A New Lower Bound for Kullback-Leibler Divergence Based on Hammersley-Chapman-Robbins Bound

Tomohiro Nishiyama
Email: htam0ybboh@gmail.com

Abstract

In this paper, we derive a useful lower bound for the Kullback-Leibler divergence (KL-divergence) based on the Hammersley-Chapman-Robbins bound (HCRB). The HCRB states that the variance of an estimator is bounded from below by the Chi-square divergence and the expectation value of the estimator. By using the relation between the KL-divergence and the Chi-square divergence, we show that the lower bound for the KL-divergence which only depends on the expectation value and the variance of a function we choose. This lower bound can also be derived from an information geometric approach. Furthermore, we show that the equality holds for the Bernoulli distributions and show that the inequality converges to the Cramér-Rao bound when two distributions are very close. We also describe application examples and examples of numerical calculation.

Keywords: Hammersley-Chapman-Robbins bound, Cramér-Rao bound, Kullback-Leibler divergence, Chi-square divergence, alpha-divergence, information geometry.

I. Introduction

The divergences are quantities that measure discrepancy between probability distributions and play key roles in statistics, information theory, signal processing and machine learning. For two probability distributions P and Q, divergences satisfy the following properties.

$D(P\|Q) \geq 0$ with equality if and only if $P = Q$.

The Kullback-Leibler divergence (KL-dievergence) [9], the Hellinger distance, and the χ^2-divergence are well-known divergences which are a type of the α-divergence [4]. In particular, the KL-divergence is often used and has good properties that it is invariant under parameter transformations and is compatible with maximum likelihood estimation.

On the other hand, the Hammersley-Chapman-Robbins bound (HCRB) [10], [3], [7] states that the variance of an estimator is bounded from below by the χ^2-divergence and the expected value of the estimator. From another point of view, the χ^2-divergence is bounded by the expected value and the variance of a function of interest (FoI). Recently, the similar bound for the Hellinger distance was derived in [8], [6]

In this paper, we derive a new lower bound for the KL-divergence using the expected value and variance of the FoI. We show that the equality holds for the Bernoulli distributions and show that we obtain the Cramér-Rao bound [5], [13] when two distributions are very close. Furthermore, we give an another proof by information geometric approach [1], [2]. One of the important applications of our inequality is to estimate the lower bound of the KL-divergence between two given data. There are several methods to estimate divergence between given data [11], [14], our lower bound can be a criterion of validity of estimation results and convergence. Finally, we describe some examples of numerical calculation. [1]

II. New bound for the KL-divergence

A. Main results

Let us consider two probability distributions (measures) P and Q with the same support $\omega \subseteq \mathbb{R}^n$. In the following, the integrals is always defined on ω.

---

[1] https://github.com/nissy220/KL_divergence

The α-divergence between P and Q is defined as

$$D_\alpha(P\|Q) := \begin{cases} \frac{1}{\alpha(\alpha-1)}\left(\int p(x)^\alpha q(x)^{1-\alpha}d\lambda(x) - 1\right) & (\alpha \neq 0, 1) \\ \int q(x)\log\frac{q(x)}{p(x)}d\lambda(x) & (\alpha = 0) \\ \int p(x)\log\frac{p(x)}{q(x)}d\lambda(x) & (\alpha = 1), \end{cases}$$

where p and q denote the Radon-Nikodym derivatives of probability measures P and Q respectively and λ denotes the Lebesgue measure or the counting measure.

