Supplementary Material for "Calibrated Surrogate Maximization of Linear-fractional Utility in Binary Classification"

A Calibration Analysis and Deferred Proofs from Section 4

In this section, we analyze calibration of the surrogate utility. Before proceeding, we need to describe Bayes optimal classifier for a given metric.

Definition 13. *Given a linear-fractional utility* $\mathcal{U}$*,* Bayes optimal *set* $\mathcal{B} \subset \mathbb{R}^{\mathcal{X}}$ *is a set of functions that achieve the supremum of* $\mathcal{U}$*, that is,* $\mathcal{B} \doteq \{f \mid \mathcal{U}(f) = \mathcal{U}^\dagger = \sup_{f'} \mathcal{U}(f')\}$*.*

Classifiers in $\mathcal{B}$ are referred to as *Bayes optimal classifiers*. Note that they are not necessarily unique. In this work, we assume that $\mathcal{B} \neq \emptyset$. First, we characterize Bayes optimal set $\mathcal{B}$.

Proposition 14. *Given a linear-fractional utility* $\mathcal{U}$ *in Eq. (1), the Bayes optimal set* $\mathcal{B}$ *for* $\mathcal{U}$ *is*

$$\mathcal{B} = \{f \mid f(x)\{(\Delta a_0 - \Delta a_1 \mathcal{U}(f))\eta(x) - (a_{1,-1}\mathcal{U}(f) - a_{0,-1})\} > 0 \ \forall x \in \mathcal{X}\},$$

where $\Delta a_0 \doteq a_{0,+1} - a_{0,-1}$ *and* $\Delta a_1 \doteq a_{1,+1} - a_{1,-1}$*.*

Proof. The maximization problem in Eq. (2) can be restated as follows.

$$\max_{\lambda \in \Lambda} \bar{\mathcal{U}}(\lambda); \quad \bar{\mathcal{U}}(\lambda) \doteq \frac{\mathbb{E}_X[a_{0,+1}\lambda(X)\eta(X) + a_{0,-1}\lambda(X)(1-\eta(X)) + b_0]}{\mathbb{E}_X[a_{1,+1}\lambda(X)\eta(X) + a_{1,-1}\lambda(X)(1-\eta(X)) + b_1]},$$

where $\Lambda \doteq \{x \mapsto \ell(-f(x)) \mid f \in \mathcal{F}\} \subset \mathbb{R}^{\mathcal{X}}$. First, the Fréchet derivative of $\bar{\mathcal{U}}$ evaluated at x is obtained as follows.

$$\begin{aligned}
[\nabla_\lambda \bar{\mathcal{U}}(\lambda)]_x &= \frac{(\Delta a_0 \eta(x) + a_{0,-1})\mathbb{E}[W_1] - (\Delta a_1 \eta(x) + a_{1,-1})\mathbb{E}[W_0]}{\mathbb{E}[W_1]^2} p(x) \\
&= \frac{p(x)}{\mathbb{E}[W_1]} \left\{ \left(\Delta a_0 - \Delta a_1 \frac{\mathbb{E}[W_0]}{\mathbb{E}[W_1]}\right)\eta(x) - \left(a_{1,-1}\frac{\mathbb{E}[W_0]}{\mathbb{E}[W_1]} - a_{0,-1}\right)\right\} \\
&= \frac{p(x)}{\mathbb{E}[W_1]} \left\{(\Delta a_0 - \Delta a_1 \bar{\mathcal{U}}(\lambda))\eta(x) - (a_{1,-1}\bar{\mathcal{U}}(\lambda) - a_{0,-1})\right\}
\end{aligned}$$

Let $f^\dagger \in \mathcal{F}$ be a function that maximizes $\mathcal{U}$, and $\lambda^\dagger \doteq \ell(-f^\dagger(\cdot))$. Then, $\lambda^\dagger$ maximizes $\bar{\mathcal{U}}$, and it satisfies (Koyejo et al., 2014, lemma 12)

$$\int_{\mathcal{X}} [\nabla_\lambda \bar{\mathcal{U}}(\lambda^\dagger)]_x \lambda^\dagger(x) dx \geq \int_{\mathcal{X}} [\nabla_\lambda \bar{\mathcal{U}}(\lambda^\dagger)]_x \lambda(x) dx \quad \forall \lambda \in \Lambda.$$

Thus, the necessary condition for local optimality is that $\mathrm{sgn}(\lambda^\dagger(x)) = \mathrm{sgn}([\nabla_\lambda \bar{\mathcal{U}}(\lambda^\dagger)]_x)$ for all $x \in \mathcal{X}$.[5] Since $\mathrm{sgn}(\lambda^\dagger(x)) = \mathrm{sgn}(\ell(-f^\dagger(x))) = \mathrm{sgn}(f^\dagger(x))$, the above condition is $\mathrm{sgn}(f^\dagger(x)) = \mathrm{sgn}([\nabla_\lambda \bar{\mathcal{U}}(\lambda^\dagger)]_x)$ for all $x \in \mathcal{X}$, which is equivalent to the condition $f^\dagger(x)\{(\Delta a_0 - \Delta a_1 \mathcal{U}(f^\dagger))\eta(x) - (a_{1,-1}\mathcal{U}(f^\dagger) - a_{0,-1})\} > 0$ for all $x \in \mathcal{X}$. This concludes the proof. Note that $p(x)/\mathbb{E}[W_1]$ is a positive value, and $\bar{\mathcal{U}}(\lambda^\dagger) = \mathcal{U}(f^\dagger)$. □

You may confirm that Proposition 14 is consistent with Bayes optimal classifier in the classical case, accuracy (Bartlett et al., 2006): a Bayes optimal classifier $f^\dagger$ should satisfy $f^\dagger(x)(2\eta(x) - 1) > 0$ for all $x \in \mathcal{X}$, since $a_{0,+1} = 1$, $a_{0,-1} = -1$, $a_{1,+1} = a_{1,-1} = b_0 = b_1 = 0$.

Next, we state a proposition which gives a direction to prove the surrogate calibration of a surrogate utility. This proposition follows a latter half of Gao and Zhou (2015, Theorem 2).

Proposition 15. *Fix a true utility* $\mathcal{U}$*, a surrogate utility* $\mathcal{U}_\phi$*, and let* $\mathcal{B}$ *a Bayes optimal set corresponding to the utility* $\mathcal{U}$*. Assume that*

$$\sup_{f \notin \mathcal{B}} \mathcal{U}_\phi(f) < \sup_f \mathcal{U}_\phi(f). \tag{8}$$

Then, the surrogate utility $\mathcal{U}_\phi$ *is* $\mathcal{U}$*-calibrated.*

[5] This can be confirmed in a similar manner to the proof of Yan et al. (2018, Theorem 3.1).

Proof. Remind that $\mathcal{U}_\phi \doteq \sup_f \mathcal{U}_\phi(f)$ and let

$$\delta \doteq \mathcal{U}_\phi^* - \sup_{f \notin \mathcal{B}} \mathcal{U}_\phi(f) > 0,$$

and $\{f_l\}_{l \geq 1}$ be any sequence such that $\mathcal{U}_\phi(f_l) \stackrel{l \to \infty}{\longrightarrow} \mathcal{U}_\phi^*$. Then, for any $\varepsilon > 0$, there exists $l_0 \in \mathbb{Z}$ such that $\mathcal{U}_\phi^* - \mathcal{U}_\phi(f_l) < \varepsilon$ for $l \geq l_0$. Here we set $\varepsilon = \frac{\delta}{2}$: $\mathcal{U}_\phi^* - \mathcal{U}_\phi(f_l) < \frac{\delta}{2}$ for $l \geq l_0$. If we assume that $f_l \notin \mathcal{B}$, this contradicts with the following facts: for a function $f \notin \mathcal{B}$,

$$\mathcal{U}_\phi^* - \mathcal{U}_\phi(f) = \underbrace{\mathcal{U}_\phi^* - \sup_{f' \notin \mathcal{B}} \mathcal{U}_\phi(f')}_{=\delta} + \underbrace{\sup_{f' \notin \mathcal{B}} \mathcal{U}_\phi(f') - \mathcal{U}_\phi(f)}_{\geq 0} \geq \delta.$$

Thus, it holds that $f_l \in \mathcal{B}$ for $l \geq l_0$, that is, $\mathcal{U}(f_l) = \mathcal{U}^\dagger$, which indicates $\mathcal{U}$-calibration. □

Thus, the proof of $\mathcal{U}$-calibration of $\mathcal{U}_\phi$ is reduced to show the condition (8). Below, we show the surrogate calibration for the F_β-measure and Jaccard index utilizing Propositions 14 and 15. The proofs are based on the above propositions, Gao and Zhou (2015, Lemma 6) and Charoenphakdee et al. (2019, Theorem 11).

Throughout the proofs, we assume that for the critical set $\mathcal{C}^\dagger \doteq \{x \mid (\Delta a_0 - \Delta a_1 \mathcal{U}(f^\dagger))\eta(x) - (a_{1,-1}\mathcal{U}(f^\dagger) - a_{0,-1}) = 0\}$, $\mathbb{P}(\mathcal{C}^\dagger) = 0$, where $f^\dagger$ is the classifier attaining the supremum of $\mathcal{U}$. For example, this holds for any η-*continuous* distribution (Yan et al., 2018, Assumption 2).

A.1 Proof of Theorem 9

As a surrogate utility of the F_β-measure following Eq. (4), we have

$$\begin{aligned}
\mathcal{U}_\phi^{\mathsf{F}_\beta}(f) &= \frac{\int_{\mathcal{X}} \{(1+\beta^2)(1-\phi(f(x)))\eta(x)\} p(x) dx}{\int_{\mathcal{X}} \{(1+\phi(f(x)))\eta(x) + \phi(-f(x))(1-\eta(x)) + \beta^2\pi\} p(x) dx} \\
&\doteq \frac{\mathbb{E}_X[W_{0,\phi}^{\mathsf{F}_\beta}(f(X), \eta(X))]}{\mathbb{E}_X[W_{1,\phi}^{\mathsf{F}_\beta}(f(X), \eta(X))]},
\end{aligned}$$

where

$$\begin{aligned}
W_{0,\phi}^{\mathsf{F}_\beta}(\xi, q) &\doteq (1+\beta^2)(1-\phi(\xi))q, \\
W_{1,\phi}^{\mathsf{F}_\beta}(\xi, q) &\doteq (1+\phi(\xi))q + \phi(-\xi)(1-q) + \beta^2\pi.
\end{aligned}$$

From Proposition 14, the Bayes optimal set $\mathcal{B}^{\mathsf{F}_\beta}$ for the F_β-measure is

$$\mathcal{B}^{\mathsf{F}_\beta} \doteq \{f \mid f(x)((1+\beta^2)\eta(x) - \mathcal{U}^{\mathsf{F}_\beta}(f)) > 0 \quad \forall x \in \mathcal{X}\}.$$

We will show F_β-calibration by utilizing Proposition 15, which casts our proof target into showing Eq. (8). We prove it by contradiction. Assume that

$$\sup_{f \notin \mathcal{B}^{\mathsf{F}_\beta}} \mathcal{U}_\phi^{\mathsf{F}_\beta}(f) = \sup_f \mathcal{U}_\phi^{\mathsf{F}_\beta}(f).$$

This implies that there exists an optimal function $f^* \notin \mathcal{B}^{\mathsf{F}_\beta}$ that achieves $\mathcal{U}_\phi^{\mathsf{F}_\beta}(f^*) = \sup_f \mathcal{U}_\phi^{\mathsf{F}_\beta}(f) \doteq (\mathcal{U}_\phi^{\mathsf{F}_\beta})^*$, that is, $\mathcal{U}_\phi^{\mathsf{F}_\beta}(f^*) = (\mathcal{U}_\phi^{\mathsf{F}_\beta})^*$ and $f^*(\bar{x})((1+\beta^2)\eta(\bar{x}) - \mathcal{U}^{\mathsf{F}_\beta}(f^*)) \leq 0$ for some $\bar{x} \in \mathcal{X}$.

Let us describe the *stationary condition* of f^*. We introduce a function δf:

$$\delta f(x) \doteq \begin{cases} 1 & \text{if } x = \bar{x}, \\ 0 & \text{if } x \neq \bar{x}. \end{cases}$$

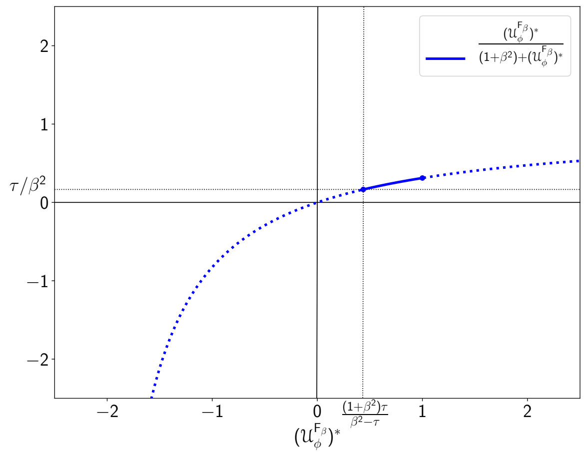

Figure 6: The range of $\frac{(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*}{(1+\beta^2)+(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*}$ in $\frac{(1+\beta^2)\tau}{\beta^2-\tau} \leq (\mathcal{U}_\phi^{\mathsf{F}_\beta})^* \leq 1$.

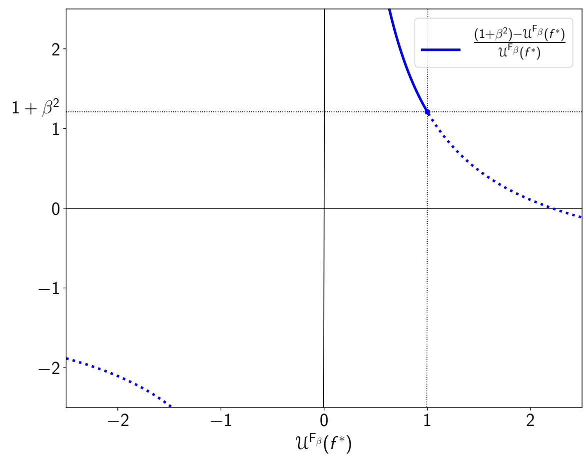

Figure 7: The range of $\frac{(1+\beta^2)-\mathcal{U}^{\mathsf{F}_\beta}(f^*)}{\mathcal{U}^{\mathsf{F}_\beta}(f^*)}$ in $0 < \mathcal{U}^{\mathsf{F}_\beta}(f^*) \leq 1$.

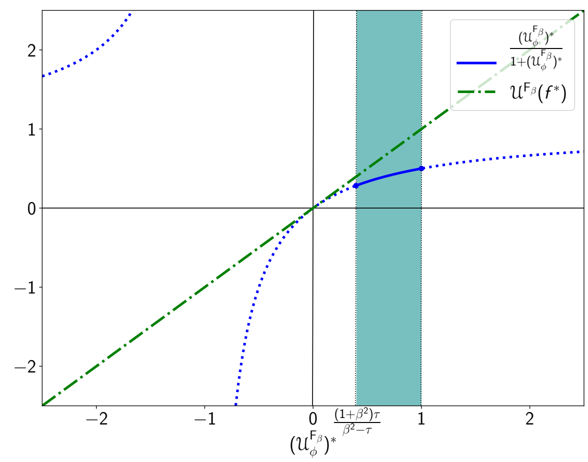

Figure 8: If $\frac{(1+\beta^2)\tau}{\beta^2-\tau} \leq (\mathcal{U}_\phi^{\mathsf{F}_\beta})^* \leq 1$, then $\frac{(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*}{1+(\mathcal{U}_\phi^{\mathsf{F}_\beta})^2} \leq \mathcal{U}^{\mathsf{F}_\beta}(f^*)$.

