Lecture 9: Mixture vs mixture, orthogonal polynomials, and moment matching

Lecturer: Yanjun Han

April 26, 2021

Today's plan

Test between two composite hypotheses:

- same two-point lower bound
- a unified view of Hermite and Charlier polynomials
- upper bounds on TV and χ^2 divergence
- duality between moment matching and best polynomial approximation
- Gaussian examples: Gaussian mixture, ℓ_1 norm estimation
- Poisson examples: generalized uniformity testing, entropy estimation

Generalized two-point method

Generalized two-point method

Fix any $\Theta_0 \subseteq \Theta$ and $\Theta_1 \subseteq \Theta$. Suppose that the following separation condition holds:

$$\min_{a \in \mathcal{A}} L(\theta_0, a) + L(\theta_1, a) \geq \Delta > 0, \quad \forall \theta_0 \in \Theta_0, \theta_1 \in \Theta_1.$$

Then for probability distributions π_0 and π_1,

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

$\pi_0'(\cdot) = \frac{\pi_0(\cdot \cap \Theta_0)}{\pi_0(\Theta_0)}$ $\quad \| \mathbb{E}_{\pi_0'}[P_{\theta_0}] - \mathbb{E}_{\pi_1'}[P_{\theta_1}] \|_{\mathrm{TV}}$

$\| \mathbb{E}_{\pi_0}[P_{\theta_0}] - \mathbb{E}_{\pi_0'}[P_{\theta_0}] \|_{\mathrm{TV}} \leq \| \pi_0 - \pi_0' \|_{\mathrm{TV}} = \pi_0(\Theta_0^c)$

$\uparrow$ data-processing

$\pi \longmapsto \mathbb{E}_\pi[P_\theta]$

$\Theta \longrightarrow X$

Orthogonal polynomials

- let $(P_\theta)_{\theta \in [\theta_0-\varepsilon, \theta_0+\varepsilon] \subseteq \mathbb{R}}$ be a 1-D family of distributions
- assume that the following local expansion holds:

$$\frac{dP_{\theta_0+u}}{dP_{\theta_0}}(x) = \sum_{m=0}^{\infty} p_m(x;\theta_0)\frac{u^m}{m!}, \qquad \forall |u| \le \varepsilon.$$

Lemma

Assume that for all $u, v \in [-\varepsilon, \varepsilon]$, the quantity

$$\sum_{x \in \mathcal{X}} \frac{P_{\theta_0+u}(x) P_{\theta_0+v}(x)}{P_{\theta_0}(x)^2} \cdot P_{\theta_0}(x) = \sum_{m,n} \mathbb{E}_{P_{\theta_0}}\left[p_m(X;\theta_0) p_n(X;\theta_0)\right] \cdot \frac{u^m v^n}{m!\, n!}$$

depends only on $(\theta_0, u \cdot v)$, then $\{p_m(x;\theta_0)\}_{m \ge 0}$ is orthogonal under P_{θ_0}.

Gaussian location model: Hermite polynomial

- Gaussian location model: $P_\theta = \mathcal{N}(\theta, 1)$
- premise of the lemma:

$$\int_{\mathbb{R}} \frac{\mathcal{N}(u,1)(x) \cdot \mathcal{N}(v,1)(x)}{\mathcal{N}(0,1)(x)} dx = \exp(uv)$$

$= \mathbb{E}_{X\sim N(0,1)}\left[\exp\left(X\cdot(u+v) - \frac{u^2+v^2}{2}\right)\right] = \exp\left(\frac{(u+v)^2}{2} - \frac{u^2+v^2}{2}\right) = \exp(uv)$

- Hermite polynomial: $H_m(x) = m! \sum_{k=0}^{\lfloor m/2 \rfloor} \frac{(-1)^k}{k!(m-2k)!} \frac{x^{m-2k}}{2^k}$
- $H_0(x) = 1, H_1(x) = x, H_2(x) = x^2 - 1, H_3(x) = x^3 - 3x, \ldots$
- orthogonality property:

$$\mathbb{E}_{X\sim\mathcal{N}(0,1)}[H_m(X)H_n(X)] = n! \cdot 1(m = n)$$

Poisson model: Charlier polynomial

- Poisson model: $P_\theta = \mathrm{Poi}(\theta)$
- premise of the lemma:

$$\sum_{x\in\mathbb{N}} \frac{\mathrm{Poi}(\lambda+u)(x)\cdot \mathrm{Poi}(\lambda+v)(x)}{\mathrm{Poi}(\lambda)(x)} = \exp\left(\frac{uv}{\lambda}\right)$$

$$= \sum_{x\in\mathbb{N}} e^{-\lambda-u-v}\left(\frac{(\lambda+u)(\lambda+v)}{\lambda}\right)^x \Big/ x! = e^{-\lambda-u-v+\frac{(\lambda+u)(\lambda+v)}{\lambda}} = \exp\left(\frac{uv}{\lambda}\right)$$