The KL-divergence, the squared Hellinger distance and the χ^2-divergence are the cases for $\alpha = 1, \frac{1}{2}, 2$.

$$\mathrm{KL}(P\|Q) := \int p(x)\log\frac{p(x)}{q(x)}d\lambda(x) = D_1(P\|Q) \tag{1}$$

$$\mathrm{Hel}^2(P, Q) := \int(\sqrt{q(x)} - \sqrt{p(x)})^2 d\lambda(x) = \frac{1}{2}D_{\frac{1}{2}}(P\|Q) \tag{2}$$

$$\chi^2(P\|Q) := \int\frac{(q(x) - p(x))^2}{q(x)}d\lambda(x) = 2D_2(P\|Q). \tag{3}$$

Next, we consider a function of interest (FoI) $f : \omega \to \mathbb{R}$ and consider an expected value and a variance of the FoI for P. These quantities are defined as $E_P[f] := \int p(x)f(x)d\lambda(x)$ and $\mathrm{Var}_P(f) := E_P[f^2] - E_P[f]^2$.

The χ^2-divergence and the squared Hellinger distance satisfy the following inequalities.

$$\chi^2(P\|Q) \geq \frac{(E_Q - E_P)^2}{V_Q} \tag{4}$$

$$\mathrm{Hel}^2(P, Q) \geq \frac{(E_Q - E_P)^2}{2\left(V_P + V_Q + \frac{1}{2}(E_Q - E_P)^2\right)}, \tag{5}$$

where E_P, E_Q and V_P, V_Q denote $E_P[f], E_Q[f]$ and $\mathrm{Var}_P(f), \mathrm{Var}_Q(f)$, respectively. The inequality (4) is the HCRB for general probability densities.

The main purpose of this section is to derive a similar lower bound which only depends on E_P, E_Q, V_P, V_Q for the KL-divergence. We first show the result.

Theorem 1 (New lower bound for the KL-divergence)**.** *Let P and Q are probability distributions with the same support ω.*

Let $f : \omega \to \mathbb{R}$ be a FoI with finite E_P, E_Q, V_P, V_Q and $V_P, V_Q > 0$, where E_P, E_Q and V_P, V_Q denote $E_P[f], E_Q[f]$ and $\mathrm{Var}_P(f), \mathrm{Var}_Q(f)$, respectively.

Then,

$$\mathrm{KL}(P\|Q) \geq \frac{(A - 2V_P)}{D}\tanh^{-1}\frac{D}{A} + \frac{1}{2}\log\frac{V_P}{V_Q}, \tag{6}$$

where $A := (E_Q - E_P)^2 + V_P + V_Q$, $D := \sqrt{A^2 - 4V_PV_Q}$. The equality holds if and only if there exists a function $C(t)$ which satisfies

$$\frac{q(x) - p(x)}{p(x) + t(q(x) - p(x))} = \big(f(x) - E_P - t(E_Q - E_P)\big)C(t) \tag{7}$$

for all $x \in \omega$ and $t \in [0, 1]$, where p and q denote the Radon-Nikodym derivatives of P and Q, respectively.

The basic idea to show this inequality is to use a relation between D_α and $D_{\alpha+1}$. Considering a mixture distribution $r(x; t) = p(x) + t(q(x) - p(x))$ for $t \in [0, 1]$, the derivative of $D_\alpha(P\|R(t))$ with respect to t can be represented using $D_\alpha(P\|Q)$ and $D_{\alpha+1}(P\|Q)$ as shown later. Since the KL-divergence and the χ^2-divergence are the case for $\alpha = 1$ and $\alpha = 2$ respectively, by applying the HCRB and integrating from 0 to 1 with respect to t, we can derive the inequality. We show details below.

Lemma 1. *Let P and Q are probability distributions with the same support ω.*

Let $r(x; t)$ be the Radon-Nikodym derivative of a probability distribution $R(t)$ and $r(x; t) := p(x) + t(q(x) - p(x))$ for $t \in [0, 1]$.

