Dualizing Le Cam's method

David Lin

May 3, 2022

In this paper, I will review the main results in “Dualizing Le Cam's method for functional estimation, with applications to estimating the unseens” by Polyanskiy and Wu [2021].

1 Motivation

First we present the usual setting of Le Cam's two-point method. Let $\{P_\theta\}_{\theta\in\Theta}$ be a class of probability distributions over the parameter set Θ. We would like to estimate some function $T : \Theta \to \mathbb{R}$ using to squared loss $L(\theta, a) = (T(\theta) - a)^2$. We have the separation parameter

$$\Delta \triangleq \min_a L(\theta_0, a) + L(\theta_1, a) = \frac{1}{2}(T(\theta_0) - T(\theta_1))^2. \tag{1}$$

Hence, given n independent samples drawn from either P_{θ_0} or P_{θ_1}, the minimax risk is given by

$$\begin{aligned} R^*(n) &\triangleq \inf_{\widehat{T}} \sup_{\theta\in\{\theta_0,\theta_1\}} \mathbb{E}_\theta[(\widehat{T} - T(\theta)^2] & (2)\\ &\geq \frac{\Delta}{2}\cdot(1 - \|P_{\theta_0} - P_{\theta_1}\|_{\mathrm{TV}}) & (3)\\ &\geq \frac{\Delta}{4}\sqrt{\exp\left(-D_{\mathrm{KL}}(P_{\theta_0}^{\otimes n}\|P_{\theta_1}^{\otimes n})\right)} & (4)\\ &\geq \frac{\Delta}{4}\cdot\frac{1}{\sqrt{1+\chi^2(P_{\theta_0}^{\otimes n}\|P_{\theta_1}^{\otimes n})}} & (5)\\ &= \frac{(T(\theta_0) - T(\theta_1))^2}{16(1+\chi^2(P_{\theta_0}\|P_{\theta_1}))^{n/2}} & (6) \end{aligned}$$

Now, we would like to choose θ_0, θ_1 for the conclusion to be meaningful. Notice that the denominator is essentially exponential in $n\chi^2$, so $\chi^2 \gg 1/n$ gives very weak lower bound. On the other hand, $\chi^2 \ll 1/n$ only improves the denominator by a constant factor. Hence, it is reasonable to try to control for $\chi^2 \approx 1/n$, and we have the bound

$$R^*(n) \leq \frac{1}{8\sqrt{e}} \sup_{\theta_0,\theta_1:\chi^2\leq 1/n} (T(\theta_0) - T(\theta_1))^2. \tag{7}$$

The main theorems suggests that under some technical conditions, this bound is tight up to constant factors.

2 Main Theorem, i.i.d. case

Now we give provide the full setting. We interpret the original hypothesis set $\{P_\theta\}_{\theta\in\Theta}$ as a stochastic kernel $P : \Theta \to \mathcal{X}$, so the class of hypotheses is extended to mixtures of the form πP for $\pi \in \Pi$, where Π is a convex set of distributions over Θ. We would like to estimate an *affine functional* $T(\pi)$ of the parameter distribution. Examples of such functions include (but are not limited to) $T(\pi) = \mathbb{E}_{\theta\sim\pi}[h(\theta)]$ for some $h : \Theta \to \mathbb{R}$.

The argument in the first section suggests defining the following notion:

Definition 1 (Modulus of continuity)

For an affine[a] functional $T : \Pi \to \mathbb{R}$ where $\Pi \subset \mathcal{P}(\Theta)$ is a convex set of probability distributions, define the χ^2*-modulus of continuity* as follows:

$$\delta_{\chi^2}(t) \triangleq \sup\{T(\pi') - T(\pi) : \chi^2(\pi' P \| \pi P) \le t^2, \pi, \pi' \in \Pi\}. \tag{8}$$

In other words, a χ^2-divergence of t^2 translates to an at most $\delta_{\chi^2}(t)$ difference in T (somewhat akin to Lipschitzness).

[a] This just means that $T(\lambda\pi + (1-\lambda)\pi') = \lambda T(\pi) + (1-\lambda)T(\pi')$.