Let $G(\gamma) \doteq \mathcal{U}_\phi^{\mathsf{F}_\beta}(f^* + \gamma\delta f)$. Since $\mathcal{U}_\phi^{\mathsf{F}_\beta}$ is Gâteaux differentiable[6] and its Gâteaux derivative at f^* must be zero in any direction, we claim that $G'(0) = 0$, where $G'(0)$ corresponds to Gâteaux derivative of $\mathcal{U}_\phi^{\mathsf{F}_\beta}$ at f^* in the

[6] Fréchet differentiability implies Gâteaux differentiability.

direction of δf. Here, $G'(0)$ is computed as

$$\begin{aligned}
G'(0) &= \frac{1}{\mathbb{E}[W_{1,\phi}^{\mathsf{F}_\beta}(f^*)]^2}\Bigg\{\mathbb{E}[W_{1,\phi}^{\mathsf{F}_\beta}(f^*)]\int_{\mathcal{X}}\{-(1+\beta^2)\phi'(f^*(x))\delta f(x)\eta(x)\}p(x)dx \\
&\quad -\mathbb{E}[W_{0,\phi}^{\mathsf{F}_\beta}(f^*)]\int_{\mathcal{X}}\{\phi'(f^*(x))\delta f(x)\eta(x)-\phi'(-f^*(x))\delta f(x)(1-\eta(x))\}p(x)dx\Bigg\} \\
&= \frac{1}{\mathbb{E}[W_{1,\phi}^{\mathsf{F}_\beta}(f^*)]}\Bigg\{\int_{\mathcal{X}}\{-(1+\beta^2)\phi'(f^*(x))\delta f(x)\eta(x)\}p(x)dx \\
&\quad -(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*\int_{\mathcal{X}}\{\phi'(f^*(x))\delta f(x)\eta(x)-\phi'(-f^*(x))\delta f(x)(1-\eta(x))\}p(x)dx\Bigg\} \\
&= \frac{\{-(1+\beta^2)\phi'(f^*(\bar{x}))\eta(\bar{x})-\phi'(f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*\eta(\bar{x})+\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*(1-\eta(\bar{x}))\}p(\bar{x})}{\mathbb{E}[W_{1,\phi}^{\mathsf{F}_\beta}(f^*)]},
\end{aligned}$$

where $\mathbb{E}[W_{0,\phi}^{\mathsf{F}_\beta}(f^*)] = \mathbb{E}_X[W_{0,\phi}^{\mathsf{F}_\beta}(f^*(X),\eta(X))]$ and $\mathbb{E}[W_{1,\phi}^{\mathsf{F}_\beta}(f^*)] = \mathbb{E}_X[W_{1,\phi}^{\mathsf{F}_\beta}(f^*(X),\eta(X))]$. Thus, the stationary condition is

$$\begin{aligned}
-(1+\beta^2)\phi'(f^*(\bar{x}))\eta(\bar{x})-\phi'(f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*\eta(\bar{x})+\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*(1-\eta(\bar{x})) = 0 \\
\left\{-(1+\beta^2)\phi'(f^*(\bar{x}))-\phi'(f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*-\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*\right\}\eta(\bar{x})+\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^* = 0. \qquad (9)
\end{aligned}$$

Since $\phi'(\pm f^*(\bar{x})) < 0$, we have $-(1+\beta^2)\phi'(f^*(\bar{x}))-\phi'(f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*-\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^* > 0$. Thus, the condition (9) becomes

$$\eta(\bar{x}) = \frac{\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*}{(1+\beta^2)\phi'(f^*(\bar{x}))+(\phi'(f^*(\bar{x}))+\phi'(-f^*(\bar{x})))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*}. \qquad (10)$$

From now on, we divide the cases to take care of the Bayes optimal condition $f^*(\bar{x})((1+\beta^2)\eta(\bar{x})-\mathcal{U}^{\mathsf{F}_\beta}(f^*)) \ge 0$.

1) <u>If $f^*(\bar{x}) > 0$ and $\eta(\bar{x}) < \frac{1}{1+\beta^2}\mathcal{U}^{\mathsf{F}_\beta}(f^*)$:</u> We show

$$\frac{\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*}{(1+\beta^2)\phi'(f^*(\bar{x}))+(\phi'(f^*(\bar{x}))+\phi'(-f^*(\bar{x})))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*} \ge \frac{\mathcal{U}^{\mathsf{F}_\beta}(f^*)}{1+\beta^2}. \qquad (11)$$

Take the difference of the left-hand side and the right-hand side:

$$\begin{aligned}
&\frac{\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*}{(1+\beta^2)\phi'(f^*(\bar{x}))+(\phi'(f^*(\bar{x}))+\phi'(-f^*(\bar{x})))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*} - \frac{\mathcal{U}^{\mathsf{F}_\beta}(f^*)}{1+\beta^2} \\
&= \frac{(1+\beta^2)\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*-(1+\beta^2)\phi'(f^*(\bar{x}))\mathcal{U}^{\mathsf{F}_\beta}(f^*)-(\phi'(f^*(\bar{x}))+\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*\mathcal{U}^{\mathsf{F}_\beta}(f^*)}{(1+\beta^2)((1+\beta^2)\phi'(f^*(\bar{x}))+(\phi'(f^*(\bar{x}))+\phi'(-f^*(\bar{x})))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*)},
\end{aligned}$$

where the denominator is always negative, which reduces to show the numerator is always negative, too:

$$\begin{aligned}
&(1+\beta^2)\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*-(1+\beta^2)\phi'(f^*(\bar{x}))\mathcal{U}^{\mathsf{F}_\beta}(f^*)-(\phi'(f^*(\bar{x}))+\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*\mathcal{U}^{\mathsf{F}_\beta}(f^*) \\
&\quad = \mathcal{U}^{\mathsf{F}_\beta}(f^*)((1+\beta^2)+(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*)\left(\underbrace{\frac{(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*}{(1+\beta^2)+(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*}}_{\ge\tau/\beta^2}\underbrace{\frac{(1+\beta^2)-\mathcal{U}^{\mathsf{F}_\beta}(f^*)}{\mathcal{U}^{\mathsf{F}_\beta}(f^*)}}_{\ge\beta^2}\phi'(-f^*(\bar{x}))-\phi'(f^*(\bar{x}))\right) \\
&\quad \le \mathcal{U}^{\mathsf{F}_\beta}(f^*)((1+\beta^2)+(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*)\left(\tau\phi'(-f^*(\bar{x}))-\phi'(f^*(\bar{x}))\right) \\
&\quad \le 0,
\end{aligned}$$

where the first inequality holds because $\frac{(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*}{(1+\beta^2)+(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*} \geq \frac{\tau}{\beta^2}$ when $\frac{(1+\beta^2)\tau}{\beta^2-\tau} \leq (\mathcal{U}_\phi^{\mathsf{F}_\beta})^* \leq 1$ (see Figure 6) and $\frac{(1+\beta^2)-\mathcal{U}^{\mathsf{F}_\beta}(f^*)}{\mathcal{U}^{\mathsf{F}_\beta}(f^*)} \geq \beta^2$ when $0 \leq \mathcal{U}^{\mathsf{F}_\beta}(f^*) \leq 1$ (see Figure 7). Note that $\phi'(-f^*(\bar{x})) < 0$. The second inequality holds because of the assumption that $\lim_{m\searrow 0}\phi'(m) \geq \tau \lim_{m\nearrow 0}\phi'(m)$ and ϕ is convex, which implies $\tau\phi'(-m) - \phi'(m) \leq 0$ for every $m > 0$.

Thus, the inequality (11) holds, which implies the following contradiction.

$$\eta(\bar{x}) = \frac{\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*}{(1+\beta^2)\phi'(f^*(\bar{x})) + (\phi'(f^*(\bar{x})) + \phi'(-f^*(\bar{x})))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*} \geq \frac{\mathcal{U}^{\mathsf{F}_\beta}(f^*)}{1+\beta^2} > \eta(\bar{x}).$$

2) If $f^*(\bar{x}) \leq 0$ and $\eta(\bar{x}) > \frac{1}{1+\beta^2}\mathcal{U}^{\mathsf{F}_\beta}(f^*)$: As well as the previous case, we begin from the stationary condition (10). If $\phi'(-f^*(\bar{x})) < 0$,

$$\begin{aligned}
\eta(\bar{x}) &= \frac{\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*}{(1+\beta^2)\phi'(f^*(\bar{x})) + (\phi'(f^*(\bar{x})) + \phi'(-f^*(\bar{x})))(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*} \\
&= \frac{(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*}{(1+\beta^2)\frac{\phi'(f^*(\bar{x}))}{\phi'(-f^*(\bar{x}))} + \left(\frac{\phi'(f^*(\bar{x}))}{\phi'(-f^*(\bar{x}))} + 1\right)(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*} \\
&\leq \frac{1}{1+\beta^2}\frac{(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*}{1+(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*} \\
&\leq \frac{1}{1+\beta^2}\mathcal{U}^{\mathsf{F}_\beta}(f^*) \\
&< \eta(\bar{x}), \qquad \text{(contradiction)}
\end{aligned}$$

where the first inequality holds because $\frac{\phi'(-m)}{\phi'(m)} \geq 1$ for every $m \geq 0$ and $f^*(\bar{x}) \leq 0$, and the second inequality holds because $\mathcal{U}_\phi^{\mathsf{F}_\beta}(f) \leq \mathcal{U}^{\mathsf{F}_\beta}(f)$ $(\forall f)$ implies $\frac{(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*}{1+(\mathcal{U}_\phi^{\mathsf{F}_\beta})^*} \leq \mathcal{U}^{\mathsf{F}_\beta}(f^*)$ when $\frac{(1+\beta^2)\tau}{\beta^2-\tau} \leq (\mathcal{U}_\phi^{\mathsf{F}_\beta})^* \leq 1$ (see Figure 8). If $\phi'(-f^*(\bar{x})) = 0$, it is easy to see the contradiction.

Combining the above cases, it follows that

$$\sup_{f \notin \mathcal{B}^{\mathsf{F}_\beta}} \mathcal{U}_\phi^{\mathsf{F}_\beta}(f) < \sup_f \mathcal{U}_\phi^{\mathsf{F}_\beta}(f).$$

Eventually, we claim that $\mathcal{U}_\phi^{\mathsf{F}_\beta}$ is F_β-calibrated by using Proposition 15. □

A.2 Proof of Theorem 10

As a surrogate utility of the Jaccard index following Eq. (4), we have

$$\mathcal{U}_\phi^{\mathsf{Jac}}(f) = \frac{\int_{\mathcal{X}}(1-\phi(f(x)))\eta(x)p(x)dx}{\int_{\mathcal{X}}\{\phi(-f(x))(1-\eta(x)) + \pi\}p(x)dx},$$

and we have the Bayes optimal set $\mathcal{B}^{\mathsf{Jac}}$ for the Jaccard index such as

$$\mathcal{B}^{\mathsf{Jac}} \doteq \left\{f \mid f(x)\{(1+\mathcal{U}^{\mathsf{Jac}}(f))\eta(x) - \mathcal{U}^{\mathsf{Jac}}(f)\} > 0 \quad \forall x \in \mathcal{X}\right\},$$

utilizing Proposition 14. We follow the same proof technique, proof by contradiction, as we use in the proof of Theorem 9. Assume that

$$\sup_{f \notin \mathcal{B}^{\mathsf{Jac}}} \mathcal{U}_\phi^{\mathsf{Jac}}(f) = \sup_f \mathcal{U}_\phi^{\mathsf{Jac}}(f),$$

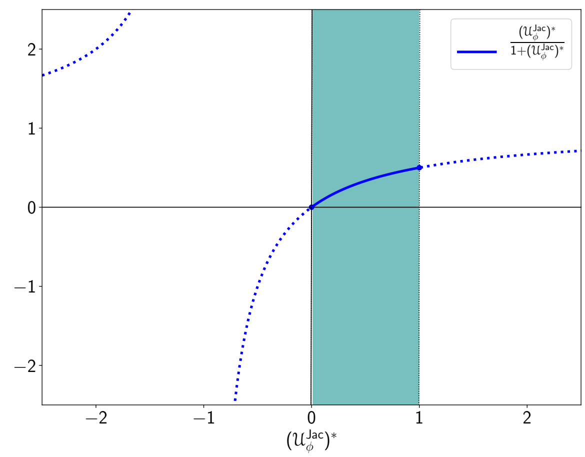

Figure 9: $\frac{(\mathcal{U}_\phi^{\mathsf{Jac}})^*}{1+(\mathcal{U}_\phi^{\mathsf{Jac}})^*}$ is monotonically increasing if $0 \le (\mathcal{U}_\phi^{\mathsf{Jac}})^* \le 1$.

which implies that there exits an optimal function $f^* \notin \mathcal{B}^{\mathsf{Jac}}$ that achieves $\mathcal{U}_\phi^{\mathsf{Jac}}(f^*) = \sup_f \mathcal{U}_\phi^{\mathsf{Jac}}(f) \doteq (\mathcal{U}_\phi^{\mathsf{Jac}})^*$, that is, $\mathcal{U}_\phi^{\mathsf{Jac}}(f^*) = (\mathcal{U}_\phi^{\mathsf{Jac}})^*$ and $f^*(\bar{x})\{(1+\mathcal{U}^{\mathsf{Jac}}(f^*))\eta(\bar{x}) - \mathcal{U}^{\mathsf{Jac}}(f^*)\} \le 0$ for some $\bar{x} \in \mathcal{X}$.