Then,

$$\frac{d}{dt}D_\alpha(P\|R(t)) = \frac{(1-\alpha)}{t}D_\alpha(P\|R(t)) + \frac{(1+\alpha)}{t}D_{\alpha+1}(P\|R(t)). \tag{8}$$

Proof. For $\alpha \neq 0, 1$,

$$\frac{d}{dt}D_\alpha(P\|R(t)) = -\frac{1}{\alpha}\int (q(x)-p(x))p(x)^\alpha r(x;t)^{-\alpha}d\lambda(x) = -\frac{1}{\alpha}\int \frac{(r(x;t)-p(x))}{t}p(x)^\alpha r(x;t)^{-\alpha}d\lambda(x) \tag{9}$$

$$= \frac{(1-\alpha)}{t}D_\alpha(P\|R(t)) + \frac{(1+\alpha)}{t}D_{\alpha+1}(P\|R(t)). \tag{10}$$

For $\alpha = 1$,

$$\frac{d}{dt}D_1(P\|R(t)) = -\int \frac{p(x)(q(x)-p(x))}{r(x;t)}d\lambda(x) \tag{11}$$

$$= \frac{1}{t}\int \frac{p(x)(p(x)-r(x;t))}{r(x;t)}d\lambda(x) = \frac{1}{t}\int \frac{(r(x;t)-p(x))^2}{r(x;t)}d\lambda = \frac{2}{t}D_2(P\|R(t)). \tag{12}$$

We can show the inequality for $\alpha = 0$ in the same way. Hence, the result follows. $\square$

Lemma 2 (HCRB). *Let P and Q are probability distributions with the same support ω.*
Let $f : \omega \to \mathbb{R}$ be a FoI with finite E_P, E_Q, V_P, V_Q and $V_Q > 0$.
Then,

$$\chi^2(P\|Q) \geq \frac{(E_Q - E_P)^2}{V_Q}, \tag{13}$$

with equality if and only if there exists a constant C which satisfies $\frac{q(x)-p(x)}{q(x)} = C(f(x) - E_Q)$ for all $x \in \omega$.

Proof. Consider the following quantity.

$$E_Q - E_P = \int (f(x) - E_Q)(q(x)-p(x))d\lambda(x) = \int \sqrt{q(x)}(f(x)-E_Q)\frac{(q(x)-p(x))}{\sqrt{q(x)}}d\lambda(x). \tag{14}$$

Applying the Cauchy-Schwartz inequality yields

$$(E_Q - E_P)^2 \leq V_Q \int \frac{(q(x)-p(x))^2}{q(x)}d\lambda(x) = V_Q\chi^2(P\|Q). \tag{15}$$

Since $V_Q > 0$, we obtain the desired inequality. Since the FoI isn't a constant from the assumption $V_Q > 0$, the equality holds if and only if there exists a constant C which satisfies $\frac{q(x)-p(x)}{\sqrt{q(x)}} = C\sqrt{q(x)}(f(x) - E_Q)$ for all $x \in \omega$. Hence, the result follows. $\square$

Proof of Theorem 1. Applying Lemma 1 for $\alpha = 1$ yields

$$\frac{d}{dt}\mathrm{KL}(P\|R(t)) = \frac{\chi^2(P\|R(t))}{t}. \tag{16}$$

Applying Lemma 2, we have

$$\frac{d}{dt}\mathrm{KL}(P\|R(t)) \geq \frac{(E_{R(t)} - E_P)^2}{t\mathrm{V}_{R(t)}}. \tag{17}$$

Since $E_{R(t)}[f] = tE_Q[f] + (1-t)E_P[f]$ and $E_{R(t)}(f^2) = tE_Q[f^2] + (1-t)E_P[f^2]$, we have $\mathrm{Var}_{R(t)}(f) = E_{R(t)}(f^2) - E_{R(t)}[f]^2 = tV_Q + (1-t)V_P + t(1-t)(E_Q - E_P)^2$. Substituting this equality into (17), we have

$$\frac{d}{dt}\mathrm{KL}(P\|R(t)) \geq \frac{t(E_Q - E_P)^2}{t(1-t)(E_Q - E_P)^2 + (1-t)V_P + tV_Q}. \tag{18}$$

