Calibrated Surrogate Maximization of Linear-fractional Utility

07th Feb.

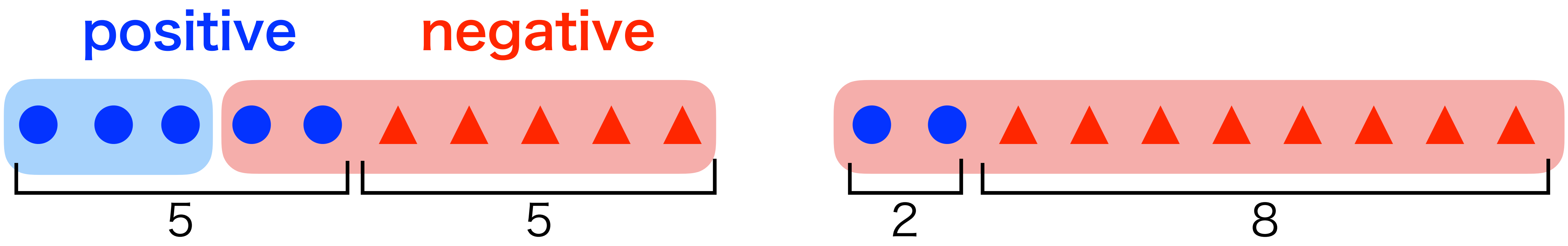
Han Bao (The University of Tokyo / RIKEN AIP)

東京大学
THE UNIVERSITY OF TOKYO

RIKEN

Is accuracy appropriate?

- Our focus: **binary classification**

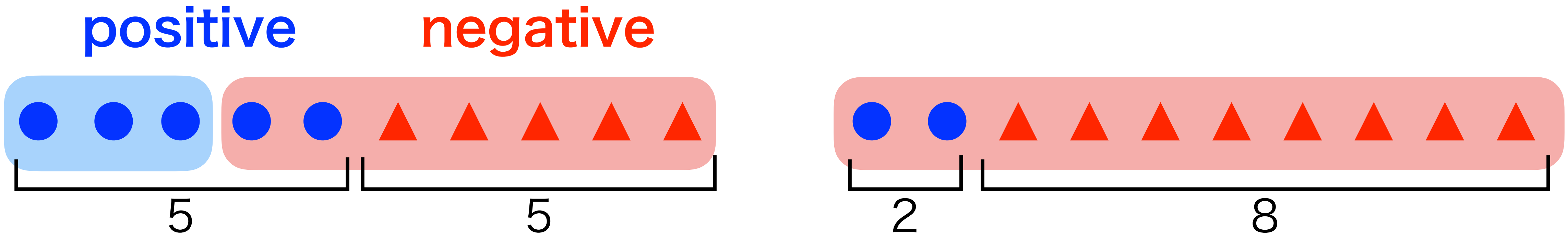


accuracy: **0.8**

accuracy: **0.8**

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-measure $$\mathrm{F}_1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

$$\mathrm{TP} = \mathbb{E}_{X,Y=+1}[1_{\{f(X)>0\}}] \qquad \mathrm{TN} = \mathbb{E}_{X,Y=-1}[1_{\{f(X)<0\}}]$$

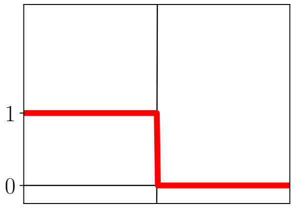
$$\mathrm{FP} = \mathbb{E}_{X,Y=-1}[1_{\{f(X)>0\}}] \qquad \mathrm{FN} = \mathbb{E}_{X,Y=+1}[1_{\{f(X)<0\}}]$$

Training and Evaluation

- Usual empirical risk minimization (ERM)

training

minimizing 0/1-error



compatible

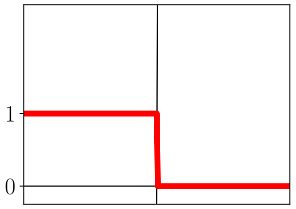
evaluation

$$\mathrm{Acc} = \mathrm{TP} + \mathrm{TN}$$
$$= 1 - (0/1\text{-risk})$$

- Training with accuracy but evaluate with F1

training

minimizing 0/1-error



incompatible

evaluation

$$F_1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

- Why not?

Direct Optimization

training

? ? ?

compatible

evaluation

$$F_1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

Weighted Accuracy

$$\mathrm{WAcc} = \frac{w_1\mathrm{TP} + w_2\mathrm{TN}}{w_1\mathrm{TP} + w_2\mathrm{TN} + w_3\mathrm{FP} + w_4\mathrm{FN}}$$

Fowlkes-Mallows index

$$\mathrm{FMI} = \frac{\mathrm{TP}}{\pi}\sqrt{\frac{1}{\mathrm{TP} + \mathrm{FP}}}$$

F-measure

$$\mathrm{F}_1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

Ac

Wanna Unify!!

alanced Error Rate

$$\mathrm{BER} = \frac{1}{\pi}\mathrm{FN} + \frac{1}{1-\pi}\mathrm{FP}$$

Jaccard ind

$$\mathrm{Jac} = \frac{}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

Matthews Correlation Coefficient

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{\pi(1-\pi)(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

Gower-Legendre index

$$\mathrm{GLI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \alpha(\mathrm{FP} + \mathrm{FN}) + \mathrm{TN}}$$

Unification of Metrics

Note:

