Calibrated Surrogate Maximization of Linear-fractional Utility

07th Feb.

Han Bao (The University of Tokyo / RIKEN AIP)
Is accuracy appropriate?

- **Our focus:** **binary classification**

![Diagram](image)

- **positive:** 5 successes, 5 failures
- **negative:** 2 successes, 8 failures

**accuracy:** 0.8 for both categories

**May cause severe issues! (e.g. in medical diagnosis)**
Is accuracy appropriate?

accuracy: 0.8
F-measure: 0.75

accuracy: 0.8
F-measure: 0

\[
F_1 = \frac{2TP}{2TP + FP + FN}
\]

\[
TP = \mathbb{E}_{X,Y=+1}[1\{f(X)>0\}]
\]

\[
FP = \mathbb{E}_{X,Y=-1}[1\{f(X)>0\}]
\]

\[
TN = \mathbb{E}_{X,Y=-1}[1\{f(X)<0\}]
\]

\[
FN = \mathbb{E}_{X,Y=+1}[1\{f(X)<0\}]
\]
Training and Evaluation

- Usual empirical risk minimization (ERM)
  - Training: minimizing 0/1-error
  - Evaluation: $\text{Acc} = \frac{TP + TN}{2}$
    - $1 - (0/1$-risk

- Training with accuracy but evaluate with $F_1$
  - Training: minimizing 0/1-error
  - Evaluation: $F_1 = \frac{2TP}{2TP + FP + FN}$

- Why not?
  - Direct Optimization
  - Training: ???
  - Evaluation: $F_1 = \frac{2TP}{2TP + FP + FN}$
Unification of Metrics

Actual Metrics

\[ F_1 = \frac{2TP}{2TP + FP + FN} \]

\[ \text{Jac} = \frac{TP}{TP + FP + FN} \]

Note:

\[ TN = \mathbb{P}(Y = -1) - FP \]
\[ FN = \mathbb{P}(Y = +1) - TP \]

linear-fraction

\[ U(f) = \frac{a_0 TP + b_0 FP + c_0}{a_1 TP + b_1 FP + c_1} \]

\[ a_k, b_k, c_k : \text{constants} \]
Unification of Metrics

\[ U(f) = \frac{a_0 TP + b_0 FP + c_0}{a_1 TP + b_1 FP + c_1} \]

\[ = \frac{a_0 \mathbb{E}_P}{a_1 \mathbb{E}_P} + b_0 \mathbb{E}_N + c_0 \]

\[ = \frac{\mathbb{E}_X[W_0(f(X))]}{\mathbb{E}_X[W_1(f(X))]} \]

- TP, FP = expectation of 0/1-loss
  - e.g. TP = \( \mathbb{P}(Y = +1, f(X) > 0) = \mathbb{E}_{X,Y=+1}[1\{f(X)>0\}] \)
Goal of This Talk

Given a metric (utility) \( U(f) = \frac{a_0 \text{TP} + b_0 \text{FP} + c_0}{a_1 \text{TP} + b_1 \text{FP} + c_1} \)

Q. How to optimize \( U(f) \) directly?

- without estimating class-posterior probability

labeled sample \( \{(x_i, y_i)\}_{i=1}^{n} \) i.i.d. \( \mathcal{P} \)

metric \( U \)

classifier \( f : \mathcal{X} \rightarrow \mathbb{R} \)

s.t. \( U(f) = \sup_{f'} U(f') \)
Outline

- Introduction

- Preliminary
  - Convex Risk Minimization
  - Plug-in Principle vs. Cost-sensitive Learning

- Key Idea
  - Quasi-concave Surrogate

- Calibration Analysis & Experiments
Formulation of Classification

- Goal of classification: maximize accuracy = minimize mis-classification rate

\[ \hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} 1[y_i \neq \text{sign}(f(x_i))] \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i f(x_i)) \]

- Make 0/1 loss smoother

(Empirical) Surrogate Risk

\[ \hat{R}_\phi(f) = \frac{1}{n} \sum_{i=1}^{n} \phi(y_i f(x_i)) \]

- Convex in \( f \)!