Integrating from 0 to 1 with respect to t, we have

$$\mathrm{KL}(P\|Q) \geq \int_0^1 \frac{t(E_Q - E_P)^2}{t(1-t)(E_Q - E_P)^2 + (1-t)V_P + tV_Q}dt. \tag{19}$$

Using the following formula

$$\int_0^x \frac{at}{-at^2 + (a+b)t + c} dt \qquad (20)$$

$$= \frac{a+b}{\sqrt{a^2 + 2a(b+2c) + b^2}} \tanh^{-1}\left(\frac{a(2x-1) - b}{\sqrt{a^2 + 2a(b+2c) + b^2}}\right) - \frac{1}{2}\log(x(-ax + a + b) + c) + \text{const.}$$

and putting $a = (E_Q - E_P)^2$, $b = V_Q - V_P$, $c = V_P$ and using the definition $A = (E_Q - E_P)^2 + V_P + V_Q$ and $D = \sqrt{A^2 - 4V_P V_Q}$, we have

$$\int_0^1 \frac{t(E_Q - E_P)^2}{t(1-t)(E_Q - E_P)^2 + (1-t)V_P + tV_Q} dt \qquad (21)$$

$$= \frac{(A - 2V_P)}{D}\left(\tanh^{-1}\left(\frac{a-b}{D}\right) + \tanh^{-1}\left(\frac{a+b}{D}\right)\right) - \frac{1}{2}\log\frac{V_Q}{V_P}.$$

By applying the addition formula of the inverse hyperbolic function $\tanh^{-1} u + \tanh^{-1} v = \tanh^{-1}\left(\frac{u+v}{1+uv}\right)$ and $D^2 + a^2 - b^2 = 2aA$, we have

$$\int_0^1 \frac{t(E_Q - E_P)^2}{t(1-t)(E_Q - E_P)^2 + (1-t)V_P + tV_Q} dt = \frac{(A - 2V_P)}{D}\tanh^{-1}\left(\frac{D}{A}\right) + \frac{1}{2}\log\frac{V_P}{V_Q}. \qquad (22)$$

By combining with (19), we obtain the desired inequality. From Lemma 2, the equality holds if there exists a function $\tilde{C}(t)$ which satisfies

$$\frac{r(x;t) - p(x)}{r(t;x)} = \tilde{C}(t)(f(x) - E_{R(t)}) \qquad (23)$$

for all $x \in \omega$ and $t \in [0, 1]$. By substituting $r(x;t) = p(x) + t(q(x) - p(x))$ and $E_{R(t)} = E_P + t(E_Q - E_P)$ into this equality and putting $C(t) = \frac{\tilde{C}(t)}{t}$, we obtain the equality condition. $\square$

Proposition 1. *When $\omega = \{0, 1\}$ and the FoI is x, the Bernoulli distributions satisfy the equality condition of Theorem 1.*

Proof. Let $P(x)$ and $Q(x)$ be probability mass functions of the Bernoulli distributions and $P(0) = 1 - p, P(1) = p$ and $Q(0) = 1 - q, Q(1) = q$. We choose a function $C(t)$ which satisfies

$$\frac{Q(0) - P(0)}{P(0) + t(Q(0) - P(0))} = -(E_P + t(E_Q - E_P))C(t). \qquad (24)$$

This is the equality condition of Theorem 1 for $x = 0$. Since $E_P + t(E_Q - E_P) = P(1) + t(Q(1) - P(1))$, $P(0) + t(Q(0) - P(0)) = 1 - (E_P + t(E_Q - E_P))$ and $Q(0) - P(0) = -(Q(1) - P(1))$, we have

$$\frac{Q(1) - P(1)}{P(1) + t(Q(1) - P(1))} = (1 - (E_P + t(E_Q - E_P))C(t). \qquad (25)$$