$\mathrm{TN} = \mathbb{P}(Y = -1) - \mathrm{FP}$

$\mathrm{FN} = \mathbb{P}(Y = +1) - \mathrm{TP}$

Actual Metrics

$$\mathrm{F}_1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

$$\mathrm{Jac} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

linear-fraction

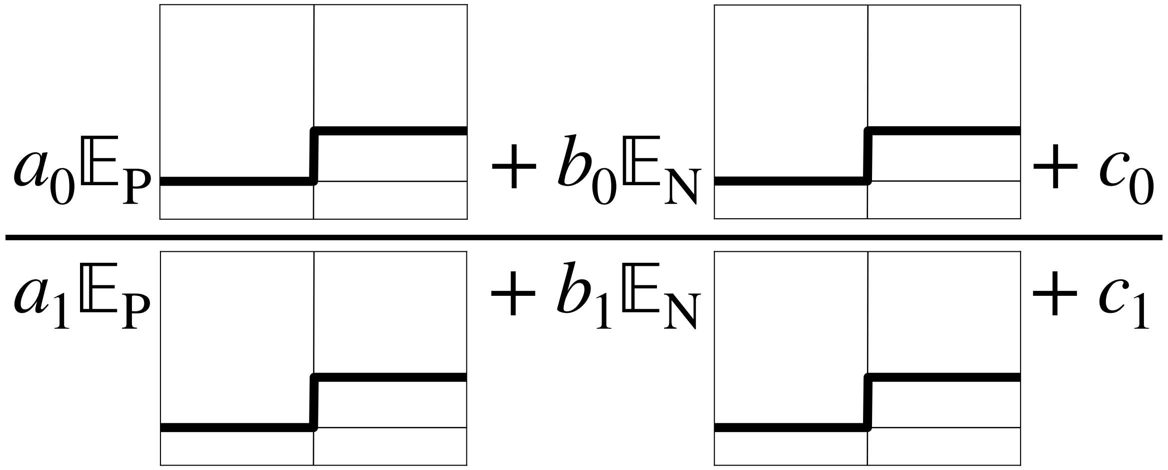
$$U(f) = \frac{a_0\mathrm{TP} + b_0\mathrm{FP} + c_0}{a_1\mathrm{TP} + b_1\mathrm{FP} + c_1}$$

a_k, b_k, c_k : constants

Unification of Metrics

linear-fraction

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



$$= \frac{a_0\mathbb{E}_{\mathrm{P}} \quad + b_0\mathbb{E}_{\mathrm{N}} \quad + c_0}{a_1\mathbb{E}_{\mathrm{P}} \quad + b_1\mathbb{E}_{\mathrm{N}} \quad + c_1}$$

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

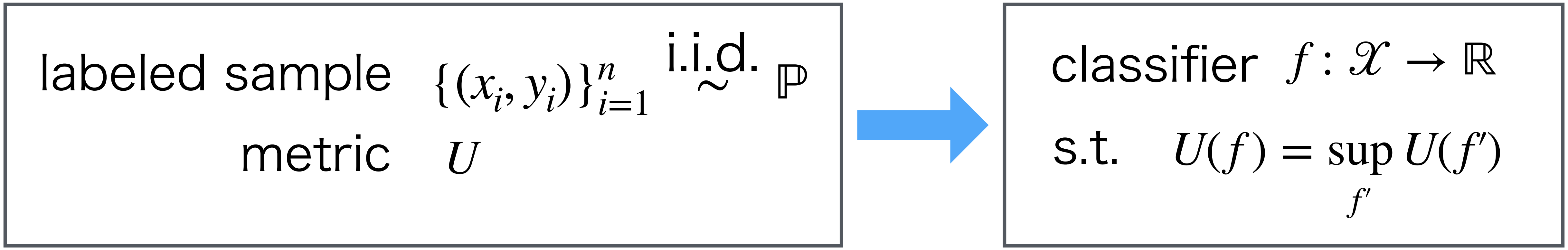
- TP, FP = expectation of 0/1-loss
  - e.g. $\mathrm{TP} = \mathbb{P}(Y = +1, f(X) > 0) = \mathbb{E}_{X,Y=+1}[1_{\{f(X)>0\}}]$

Goal of This Talk

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

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

- without estimating class-posterior probability



labeled sample $\{(x_i, y_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} \mathbb{P}$

metric U

classifier $f : \mathcal{X} \to \mathbb{R}$

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

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

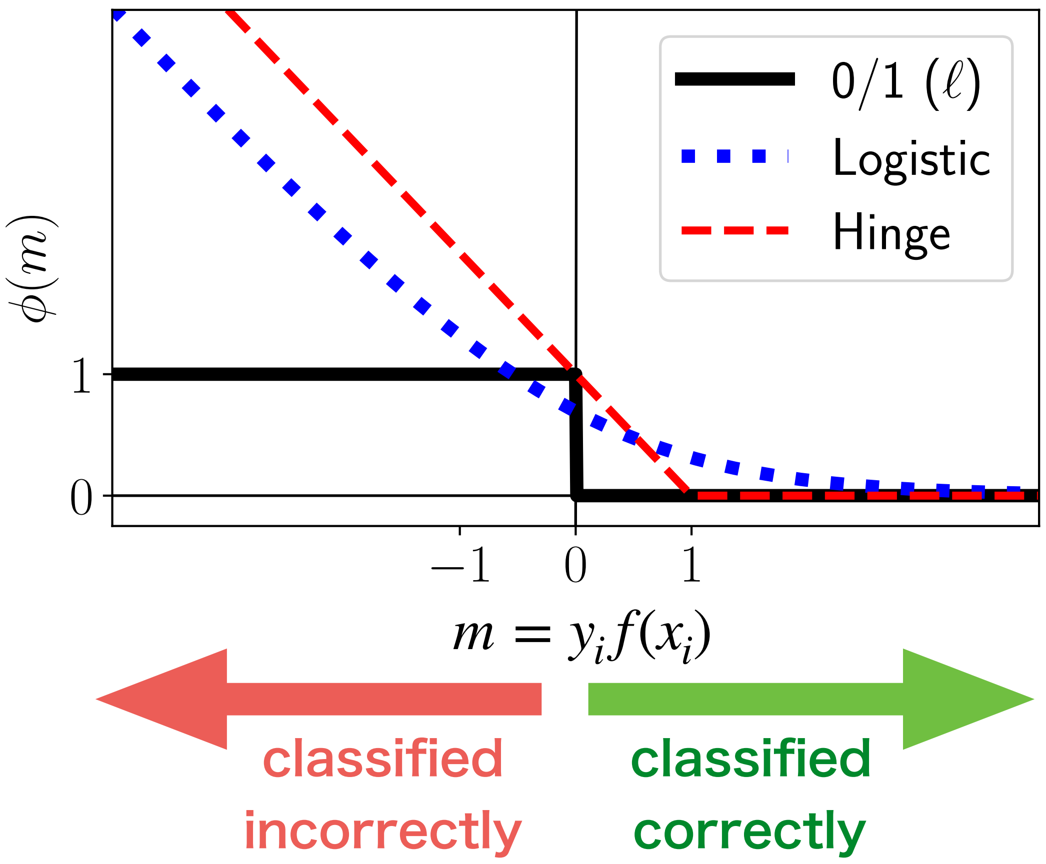
$$= \frac{1}{n}\sum_{i=1}^{n} \ell(y_i f(x_i))$$