The stationary condition of $\mathcal{U}_\phi^{\mathsf{Jac}}$ around f^* can be stated as well as Eq. (10) in Theorem 9:

$$\eta(\bar{x}) = \frac{\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{Jac}})^*}{\phi'(f^*(\bar{x})) + \phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{Jac}})^*}. \tag{12}$$

1) <u>If $f^*(\bar{x}) > 0$ and $\eta(\bar{x}) < \frac{\mathcal{U}^{\mathsf{Jac}}(f^*)}{1+\mathcal{U}^{\mathsf{Jac}}(f^*)}$:</u> We show

$$\frac{\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{Jac}})^*}{\phi'(f^*(\bar{x})) + \phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{Jac}})^*} \ge \frac{\mathcal{U}^{\mathsf{Jac}}(f^*)}{1+\mathcal{U}^{\mathsf{Jac}}(f^*)}.$$

First, take the difference of the left-hand side and the right-hand side.

$$\begin{aligned}
&\frac{\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{Jac}})^*}{\phi'(f^*(\bar{x})) + \phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{Jac}})^*} - \frac{\mathcal{U}^{\mathsf{Jac}}(f^*)}{1+\mathcal{U}^{\mathsf{Jac}}(f^*)} \\
&\quad = \frac{\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{Jac}})^* - \phi'(f^*(\bar{x}))\mathcal{U}^{\mathsf{Jac}}(f^*)}{(\phi'(f^*(\bar{x})) + \phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{Jac}})^*)(1+\mathcal{U}^{\mathsf{Jac}}(f^*))},
\end{aligned}$$

where the denominator is always negative, which reduces to show the numerator is always negative, too. If $\phi'(-f^*(\bar{x})) < 0$,

$$\begin{aligned}
&\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{Jac}})^* - \phi'(f^*(\bar{x}))\mathcal{U}^{\mathsf{Jac}}(f^*) \\
&\quad = \phi'(-f^*(\bar{x}))\left((\mathcal{U}_\phi^{\mathsf{Jac}})^* - \frac{\phi'(f^*(\bar{x}))}{\phi'(-f^*(\bar{x}))}\mathcal{U}^{\mathsf{Jac}}(f^*)\right) \\
&\quad \le \phi'(-f^*(\bar{x}))\left((\mathcal{U}_\phi^{\mathsf{Jac}})^* - \frac{\phi'(f^*(\bar{x}))}{\phi'(-f^*(\bar{x}))}\right) && (\because \mathcal{U}^{\mathsf{Jac}}(f^*) \le 1) \\
&\quad \le \phi'(-f^*(\bar{x}))((\mathcal{U}_\phi^{\mathsf{Jac}})^* - \tau) \\
&\quad \le 0, && (\because (\mathcal{U}_\phi^{\mathsf{Jac}})^* \ge \tau)
\end{aligned}$$

where the second inequality holds because of the assumption that $\lim_{m\searrow 0}\phi'(m) \ge \tau \lim_{m\nearrow 0}\phi'(m)$ for every $m > 0$, and ϕ is convex. Thus, we admit the contradiction.

$$\eta(\bar{x}) = \frac{\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{Jac}})^*}{\phi'(f^*(\bar{x})) + \phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{Jac}})^*} > \frac{\mathcal{U}^{\mathsf{Jac}}(f^*)}{1+\mathcal{U}^{\mathsf{Jac}}(f^*)} > \eta(\bar{x}).$$

If $\phi'(-f^*(\bar{x}))=0$, then $\phi'(f^*(\bar{x}))=0$ from the assumption $\lim_{m\searrow 0}\phi'(m) \geq \tau \lim_{m\nearrow 0}\phi'(m)$, which immediately results in the contradiction.

2) If $f^*(\bar{x}) \leq 0$ and $\eta(\bar{x}) > \frac{\mathcal{U}^{\mathsf{Jac}}(f^*)}{1+\mathcal{U}^{\mathsf{Jac}}(f^*)}$: We begin from the stationary condition in Eq. (12). If $\phi'(-f^*(\bar{x}))<0$,

$$\begin{aligned}
\eta(\bar{x}) &= \frac{\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{Jac}})^*}{\phi'(f^*(\bar{x}))+\phi'(-f^*(\bar{x}))(\mathcal{U}_\phi^{\mathsf{Jac}})^*} \\
&= \frac{(\mathcal{U}_\phi^{\mathsf{Jac}})^*}{\frac{\phi'(f^*(\bar{x}))}{\phi'(-f^*(\bar{x}))}+(\mathcal{U}_\phi^{\mathsf{Jac}})^*} \\
&\leq \frac{(\mathcal{U}_\phi^{\mathsf{Jac}})^*}{1+(\mathcal{U}_\phi^{\mathsf{Jac}})^*} && \left(\because\ \frac{\phi'(f^*(\bar{x}))}{\phi'(-f^*(\bar{x}))} \geq 1 \quad \forall f^*(\bar{x}) \leq 0\right) \\
&\leq \frac{\mathcal{U}^{\mathsf{Jac}}(f^*)}{1+\mathcal{U}^{\mathsf{Jac}}(f^*)} \\
&< \eta(\bar{x}), && \text{(contradiction)}
\end{aligned}$$

where the second inequality follows because $\mathcal{U}_\phi^{\mathsf{Jac}}(f) \leq \mathcal{U}^{\mathsf{Jac}}(f)$ $(\forall f)$ and a function $x \mapsto \frac{x}{1+x}$ $(0 \leq x \leq 1)$ is monotonically increasing (see Figure 9).

It is easy to see contradiction in case of $\phi'(-f^*(\bar{x}))=0$.

Combining the above cases, it follows that

$$\sup_{f \notin \mathcal{B}^{\mathsf{Jac}}} \mathcal{U}_\phi^{\mathsf{Jac}}(f) < \sup_f \mathcal{U}^{\mathsf{Jac}}(f).$$

Eventually, we claim that $\mathcal{U}_\phi^{\mathsf{Jac}}$ is Jaccard-calibrated by using Proposition 15. □

A.3 Analysis of Accuracy-Calibration

In this subsection, we show accuracy-calibration conditions in the same manners as the F_β-measure and Jaccard index, and confirm that the τ-discrepancy is not necessary in this case. As the true and a surrogate utility of the accuracy following Eq. (4), define

$$\begin{aligned}
\mathcal{U}^{\mathsf{Acc}}(f) &= \int_{\mathcal{X}} \{\ell(-f(x))\eta(x) - \ell(-f(x))(1-\eta(x)) + (1-\pi)\}\, p(x)\mathrm{d}x, \\
\mathcal{U}_\phi^{\mathsf{Acc}}(f) &= \int_{\mathcal{X}} \{(1-\phi(f(x)))\eta(x) - \phi(-f(x))(1-\eta(x)) + (1-\pi)\}\, p(x)\mathrm{d}x.
\end{aligned}$$

Proposition 16 (Accuracy-calibration)**.** *Assume that a surrogate loss $\phi:\mathbb{R}\to\mathbb{R}_{\geq 0}$ is convex, differentiable almost everywhere, and $\phi'(0)<0$. Then, $\mathcal{U}_\phi^{\mathsf{Acc}}$ is accuracy-calibrated.*

Proof. We have the Bayes optimal set $\mathcal{B}^{\mathsf{Acc}}$ for the accuracy such as

$$\mathcal{B}^{\mathsf{Acc}} \doteq \{f \mid f(x)(2\eta(x)-1)>0 \quad \forall x \in \mathcal{X}\}$$

utilizing Proposition 14. In the same manner as the proofs of Theorems 9 and 10, assume that

$$\sup_{f \notin \mathcal{B}^{\mathsf{Acc}}} \mathcal{U}_\phi^{\mathsf{Acc}}(f) = \sup_f \mathcal{U}_\phi^{\mathsf{Acc}}(f),$$

and we prove by contradiction. The above assumption implies that there exists an optimal function $f^* \notin \mathcal{B}^{\mathsf{Acc}}$ such that $\mathcal{U}_\phi^{\mathsf{Acc}}(f^*) = \sup_f \mathcal{U}_\phi^{\mathsf{Acc}}(f) \doteq (\mathcal{U}_\phi^{\mathsf{Acc}})^*$, that is, $\mathcal{U}_\phi^{\mathsf{Acc}}(f^*) = (\mathcal{U}_\phi^{\mathsf{Acc}})^*$ and $f^*(\bar{x})(2\eta(\bar{x})-1) \leq 0$ for some $\bar{x} \in \mathcal{X}$.

The stationary condition of $\mathcal{U}_\phi^{\mathsf{Acc}}$ around f^* can be stated in the same way as Eq. (9):

$$(\phi'(f^*(\bar{x}))+\phi'(-f^*(\bar{x})))\eta(\bar{x}) - \phi'(-f^*(\bar{x})) = 0. \tag{13}$$

We divide the cases based on the sign of $f^*(\bar{x})$.

1) $f^*(\bar{x}) > 0$ and $\eta(\bar{x}) < \frac{1}{2}$: Since $\phi'(-f^*(\bar{x})) < \phi'(f^*(\bar{x})) < 0$ because of the convexity of ϕ,

$$\eta(\bar{x}) = \frac{1}{1 + \frac{\phi'(f^*(\bar{x}))}{\phi'(-f^*(\bar{x}))}} > \frac{1}{2},$$

which contradicts with $\eta(\bar{x}) < \frac{1}{2}$. Note that $\frac{\phi'(f^*(\bar{x}))}{\phi'(-f^*(\bar{x}))} \in (0,1)$ when $\phi'(-f^*(\bar{x})) < \phi'(f^*(\bar{x})) < 0$.

2) $f^*(\bar{x}) < 0$ and $\eta(\bar{x}) > \frac{1}{2}$: Since $\phi'(f^*(\bar{x})) < \phi'(-f^*(\bar{x})) < 0$ because of the convexity of ϕ,

$$\eta(\bar{x}) = \frac{1}{1 - \frac{\phi'(f^*(\bar{x}))}{\phi'(-f^*(\bar{x}))}} < \frac{1}{2},$$

which contradicts with $\eta(\bar{x}) > \frac{1}{2}$. Note that $\frac{\phi'(f^*(\bar{x}))}{\phi'(-f^*(\bar{x}))} > 1$ when $\phi'(f^*(\bar{x})) < \phi'(-f^*(\bar{x})) < 0$.

3) $f^*(\bar{x}) = 0$: Since $\phi'(f^*(\bar{x})) = \phi'(-f^*(\bar{x})) = \phi'(0) < 0$, the stationary condition (13) reduces to $\phi'(-f^*(\bar{x})) = 0$, which contradicts with $\phi'(-f^*(\bar{x})) = \phi'(0) < 0$.

Thus, it follows that $\sup_{f \notin \mathcal{U}^{\mathsf{Acc}}} \mathcal{U}_\phi^{\mathsf{Acc}}(f) < \sup_f \mathcal{U}_\phi^{\mathsf{Acc}}(f)$. Eventually, we claim that $\mathcal{U}_\phi^{\mathsf{Acc}}$ is accuracy-calibrated by using Proposition 15. □

As we can see from Proposition 16, our surrogate calibration analysis can also be applied to the classification accuracy. In addition, the τ-discrepancy condition disappears from assumptions in the accuracy case, which recovers the conditions Theorem 6 in Bartlett et al. (2006). Even so, our analysis still remains to be sufficient conditions. Further analysis towards the necessary and sufficient conditions in the general calibration analysis is left as an future work.

A.4 Calibration Analysis of General Linear-fractional Metrics

So far, we analyze the surrogate calibration for the F_β-measure in Theorem 9, and Jaccard index in Theorem 10. In addition, we take a look at how our analysis goes for the classification accuracy in Theorem 16. Now, we move on to the generalized result of the surrogate calibration which encompasses the entire linear-fractional metrics. Let us consider the maximization of the true utility $\mathcal{U}$ in Eq. (1), and the maximization of the corresponding surrogate utility $\mathcal{U}_\phi$ in Eq. (4).

Theorem 17 ($\mathcal{U}$-calibration in general case)**.** *Let f^* be a measurable function that achieves $\mathcal{U}_\phi(f^*) = \sup_f \mathcal{U}_\phi(f) \doteq \mathcal{U}_\phi^*$. Assume that a surrogate loss $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is convex, non-increasing, and differentiable almost everywhere. On the true utility, we assume the following conditions.*

(1) $\Delta a_0 > 0$.

(2) $\Delta a_1 \leq 0$.

(3) $a_{0,-1} \neq 0$ *or* $a_{1,-1} \neq 0$.

(4) $a_{1,-1} + a_{0,-1} \neq 0$.

(5) $a_{0,+1}a_{1,-1} + a_{0,-1}a_{1,+1} > 0$.

(6) *If* $a_{1,-1} > 0$*, then* $\mathcal{U}(f^*) > -\frac{a_{0,-1}}{a_{1,-1}}$.

Moreover, assume that there exists $\tau \in (0,1)$ such that τ satisfies the following conditions.

(a) ϕ *is τ-discrepant.*

(b) $\mathcal{U}_\phi^*$ *satisfies*

$$\tau \leq \frac{a_{0,+1} - a_{1,+1}}{a_{1,-1} + a_{0,-1}} \cdot \frac{-a_{0,-1} + a_{1,-1}\mathcal{U}_\phi^*}{a_{0,+1} + a_{1,+1}\mathcal{U}_\phi^*}.$$

(c) $\mathcal{U}_\phi^*$ *and* $\mathcal{U}(f^*)$ *satisfy*

$$\tau \leq \frac{a_{0,-1} - a_{1,-1}\mathcal{U}(f^*)}{a_{1,+1}\mathcal{U}(f^*) - a_{0,+1}} \cdot \frac{a_{0,+1} + a_{1,+1}\mathcal{U}_\phi^*}{-a_{0,-1} + a_{1,-1}\mathcal{U}_\phi^*}.$$

Then, the surrogate utility $\mathcal{U}_\phi$ *is* $\mathcal{U}$*-calibrated.*

The conditions (1), (2), (3), (4), and (5) exclude *pathological* true utilities which cannot be handled by the Bayes optimal analysis. For instance, the Bayes optimal rule would be a classifier that always outputs positive values without the conditions (1) and (2); on the other hand, the Bayes optimal rule would be a classifier that always outputs negative values without the condition (3). The conditions (6), (a), (b), and (c) force the surrogate utility $\mathcal{U}_\phi$ to be calibrated to $\mathcal{U}$.

Below, we give the proof of Theorem 17.

Proof of Theorem 17. We focus on the following surrogate utility $\mathcal{U}_\phi$ as in Eq. (4):

$$\begin{aligned}\mathcal{U}_\phi(f) &= \frac{\int_{\mathcal{X}} \{a_{0,+1}(1-\phi(-f(x)))\eta(x) + a_{0,-1}\phi(f(x))(1-\eta(x)) + b_0\} p(x)\mathrm{d}x}{\int_{\mathcal{X}} \{a_{1,+1}(1+\phi(-f(x)))\eta(x) + a_{1,-1}\phi(f(x))(1-\eta(x)) + b_1\} p(x)\mathrm{d}x} \\ &= \frac{\mathbb{E}_X[W_{0,\phi}(f(X),\eta(X))]}{\mathbb{E}_X[W_{1,\phi}(f(X),\eta(X))]},\end{aligned}$$

where

$$\begin{aligned}W_{0,\phi}(\xi,q) &\doteq a_{0,+1}(1-\phi(-\xi))q + a_{0,-1}\phi(\xi)(1-q) + b_0, \\ W_{1,\phi}(\xi,q) &\doteq a_{1,+1}(1+\phi(-\xi))q + a_{1,-1}\phi(\xi)(1-q) + b_1.\end{aligned}$$

Proposition 14 tells us that the Bayes optimal set $\mathcal{B}$ for the utility $\mathcal{U}$ is

$$\mathcal{B} = \{f \mid f(x)\{(\Delta a_0 - \Delta a_1\mathcal{U}(f))\eta(x) - (a_{1,-1}\mathcal{U}(f) - a_{0,-1})\} > 0 \ \forall x \in \mathcal{X}\}.$$

We prove $\mathcal{U}$-calibration by contradiction. Assume that

$$\sup_{f \notin \mathcal{B}} \mathcal{U}_\phi(f) = \sup_f \mathcal{U}_\phi(f).$$

This implies that there exists $\bar{x} \in \mathcal{X}$ such that $f^*(\bar{x})\{(\Delta a_0 - \Delta a_1\mathcal{U}(f^*))\eta(\bar{x}) - (a_{1,-1}\mathcal{U}(f^*) - a_{0,-1})\} \geq 0$.

