Correspondance between f-divergence and Surrogate Loss in Binary Classification

Yifeng Chu

1 introduction

1.1 Overview

This report closely follows Nguyen *et al.*'s work [4], where they establish the Correspondance between loss functions used for 0-1 loss in binary Classification and f-divergence of label-related distributions induced by dimension-reducing quantizers for feature vectors. The existence of quantizers change turn the classical problem of learning a discriminant function to a jointly estimation of optimal quantizer and discriminant function. Based on this setting, they investigate the necessary and sufficient conditions for surrogate loss function that yield Bayes consistency for empirical risk minimizer.

Note that no extended work or generalized results are shown in this report. Instead, it may server as a supplementary materials to understand the original paper as we fill in extra details and more accessible explanations. The following sections will discuss problem formulations, main theorems in Nguyen's paper and related works. Furthermore, we will focus our attention on one of major results where it includes beautiful constructive proof and briefly mention the other two due to limited space.

1.2 Problem Formulation and Notations

The notations used throughout the report will be exactly the same with the original paper for consistency. Consider the classical setting of binary classification problem: let $\mathcal{X}$ be Borel subset of $\mathbb{R}^d$, $\mathcal{Y}$ be the set $\{-1, +1\}$, and $\mathcal{P}$ be the joint law of the pair of random variable (X, Y) that takes values in $\mathcal{X}\times\mathcal{Y}$. Let $\mathcal{Q}$ be a collection of stochastic transformation $Q : \mathcal{X} \to \mathcal{Z}$ such that $\mathcal{Z}$ takes values in discrete vectors of dimension m (dimension does not matter for the purpose of discussion). Let $\gamma : \mathcal{Z} \to \mathbb{R}$ be discriminant function. Given training samples $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, the goal is to determine a pair of (γ, Q) such that Bayes risk $R_{\text{Bayes}}(\gamma, Q) := \mathbf{P}[Y \neq \text{sgn}(\gamma(Z))] = \mathbf{E}\left[\mathbf{1}_{\{\text{sgn}(Y\gamma(X))\geq 0\}}\right]$ is minimized. Since 0-1 loss is nonconvex, we typically consider a relaxation of 0-1 loss: ϕ such that $R_\phi(\gamma, Q) := \mathbf{E}[\phi(Y\gamma(X))]$.

2 Summary of Main Results and Related Work

2.1 First Result

Let optimal ϕ-risk be

$$R_\phi(Q) := \inf_{\gamma \in \Gamma} R_\phi(Q, \gamma).$$

where Γ is the set of all measurable functions of $\mathcal{Z}$. Let label-related distribution induced by quantizer $Q(z|x)$ be

$$\mu(z) := \mathbf{P}[Y = 1, Z = z] = p \int_{\mathcal{X}} Q(z|x) \mathrm{d}\mathbf{P}[x|Y = 1] \tag{1}$$

$$\pi(z) := \mathbf{P}[Y = -1, Z = z] = q \int_{\mathcal{X}} Q(z|x) \mathrm{d}\mathbf{P}[x|Y = -1] \tag{2}$$

$$\tag{3}$$

where $p = \mathbf{P}[Y = 1], q = \mathbf{P}[Y = -1]$. Let f-divergence between μ and π be

$$I_f(\mu, \pi) := \sum_{z \in \mathcal{Z}} \pi(z) f\left(\frac{\mu(z)}{\pi(z)}\right)$$

where $f : [0, +\infty] \to \mathbb{R} \cup \{+\infty\}$ is a continuous convex function. The brief discription of the first main result would be the following relation

$$R_\phi(Q) = -I_f(\mu, \pi) \tag{4}$$

under certain regularity conditions. Since we will have a thorough discussion on this result, details will be presented later.

