Fenchel-Young Losses with Skewed Entropies for Class-posterior Probability Estimation (Supplementary Material)

A Proofs

A.1 Proof of Lemma 2

Proof. It is easy to show $\eta \in \mathsf{Im}(F)$ for all $\eta \in (0,1)$. Indeed, this is an immediate corollary of the intermediate value theorem (Apostol, 1991) over an open interval. Note that $F(\theta) \to 1$ as $\theta \to \infty$ and $F(\theta) \to 0$ as $\theta \to -\infty$.

We show F is strictly increasing over $(\underline{\theta}_F, \overline{\theta}_F)$. Fix $\widetilde{\theta} \in (\underline{\theta}_F, \overline{\theta}_F)$. Let $t \in (0,1)$ be a constant such that $\widetilde{\theta} = t\underline{\theta}_F + (1-t)\overline{\theta}_F$. If $\underline{\theta}_F$ is $-\infty$, then we define by $\widetilde{\theta} = t\underline{C}_F + (1-t)\overline{\theta}_F$ for small enough $\underline{C}_F$. Subsequently, we consistently write $\underline{C}_F$ instead of $\underline{\theta}_F$ regardless of the boundedness of $\underline{\theta}_F$. We treat the case where $\overline{\theta}_F$ is ∞ in the same manner by large enough $\overline{C}_F$. Since $\mathrm{supp}(F)$ is convex and $\underline{C}_F, \overline{C}_F \in \mathrm{supp}(F)$, $\widetilde{\theta} \in \mathrm{supp}(F)$. This implies that $F(\widetilde{\theta} + \varepsilon) > F(\widetilde{\theta} - \varepsilon)$ for all $\varepsilon > 0$. It immediately follows that F is strictly increasing over $(\underline{\theta}_F, \overline{\theta}_F)$. □

A.2 Proof of Proposition 4

In order to show Proposition 4, we need the following lemma.

Lemma 8. *ℓ_Ω has the PSM if and only if there exists $m > 0$ such that $-m \in \partial\Omega(0)$ and $\partial\Omega(1) = \emptyset$.*

Proof of Lemma 8. The proof is a simple corollary of Blondel et al. (2020, Proposition 5). We provide the proof here for the completeness.

[⇒] If ℓ_Ω has the PSM, there exists $m > 0$ such that $-\theta \geq m$ implies $\ell_\Omega(\theta; 0) = 0$ (part (a)). Let us fix $\theta = -m$, then $\ell_\Omega(\theta; 0) = 0$. By Proposition 1, we have $\theta \in \partial\Omega(0)$. On the other hand, if ℓ_Ω has the PSM, for all $m > 0$, there exists $\theta > m$ such that $\ell(\theta; 1) > 0$ (part (b)). By Proposition 1, we have $\theta \notin \partial\Omega(1)$ for such θ. If $\ell_\Omega(\cdot; 1)$ is non-increasing, then we know $\ell(m; 1) > \ell(\theta; 1)$ implying that $m \notin \partial\Omega(1)$. We can confirm $\ell_\Omega(\cdot; 1)$ is non-increasing because $\nabla_\theta \ell_\Omega(\theta; 1) = F(\theta) - 1 \leq 0$, which holds from Assumption A and Proposition 1. Hence, it holds that $m \notin \partial\Omega(1)$ for all $m > 0$, which is equivalent to $\partial\Omega(1) = \emptyset$.

[⇐] First, fix $\theta \leq -m$. Since Ω is convex, we have $s \in \partial\Omega(0)$ for all $s \leq -m$, implying that $\theta \in \partial\Omega(0)$. By Proposition 1, $\ell(\theta; 0) = 0$ (part (a)). Next, for a given $m > 0$, fix an arbitrary $\theta > m$. Since $\partial\Omega(1) = \emptyset$, $\theta \notin \partial\Omega(1)$, implying $\ell_\Omega(\theta; 1) > 0$ by Proposition 1 (part (b)). □

We are now ready to show Proposition 4.

