Fenchel-Young Losses with Skewed Entropies for Class-posterior Probability Estimation

Han Bao
The University of Tokyo
RIKEN AIP
tsutsumi@ms.k.u-tokyo.ac.jp

Masashi Sugiyama
RIKEN AIP
The University of Tokyo
sugi@k.u-tokyo.ac.jp

Abstract

We study class-posterior probability estimation (CPE) for binary responses where one class has much fewer data than the other. For example, *events* such as species co-occurrence in ecology and wars in political science are often much rarer than *non-events*. Logistic regression has been widely used for CPE, while it tends to underestimate the probability of rare events. Its main drawback is symmetry of the logit link—symmetric links can be misled by small and imbalanced samples because it is more incentivized to overestimate the majority class with finite samples. Parametric skewed links have been proposed to overcome this limitation, but their estimation usually results in nonconvex optimization unlike the logit link. Such nonconvexity is knotty not only from the computational viewpoint but also in terms of the parameter identifiability. In this paper, we provide a procedure to derive a convex loss for a skewed link based on the recently proposed Fenchel-Young losses. The derived losses are always convex and have a nice property suitable for class imbalance. The simulation shows the practicality of the derived losses.

1 Introduction

Modeling and estimating class-posterior probabilities of binary responses is a fundamental problem in many fields, where a large gap often exists between the observed numbers of events and nonevents. For instance, species co-occurrence in ecology (Jiang et al., 2013) and dyads of countries at war in political science (King and Zeng, 2001) are much fewer than nonevents because of combinatorics. Such rarity also emerges in market promotion (Wang and Dey, 2010) and infection in epidemiology (Breslow, 1996). In these fields, the modeling perspective helps to grasp the underlying mechanisms, and the class-posterior probability plays a key role in supporting a political decision and scientific claim. Hence, the class-posterior probability estimation (CPE) problem with binary responses has been studied extensively (Buja et al., 2005).

Logistic regression is one of the common CPE approaches, where parameters of a logit model are estimated via maximum likelihood estimation (MLE) from the generalized linear model perspective (McCullagh and Nelder, 1989)—equivalently, the parameters are estimated by minimizing the log loss from the machine learning perspective (Buja et al., 2005). A number of studies have revealed that logistic regression would underestimate class-posterior probability of rare events under imbalanced and limited samples (Czado and Santner, 1992; King and Zeng, 2001; Wang and Dey, 2010; Menon et al., 2012). One of the major reasons is that the logit link is essentially symmetric in responses, while they are often distributed differently—the link is *misspecified* there. While several researches have tackled with Jeffreys' prior (Firth, 1993), undersampling (Wallace and Dahabreh, 2012), and cost-sensitive learning (King and Zeng, 2001), these approaches do not address the misspecification.

The other approaches is to replace the logit link with a skewed link. This line includes the generalized logit link (Stukel, 1988), skewed logit link (Chen et al., 1999), skewed generalized t-link (Kim et al., 2008), generalized extreme value (GEV) link (Wang and Dey, 2010), and symmetric power link (Jiang et al., 2013). These links entail wide ranges of skewness controlled by hyperparameters and the misspecification issue is resolved by model selection. Among them, we mainly

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

focus on the GEV link due to its simplicity. Wang and Dey (2010) applied the GEV link with MLE. This is equivalent to the combination of the log loss and GEV link, resulting in nonconvex optimization. Agarwal et al. (2014) derived *canonical proper losses* (Buja et al., 2005; Reid and Williamson, 2010) with the GEV link. Yet their loss is still not always convex because of the bounded support of the GEV distribution.

To overcome the nonconvexity, we utilize *Fenchel-Young losses* (Blondel et al., 2020) to derive a convex loss. Fenchel-Young losses provide a general recipe for a convex loss given an entropic regularizer of the prediction function, and have achieved notable success in structured prediction (Martins and Astudillo, 2016; Niculae et al., 2018; Nowak-Vila et al., 2020). While Blondel et al. (2020) derived a loss from an entropic regularizer, we instantiate Fenchel-Young losses for binary CPE from skewed link families such as the GEV link. The derived losses are always convex unlike the canonical proper loss and have a nice separation margin property which allows us to penalize predictions of the rare class more. Our experiments demonstrate that the proposed loss can provide more accurate class-posterior probability estimates with small and imbalanced samples than the existing methods.

Related work. Imbalanced classification has been tackled by many approaches such as sampling (Chawla et al., 2002; Dal Pozzolo et al., 2015), cost-sensitive learning (Elkan, 2001), modified losses (Lin et al., 2017; Cao et al., 2019; Cui et al., 2019; Charoenphakdee et al., 2020), structural surrogates based on the F-measure (Joachims, 2005; Eban et al., 2017; Bao and Sugiyama, 2020), Jaccard (Yu and Blaschko, 2015; Berman et al., 2018), and Dice index (Milletari et al., 2016; Li et al., 2020; Nordström et al., 2020). Note that their goal is not CPE but to predict class labels. CPE has often been applied to the F-measure maximization (Ye et al., 2012; Koyejo et al., 2014) without addressing the difficulty of imbalanced CPE.

It was not until recently that CPE has been studied theoretically. Telgarsky et al. (2015) provides CPE error bounds for convex risk minimizers, while Mey and Loog (2021) provides bounds for proper losses.

2 Background: Fenchel-Young Losses from Entropies

The basics of Fenchel-Young losses are described here.

Notation. The extended real line is denoted by $\overline{\mathbb{R}} \stackrel{\text{def}}{=} \mathbb{R} \cup \{\pm\infty\}$. The domain of a function $\Omega : \mathbb{R} \to \overline{\mathbb{R}}$ is denoted by $\mathsf{dom}(\Omega) \stackrel{\text{def}}{=} \{\eta \in \mathbb{R} \mid \Omega(\eta) < \infty\}$. We write the convex dual of Ω by $\Omega^\star(\theta) \stackrel{\text{def}}{=} \sup_{\eta \in \mathsf{dom}(\Omega)} \theta\eta - \Omega(\eta)$. We write $[x]_+ \stackrel{\text{def}}{=} \max\{x, 0\}$. The indicator function of a set $\mathcal{C}$ is denoted by $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$, taking 0 if $\boldsymbol{x} \in \mathcal{C}$ and ∞ otherwise. The support of a continuous probability distribution on $\mathbb{R}$ with the cumulative distribution function (CDF) F is denoted by $\mathrm{supp}(F) \stackrel{\text{def}}{=} \{\theta \in \mathbb{R} \mid F(\theta+\varepsilon) - F(\theta-\varepsilon) > 0 \ \forall \varepsilon > 0\}$.[1]

Problem setup. We consider binary supervised learning with input variable $\boldsymbol{x} \in \mathcal{X}$ and associated outcome $y \in \mathcal{Y} = \{0, 1\}$. We assume an underlying distribution $\mathbb{P}$ on $\mathcal{X} \times \mathcal{Y}$ from which both training and test examples are drawn independently and identically. Denote the class-posterior probability associated with $\mathbb{P}$ by $\eta(\boldsymbol{x}) \stackrel{\text{def}}{=} \mathbb{P}(Y = 1 | X = \boldsymbol{x})$. Given an i.i.d sample $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}^n$, our aim is to learn a CPE model $\widehat{\eta} : \mathcal{X} \to [0, 1]$ that is as close to η as possible.