2.2 The other two results

Note that for the other two results, only informal demonstration are introduced to briefly explain the results, essential notions included to avoid unnecessary ambiguities. For more precise statements, refer to [4]. Conisder specific class of f that has the expression

$$f(u) = c \min\{u, 1\} + au + b \tag{5}$$

where a, b, and c are scalars. Consider the sequence of classes of functions and quantizers,

$$\mathcal{C}_1 \subset \mathcal{C}_2 \subset \ldots \subset \Gamma \quad \text{and} \quad \mathcal{D}_1 \subset \mathcal{D}_2 \subset \ldots \subset \mathcal{Q}$$

where Γ is the set of all measurable functions of $\mathcal{Z}$ and $\mathcal{Q}$ is the class of constrained quantizers with restriction that μ and π are strictly positive measures. Let empirical ϕ-risk and corresponding risk minimizer be

$$\hat{R}_\phi(\gamma, Q) = \min_{(\gamma,Q)\in(\mathcal{C}_n,\mathcal{D}_n)} \frac{1}{n}\sum_{i=1}^{n}\sum_{z\in\mathcal{Z}} \phi(Y_i\gamma(z))Q(z|X_i)$$
$$(\gamma_n^*, Q_n^*) = \arg\min_{\gamma\in\Gamma} \hat{R}_\phi(\gamma, Q)$$

Let minimum Bayes risk be

$$R^*_{\text{Bayes}} := \inf_{(\gamma,Q)\in(\Gamma,\mathcal{Q})} R_{\text{Bayes}}(\gamma, Q) \tag{6}$$

Theorem 2 states if f have the expression in Equation (5) for all $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, we have

$$\lim_{n\to\infty} R_{\text{Bayes}}(\gamma_n^*, Q_n^*) = R^*_{\text{Bayes}} \quad \text{in probability} \tag{7}$$

under certain regularity conditions. In other words, theorem 2 provides a sufficient condition on surrogate loss function ϕ where ERM algorithm can yield Bayes consistency.

Then, a natural question to ask would be whether there exist other classes of surrogate loss function ϕ so that ERM algorithm is also Bayes consistent. The brief answer is negative. Qualitively speaking, we need to group loss functions based on certain relations that suffices to dichotomy the collections with respect to Bayes consistency. Nguyen *et al.* approached this problem by the following steps:

1. Establish the partial ordering of quantizers based on optimal risk incurred by using loss function ϕ (the partial ordering is extremely similar to the concept *more informative* experiment in Blackwell's work [2])

2. Define equivalent class of loss functions by grouping loss functions that result in the same ordering between any two quantizers under any possible distributions, which is defined as *universal equivalence* by Nguyen *et al.* [4].

3. First find rules (rules are statement in Theorem 3) to group equivalent f instead of rules to group ϕ. Then group f by suing the correspondance between f and ϕ in Equation (4).

4. Use the facts

   (a) The necessary condition for ϕ such that the pair (γ, Q) that yield Bayes consistency has to minimize Bayes risk in population level. In other words, f has to induce total variation distance via correspondance between ϕ and f in Equation (4) since divergence functional induced by 0-1 loss differs total variation distance only by constant.

(b) f such that I_f is universal is equivalent to the family in Equation (5) by rules in step 3.

To summarize, Nguyen *et al.* group ϕ (or compare ϕ) in terms of f. The motivation of doing so is not explicitly mentioned in the paper. However, a reasonable explanation is related to the fact that there is more degree of freedom to construct loss function ϕ that results in same f than the other way around. More details will be given in the Section 3. Although a quick example given in [4] can be used to demonstrate the point. Consider

2.3 Related Works

Nguyen *et al.* mentioned more than once that, the discussion related to loss function and divergence between label-related distribution first appeared in Blackwell's work on comparison of experiment [2]. Even though the phrases in two papers are not exactly the same, the given quantizer can induce an experiment (experiment is defined to be a finite collection of possible distribution on observations) and minimum loss that actions can incur given specific experiment corresponds to the f-divergence. An further generalization to multiclass classification is given by [**?**]

Also, when discussing Bayes consistency, Nguyen *et al.* considered the setting that ERM is minimizing over all measurable functions which may not necessarily the case. For most of time, we always constrain the hypothesis set to a certain class of function. Sriperumbudur *et al.* [5] has some work on similar subject except in their case, another distance measure on probability, named *the integral probability metrics*, are considered.

On application aspect, establishing the correspondance between discriminant function and f-divergence provides a way to estimate divergence by convex risk minimization [3].