Proof of Proposition 4. <u>Part (a) $\implies$ Part (b)</u>: Lemma 3 ensures the differentiability of Ω hence $F = \nabla\Omega^\star$ from Figure 1. <u>Our goal is to show $\nabla\Omega^\star(\mathbb{R}) = [0,1)$.</u> To do so, we need to show

1. <u>For all $\eta \in [0,1)$, there exists $\theta \in \mathbb{R}$ such that $\eta \in \operatorname{arg\,min}_{q \in [0,1]} \Omega(q) - \theta q$.</u> Since the Lagrangian associated this optimization problem is $\mathcal{L}(q, \mu, \lambda) \stackrel{\text{def}}{=} \Omega(q) - \theta q - \mu q + \lambda(q-1)$, where μ and λ are KKT multipliers, the KKT conditions are
   - $0 \in \partial_q \mathcal{L}(\eta, \mu, \lambda) = \partial\Omega(\eta) - \theta - \mu + \lambda$,
   - $0 \leq \eta \leq 1$,
   - $\mu\eta = 0$ and $\mu \geq 0$,
   - $\lambda(\eta - 1) = 0$ and $\lambda \geq 0$.

   By substituting $\mu = \lambda = 0$, we obtain $\theta \in \partial\Omega(\eta)$. Since $\partial\Omega(\eta) \neq \emptyset$ for $\eta \in [0,1)$ is assumed, this θ is a feasible solution. Hence, such θ does exist.

2. For all $\theta \in \mathbb{R}$, $1 \notin \mathsf{arg\,min}_{q \in [0,1]} \Omega(q) - \theta q$. In the same way, we have the KKT conditions such that $0 \in \partial_q \mathcal{L}(1, 0, \lambda) = \partial\Omega(1) - \theta + \lambda$ and $\lambda \geq 0$. Since $\partial\Omega(1) = \emptyset$, such θ does not exist.

Part (b) $\Longrightarrow$ Part (a): The assumption implies $\partial\Omega^\star(\mathbb{R}) = [0, 1)$, meaning that for all $\eta \in [0, 1)$, the following two facts hold.

1. There exists $\theta \in \mathbb{R}$ such that $\eta \in \partial\Omega^\star(\theta)$. By Danskin's theorem, this implies $\theta \in \partial\Omega(\eta)$, which concludes that $\partial\Omega(\eta) \neq \emptyset$.
2. There does not exist $\theta \in \mathbb{R}$ such that $1 \in \partial\Omega^\star(\theta)$. In the same way, this implies that $\partial\Omega(1) = \emptyset$.

Part (a) $\Longrightarrow$ Part (c): By Lemma 8, we need to show that there exists $m > 0$ such that $-m \in \partial\Omega(0)$ and $\partial\Omega(1) = \emptyset$, where the latter immediately follows from the assumption. For the latter, we first prove by contradiction that there exists $\theta \in \partial\Omega(0)$ such that $\theta < 0$. Assume that for all $\eta' \in [0, 1]$ and $g \geq 0$, $\Omega(\eta') \geq \Omega(0) + g\eta'$. It is equivalent to $\Omega(\eta') \geq g\eta'$ because $\Omega(0) = 0$ by the construction of the induced entropy (4). Since we can take arbitrary $g \geq 0$, it implies $\Omega(\eta') = \infty$, contradicting with $\mathsf{dom}(\Omega) = [0, 1]$. Hence, there exists $\theta \in \partial\Omega(0)$ such that $\theta < 0$. By Lemma 8, it is concluded that ℓ_Ω has the PSM.

Part (c) $\Longrightarrow$ Part (a): Our goal is to show that there exists $g \in \mathbb{R}$ such that $\Omega(\eta') \geq \Omega(\eta) + g(\eta' - \eta)$ for all $\eta' \in [0, 1]$ ($\partial\Omega(1) = \emptyset$ immediately follows from the PSM). By Lemma 8, there exists $\theta \in \partial\Omega(0)$ such that $\theta < 0$. To show the above claim by contradiction, assume that for all $g \in \mathbb{R}$, there exists $\eta' \in [0, 1]$ such that $\Omega(\eta') < \Omega(\eta) + g(\eta' - \eta)$. Since $\Omega(\eta') \geq \Omega(0) + \theta(\eta' - 0) = \theta\eta'$ from $\theta \in \partial\Omega(0)$, we have $\Omega(\eta) > g\eta + (\theta - g)\eta'$. Since $\eta' = \eta$ contradicts with the assumption $\Omega(\eta') < \Omega(\eta) + g(\eta' - \eta) = \Omega(\eta)$, we may naturally assume $\eta' \neq \eta$. Then, $\Omega(\eta) > \theta\eta' - g(\eta' - \eta)$ for all $g \in \mathbb{R}$ implies that $\Omega(\eta) = \infty$, contradicting with $\mathsf{dom}(\Omega) = [0, 1]$. Hence, we have $\partial\Omega(\eta) \neq \emptyset$. □

