Lecture 16: Distribution Learning

Han Shao

Readings: Devroye and Lugosi, Sections 5.2, 6.1, and 6.6
Bousquet, Kane, and Moran, Section 5

1 The Distribution-Learning Problem

So far, our hypotheses have usually been classifiers. We now consider *distribution learning*: the observations are unlabeled samples, and the hypothesis is itself a probability distribution.
Let $\mathcal{D}$ be an unknown distribution over a domain $\mathcal{X}$. We observe

$$S = (x_1, \dots, x_m) \sim \mathcal{D}^m$$

and want to return a distribution $Q_S = \mathcal{A}(S)$ that approximates $\mathcal{D}$. The domain may be finite, countably infinite, or a continuous domain $\mathcal{X} \subseteq \mathbb{R}^d$. All sets and functions below are assumed measurable.
We are given a finite, nonempty class of candidate distributions

$$\mathcal{H} = \{Q_1, \dots, Q_N\}.$$

The candidates are known and fixed before the sample is drawn; only the target $\mathcal{D}$ is unknown. In the continuous case, we assume the candidates have probability density functions. We first consider *proper* learners, which must output $Q_S \in \mathcal{H}$.

Measuring Error: Total Variation

Definition 1 (Total variation distance)**.** *For distributions P, Q over the same domain, their total variation distance is*

$$\mathrm{TV}(P, Q) = \sup_{A \subseteq \mathcal{X}} |P(A) - Q(A)|,$$

where the supremum ranges over measurable sets.

Thus $\mathrm{TV}(P, Q) \le \varepsilon$ means that the probabilities assigned to *every event* differ by at most ε. This distance lies in $[0, 1]$ and satisfies the triangle inequality: for every P, Q, R,

$$\mathrm{TV}(P, R) \le \mathrm{TV}(P, Q) + \mathrm{TV}(Q, R).$$

Indeed, apply $|P(A) - R(A)| \le |P(A) - Q(A)| + |Q(A) - R(A)|$ to each event A, and then take the supremum.
If P, Q have probability density functions p, q, define $A_{P,Q} = \{x \in \mathcal{X} : p(x) > q(x)\}$. Then

$$\mathrm{TV}(P, Q) = P(A_{P,Q}) - Q(A_{P,Q}) = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| \, dx.$$

For a finite or countable domain, use the probability mass functions $p(x) = P(\{x\})$ and $q(x) = Q(\{x\})$ instead. The same set attains total variation, with the integral replaced by a sum:

$$\mathrm{TV}(P, Q) = P(A_{P,Q}) - Q(A_{P,Q}) = \frac{1}{2} \sum_{x \in \mathcal{X}} |p(x) - q(x)|.$$

Realizable and Agnostic Learning

In the *realizable* setting, $\mathcal{D} \in \mathcal{H}$, and we seek $\mathrm{TV}(\mathcal{D}, Q_S) \le \varepsilon$ with probability at least $1 - \delta$. In the *agnostic* setting, we do not assume that $\mathcal{D}$ belongs to $\mathcal{H}$. Define the approximation error

$$\mathrm{OPT} = \min_{Q \in \mathcal{H}} \mathrm{TV}(\mathcal{D}, Q).$$

We seek a guarantee of the form

$$\Pr_{S,\mathcal{A}}[\mathrm{TV}(\mathcal{D}, Q_S) \le \alpha\, \mathrm{OPT} + \varepsilon] \ge 1 - \delta, \tag{1}$$

where $\alpha \ge 1$ is the multiplicative approximation factor. The probability includes the learner's internal randomness when it is randomized. Its random bits are independent of the sample.
Our aim is a sample bound depending on N, ε, δ, but not on $|\mathcal{X}|$. We will prove that $\alpha = 3$ is achievable using

$$m = O\left(\frac{\ln N + \ln(1/\delta)}{\varepsilon^2}\right).$$

In particular, for two candidates and constant confidence, this is $O(1/\varepsilon^2)$.

Why not estimate the entire target first? Let $\widehat{\mathcal{D}}_S$ be the empirical distribution, defined by

$$\widehat{\mathcal{D}}_S(A) = \frac{1}{m} \sum_{r=1}^{m} \mathbb{1}\{x_r \in A\}.$$

It need not approximate $\mathcal{D}$ in total variation. For example, when $\mathcal{D}$ has a density on $[0, 1]$, the finite set of sampled points has empirical mass 1 but true mass 0, so $\mathrm{TV}(\widehat{\mathcal{D}}_S, \mathcal{D}) = 1$. Instead of estimating the probabilities of all events, we will estimate only a finite collection of events chosen using the candidates.