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 ϕ

- logistic loss
- hinge loss ⇒ SVM
- exponential loss ⇒ AdaBoost

3 Actors in Risk Minimization

- Minimize classification risk (= 1 - Accuracy)

$$R(f) = \mathbb{E}[\ \underbrace{\ell}_{\text{0/1-loss}}(\ \underbrace{Yf(X)}_{\text{prediction margin}}\ )\ ]$$

0/1-loss represents if X is correctly classified by f

- Surrogate loss makes tractable

differentiable upper bound of 0/1-loss

(surrogate risk)

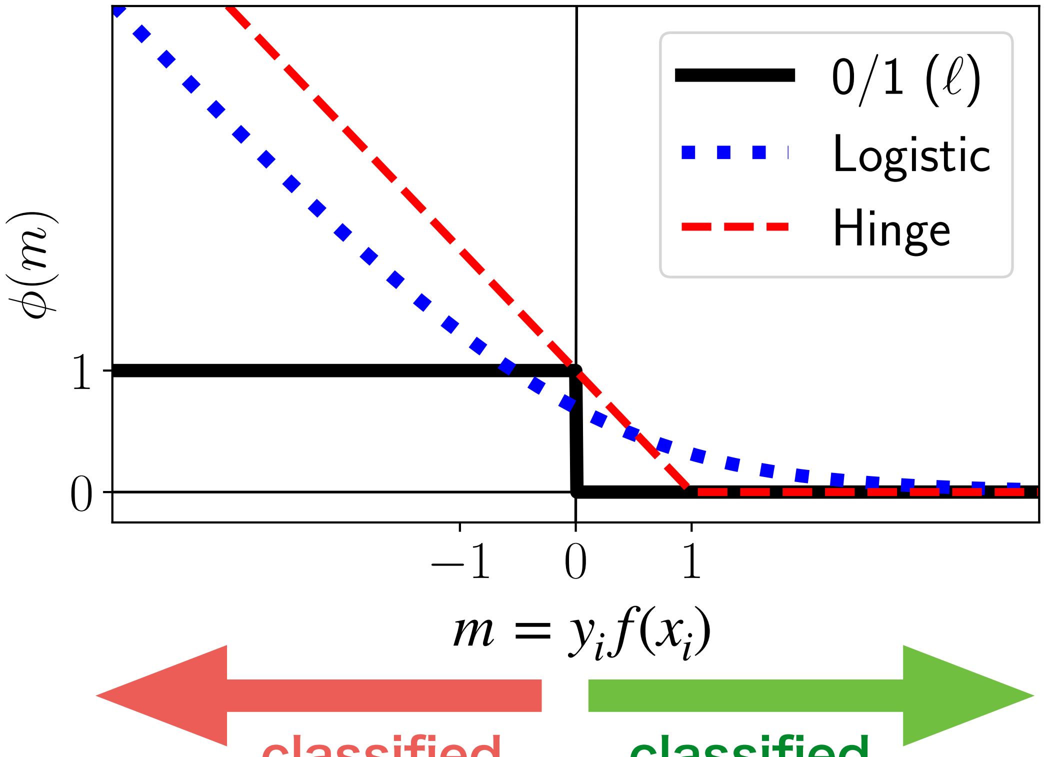
$$R_\phi(f) = \mathbb{E}[\ \underbrace{\phi}_{\text{surrogate loss}}\ (Yf(X))]$$

- Sample approximation (M-estimation)

(empirical (surrogate) risk)