A.3 Proof pf Proposition 5

Proof. By Proposition 4 (part (a)), there exists $\theta \in \mathbb{R}$ such that $\theta \in \partial\Omega(0)$ if ℓ_Ω has PSM. That is, for some $\theta \in \mathbb{R}$,

$$\begin{aligned}
\theta \in \partial\Omega(0) &\iff \Omega(\eta) \geq \Omega(0) + \theta(\eta - 0) && \forall \eta \in [0, 1], \\
&\iff \theta \leq \frac{\Omega(\eta)}{\eta} && \forall \eta \in (0, 1].
\end{aligned}$$

Note that Lemma 3 ensures the differentiability of Ω. The margin is the smallest $-\theta$ satisfying the above, which is

$$\begin{aligned}
\sup_{\eta \in (0,1]} -\frac{\Omega(\eta)}{\eta} &= -\lim_{\eta \searrow 0} \frac{\Omega(\eta)}{\eta} && \triangleleft\ \Omega(\eta)/\eta \text{ is non-decreasing} \\
&= -\lim_{\eta \searrow 0} \frac{\nabla\Omega(\eta)}{1} && \triangleleft\ \text{L'Hôpital's rule} \\
&= -\nabla\Omega(0) \\
&= -F^{-1}(0) \\
&= -\inf \operatorname{supp}(F).
\end{aligned}$$

$\Omega(\eta)/\eta$ is non-decreasing because

$$\begin{aligned}
\frac{\mathrm{d}\left(\frac{\Omega(\eta)}{\eta}\right)}{\mathrm{d}\eta} &= \frac{\eta\nabla\Omega(\eta) - \Omega(\eta)}{\eta^2} \\
&= \frac{\Omega^\star(\nabla\Omega(\eta))}{\eta^2} && \triangleleft\ \text{Danskin's theorem} \\
&\geq 0. && \triangleleft\ F \geq 0
\end{aligned}$$

□

B Derivations

B.1 GEV-Fenchel-Young Loss

Recall that the entropy Ω and its dual $\Omega^\star$ is given as

$$\Omega(\eta) = \int_0^\eta F_\xi^{-1}(q)\mathrm{d}q, \qquad \Omega^\star(\theta) = \int_{-\infty}^\theta F_\xi(s)\mathrm{d}s.$$

In this section, we derive closed-forms of Ω and $\Omega^\star$ depending on the value of ξ.

(Case A) When $\xi = 0$: The CDF and its inverse are

$$F_\xi(\theta) = \exp(-\exp(-\theta)), \qquad F_\xi^{-1}(\eta) = -\log(-\log(\eta)).$$

Then, the dual entropy is calculated as

$$\Omega^\star(\theta) = \int_{-\infty}^\theta \exp(-\exp(-s))\mathrm{d}s = \int_\infty^{e^{-\theta}} -\frac{e^{-t}}{t}\mathrm{d}t = \int_{e^{-\theta}}^\infty \frac{e^{-t}}{t}\mathrm{d}t = \Gamma(0, \exp(-\theta)),$$

where change of the variable $t \stackrel{\text{def}}{=} \exp(-s)$ is applied at the second identity, and the last identity follows from the definition of the upper incomplete gamma function

$$\Gamma(a, x) = \int_x^\infty t^{a-1}e^{-t}\mathrm{d}t.$$

The entropy is calculated as

$$\begin{aligned}
\Omega(\eta) &= \int_0^\eta -\log(-\log q)\mathrm{d}q \\
&= [-q\log(-\log q)]_0^\eta - \int_0^\eta -q \cdot \frac{1}{-\log q} \cdot \frac{-1}{q}\mathrm{d}q && \triangleleft \text{ integral by parts} \\
&= [-q\log(-\log q)]_0^\eta + \int_0^\eta (\log q)^{-1}dq \\
&= -\eta\log(-\log \eta) + \int_0^\eta (\log q)^{-1}dq && \triangleleft\ -q\log(-\log q) \stackrel{q\to 0}{\to} 0 \\
&= -\eta\log(-\log \eta) + \int_\infty^{-\log\eta} \frac{1}{-t} \cdot (-e^{-t})\mathrm{d}t && \triangleleft \text{ change of variable } q \stackrel{\text{def}}{=} e^{-t} \\
&= -\eta\log(-\log \eta) - \int_{-\log\eta}^\infty t^{-1}e^{-t}\mathrm{d}t \\
&= -\eta\log(-\log \eta) - \Gamma(0, -\log\eta) \\
&= -\eta\log(-\log \eta) + \mathsf{Ei}(\log\eta).
\end{aligned}$$