Example of \( \phi \)

- Logistic loss
- Hinge loss \( \Rightarrow \) SVM
- Exponential loss \( \Rightarrow \) AdaBoost
3 Actors in Risk Minimization

- Minimize **classification risk** (= 1 - Accuracy)
  \[ R(f) = \mathbb{E}[\ell(Yf(X)) ] \]
  0/1-loss \hspace{1cm} prediction margin
  0/1-loss represents if \(X\) is correctly classified by \(f\)

- **Surrogate loss** makes tractable
  differentiable upper bound of 0/1-loss
  (surrogate risk)
  \[ R_{\phi}(f) = \mathbb{E}[\phi(Yf(X))] \]
  surrogate loss

- **Sample approximation** (M-estimation)
  (empirical (surrogate) risk)
  \[ \hat{R}_{\phi}(f) = \frac{1}{n} \sum_{i=1}^{n} \phi(y_if(x_i)) \]
  what we actually minimize
Convexity & Statistical Property

Tractable (Convex)

\[ \hat{R}_\phi(f) = \frac{1}{n} \sum_{i=1}^{n} \phi(y_i f(x_i)) \]

- Generalize

\[ R_\phi(f) = \mathbb{E}[\phi(Y f(X))] \]

Intractable

\[ R(f) = \mathbb{E}[\ell(Y f(X))] \]

Q. \( \text{argmin} \ R_\phi = \text{argmin} \ R \) ?

A. Yes, w/ calibrated surrogate

**Theorem.** [Bartlett+ 2006]
Assume \( \phi \): convex.

Then, \( \text{argmin}_f R_\phi(f) = \text{argmin}_f R(f) \) iff \( \phi'(0) < 0 \).

(informal)

Related Work: Plug-in Rule

Classifier based on class-posterior probability

Bayes-optimal classifier (accuracy): \( \mathbb{P}(Y = +1 \mid x) \geq \frac{1}{2} \)

\[
\begin{array}{ccc}
Y = -1 & \hspace{1cm} & Y = +1 \\
0 & \hspace{1cm} & \frac{1}{2} \\
\end{array}
\]

\( \mathbb{P}(Y = +1 \mid X) \)

Bayes-optimal classifier (general case): \( \mathbb{P}(Y = +1 \mid x) - \delta^* \)

\[
\begin{array}{ccc}
Y = -1 & \hspace{1cm} & Y = +1 \\
0 & \hspace{1cm} & \delta^* \\
\end{array}
\]

\( \mathbb{P}(Y = +1 \mid X) \)

\( \Rightarrow \) estimate \( \mathbb{P}(Y = +1 \mid x) \) and \( \delta^* \) independently

---


Outline

- Introduction

- Preliminary
  - Convex Risk Minimization
  - Plug-in Principle vs. Cost-sensitive Learning

- Key Idea
  - Quasi-concave Surrogate

- Calibration Analysis & Experiments
Convexity & Statistical Property

\[
\hat{R}_\phi(f) = \frac{1}{n} \sum_{i=1}^{n} \phi(y_i f(x_i))
\]

\[
R_\phi(f) = \mathbb{E}[\phi(Yf(X))]
\]

\[
R(f) = \mathbb{E}[\ell(Yf(X))]
\]

Q. tractable & calibrated objective?

\[
U(f) = \frac{\mathbb{E}_X[W_0(f(X))]}{\mathbb{E}_X[W_1(f(X))]}, \text{ intractable}
\]

argmin \( R_\phi \) = argmin \( R \)

ttractable (convex)
Non-concave, but quasi-concave

Idea: concave / convex = quasi-concave

\[ \frac{f(x)}{g(x)} \text{ is quasi-concave} \]
if \( f : \text{concave}, g : \text{convex}, \)
\[ f(x) \geq 0 \text{ and } g(x) > 0 \text{ for } \forall x \]

(proof) Show \( \{x | f/g \geq \alpha\} \) is convex.
\[ \frac{f(x)}{g(x)} \geq \alpha \iff f(x) - \alpha g(x) \geq 0 \]
\( \text{concave} \)

NB: super-level set of concave func.
is convex
\( \therefore \{x | f/g \geq \alpha\} \text{ is convex for } \forall \alpha \geq 0 \)

non-concave, but unimodal \( \Rightarrow \) efficiently optimized

- quasi-concave \( \supseteq \) concave
- super-levels are convex
Surrogate Utility