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

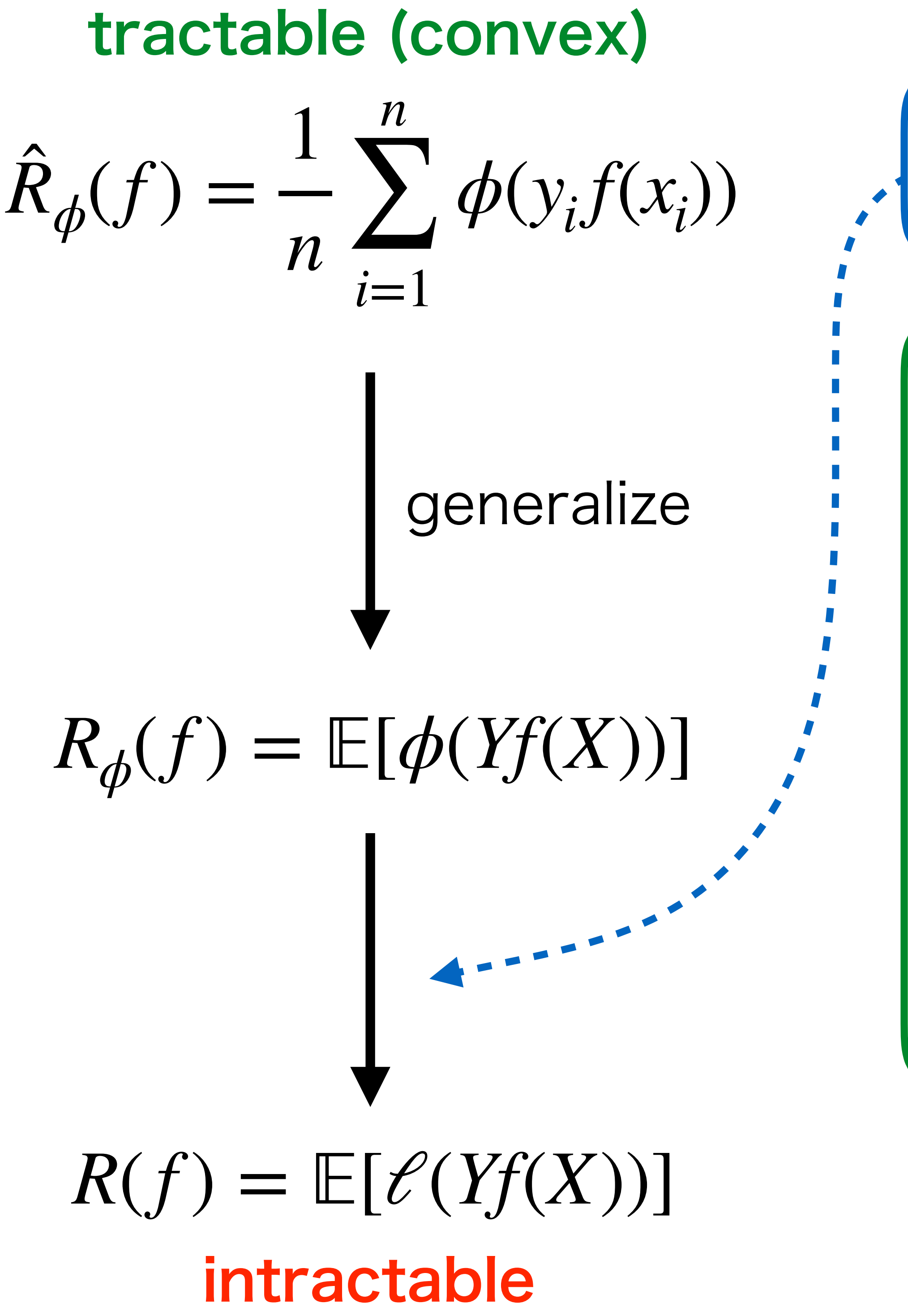
what we actually minimize



classified incorrectly

classified correctly

Convexity & Statistical Property



tractable (convex)

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

generalize

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

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

intractable

Q. argmin R_ϕ = argmin R ?

A. Yes, w/ calibrated surrogate

Theorem. [Bartlett+ 2006]

Assume ϕ: convex.

Then, $\mathrm{argmin}_f\, R_\phi(f) = \mathrm{argmin}_f\, R(f)$

iff $\phi'(0) < 0$.

(informal)

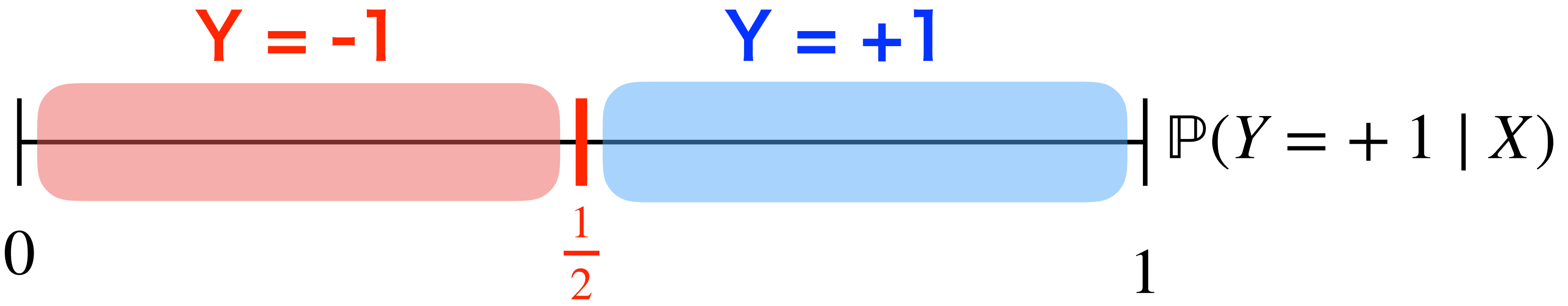
P. L. Bartlett, M. I. Jordan, & J. D. McAuliffe. (2006).
Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473), 138-156.

Related Work: Plug-in Rule

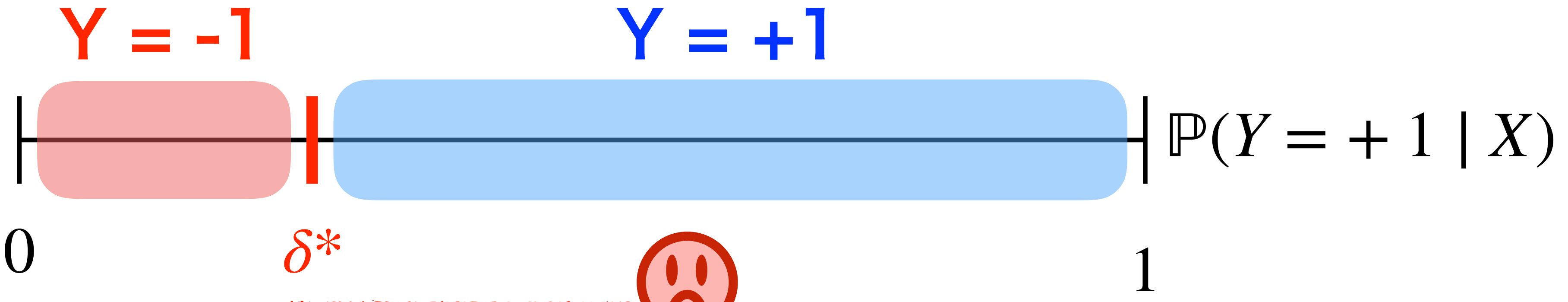
[Koyejo+ NIPS2014; Yan+ ICML2018]