2 The Yatracos/Scheffé-Style Algorithm

We use the *minimum-distance* version of the Yatracos/Scheffé method (Yatracos, 1985; Devroye and Lugosi, 2001; Bousquet et al., 2019). Write q_i for the density or probability mass function of Q_i. The pairwise comparison sets form the family

$$\mathcal{Y}_{\mathcal{H}} = \big\{\{x \in \mathcal{X} : q_i(x) > q_j(x)\} : i, j \in [N]\big\}, \qquad |\mathcal{Y}_{\mathcal{H}}| \le N^2.$$

For a candidate $Q \in \mathcal{H}$, define its empirical discrepancy on these sets by

$$\Delta_S(Q) = \max_{A \in \mathcal{Y}_{\mathcal{H}}} \big|Q(A) - \widehat{\mathcal{D}}_S(A)\big|.$$

The algorithm returns a candidate whose probabilities on these tests best match their empirical frequencies:

$$Q_S \in \operatorname*{argmin}_{Q \in \mathcal{H}} \Delta_S(Q). \tag{2}$$

Algorithm 1 Yatracos/Scheffé-style minimum-distance selection

Require: Candidates $\mathcal{H} = \{Q_1, \dots, Q_N\}$ and an i.i.d. sample $S = (x_1, \dots, x_m)$ from $\mathcal{D}$.

1: Form the pairwise comparison family $\mathcal{Y}_{\mathcal{H}}$ defined above.

2: For each $A \in \mathcal{Y}_{\mathcal{H}}$, compute $\widehat{\mathcal{D}}_S(A) = \dfrac{1}{m}\sum_{r=1}^{m} \mathbb{1}\{x_r \in A\}$.

3: For each $i \in [N]$, compute $\Delta_S(Q_i) = \max_{A \in \mathcal{Y}_{\mathcal{H}}} |Q_i(A) - \widehat{\mathcal{D}}_S(A)|$.

4: Return $Q_S \in \operatorname{argmin}_{Q \in \mathcal{H}} \Delta_S(Q)$.

Ties may be broken by the smallest candidate index.
We assume access to the candidate densities or mass functions and the probabilities $Q_i(A)$ appearing in the algorithm. The result below concerns sample complexity; it does not assert that these computations are efficient for every representation of the candidates.

Theorem 1 (Proper 3-approximate distribution learning)**.** *Let $\mathcal{H}$ be a fixed finite class of $N \geq 1$ distributions. For every target distribution $\mathcal{D}$ and every $\varepsilon, \delta \in (0,1)$, algorithm 1 returns $Q_S \in \mathcal{H}$ satisfying*

$$\mathrm{TV}(\mathcal{D}, Q_S) \leq 3 \min_{Q \in \mathcal{H}} \mathrm{TV}(\mathcal{D}, Q) + \varepsilon$$

with probability at least $1 - \delta$, provided that

$$m \geq \frac{2}{\varepsilon^2} \ln\left(\frac{2N^2}{\delta}\right).$$

This sample bound has no dependence on the size of the domain.

Proof. The proof separates a deterministic comparison argument from a uniform-convergence bound. Define

$$\varepsilon_S = \max_{A \in \mathcal{Y}_{\mathcal{H}}} \left|\mathcal{D}(A) - \widehat{\mathcal{D}}_S(A)\right|.$$

Step 1: the best candidate has small discrepancy. Choose

$$Q^\star \in \operatorname*{argmin}_{Q \in \mathcal{H}} \mathrm{TV}(\mathcal{D}, Q), \qquad \mathrm{TV}(\mathcal{D}, Q^\star) = \mathrm{OPT}.$$

For every $A \in \mathcal{Y}_{\mathcal{H}}$,

$$\begin{aligned} |Q^\star(A) - \widehat{\mathcal{D}}_S(A)| &\leq |Q^\star(A) - \mathcal{D}(A)| + |\mathcal{D}(A) - \widehat{\mathcal{D}}_S(A)| \\ &\leq \mathrm{OPT} + \varepsilon_S. \end{aligned}$$

Consequently,

$$\Delta_S(Q_S) \leq \Delta_S(Q^\star) \leq \mathrm{OPT} + \varepsilon_S, \tag{3}$$

where the first inequality is the selection rule.