We are thus ready to state the main theorem in the i.i.d. case:

Theorem 1 (Main Theorem, i.i.d. case [Polyanskiy and Wu, 2021])

Under some technical (mostly topological/measure-theoretic) assumptions listed later, we have

$$\frac{1}{(1+\sqrt{e})^2}\delta_{\chi^2}(\tfrac{1}{\sqrt{n}})^2 \le R^*(n) \le \delta_{\chi^2}(\tfrac{1}{\sqrt{n}})^2, \tag{9}$$

so the minimax rate given by the two-point method is tight up to constant factors. The assumptions are that:

A1. Π is convex.

A2. T is affine.

A3. There exists a vector space of functions $\mathcal{F}$ on the observation space $\mathcal{X}$ such that (1) $\mathcal{F}$ contains constant and (2) $\mathcal{F}$ is dense in $L_2(\mathcal{X}, \pi P)$ for every $\pi \in \Pi$.

A4. There exists a topology on Π coarse enough that Π is compact but fine enough that $T(\pi), \pi P f, \pi P(f^2)$ are continuous for all $\pi \in P, f \in \mathcal{F}$.

Furthermore, the upper bound is attained by an estimator of the form

$$\widehat{T}_g = \frac{1}{n}\sum_{i=1}^{n} g(X_i). \tag{10}$$

For the sake of simplicity, we will ignore the specific constant factors and focus on proving the same asymptotic rate given in the theorem.

Before giving the proof, we make a quick comment about the choice of g. In particular, we will only need to consider two types of g (up to scalars): constants, and a g that approaches the supremum in the variational representation of the χ^2 divergence between $\pi P, \pi' P$:

$$\chi^2(\pi P \| \pi' P) = \sup_g \left\{ \frac{(\mathbb{E}_{\pi P}[g] - \mathbb{E}_{\pi' P}[g])^2}{\mathrm{Var}_{\pi' P}[g]} \right\}. \tag{11}$$

This is reflected in our assumptions about the function class $\mathcal{F}$: the supremum will be attained (in the limit) over any $L_2(\mathcal{X}, \pi' P)$-dense subset.

Proof. The lower bound follows the usual two-point argument gives the lower bound, so we will focus on showing the upper bound instead.

The analysis begins by splitting the error (from using $\widehat{T}_g$ of the suggested form) into the bias and variance components:

$$\mathbb{E}_{X_i \sim \pi P}[(\hat{T}_g - T(\pi))^2] \le (\mathbb{E}_{\pi P}[g] - T(\pi))^2 + \frac{1}{n}\mathrm{Var}_{\pi P}[g] \tag{12}$$

$$\le \left(|\mathbb{E}_{\pi P}[g] - T(\pi)| + \frac{1}{\sqrt{n}}\sqrt{\mathrm{Var}_{\pi P}[g]} \right)^2 \tag{13}$$

Hence,

$$\sqrt{R^*(n)} \le \inf_g \sup_{\pi \in \Pi} \left\{ |\mathbb{E}_{\pi P} - T(\pi)| + \frac{1}{\sqrt{n}} \operatorname{Var}_{\pi P}[g] \right\} = \delta_{bv}(\tfrac{1}{\sqrt{n}}). \tag{14}$$

Thus, the claim reduces to showing that for all $t \ge 0$,

$$\delta_{\chi^2}(t) \ge \delta_{bv}(t) \triangleq \inf_g \sup_\pi \{ |\mathbb{E}_{\pi P}[g] - T(\pi)| + t \operatorname{Var}_{\pi P}[g] \} \tag{15}$$

Pretend for a moment that we can swap the inf and the sup. Then, the answer becomes somewhat trivial:

$$\begin{aligned} \inf_g \sup_\pi \{ |\mathbb{E}_{\pi P}[g] - T(\pi)| + t \operatorname{Var}_{\pi P}[g] \} &\overset{?}{=} \sup_\pi \inf_g \left\{ |\mathbb{E}_{\pi P}[g] - T(\pi)| + t \sqrt{\operatorname{Var}_{\pi P}[g]} \right\} && (16) \\ &= 0 && (17) \end{aligned}$$