- Classifier based on class-posterior probability

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



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



⇒ estimate $\boxed{\mathbb{P}(Y=+1\,|\,x)}$ and δ^* independently

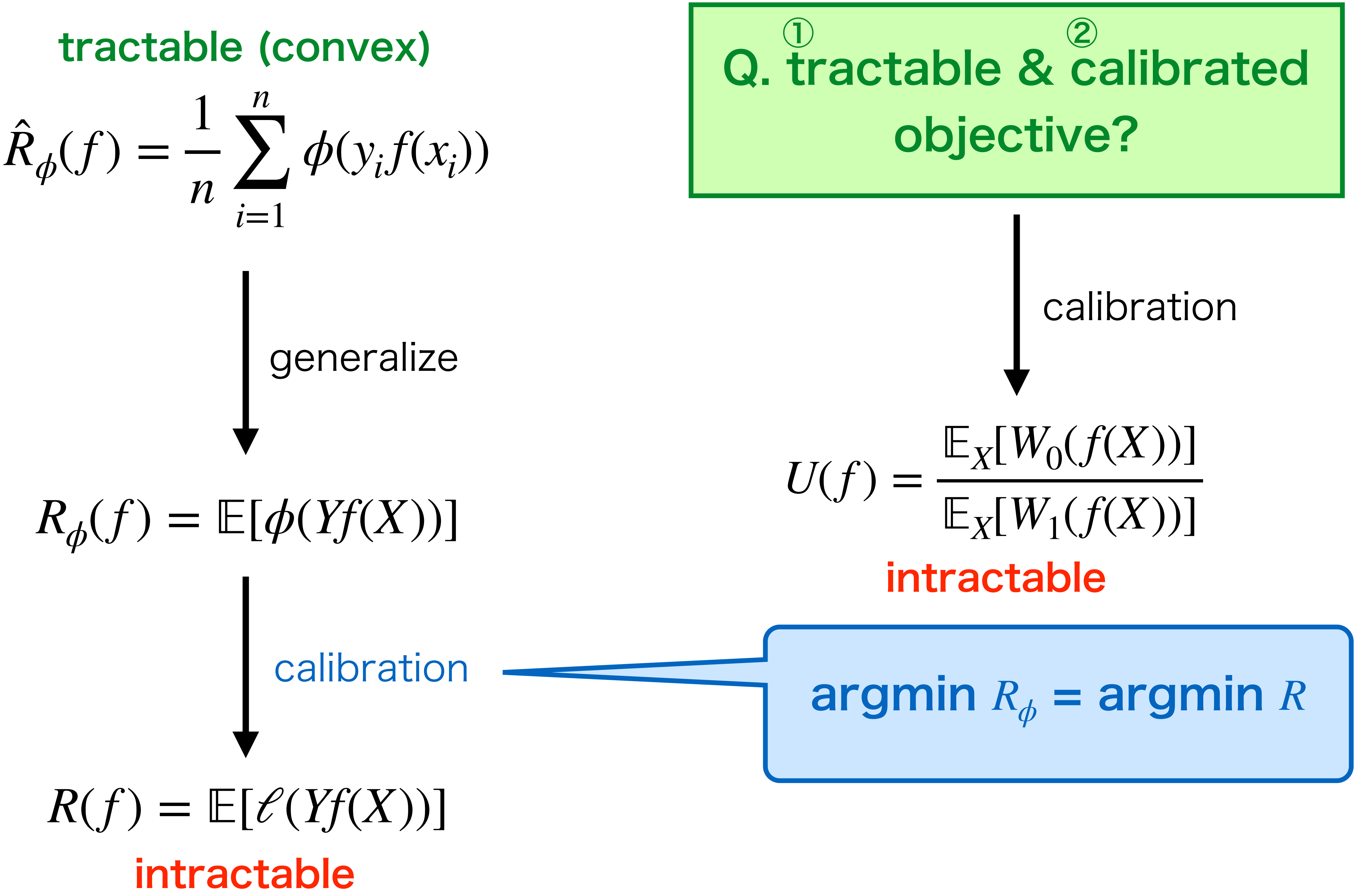
O. O. Koyejo, N. Natarajan, P. K. Ravikumar, & I. S. Dhillon.
Consistent binary classification with generalized performance metrics. In *NIPS*, 2014.

B. Yan, O. Koyejo, K. Zhong, & P. Ravikumar.
Binary classification with Karmic, threshold-quasi-concave metrics. In *ICML*, 2018.

Outline

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

Convexity & Statistical Property



tractable (convex)

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

generalize

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

calibration

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

intractable

Q. ①tractable & ②calibrated objective?

calibration

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

intractable

argmin R_ϕ = argmin R

Non-concave, but quasi-concave

Idea: concave / convex = quasi-concave

$\dfrac{f(x)}{g(x)}$ is quasi-concave

if f : concave, g : convex,

$f(x) \geq 0$ and $g(x) > 0$ for $\forall x$

(proof) Show $\{x \,|\, f/g \geq \alpha\}$ is convex.

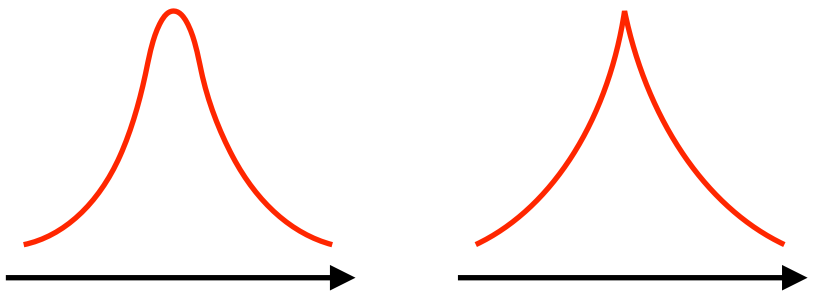
$$\frac{f(x)}{g(x)} \geq \alpha \iff \underbrace{f(x) - \alpha g(x)}_{\text{concave}} \geq 0$$

NB: super-level set of concave func. is convex

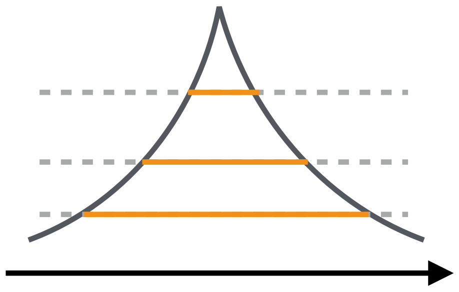
$\therefore \{x \,|\, f/g \geq \alpha\}$ is convex for $\forall \alpha \geq 0$