- Charlier polynomial: $c_m(x;\lambda) = \sum_{k=0}^m (-1)^{m-k}\binom{m}{k}\frac{(x)_k}{\lambda^k}$
- $c_0(x;\lambda) = 1, c_1(x;\lambda) = \frac{x}{\lambda} - 1, c_2(x;\lambda) = \frac{x(x-1)}{\lambda^2} - \frac{2x}{\lambda} + 1, \ldots$
- orthogonality property:

$$\mathbb{E}_{X\sim\mathrm{Poi}(\lambda)}[c_m(X;\lambda)c_n(X;\lambda)] = \frac{n!}{\lambda^n}\cdot 1(m=n)$$

Gaussian model: upper bounds on TV

Theorem

For any $\mu \in \mathbb{R}$ and r.v.s U, V, it holds that

$$\|\mathbb{E}[\mathcal{N}(\mu + U, 1)] - \mathbb{E}[\mathcal{N}(\mu + V, 1)]\|_{\mathrm{TV}} \leq \frac{1}{2}\left(\sum_{m=0}^{\infty} \frac{|\mathbb{E}[U^m] - \mathbb{E}[V^m]|^2}{m!}\right)^{\frac{1}{2}}.$$

$= U * N(\mu, 1)$; Δ_m

$\phi(x)$: pdf of $N(0,1)$

$$LHS = \frac{1}{2}\int \left|\mathbb{E}[\phi(x-U)] - \mathbb{E}[\phi(x-V)]\right| dx$$

$$\phi(x-U) = \phi(x)\cdot\sum_{m=0}^{\infty} H_m(x)\cdot\frac{(-U)^m}{m!}$$

$$= \frac{1}{2}\int \left|\phi(x)\cdot\sum_m H_m(x)\cdot\frac{\mathbb{E}[(-U)^m] - \mathbb{E}[(-V)^m]}{m!}\right| dx$$

$$\overset{(-S}{\leq} \frac{1}{2}\left(\int \left|\sum_m H_m(x)\cdot\frac{\Delta_m}{m!}\right|^2 \phi(x)\,dx\right)^{\frac{1}{2}}\underbrace{\left(\int \phi(x)\,dx\right)^{\frac{1}{2}}}_{=1}$$

$$= \frac{1}{2}\left(\sum_m \frac{\Delta_m^2}{m!}\right)^{1/2}.$$

Gaussian model: upper bounds on χ^2

Theorem

If additionally $\mathbb{E}[V] = 0, \mathbb{E}[V^2] \leq M$, it holds that

$$\chi^2(\mathbb{E}[\mathcal{N}(\mu + U, 1)], \mathbb{E}[\mathcal{N}(\mu + V, 1)]) \leq e^{M^2/2} \cdot \sum_{m=0}^{\infty} \frac{|\mathbb{E}[U^m] - \mathbb{E}[V^m]|^2}{m!}.$$

$$LHS = \int \frac{(\mathbb{E}[\phi(x-U)] - \mathbb{E}[\phi(x-V)])^2}{\mathbb{E}[\phi(x-V)]}\, dx$$

$$\mathbb{E}[\phi(x-V)] = \phi(x) \cdot \mathbb{E}\left[\exp\left(x \cdot V - \frac{V^2}{2}\right)\right] \geq \phi(x) \cdot \exp\left(x \cdot 0 - \frac{M^2}{2}\right)$$

$$\leq e^{\frac{M^2}{2}} \int \frac{\left(\sum_m \phi(x) \cdot H_m(x) \cdot \frac{\Delta_m}{m!}\right)^2}{\phi(x)}\, dx$$

$$= e^{M^2/2} \int \phi(x) \cdot \left(\sum_m H_m(x) \cdot \frac{\Delta_m}{m!}\right)^2 dx$$

$$= RHS$$

Poisson model: upper bounds on TV and χ^2

Theorem

For any $\lambda > 0$ and r.v.s U, V supported on $[-\lambda, \infty)$, it holds that

$$\|\mathbb{E}[\mathrm{Poi}(\lambda + U)] - \mathbb{E}[\mathrm{Poi}(\lambda + V)]\|_{\mathrm{TV}} \le \frac{1}{2} \left(\sum_{m=0}^{\infty} \frac{|\mathbb{E}[U^m] - \mathbb{E}[V^m]|^2}{m! \lambda^m} \right)^{\frac{1}{2}}.$$