(Case B) When $\xi > 0$: The CDF and its inverse are

$$F_\xi(\theta) = \begin{cases} 0 & \text{if } \theta < -1/\xi, \\ \exp(-(1+\xi\theta)^{-1/\xi}) & \text{if } \theta \geq -1/\xi. \end{cases}, \qquad F_\xi^{-1}(\eta) = \frac{1}{\xi}\left(\frac{1}{(-\log\eta)^\xi} - 1\right).$$

The entropy is calculated as

$$\begin{aligned}
\Omega(\eta) &= \int_0^\eta \frac{1}{\xi}\left(\frac{1}{(-\log q)^\xi} - 1\right)\mathrm{d}q \\
&= \frac{1}{\xi}\int_\infty^{-\log\eta} t^{-\xi} \cdot (-e^{-t})\mathrm{d}t - \frac{1}{\xi}\eta && \triangleleft \text{ change of variable } q \stackrel{\text{def}}{=} e^{-t} \\
&= \frac{1}{\xi}\int_{-\log\eta}^\infty t^{-\xi}e^{-t}\mathrm{d}t - \frac{1}{\xi}\eta \\
&= \frac{1}{\xi}\left\{\Gamma(1-\xi, -\log\eta) - \eta\right\}.
\end{aligned}$$

Note that $\Omega(1)$ is defined only for $0 < \xi < 1$ otherwise it diverges because the recurrence relationship $\Gamma(s+1, x) = s\Gamma(s, x) + x^s e^{-x}$ can be used only when $s > 0$ or $x > 0$.

In order to calculate the dual entropy, we divide the cases. When $\theta \geq -1/\xi$,

$$\begin{aligned}\Omega^\star(\theta) &= \int_{-1/\xi}^{\theta} \exp(-(1+\xi s)^{-1/\xi})\mathrm{d}s \\ &= \int_{\infty}^{(1+\xi\theta)^{-1/\xi}} -\frac{e^{-t}}{t^{1+\xi}}\mathrm{d}t && \triangleleft \text{ change of variable } t \stackrel{\text{def}}{=} (1+\xi s)^{-1/\xi} \\ &= \int_{(1+\xi\theta)^{-1/\xi}}^{\infty} t^{-\xi-1}e^{-t}\mathrm{d}t \\ &= \Gamma(-\xi, (1+\xi\theta)^{-1/\xi}).\end{aligned}$$

When $\theta < -1/\xi$,

$$\Omega^\star(\theta) = \lim_{\theta \searrow -1/\xi} \Gamma(-\xi, (1+\xi\theta)^{-1/\xi}) = 0.$$

(Case C) When $\xi < 0$: The CDF and its inverse are

$$F_\xi(\theta) = \begin{cases} 1 & \text{if } \theta > -1/\xi, \\ \exp(-(1+\xi\theta)^{-1/\xi}) & \text{if } \theta \leq -1/\xi \end{cases}, \qquad F_\xi^{-1}(\eta) = \frac{1}{\xi}\left(\frac{1}{(-\log\eta)^\xi} - 1\right).$$

The entropy is the same as when $\xi > 0$: $\Omega(\eta) = \frac{1}{\xi}\{\Gamma(1-\xi, -\log\eta) - \eta\}$. In order to calculate the dual entropy, we divide the cases. When $\theta \leq -1/\xi$,

$$\begin{aligned}\Omega^\star(\theta) &= \int_{-\infty}^{\theta} \exp(-(1+\xi s)^{-1/\xi})\mathrm{d}s \\ &= \int_{\infty}^{(1+\xi\theta)^{-1/\xi}} -\frac{e^{-t}}{t^{1+\xi}}\mathrm{d}t && \triangleleft \text{ change of variable } t \stackrel{\text{def}}{=} (1+\xi s)^{-1/\xi} \\ &= \int_{(1+\xi\theta)^{-1/\xi}}^{\infty} t^{-\xi-1}e^{-t}\mathrm{d}t \\ &= \Gamma(-\xi, (1+\xi\theta)^{-1/\xi}),\end{aligned}$$

When $\theta > -1/\xi$,

$$\begin{aligned}\Omega^\star(\theta) &= \underbrace{\int_{-\infty}^{-1/\xi} \exp(-(1+\xi s)^{-1/\xi})\mathrm{d}s}_{=\Gamma(-\xi,0)=\Gamma(-\xi)} + \int_{-1/\xi}^{\theta} \mathrm{d}s \\ &= \Gamma(-\xi) + [s]_{-1/\xi}^{\theta} \\ &= \theta + \Gamma(-\xi) + \xi^{-1},\end{aligned}$$

where $\Gamma(a) \stackrel{\text{def}}{=} \Gamma(a, 0)$ is the (complete) gamma function.