Regularized prediction. Blondel et al. (2020) provided a generic framework to construct loss functions from a *prediction regularizer*. We describe the definition in the binary classification case where $\mathcal{Y} = \{0, 1\}$.

In binary classification, our strategy is to first fit a model $g : \mathcal{X} \to \mathbb{R}$ producing a prediction score $\theta = g(\boldsymbol{x})$. The class y is then predicted: $y = 1$ if $\theta > 0$ and $y = 0$ if $\theta \leq 0$. This is equivalent[2] to the prediction function $\widehat{y}(\theta) = \mathsf{arg\,max}_{y \in \mathcal{Y}} \theta y$. It is smoothed due to its non-differentiability by a convex regularizer.

Definition 1 (Regularized prediction function (Blondel et al., 2020))**.** Let $\Omega : \mathbb{R} \to \overline{\mathbb{R}}$ be a regularizer. The *prediction function* regularized by Ω is defined as

$$\widehat{y}_\Omega(\theta) \in \mathop{\mathsf{arg\,max}}_{\eta \in \mathsf{dom}(\Omega)} \theta\eta - \Omega(\eta). \tag{1}$$

For example, when $\Omega = -\mathsf{H}_\mathrm{S} + \mathbb{1}_{[0,1]}$, where $\mathsf{H}_\mathrm{S}(\eta) \stackrel{\text{def}}{=} -\eta \log \eta - (1-\eta)\log(1-\eta)$ is the Shannon entropy, $\widehat{y}_\Omega$ recovers the inverse logit $\widehat{y}_\Omega(\theta) = \frac{1}{1+\exp(-\theta)}$. When $\Omega = -\mathsf{H}_\mathrm{T} + \mathbb{1}_{[0,1]}$ with $\mathsf{H}_\mathrm{T}(\eta) \stackrel{\text{def}}{=} \frac{1}{2}\eta^2$, $\widehat{y}_\Omega$ is the Euclidean projection $\widehat{y}_\Omega(\theta) = \mathsf{arg\,min}_{\eta \in [0,1]} (\eta - \theta)^2$.

Fenchel-Young losses. Given a regularized predictor $\widehat{y}_\Omega$, we are interested in an appropriate choice of loss functions. Fenchel-Young losses admit good properties. First, we introduce Fenchel-Young losses.

Definition 2 (Fenchel-Young loss (Blondel et al., 2020))**.** Let $\Omega : \mathbb{R} \to \overline{\mathbb{R}}$ be a regularizer, $y \in \mathsf{dom}(\Omega)$ be a target, and $\theta \in \mathbb{R} (= \mathsf{dom}(\Omega^\star))$ be a prediction score.

[1] The support of a probability measure ν is the set of points whose any open neighbors have positive measure. Here, the CDF is defined as $F(\theta) = \nu([-\infty, \theta])$.

[2] We break the tie with $\mathsf{arg\,max}_{y \in \mathcal{Y}} \theta y = 0$ for $\theta = 0$, which is not important for subsequent discussion.

The *Fenchel-Young loss* $\ell_\Omega : \mathsf{dom}(\Omega^\star) \times \mathsf{dom}(\Omega) \to \mathbb{R}_{\geq 0}$ generated by Ω is defined as

$$\ell_\Omega(\theta; y) \stackrel{\mathrm{def}}{=} \Omega^\star(\theta) + \Omega(y) - \theta y. \tag{2}$$

Fenchel-Young losses possess several nice properties.

Proposition 1 (Blondel et al. (2020))**.** *Fenchel-Young losses generated by Ω have the following properties.*

(a) Non-negativity. $\ell_\Omega(\theta; y) \geq 0$ holds for any $\theta \in \mathsf{dom}(\Omega^\star) = \mathbb{R}$ and $y \in \mathcal{Y} \subseteq \mathsf{dom}(\Omega)$.

(b) Zero loss. If Ω is a lower semi-continuous proper convex function, then $\min_\theta \ell_\Omega(\theta; y) = 0$, and $\ell_\Omega(\theta; y) = 0$ iff $y \in \partial\Omega^\star(\theta)$. If Ω is strictly convex, then $\ell_\Omega(\theta; y) = 0$ iff $y = \widehat{y}_\Omega(\theta) = \nabla\Omega^\star(\theta) = \mathsf{arg\,min}_{\theta \in \mathbb{R}} \ell_\Omega(\theta; y)$.

(c) Convexity. $\ell_\Omega(\theta; y)$ is convex in θ.

(d) If Ω is strictly convex, $\ell_\Omega(\theta; y)$ is differentiable in θ and $\nabla_\theta \ell_\Omega(\theta; y) = \widehat{y}_\Omega(\theta) - y$. If Ω is strongly convex, ℓ_Ω is smooth.

For example, the logistic loss $\ell_\Omega(\theta; y) = -\theta y + \log(1 + e^\theta)$ is recovered by $\Omega = -\mathsf{H}_\mathrm{S} + \mathbb{1}_{[0,1]}$, while $\Omega = -\mathsf{H}_\mathrm{T} + \mathbb{1}_{[0,1]}$ recovers the modified Huber loss (Zhang, 2004).

The original Fenchel-Young loss framework that we review in this section constructs a loss function given a convex regularizer Ω, or a (negative) entropy.[3] Once we pick an entropy Ω, a loss function ℓ_Ω and a regularized prediction function $\widehat{y}_\Omega$ are immediately induced by convex duality (as seen in (2) and (1), resp.). This perspective is useful to derive loss functions from well-known entropies such as $-\mathsf{H}_\mathrm{S}$ and $-\mathsf{H}_\mathrm{T}$ because we know the derived losses are endowed with differentiability (induced by $-\mathsf{H}_\mathrm{S}$) and output sparsity (induced by $-\mathsf{H}_\mathrm{T}$) (Martins and Astudillo, 2016). However, we do not know what entropy will induce a loss property we want, e.g., robustness to class-imbalance.

3 Yet Another Way: Fenchel-Young Losses from Inverse Link Functions

In this section,[4] we propose an alternative framework to *derive Fenchel-Young losses from inverse link functions*, in order to provide a more intuitive procedure of loss design than the design of entropic regularizers.

Inverse link functions map prediction scores into probability estimates, to which we can grant flexible CPE properties more directly. To make induction viable, we put regularity conditions on inverse link F.