Step 2: compare the selected candidate with the best one. Let $q_S, q^\star$ denote the densities or mass functions of $Q_S, Q^\star$, and set $A = \{x : q_S(x) > q^\star(x)\}$. Because both distributions belong to $\mathcal{H}$, this set belongs to $\mathcal{Y}_{\mathcal{H}}$, even though Q_S depends on the sample. By the total-variation identity above,

$$\mathrm{TV}(Q_S, Q^\star) = Q_S(A) - Q^\star(A)$$

$$
\begin{aligned}
&\le |Q_S(A) - \widehat{\mathcal{D}}_S(A)| + |\widehat{\mathcal{D}}_S(A) - Q^\star(A)| \\
&\le \Delta_S(Q_S) + \Delta_S(Q^\star) \\
&\le 2\,\mathrm{OPT} + 2\varepsilon_S.
\end{aligned}
$$

One more application of the triangle inequality gives the deterministic bound

$$
\mathrm{TV}(\mathcal{D}, Q_S) \le \mathrm{TV}(\mathcal{D}, Q^\star) + \mathrm{TV}(Q^\star, Q_S) \le 3\,\mathrm{OPT} + 2\varepsilon_S. \tag{4}
$$

Step 3: uniformly estimate the pairwise tests. For any fixed $A \in \mathcal{Y}_\mathcal{H}$, the variables $\mathbb{1}\{x_r \in A\}$ are independent Bernoulli random variables with expectation $\mathcal{D}(A)$. Hoeffding's inequality gives

$$
\Pr_S\left[|\widehat{\mathcal{D}}_S(A) - \mathcal{D}(A)| > \frac{\varepsilon}{2}\right] \le 2\exp\left(-\frac{m\varepsilon^2}{2}\right).
$$

The family $\mathcal{Y}_\mathcal{H}$ is fixed before sampling and has at most N^2 members. A union bound therefore yields

$$
\Pr_S\left[\varepsilon_S > \frac{\varepsilon}{2}\right] \le 2N^2\exp\left(-\frac{m\varepsilon^2}{2}\right) \le \delta.
$$

On the complementary event, substitute $2\varepsilon_S \le \varepsilon$ into (4) to conclude the proof. □

The proof does not estimate $\mathrm{TV}(\mathcal{D}, Q)$ accurately for every candidate. Instead, it estimates only the candidate-dependent event probabilities needed to compare Q_S and $Q^\star$. Under realizability, $\mathrm{OPT} = 0$, so the same bound guarantees $\mathrm{TV}(\mathcal{D}, Q_S) \le \varepsilon$.

3 Why the Factor Three Is Necessary for Proper Learning

The preceding theorem gives a factor-3 guarantee independent of the domain size. The next result shows that this factor cannot be improved uniformly for proper learners, even with only two candidates (Bousquet et al., 2019).

Theorem 2 (Proper-learning lower bound)**.** *For every $\alpha \in [1, 3)$, there exist two fixed distributions $\mathcal{H} = \{Q_1, Q_2\}$ on $[0, 1]$ and a constant $\varepsilon_0 > 0$ such that, for every finite sample size m and every possibly randomized proper learner $\mathcal{A}$, some target distribution $\mathcal{D}$ satisfies*

$$
\Pr_{S,\mathcal{A}}\left[\mathrm{TV}(\mathcal{D}, \mathcal{A}(S)) > \alpha \min_{Q\in\mathcal{H}} \mathrm{TV}(\mathcal{D}, Q) + \varepsilon_0\right] \ge \frac{1}{3}, \qquad S \sim \mathcal{D}^m.
$$

The candidates and ε_0 are fixed independently of m; the hard target may depend on m and the learner. Thus no finite sample bound can provide the stated approximation uniformly over targets with failure probability less than $1/3$.

Proof sketch. We describe the construction and the indistinguishability argument; the routine density calculations are summarized below.

Two candidates and hidden local modifications. Choose a small $\beta = 1/(k-1) > 0$, with integer $k \geq 3$, and define the candidate densities by

	$[0, 1/2)$	$[1/2, 1]$
q_1	$1-\beta$	$1+\beta$
q_2	$1+\beta$	$1-\beta$

Partition $[0, 1]$ into $2r$ equal blocks, each containing k equal subintervals. Choose one special subinterval uniformly and independently in each block. A hidden index $b \in \{1, 2\}$ selects the target density pattern:

	Left half		Right half	
	Ordinary	Special	Ordinary	Special
$p_{1,\theta}$	$1-\beta$	2	$1+\beta$	0
$p_{2,\theta}$	$1+\beta$	0	$1-\beta$	2

Here $\theta \in [k]^{2r}$ records the special subintervals, and $\mathcal{D}_{b,\theta}$ denotes the distribution with density $p_{b,\theta}$. Each target assigns mass $1/(2r)$ to every block. Direct calculation gives

$$\mathrm{TV}(\mathcal{D}_{b,\theta}, Q_b) = \frac{\beta}{2}, \qquad \mathrm{TV}(\mathcal{D}_{b,\theta}, Q_{3-b}) = \frac{3\beta}{2} - \frac{2\beta^2}{1+\beta}.$$

Thus the ratio of the wrong candidate's error to OPT is $3 - 4\beta/(1+\beta)$, approaching 3 as $\beta \to 0$.