B.2 GEV-Canonical Loss

The canonical proper losses associated with the GEV link are shown in Table 5. We will show their derivations subsequently.

As a corollary of Theorem 6, the canonical proper loss with a link ψ given a weight function $w : (0,1) \to \mathbb{R}_{\geq 0}$ such that $w = \nabla\psi$ is given as

$$\ell(\widehat{\eta}; 1) = \int_{\widehat{\eta}}^{1} (1-q)w(q)\mathrm{d}q, \qquad \ell(\widehat{\eta}; 0) = \int_{0}^{\widehat{\eta}} qw(q)\mathrm{d}q.$$

Table 5: GEV canonical proper losses. The explicit form of Ω is provided in Table 1.

ξ	$\ell_{F_\xi^{-1}}(\theta; 1)$	$\ell_{F_\xi^{-1}}(\theta; 0)$
$\xi < 0$	$\Gamma(-\xi, (1+\xi\theta)_+^{-1/\xi}) - \min\{\theta, -1/\xi\} + \Omega(1)$	$\Gamma(-\xi, (1+\xi\theta)_+^{-1/\xi})$
$\xi = 0$	$\Gamma(0, e^{-\theta}) - \theta + \Omega(1)$	$\Gamma(0, e^{-\theta})$
$0 < \xi < 1$	$\Gamma(-\xi, (1+\xi\theta)_+^{-1/\xi}) - \max\{\theta, -1/\xi\} + \Omega(1)$	$\Gamma(-\xi, (1+\xi\theta)_+^{-1/\xi})$

In addition, we have a relationship between a weight function and partial losses such that

$$w(\eta) = \nabla_\eta \ell(\eta; 0) - \nabla_\eta \ell(\eta; 1).$$

In this section, we derive the canonical proper loss and composite loss.

The link function of the GEV-canonical loss is defined as

$$\psi(\eta) = F_\xi^{-1}(\eta) = \begin{cases} -\log(-\log \eta) & \text{if } \xi = 0, \\ \frac{1}{\xi}\left(\frac{1}{(-\log \eta)^\xi} - 1\right) & \text{if } \xi \neq 0, \end{cases}$$

and the associated canonical weight function for $\xi \neq 0$ is

$$w(\eta) = \frac{\mathrm{d}\psi(\eta)}{\mathrm{d}\eta} = \frac{1}{\eta(-\log \eta)^{1+\xi}},$$

which is general enough to cover the case $\xi = 0$, $w(\eta) = -1/(\eta \log \eta)$.

First, we derive the canonical proper loss. The partial loss for $y = 0$ is

$$\begin{aligned} \ell(\widehat{\eta}; 0) &= \int_0^{\widehat{\eta}} \frac{1}{(-\log q)^{1+\xi}} \mathrm{d}q \\ &= \int_\infty^{-\log \widehat{\eta}} \frac{-e^{-t}}{t^{1+\xi}} \mathrm{d}t && \triangleleft \text{ change of variable } t \overset{\text{def}}{=} -\log q \\ &= \int_{-\log \widehat{\eta}}^\infty t^{-\xi-1} e^{-t} \mathrm{d}t \\ &= \Gamma(-\xi, -\log \widehat{\eta}). \end{aligned}$$

The partial loss for $y = 1$ is derived from the relationship $w(\eta) = \nabla_\eta \ell(\eta; 0) - \nabla_\eta \ell(\eta; 1)$. By integrating the both sides, we have $\psi(\eta) = \ell(\eta; 0) - \ell(\eta; 1) + C$, where C is an integration constant. Hence,

$$\begin{aligned} \ell(\widehat{\eta}; 1) &= \ell(\widehat{\eta}; 0) - \psi(\widehat{\eta}) + C \\ &= C + \begin{cases} \Gamma(0, -\log \widehat{\eta}) + \log(-\log \widehat{\eta}) & \text{if } \xi = 0, \\ \Gamma(-\xi, -\log \widehat{\eta}) - \frac{1}{\xi}\left(\frac{1}{(-\log \widehat{\eta})^\xi} - 1\right) & \text{if } \xi \neq 0. \end{cases} \end{aligned}$$