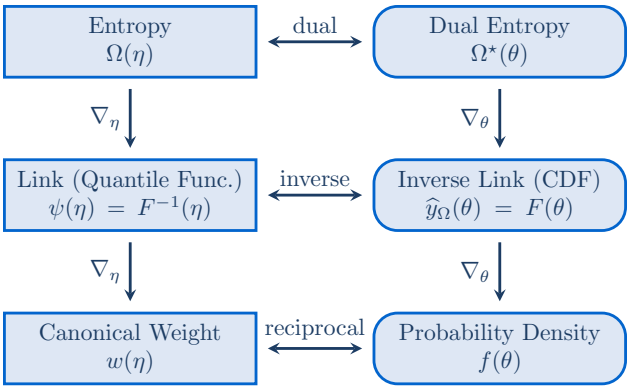

Figure 1: Relationship among entropies, link functions, the CDF, and canonical weight functions. The weight function is described in §5. The reciprocal relationship $w(\eta) \cdot f(\theta) = 1$ is known as the Crouzeix (1977) identity.

Assumption A. Given a non-decreasing function $F : \mathbb{R} \to [0, 1]$, we assume that $\eta \in \mathsf{Im}(F)$ for all $\eta \in (0, 1)$, and that F is strictly increasing over $(\underline{\theta}_F, \overline{\theta}_F)$ for $\underline{\theta}_F \stackrel{\mathrm{def}}{=} \inf\{\theta \mid F(\theta) > 0\}$ and $\overline{\theta}_F \stackrel{\mathrm{def}}{=} \sup\{\theta \mid F(\theta) < 1\}$.

Assumption A is required to ensure that the induced entropy we derive later indeed satisfies convexity. This assumption is satisfied by usual inverse links such as the logistic $F(\theta) = \frac{1}{1+\exp(-\theta)}$ and the probit Φ, i.e., the CDF of the standard normal. In general, the CDF of a regular distribution supported on a connected interval satisfies Assumption A.

Lemma 2. *Given a CDF F, assume that F is continuous and $\mathrm{supp}(F)$ is a convex set. Then, F satisfies Assumption A.*

Lemma 2 justifies our usage of CDFs as inverse links. For F satisfying Assumption A, we define its inverse $F^{-1} : [0, 1] \to \overline{\mathbb{R}}$ as

$$F^{-1}(\eta) \stackrel{\mathrm{def}}{=} \begin{cases} \underline{\theta}_F & \text{if } \eta = 0, \\ \theta \quad \text{s.t. } F(\theta) = \eta & \text{if } \eta \in (0, 1), \\ \overline{\theta}_F & \text{if } \eta = 1. \end{cases} \tag{3}$$

Loss and entropy from inverse link. Figure 1 is an important blueprint, which can be derived from classical convex analysis (Rockafellar, 1970), to connect an inverse link with an entropy. Once we fix an inverse link F, it is identified with the regularized predictor $\widehat{y}_\Omega$. Then, an entropy Ω_F is induced, which is used to derive the Fenchel-Young loss via (2). Assumption A ensures convexity of the induced entropy.

Lemma 3. *Under Assumption A, define $\Omega_F(\eta) \stackrel{\mathrm{def}}{=} \int_0^\eta F^{-1}(q)\mathrm{d}q$. Then, Ω_F is strongly convex over $(0, 1)$.*

[3] In the classical sense, $-\Omega$ is regarded as a (generalized) entropy. We occasionally omit the word "negative" to simply call Ω an entropy when it is clear from the context.

[4] All proofs missing in this section are deferred to §A of the supplementary material.

Lemma 3 immediately follows from the fact that F^{-1} is strictly increasing hence $\nabla(F^{-1}) > 0$. If Ω_F is strongly convex, we know $(F(\theta) =)\ \widehat{y}_{\Omega_F}(\theta) = \nabla\Omega_F^\star(\theta)$ from Proposition 1. Hence, the entropy Ω_F and its dual are well-defined:

$$\Omega_F(\eta) = \int_0^\eta F^{-1}(q)\mathrm{d}q, \quad \Omega_F^\star(\theta) = \int_{-\infty}^\theta F(s)\mathrm{d}s. \quad (4)$$

By substituting the induced entropy and its dual (4) into (2), we can generate a Fenchel-Young loss from the inverse link function.

$$\ell_{\Omega_F}(\theta; y) = \int_{-\infty}^\theta F(s)\mathrm{d}s - y\theta + \Omega_F(1). \quad (5)$$

We drop the subscript F in Ω_F if clear. Subsequently, we scrutinize properties that the loss (5) has.

Bayes risk. We will see the relationship between the Bayes risk and entropy. Given a Fenchel-Young loss ℓ_Ω associated with an inverse link F, the conditional risk of a prediction θ at the true η is denoted by

$$L_\Omega(\theta; \eta) \stackrel{\text{def}}{=} \mathbb{E}_{\mathsf{Y}\sim\eta}[\ell_\Omega(\theta; \mathsf{Y})], \quad (6)$$

where $\mathbb{E}_{\mathsf{Y}\sim\eta}$ means $\mathsf{Y} \sim \mathsf{Bernoulli}(\eta)$, and the pointwise Bayes risk as $\underline{L}_\Omega(\eta) \stackrel{\text{def}}{=} \inf_{\theta\in\mathbb{R}} L(\theta; \eta)$. Since the conditional risk of (5) is[5]

$$\underbrace{L_\Omega(\theta; \eta)}_{\text{cross-entropy}} = \eta\ell_\Omega(\theta; 1) + (1-\eta)\ell_\Omega(\theta; 0)$$
$$= \underbrace{\ell_\Omega(\theta; \eta)}_{\text{Bregman divergence}} \underbrace{-\Omega(\eta) + \eta\Omega(1)}_{\text{entropy}=\underline{L}_\Omega(\eta)},$$

the Bayes risk is $\underline{L}_\Omega(\eta) = -\Omega(\eta) + \eta\Omega(1)$. The linear term $\eta\Omega(1)$ in the Bayes risk is a modifier for skewed entropies, in order to ensure non-negativity of entropies. This fact will be seen in §4 with an example of the GEV link. Note that $\Omega(1) = \Omega(0) = 0$ for the symmetric inverse link thereby the modifier vanishes.

The Bayes risk is attained if and only if $\widehat{y}_\Omega(\theta) = \eta$ from Proposition 1, meaning that the Fenchel-Young loss minimizer is a Fisher consistent estimator of the class-posterior probability (Blondel et al., 2020). $\widehat{y}_\Omega$ is justified as a CPE model in this sense.

Partial separation margin. If a loss entails the separation margin property (Blondel et al., 2020), some finite scores achieve the zero-loss. Hence, the loss will not penalize correct predictions with large enough margins. Here, we introduce a notion *partial separation margin property* useful for the imbalanced case.

[5]We can generalize the formula (KL-divergence) = (cross-entropy) − (entropy) to general Bregman divergences (Nielsen and Nock, 2010). Eventually, we can identify the conditional/Bayes risk as (generalized) cross-entropy/entropy, respectively.

