Sharp Inequalities between Total Variation and Hellinger Distances for Gaussian Mixtures

Joonhyuk Jung [1] Chao Gao [1]

[1]University of Chicago

ICML International Conference On Machine Learning

Introduction

Given a probability measure π supported on $\mathbb{R}^d$, we define the Gaussian mixture density by

$$f_\pi(x) := \int_{\mathbb{R}^d} \phi_d(x-\theta)\, d\pi(\theta),$$

where ϕ_d is the d-dimensional standard Gaussian density function. We study the relation between the total variation distance $\mathrm{TV}(p,q) := \frac{1}{2}\int |p-q|$ and the Hellinger distance $H(p,q) := \sqrt{\frac{1}{2}\int(\sqrt{p}-\sqrt{q})^2}$ of two Gaussian mixture densities.

For mixing distributions π and η supported on a bounded Euclidean ball $\{\theta \in \mathbb{R}^d : \|\theta\|_2 \le M\}$, it was proved by Jia et al. [4] that $H^2(f_\pi, f_\eta) \asymp \mathrm{KL}(f_\pi \| f_\eta)$ holds up to constant factors depending on M and d. However, whether

$$\mathrm{TV}(f_\pi, f_\eta) \asymp H(f_\pi, f_\eta)$$

holds was explicitly listed as an open question in the paper. We disprove the linear comparability by showing

$$H(f_\pi, f_\eta) \le \mathrm{TV}^{1-o(1)}(f_\pi, f_\eta). \tag{1}$$

In Theorem 1, we provide the exact form of the $o(1)$ exponent. In Theorem 2, in addition, we show that the $o(1)$ term is indeed necessary.

Main Result and Sharpness

Theorem 1 (Inequality between TV distance and χ^2-divergence)

Let π and η be probability measures supported on the d-dimensional cube $[-M,M]^d$. Let $\delta > 0$. Then, there exists $C_0 = C_0(\delta, M, d) > 0$, not depending on π or η, such that

$$\sqrt{\chi^2(f_\pi \| f_\eta)} \le \left(C_0 \vee \mathrm{TV}^{-\alpha(\mathrm{TV}(f_\pi,f_\eta))}(f_\pi, f_\eta)\right) \mathrm{TV}(f_\pi, f_\eta),$$

where we define

$$\alpha(t) := \frac{2+\delta}{\log\left(\log(1/t) \vee e\right)}, \qquad t > 0.$$

Theorem 2 (Sharpness)

There exist two sequences of probability measures $\{\pi_n\}$ and $\{\eta_n\}$ supported on $[-M,M]$ such that, if we define

$$\mathrm{TV}_n := \mathrm{TV}(f_{\pi_n}, f_{\eta_n}), \qquad H_n := H(f_{\pi_n}, f_{\eta_n}),$$

then $\mathrm{TV}_n \downarrow 0$ as $n \to \infty$, and moreover it holds for all n that

$$H_n \ge \mathrm{TV}_n^{1-\alpha^*(\mathrm{TV}_n)}, \tag{2}$$

where we define

$$\alpha^*(t) := \frac{0.33}{\log\log(1/t)}, \qquad t > 0.$$

Remarks

- The Theorem 1 immediately implies the upper bound (1) on the Hellinger distance as $H^2(p,q) \le \chi^2(p\|q)$ holds in general.
- We can also construct d-dimensional probability measures π_n and η_n satisfying (2) because we have $\mathrm{TV}(f_\pi, f_\eta) = \mathrm{TV}(f_{\pi^\star}, f_{\eta^\star})$ and $H(f_\pi, f_\eta) = H(f_{\pi^\star}, f_{\eta^\star})$ for

$$\pi = \pi^\star \otimes \delta_0 \otimes \cdots \otimes \delta_0, \qquad \eta = \eta^\star \otimes \delta_0 \otimes \cdots \otimes \delta_0,$$

where δ_0 denotes the point mass at zero and $\otimes$ the product measure.