because we can set g to be the constant $T(\pi)$ (with zero variance). This is, of course, too simplistic to be correct, but there is shred of truth in this. One possible criterion[1] for swapping the inf and sup requires that the inner functional be convex in g and concave in π, but the term $|\mathbb{E}_{\pi P}[g] - T(\pi)|$ is *convex* in π. To remedy this, we define the following "affine relaxation":

$$\begin{aligned} |\mathbb{E}_{\pi P}[g] - T(\pi)| &= \max_{\xi \in \{0,2\}} \mathbb{E}_{\pi P}[g] - T(\pi) - \xi(\mathbb{E}_{\pi P}[g] - T(\pi)) && (18) \\ &\le \max_{\xi \in \{0,2\}, \pi' \in \Pi} \mathbb{E}_{\pi P}[g] - T(\pi) - \xi(\mathbb{E}_{\pi' P}[g] - T(\pi')) && (19) \end{aligned}$$

The resulting functional is affine in π, π' and ξ, so we repeat the same argument (correctly, this time):

$$\begin{aligned} &\inf_g \sup_\pi \left\{ |\mathbb{E}_{\pi P}[g] - T(\pi)| + t \sqrt{\operatorname{Var}_{\pi P}[g]} \right\} && (20) \\ &\le \inf_g \max_{\pi, \pi', \xi \in [0,2]} \mathbb{E}_{\pi P}[g] - T(\pi) - \xi(\mathbb{E}_{\pi' P}[g] - T(\pi')) + t \sqrt{\operatorname{Var}_{\pi P}[g]} && (21) \\ &= \max_{\pi, \pi', \xi \in [0,2]} \inf_g \mathbb{E}_{\pi P}[g] - T(\pi) - \xi(\mathbb{E}_{\pi' P}[g] - T(\pi')) + t \sqrt{\operatorname{Var}_{\pi P}[g]} && (22) \\ &= \max_{\pi, \pi', \xi \in [0,2]} \left\{ T(\pi) - \xi(T(\pi')) + \inf_g \mathbb{E}_{\pi P}[g] - \xi \mathbb{E}_{\pi' P}[g] + t \sqrt{\operatorname{Var}_{\pi P}[g]} \right\} && (23) \end{aligned}$$

Now we do the same trick: if $\xi \ne 1$, then by setting $g \equiv c \to \pm\infty$, the inner term goes to $-\infty$, so we may assume that $\xi = 1$. Since we can always scale g, it suffices to focus on the sign of the remaining terms. Rewrite it as:

$$\begin{aligned} \inf_g \mathbb{E}_{\pi P}[g] - \mathbb{E}_{\pi' P}[g] + t \sqrt{\operatorname{Var}_{\pi P}[g]} &= \inf_g \left(\frac{\mathbb{E}_{\pi P}[g] - \mathbb{E}_{\pi' P}[g]}{\sqrt{\operatorname{Var}_{\pi P}[g]}} + t \right) \cdot \sqrt{\operatorname{Var}_{\pi P}[g]} && (24) \\ &\le \inf_g \left(-\sqrt{\chi^2(\pi P \| \pi' P)} + t + \epsilon \right) \sqrt{\operatorname{Var}_{\pi P}[g^*]} && (25) \end{aligned}$$

where $\epsilon > 0$ and g^* attains the supremum in the variational representation of the χ^2-divergence to within ϵ (with the appropriate choice of signs). This means that if $\chi^2(\pi P || \pi' P) > t^2$, then the infimum term goes to $-\infty$ as we scale $g \to \infty$.

Finally, when $\chi^2(\pi P \| \pi' P) \le t^2$, the term in the inner infimum (over g) is positive, so the minimum is 0. Thus, we obtain

$$\inf_g \sup_\pi \{ |\mathbb{E}_{\pi P}[g] - T(\pi)| + t \operatorname{Var}_{\pi P}[g] \} \le \max_{\pi, \pi' \in \Pi} \{ T(\pi) - T(\pi') : \chi^2(\pi P \| \pi' P) \le t^2 \} \tag{26}$$

as desired.

Remark. We recap where the various assumptions were used:

[1] Ky Fan's theorem (Theorem 5, restated in Polyanskiy and Wu [2021]) implies that for a continuous functional f on $X \times Y$ with X compact, and f is concave in X and convex in Y, then $\max_X \inf_Y f = \inf_Y \max_X f$.