Definition 3 (Partial separation margin (PSM))**.** A binary loss $\ell(\theta; y)$ $(\theta \in \mathbb{R},\ y \in \{0, 1\})$ is said to have the (negative) *partial separation margin* if it satisfies

(a) $\exists m > 0$ s.t. $-\theta \geq m \implies \ell(\theta; 0) = 0$, and
(b) $\forall m > 0, \exists\theta > m$ and $\ell(\theta; 1) > 0$.

The smallest m for (a) is called *(negative) margin*.

In other words, a loss with PSM always penalizes predictions for the positive (rare) class more heavily than the negative (majority) class. Next, we characterize necessary and sufficient conditions for PSM.

Proposition 4. *Let ℓ_Ω be a Fenchel-Young loss with a convex regularizer Ω induced from the CDF F by (4). The following statements are equivalent.*

(a) $\forall\eta \in [0, 1)$, $\partial\Omega(\eta) \neq \emptyset$ *and* $\partial\Omega(1) = \emptyset$.
(b) $F(\mathbb{R}) = [0, 1)$.
(c) ℓ_Ω *has the partial separation margin property.*

Proposition 4 (a) claims that a *partially sparse* entropy such that it does not have gradients only at the right end is required for PSM. Since this does not hold for any symmetric entropies such as $-\mathsf{H}_\mathrm{S}$ and $-\mathsf{H}_\mathrm{T}$, we need skewed entropies. To grant PSM to the induced loss, Proposition 4 (b) supports that a probability distribution with the support bounded at the left end is sufficient. An example is the induced entropy from the GEV link, which we will see in §4.

We also have a closed form of the negative margin.

Proposition 5. *Let ℓ_Ω be a Fenchel-Young loss induced from a link with CDF F. If ℓ_Ω has PSM, then its negative margin is $-\inf\mathrm{supp}(F)$.*

To sum up this section, we provide a new recipe for Fenchel-Young losses from inverse links to avoid designing entropies. We will see an example in §4.

4 Fenchel-Young Loss from GEV Link

The GEV distributions have been widely used in modeling binary responses of rare events (Kotz and Nadarajah, 2000). It is more advantageous as a link than the others because of the simplicity (with only a single shape parameter), flexibility to cover a wide enough range of skewness, and identifiability (Wang and Dey, 2010). In this section, we derive a convex binary CPE loss associated with the GEV distribution to mitigate the influence of class-imbalance.

The CDF of the GEV distribution with shape parameter $\xi \in \mathbb{R}$ is defined as $F_\xi(\theta) = \exp(-[1+\xi\theta]_+^{-1/\xi})$. It is supported on $[-1/\xi, \infty)$ for $\xi > 0$, $(-\infty, -1/\xi]$ for $\xi < 0$,

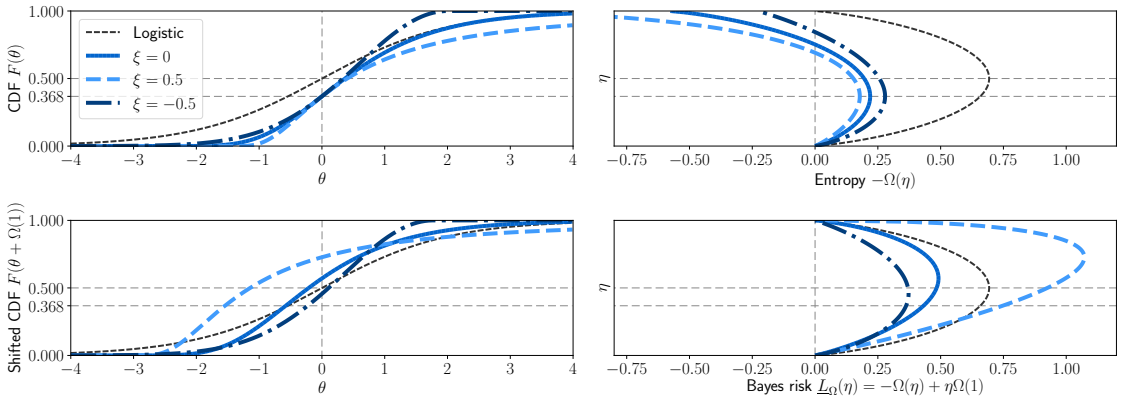

Figure 2: Illustrations of the GEV CDFs and induced entropies. The logistic loss and the corresponding Shannon entropy are plotted together for comparison. Please note that the Bayes risk is non-negative (right-bottom) while the induced entropy is not (right-top).

and $\mathbb{R}$ for $\xi = 0$. The standard Gumbel distribution is recovered when $\xi \to 0$: $F_0(\theta) = \exp(-\exp(-\theta))$. See Figure 2 (left-top).

Wang and Dey (2010) applied Bayesian inference to estimate model parameters and the GEV link parameters. Agarwal et al. (2014) derived the canonical proper (hence convex) loss corresponding to the GEV link. Both suffer from heavy computation costs because the former is nonconvex for $\xi \notin [-1, 0.1)$ while the canonical loss provided by the latter is not defined over the entire $\mathbb{R}$ therefore predictions must be clipped.

We derive a Fenchel-Young loss with the GEV link, called the GEV-Fenchel-Young loss, as a convex binary CPE loss. First, we derive the entropy Ω and its dual $\Omega^\star$ from the GEV CDF F_ξ using (4), which are shown in Table 1.[6] Then, we can obtain the GEV-Fenchel-Young loss via (5). The derived loss is plotted in Figure 3. Here, $\mathsf{Ei}(x)$ and $\Gamma(a, x)$ are the exponential integral function and the (upper) incomplete gamma function defined as follows (for $x \neq 0$).[7]

$$\mathsf{Ei}(x) \stackrel{\mathrm{def}}{=} -\int_{-x}^{\infty} \frac{e^{-t}}{t} \mathrm{d}t, \quad \Gamma(a, x) \stackrel{\mathrm{def}}{=} \int_{x}^{\infty} t^{a-1} e^{-t} \mathrm{d}t.$$

$\Gamma(a) \stackrel{\mathrm{def}}{=} \Gamma(a, 0)$ is the (complete) gamma function. Since $\Gamma(1-\xi, -\log\eta)$ is not finite at $\eta = 1$ when $\xi \geq 1$, Ω can have finite values only when $\xi < 1$; that is, the GEV-Fenchel-Young loss is only defined for $\xi < 1$.

Bayes risk. We investigate the conditional risk minimizer of the GEV-Fenchel-Young loss. This brings us better understanding of how the loss and its minimizer behaves, leading to intuitive design of the loss.

Since the GEV-Fenchel-Young loss is minimized at $F_\xi(\theta) = \eta$, the conditional risk minimizer is $\theta^*(\eta) = F_\xi^{-1}(\eta)$. The explicit form of the minimizer is given in Table 1. Note that $F_\xi(0) = 1/e$ for any $\xi \in \mathbb{R}$ (see left-top of Figure 2). The risk minimizer $\theta^*(\eta)$ is consistent with respect to $\mathrm{sgn}(\eta - 1/e) \approx \mathrm{sgn}(\eta - 0.368)$, instead of the Bayes rule $\mathrm{sgn}(\eta - 1/2)$. This is another evidence that the GEV-Fenchel-Young loss focuses the rare class. We may also obtain the loss consistent to the usual Bayes rule by shifting F_ξ.