The integration constant C can be determined with the constraint $\ell(1; 1) = \lim_{\eta \nearrow 1} \int_\eta^1 q w(q) \mathrm{d}q = 0$. When $\xi \neq 0$

and $\xi < 1$,

$$\begin{aligned}
C &= -\lim_{\eta \nearrow 1} \left\{ \Gamma(-\xi, -\log\eta) - \frac{1}{\xi}\left(\frac{1}{(-\log\eta)^\xi} - 1\right) \right\} \\
&\overset{(\clubsuit)}{=} -\lim_{\eta \nearrow 1} \left\{ \frac{1}{\xi}\left(-\Gamma(1-\xi, -\log\eta) + \frac{\eta}{(-\log\eta)^\xi}\right) - \frac{1}{\xi}\left(\frac{1}{(-\log\eta)^\xi} - 1\right) \right\} \\
&= \lim_{\eta \nearrow 1} \left\{ \underbrace{\frac{1}{\xi}(\Gamma(1-\xi, -\log\eta) - p)}_{=\Omega(\eta) \text{ from §B.1}} + \frac{1}{\xi} \cdot \frac{1-\eta}{(-\log\eta)^\xi} - \underbrace{\frac{1-\eta}{\xi}}_{\to 0} \right\} \\
&= \Omega(1) + \frac{1}{\xi}\lim_{\eta \nearrow 1} \frac{1-\eta}{(-\log\eta)^\xi} \\
&= \Omega(1) + \frac{1}{\xi}\lim_{\eta \nearrow 1} \frac{-1}{\xi(-\log\eta)^{\xi-1} \cdot (-\frac{1}{\eta})} && \triangleleft \text{ L'Hôpital's rule} \\
&= \Omega(1) + \frac{1}{\xi^2}\lim_{\eta \nearrow 1} p(-\log\eta)^{1-\xi} \\
&= \Omega(1), && \triangleleft\ x^{1-\xi} \overset{x \to 0}{\to} 0
\end{aligned}$$

where we used the recurrence relationship $\Gamma(s+1, x) = s\Gamma(s, x) + x^s e^{-x}$ at the identity ($\clubsuit$). It is elementary to check that $C = \Omega(1)$ for $\xi = 0$. Note that the integration constant C diverges for $\xi \geq 1$ hence the partial loss $\ell(\widehat{\eta}; 1)$ cannot be defined.

Then, we derive the composite loss: $\ell_{F_\xi^{-1}}(\theta; y) = \ell(F_\xi(\theta); y)$.

(Case A) When $\xi = 0$: Since the inverse link function (CDF) is $F_\xi(\theta) = \exp(-\exp(-\theta))$, the composite loss is

$$\ell_{F_\xi^{-1}}(\theta; y) = \begin{cases} \Gamma(0, e^{-\theta}) - \theta + \Omega(1) & \text{if } y = 1, \\ \Gamma(0, e^{-\theta}) & \text{if } y = 0. \end{cases}$$

(Case B) When $0 < \xi < 1$: Since the inverse link function (CDF) is

$$F_\xi(\theta) = \begin{cases} 0 & \text{if } \theta < -1/\xi, \\ \exp(-(1+\xi\theta)^{-1/\xi}) & \text{if } \theta \geq -1/\xi, \end{cases}$$

we have $F_\xi^{-1}(F_\xi(\theta)) = \max\{\theta, -1/\xi\}$[a]. Hence,

$$\ell_{F_\xi^{-1}}(\theta; y) = \begin{cases} \Gamma(-\xi, (1+\xi\theta)_+^{-1/\xi}) - \max\{\theta, -1/\xi\} + \Omega(1) & \text{if } y = 1, \\ \Gamma(-\xi, (1+\xi\theta)_+^{-1/\xi}) & \text{if } y = 0. \end{cases}$$

(Case C) When $\xi < 0$: Since the inverse link function (CDF) is

$$F_\xi(\theta) = \begin{cases} 1 & \text{if } \theta > -1/\xi, \\ \exp(-(1+\xi\theta)^{-1/\xi}) & \text{if } \theta \leq -1/\xi, \end{cases}$$

we have $F_\xi^{-1}(F_\xi(\theta)) = \min\{\theta, -1/\xi\}$. Hence,

$$\ell_{F_\xi^{-1}}(\theta; y) = \begin{cases} \Gamma(-\xi, (1+\xi\theta)_+^{-1/\xi}) - \min\{\theta, -1/\xi\} + \Omega(1) & \text{if } y = 1, \\ \Gamma(-\xi, (1+\xi\theta)_+^{-1/\xi}) & \text{if } y = 0. \end{cases}$$

[a] This operation is the source of nonconvexity of the canonical proper loss.