In addition, if $\mathbb{E}[V] = 0$ and $|V| \le M$ almost surely, then

$$\chi^2(\mathbb{E}[\mathrm{Poi}(\lambda + U)], \mathbb{E}[\mathrm{Poi}(\lambda + V)]) \le e^M \cdot \sum_{m=0}^{\infty} \frac{|\mathbb{E}[U^m] - \mathbb{E}[V^m]|^2}{m! \lambda^m}.$$

$$\begin{aligned}
\mathbb{E}[\mathrm{Poi}(\lambda+V)^{(x)}] &= \mathbb{E}\left[e^{-\lambda - V} \cdot \frac{(\lambda+V)^x}{x!} \right] \\
&\ge e^{-\lambda - M} \, \mathbb{E}\left[\frac{(\lambda+V)^x}{x!} \right] \\
&\ge e^{-\lambda - M} \frac{\lambda^x}{x!} \\
&= e^{-M} \cdot \mathbb{P}(\mathrm{Poi}(\lambda) = x).
\end{aligned}$$

Example I: Gaussian mixture model

- model: $X_1, \cdots, X_n \sim p\mathcal{N}(\mu_1, \sigma_1^2) + (1-p)\mathcal{N}(\mu_2, \sigma_2^2)$
- estimating unknown parameters: $p \in [0.01, 0.99], \mu_1, \mu_2, \sigma_1, \sigma_2$
- assumption: overall variance of mixture is at most $\sigma^2 = \Omega(1)$

Theorem (Hardt and Price, 2015)

Sample complexity of estimating all parameters within accuracy $O(1)$ is

$$n^\star = \Theta(\sigma^{12}).$$

Proof of lower bound

- target: find two pairs of $(p, \mu_1, \mu_2, \sigma_1, \sigma_2)$ which are $\Omega(1)$-apart, while minimize the χ^2-divergence between these mixtures
- observation: $\mathcal{N}(\mu_1, \sigma_1^2) * \mathcal{N}(0, \sigma^2) = \mathcal{N}(\mu_1, \sigma_1^2 + \sigma^2)$
- suffice to find Gaussian mixtures U, V to minimize

$$\chi^2(U * \mathcal{N}(0, \sigma^2), V * \mathcal{N}(0, \sigma^2))$$

- 5 parameters, so matching first 5 moments of (U, V)

$$U \sim 0.5 \cdot \mathcal{N}(-1, 1) + 0.5 \cdot \mathcal{N}(1, 2)$$
$$V \sim 0.2968 \cdot \mathcal{N}(-1.2257, 0.6100) + 0.7032 \cdot \mathcal{N}(0.5173, 2.3960)$$

- χ^2 divergence:

$$\chi^2(U * \mathcal{N}(0, \sigma^2), V * \mathcal{N}(0, \sigma^2)) \le e^{O(1)/\sigma^2} \cdot \sum_{m=6}^{\infty} \frac{\Delta_m^2}{m! \sigma^{2m}} = O\left(\frac{1}{\sigma^{12}}\right)$$

Example II: ℓ_1 norm estimation

- model: $X \sim \mathcal{N}(\theta, I_p)$ with $\|\theta\|_\infty \leq 1$
- loss function: $L(\theta, T) = |T - \|\theta\|_1|$

Theorem (Cai and Low, 2011)

$$R_p^\star = \Theta\left(p \cdot \frac{\log\log p}{\log p}\right)$$

Failure of point vs. mixture

- a tempting approach: test between $H_0 : \theta = 0$ and $H_1 : \|\theta\|_1 \geq \rho$
- consider any mixture π on H_1, then

$$\chi^2(\mathbb{E}_\pi[P_\theta], P_0) = \mathbb{E}_{\theta,\theta' \sim \pi}[\exp(\theta^\top \theta')] - 1$$