Let us describe the *stationary condition* of $\mathcal{U}_\phi$ at f^* in the same manner as the proof of Theorem 9. We introduce a function δf:

$$\delta f(x) \doteq \begin{cases} 1 & \text{if } x = \bar{x}, \\ 0 & \text{if } x \neq \bar{x}. \end{cases}$$

Let $G(\gamma) \doteq \mathcal{U}_\phi(f^* + \gamma\delta f)$, then the stationary condition is $G'(0) = 0$. Here, $G'(0)$ is computed as

$$\begin{aligned}G'(0) &= \frac{1}{\mathbb{E}[W_{1,\phi}(f^*)]^2} \\ &\quad \cdot \Big\{\mathbb{E}[W_{1,\phi}(f^*)] \int_{\mathcal{X}} (-a_{0,+1}\phi'(f^*(x))\eta(x) - a_{0,-1}\phi'(-f^*(x))(1-\eta(x)))\delta f(x)p(x)\mathrm{d}x \\ &\qquad - \mathbb{E}[W_{0,\phi}(f^*)] \int_{\mathcal{X}} (a_{1,+1}\phi'(f^*(x))\eta(x) - a_{1,-1}\phi'(-f^*(x))(1-\eta(x)))\delta f(x)p(x)\mathrm{d}x\Big\} \\ &= \frac{1}{\mathbb{E}[W_{1,\phi}(f^*)]} \Big\{\int_{\mathcal{X}} (-a_{0,+1}\phi'(f^*(x))\eta(x) - a_{0,-1}\phi'(-f^*(x))(1-\eta(x)))\delta f(x)p(x)\mathrm{d}x \\ &\qquad - \mathcal{U}_\phi^* \int_{\mathcal{X}} (a_{1,+1}\phi'(f^*(x))\eta(x) - a_{1,-1}\phi'(-f^*(x))(1-\eta(x)))\delta f(x)p(x)\mathrm{d}x\Big\} \\ &= \frac{1}{\mathbb{E}[W_{1,\phi}(f^*)]} \{(-a_{0,+1}\phi'(f^*(\bar{x}))\eta(\bar{x}) - a_{0,-1}\phi'(-f^*(\bar{x}))(1-\eta(\bar{x}))) \\ &\qquad - \mathcal{U}_\phi^*(a_{1,+1}\phi'(f^*(\bar{x}))\eta(\bar{x}) - a_{1,-1}\phi'(-f^*(\bar{x}))(1-\eta(\bar{x})))\},\end{aligned}$$

where $\mathbb{E}[W_{0,\phi}(f^*)] = \mathbb{E}_X[W_{0,\phi}(f^*(X), \eta(X))]$ and $\mathbb{E}[W_{1,\phi}(f^*)] = \mathbb{E}_X[W_{1,\phi}(f^*(X), \eta(X))]$. Thus, the stationary condition is

$$\begin{aligned} &-a_{0,+1}\phi'(f^*(\bar{x}))\eta(\bar{x}) - a_{0,-1}\phi'(-f^*(\bar{x}))(1-\eta(\bar{x})) \\ &-a_{1,+1}\phi'(f^*(\bar{x}))\eta(\bar{x})\mathcal{U}_\phi^* + a_{1,-1}\phi'(-f^*(\bar{x}))(1-\eta(\bar{x})) = 0, \end{aligned}$$

which is equivalent to

$$\eta(\bar{x}) = \frac{(-a_{0,-1} + a_{1,-1}\mathcal{U}_\phi^*)\phi'(-f^*(\bar{x}))}{(a_{0,+1} - a_{1,+1}\mathcal{U}_\phi^*)\phi'(f^*(\bar{x})) + (-a_{0,-1} + a_{1,-1}\mathcal{U}_\phi^*)\phi'(-f^*(\bar{x}))} \quad (\doteq \bar{\eta}_{\mathsf{STA}}). \tag{14}$$

From now on, we divide the cases to take care of the Bayes optimal condition $f^*(\bar{x})\{(\Delta a_0 - \Delta a_1 \mathcal{U}(f^*))\eta(\bar{x}) - (a_{1,-1}\mathcal{U}(f^*) - a_{0,-1})\} \geq 0$. Since $\Delta a_0 - \Delta a_1\mathcal{U}(f^*) > 0$ due to $\Delta a_0 > 0$ and $\Delta a_1 \leq 0$, the Bayes optimal condition can be rewritten as $f^*(\bar{x})\{\eta(\bar{x}) - \frac{a_{1,-1}\mathcal{U}(f^*) - a_{0,-1}}{\Delta a_0 - \Delta a_1 \mathcal{U}(f^*)}\} \geq 0$.

1) <u>If $f^*(\bar{x}) > 0$ and $\eta(\bar{x}) < \frac{a_{1,-1}\mathcal{U}(f^*) - a_{0,-1}}{\Delta a_0 - \Delta a_1 \mathcal{U}(f^*)}$:</u>

Let $\bar{\eta}_{\mathsf{OPT}} \doteq \frac{a_{1,-1}\mathcal{U}(f^*) - a_{0,-1}}{\Delta a_0 - \Delta a_1 \mathcal{U}(f^*)}$. Note that $a_{0,-1} - a_{1,-1}\mathcal{U}(f^*) < 0$ and $-a_{0,-1} + a_{1,-1}\mathcal{U}_\phi^* > 0$ since $a_{0,-1} \leq 0$, $a_{1,-1} \geq 0$, and either $a_{0,-1}$ or $a_{1,-1}$ is non-zero (condition (3)). We show the contradiction $\bar{\eta}_{\mathsf{OPT}} \leq \bar{\eta}_{\mathsf{STA}}$, which can be transformed as follows.

$$\begin{aligned} \bar{\eta}_{\mathsf{OPT}} \leq \bar{\eta}_{\mathsf{STA}} &\iff \frac{1}{1 + \frac{a_{1,+1}\mathcal{U}(f^*) - a_{0,+1}}{a_{0,-1} - a_{1,-1}\mathcal{U}(f^*)}} \leq \frac{1}{1 + \frac{a_{0,+1} + a_{1,+1}\mathcal{U}_\phi^*}{-a_{0,-1} + a_{1,-1}\mathcal{U}_\phi^*} \cdot \frac{\phi'(f^*(\bar{x}))}{\phi'(-f^*(\bar{x}))}}, \\ &\iff \frac{a_{1,+1}\mathcal{U}(f^*) - a_{0,+1}}{a_{0,-1} - a_{1,-1}\mathcal{U}(f^*)} \geq \underbrace{\frac{a_{0,+1} + a_{1,+1}\mathcal{U}_\phi^*}{-a_{0,-1} + a_{1,-1}\mathcal{U}_\phi^*}}_{>0} \cdot \frac{\phi'(f^*(\bar{x}))}{\phi'(-f^*(\bar{x}))}. \\ &\iff \underbrace{\frac{a_{1,+1}\mathcal{U}(f^*) - a_{0,+1}}{a_{0,-1} - a_{1,-1}\mathcal{U}(f^*)}}_{\doteq H(\mathcal{U}(f^*))} \cdot \frac{-a_{0,-1} + a_{1,-1}\mathcal{U}_\phi^*}{a_{0,+1} + a_{1,+1}\mathcal{U}_\phi^*} \geq \frac{\phi'(f^*(\bar{x}))}{\phi'(-f^*(\bar{x}))}. \end{aligned} \tag{15}$$

If $a_{1,-1} \neq 0$, we have

$$H(t) = \frac{\frac{a_{0,+1}a_{1,-1} + a_{0,-1}a_{1,+1}}{a_{1,-1}^2}}{t + \frac{a_{0,-1}}{a_{1,-1}}} - \frac{a_{1,+1}}{a_{1,-1}}.$$

Since $\frac{a_{0,+1}a_{1,-1} + a_{0,-1}a_{1,+1}}{a_{1,-1}^2} > 0$, H is monotonically decreasing on $-\frac{a_{0,-1}}{a_{1,-1}} < t \leq 1$. Together with the assumption $\mathcal{U}(f^*) > -\frac{a_{0,-1}}{a_{1,-1}}$, we have $H(\mathcal{U}(f^*)) \geq H(1) = \frac{a_{0,+1} - a_{1,+1}}{a_{1,-1} - a_{0,-1}}$.

If $a_{1,-1} = 0$, $H(t) = \frac{a_{1,+1}}{a_{0,-1}}t - \frac{a_{0,+1}}{a_{0,-1}}$, noting that either $a_{0,-1}$ or $a_{a,-1}$ is non-zero (condition (3)). Here, we have $H(\mathcal{U}(f^*)) \geq H(1)$ as well since H is a decreasing linear function.

Since ϕ is τ-discrepant and τ satisfies the condition (b),

$$\begin{aligned} \frac{\phi'(f^*(\bar{x}))}{\phi'(-f^*(\bar{x}))} &\leq \tau \\ &\leq \underbrace{\frac{a_{0,+1} - a_{1,+1}}{a_{1,-1} + a_{0,-1}}}_{=H(1)} \cdot \frac{-a_{0,-1} + a_{1,-1}\mathcal{U}_\phi^*}{a_{0,+1} + a_{1,+1}\mathcal{U}_\phi^*} \qquad \text{(using (b))} \\ &\leq \underbrace{\frac{a_{1,+1}\mathcal{U}(f^*) - a_{0,+1}}{a_{0,-1} - a_{1,-1}\mathcal{U}(f^*)}}_{=H(\mathcal{U}(f^*))} \cdot \frac{-a_{0,-1} + a_{1,-1}\mathcal{U}_\phi^*}{a_{0,+1} + a_{1,+1}\mathcal{U}_\phi^*}, \end{aligned}$$

which concludes Eq. (15) and $\bar{\eta}_{\mathsf{OPT}} \leq \bar{\eta}_{\mathsf{STA}}$ (contradiction).

2) If $f^*(\bar{x}) \le 0$ and $\eta(\bar{x}) > \frac{a_{1,-1}\mathcal{U}(f^*) - a_{0,-1}}{\Delta a_0 - \Delta a_1 \mathcal{U}(f^*)}$:

We show the contradiction $\bar{\eta}_{\mathsf{OPT}} \ge \bar{\eta}_{\mathsf{STA}}$, which can be transformed in the same way as Eq. (15) as follows.

$$\begin{aligned} &\frac{a_{1,+1}\mathcal{U}(f^*) - a_{0,+1}}{a_{0,-1} - a_{1,-1}\mathcal{U}(f^*)} \cdot \frac{-a_{0,-1} + a_{1,-1}\mathcal{U}^*_\phi}{a_{0,+1} + a_{1,+1}\mathcal{U}^*_\phi} \le \frac{\phi'(f^*(\bar{x}))}{\phi'(-f^*(\bar{x}))}. \\ \iff &\frac{a_{0,-1} - a_{1,-1}\mathcal{U}(f^*)}{a_{1,+1}\mathcal{U}(f^*) - a_{0,+1}} \cdot \frac{a_{0,+1} + a_{1,+1}\mathcal{U}^*_\phi}{-a_{0,-1} + a_{1,-1}\mathcal{U}^*_\phi} \ge \frac{\phi'(-f^*(\bar{x}))}{\phi'(f^*(\bar{x}))}. \end{aligned} \tag{16}$$

Note that $a_{1,+1}\mathcal{U}(f^*) - a_{0,+1} > 0$ and $-a_{0,-1} + a_{1,-1}\mathcal{U}^*_\phi > 0$ since $a_{0,+1} \ge 0$, $a_{0,-1} \le 0$, $a_{1,+1} \ge 0$, and $a_{1,-1} \ge 0$. Since ϕ is τ-discrepant and τ satisfies the condition (c),

$$\frac{\phi'(-f^*(\bar{x}))}{\phi'(f^*(\bar{x}))} \le \tau \le \frac{a_{0,-1} - a_{1,-1}\mathcal{U}(f^*)}{a_{1,+1}\mathcal{U}(f^*) - a_{0,+1}} \cdot \frac{a_{0,+1} + a_{1,+1}\mathcal{U}^*_\phi}{-a_{0,-1} + a_{1,-1}\mathcal{U}^*_\phi}, \qquad \text{(using (c))}$$

which concludes Eq. (16) and $\bar{\eta}_{\mathsf{OPT}} \ge \bar{\eta}_{\mathsf{STA}}$ (contradiction).

Combining the above cases, it follows that

$$\sup_{f \notin \mathcal{B}} \mathcal{U}_\phi(f) < \sup_f \mathcal{U}_\phi(f).$$

Eventually, we claim that $\mathcal{U}_\phi$ is $\mathcal{U}$-calibrated using Proposition 14. □

A.5 Non-negativity of Optimal Surrogate Utilities

Here, we briefly discuss that the optimal surrogate utilities are non-negative even though the numerator can be negative. Let us focus on the F_β case:

$$\begin{aligned} W_{0,\phi}(\xi, q) &= (1+\beta^2)(1-\phi(\xi))q, \\ W_{1,\phi}(\xi, q) &= (1+\phi(\xi))q + \phi(-\xi)(1-q) + \beta^2\pi, \\ \mathcal{U}_\phi(f) &= \frac{\mathbb{E}_X W_{0,\phi}(f(X), \eta(X))}{\mathbb{E}_X W_{1,\phi}(f(X), \eta(X))}, \end{aligned}$$

and let f^* and $\check{f}$ be suprema of $\mathcal{U}_\phi$ and $\mathbb{E}_X[W_0(f(X), \eta(X))]$ in f within all measurable functions, respectively. Then,

$$\begin{aligned} \mathcal{U}_\phi(f^*) \ge \mathcal{U}_\phi(\check{f}) &= \frac{\sup_{f'} \mathbb{E}_X W_{0,\phi}(f'(X), \eta(X))}{\mathbb{E}_X[W_{1,\phi}(\check{f}(X), \eta(X))]} \\ &\overset{(a)}{=} \frac{\mathbb{E}_X[H_{0,\phi}(\eta(X))]}{\mathbb{E}_X[W_{1,\phi}(\check{f}(X), \eta(X))]} \\ &= \frac{(1+\beta^2)\pi}{\mathbb{E}_X[W_{1,\phi}(\check{f}(X), \eta(X))]} \\ &\ge 0, \end{aligned}$$

where $H_{0,\phi}(q) \doteq \sup_{\xi \in \mathbb{R}} W_{0,\phi}(\xi, q)$. The equality (a) holds under a certain regularity condition (Steinwart, 2007, Lemma 2.5). Hence, we confirm that the optimal value of $\mathcal{U}_\phi$ is non-negative.

The same discussion holds for the Jaccard case.

B Proof of Quasi-concavity of the Surrogate Utility

Proof of Lemma 5. Define an α-super-level set of $\widehat{\mathcal{U}}_\phi$ restricted in $\bar{\mathcal{F}}$ as $\mathcal{A}_\alpha \doteq \{f \in \bar{\mathcal{F}} \mid \widehat{\mathcal{U}}_\phi(f) \geq \alpha\}$. It is enough to show $\mathcal{A}_\alpha$ is a convex set for any $\alpha \geq 0$ owing to $f \in \bar{\mathcal{F}}$.