C Loss Interpretation from Bregman Divergence Perspective

In this section, we show how a gap between the Fenchel-Young loss and canonical composite loss arises. Let ℓ_ψ be a canonical composite loss with a link $\psi : [0,1] \to \mathbb{R}$ and a proper loss ℓ such that $\nabla\psi = w$, where w is a weight function of ℓ. Let ℓ_{Ω_F} be a Fenchel-Young loss associated with a CDF $F : \mathbb{R} \to [0,1]$. We specify a constraint $F(\theta) = \psi^{-1}(\theta)$ for $\theta \in \mathsf{Im}(\psi)$ to see difference between ℓ_ψ and ℓ_{Ω_F}. To map a real-valued prediction score $\theta \in \mathbb{R}$ to a probability estimate $\widehat{\eta} \in [0,1]$, the regularized predictor $\widehat{\eta} = \widehat{y}_{\Omega_F}(\theta) = F(\theta)$ and the inverse link function $\widehat{\eta} = \psi^{-1}(\theta)$ are used. Note that ψ^{-1} can only take $\theta \in \mathsf{Im}(\psi) = \mathrm{supp}(F)$ as an input.

Gap between Two Losses. The pointwise regret of the proper loss ℓ is

$$\begin{aligned} L(\widehat{\eta};\eta) - \underline{L}(\eta) &= \underbrace{\underline{L}(\widehat{\eta}) + \underline{L}'(\widehat{\eta})(\eta - \widehat{\eta})}_{\text{by Savage's representation (Savage, 1971)}} - \underline{L}(\eta) \\ &= (-\underline{L})(\eta) - (-\underline{L})(\widehat{\eta}) - (-\underline{L}')(\widehat{\eta})(\eta - \widehat{\eta}) \\ &= B_{-\underline{L}}(\eta \| \widehat{\eta}). \end{aligned}$$

On the other hand, the pointwise regret of the Fenchel-Young loss ℓ_{Ω_F} is

$$\begin{aligned} L_{\Omega_F}(\theta;\eta) - \underline{L}_{\Omega_F}(\eta) &= \ell_\Omega(\theta;\eta) \\ &= \Omega_F^\star(\theta) + \Omega_F(\eta) - \theta\eta. \end{aligned}$$

By using Corollary 7, we have

$$\underline{L}(\eta) = \int_0^\eta -\psi(q)\mathrm{d}q = \int_0^\eta -F^{-1}(q)\mathrm{d}q \overset{\text{by (4)}}{=} -\Omega_F(\eta),$$

which results in the gap Δ between the Fenchel-Young loss and canonical composite loss such that

$$\begin{aligned} \Delta &\overset{\text{def}}{=} \ell_{\Omega_F}(\theta;\eta) - B_{-\underline{L}}(\eta \| \widehat{\eta}) \\ &= \{\Omega_F^\star(\theta) + \Omega_F(\eta) - \theta\eta\} - \{\Omega_F(\eta) - \Omega_F(\widehat{\eta}) - \psi(\widehat{\eta})(\eta - \widehat{\eta})\} \\ &= \Omega_F^\star(\theta) + \Omega_F(\widehat{\eta}) - \theta\eta + \psi(\widehat{\eta})(\eta - \widehat{\eta}). \end{aligned}$$

Here, we notice that the probability estimate $\widehat{\eta}$ for the composite loss is obtained by $\widehat{\eta} = \psi^{-1}(\theta)$, to see the gap Δ. We divide the cases depending on a space where θ lives in.

If $0 < F(\theta) < 1$, then $\widehat{\eta} = F(\theta)$ from (3) and $\Delta = \Omega_F^\star(\theta) + \Omega_F(\widehat{\eta}) - \theta\widehat{\eta} = 0$, where the last identity is obtained via Danskin's theorem, that is, the Fenchel-Young gap is filled if and only if $\widehat{\eta} = \nabla\Omega_F^\star(\theta) = F(\theta)$. This is always implied by the assumption $0 < F(\theta) < 1$.