$e^x = 1 + x + \frac{x^2}{2} + \cdots$

$\mathbb{E}[(\theta^\top \theta')^k] \geq 0$

$$\geq \frac{1}{2}\mathbb{E}_{\theta,\theta'}\left[(\theta^\top\theta')^2\right]$$
$$= \frac{1}{2}\mathbb{E}\left[(\theta_1\theta_1' + \cdots + \theta_p\theta_p')^2\right]$$
$$= \frac{1}{2}\sum_i \left(\mathbb{E}[\theta_i^2]\right)^2 + \underbrace{\geq 0}$$
$$\geq \frac{1}{2p}\left(\sum_i \mathbb{E}[\theta_i^2]\right)^2$$
$$\geq \frac{1}{2p}\left(\frac{1}{p}\cdot\rho^2\right)^2 = \frac{\rho^4}{2p^3}$$

- best choice of ρ: $\rho = O(p^{3/4})$, also the right radius for the testing problem

$T = \|X\|^2 - p$

Idea: mixture vs. mixture

Idea: test between $H_0 : \|\theta\|_1 \le \rho_0$ and $H_1 : \|\theta\|_1 \ge \rho_1$

- properly choose probability measures μ_0, μ_1 on $[-1, 1]$
- try $\pi_0 = \mu_0^{\otimes p}, \pi_1 = \mu_1^{\otimes p}$

Targets:

- indistinguishability: $\chi^2(\mu_0 * \mathcal{N}(0,1), \mu_1 * \mathcal{N}(0,1)) = O(1/p)$
- support: $\pi_0(H_0^c) + \pi_1(H_1^c) = o(1)$
- separation: $\rho_1 - \rho_0$ as large as possible

Target I: indistinguishability

- idea: match the first K moments of μ_0 and μ_1
- χ^2-divergence:

$$\chi^2(\mu_0 * \mathcal{N}(0,1), \mu_1 * \mathcal{N}(0,1)) \leq e^2 \cdot \sum_{m=K+1}^{\infty} \frac{2^{m+1}}{m!} = O\left(\left(\frac{2e}{K}\right)^K\right)$$

$= \frac{1}{p}$

- choice of K: $K \asymp \log p / \log\log p$

Target II: support

- target: $\|\theta\|_1 \le \rho_0$ w.h.p. under $\theta \sim \pi_0 = \mu_0^{\otimes p}$, and $\|\theta\|_1 \ge \rho_1$ w.h.p. under $\theta \sim \pi_1 = \mu_1^{\otimes p}$
- idea: under π_i, $\|\theta\|_1$ concentrates around $p \cdot \mathbb{E}_{\theta\sim\mu_i}[|\theta|]$ with fluctuation $O(\sqrt{p})$
- choice of ρ_0 and ρ_1:

$$\rho_0 = p \cdot \mathbb{E}_{\theta\sim\mu_0}[|\theta|] + C\sqrt{p},$$
$$\rho_1 = p \cdot \mathbb{E}_{\theta\sim\mu_1}[|\theta|] - C\sqrt{p}.$$

Target III: separation

- maximize $\mathbb{E}_{\theta\sim\mu_1}[|\theta|] - \mathbb{E}_{\theta\sim\mu_0}[|\theta|]$ subject to μ_0 and μ_1 supported on $[-1, 1]$ and have matching first K moments
- key result: duality between moment matching and best polynomial approximation

Theorem

For bounded interval $I \subseteq \mathbb{R}$ and continuous function f on I, let $S^\star$ be the objective value of

$$\begin{aligned} \text{maximize} \quad & \mathbb{E}_{\theta\sim\mu_1}[f(\theta)] - \mathbb{E}_{\theta\sim\mu_0}[f(\theta)] \\ \text{subject to} \quad & \mu_0, \mu_1 \text{ supported on } I, \text{ with matching } K \text{ first moments} \end{aligned}$$

We have

For $f(x) = |x|$, $E_k = \frac{1}{k}$

$$S^\star = 2 \cdot \inf_{a_0, a_1, \cdots, a_K} \sup_{x \in I} \left| f(x) - \sum_{k=0}^{K} a_k x^k \right|$$