This is the equality condition for $x = 1$. Hence, the result follows. $\square$

Proposition 2. *For a parameter $\theta \in \mathbb{R}$, let $p(x; \theta)$ and $p(x; \theta + \delta\theta)$ be probability density functions of probability distributions P and Q, respectively. When $\delta\theta \to 0$, the inequality of Theorem 1 converges to the Cramér-Rao bound.*

The Cramér-Rao bound is

$$\text{Var}_P(f) \geq \frac{\psi'(\theta)^2}{I(p)}, \qquad (26)$$

where $I(p) := \int \frac{p'(x,\theta)^2}{p(x,\theta)} d\lambda(x)$ is the Fisher information, $\psi(\theta) := E_P[f](\theta)$ and $'$ denotes the derivative with respect to θ.

Proof. Substituting $E_Q - E_P = (E_P[f])'(\theta)\delta\theta + O(\delta\theta^2)$ and $V_P = V_Q + O(\delta\theta)$ into (19) yields

$$\text{KL}(P\|Q) \geq \delta\theta^2 \int_0^1 \frac{t\psi'(\theta)^2}{V_P} dt + O(\delta\theta^3) = \delta\theta^2 \frac{\psi'(\theta)^2}{2\text{Var}_P(f)} + O(\delta\theta^3). \qquad (27)$$

On the other hand, since $\mathrm{KL}(P\|Q) = \frac{1}{2}I(p)\delta\theta^2 + O(\delta\theta^3)$ holds (see [9]), in the limit $\delta\theta \to 0$, we have

$$I(p) \geq \frac{\psi'(\theta)^2}{\mathrm{Var}_P(f)}. \tag{28}$$

Hence, the result follows. $\square$

One of the important application of our lower bound is divergence estimation between two given data. Since we can approximately calculate the expected value of the FoI of the data $\{x_i\}$ such as $E[f] = \frac{1}{N}\sum_i f(x_i)$ and $\mathrm{Var}(f) = \frac{1}{N}\sum_i f(x_i)^2 - E[f]^2$, we can easily calculate the lower bound of the KL-divergence by using Theorem 1 even if the true distributions is unknown. Hence, Our lower bound can be a criteria to judge the validity of estimation results and to abort the iteration.

B. Another proof of Theorem 1

In this subsection, we prove Theorem 1 by the information geometric approach. First, we derive the relation between the derivative of the canonical divergence along the geodesic on the dually flat manifold and the Fisher information metric. Next, by applying the Cramér-Rao bound to this relation, we prove Theorem 1.

Let $P_i(i = 0, 1, \ldots, d)$ be probability distributions (measures) on the same support $\omega \subseteq \mathbb{R}^n$ and $S := \{\boldsymbol{\eta}|\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_d) \in \mathbb{R}^d, \eta_i \geq 0, \sum_{i=1}^d \eta_i \leq 1\}$. Consider mixture distributions $r(x; \boldsymbol{\eta}) := (1 - \sum_{i=1}^d \eta_i)p_0(x) + \sum_{i=1}^d \eta_i p_i(x)$ for $\boldsymbol{\eta} \in S$ and the function called potential $F(\boldsymbol{\eta}) := E_{r(x;\boldsymbol{\eta})}[\log r(x; \boldsymbol{\eta})]$, where p_i denote the Radon-Nikodym derivatives of probability measures P_i and $E_{r(x;\boldsymbol{\eta})}$ denotes the expected value of $r(x; \boldsymbol{\eta})$. By the definition of F, F is the convex function and the negative Shannon entropy. The potential F and the coordinate $\boldsymbol{\eta}$ define the dually flat structure in the manifold $M = \{r(x; \boldsymbol{\eta})|\boldsymbol{\eta} \in S\}$. We can introduce the canonical Bregman divergence [12] on M.