In contrast, if $\underline{\theta}_F > -\infty$ and $F(\theta) = 0$, then $\theta \le \underline{\theta}_F$ (see the definition of $\underline{\theta}_F$ in Assumption A) and $\Delta = \Omega_F^\star(\theta) + \eta(\underline{\theta}_F - \theta) = \eta(\underline{\theta}_F - \theta) \ge 0$. The gap is filled if and only if $\theta = \underline{\theta}_F$, that is, $\theta = \psi(0)$. Note that such θ exists for ψ with $\inf \mathsf{Im}(\psi) > -\infty$ (equivalently, $\inf \mathrm{supp}(F) > -\infty$). If $\theta < \underline{\theta}_F$, then $\theta < \psi(0)$, meaning that the gap Δ is strictly larger than 0.

If $\overline{\theta}_F < +\infty$ and $F(\theta) = 1$, it is confirmed in the same manner that the gap Δ is filled if and only if $\theta = \psi(1)$, otherwise $\Delta > 0$.

To summarize, the canonical composite loss and Fenchel-Young loss is the same for $\theta \in \mathsf{Im}(\psi) = \mathrm{supp}(F)$; otherwise, the Fenchel-Young loss is an upper bound of the canonical proper loss.

Nonconvexity of Canonical Composite Loss. We only confirm the nonconvexity of $\ell_\psi(\cdot;1)$ in the case where the support of F is bounded at the left end. The nonconvexity of $\ell_\psi(\cdot;0)$ can be confirmed in the same way as well. Take $\theta_0 < \underline{\theta}_F$, $\widetilde{\theta} = \underline{\theta}_F$, and $\theta_1 = 0$. We write $\ell_\psi(\cdot)$ for $\ell_\psi(\cdot;1)$ to make notation simpler. We show that the line segment is not always above $\ell_\psi(\cdot)$ on (θ_0, θ_1). Indeed, the line through $(\theta_0, \ell_\psi(\theta_0))$ and $(\theta_1, \ell_\psi(\theta_1))$ is

$$\frac{\ell_\psi(\theta_0) - \ell_\psi(\theta_1)}{\theta_0 - \theta_1}(\theta - \theta_0) + \ell_\psi(\theta_0) = (\ell(0) - \ell_\psi(0))\left(\frac{\theta}{\theta_0} - 1\right) + \ell(0),$$

because $\ell_\psi(\theta_0) = \ell(0)$ by the clipping of composite loss, and its value at $\theta = \widetilde{\theta}$ is $(\ell(0) - \ell_\psi(0))(\underline{\theta}_F/\theta_0 - 1) + \ell(0) < \ell(0) = \ell_\psi(\widetilde{\theta})$. Hence, the line segment is strictly below ℓ_ψ at $\theta = \widetilde{\theta}$, meaning that ℓ_ψ is nonconvex.

D Detail of Experiments

We describe the detail of baselines used in §6 here.

- **GEV-Can** is the canonical loss of the GEV link proposed by Agarwal et al. (2014).
- **GEV-Log** is a composite loss of the log loss and the GEV link, which is a equivalent formulation to Wang and Dey (2010) with a fixed shape parameter ξ.
- **Log** is ℓ_2-regularized logistic regression.
- **Platt** calibrates a classifier trained with the hinge loss with Platt's scaling (Platt, 1999), which performs post-hoc logistic regression on outputs of the trained classifier.
- **Isotonic** calibrates a ranking function trained with the logistic loss with isotonic regression (Menon et al., 2012). For probability calibration methods (Platt, Isotonic), we split the training set into the same-sized two sets and use the former for training the base classifier and the latter for probability calibration.
- **Weight** adopts cost-sensitive logistic regression, weighting the positive class with $1/\mathbb{P}(\mathsf{Y}=1) - 1$ to balance the both class.
- **Bagging** is a methodology to combine the undersampling and bagging, adopted by Wallace and Dahabreh (2012). We undersample the majority class to balance the dataset. To reduce the variance, we take average of logistic regressors trained with different 10 subsamples.

All methods including the proposed method (**GEV-FY**) are optimized by Adam (Kingma and Ba, 2015) with batch size 256, 3,000 epochs, and the ℓ_2-regularization. Both the learning rate and regularization parameter are chosen from $\{10^{-1}, 10^{-3}, 10^{-5}\}$. The early stopping is applied with the relative error tolerance 1.0×10^{-4} on training losses and 10 epochs patience. For GEV-FY, GEV-Can, GEV-Log, the shape parameter ξ is chosen from $\{-1, -0.75, -0.5, \dots, 0.5, 0.75\}$. Hyperparameters are chosen with the 5-fold cross validation.