non-concave, but unimodal

⇒ efficiently optimized



- quasi-concave ⊇ concave
- super-levels are convex



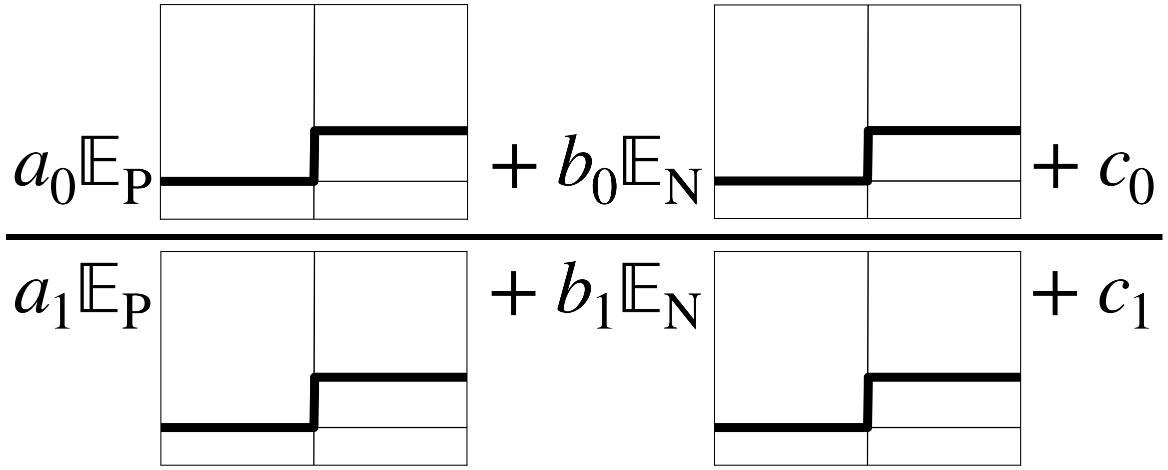
Surrogate Utility

- Idea: bound true utility from below

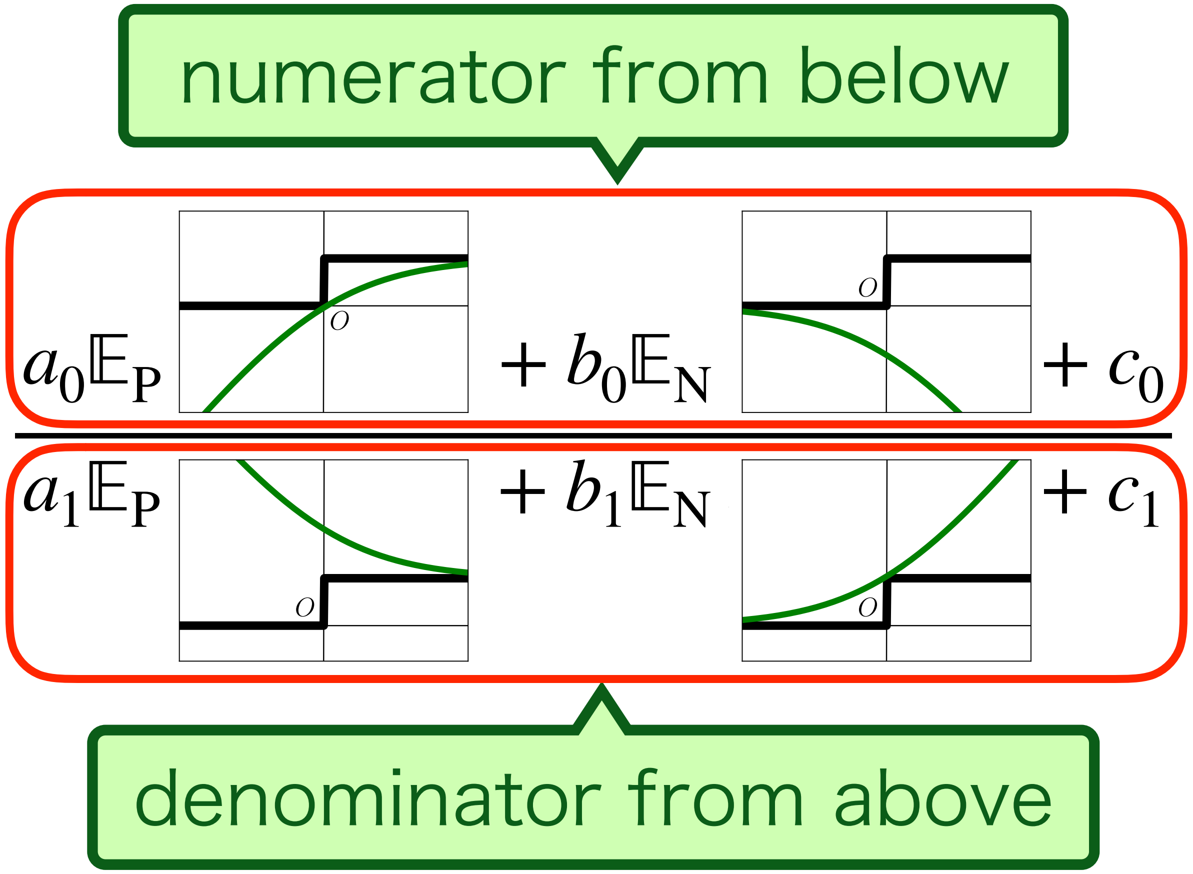
linear-fraction

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

=



≥



non-negative sum of concave
⇒ concave

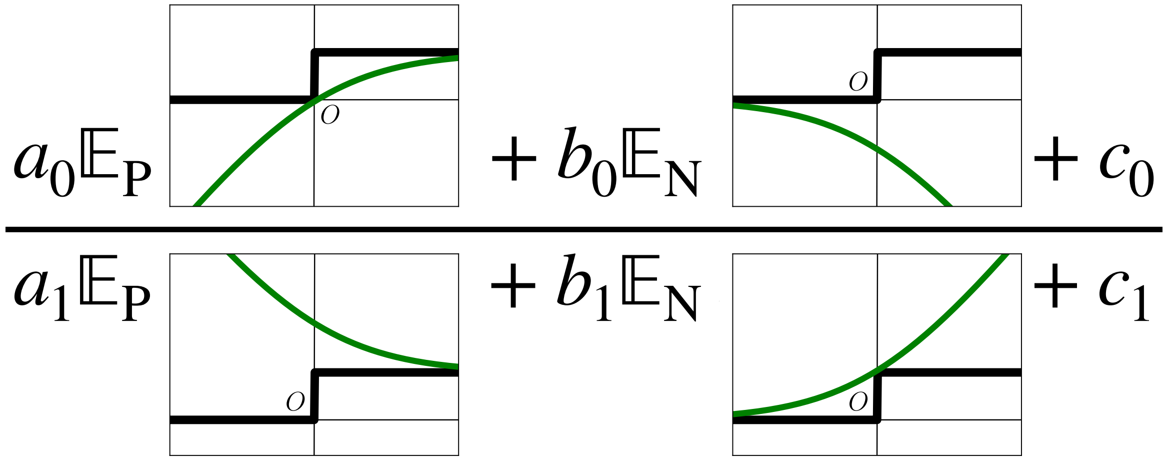
non-negative sum of convex
⇒ convex

Surrogate Utility

- Idea: bound true utility from below

linear-fraction

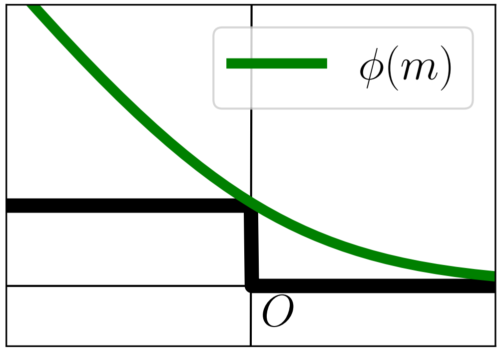
$$U(f) = \frac{a_0\mathrm{TP} + b_0\mathrm{FP} + c_0}{a_1\mathrm{TP} + b_1\mathrm{FP} + c_1}$$