3 Discussions

First, we give precise statement of theorem 1 and helper lemma under certain assumptions. Since the proofs of the following Lemma 1, Lemma 2 are very straightforward, it is not necessary to rewrite it here again. Instead, we illustrate the results by some imaginary plots by describing procedure in steps to provide geometric intuitions. As for Theorem 1, we will rewrite the proof and fill in more details.

Assume the following,

Assumption 1. *Assume the following:*

A1. ϕ *is convex and differentiable at 0 and* $\phi'(0) < 0$*;*

A2. ϕ *is continuous;*

A3. Let $\alpha^* = \inf\{\alpha \in \mathbb{R} \cup \{+\infty\} : \phi(\alpha) = \inf \phi\}$. *If* $\alpha^* < +\infty$, *then for any* $\varepsilon > 0$

$$\phi(\alpha^* - \varepsilon) \geq \phi(\alpha^* + \varepsilon) \tag{8}$$

Remark. *In the original paper, the statement of A1 is:* ϕ *is classification-calibrated, which is also known as Fisher consistency that origins from classical parameter estimation setting. Informally, we say* ϕ *is classification-calibrated if* $\gamma^* = \arg\min_{\gamma\in\Gamma} \mathbf{E}[\phi(Y\gamma(X))]$ *is such that* $\operatorname{sgn}(\gamma^*) = f^*$, *where* f^* *is optimal Bayes decision rule. Nguyen* et al. *used a definition tailored to binary classification case and gave a necessary and sufficient condition for* ϕ *to be classification-calibrated when* ϕ *is convex (same as A1 stated here). Proof can be found in [1].*

For a lower semicontinuous convex function, $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$, convex conjugate of f is defined to be: $f^*(u) = \sup_{v\in\mathbb{R}}(uv - f(v))$. Consider

$$\Psi(\beta) = f^*(-\beta) \tag{9}$$

. Let $\beta_1 := \inf\{\beta : \Psi(\beta) < +\infty\}$ and $\beta_2 := \inf\{\beta : \Psi(\beta) \leq \inf \Psi\}$.

Theorem 1. *[4]*

(a) For any margin-based surrogate loss function ϕ, *there is an* f*-divergence such that* $R_\phi(Q) = -I_f(\mu, \pi)$ *for some lower semicontinuous convex function* f *In addition, if* ϕ *is a decreasing convex loss function that satises conditions A1, A2 and A3, then the following properties hold:*

(i) Ψ *is a decreasing and convex function;*

(ii) $\Psi(\Psi(\beta)) = \beta$ *for all* $\beta \in (\beta_1, \beta_2)$

(iii) there exists a point $u^* \in (\beta_1, \beta_2)$ *such that* $\Psi(u^*) = u^*$

(b) Conversely, if f *is a lower semicontinuous convex function satisfying all conditions (i)–(iii), there exists a decreasing convex surrogate loss* ϕ *that induces the* f*-divergence in the sense of following two equations.*

$$f(u) := -\inf_{\alpha}(\phi(-\alpha) + \phi(\alpha)u) \tag{10}$$

$$R_\phi(Q) = -I_f(\mu, \pi) \tag{11}$$

The proof of Theorem 1 mainly relies on exploiting the properties of constructed intermediate functions, which are presented as follows. Define

$$\phi^{-1}(\beta) := \inf\{\alpha : \phi(\alpha) \leq \beta\} \tag{12}$$

where $\inf \varnothing := +\infty$.

Lemma 1 (Properties of ϕ^{-1})**.** *[4] Suppose that ϕ is a convex loss satisfying assumptions A1, A2 and A3.*

(a) For all $\beta \in \mathbb{R}$ such that $\phi^{-1}(\beta) < +\infty$, the inequality $\phi(\phi^{-1}(\beta)) \leq \beta$ holds. Furthermore, equality occurs when ϕ is continuous at $\phi^{-1}(\beta)$.

(b) The function $\phi^{-1} : \mathbb{R} \to \overline{\mathbb{R}}$ is strictly decreasing and convex.