The corresponding Bayes risk is

$$\underline{L}_\Omega(\eta) = \begin{cases} -\mathsf{Ei}(\log\eta) + \eta\log(-\log\eta) + \gamma\eta & \text{if } \xi = 0, \\ \frac{1}{\xi}(\Gamma(1-\xi)\eta - \Gamma(1-\xi, -\log\eta)) & \text{if } \xi \neq 0, \end{cases}$$

where γ is the Euler's constant.[8] The Bayes risk is plotted in Figure 2 (right-bottom). We see that the modifier $+\eta\Omega(1)$ ensures non-negativity of the Bayes risk $\underline{L}_\Omega(\eta) \geq 0$, unlike the induced entropy $-\Omega(\eta)$.

Partial separation margin. From Proposition 4 (b), it is sufficient to confirm PSM by checking the support of F_ξ. Since the GEV distribution is supported on $[-1/\xi, \infty)$ with $0 < \xi < 1$, we conclude that the GEV-Fenchel-Young loss has PSM then and penalizes predictions of the rare class more. Proposition 5 ensures that the negative margin is $-\inf \mathrm{supp}(F) = 1/\xi$.

In summary, we derived the Fenchel-Young loss associated with the GEV link as a viable example of loss design from an inverse link provided in §3. The GEV-Fenchel-Young loss superior in terms of the conditional risk minimizer and partial separation margin.

[6] It is derived in §B.1 of the supplementary material.

[7] To compute the incomplete gamma function $\Gamma(a, x)$ for $a < 0$, it is convenient to use the recurrent formula $\Gamma(a+1, x) = a\Gamma(a, x) + x^a e^{-x}$ and $\Gamma(0, x) = -\mathsf{Ei}(-x)$ (Abramowitz and Stegun, 1948).

[8] To obtain the expression for $\xi = 0$, we need to know the value $\Omega(1)$, which can be evaluated as $\Omega(1) = -\Gamma'(1) = \gamma$ by using the L'Hôpital's rule (Apostol, 1991). Note that the Euler's constant $\gamma(\approx 0.577)$ is the negative of the digamma function at 1 (Abramowitz and Stegun, 1948).

Table 1: The entropy and its dual of the GEV CDF, and the class-conditional risk minimizer of the GEV-Fenchel-Young loss. Note that Ω is defined over the entire $[0,1]$ only for $\xi < 1$ (see §4).

ξ	$\Omega(\eta)$	$\Omega^\star(\theta)$	$\theta^*(\eta)$
$\xi < 0$	$\frac{1}{\xi}\left(\Gamma(1-\xi, -\log\eta) - \eta\right)$	$\begin{cases}\Gamma(-\xi, (1+\xi\theta)^{-1/\xi}) & \text{if } \theta \le -1/\xi,\\ \theta + \Gamma(-\xi) + \xi^{-1} & \text{if } \theta > -1/\xi.\end{cases}$	$\begin{cases}\frac{1}{\xi}\left(\frac{1}{(-\ln\eta)^\xi} - 1\right) & \text{if } \eta \in [0,1),\\ -1/\xi & \text{if } \eta = 1.\end{cases}$
$\xi = 0$	$\mathsf{Ei}(\log\eta) - \eta\log(-\log\eta)$	$-\mathsf{Ei}(-e^{-\theta})$	$-\log(-\log\eta)$
$0 < \xi < 1$	$\frac{1}{\xi}\left(\Gamma(1-\xi, -\log\eta) - \eta\right)$	$\begin{cases}\Gamma(-\xi, (1+\xi\theta)^{-1/\xi}) & \text{if } \theta \le -1/\xi,\\ \theta + \Gamma(-\xi) + \xi^{-1} & \text{if } \theta > -1/\xi.\end{cases}$	$\begin{cases}\frac{1}{\xi}\left(\frac{1}{(-\ln\eta)^\xi} - 1\right) & \text{if } \eta \in (0,1],\\ -1/\xi & \text{if } \eta = 0.\end{cases}$

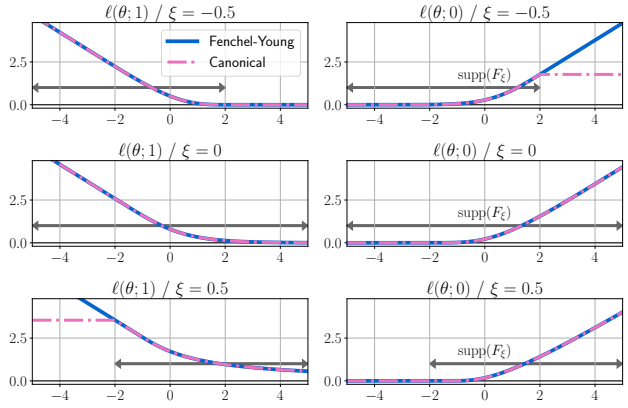

Figure 3: The GEV-Fenchel-Young and GEV-canonical losses. When $\xi = 0$, both losses match exactly. When $\xi \neq 0$, the positive (negative, resp.) partial loss differs for positive (negative, resp.) ξ. In those cases, the canonical losses are nonconvex for $\theta \notin \mathrm{supp}(F_\xi)$, while the Fenchel-Young losses smoothly extrapolate with linear lines. It is also interesting to mention that $\ell(\theta; 1)$ ($\xi = -0.5$) and $\ell(\theta; 0)$ ($\xi = 0.5$) take values 0 for $\theta \notin \mathrm{supp}(F_\xi)$, which is characterized by the separation margin (§3).

5 Relation to Canonical Proper Losses

Canonical proper losses are another CPE loss class. Agarwal et al. (2014) tried to derive convex GEV losses via canonical proper losses. Yet they are not always convex unlike the GEV-Fenchel-Young loss. Since both are closely related, we scrutinize their difference to see why canonical proper losses sometimes result in nonconvex losses, seeking insight for loss design.

Proper loss. We first introduce proper losses (Buja et al., 2005): $\ell(\widehat{\eta}; y)$ penalizes a probability estimate $\widehat{\eta} \in [0,1]$ given a label $y \in \mathcal{Y}$. The conditional risk of prediction $\widehat{\eta} \in [0,1]$ at ground-truth $\eta \in [0,1]$ is denoted as[9]

$$L(\widehat{\eta}; \eta) \stackrel{\mathrm{def}}{=} \mathbb{E}_{\mathsf{Y}\sim\eta}[\ell(\widehat{\eta}; \mathsf{Y})] = \eta\ell(\widehat{\eta}; 1) + (1-\eta)\ell(\widehat{\eta}; 0). \quad (7)$$

[9]Note that the conditional risk of proper losses (7) is defined in a slightly different way from Fenchel-Young loss (6) in that they measure closeness of either $\widehat{\eta} \in [0,1]$ (CPE) or $\theta \in \mathbb{R}$ (prediction score) to true η, respectively.