Fix any $\alpha \geq 0$. Then,

$$\begin{aligned}\widehat{\mathcal{U}}_\phi(f) \geq \alpha &\iff \frac{\frac{1}{m}\sum_{i=1}^m \widetilde{W}_{0,\phi}(f(x_i), y_i)}{\frac{1}{n-m}\sum_{j=m+1}^n \widetilde{W}_{1,\phi}(f(x_j), y_j)} \geq \alpha \\ &\iff \underbrace{\frac{1}{m}\sum_{i=1}^m \widetilde{W}_{0,\phi}(f(x_i), y_i) - \alpha\frac{1}{n-m}\sum_{j=m+1}^n \widetilde{W}_{1,\phi}(f(x_j), y_j)}_{(*)} \geq 0.\end{aligned}$$

Here, $\frac{1}{m}\sum_{i=1}^m \widetilde{W}_{0,\phi}(f(x_i), y_i)$ is concave in f since it is a non-negative sum of concave functions. Note that $\widetilde{W}_{0,\phi}(f(x_i), y_i)$ is concave in f for any (x_i, y_i) due to the definition of $\widetilde{W}_{0,\phi}$ in Eq. (3) and the assumption ϕ is convex. Similarly, $\frac{1}{n-m}\sum_{j=m+1}^n \widetilde{W}_{1,\phi}(f(x_j), y_j)$ is convex as well. Thus, $(*)$ is concave in f, which means that $\mathcal{A}_\alpha$ is a convex set since any super-level sets of a concave function is convex.

Hence, we confirm that $\mathcal{A}_\alpha$ is convex for any $\alpha \geq 0$. □

C Proof of Uniform Convergence

First, we need carefully analyze our *non-smooth* surrogate loss to take handle of the Rademacher complexity (Bartlett and Mendelson, 2002), which is defined as follows.

Definition 18 (Rademacher complexity)**.** *Let $\mathcal{S} \doteq \{z_1, \ldots, z_n\}$ be a sample with size n. Let $\mathcal{G} \doteq \{g \mid \mathcal{Z} \to \mathbb{R}\}$ be a class of measurable functions, and $\sigma \doteq (\sigma_1, \ldots, \sigma_n)$ be the Rademacher variables, that is, random variables taking $+1$ and -1 with even probabilities. Then, the Rademacher complexity of $\mathcal{G}$ of the sample size n is defined as*

$$\mathfrak{R}_n(\mathcal{G}) \doteq \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(Z_i)\right|\right].$$

Usually, we analyze the Rademacher complexity of the composite function class $\phi \circ \mathcal{F} \doteq \{(x, y) \mapsto \phi(yf(x)) \mid f \in \mathcal{F}\}$ by using the Ledoux-Talagrand's contraction inequality (Ledoux and Talagrand, 1991) when the surrogate ϕ is Lipschitz continuous: $\mathfrak{R}_n(\phi \circ \mathcal{F}) \le 2\rho_\phi \mathfrak{R}_n(\mathcal{F})$, where ρ_ϕ is the Lipschitz norm of ϕ. On the other hand, we need to deal with the case of the uniform convergence of gradients, which requires smoothness of the surrogate, while τ-discrepant loss is non-smooth surrogates. Thus, we need an alternative analysis.

Lemma 19. *Assume that ϕ is τ-discrepant and can be decomposed as $\phi(m) = \phi_{+1}(m)\mathbb{1}_{\{m>0\}} + \phi_{-1}(m)\mathbb{1}_{\{m\le 0\}}$. For $k = 0, 1$, denote $\widetilde{W}'_{k,\phi} \circ \mathcal{F} \doteq \{(x, y) \mapsto \widetilde{W}'_{k,\phi}(f(x), y) \mid f \in \mathcal{F}\}$. Then,*

$$\mathfrak{R}_n(\widetilde{W}'_{k,\phi} \circ \mathcal{F}) \le 2(\gamma_{+1} + \gamma_{-1})\mathfrak{R}_n(\mathcal{F}).$$

Proof. First, we prove for $k = 0$. Note that $\widetilde{W}'_{0,\phi}(f(x), +1) = (1 - \phi(f(x)))' = -\phi'(f(x))$, and that $\widetilde{W}'_{0,\phi}(f(x), -1) = (-\phi(-f(x)))' = \phi'(-f(x))$, thus, $\widetilde{W}'_{0,\phi}(f(x), y) = -y\phi'(yf(x))$.

$$\begin{aligned}
&\mathfrak{R}_n(\widetilde{W}'_{0,\phi} \circ \mathcal{F}) \\
&\quad= \mathbb{E}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\widetilde{W}'_{0,\phi}(f(x_i), y_i)\right|\right] \\
&\quad= \mathbb{E}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i(-y_i\phi'(y_i f(x_i)))\right|\right] \\
&\quad= \mathbb{E}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\phi'(y_i f(x_i))\right|\right] \\
&\qquad\quad (\because \sigma_i \text{ and } -\sigma_i y_i \text{ are distributed in the same way for a fixed } y_i) \\
&\quad= \mathbb{E}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left\{\phi'_{-1}(y_i f(x_i))\mathbb{1}_{\{y_i f(x_i)\le 0\}} + \phi'_{+1}(y_i f(x_i))\mathbb{1}_{\{y_i f(x_i)>0\}}\right\}\right|\right] \\
&\quad\le \underbrace{\mathbb{E}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\phi'_{-1}(y_i f(x_i))\mathbb{1}_{\{y_i f(x_i)\le 0\}}\right|\right]}_{\text{(A)}} \\
&\qquad\quad + \underbrace{\mathbb{E}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\phi'_{+1}(y_i f(x_i))\mathbb{1}_{\{y_i f(x_i)>0\}}\right|\right]}_{\text{(B)}},
\end{aligned}$$

where the last inequality is just the triangular inequality. For (A), let $\psi_{-1}(m) \doteq \phi'_{-1}(m)\frac{m}{|m|}$ if $m \ne 0$, and

$\psi_{-1}(0) \doteq 0$. Since

$$\begin{aligned}
\psi'_{-1}(m) &= \frac{(\psi'_{-1}(m)m)'|m| - \phi'_{-1}(m)m \cdot (|m|)'}{m^2} \\
&= \frac{\phi''_{-1}(m)m|m| + \phi'_{-1}(m)|m| - \phi'_{-1}(m)m \cdot \frac{m}{|m|}}{m^2} \\
&= \frac{\phi''_{-1}(m)m^3 + \phi'_{-1}(m)m^2 - \phi'_{-1}(m)m^2}{m^2|m|} \\
&= \phi''_{-1}(m)\frac{m}{|m|},
\end{aligned}$$

the Lipschitz norm of ψ_{-1} can be computed as

$$\begin{aligned}
\sup_{f\in\mathcal{F},(x,y)\in\mathcal{X}\times\mathcal{Y}} |\psi'_{-1}(f(x))| &= \sup_{f,x,y}\left|\phi''_{-1}(yf(x))\frac{yf(x)}{|yf(x)|}\right| \\
&= \sup_{f,x,y}|\phi''_{-1}(yf(x))| \cdot \sup_{f,x,y}\left|\frac{yf(x)}{|yf(x)|}\right| \\
&= \gamma_{-1}.
\end{aligned}$$

Note that the Lipschitz norm of ϕ'_{-1} is γ_{-1} because ϕ_{-1} is γ_{-1}-smooth. Then, we further bound (A) by using the fact $\mathbb{1}_{\{y_i f(x_i)\le 0\}} = \frac{1-\frac{y_if(x_i)}{|y_if(x_i)|}}{2}$.

$$\begin{aligned}
\text{(A)} &= \mathbb{E}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\phi'_{-1}(y_if(x_i))\frac{1-\frac{y_if(x_i)}{|y_if(x_i)|}}{2}\right|\right] \\
&\le \frac{1}{2}\mathbb{E}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\phi'_{-1}(y_if(x_i))\right|\right] + \frac{1}{2}\mathbb{E}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\phi'_{-1}(y_if(x_i))\frac{y_if(x_i)}{|y_if(x_i)|}\right|\right] \\
&\qquad (\because \text{triangular inequality}) \\
&= \frac{1}{2}\mathbb{E}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\phi'_{-1}(y_if(x_i))\right|\right] + \frac{1}{2}\mathbb{E}_{\mathcal{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\psi_{-1}(y_if(x_i))\frac{y_if(x_i)}{|y_if(x_i)|}\right|\right] \\
&= \frac{1}{2}\mathfrak{R}_n(\phi'_{-1}\circ\mathcal{F}) + \frac{1}{2}\mathfrak{R}_n(\psi_{-1}\circ\mathcal{F}) \\
&\le \frac{1}{2}\cdot 2\gamma_{-1}\mathfrak{R}_n(\mathcal{F}) + \frac{1}{2}\cdot 2\gamma_{-1}\mathfrak{R}_n(\mathcal{F}) \\
&= 2\gamma_{-1}\mathfrak{R}_n(\mathcal{F}),
\end{aligned}$$

where the inequality is the result of the Ledoux-Talagrand's contraction inequality (Ledoux and Talagrand, 1991, Theorem 4.12). Note that both ϕ'_{-1} and ψ_{-1} are γ_{-1}-Lipschitz. We can prove that (B) is bounded by $\gamma_{+1}\mathfrak{R}_n(\mathcal{F})$ from the above as well. Therefore, the claim is supported. We can prove the case $k = 1$ in the same manner. □

Now, we move on to the proof of Lemma 11.

Proof of Lemma 11. We write $\mathcal{V}_\phi(f_\theta)$ as $\mathcal{V}_\phi(\theta)$. If we explicit note for which sample we take the empirical average in $\widehat{\mathcal{V}}_\phi(\theta)$, let us write $\widehat{\mathcal{V}}_\phi(\theta;\mathcal{S})$. Let $\mathcal{E}(\mathcal{S}) \doteq \sup_{\theta\in\Theta}\|\widehat{\mathcal{V}}_\phi(\theta;\mathcal{S}) - \mathcal{V}_\phi(\theta)\|$. For simplicity, we write $z_i \doteq (x_i, y_i)$ and $\widetilde{W}_{0,\phi}(\theta; z_i) \doteq \widetilde{W}_{0,\phi}(f_\theta(x_i), y_i)$. First, we observe $\mathcal{E}(\mathcal{S})$ admits the bounded difference property (McDiarmid, 1989).

Denote that $\mathcal{S} \doteq \{z_i\}_{i=1}^n$ and $\mathcal{S}' \doteq \{z_1, \dots, z_k', \dots, z_n\}$. If $1 \le k \le m$,

$$
\begin{aligned}
&\sup_{\mathcal{S} \subset \mathcal{X} \times \mathcal{Y}, z_k' \in \mathcal{X} \times \mathcal{Y}} |\mathcal{E}(\mathcal{S}) - \mathcal{E}(\mathcal{S}')| \\
&\doteq \sup_{\mathcal{S}, z_k'} \left| \sup_{\theta \in \Theta} \|\widehat{\mathcal{V}}_\phi(\theta; \mathcal{S}) - \mathcal{V}_\phi(\theta)\| - \sup_{\theta \in \Theta} \|\widehat{\mathcal{V}}_\phi(\theta; \mathcal{S}') - \mathcal{V}_\phi(\theta)\| \right| \\
&\le \sup_{\mathcal{S}, z_k', \theta} \|\widehat{\mathcal{V}}_\phi(\theta; \mathcal{S}) - \widehat{\mathcal{V}}_\phi(\theta; \mathcal{S}')\| \qquad (\because \text{triangular inequality}) \\
&= \frac{1}{m(n-m)} \sup_{\mathcal{S}, z_k', \theta} \Bigg\| \{\nabla \widetilde{W}_{0,\phi}(\theta; z_k) - \nabla \widetilde{W}_{0,\phi}(\theta; z_k')\} \sum_{j=m+1}^n \widetilde{W}_{1,\phi}(\theta; z_j) \\
&\qquad - \{\widetilde{W}_{0,\phi}(\theta; z_k) - \widetilde{W}_{0,\phi}(\theta; z_k')\} \sum_{j=m+1}^n \nabla \widetilde{W}_{1,\phi}(\theta; z_j) \Bigg\| \\
&\le \frac{1}{m(n-m)} \sup_{\mathcal{S}, z_k', \theta} \Bigg\{ \left(\|\nabla \widetilde{W}_{0,\phi}(\theta; z_k)\| + \|\nabla \widetilde{W}_{0,\phi}(\theta; z_k')\| \right) \sum_{j=m+1}^n |\widetilde{W}_{1,\phi}(\theta; z_j)| \\
&\qquad + \left(|\widetilde{W}_{0,\phi}(\theta; z_k)| + |\widetilde{W}_{0,\phi}(\theta; z_k')| \right) \sum_{j=m+1}^n \|\nabla \widetilde{W}_{1,\phi}(\theta; z_j)\| \Bigg\} \\
&\le \frac{2\rho_0 c_{\mathcal{X}} \cdot (n-m) c_1 + 2c_0 \cdot (n-m) \rho_1 c_{\mathcal{X}}}{m(n-m)} \\
&= \frac{4c_{\mathcal{X}}(\rho_1 c_0 + \rho_0 c_1)}{n},
\end{aligned}
$$

where the second inequality also holds due to the triangular inequality, and the last inequality follows from the fact that $\widetilde{W}_{0,\phi}$ and $\widetilde{W}_{1,\phi}$ are ρ_0-/ρ_1-Lipschitz and bounded by c_0 and c_1, respectively. The same holds for the case $m+1 \le k \le n$.

Thus, $\mathcal{E}$ is the bounded difference with a constant $(4c_{\mathcal{X}}(\rho_1 c_0 + \rho_0 c_1))/n$ for each index, and we can obtain the following inequality by McDiarmid's inequality (McDiarmid, 1989):

$$
\mathbb{P}[\mathcal{E}(\mathcal{S}) - \mathbb{E}_{\mathcal{S}}[\mathcal{E}(\mathcal{S})] > \epsilon] \le 2\exp\left(-\frac{n\epsilon^2}{8c_{\mathcal{X}}^2(\rho_1 c_0 + \rho_0 c_1)^2} \right),
$$

which is equivalent to

$$
\mathcal{E}(\mathcal{S}) - \mathbb{E}_{\mathcal{S}}[\mathcal{E}(\mathcal{S})] \le \sqrt{\frac{8c_{\mathcal{X}}^2(\rho_1 c_0 + \rho_0 c_1)^2 \log \frac{2}{\delta}}{n}},
$$

with probability at least $1-\delta$.