Application 1: Learning Gaussian Mixtures in TV

In this section, we consider the problem of estimating a Gaussian mixture $P \in \mathcal{P}_{M,d}$ based on i.i.d. samples drawn from P, where $\mathcal{P}_{M,d}$ denotes the collection of d-dimensional Gaussian mixtures where the mixing distributions are supported on the d-dimensional cube $[-M,M]^d$.

Theorem 3 (Learning Gaussian mixtures in TV distance)

Suppose $\mathcal{P}$ is a compact subset of $\mathcal{P}_{M,d}$. Then, we have

$$\epsilon_n^{2\left(1+\frac{\Theta(1)}{\log(\log(1/\epsilon_n)\vee e)}\right)} \lesssim \inf_{\widehat{P}} \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[\mathrm{TV}^2\left(P, \widehat{P}\right)\right] \lesssim \epsilon_n^2,$$

where

$$\epsilon_n^2 \asymp \inf_{\epsilon>0}\left(\epsilon^2 + \frac{1}{n}\log N_{H,loc}(\mathcal{P}, \epsilon)\right),$$

and $N_{H,loc}(\mathcal{P}, \epsilon)$ is the local Hellinger covering number of $\mathcal{P}$.

Application 2: Robust Density Estimation in Hellinger

In this section, we consider the problem of estimating a Gaussian mixture with contaminated data,

$$X_1, \dots, X_n \overset{i.i.d.}{\sim} P := (1-\epsilon)P_{f_\pi} + \epsilon Q, \tag{3}$$

where the distribution $P_{f_\pi} \in \mathcal{P}_{M,d}$ has density function f_π and Q is an arbitrary distribution of contamination. The data generating process in (3) is recognized as Huber's contamination model [3].

Theorem 4 (Robust density estimation in Hellinger distance)

Consider the data generating process in (3). Then, we have

$$\inf_{\widehat{f}} \sup_{\pi,Q} \mathbb{E}\left[H^2\left(f_\pi, \widehat{f}\right)\right] \asymp \epsilon^{2\left(1-\frac{\Theta(1)}{\log(\log(1/\epsilon)\vee e)}\right)}, \tag{4}$$

provided that $n \ge \mathrm{poly}(1/\epsilon)$, where the expectation is under (3) and the supremum is taken over all Q and π such that $\mathrm{supp}(\pi) \subseteq [-M,M]^d$.

Theorem 5 (Robust regret bound)

Consider the data generating process in (3). Suppose $\widehat{\theta}^\star(\cdot)$ is the oracle Bayes estimator given by Tweedie's formula [2]. Then,

$$\inf_{\widehat{\theta}} \sup_{\pi,Q} \mathbb{E}\left[\mathbb{E}_{X\sim f_\pi}\left\|\widehat{\theta}(X) - \widehat{\theta}^\star(X)\right\|^2\right] \lesssim \epsilon^{2\left(1-\frac{\Theta(1)}{\log(\log(1/\epsilon)\vee e)}\right)} + \frac{1}{n^{1-o(1)}},$$

where the outer expectation is under (3) and the supremum is taken over all Q and π such that $\mathrm{supp}(\pi) \subseteq [-M,M]^d$.

Remarks

- Applying the same argument as in Chen et al. [1], the Theorem 2 (Sharpness) immediately implies the lower bound in (4).
- We also prove that Yatacos' estimator [5] attains the upper bound in (4).
- Dependency of the minimax rate (4) on n is a long-standing open question in the clean data setting ($\epsilon = 0$), and it remains open in the "small ϵ" regime.

References

[1] Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.

[2] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106 (496):1602–1614, 2011.

[3] Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[4] Zeyu Jia, Yury Polyanskiy, and Yihong Wu. Entropic characterization of optimal rates for learning Gaussian mixtures. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4296–4335. PMLR, 2023.

[5] Yannis G Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13(2):768–774, 1985.

https://joonhyuk.com joonhyukjung@uchicago.edu