The pointwise Bayes risk is denoted as $\underline{L}(\eta) \stackrel{\mathrm{def}}{=} \inf_{\widehat{\eta}\in[0,1]} L(\widehat{\eta}; \eta)$. Losses ℓ are said to be *(strictly) proper*—equivalent to Fisher consistency—if $L(\widehat{\eta}; \eta)$ is minimized uniquely by $\widehat{\eta} = \eta$ for all $\widehat{\eta} \in [0,1]$.

Although partial losses $\ell(\widehat{\eta}; 1)$ and $\ell(\widehat{\eta}; 0)$ can be asymmetric unlike the log loss (Scott, 2012), they can be characterized by single *weight functions*.

Theorem 6 (Shuford et al. (1966))**.** *Suppose that $\ell(\widehat{\eta}; \eta)$ is differentiable in $\widehat{\eta}$. Then, ℓ is proper if and only if for all $\widehat{\eta} \in (0,1)$,*

$$\frac{-\nabla_{\widehat{\eta}}\ell(\widehat{\eta}; 1)}{1-\widehat{\eta}} = \frac{\nabla_{\widehat{\eta}}\ell(\widehat{\eta}; 0)}{\widehat{\eta}} = w(\widehat{\eta}) \quad (8)$$

for some weight function $w : (0,1) \to \mathbb{R}_{\ge 0}$ *such that* $\int_\varepsilon^{1-\varepsilon} w(q)\mathrm{d}q < \infty$ *for all* $\varepsilon > 0$.

Corollary 7 (Reid and Williamson (2010))**.** *Let ℓ be a twice differentiable proper loss with a weight function w defined in (8). Then, for all $\eta \in (0,1)$, its pointwise Bayes risk $\underline{L}$ satisfies $w(\eta) = -\nabla^2\underline{L}(\eta)$.*

Canonical composite loss. Proper losses are usually used with a link function $\psi : [0,1] \to \mathbb{R}$, which connects a probability to a real-valued score. Given a proper loss ℓ and an invertible link ψ, a loss on real-valued scores $\ell_\psi(\theta; y) \stackrel{\mathrm{def}}{=} \ell(\psi^{-1}(\theta); y)$ is defined, called a *composite loss*. Buja et al. (2005) introduced a *canonical link* defined by $\nabla\psi = w$. Composite proper losses $\ell_\psi(\theta; y)$ with canonical links $\nabla\psi = w$ are always convex in $\theta \in \mathsf{Im}(\psi)$ (Reid and Williamson, 2010).

Composite losses ℓ_ψ cannot be defined for $\theta \notin \mathsf{Im}(\psi)$ since they assume that ψ is invertible. We compromise therein by clipping ψ such that $\ell_\psi(\theta; y) \stackrel{\mathrm{def}}{=} \ell(1; y)$ for $\theta > \sup\mathsf{Im}(\psi)$ and $\ell_\psi(\theta; y) \stackrel{\mathrm{def}}{=} \ell(0; y)$ for $\theta < \inf\mathsf{Im}(\psi)$.

Relationship to Bregman divergence. Canonical proper composite losses resemble Fenchel-Young losses in that they provide a CPE model via ψ^{-1} and $\widehat{y}_\Omega$, respectively. Here, we compare a canonical proper composite loss ℓ_ψ (with a loss ℓ and link ψ), and a Fenchel-Young loss ℓ_{Ω_F} associated with a CDF F.

Table 2: Simulation results with two normals. The RMSE to the true $\eta(\boldsymbol{x})$ is reported (lower is better).

n	π	GEV-FY	Log
3,000	0.005	$\mathbf{2.22 \times 10^{-2}}$	3.14×10^{-2}
3,000	0.001	$\mathbf{2.09 \times 10^{-2}}$	2.48×10^{-2}
10,000	0.001	$\mathbf{1.16 \times 10^{-2}}$	1.39×10^{-2}

The keystone is the Bregman divergence (Bregman, 1967), defined with a convex generator φ as

$$B_\varphi(z_1 \| z_0) \stackrel{\mathrm{def}}{=} \varphi(z_1) - \varphi(z_0) - \nabla\varphi(z_0)(z_1 - z_0).$$

Both ℓ_ψ and ℓ_{Ω_F} have relation to the Bregman divergence (Bregman, 1967). Indeed, we see that the pointwise regret of the proper loss ℓ is the Bregman divergence $L(\widehat{\eta}; \eta) - \underline{L}(\eta) = B_{-\underline{L}}(\eta \| \widehat{\eta})$.[10] In contrast, the pointwise regret of the Fenchel-Young loss $\ell_{\Omega_F}(\theta; \eta)$ is an upper bound of $B_{-\underline{L}}(\eta \| \widehat{\eta})$. This gap is filled if and only if for $\theta \in \mathsf{Im}(\psi) = \mathrm{supp}(F)$. If the associated CDF F has the bounded support, the gap persists for $\theta \notin \mathbb{R} \setminus \mathrm{supp}(F)$. This is the source of nonconvexity of canonical composite losses. Figure 3 illustrates it with the GEV link. The detailed discussion is deferred to §C of the supplementary material.

Comparison of GEV losses. The GEV-canonical loss $\ell_{F_\xi^{-1}}(\theta; y)$, the proper composite loss with the link F_ξ^{-1} and the weight $w = \nabla(F_\xi^{-1})$, is derived via Theorem 6, as shown in §B.2 of the supplementary material.

Figure 3 shows the GEV-Fenchel-Young loss and GEV-canonical loss for different values of ξ. The GEV-canonical loss is essentially defined over $\mathrm{supp}(F_\xi)$, and the loss value outside the domain is clipped, which induces nonconvexity. On the other hand, the GEV-Fenchel-Young loss is defined over the entire $\mathbb{R}$ and convex by extrapolating the canonical loss with the linear line for $\theta \notin \mathrm{supp}(F_\xi)$. Needless to say, it is beneficial from the computational perspective to avoid nonconvexity via such an extrapolation. Despite that the extrapolation is not unique, the Fenchel-Young loss provides a systematic way to extrapolate.

6 Experiments

We evaluate CPE with the linear model. The implementation is available at https://github.com/levelfour/GEV_Fenchel_Young_Loss.

Synthetic dataset. To compare the logistic regression and GEV-Fenchel-Young loss with different $\pi \stackrel{\mathrm{def}}{=} \mathbb{P}(\mathsf{Y} = 1)$ and sample sizes n, we first use a synthetic dataset with two one-dimensional normals: $\mathsf{X}|\mathsf{Y} = 1 \sim \mathcal{N}(+1, 0.4)$ and $\mathsf{X}|\mathsf{Y} = 0 \sim \mathcal{N}(-1, 0.4)$. 3,000 training data are generated with different π and n. Losses are optimized for 100 epochs with Adam (Kingma and Ba, 2015) and learning rate 1.0 and without regularization. In this simulation, ξ is fixed with $\xi = 0.5$ for the GEV link. Since the true CPE model is known here, the RMSE is reported.