Next, we bound $\mathbb{E}_{\mathcal{S}}[\mathcal{E}(\mathcal{S})]$ by the *symmetrization device* (Ledoux and Talagrand, 1991, Lemma 6.3).

$$
\begin{aligned}
\mathbb{E}_{\mathcal{S}}[\mathcal{E}(\mathcal{S})] &= \mathbb{E}_{\mathcal{S}}\left[\sup_{\theta \in \Theta} \|\widehat{\mathcal{V}}_\phi(\theta; \mathcal{S}) - \mathcal{V}_\phi(\theta)\| \right] \\
&\le \underbrace{\mathbb{E}_{\mathcal{S}} \sup_\theta \left\| \frac{1}{m(n-m)} \sum_{i=1}^m \sum_{j=m+1}^n \widetilde{W}_{1,\phi}(\theta; z_j) \nabla \widetilde{W}_{0,\phi}(\theta; z_i) - \mathbb{E}[W_{1,\phi} \nabla W_{0,\phi}] \right\|}_{\text{(A)}} \\
&\quad + \underbrace{\mathbb{E}_{\mathcal{S}} \sup_\theta \left\| \frac{1}{m(n-m)} \sum_{i=1}^m \sum_{j=m+1}^n \widetilde{W}_{0,\phi}(\theta; z_i) \nabla \widetilde{W}_{1,\phi}(\theta; z_j) - \mathbb{E}[W_{0,\phi} \nabla W_{1,\phi}] \right\|}_{\text{(B)}}, \qquad (17)
\end{aligned}
$$

where the second line is the result of the triangular inequality, and

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}}[(\mathrm{A})] \\
&= \mathbb{E}_{\mathcal{S}} \sup_{\theta} \left\| \frac{1}{m(n-m)} \sum_{i=1}^{m} \sum_{j=m+1}^{n} \widetilde{W}_{1,\phi}(\theta; z_j) \left(\nabla \widetilde{W}_{0,\phi}(\theta; z_i) - \mathbb{E}[\nabla W_{0,\phi}] \right) \right. \\
&\qquad \left. + \frac{1}{m(n-m)} \sum_{i=1}^{m} \sum_{j=m+1}^{n} \mathbb{E}[\nabla W_{0,\phi}] \left(\widetilde{W}_{1,\phi}(\theta; z_j) - \mathbb{E}[W_{1,\phi}] \right) \right\| \\
&\le \mathbb{E}_{\mathcal{S}} \sup_{\theta} \left\{ \frac{1}{m(n-m)} \sum_{j=m+1}^{n} |\widetilde{W}_{1,\phi}(\theta; z_j)| \cdot \left\| \sum_{i=1}^{m} \nabla \widetilde{W}_{0,\phi}(\theta; z_i) - \mathbb{E}[\nabla W_{0,\phi}] \right\| \right. \\
&\qquad \left. + \frac{1}{m(n-m)} \sum_{i=1}^{m} \|\mathbb{E}[\nabla W_{0,\phi}]\| \cdot \left| \sum_{j=m+1}^{n} \widetilde{W}_{1,\phi}(\theta; z_j) - \mathbb{E}[W_{1,\phi}] \right| \right\} \\
&\le \mathbb{E}_{\mathcal{S}} \left[\sup_{\theta} \left\{ c_1 \left\| \frac{1}{m} \sum_{i=1}^{m} \nabla \widetilde{W}_{0,\phi}(\theta; z_i) - \mathbb{E}[\nabla W_{0,\phi}] \right\| + \rho_0 c_{\mathcal{X}} \left| \frac{1}{n-m} \sum_{j=m+1}^{n} \widetilde{W}_{1,\phi}(\theta; z_j) - \mathbb{E}[W_{1,\phi}] \right| \right\} \right] \\
&= \underbrace{c_1 \, \mathbb{E}_{\mathcal{S}} \left[\sup_{\theta} \left\| \frac{1}{m} \sum_{i=1}^{m} \nabla \widetilde{W}_{0,\phi}(\theta; z_i) - \mathbb{E}[\nabla W_{0,\phi}] \right\| \right]}_{(\mathrm{A'})} + \rho_0 c_{\mathcal{X}} \underbrace{\mathbb{E}_{\mathcal{S}} \left[\sup_{\theta} \left| \frac{1}{n-m} \sum_{j=m+1}^{n} \widetilde{W}_{1,\phi}(\theta; z_j) - \mathbb{E}[W_{1,\phi}] \right| \right]}_{(\mathrm{A''})},
\end{aligned}
$$

where the first inequality is the triangular inequality. Now we introduce the Rademacher random variables $\sigma_{1:n} \doteq \{\sigma_1, \ldots, \sigma_n\}$ that are independently and uniformly distributed on $\{+1, -1\}$.

- For (A'), we can bound it from the above by the symmetrization device and the fact that $\|\cdot\|_2 \le \|\cdot\|_1$.

$$
\begin{aligned}
(\mathrm{A'}) &= \mathbb{E}_{\mathcal{S}} \left[\sup_{\theta} \left\| \frac{1}{m} \sum_{i=1}^{m} \nabla \widetilde{W}_{0,\phi}(\theta; z_i) - \mathbb{E}[\nabla W_{0,\phi}] \right\| \right] \\
&\le \mathbb{E}_{\mathcal{S}, \sigma_{1:m}} \left[\sup_{\theta} \sum_{l=1}^{d} \left| \frac{1}{n} \sum_{i=1}^{m} \nabla_{\theta_l} \widetilde{W}_{0,\phi}(\theta; z_i) - \mathbb{E}[\nabla_{\theta_l} W_{0,\phi}] \right| \right] && (\|\cdot\|_2 \le \|\cdot\|_1) \\
&\le \sum_{l=1}^{d} \mathbb{E}_{\mathcal{S}, \sigma_{1:m}} \left[\sup_{\theta} \left| \frac{2}{m} \sum_{i=1}^{m} \sigma_i \nabla_{\theta_l} \widetilde{W}_{0,\phi}(\theta; z_i) \right| \right] && (\text{symmetrization device}) \\
&= \sum_{l=1}^{d} 2 \mathbb{E}_{\mathcal{S}, \sigma_{1:m}} \left[\sup_{\theta} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \widetilde{W}'_{0,\phi}(\theta; z_i) \cdot x_l \right| \right] && (f_\theta(x) = \theta^\top x) \\
&\le \sum_{l=1}^{d} 2 \mathbb{E}_{\mathcal{S}, \sigma_{1:m}} \left[\sup_{\theta} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \widetilde{W}'_{0,\phi}(\theta; z_i) \right| \cdot c_{\mathcal{X}} \right] && (|x_l| \le \|x\| \le c_{\mathcal{X}} \ \forall x \in \mathcal{X}) \\
&\le 4 d c_{\mathcal{X}} (\gamma_{+1} + \gamma_{-1}) \mathfrak{R}_m(\mathcal{F}) \\
&= 4 d c_{\mathcal{X}} (\gamma_{+1} + \gamma_{-1}) \mathfrak{R}_{n/2}(\mathcal{F}),
\end{aligned}
$$

where the last inequality uses Lemma 19.

- For (A”), we can bound it from the above by the symmetrization device.

$$
\begin{aligned}
\text{(A”)} &= \mathbb{E}_{\mathcal{S}}\left[\sup_{\theta}\left|\frac{1}{n-m}\sum_{j=m+1}^{n}\widetilde{W}_{1,\phi}(\theta;z_j) - \mathbb{E}[W_{1,\phi}]\right|\right] \\
&\leq \mathbb{E}_{\mathcal{S},\sigma_{m:n-m}}\left[\sup_{\theta}\left|\frac{2}{n-m}\sum_{j=m+1}^{n}\sigma_j\widetilde{W}_{1,\phi}(\theta;z_i)\right|\right] && \text{(symmetrization device)} \\
&\leq 4\rho_1\mathfrak{R}_{n-m}(\mathcal{F}) && \text{(contraction inequality)} \\
&= 4\rho_1\mathfrak{R}_{n/2}(\mathcal{F}),
\end{aligned}
$$

where the second inequality uses the Ledoux-Talagrand's contraction inequality (Ledoux and Talagrand, 1991, Theorem 4.12), together with the fact that $\widetilde{W}_{1,\phi}$ is ρ_1-Lipschitz continuous.

Thus, Eq. (17) can be bounded as follows.

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}}[\mathcal{E}(\mathcal{S})] \\
&\leq c_1\text{(A')} + \rho_0 c_{\mathcal{X}}\text{(A”)} + \mathbb{E}_{\mathcal{S}}[\text{(B)}] \\
&\leq 4dc_{\mathcal{X}}c_1(\gamma_{+1}+\gamma_{-1})\mathfrak{R}_{n/2}(\mathcal{F}) + 4\rho_0\rho_1 c_{\mathcal{X}}\mathfrak{R}_{n/2}(\mathcal{F}) + \underbrace{4dc_{\mathcal{X}}c_0(\gamma_{+1}+\gamma_{-1})\mathfrak{R}_{n/2}(\mathcal{F}) + 4\rho_1\rho_0 c_{\mathcal{X}}\mathfrak{R}_{n/2}(\mathcal{F})}_{\text{can be proven in the same manner as (A)}} \\
&= (4c_{\mathcal{X}}c_0 d\gamma + 4c_{\mathcal{X}}c_1 d\gamma + 8\rho_0\rho_1 c_{\mathcal{X}})\mathfrak{R}_{n/2}(\mathcal{F}) \qquad (\gamma \doteq \gamma_{+1}+\gamma_{-1}) \\
&\leq (4c_{\mathcal{X}}c_0 d\gamma + 4c_{\mathcal{X}}c_1 d\gamma + 8\rho_0\rho_1 c_{\mathcal{X}})\frac{\sqrt{2}c_{\mathcal{X}}c_{\Theta}}{\sqrt{n}},
\end{aligned}
$$

where the last inequality comes from Mohri et al. (2012, Theorem 4.3), which results in $\mathfrak{R}_n(\mathcal{F}) = \mathfrak{R}_n(\{x \mapsto \theta^\top x \mid \theta \in \Theta\}) \leq c_{\mathcal{X}}c_{\Theta}/\sqrt{n}$.

After all, we obtain the desired uniform bound: with probability at least $1-\delta$,

$$
\begin{aligned}
&\sup_{\theta\in\Theta}\|\widehat{\mathcal{V}}_\phi(\theta;\mathcal{S}) - \mathcal{V}_\phi(\theta)\| = \mathcal{E}(\mathcal{S}) \\
&\qquad\leq \mathbb{E}_{\mathcal{S}}[\mathcal{E}(\mathcal{S})] + \frac{\sqrt{8}c_{\mathcal{X}}(\rho_1 c_0 + \rho_0 c_1)\sqrt{\log\frac{2}{\delta}}}{\sqrt{n}} \\
&\qquad\leq \frac{(4c_{\mathcal{X}}c_0 d\gamma + 4c_{\mathcal{X}}c_1 d\gamma + 8\rho_0\rho_1 c_{\mathcal{X}}) + \sqrt{8}c_{\mathcal{X}}(\rho_1 c_0 + \rho_0 c_1)\sqrt{\log\frac{2}{\delta}}}{\sqrt{n}}.
\end{aligned}
$$

□

D Experimental Results

D.1 Details of Datasets

Datasets that we use throughout this section are obtained from the *UCI Machine Learning Repository* (Lichman, 2013) and the *LIBSVM* (Chang and Lin, 2011). For those which have independent training data, validation data, and test data, all of them are merged into one dataset. We randomly split the original data with the ratio 8 : 2, and the former is used for training while the latter is used for evaluation. Each feature value is scaled between zero and one.

Table 4: Details of datasets.

Dataset	dimension	sample size	class-prior
adult	123	48842	0.239
australian	14	690	0.445
breast-cancer	10	683	0.350
cod-rna	8	331152	0.333
diabetes	8	768	0.651
german.numer	24	1000	0.300
heart	13	270	0.444
ionosphere	34	351	0.641
mushrooms	112	8124	0.482
phishing	68	11055	0.557
phoneme	5	5404	0.293
skin_nonskin	3	245057	0.208
sonar	60	208	0.394
spambase	57	4601	0.394
splice	60	1000	0.517
w8a	300	64700	0.030

D.2 Details of Baseline Methods

We describe the details of baseline methods. Baselines 2 and 3 are also mentioned in Sec. 6.

Baseline 1 (ERM): The first baseline is the vanilla empirical risk minimization, which does not optimize the metric of our interest but accuracy. The hinge loss and ℓ_2-regularization are employed with the regularization parameter 10^{-2}.

Baseline 2 (W-ERM): Weighted empirical risk minimization, or cost-sensitive empirical risk minimization, is often used to optimize non-linear performance metrics (Koyejo et al., 2014; Narasimhan et al., 2014; Parambath et al., 2014). Here, we applied a simple approach: Find a cost parameter from a given cost parameter space, which gives the maximum validation performance of a classifier trained by the cost-sensitive empirical risk minimization (Scott, 2012). The training dataset is split to 4 to 1 at random, and the latter is saved for validation of a regularization parameter. The former set is further split to 9 to 1 at random, and the former 90% is used for training the base classifier, while the latter 10% is used for the validation. As the base cost-sensitive learner, we use the hinge loss minimizer with ℓ_2-regularization (a regularization parameter is chosen from $\{10^{-1}, 10^{-3}, 10^{-5}\}$ by cross validation). The cost parameter is chosen from the range $[10^{-3}, 1-10^{-3}]$ evenly split to 20 ranges, that is, $\left\{10^{-3} + \frac{1-2\cdot10^{-3}}{20}i \mid i = 1, \ldots, 20\right\}$.

Baseline 3 (Plug-in): Plug-in estimator is one of the other common methods to optimize the non-linear performance metrics (Koyejo et al., 2014; Yan et al., 2018), which is the two-step method: To estimate the class posterior probability $\widehat{\eta}(x) = p(y = +1|x)$ first, and then to decide the optimal threshold $\widehat{\delta}$. The classifier is constructed as $x \mapsto \mathrm{sgn}(\widehat{\eta}(x) - \widehat{\delta})$. The training dataset is split to 4 to 1 at random, and the latter is saved for validation of a regularization parameter. The former set is further split to 9 to 1 at random, and they are independently used for the first and second step. For estimating $p(y = +1|x)$ (the first step), the logistic regression is used (Reid and Williamson, 2009), with ℓ_2-regularization (a regularization parameter is chosen from $\{10^{-1}, 10^{-3}, 10^{-5}\}$ by cross validation). For deciding $\widehat{\delta}$, we pick a threshold with the highest validation metric from $\left\{10^{-3} + \frac{1-2\cdot10^{-3}}{20}i \mid i = 1, \ldots, 20\right\}$.

D.3 Convergence Comparison

Figures 10 and 11 are the full version of the convergence comparison of U-GD and U-BFGS. Figure 10 shows the result of F_1-measure, and Figure 11 shows the result of Jaccard index. The vertical axes show test metric values, where the higher the better. Note that both F_1-measure and Jaccard index ranges over zero to one. The horizontal axes show the number of iterations. For each dataset, metric, and method, we ran 300 iterations to see their convergence behaviors.

Overall, U-BFGS shows faster convergence than U-GD in terms of the number of iterations. In almost all cases, U-BFGS converges within 30 iterations, except german.numer and mushrooms in Jaccard case. Moreover, it usually achieves higher performance than U-GD. U-GD convergences require at least around 100 iterations (mushrooms and phishing in F_1 case), and sometimes it does not converge even within 300 iterations such as heart and ionosphere in F_1 and Jaccard cases.

D.4 Performance Comparison with Benchmark Data

Benchmark results are shown in Tabs. 5 and 6. Each entry shows its final metric value for either F_1-measure or Jaccard index. For each dataset, we first picked the method with the highest test performance as a outperforming method within that dataset, then conducted one-sided t-test with the significant level 5%, and they are also regarded as outperforming methods if the performance differences are not significant as a result of hypothesis tests. Outperforming methods are indicated in bold-faces.

As general tendencies, we observe that U-BFGS and Plug-in work well for both F_1-measure and Jaccard index. As for F_1-measure, their performances are competitive, while U-BFGS is better as for Jaccard index. In practice, both U-BFGS and Plug-in are worth being tried.

As for other methods: ERM does not work good as we expect, because it does not optimize the metrics of our interests, F_1-measure and Jaccard index, at all. W-ERM does not work as well as Plug-in even though both of them are known to be consistent to the linear-fractional utilities. We may need more finer split of the threshold search space, or try a binary-search-type algorithm provided by recent work (Yan et al., 2018). U-GD does not work as well as U-BFGS contrary to our expectation. We may need more iterations to make U-GD converge, as we see in Figures 10 and 11. Note that we ran 100 iterations for both U-GD and U-BFGS for the results shown in Tabs. 5 and 6.