Remark. *(a) The results still hold without assuming A1-A3 as long as ϕ is convex.*

(b) When ϕ is convex, what prevents us to define inverse of ϕ , for certain β is that there may exist two distinct α such that $\phi(\alpha) = \beta$. One solution is to project increasing part of curve (if it exists) to decreasing part to make ϕ monotone on $\Lambda = \{\alpha : \phi^{-1}(\beta) \in \mathbb{R}\}$, which is roughly geometry interpretation of the effect of ϕ^{-1}. In this way, we can guarantee inverse of ϕ is well-defined.

(c) By plotting out the figure, we can easily obtain statement in part (a) of the lemma. As for part (b), since plot of inverse can be obtained via mirroring plot of ϕ by line $\alpha = \beta$, it can be observed directly ϕ^{-1} is convex if ϕ is monotonically decreasing.

(d) Based on the definition of $\phi(\phi^{-1}(\beta))$, we can claim all functions have stair shape where lines that are not lying along $\alpha = \beta$ must be flat. Indeed, all discontinuities in $\phi(\phi^{-1}(\beta))$ must sits left to or at α^*.

Based on function ϕ^{-1}, we define $\tilde{\Psi} : \mathbb{R} \to \overline{\mathbb{R}}$ to be:

$$\tilde{\Psi}(\beta) := \begin{cases} \phi\big(-\phi^{-1}(\beta)\big), & \text{if } \phi^{-1}(\beta) \in \mathbb{R} \\ +\infty, & \text{otherwise} \end{cases} \tag{13}$$

Lemma 2 (Properties of $\tilde{\Psi}$)**.** *[4] Suppose that ϕ is a convex loss satisfying assumptions A1, A2 and A3. We have:*

(a) $\tilde{\Psi}$ is strictly decreasing in the interval $(\tilde{\beta}_1, \tilde{\beta}_2)$. If ϕ is decreasing, then $\tilde{\Psi}$ is also decreasing in $-\infty, +\infty$. In addition, $\tilde{\Psi} = +\infty$ for $\beta < \tilde{\beta}_1$

(b) $\tilde{\Psi}$ in convex in $(-\infty, \tilde{\beta}_2)$. If ϕ is a decreasing function, then $\tilde{\Psi}$ is convex in $(-\infty, +\infty)$

(c) $\tilde{\Psi}$ is lower semi-continuous, and continuous in its domain.

(d) For any $\alpha \geq 0$, $\phi(\alpha) = \tilde{\Psi}(\phi(-\alpha))$. In particular, there exists $u^* \in (\tilde{\beta}_1, \tilde{\beta}_2)$ such that $\tilde{\Psi}(u^*) = u^*$.

(e) The function $\tilde{\Psi}$ satisfies $\tilde{\Psi}(\tilde{\Psi}(\beta)) < \beta$ for all $\beta \in \mathrm{dom}(\tilde{\Psi})$. Moreover, if ϕ is a continuous function on its domain $\{\alpha \in \mathbb{R} : \phi(\alpha) < +\infty\}$, then $\tilde{\Psi}(\tilde{\Psi}(\beta)) = \beta$ for all $\beta \in (\tilde{\beta}_1, \tilde{\beta}_2)$

Remark. *(a) Same as Lemma 1. A1-A3 assumptions are not necessarily needed for several results.*

(b) Based on notation, the effect of performing $\tilde{\Psi}$ in geometry is when β is specified, find the leftmost α such that $\phi(\alpha) \geq \beta$, then flip the sign of α to get the corresponding function value. part (d) in Lemma 2 has very straightforward interpretation of this procedure. Indeed, we fix $\alpha > 0$ first, let $\beta' = \phi(\alpha)$. For the convenience of discussion, consider the Since ϕ on $(-\infty, 0)$ is decreasing, thus inverse is well-defined. We have $\phi^{-1}(\beta') = \alpha$. Therefore $\phi(\alpha) = \tilde{\Psi}(\phi(-\alpha))$.