The quantitative results are shown in Table 2. We see GEV-Fenchel-Young loss consistently outperforms. The estimated class-probability curves are shown in Figure 4. Notably, the bias of the GEV curve under $\pi = 0.001$ almost vanishes with $n = 10{,}000$ while the logistic curve is still biased. When comparing the left and middle, it is confirmed that the heavier imbalance makes both biased, but the GEV curve is less affected.

Benchmark results. We use UCI datasets (Dua and Graff, 2017) to run benchmarks. Datasets are divided into training and test sets with the ratio 4 to 1 randomly, and we report average results over 10 random splits. Since we do not have true class probabilities, Brier score $\widehat{\mathsf{BS}}$ (Brier, 1950) and stratified Brier score $\widehat{\mathsf{sBS}}$ (Wallace and Dahabreh, 2012)[11]

$$\widehat{\mathsf{BS}}(\widehat{\eta}) = \frac{1}{n}\sum_{i=1}^{n} \left(\widehat{\eta}(\boldsymbol{x}_i) - [\![y_i = 1]\!]\right)^2,$$

$$\widehat{\mathsf{sBS}}(\widehat{\eta}) = \frac{1}{2n_1}\sum_{i:y_i=1} \left(\widehat{\eta}(\boldsymbol{x}_i) - 1\right)^2 + \frac{1}{2n_0}\sum_{i:y_i=0} \widehat{\eta}(\boldsymbol{x}_i)^2$$

are used to evaluate a CPE model $\widehat{\eta}$, where n_y is the numbers of data with $y_i = y$. The stratified one tends to focus on the minority, while the vanilla one treats the entire range $[0, 1]$ equally. Due to this, the stratified one is used as the validation score. The vanilla one is used as the test score because we want good class probability estimates for the entire $[0, 1]$ even under class-imbalance. We compare the proposed GEV-Fenchel-Young loss (**GEV-FY**) with the following baselines: GEV-canonical loss (**GEV-Can**) (Agarwal et al., 2014), GEV-log loss (**GEV-Log**) (Wang and Dey, 2010), logistic regression (**Log**), Platt's scaling (**Platt**) (Platt, 1999), probability calibration with isotonic regression (**Isotonic**) (Menon et al., 2012), balanced logistic regression (**Weight**), and undersampling with bagging (**Bagging**) (Wallace and Dahabreh, 2012). More details are described in §D of the supplementary material.

The results of vanilla Brier score are shown in Table 3.

[10] Due to $L(\widehat{\eta}; \eta) = \underline{L}(\widehat{\eta}) + \underline{L}'(\widehat{\eta})(\eta - \widehat{\eta})$ (Savage, 1971).

[11] The choice of the evaluation metric is an arguably crucial. We used the Brier score since it is commonly used (King and Zeng, 2001; Wallace and Dahabreh, 2012), and different from the associated score with the log and GEV loss, which is considered to be fair.

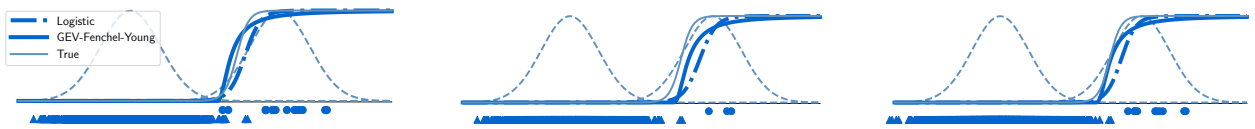

Figure 4: Simulation with two normals: $(n, \pi) = \{$ left: $(3{,}000, 0.005)$, middle: $(3{,}000, 0.001)$, right: $(10{,}000, 0.001)\}$.

Table 3: Benchmark results on UCI datasets in Brier score (lower is better). Bold faces indicate the best and statistically insignificantly different methods from the best in each row, using one-sided t-test with 95% confidence. $\pi = \mathbb{P}(\mathsf{Y} = 1)$.

Dataset	n	π	GEV-FY	GEV-Can	GEV-Log	Log	Platt	Isotonic	Weight	Bagging
car	1728	0.038	**0.0162**	0.0166	**0.0154**	**0.0155**	0.0277	**0.0138**	0.0604	0.0501
ecoli	336	0.107	**0.0601**	**0.0567**	**0.0549**	0.2201	0.0743	0.0867	0.0973	0.0801
glass	214	0.079	**0.0845**	**0.2468**	0.2694	**0.0710**	**0.0711**	**0.0691**	0.2084	0.2382
haberman	306	0.265	**0.1882**	**0.2749**	0.4129	**0.1954**	**0.1901**	**0.1863**	0.2314	0.2293
nursery	12960	0.025	**0.0136**	**0.0137**	0.0134	**0.0133**	**0.0131**	**0.0131**	0.0489	0.0493
yeast	1484	0.289	**0.1630**	**0.1632**	**0.1648**	**0.1656**	0.1665	**0.1598**	0.1903	0.1891

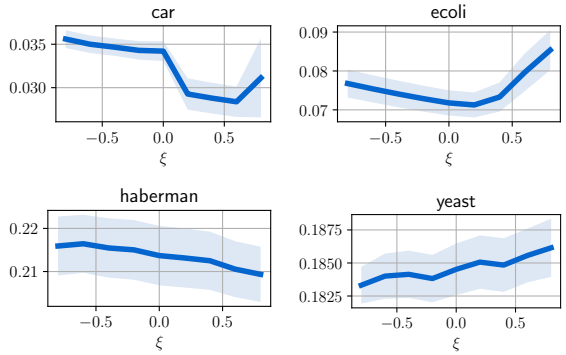

Figure 5: Sensitivity to shape parameter ξ. The lines show means of Brier score with standard errors of 50 runs.

Overall, we see that GEV-FY outperforms the other baselines, or performs at least comparably.

Sensitivity of shape parameter. We simulate the GEV-Fenchel-Young loss to see sensitivity to the shape parameter ξ. From §4, we expect better performance with positive ξ when the positive class is rare. We ran 50 trials of random splitting of the datasets with the same setting as the previous benchmarks except that ξ is fixed to $\{-0.8, -0.6, \ldots, 0.8\}$. The results are shown in Figure 5, from which we confirm that $\xi > 0$ indeed performs better with car and ecoli. The results with haberman does not seem significantly better, and yeast seems slightly worse with $\xi > 0$. We conjecture this is because π is not extremely small (haberman: 0.265, yeast: 0.289) comparing with the former two (car: 0.040, ecoli: 0.104). Hence, the PSM is thought to be useful for heavily-imbalanced data.