$$D(\boldsymbol{\eta}_1\|\boldsymbol{\eta}_2) := F(\boldsymbol{\eta}_1) - F(\boldsymbol{\eta}_2) - \sum_i \partial^i F(\boldsymbol{\eta}_2)(\eta_{1i} - \eta_{2i}), \tag{29}$$

where $\partial^i := \frac{\partial}{\partial \eta_i}$. The Riemannian metric g^{ij} on M is $g^{ij} = \partial^i\partial^j F$ and g^{ij} is equal to the Fisher information metric I^{ij}.

$$g^{ij}(\boldsymbol{\eta}) = \partial^i\partial^j F(\boldsymbol{\eta}) = I^{ij}(\boldsymbol{\eta}) := E_{r(x;\boldsymbol{\eta})}[\partial^i l_{\boldsymbol{\eta}}\partial^j l_{\boldsymbol{\eta}}], \tag{30}$$

where $l_{\boldsymbol{\eta}}$ denotes $\log r(x; \boldsymbol{\eta})$. We can easily confirm this equality by the definition of the potential F and $r(x; \boldsymbol{\eta})$. In the same way, we have

$$D(\boldsymbol{\eta}_1\|\boldsymbol{\eta}_2) = \mathrm{KL}(R(\boldsymbol{\eta}_1)\|R(\boldsymbol{\eta}_2)), \tag{31}$$

where $R(\boldsymbol{\eta})$ denotes probability distributions which correspond to $r(x; \boldsymbol{\eta})$. The geodesic on M can be written as $\boldsymbol{\eta}(t) = \boldsymbol{a}t + \boldsymbol{b}$ for the constant vector $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$ and a parameter $t \in \mathbb{R}$.

From the definition of the canonical divergence, the derivative of the canonical divergence along the geodesic is

$$\frac{d}{dt}D(\boldsymbol{\eta}(0)\|\boldsymbol{\eta}(t)) = t\sum_{i,j}\partial^i\partial^j F(\boldsymbol{\eta}(t))a_i a_j = t\sum_{i,j}g^{ij}(\boldsymbol{\eta}(t))a_i a_j = t\sum_{i,j}I^{ij}(\boldsymbol{\eta}(t))a_i a_j, \tag{32}$$

where we use (30). This equality represents the relation between the derivative of the canonical divergence along the geodesic and the Fisher information metric.

In the case M is 1-dimensional and $\eta_1(t) = t$, (32) is simplified as

$$\frac{d}{dt}D(0\|t) = tI(t). \tag{33}$$

By putting $p_0(x) = p(x)$, $p_1(x) = q(x)$, $r(x; t) = p(x) + t(q(x) - p(x))$ and applying the Cramér-Rao bound (26), we have

$$V_{R(t)} = \mathrm{Var}_{R(t)}(f) \geq \frac{(E_Q - E_P)^2}{I(t)}, \tag{34}$$

where we use $\psi(t) = E_{R(t)}[f] = (E_Q - E_P)t + E_P$. From (31), (33) and (34), we have

$$\frac{d}{dt}\mathrm{KL}(P\|R(t)) \geq \frac{t(E_Q - E_P)^2}{V_{R(t)}}. \tag{35}$$

By using $\mathrm{Var}_{R(t)}(f) = E_{R(t)}(f^2) - E_{R(t)}[f]^2 = tV_Q + (1-t)V_P + t(1-t)(E_Q - E_P)^2$, we have the same inequality as (18). The rest of the proof is the same as the previous subsection.

III. EXAMPLES

We show four examples using Theorem 1.

A. Simple discrete distribution

We consider the case $\omega = \{1, 2, 3, 4\}$, $P(1) = P(2) = P(3) = P(4) = 0.25$ and $Q(1) = 0.1, Q(2) = 0.2, Q(3) = 0.3, Q(4) = 0.4$. Then, we have $E_P = 2.5$, $E_Q = 3.0$ and $V_P = 1.25, V_Q = 1.0$. The KL-divergence is equal to 0.121777274287 and our lower bound is equal to 0.111571775657.