(c) part (a) and (e) are good examples where we can see how A1, A3 plays its role here. The reason why $\tilde{\Psi}$ always decreases (or never increases) the input's value is due to the facts that if $\alpha^* \in \mathbb{R}$, right side is always flatter than right side since for the same deviation from α^*, the left side alpha incurs larger loss and another fact is that α^* is always larger than 0 since point 0 is at decreasing side, that indicates whenever the points left from a^* flip the sign to $-\alpha$, the resulted $\phi(-alpha)$ is always no greater than original $\phi(\alpha)$.

proof of Theorem 1. For part (a), as we can observe from the results, even when ϕ is not convex, f can also be specified. Indeed, Then the optimal ϕ-risk is:

$$
\begin{aligned}
R_\phi(Q) &= \inf_\gamma \mathbb{E}\phi(Y\gamma(Z)) && (14)\\
&= \inf_\gamma \sum_z \phi(\gamma(z))\mu(z) + \phi(-\gamma(z))\pi(z) && (15)\\
&= \sum_{z\in Z} \inf_\alpha (\phi(\alpha)\mu(z) + \phi(-\alpha)\pi(z)) && (16)\\
&= \sum_z \pi(z) \inf_\alpha \left(\phi(-\alpha) + \phi(\alpha)\frac{\mu(z)}{\pi(z)}\right) && (17)
\end{aligned}
$$

Equation (16) is due to for each z, $\gamma(z)$ has fixed valued. Thus dependency on γ can be suppressed. Now we show with extra assumptions on ϕ, the corresponding f will have some properties.

$$
\begin{aligned}
f(u) &= -\inf_{\alpha\in\mathbb{R}} (\phi(-\alpha) + \phi(\alpha)u) && (18)\\
&= -\inf_{\{\alpha,\beta|\phi^{-1}(\beta)\in\mathbb{R},\phi(\alpha)=\beta\}} (\phi(-\alpha) + \beta u) && (19)\\
&= -\inf_{\beta:\phi^{-1}(\beta)\in\mathbb{R}} \left(\phi\left(-\phi^{-1}(\beta)\right) + \beta u\right) && (20)\\
&= -\inf_{\beta\in\mathbb{R}} (\beta u + \tilde{\Psi}(\beta)) && (21)\\
&= \sup_{\beta\in\mathbb{R}} (-\beta u - \tilde{\Psi}(\beta)) = \tilde{\Psi}^*(-u) && (22)
\end{aligned}
$$

In Equation (19), β is introduced to substitute α. Therefore, infimum over α becomes infimum over the pair (α, β) under restriction $\phi(\alpha) = \beta$. In Equation (20), for level set $\phi(\alpha) = \beta$, it may contain more than one elements. Since A3 tells us for same deviation from α^*, deviation of left side always suffer more loss or has greater function value. Therefore, if we pick leftmost α such that $\phi(\alpha) = \beta$, we can guarantee $\phi(-\alpha) < \phi(\alpha)$. Furthermore, leftmost α is exactly how $\phi^{-1}(\beta)$ is defined. That is why we can substitute α with $\phi^{-1}(\beta)$ to get the infimum over α. (Note that decreasing allows flat components comparing to strictly decreasing)

Also, since ϕ is decreasing, by using part (b) and (c) of Lemma 2, we can express $\tilde{\Psi}$ as

$$\tilde{\Psi}(\beta) = \tilde{\Psi}^{**}(\beta) = f^*(-\beta) = \Psi(\beta) \tag{23}$$

Therefore, $\tilde{\Psi} = \Psi$, which concludes the proof of part (a).

One things that is of interest is that we can claim that any ϕ that can induce a f-divergence must have the expression:

$$\phi(\alpha) = \begin{cases} u^*, & \text{if } \alpha = 0 \\ \Psi(g(\alpha + u^*)), & \text{if } \alpha > 0 \\ g(-\alpha + u^*), & \text{if } \alpha < 0 \end{cases} \tag{24}$$

where $g : [u^*, +\infty) \to \overline{\mathbb{R}}$ is some increasing continuous and convex function. We consider α in different intervals (or point). Let $u^* := \alpha(0) \in (\beta_1, \beta_2)$. From part (d) of Lemma 2, Ψ defined via α gives $\Psi(\alpha(0)) = \alpha(0)$, which satisfies the condition $\Psi(u^*) = u^*$ for Ψ related to conjugate of f. If $\alpha > 0$, then $\alpha \mapsto \phi(-\alpha)$ is increasing and convex. Then if we use some increasing continous and convex function to construct a legal $\phi(-\alpha)$ on $(0, +\infty)$, we can use part (d) of Lemma 2 to represent $\phi(\alpha)$ using $\phi(-\alpha)$ with Ψ. And when $\alpha < 0$, we have already covered this case when considering $\alpha > 0$.