F-measure optimization. As an application of CPE, we simulate F-measure maximization. We use the plug-in approach (Koyejo et al., 2014): a CPE model is trained first, then the best threshold is sought for. We use 70% of training data to fit CPE models and 30% for the thresholds. Thresholds are picked from $\{0.05, 0.10, \ldots, 0.95\}$. All the other setting are the same as the benchmarks. The results are shown in Table 4. We can see that the proposed method (GEV-FY) performs the best with all datasets, showing the importance of better CPE under class-imbalance.

Table 4: F-measure optimization results on UCI datasets (higher is better). Bold faces are the best and statistically insignificantly different methods from the best in each row, using one-sided t-test with 95% confidence of 10 runs.

	GEV-FY	Log	Isotonic
car	**0.6455**	**0.5511**	**0.5564**
ecoli	**0.6504**	0.2475	0.3447
glass	**0.1345**	**0.1591**	**0.1359**
haberman	**0.3963**	0.3600	**0.4329**
nursery	**0.6037**	**0.6027**	**0.5962**
yeast	**0.5765**	**0.5615**	**0.5691**

7 Conclusion

We studied binary CPE under class-imbalance. The existing approaches have adopted skewed links to avoid the misspecification issue of the logit link, yet the optimization was nonconvex. We utilized Fenchel-Young losses to derive convex losses and confirmed its effectiveness by deriving a loss from the GEV CDF. Technically, this viewpoint is interesting in that we can design a convex loss based on an inverse link. This is in stark contrast to the previous roadmap requiring an entropic regularizer, which is often not intuitive to design. The simulation revealed that the GEV-Fenchel-Young loss is more robust to class-imbalance and suitable for applications such as F-measure maximization.

Acknowledgements

HB was supported by JSPS KAKENHI Grant Number 19J21094. MS was supported by JST CREST Grant Number JPMJCR18A2.

References

Abramowitz, M. and Stegun, I. A. (1948). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. US Government Printing Office.

Agarwal, A., Narasimhan, H., Kalyanakrishnan, S., and Agarwal, S. (2014). GEV-canonical regression for accurate binary class probability estimation when one class is rare. In *Proceedings of the 31th International Conference on Machine Learning*, pages 1989–1997.

Apostol, T. M. (1991). *Calculus, Volume 1*. John Wiley & Sons.

Bao, H. and Sugiyama, M. (2020). Calibrated surrogate maximization of linear-fractional utility in binary classification. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 2337–2347.

Berman, M., Rannen Triki, A., and Blaschko, M. B. (2018). The Lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4413–4421.

Blondel, M., Martins, A. F., and Niculae, V. (2020). Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 21(35):1–69.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217.

Breslow, N. E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association*, 91(433):14–28.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Meather Review*, 78(1):1–3.

Buja, A., Stuetzle, W., and Shen, Y. (2005). Loss functions for binary class probability estimation and classification: Structure and applications.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1567–1578.

Charoenphakdee, N., Vongkulbhisal, J., Chairatanakul, N., and Sugiyama, M. (2020). On focal loss for class-posterior probability estimation: A theoretical perspective. *arXiv preprint arXiv:2011.09172*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Chen, M.-H., Dey, D. K., and Shao, Q.-M. (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, 94(448):1172–1186.

Crouzeix, J.-P. (1977). A relationship between the second derivatives of a convex function and of its conjugate. *Mathematical Programming*, 13(1):364–365.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277.

Czado, C. and Santner, T. J. (1992). The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference*, 33(2):213–231.

Dal Pozzolo, A., Caelen, O., and Bontempi, G. (2015). When is undersampling effective in unbalanced classification tasks? In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 200–215. Springer.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Eban, E., Schain, M., Mackey, A., Gordon, A., Rifkin, R., and Elidan, G. (2017). Scalable learning of non-decomposable objectives. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 832–840.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pages 973–978.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38.

Jiang, X., Dey, D. K., Prunier, R., Wilson, A. M., and Holsinger, K. E. (2013). A new class of flexible link functions with application to species co-occurrence in cape floristic region. *The Annals of Applied Statistics*, pages 2180–2204.

Joachims, T. (2005). A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 377–384.

Kim, S., Chen, M.-H., and Dey, D. K. (2008). Flexible generalized t-link models for binary response data. *Biometrika*, 95(1):93–106.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9(2):137–163.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Kotz, S. and Nadarajah, S. (2000). *Extreme Value Distributions: Theory and Applications*. World Scientific.

Koyejo, O. O., Natarajan, N., Ravikumar, P. K., and Dhillon, I. S. (2014). Consistent binary classification with generalized performance metrics. In *Advances in Neural Information Processing Systems*, pages 2744–2752.

Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., and Li, J. (2020). Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

Martins, A. and Astudillo, R. (2016). From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1614–1623.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, volume 37. CRC Press.

Menon, A. K., Jiang, X. J., Vembu, S., Elkan, C., and Ohno-Machado, L. (2012). Predicting accurate probabilities with a ranking loss. In *Proceedings of the 29th International Conference on Machine Learning*, pages 703–710.

Mey, A. and Loog, M. (2021). Consistency and finite sample behavior of binary class probability estimation. In *AAAI Conference on Artificial Intelligence*. to appear.

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 4th international conference on 3D vision (3DV)*, pages 565–571. IEEE.

Niculae, V., Martins, A., Blondel, M., and Cardie, C. (2018). SparseMAP: Differentiable sparse structured inference. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3799–3808.

Nielsen, F. and Nock, R. (2010). Entropies and cross-entropies of exponential families. In *International Conference on Image Processing*, pages 3621–3624. IEEE.

Nordström, M., Bao, H., Löfman, F., Hult, H., Maki, A., and Sugiyama, M. (2020). Calibrated surrogate maximization of Dice. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 269–278. Springer.

Nowak-Vila, A., Bach, F., and Rudi, A. (2020). Consistent structured prediction with max-min margin Markov networks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7381–7391.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74.

Reid, M. D. and Williamson, R. C. (2010). Composite binary losses. *Journal of Machine Learning Research*, 11(83):2387–2422.

Rockafellar, R. T. (1970). *Convex Analysis*, volume 36. Princeton University Press.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801.

Scott, C. (2012). Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6:958–992.

Shuford, E. H., Albert, A., and Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika*, 31(2):125–145.

Stukel, T. A. (1988). Generalized logistic models. *Journal of the American Statistical Association*, 83(402):426–431.

Telgarsky, M., Dudik, M., and Schapire, R. (2015). Convex risk minimization and conditional probability estimation. In *Conference on Learning Theory*, pages 1629–1682. PMLR.

Wallace, B. C. and Dahabreh, I. J. (2012). Class probability estimates are unreliable for imbalanced data (and how to fix them). In *IEEE 12th International Conference on Data Mining*, pages 695–704. IEEE.

Wang, X. and Dey, D. K. (2010). Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption. *The Annals of Applied Statistics*, 4(4):2000–2023.

Ye, N., Chai, K. M. A., Lee, W. S., and Chieu, H. L. (2012). Optimizing F-measures: a tale of two approaches. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1555–1562.

Yu, J. and Blaschko, M. (2015). Learning submodular losses with the Lovász hinge. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1623–1631.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–85.