A1-A2. The convexity of Π and affineness of T is needed for the affine relaxation technique to work.

A3. The function space $\mathcal{F}$ where we select P had to be $L_2(\mathcal{X}, \pi P)$-dense so that the supremum in the variational representation of χ^2 was achieved.

A4. Π had to be compact so that all suprema were attainable, and the other functions had to be continuous for the swapping of the maximum and the minimum.

2.1 Toy example

Suppose we have n i.i.d. samples drawn from $\mathsf{Bern}(p)$, and we would like to estimate p. This suggests the following setup: $\Theta = \mathcal{X} = \{0, 1\}$, with trivial kernel P and distribution domain

$$\Pi = \{\mathsf{Bern}(p) : p \in [0, 1]\}.$$

Note that Π is convex, and furthermore the functional in consideration can we written as $p = \mathbb{E}_{X\sim\pi P}[X]$. It is easily checked that the remaining technical hypotheses apply, so we are left with computing $\delta_{\chi^2}(t)$. Note that

$$\chi^2(\mathsf{Bern}(p), \mathsf{Bern}(p')) = \frac{(p - p')^2}{p'(1 - p')}$$

so assuming that $t \leq 1$, $\chi^2 \leq t^2 \implies p - p' \leq t/2$, attained when $(p, p') = ((1 - t)/2, 1/2)$, so $\delta_{\chi^2}(\frac{1}{\sqrt{n}}) = \frac{1}{2\sqrt{n}}$, so $R^*(n) \asymp O(1/n)$, which is the usual parametric rate.

One thing worth considering is also the choice of estimator g suggested by this method. A function that attains the supremum in the χ^2 optimization is the identity $g(X) = X$, so this corresponds to the fact that the mean gives the parametric rate above .

3 Main Theorem, deterministic case

Now, we suppose instead that the samples $\boldsymbol{X} = (X_1, \ldots, X_n)$ are drawn independently with respect to a deterministic parameter set $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ (i.e. $X_i \sim P_{\theta_i}$), and we impose a cost constraint

$$\boldsymbol{\Theta}_c = \left\{\theta \in \Theta^{\otimes n} : \frac{1}{n}\sum_{i=1}^{n} c(\theta_i) \leq 1\right\}$$

for some cost function $c : \Theta \to \mathbb{R}$. The goal is to estimate a functional T of the empirical distribution $\pi_{\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{\theta_i}$, defined by

$$T(\pi_\theta) = \frac{1}{n}\sum_{i=1}^{n} h(\theta_i), \tag{27}$$

up to the usual quadratic loss, for some $h : \Theta \to \mathbb{R}$. (We will also define, more generally, that $T(\pi) = \mathbb{E}_{\theta\sim\pi}[h]$ for any $\pi \in \Pi = \{\pi : \mathbb{E}_{\theta\sim\pi}[c] \leq 1\}$.) The main overall difference here is that we are estimating a linear functional of an empirical distribution, and our observations are not identically distributed, and it might be useful to refer to Section 4.2 to see how the cost condition is relevant in applications.

There is some intuition based on i.i.d. case that hints at the result we should expect here. Suppose instead that π was a distribution satisfying the cost condition $\mathbb{E}_{\theta\sim\pi}[c(\theta)] \leq 1$, then by concentration we expect that the constraint $\frac{1}{n}\sum_{i=1}^{n} c(\theta_i) \leq 1$ to be fulfilled approximately (and this can be made absolute by a trunctation argument). On the other hand, we also expect $T(\pi_{\boldsymbol{\theta}})$ to concentrate about its mean $\mathbb{E}_{\theta\sim\pi}[h(\theta)]$, which is an affine functional of π and thus falls under the purview of the previous theorem. Thus, we expect the minimax rate to have the same rate $R^*_{\text{det}}(n) \asymp \delta_{\chi^2}(\frac{1}{\sqrt{n}})^2$.