Proof of duality result

- First proof: minimax theorem

$$\inf_{a_0,\cdots,a_K} \sup_x \left| f(x) - \sum_k a_k x^k \right|$$

$$= \inf_{a_0,\cdots,a_K} \sup_{\mu : \|\mu\|_{TV} \le 1} \int_I \left(f(x) - \sum_k a_k x^k \right) \mu(dx)$$

$$= \sup_{\mu} \inf_{a_0,\cdots,a_K} \cdots$$

finite only if $\int x^k \mu(dx) = 0$ for $k = 0, 1, \cdots, K$

$$= \int f(x)\, \mu(dx)$$

$\mu_+(I) \le \frac{1}{2}$, $\mu_-(I) \le \frac{1}{2}$ $\hookrightarrow$ $= \sup_{\mu_+, \mu_-} \int f(x) \mu_+(dx) - \int f(x) \mu_-(dx)$ $\qquad \mu = \mu_+ - \mu_-$

- Second proof: Chebyshev alternation theorem

Technical detail I: Poissonization

- multinomial model: $X_1, \cdots, X_n \sim P = (p_1, \cdots, p_k)$
- Poissonized model: $h_j \sim \text{Poi}(np_j), j \in [k]$
- HW1: Le Cam's distance does not vanish for large k
- fix: require similar order instead of a small difference

Lemma

Let R_n and $R_n^\star$ be the respective minimax risks under multinomial and Poissonized model, respectively, under the same loss. Then

$$\frac{1}{2} R_{2n} \le R_n^\star \le R_{n/2} + R_0 e^{-n/8}$$

$R_n^*(\pi) = \sum_m R_m(\pi) \cdot P(\text{Poi}(n) = m)$

$n \mapsto R_n(\pi)$ non-increasing

Technical detail II: approximate distribution

- motivation: $\sum_{j=1}^{k} p_j = 1$ typically violated under product prior
- Poissonized model with approximate distribution: for a non-negative vector $(p_1, \cdots, p_k)$, observe $h_j \sim \mathrm{Poi}(np_j)$
- claim: suffice to consider the above model for lower bounds
- reduction idea: in the Poissonized model, directly use the estimator under the multinomial model with sample size $\sum_{j=1}^{k} h_j$

$$(h_1, \cdots, h_k) \mid h_1 + \cdots + h_k = n \sim \text{multinomial}(n; P/\|P\|_1)$$

- $\|P\|_1 \approx 1$ w.h.p. thanks to concentration
- details vary from example to example

Example III: generalized uniformity testing

- model: $X_1, \cdots, X_n \sim P = (p_1, \cdots, p_k)$
- target: test between $H_0 : P = U_S$ for some $S \subseteq [k]$ and $H_1 : P$ is ε-far from any uniform distribution under TV

Theorem (Batu and Canonne, 2017; Diakonikolas, Kane, and Stewart, 2018)

The optimal sample complexity of generalized uniformity testing is

$$n^\star = \Theta\left(\frac{\sqrt{k}}{\varepsilon^2} + \frac{k^{2/3}}{\varepsilon^{4/3}}\right)$$

Proof of lower bound

- idea: assign to $(p_1, \cdots, p_k)$ product priors with

$$H_0: \quad U = \begin{cases} 0 & \text{w.p. } \varepsilon^2/(1+\varepsilon^2) \\ (1+\varepsilon^2)/k & \text{w.p. } 1/(1+\varepsilon^2) \end{cases},$$

$$H_1: \quad V = \begin{cases} (1-\varepsilon)/k & \text{w.p. } 1/2 \\ (1+\varepsilon)/k & \text{w.p. } 1/2 \end{cases}.$$

- property: $\mathbb{E}[U] = \mathbb{E}[V] = 1/k$, matching first two moments, and

$$|\mathbb{E}[(U-1/k)^m] - \mathbb{E}[(V-1/k)^m]| \le \frac{2\varepsilon^2}{k^m}, \quad m \ge 3.$$

- χ^2-divergence:

$$\chi^2(\mathbb{E}[\mathrm{Poi}(nU)], \mathbb{E}[\mathrm{Poi}(nV)]) \le e^{n\varepsilon/k} \sum_{m=3}^{\infty} \frac{4\varepsilon^4 (n/k)^{2m}}{m!(n/k)^m} = O\left(\frac{n^3\varepsilon^4}{k^3}\right)$$

Why only two moments?

- uniformity testing: only match the first moment
- generalized uniformity testing: match the first and second moments
- can we match more?