- Idea: bound true utility from below

\[ U(f) = \frac{a_0 TP + b_0 FP + c_0}{a_1 TP + b_1 FP + c_1} \]

\[ \geq \frac{a_0 \mathbb{E}_P + b_0 \mathbb{E}_N + c_0}{a_1 \mathbb{E}_P + b_1 \mathbb{E}_N + c_1} \]

non-negative sum of concave
⇒ concave

non-negative sum of convex
⇒ convex

numerator from below

denominator from above
Surrogate Utility

- Idea: bound true utility from below

\[
U(f) = \frac{a_0TP + b_0FP + c_0}{a_1TP + b_1FP + c_1} \geq \frac{a_0\mathbb{E}_P}{a_1\mathbb{E}_P} + \frac{b_0\mathbb{E}_N}{b_1\mathbb{E}_N} + c_0 + c_1
\]

\[
U_\phi(f) = \frac{a_0\mathbb{E}_P[1 - \phi(f(X))] + b_0\mathbb{E}_N[-\phi(-f(X))] + c_0}{a_1\mathbb{E}_P[1 + \phi(f(X))] + b_1\mathbb{E}_N[\phi(-f(X))]} + c_1
\]

\[
:= \frac{\mathbb{E}[W_{0,\phi}]}{\mathbb{E}[W_{1,\phi}]} : \text{Surrogate Utility}
\]
Hybrid Optimization Strategy

\[ U_\phi(f) = \frac{a_0 E_P}{a_1 E_P} + b_0 E_N + b_1 E_N + c_0 + c_1 = \]

- Note: numerator can be negative
  - \( U_\phi \) isn’t quasi-concave if numerator < 0
  - maximize numerator first (concave), then maximize fractional form (quasi-concave)
Hybrid Optimization Strategy

**Algorithm 1: Hybrid Optimization Algorithm**

**Input**: $\phi$ convex loss, $\theta$ initial classifier parameter

repeat

| $g^n \leftarrow \nabla_\theta \hat{U}_\phi^n(f_\theta)$ |
| $\theta \leftarrow \text{gradient\_based\_update}(\theta, g^n)$ |

until $\hat{U}_\phi^n(f_\theta) \leq 0$

repeat

| $g \leftarrow \nabla_\theta \hat{U}_\phi(f_\theta), \hat{g} = g / \|g\|$ |
| $\theta \leftarrow \text{gradient\_based\_update}(\theta, \hat{g})$ |

until stopping criterion is satisfied

**Output**: maximizer $f_\theta$

Outline

- Introduction
- Preliminary
  - Convex Risk Minimization
  - Plug-in Principle vs. Cost-sensitive Learning
- Key Idea
  - Quasi-concave Surrogate
- Calibration Analysis & Experiments
For classification risk:

- Surrogate risk:
  \[ R_\phi(f) = \mathbb{E}[\phi(Yf(X))] \]

- Classification risk:
  \[ R(f) = \mathbb{E}[\ell(Yf(X))] \]

If \( \phi \) is classification-calibrated loss, \([\text{Bartlett+ 2006}]\):

\[ R_\phi(f_n) \xrightarrow{n \to \infty} 0 \implies R(f_n) \xrightarrow{n \to \infty} 0 \quad \forall \{f_n\} \]

Note: informal

For fractional utility:

- Surrogate utility:
  \[ U_\phi(f) = \frac{\mathbb{E}_X[W_0,\phi(f(X))]}{\mathbb{E}_X[W_1,\phi(f(X))]} \]

- True utility:
  \[ U(f) = \frac{\mathbb{E}_X[W_0(f(X))]}{\mathbb{E}_X[W_1(f(X))]} \]

**Q.** What kind of conditions are needed for \( \phi \) to satisfy

\[ U_\phi(f_n) \xrightarrow{n \to \infty} 1 \implies U(f_n) \xrightarrow{n \to \infty} 1 \quad \forall \{f_n\} \]?

Special Case: $F_1$-measure

**Theorem**

\[ U_\phi(f_n)^n \to 1 \implies U(f_n)^n \to 1 \quad \forall \{f_n\} \]

if $\phi$ satisfies

- $\exists c \in (0,1)$ s.t. $\sup_f U_\phi(f) \geq \frac{2c}{1-c}$, $\lim_{m \to +0} \phi'(m) \geq c \lim_{m \to -0} \phi'(m)$
- $\phi$ is non-increasing
- $\phi$ is convex

*Note: informal*

**Example**

\[
\phi_{-1}(m) = \log(1 + e^{-m})
\]

\[
\phi_{+1}(m) = \log(1 + e^{-cm})
\]

non-differentiable at $m=0$
## Experiment: F\textsubscript{1}-measure

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

(F\textsubscript{1}-measure is shown)

model: linear-in-parameter

surrogate loss: \( \phi(m) = \max\{\log(1 + e^{-m}), \log(1 + e^{\frac{-m}{3}})\} \)
## Experiment: Jaccard index