B. Normal distribution

The probability density function of the normal distribution is

$$p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{36}$$

and the KL-divergence is

$$KL(\mu_P, \sigma_P \| \mu_Q, \sigma_Q) = \frac{(\mu_Q - \mu_P)^2}{2\sigma_Q^2} + \frac{\sigma_P^2}{2\sigma_Q^2} - \frac{1}{2} + \log\frac{\sigma_Q}{\sigma_P}. \tag{37}$$

1) We consider the case $\sigma_P = \sigma_Q = \sigma$. In this case, the KL-divergence is simplified as $KL(\mu_P, \sigma_P\|\mu_Q, \sigma_Q) = \frac{(\mu_Q - \mu_P)^2}{2\sigma^2}$. Choosing the foI as $f(x) = x$, we have $E_P = \mu_P$, $E_Q = \mu_Q$ and $V_P = V_Q = \sigma^2$. Figure 1. displays the result of comparison of the KL-divergence and our lower bound for $\beta := \frac{\mu_Q - \mu_P}{\sigma}$.
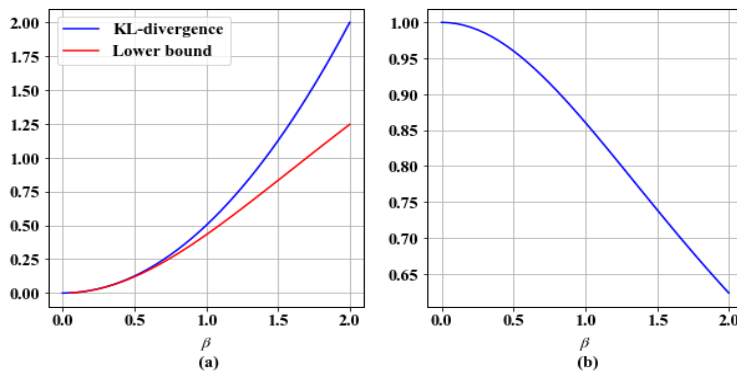

Fig. 1. (a) Comparison of the KL-divergence (blue line) and our lower bound (red line) for the normal distribution. (b) The ratio of our lower bound to the KL-divergence.

2) We consider the case $\mu_P = \mu_Q = 0$. In this case, the KL-divergence is simplified as $\frac{\sigma_P^2}{2\sigma_Q^2} - \frac{1}{2} + \log\frac{\sigma_Q}{\sigma_P}$. If we choose the foI as $f(x) = x$, the lower bound is trivial from (19). Hence, we choose the foI as $f(x) = x^2$ and we have $E_P = \sigma_P^2$, $E_Q = \sigma_Q^2$ and $V_P = 2\sigma_P^4$, $V_Q = 2\sigma_Q^4$. Figure 2. displays the result of comparison of the KL-divergence and our lower bound for $\beta := \frac{\sigma_Q}{\sigma_P}$.
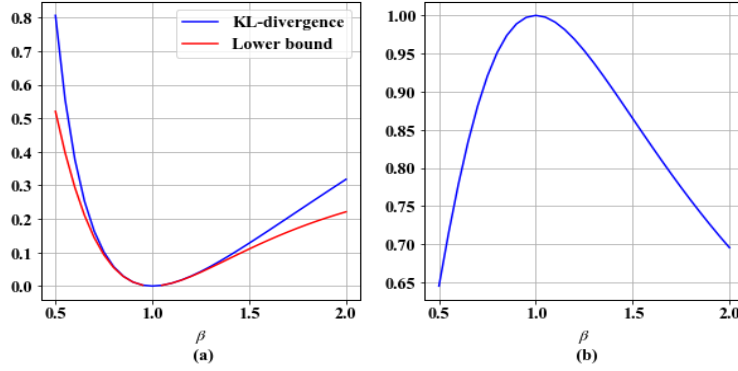
Fig. 2. (a) Comparison of the KL-divergence (blue line) and our lower bound (red line) for the normal distribution. (b) The ratio of our lower bound to the KL-divergence.