Lemma

Let μ be a probability measure supported on k elements of $[0, \infty)$, one of which is zero. If ν is another probability measure matching the first $2k-1$ moments of μ, then $\nu = \mu$.

support of μ: $0, x_1, \cdots, x_{k-1}$

$Q(x) = x(x-x_1)^2(x-x_2)^2 \cdots (x-x_{k-1})^2$ $\quad \deg(Q) = 2k-1$

$0 = \mathbb{E}_{x\sim\mu}[Q(x)] = \mathbb{E}_{x\sim\nu}[Q(x)] \geq 0$

Example IV: entropy estimation

- model: $X_1, \cdots, X_n \sim P$
- target: estimate the Shannon entropy $H(P) = \sum_{i=1}^{k} -p_i \log p_i$ with loss $L(P, T) = |T - H(P)|^2$

Theorem (Jiao, Venkat, Han, and Weissman, 2015; Wu and Yang, 2016)

$$R^{\star}_{n,k} = \Theta\left(\left(\frac{k}{n \log n}\right)^2 + \frac{(\log k)^2}{n}\right), \quad n = \Omega\left(\frac{k}{\log k}\right).$$

Roadmap of proof

Apply product distribution $\mu_0^{\otimes k}$ and $\mu_1^{\otimes k}$ to $(p_1, \cdots, p_k)$ such that:

- indistinguishability: μ_0 and μ_1 have matching first K moments
- separation: $\Delta = \mathbb{E}_{p\sim\mu_1}[-p\log p] - \mathbb{E}_{p\sim\mu_0}[-p\log p]$ is large
- support: the random fluctuation of $H(P)$ under each product distribution has magnitude smaller than Δ
- mean value: $\mathbb{E}_{p\sim\mu_1}[p] = O((n\log n)^{-1})$

Separation and indistinguishability

- suppose that μ_0, μ_1 supported on $[0, M]$, with matching K moments
- indistinguishability:

$$\|\mathbb{E}_{p\sim\mu_0}[\mathrm{Poi}(np)] - \mathbb{E}_{p\sim\mu_1}[\mathrm{Poi}(np)]\|_{\mathrm{TV}}^2 \le \sum_{m=K+1}^{\infty} \frac{(nM)^{2m}}{m!(nM)^m} = \frac{1}{\mathrm{poly}(n)}$$

  if $K = \Omega(nM + \log n)$
- separation: the best polynomial approximation error of $-x \log x$ on $x \in [0, M]$ is $\Delta_0 \asymp M/K^2$
- optimal choices of parameters: $M \asymp \log n/n, K \asymp \log n$, so that $\Delta_0 \asymp 1/(n \log n)$

Mean constraint: change-of-measure trick

- problem: previous construction $p \in [0, M]$ with $M \asymp \log n/n$ may not fulfill the mean constraint $\mathbb{E}_{p\sim\mu_0}[p] = O((n \log n)^{-1})$
- idea: construct measures ν_0, ν_1 supported on $[1/(n \log n), M]$ with first matching K moments, and

$$\mathbb{E}_{p\sim\nu_1}[-\log p] - \mathbb{E}_{p\sim\nu_0}[-\log p] = \Omega(1)$$

- change-of-measure trick:

$$\mu_i(dx) = \left(1 - \mathbb{E}_{X\sim\nu_i}\left[\frac{1}{nX \log n}\right]\right)\delta_0(dx) + \frac{\nu_i(dx)}{nx \log n}$$

- property: matching moments up to order $K + 1$, and

$$\mathbb{E}_{p\sim\mu_0}[p] = \frac{1}{n \log n}$$

References

- Moritz Hardt and Eric Price. "Tight bounds for learning a mixture of two Gaussians." *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. ACM, 2015.
- Yihong Wu and Pengkun Yang. "Optimal estimation of Gaussian mixtures via denoised method of moments." *Annals of Statistics* 48.4 (2020): 1981–2007.
- Yihong Wu and Pengkun Yang. "Minimax rates of entropy estimation on large alphabets via best polynomial approximation." *IEEE Transactions on Information Theory* 62.6 (2016): 3702–3720.
- Yanjun Han, Jiantao Jiao, Tsachy Weissman, and Yihong Wu. "Optimal rates of entropy estimation over lipschitz balls." *Annals of Statistics*, 48(6): 3228–3250, 2020.

Next lecture: testing multiple hypotheses, Fano and Assouad