This is unfortunately wrong, due to the following counterexample given in the paper:

Example 2 (Counterexample to deterministic case)

Again, we consider $\Theta = \mathcal{X} = \{0, 1\}$ with the trivial cost function $c \equiv 0$, which gives $\Pi = \{\mathsf{Bern}(p) : p \in [0, 1]\}$. Let the transition kernel $P : \Theta \to \mathcal{X}$ be the *binary symmetric channel*:

$$P(x|\theta) = \begin{cases} \theta & \text{with probability } \tau \\ 1-\theta & \text{with probability } 1-\tau \end{cases} \tag{28}$$

We set $h(\theta) = \theta$ (so we would like to estimate $\frac{1}{n}\sum_{i=1}^n \theta_i$), so $T(\mathsf{Bern}(p)) = p$. Here, we use the following lower bound (independent of τ):

$$\delta_{\chi^2}(t) \geq \frac{t}{2\sqrt{2}} \sup\{T(\pi) - T(\pi') : \pi, \pi' \in \Pi\} \tag{29}$$

$$= \frac{t}{2\sqrt{2}} \tag{30}$$

On the other hand, the unbiased estimator

$$\widehat{T}(X_1, \ldots, X_n) = \frac{1}{n(1-2\tau)} \sum_{i=1}^n (X_i - \tau) \tag{31}$$

achieves the rate

$$R^*_{\text{det}}(n) \leq \frac{\tau(1-\tau)}{(1-2\tau)^2} \cdot \frac{1}{n} \tag{32}$$

which is not uniform in τ! Setting $\tau = o(1)$ we get $R^*_{\text{det}}(n) \ll \delta_{\chi^2}(\frac{1}{\sqrt{n}})^2$.

Actually, the next theorem implies that the only counterexamples occur when we have $R^*_{\text{det}}(n) = 0$ or $R^*_{\text{det}}(n) \asymp \frac{1}{n}$ (the parametric rate)[2]. We will need extra assumptions for the lower bound to remove these cases.

Theorem 2 (Main Theorem, deterministic case [Polyanskiy and Wu, 2021])

Under the same technical assumptions as the i.i.d. case, we have the same upper bound

$$R^*_{\text{det}}(n) \leq \delta_{\chi^2}(\tfrac{1}{\sqrt{n}}).$$

On the other hand, assuming that

A5. $\text{Var}_{\theta\sim\pi}[h(\theta)]$ is uniformly bounded above by K_V for any $\pi \in \Pi$; and

A6. the cost function c satisfies $c \geq 0$ with equality holding for at least one $\theta_0 \in \Theta$,

then, we will have the lower bound

$$R^*_{\text{det}}(n) \geq \frac{1}{2400} \delta^*_{\chi^2}(\tfrac{1}{\sqrt{n}})^2 - \frac{K_V}{2n}. \tag{33}$$

In particular, if $\delta_{\chi^2}(\frac{1}{\sqrt{n}}) = \omega(\frac{1}{\sqrt{n}})$, the lower bound is matching and the minimax rate is determined asymptotically.

Sketch of proof. The upper bound is very similar to the previous case. For the lower bound, the proof can be more concisely expressed using the mixture vs. mixture version of Le Cam's method:

[2] In the paper (Polyanskiy and Wu [2021]), there is a simple argument to show why the minimax rate is either 0 or $\Omega(1/n)$.

Theorem 3 (Mixture vs. Mixture)

For $\Theta_0, \Theta_1 \subset \Theta$, suppose that the following separation holds for all $a \in \mathcal{A}, \theta_0 \in \Theta_0, \theta_1 \in \Theta_1$:

$$L(\theta_0, a) + L(\theta_1, a) \geq \Delta. \tag{34}$$

Then, for all probability distributions π_0, π_1 over Θ,

$$\inf_T \max_{\theta \in \Theta_0 \cup \Theta_1} \mathbb{E}_\theta[L(\theta, T(X))] \geq \frac{\Delta}{2} \left(1 - \|\mathbb{E}_{\pi_0}[P_{\theta_0}] - \mathbb{E}_{\pi_1}[\pi_{\theta_1}]\|_{\text{TV}} - \pi_0(\Theta_0^c) - \pi_1(\Theta_1^c)\right). \tag{35}$$