Table 5: Results of the F_1-measure: 50 trials are conducted for each pair of a method and dataset. Standard errors (multiplied by 10^4) are shown in parentheses. Bold-faces indicate outperforming methods, chosen by one-sided t-test with the significant level 5%.

(F_1-measure)	Proposed		Baselines		
Dataset	U-GD	U-BFGS	ERM	W-ERM	Plug-in
adult	0.617 (101)	0.660 (11)	0.639 (51)	0.676 (18)	**0.681 (9)**
australian	**0.843 (41)**	**0.844 (45)**	0.820 (123)	0.814 (116)	0.827 (51)
breast-cancer	**0.963 (31)**	**0.960 (32)**	0.950 (37)	0.948 (44)	0.953 (40)
cod-rna	0.802 (231)	0.594 (4)	0.927 (7)	0.927 (6)	**0.930 (2)**
diabetes	**0.834 (32)**	**0.828 (31)**	0.817 (50)	0.821 (40)	0.820 (42)
fourclass	**0.638 (70)**	**0.638 (64)**	0.601 (124)	0.591 (212)	0.618 (64)
german.numer	0.561 (102)	**0.580 (74)**	0.492 (188)	0.560 (107)	**0.589 (73)**
heart	**0.796 (101)**	**0.802 (99)**	**0.792 (80)**	0.764 (151)	0.764 (137)
ionosphere	**0.908 (49)**	**0.901 (43)**	0.883 (104)	0.842 (217)	**0.897 (54)**
madelon	**0.666 (19)**	0.632 (67)	0.491 (293)	0.639 (110)	**0.663 (24)**
mushrooms	1.000 (1)	0.997 (7)	**1.000 (1)**	1.000 (2)	0.999 (4)
phishing	0.937 (29)	**0.943 (7)**	**0.944 (8)**	0.940 (12)	**0.944 (8)**
phoneme	**0.648 (27)**	0.559 (22)	0.530 (201)	0.616 (135)	0.633 (35)
skin_nonskin	0.870 (3)	0.856 (4)	0.854 (7)	**0.877 (8)**	0.838 (5)
sonar	**0.735 (95)**	**0.740 (91)**	0.706 (121)	0.655 (189)	**0.721 (113)**
spambase	0.876 (27)	0.756 (61)	0.887 (42)	0.881 (58)	**0.903 (18)**
splice	0.785 (49)	**0.799 (46)**	0.785 (55)	0.771 (67)	**0.801 (45)**
w8a	0.297 (80)	0.284 (96)	0.735 (35)	**0.742 (29)**	**0.745 (26)**

Table 6: Results of the Jaccard index: 50 trials are conducted for each pair of a method and dataset. Standard errors (multiplied by 10^4) are shown in parentheses. Bold-faces indicate outperforming methods, chosen by one-sided t-test with the significant level 5%.

(Jaccard index)	Proposed		Baselines		
Dataset	U-GD	U-BFGS	ERM	W-ERM	Plug-in
adult	0.499 (44)	0.498 (11)	0.471 (51)	0.510 (20)	**0.516 (10)**
australian	**0.735 (63)**	**0.733 (59)**	0.702 (144)	0.693 (143)	0.707 (76)
breast-cancer	**0.921 (54)**	**0.918 (55)**	0.905 (66)	0.903 (78)	**0.913 (69)**
cod-rna	0.854 (3)	0.785 (8)	0.864 (11)	0.865 (9)	**0.869 (3)**
diabetes	**0.714 (44)**	0.702 (50)	0.692 (70)	0.698 (56)	0.695 (60)
fourclass	**0.469 (69)**	**0.457 (68)**	0.436 (112)	0.434 (171)	0.449 (66)
german.numer	**0.433 (64)**	**0.429 (69)**	0.335 (153)	0.391 (98)	**0.418 (71)**
heart	**0.665 (135)**	**0.675 (135)**	**0.664 (102)**	0.629 (178)	0.626 (163)
ionosphere	**0.826 (76)**	**0.829 (65)**	0.796 (134)	0.749 (245)	**0.815 (87)**
madelon	**0.495 (31)**	0.459 (69)	0.346 (225)	0.474 (100)	**0.496 (27)**
mushrooms	0.999 (2)	0.995 (4)	**1.000 (1)**	0.999 (4)	0.997 (7)
phishing	0.883 (43)	**0.893 (11)**	**0.894 (14)**	0.888 (22)	**0.894 (15)**
phoneme	0.435 (51)	0.436 (24)	0.371 (160)	**0.450 (104)**	**0.461 (34)**
skin_nonskin	0.744 (5)	0.751 (5)	0.746 (10)	**0.780 (13)**	0.722 (7)
sonar	**0.600 (125)**	**0.600 (111)**	0.552 (147)	0.495 (202)	**0.572 (134)**
spambase	**0.827 (22)**	0.708 (22)	0.798 (67)	0.790 (86)	**0.824 (31)**
splice	**0.670 (60)**	**0.672 (56)**	0.646 (71)	0.629 (84)	**0.672 (57)**
w8a	0.496 (151)	0.452 (28)	0.580 (44)	**0.590 (35)**	**0.595 (33)**

D.5 Sample Complexity

It is interesting to study the relationship between the metric performances and the size of samples, because we expect Plug-in, which requires to estimate probabilities accurately, does not work well when the size of samples is quite small. Figures 12 and 13 show the sample complexity results. Even though learning is not stable for small samples (e.g., heart and w8a), we can observe clear differences in some cases such as cod-rna, diabetes, german.numer, ionosphere, sonar, and splice in F_1-measure, and australian, cod-rna, diabetes, ionosphere, phishing, sonar, and spambase in Jaccard index, where either U-GD or U-BFGS works better than Plug-in even if sample sizes are quite small around 20 to 40. In addition, Plug-in seldom works significantly better than the gradient-based methods in the cases where sample sizes range around 100 to 400 as investigated in this section. This is contrary to the behavior shown in Tabs. 5 and 6, where the full-size datasets are used to train classifiers.

As a conclusion, it can be a good option to consider using the gradient-based methods where sample sizes are very small.

D.6 Performance Sensitivity on τ

Lastly, we see the performance sensitivity on the choices of τ. We change $\tau \in \{0.1, 0.2, \ldots, 0.9\}$ and run U-GD and U-BFGS for both the F_1-measure and Jaccard index. The results are summarized in Figures 14 and 15. From these figure, we can say there is a tendency that the performance becomes better as τ becomes closer to 1. For example, the below combinations of the datasets and metrics have such a tendency.

- australian, breast-cancer, german.numer, heart, ionosphere, mushrooms, phishing, and splice in the F_β-measure,
- australian, mushrooms, phishing, and splice in the Jaccard index.

However, there are also other cases where there exist extrema of the performance with respect to the choices of τ. For example, the below combinations of the datasets and metrics have such a tendency.

- german.numer and sonar in the F_β-measure,
- breast-cancer, heart, ionosphere and sonar in the Jaccard index.

From our theoretical results in Theorems 9 and 10, we cannot determine whether the surrogate utility is calibrated or not if τ exceeds about 0.33 for the F_β-measure, and becomes closer to 1.0 for the Jaccard index. These thresholds are not so clear in Figures 14 and 15 since the conditions on τ is merely sufficient conditions, as we explain in Sec. 4. Further analyses on the discrepancy parameter are left for future work.

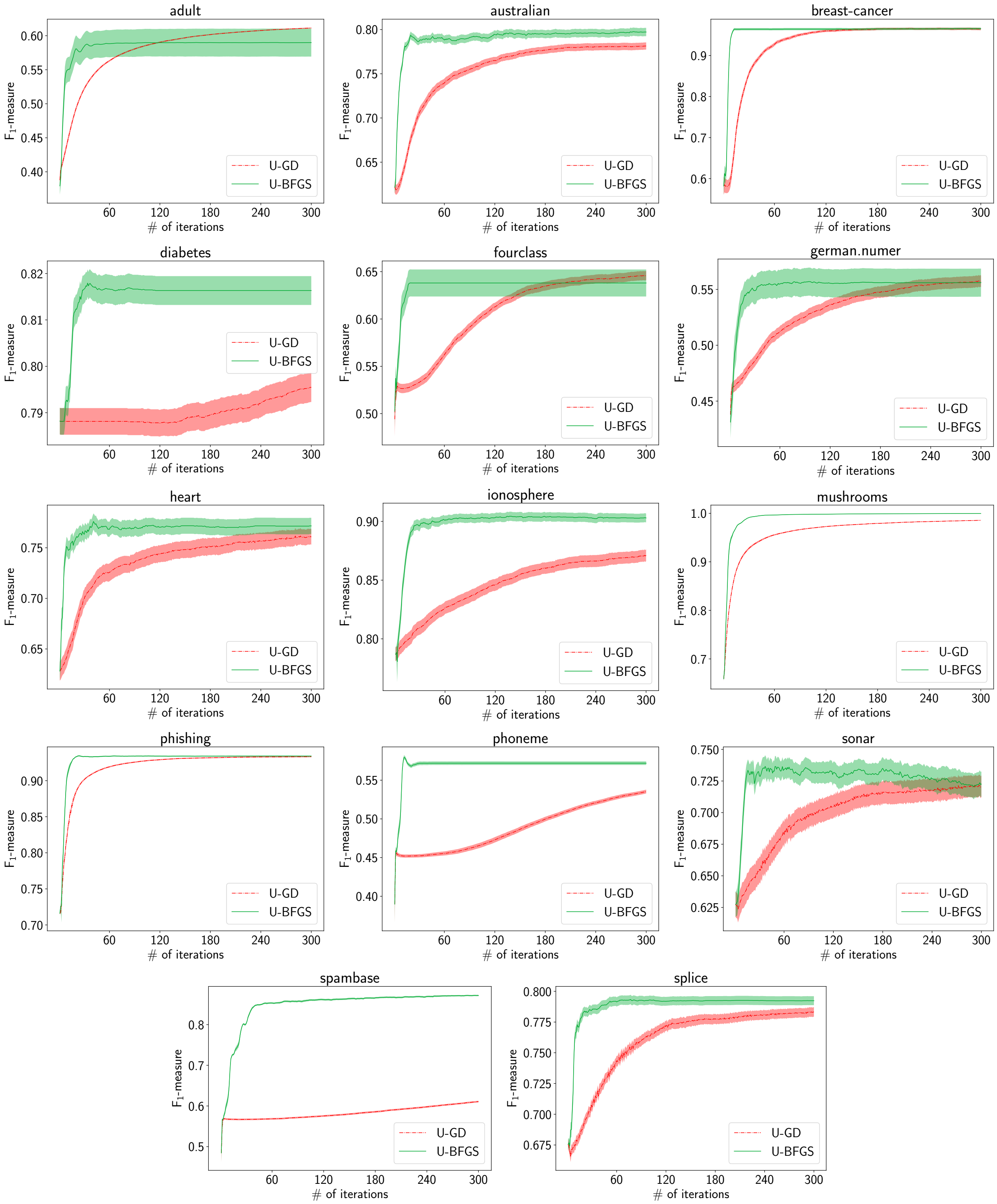

Figure 10: Convergence comparison of the F_1-measure (vertical axes). Standard errors of 50 trials are shown as shaded areas.

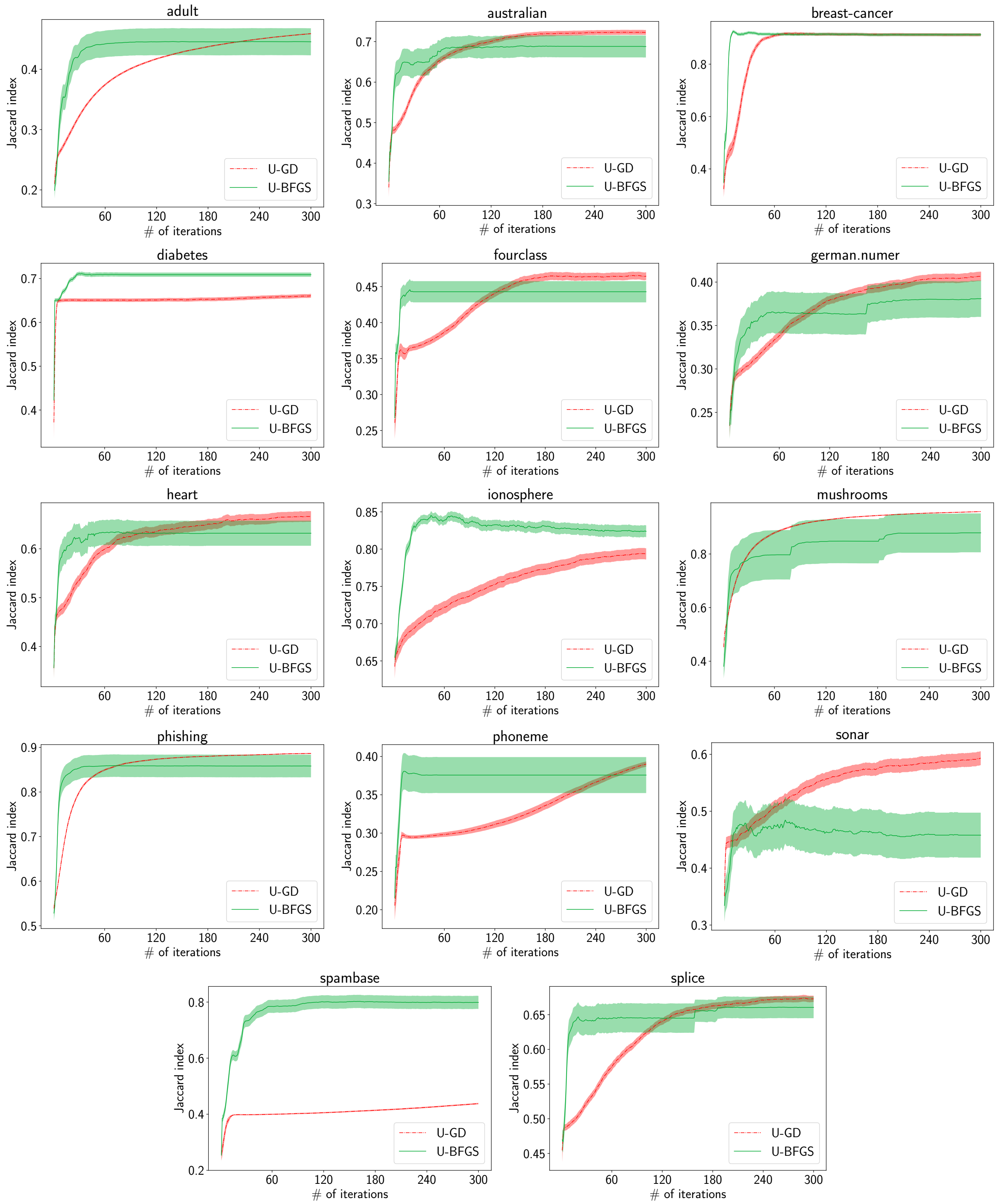

Figure 11: Convergence comparison of the Jaccard index (vertical axes). Standard errors of 50 trials are shown as shaded areas.