C. Exponential distribution

The probability density function of the exponential distribution is

$$p(x; \nu) = \frac{1}{\nu} \exp\left(-\frac{1}{\nu} x\right). \tag{38}$$

and the KL-divergence is

$$KL(\nu_P \| \nu_Q) = \frac{\nu_P}{\nu_Q} - 1 + \log \frac{\nu_Q}{\nu_P}. \tag{39}$$

Choosing the foI as $f(x) = x$, we have $E_P = \nu_P$, $E_Q = \nu_Q$ and $V_P = \nu_P^2$, $V_Q = \nu_Q^2$. Figure 3. displays the result of comparison of the KL-divergence and our lower bound for $\beta := \frac{\nu_Q}{\nu_P}$.
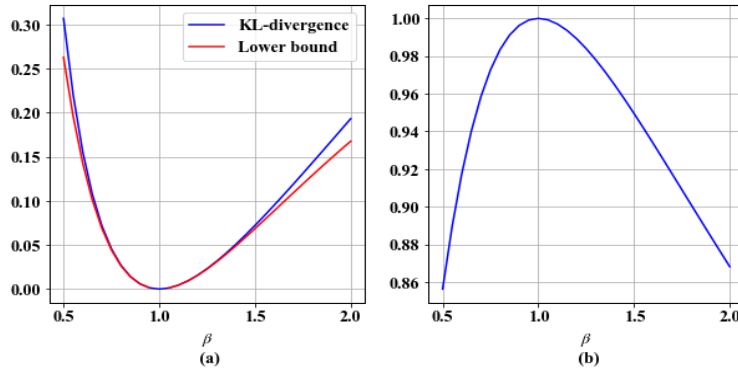

Fig. 3. (a) Comparison of the KL-divergence (blue line) and our lower bound (red line) for the exponential distribution. (b) The ratio of our lower bound to the KL-divergence.

From these examples, we can confirm that our lower bounds behave similarly to the actual value of the KL-divergence.

IV. CONCLUSION

We have derived the new lower bound for the KL-divergence which only depends on the expected value and the variance of the FoI by applying the HCRB and we have given another proof by information geometric approach. We also have shown that the equality holds for the Bernoulli distributions and we have obtained the Cramér-Rao bound from our lower bound when two distributions are very close. Furthermore, we have described that our bound behaves similarly to the KL-divergence by some examples of numerical calculation. One of the important point of our bound is easiness of calculation even if the true distribution is unknown. We expect the application range of it will expand.

References

[1] Shun-ichi Amari. *Information geometry and its applications*. Springer, 2016.

[2] Shun-ichi Amari and Andrzej Cichocki. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 58(1):183–195, 2010.

[3] Douglas G Chapman, Herbert Robbins, et al. Minimum variance estimation without regularity assumptions. *The Annals of Mathematical Statistics*, 22(4):581–586, 1951.

[4] Andrzej Cichocki and Shun-ichi Amari. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

[5] Harald Cramér. *Mathematical methods of statistics*, volume 9. Princeton university press, 1999.

[6] Masoumeh Dashti and Andrew M Stuart. The bayesian approach to inverse problems. *Handbook of Uncertainty Quantification*, pages 1–118, 2016.

[7] JM Hammersley. On estimating restricted parameters, jr statist. *Soc.(B)*, 12, 1950.

[8] Markos A Katsoulakis, Luc Rey-Bellet, and Jie Wang. Scalable information inequalities for uncertainty quantification. *Journal of Computational Physics*, 336:513–545, 2017.

[9] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[10] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[11] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[12] Frank Nielsen. An elementary introduction to information geometry. *arXiv preprint arXiv:1808.08271*, 2018.

[13] C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in statistics*, pages 235–247. Springer, 1992.

[14] Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.