Start with distributions $\nu_0, \nu_1 \in \Pi$ such that $\chi^2(\nu_1 P \| \nu_0 P) \geq \frac{1}{n}$ and suppose $\delta = T(\nu_1) - T(\nu_0) > 0$, and fix a free parameter $\gamma \in (0,1)$. The distributions of single samples will be the following weighted combinations

$$\nu_0' = \gamma \nu_0 + (1-\gamma)\delta_{\theta_*}, \quad \nu_1' = \gamma \nu_1 + (1-\gamma)\delta_{\theta_*} \tag{36}$$

where θ_* is a zero-cost parameter. The two priors we will consider are precisely $(\nu_0')^{\otimes n}$ and $(\nu_1')^{\otimes n}$.

The two separated sets will be

$$\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_c : T(\pi_{\boldsymbol{\theta}}) \leq \tfrac{2}{3} T(\nu_0') + \tfrac{1}{3} T(\nu_1')\} \tag{37}$$

$$\boldsymbol{\Theta}_1 = \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_c : T(\pi_{\boldsymbol{\theta}}) \geq \tfrac{1}{3} T(\nu_0') + \tfrac{2}{3} T(\nu_1')\} \tag{38}$$

which immediately gives $T(\pi_{\boldsymbol{\theta}_1}) - T(\pi_{\boldsymbol{\theta}_0}) \geq \frac{1}{3}(T(v_1)' - T(v_0)') = \frac{\gamma\delta}{3}$ for any $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0, \boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1$, so $\Delta = \frac{(\gamma\delta)^2}{18}$.

Now we show that the probability weight of Θ_0, Θ_1 under the respective priors is relatively small due to concentration. For Θ_0, $T(\pi_{\boldsymbol{\theta}})$ has mean $T(\nu_0')$ and variance at most K_V under the first prior, so by Chebyshev (and Markov for the cost condition[3]):

$$(\nu_0')^{\otimes n}(\boldsymbol{\Theta}_0^c) \leq \mathbb{P}_{\theta_i \sim \nu_0'}\left[\frac{1}{n}\sum_{i=1}^n c(\theta_i) > 1\right] + \mathbb{P}_{\theta_i \sim \nu_0'}\left[\frac{1}{n}\sum_{i=1}^n h(\theta_i) > \mathbb{E}h(\theta_i) + \gamma\delta/3\right] \tag{39}$$

$$\leq \gamma + \frac{9K_V}{n\gamma^2\delta^2} \tag{40}$$

Similarly, $(v_1')^{\otimes n}(\boldsymbol{\Theta}_1^c) \leq \gamma + \frac{9K_V}{n\gamma^2\delta^2}$. Finally, we upper bound the TV distance using standard techniques:

$$\left\|(\nu_1')^{\otimes n} P^{\otimes n} - (\nu_0')^{\otimes n} P^{\otimes n}\right\|_{\text{TV}} = \left\|(\nu_1' P)^{\otimes n} - (\nu_0' P)^{\otimes n}\right\|_{\text{TV}} \tag{41}$$

$$\leq \frac{1}{2}\sqrt{\chi^2((\nu_1' P)^{\otimes n} \| (\nu_0' P)^{\otimes n})} \tag{42}$$

$$= \frac{1}{2}\sqrt{(1 + \chi^2(\nu_1' P \| \nu_0' P))^n - 1} \tag{43}$$

$$\leq \frac{1}{2}\sqrt{(1 + \gamma \cdot \chi^2(\nu_1 P \| \nu_0 P))^n - 1} \tag{44}$$

$$\leq \frac{1}{2}\sqrt{(1 + \gamma/n)^n - 1} \leq \frac{1}{2}\sqrt{e^\gamma - 1} \tag{45}$$

$$\tag{46}$$

where we used the convexity of $\chi^2(\bullet \| \bullet)$. Finally, to conclude,

$$R^*_{\text{det}}(n) \geq \frac{(\gamma\delta)^2}{36} \cdot \left(1 - 2\gamma - \frac{18K_V}{n\gamma^2\delta^2} - \frac{\sqrt{e^\gamma - 1}}{2}\right) \tag{47}$$

$$= \frac{\delta^2}{36}\left(\gamma^2 - \gamma^3 - \frac{\gamma^2(e^\gamma - 1)}{2}\right) - \frac{K_V}{2n} \tag{48}$$

and maximizing γ gives the desired result.