Now we proceed to prove part (b) of Theorem 1. Recall that part (b) requires us to specify ϕ when f is given. The main point here is not only we can show the existence, but also we claim that any ϕ constructed in Equation (24) can lead to f related to Ψ.

Since f is lower semicontinuous by assumption, again we can write

$$\begin{aligned} f(u) &= f^{**}(u) = \Psi^*(-u) \\ &= \sup_{\beta \in \mathbb{R}} (-\beta u - \Psi(\beta)) = -\inf_{\beta \in \mathbb{R}} (\beta u + \Psi(\beta)) \end{aligned}$$

Also recall that $\tilde{\Psi}$ is defined via ϕ and Ψ is defined via f. The available relations we have are $f = -\inf_{\beta \in \mathbb{R}}(\beta u + \Psi(\beta))$. By substituting $\beta = \phi(\alpha)$, we have $f = -inf_{\beta,\alpha}(\phi(\alpha)u) + \Psi(\beta)$. Meanwhile, $\tilde{\Psi}(\beta) = \phi(-\phi^{-1}(\beta)) = \phi(-\alpha)$. If we can show $\tilde{\Psi} = \Psi$, then existence of ϕ can be verified. Indeed, consider $u^* \in (\beta_1, \beta_2)$. When $\beta \geq u^*$, since $g(u^*) = u^*$ and g is increasing, there exists $\alpha > 0$ such that $g(\alpha + u^*) \geq u^*$. Since from construction $\phi(-\alpha) = g(\alpha + u^*) = \beta$, to make $\phi^{-1}(\beta) = -\alpha$ be well-defined, we need to choose smallest $-\alpha$, which

is equivalent to choosing largest α. It then follows quickly that $\tilde{\Psi}(\beta) = \phi(-\phi^{-1}(\beta)) = \phi(\alpha) = \Psi(g(\alpha + u^*)) = \Psi(\beta)$, where the first equality is from definition, second from $\phi^{-1}(\beta) = \alpha$, third from construction of ϕ by choosing proper g. When $\beta < \beta_1$, then $\Psi(\beta) = +\infty$. By following the link from Ψ to ϕ to $\tilde{\Psi}$, we obtain $\tilde{\Psi}(\beta) = +\infty$. The last case where $\beta_1 \leq \beta < u^* < \beta_2$ is very similar to the first case. By the way of constructing ϕ, $\phi(\alpha) = \Psi(g(\alpha+u^*))$ for $\alpha > 0$. Choose smallest α such that $\phi(\alpha) = \beta$ such that $\alpha = \phi^{-1}(\beta)$ is well-defined. Then $\tilde{\Psi}(\beta) = \phi(-\phi^{-1}(\beta)) = \phi(-\alpha) = g(\alpha + u^*) = \Psi(\Psi(g(\alpha + u^*)))$, where first equality is by definition of $\tilde{\Psi}$, second by $\alpha = \phi^{-1}(\beta)$, thrid by construction of ϕ and last by assumption (ii) in Theorem 1 and the fact $g(\alpha + u^*) \in (\beta_1, \beta_2)$.

□

References

[1] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. 101(473):138–156.

[2] David Blackwell. COMPARISON OF EXPERIMENTS.

[3] John Duchi, Khashayar Khosravi, and Feng Ruan. Multiclass classification, information, divergence and surrogate risk. 46(6):3246–3275.

[4] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. 56(11):5847–5861.

[5] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. On surrogate loss functions and f -divergences. 37(2):876–904.

[6] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On integral probability metrics, \phi-divergences and binary classification.