Statistics. Springer-Verlag, New York, 1986. 7

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008. 2

[Pol03] David Pollard. Asymptopia. `http://www.stat.yale.edu/~pollard/Books/Asymptopia`, 2003. Manuscript. 2

[Yu97] Bin Yu. Assouad, Fano, and Le Cam. In David Pollard, Erik Torgersen, and Grace L. Yang, editors, *Festschrift for Lucien Le Cam*, pages 423–435. Springer New York, 1997. (document), 7