[3] This is imposed by our assumption that $\boldsymbol{\theta} \in \boldsymbol{\Theta}_c$ instead of $\Theta^{\otimes n}$

4 Applications

4.1 Binomial Mixtures

We first describe a useful bound for the χ^2-modulus of continuity for a particular setting, that can be applied in the next few examples.

Consider a situation where we would like to do quality control by testing a small number of samples, and then to infer the number of "bad" items among the entire inventory.

Formally, fix a sampling probability $p \in (0,1)$. Let $\mathcal{X} = \Theta = [d]$ (where θ will represent the number of "bad" items), and $\Pi = \mathcal{P}([d])$. Given observations $X \sim \mathsf{Bin}(\theta, p)$, we would like to estimate the probability mass of $\{\theta = 0\}$ under the prior (i.e. $T(\pi) = \pi(0)$).

The next result describes the modulus of continuity in this case, which we will apply in the following sections.

Theorem 4 (δ_{χ^2} for binomial mixtures, Polyanskiy and Wu [2021])
For any $t \geq 0, d \geq 1$, we have

$$\delta_{\chi^2}(t) \leq t^{1 \wedge \frac{p}{1-p}} \tag{49}$$

and in the other direction, we have

$$\delta_{\chi^2}(t) \geq \begin{cases} \frac{t}{2\sqrt{2}} & p \geq \frac{1}{2} \\ C\left(\frac{t}{\ln \frac{1}{t}}\right)^{\frac{p}{1-p}} & t \geq t_0, d \geq C \ln^2 \frac{1}{t} \end{cases} \tag{50}$$

For some t_0, C depending on p.

The difficult parts of the proof depends on two other results: Proposition 9 in Polyanskiy et al. [2020] which converts the LP into analytic properties of certain generating functions and then analyzes them using complex analysis, and Lemma 12 in Polyanskiy et al. [2020], which describes a construction.

4.2 Distinct Elements Problem

Suppose we have an urn with at most n balls, and we would like to figure out the number of different color present among the balls. However, we will only observe each ball with probability p. If $\theta_i \in \Theta = [n]$ denotes the number of balls of color $i \in [n]$, then the parameter set can be captured with the cost function $c(\theta) = \theta$:

$$\boldsymbol{\Theta}_c = \left\{ \boldsymbol{\theta} \in \Theta^{\otimes n} : \frac{1}{n} \sum_{i=1}^{n} \theta_i \leq 1 \right\}. \tag{51}$$

The observations are of the form $X_i \sim \mathsf{Bin}(\theta_i, p)$, and the goal is to estimate the (normalized) number of distinct colors

$$T(\pi_{\boldsymbol{\theta}}) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\theta_i > 1}. \tag{52}$$

Additionally, we also have $K_V \leq 1/4$ and $c(0) = 0$, so the theorem in the deterministic case applies. We can get the following result:

Theorem 5

If $p \geq \frac{1}{2}$, then $R^*(n) \asymp 1/n$. Otherwise, if $p < 1/2$, then

$$\frac{c}{\log^2 n} n^{-p/(1-p)} \leq R^*(n) \leq n^{-p/(1-p)} \tag{53}$$

for some constant $c = c(p) > 0$. Furthermore, the upper bound is attained (to within constant factors) by an estimator of the form

$$\widehat{T} = \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

with $g(0) = 0$ (making it oblivious to the total number of balls).

The case where $p \geq \frac{1}{2}$ follows immediately from applying the previous χ^2-bound to get the rate for $R^*(n)$. Some extra work is required to construct an explicit upper bound for the above form (and then to also enforce $g(0) = 0$), but it follows idea from previous work in Polyanskiy et al. [2020] (similar to the binomial mixture result from earlier).

References

Yury Polyanskiy and Yihong Wu. Dualizing le cam's method for functional estimation, with applications to estimating the unseens, 2021.

Yury Polyanskiy, Ananda Theertha Suresh, and Yihong Wu. Sample complexity of population recovery, 2020.