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Proposed</th>
<th>Baselines</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>U-GD</td>
<td>U-BFGS</td>
<td>ERM</td>
<td>W-ERM</td>
</tr>
<tr>
<td>adult</td>
<td>0.499 (44)</td>
<td>0.498 (11)</td>
<td>0.471 (51)</td>
<td>0.510 (20)</td>
</tr>
<tr>
<td>australian</td>
<td>0.735 (63)</td>
<td>0.733 (59)</td>
<td>0.702 (144)</td>
<td>0.693 (143)</td>
</tr>
<tr>
<td>breast-cancer</td>
<td>0.921 (54)</td>
<td>0.918 (55)</td>
<td>0.905 (66)</td>
<td>0.903 (78)</td>
</tr>
<tr>
<td>cod-rna</td>
<td>0.854 (3)</td>
<td>0.785 (8)</td>
<td>0.864 (11)</td>
<td>0.865 (9)</td>
</tr>
<tr>
<td>diabetes</td>
<td>0.714 (44)</td>
<td>0.702 (50)</td>
<td>0.692 (70)</td>
<td>0.698 (56)</td>
</tr>
<tr>
<td>fourclass</td>
<td>0.469 (69)</td>
<td>0.457 (68)</td>
<td>0.436 (112)</td>
<td>0.434 (171)</td>
</tr>
<tr>
<td>german_numer</td>
<td>0.433 (64)</td>
<td>0.429 (69)</td>
<td>0.335 (153)</td>
<td>0.391 (98)</td>
</tr>
<tr>
<td>heart</td>
<td>0.665 (135)</td>
<td>0.675 (135)</td>
<td>0.664 (102)</td>
<td>0.629 (178)</td>
</tr>
<tr>
<td>ionosphere</td>
<td>0.826 (76)</td>
<td>0.829 (65)</td>
<td>0.796 (134)</td>
<td>0.749 (245)</td>
</tr>
<tr>
<td>madelon</td>
<td>0.495 (31)</td>
<td>0.459 (69)</td>
<td>0.346 (225)</td>
<td>0.474 (100)</td>
</tr>
<tr>
<td>mushrooms</td>
<td>0.999 (2)</td>
<td>0.995 (4)</td>
<td>1.000 (1)</td>
<td>0.999 (4)</td>
</tr>
<tr>
<td>phishing</td>
<td>0.883 (43)</td>
<td>0.893 (11)</td>
<td>0.894 (14)</td>
<td>0.888 (22)</td>
</tr>
<tr>
<td>phoneme</td>
<td>0.435 (51)</td>
<td>0.436 (24)</td>
<td>0.371 (160)</td>
<td>0.450 (104)</td>
</tr>
<tr>
<td>skin_nonskin</td>
<td>0.744 (5)</td>
<td>0.751 (5)</td>
<td>0.746 (10)</td>
<td>0.780 (13)</td>
</tr>
<tr>
<td>sonar</td>
<td>0.600 (125)</td>
<td>0.600 (111)</td>
<td>0.552 (147)</td>
<td>0.495 (202)</td>
</tr>
<tr>
<td>spambase</td>
<td>0.827 (22)</td>
<td>0.708 (22)</td>
<td>0.798 (67)</td>
<td>0.790 (86)</td>
</tr>
<tr>
<td>splice</td>
<td>0.670 (60)</td>
<td>0.672 (56)</td>
<td>0.646 (71)</td>
<td>0.629 (84)</td>
</tr>
<tr>
<td>w8a</td>
<td>0.496 (151)</td>
<td>0.452 (28)</td>
<td>0.580 (44)</td>
<td>0.590 (35)</td>
</tr>
</tbody>
</table>

(Jaccard index is shown)

model: linear-in-parameter

surrogate loss: \[
\phi(m) = \max\{\log(1 + e^{-m}), \log(1 + e^{-\frac{3m}{4}})\}
\]
Goal

\[ U(f) = \frac{a_0 TP + b_0 FP + c_0}{a_1 TP + b_1 FP + c_1} \]

maximize linear-fractional utility

Tractable Optimization

surrogate utility

\[ \frac{a_0 \mathbb{E}_P}{a_1 \mathbb{E}_P} + b_0 \mathbb{E}_N + c_0 \]

\[ \frac{a_1 \mathbb{E}_P}{a_1 \mathbb{E}_P} + b_1 \mathbb{E}_N + c_1 \]

quasi-concave optimization

Calibrated Surrogate

If loss is like

1. grad discrepancy
2. decreasing
3. convex

then

\[ \text{argmax}_f U_\phi(f) = \text{argmax}_f U(f) \]

Open Problems

- necessary and sufficient condition of calibration
- explicit convergence rate
- theoretical comparison with probability estimation