Why a small sample cannot identify the better candidate. Draw b and θ independently and uniformly, fix the resulting target, and then draw $S \sim \mathcal{D}_{b,\theta}^m$. Let E be the event that no block is sampled twice. A union bound gives

$$\Pr[E^c] \leq \binom{m}{2} \frac{1}{2r}.$$

Averaging over the choice of special subinterval makes the density within each block uniform under either value of b. On E, different observations encounter independent hidden modifications, so the conditional sample laws for $b = 1$ and $b = 2$ coincide. A proper learner must choose one candidate; conditional on E, its average probability of choosing the wrong one is $1/2$. Choose r large enough that $\Pr[E] \geq 2/3$. Averaging then supplies a fixed target for which the wrong candidate is returned with probability at least $1/3$.
Finally, choose $\beta < (3-\alpha)/6$ and set $\varepsilon_0 = \beta^2$. Then

$$\frac{3\beta}{2} - \frac{2\beta^2}{1+\beta} > \alpha\frac{\beta}{2} + \varepsilon_0.$$

Returning the wrong candidate violates the required guarantee. □

Interpreting the lower bound. On a fixed finite domain, sufficiently many samples can recover the target accurately and allow a proper guarantee with factor 1. The obstruction is to a *domain-independent* improvement below 3: the construction can hide modifications at an arbitrarily fine scale. A finite-domain version replaces the subintervals by points; for fixed β, the same collision argument gives an $\Omega(\sqrt{|\mathcal{X}|})$ sample requirement for beating the corresponding factor below 3. The result also applies to randomized proper learners, because each realized output must still be one of the two candidates.

Improper learning: the optimal factor is two. The lower bound above relies on requiring $Q_S \in \mathcal{H}$. If the learner may instead output a distribution outside $\mathcal{H}$, the optimal universal approximation factor improves from 3 to 2. Bousquet et al. (2019) show that, for every finite candidate class $\mathcal{H}$ and every $\varepsilon, \delta \in (0,1)$, an improper learner can ensure

$$\Pr_{S,\mathcal{A}}\left[\mathrm{TV}(\mathcal{D}, Q_S) \le 2 \min_{Q \in \mathcal{H}} \mathrm{TV}(\mathcal{D}, Q) + \varepsilon\right] \ge 1 - \delta,$$

using a number of samples depending only on $|\mathcal{H}|, \varepsilon, \delta$, not on the domain size. The factor 2 is optimal for such domain-independent guarantees: in general, no smaller constant works, even with only two candidates. Thus allowing outputs outside the candidate class improves the best possible approximation factor, not merely the sample complexity.

No VC-style dimension for general distribution learning. Beyond finite candidate classes, one may ask for a VC-style combinatorial characterization of PAC learnability in total variation. For general distribution classes, no such characterization exists under the requirements described below. Specifically, Lechner and Ben-David (2024) rule out scale-invariant characterizations satisfying natural finite-witness requirements, even in the realizable setting with improper learners. These requirements capture the idea that large dimension is witnessed by a finite configuration, as in VC shattering. They also rule out a single integer-valued dimension whose finiteness characterizes learnability and which controls sample complexity. Lechner and Ben-David (2026) extend the impossibility to scale-sensitive combinatorial dimensions, which may depend on the accuracy parameter, under corresponding finite-witness requirements. These results concern universal characterizations; they do not rule out useful sufficient conditions or characterizations for restricted families of distributions. We state these results without proof.

References

Yannis G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13(2):768–774, 1985. doi:10.1214/aos/1176349553.

Luc Devroye and Gábor Lugosi. *Combinatorial Methods in Density Estimation*. Springer Series in Statistics, Springer, 2001. Book PDF.

Olivier Bousquet, Daniel Kane, and Shay Moran. The optimal approximation factor in density estimation. *Proceedings of the Thirty-Second Conference on Learning Theory*, PMLR 99:318–341, 2019. arXiv:1902.05876. The lower-bound sketch follows the corrected construction in Section 5 of the revised arXiv version.

Tosca Lechner and Shai Ben-David. Inherent limitations of dimensions for characterizing learnability of distribution classes. *Proceedings of the Thirty-Seventh Conference on Learning Theory*, PMLR 247:3353–3374, 2024. Paper.

Tosca Lechner and Shai Ben-David. No scale sensitive dimension for distribution learning. *Proceedings of the 37th International Conference on Algorithmic Learning Theory*, PMLR 313:1–27, 2026. Paper.