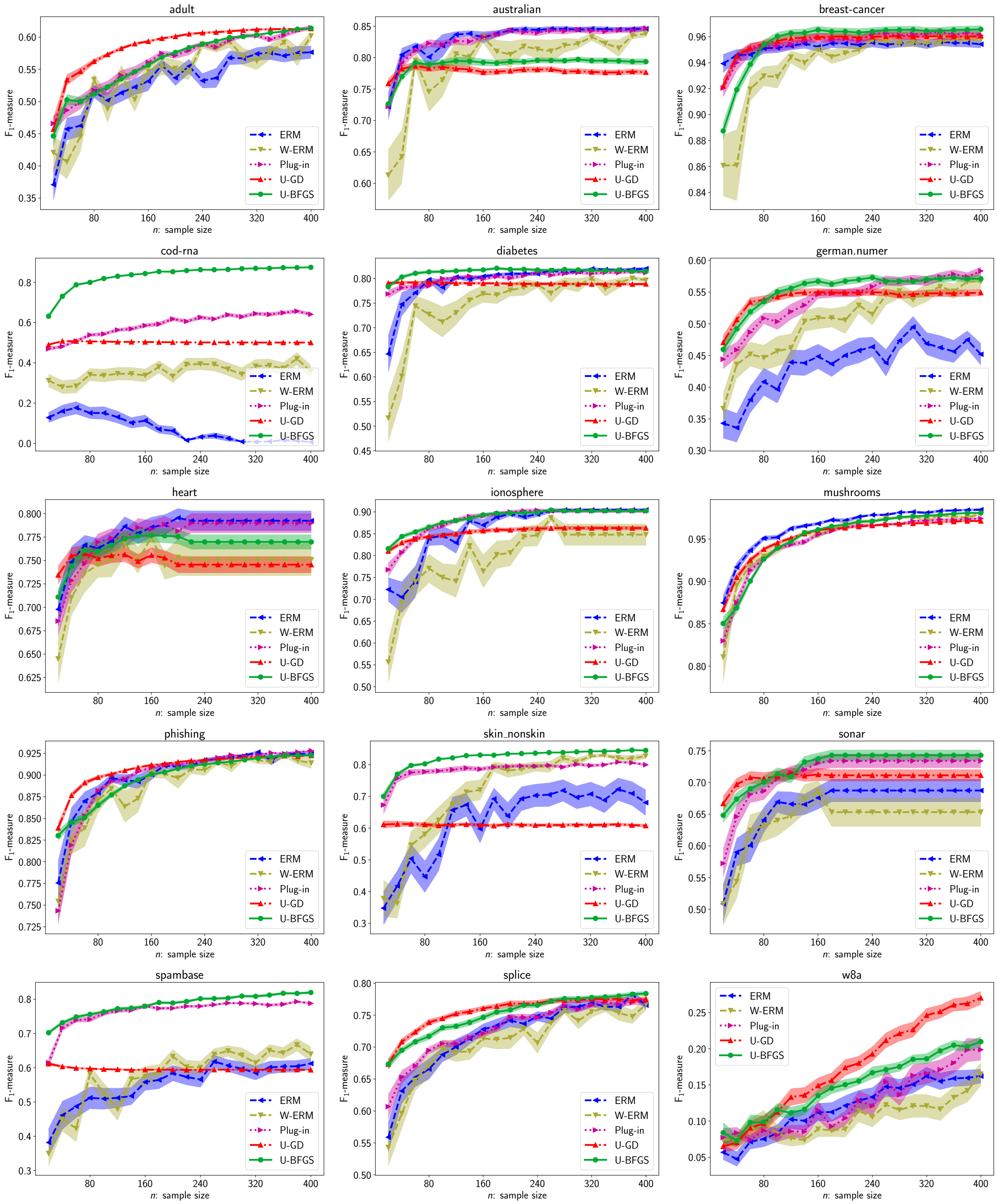

Figure 12: The relationship of the test F_1-measure (vertical axes) and sample size (horizontal axes). Standard errors of 50 trials are shown as shaded areas.

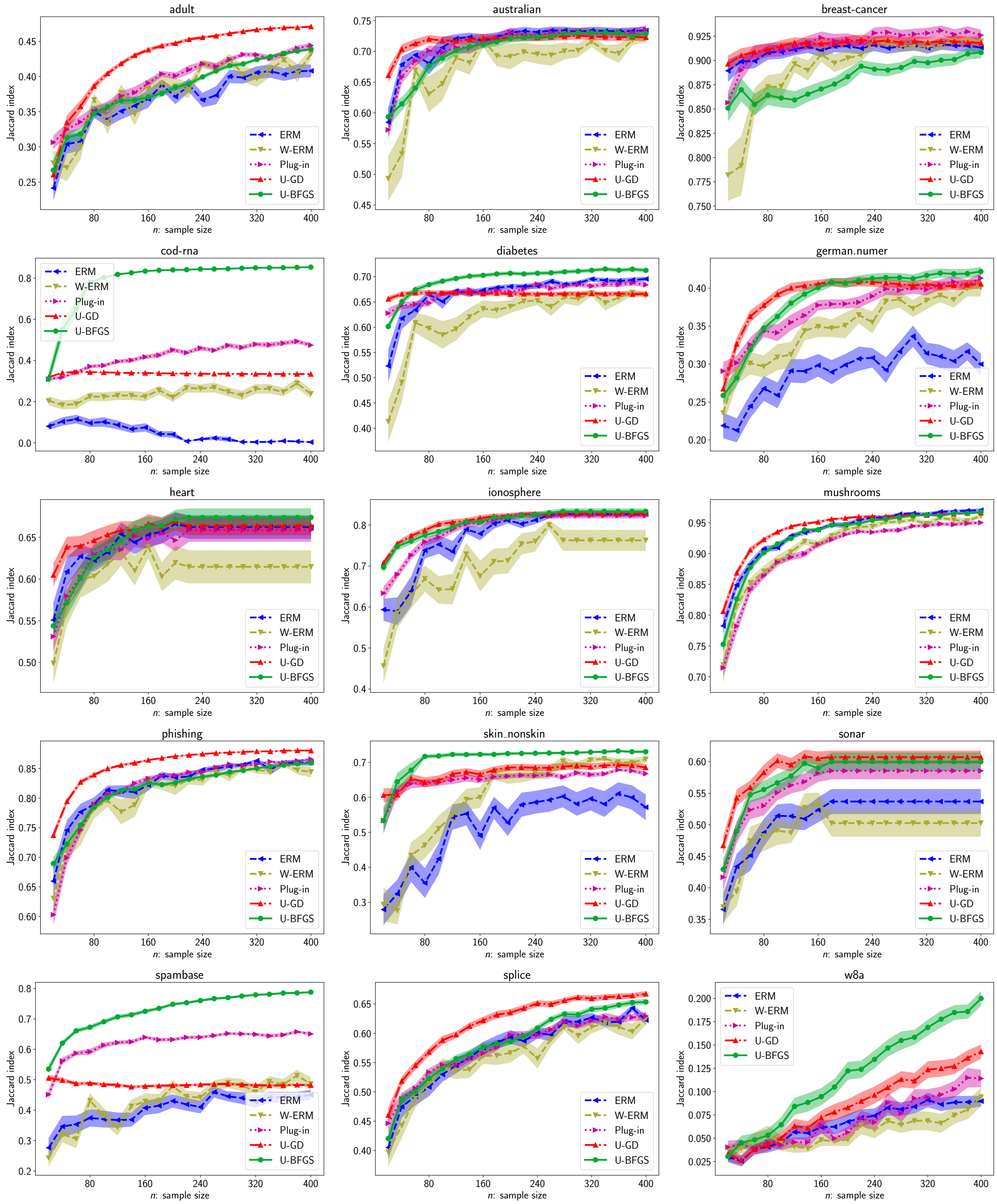

Figure 13: The relationship of the test Jaccard (vertical axes) and sample size (horizontal axes). Standard errors of 50 trials are shown as shaded areas.

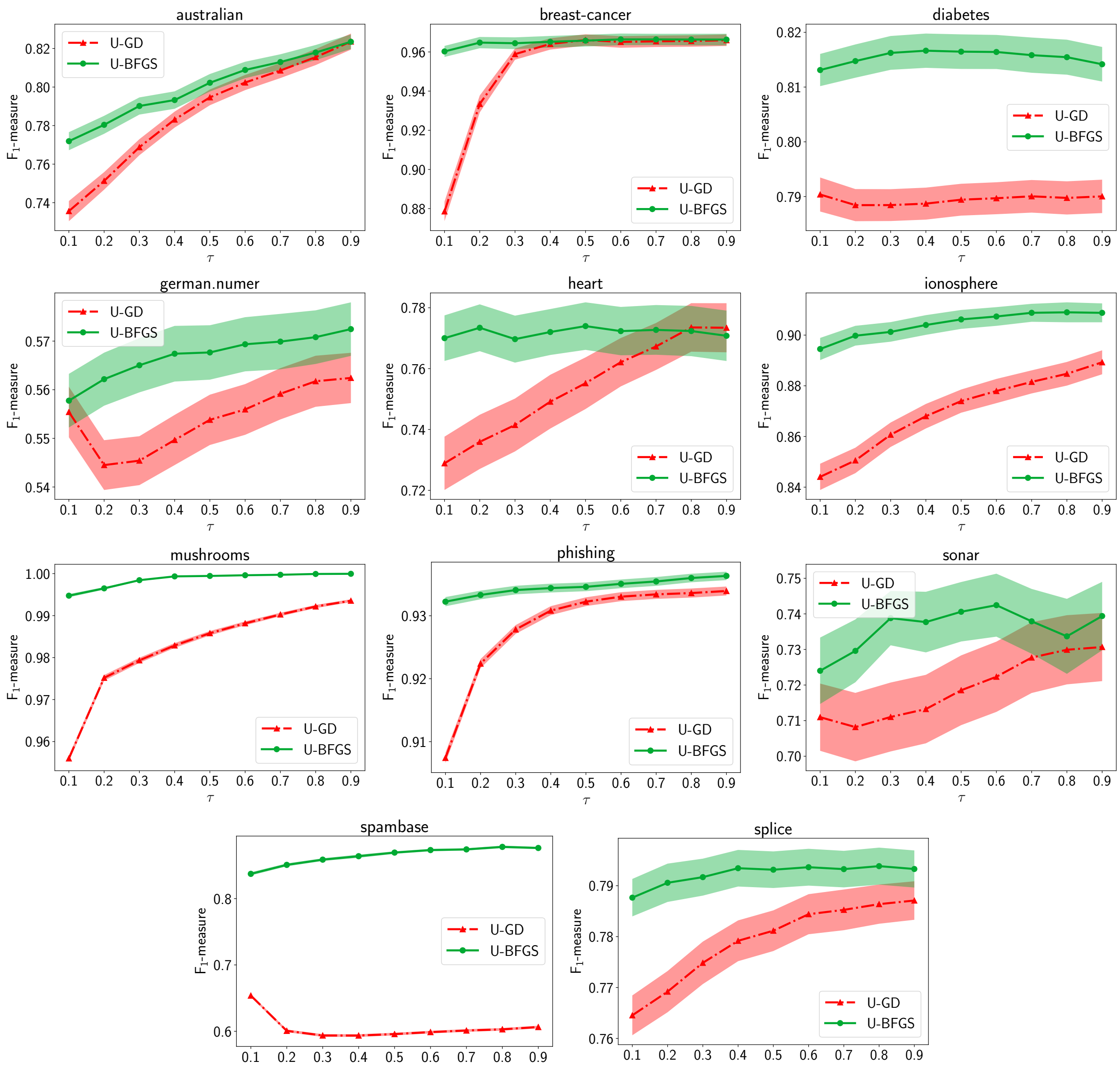

Figure 14: The relationship of the test F_1-measure (vertical axes) and the choices of τ (horizontal axes). Standard errors of 50 trials are shown as shaded areas.

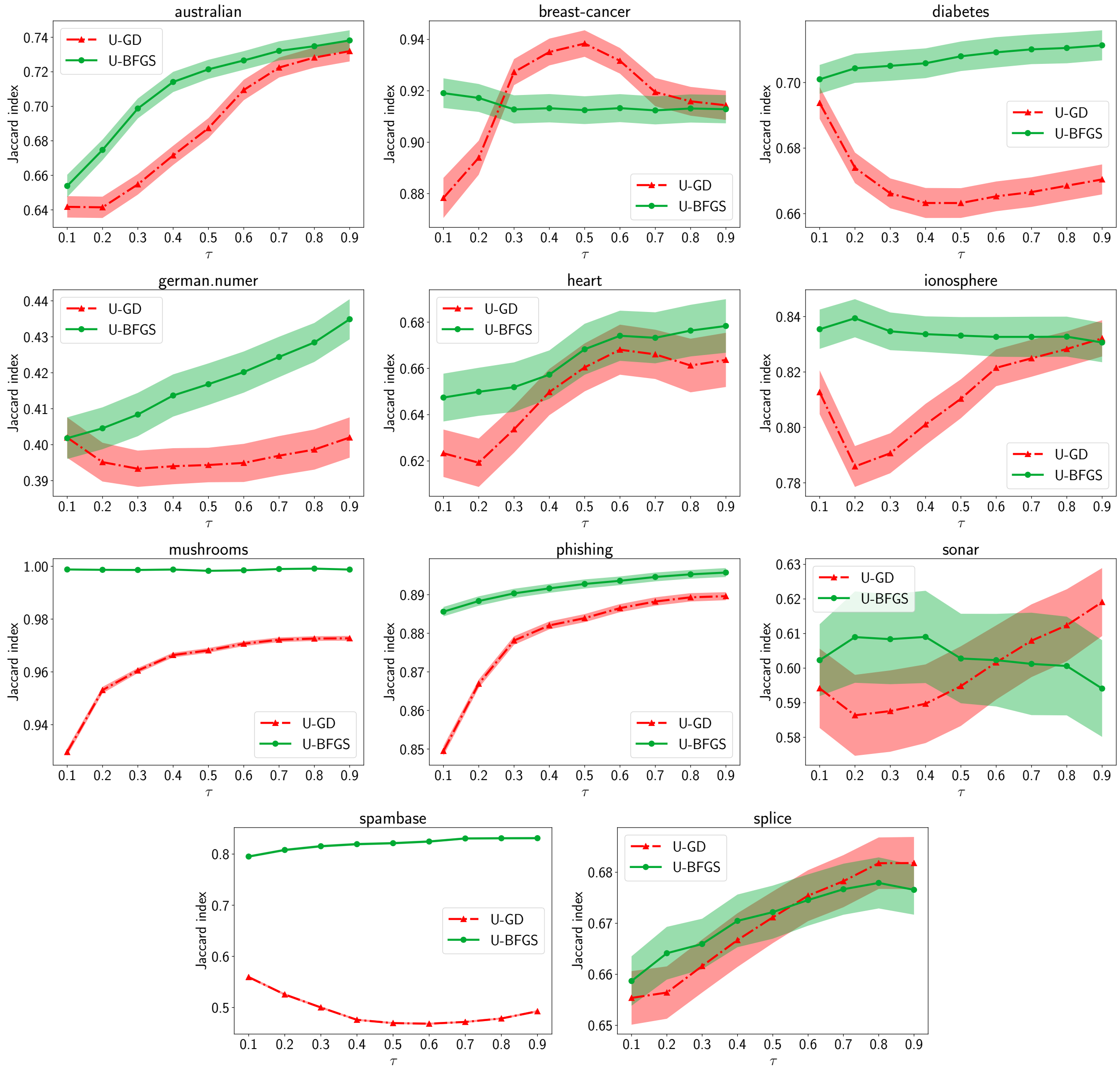

Figure 15: The relationship of the test Jaccard (vertical axes) and the choices of τ (horizontal axes). Standard errors of 50 trials are shown as shaded areas.