$$\geq \frac{a_0\mathbb{E}_\mathrm{P} \qquad + b_0\mathbb{E}_\mathrm{N} \qquad + c_0}{a_1\mathbb{E}_\mathrm{P} \qquad + b_1\mathbb{E}_\mathrm{N} \qquad + c_1}$$

$$\parallel$$

surrogate loss

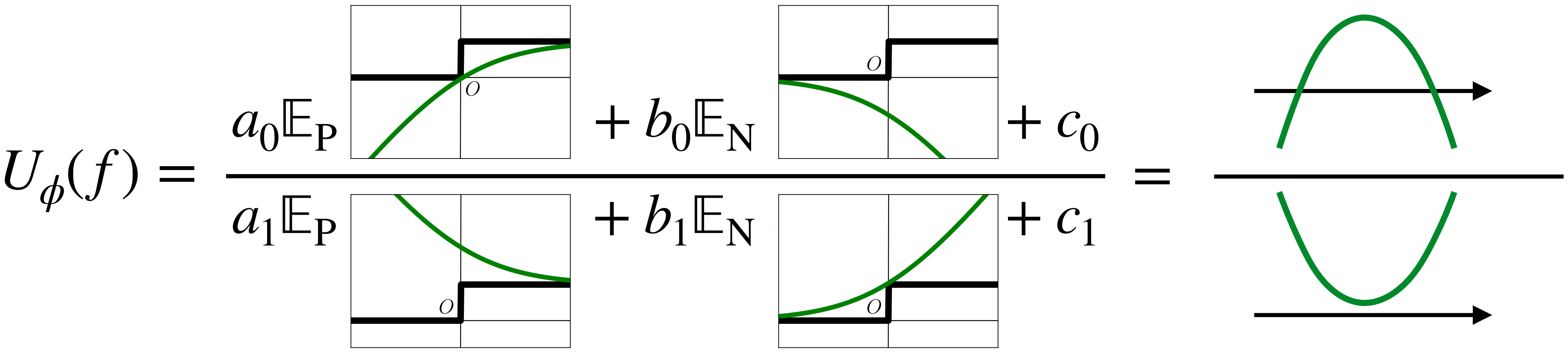


$$U_\phi(f) = \frac{a_0\mathbb{E}_\mathrm{P}\boxed{[1 - \phi(f(X))]} + b_0\mathbb{E}_\mathrm{N}\boxed{[-\phi(-f(X))]} + c_0}{a_1\mathbb{E}_\mathrm{P}\boxed{[1 + \phi(f(X))]} + b_1\mathbb{E}_\mathrm{N}\boxed{[\ \ \phi(-f(X))\ ]} + c_1}$$

$$:= \frac{\mathbb{E}[W_{0,\phi}]}{\mathbb{E}[W_{1,\phi}]}$$

: Surrogate Utility

Hybrid Optimization Strategy



$$U_\phi(f) = \frac{a_0\mathbb{E}_\mathrm{P} + b_0\mathbb{E}_\mathrm{N} + c_0}{a_1\mathbb{E}_\mathrm{P} + b_1\mathbb{E}_\mathrm{N} + c_1} =$$

- Note: numerator can be negative
  - U_ϕ 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
    $g^{\mathrm{n}} \longleftarrow \nabla_\theta \widehat{\mathfrak{U}}^{\mathrm{n}}_\phi(f_\theta)$
    $\theta \longleftarrow$ gradient_based_update$(\theta, g^{\mathrm{n}})$
until $\widehat{\mathfrak{U}}^{\mathrm{n}}_\phi(f_\theta) \leq 0$
repeat
    $g \longleftarrow \nabla_\theta \widehat{\mathfrak{U}}_\phi(f_\theta), \boxed{\widehat{g} = g/\|g\|}$
    $\theta \longleftarrow$ gradient_based_update$(\theta, \widehat{g})$
until stopping criterion is satisfied
Output: maximizer $f_\theta$
```

maximize numerator

maximize fraction

normalized gradient
for quasi-concave optimization

[Hazan+ NeurIPS2015]

Hazan, E., Levy, K., & Shalev-Shwartz, S. (2015). Beyond convexity: Stochastic quasi-convex optimization. In *Advances in Neural Information Processing Systems* (pp. 1594-1602).

Outline

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

Justify Surrogate Optimization

- For classification risk

surrogate risk

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

classification risk

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

If ϕ is **classification-calibrated** loss, [Bartlett+ 2006]

$$R_\phi(f_n) \overset{n\to\infty}{\to} 0 \Longrightarrow R(f_n) \overset{n\to\infty}{\to} 0 \quad \forall \{f_n\}$$

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 ϕ to satisfy

$$U_\phi(f_n) \overset{n\to\infty}{\to} 1 \Longrightarrow U(f_n) \overset{n\to\infty}{\to} 1 \quad \forall \{f_n\} \text{ ?}$$

P. L. Bartlett, M. I. Jordan, & J. D. McAuliffe. (2006).
Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473), 138-156.

Special Case: F$_1$-measure

merely sufficient!

Theorem

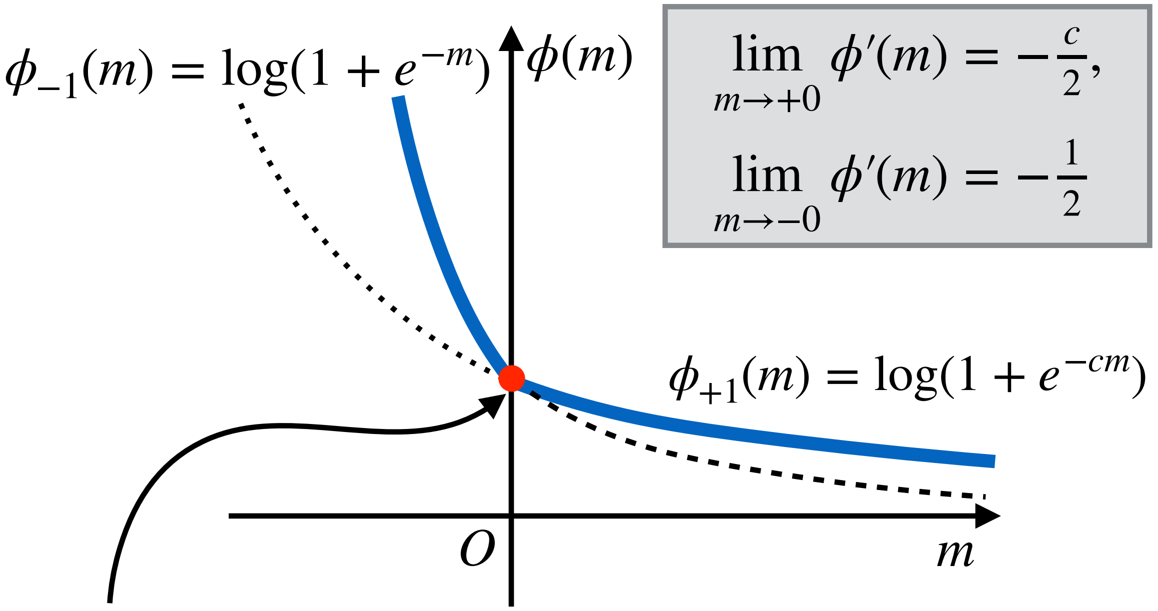
$$U_\phi(f_n) \overset{n\to\infty}{\to} 1 \Longrightarrow U(f_n) \overset{n\to\infty}{\to} 1 \quad \forall\{f_n\}$$

if ϕ satisfies

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

Note: informal

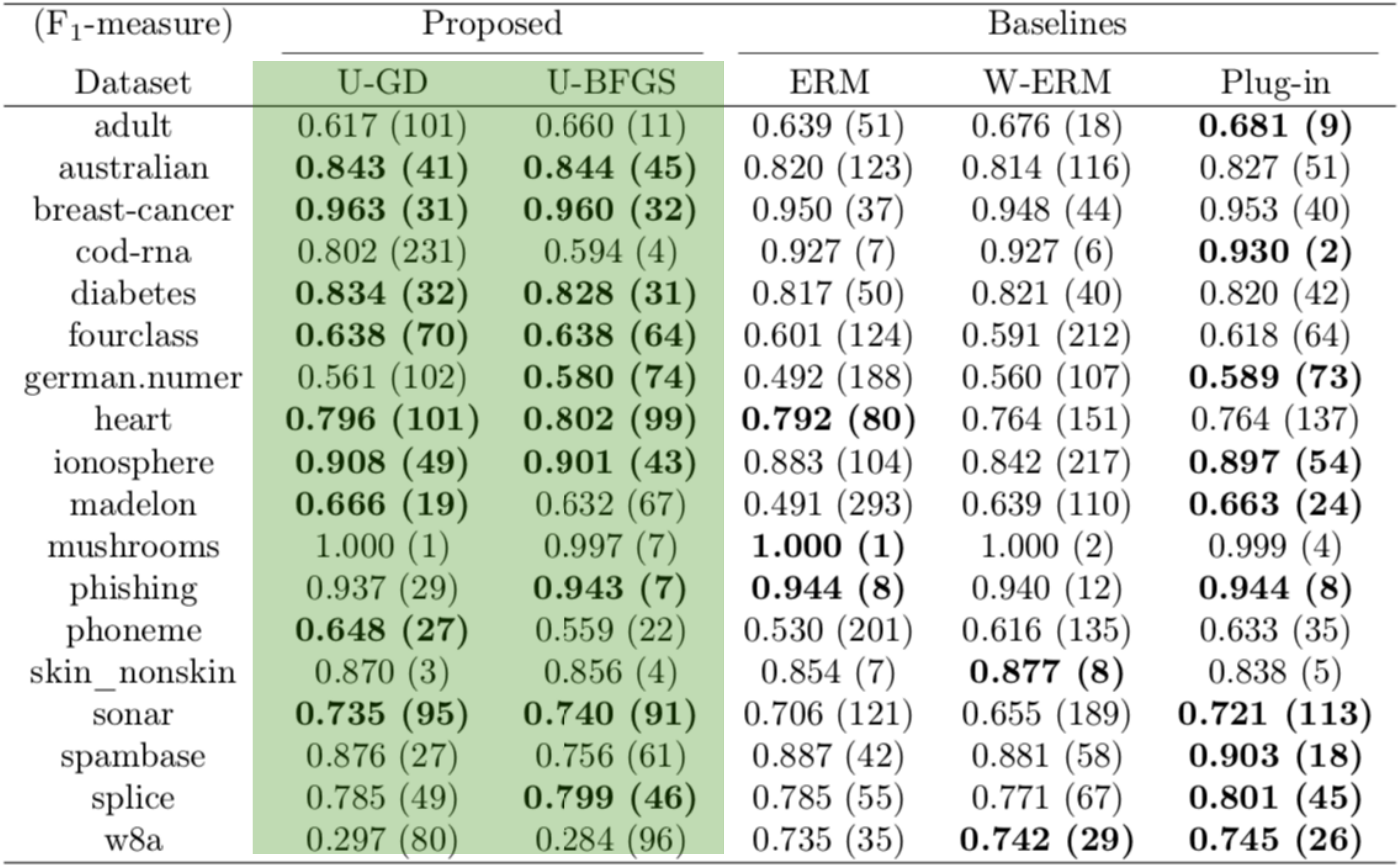
■ Example



$$\lim_{m\to+0} \phi'(m) = -\frac{c}{2}, \qquad \lim_{m\to-0} \phi'(m) = -\frac{1}{2}$$

non-differentiable at m=0

Experiment: F₁-measure



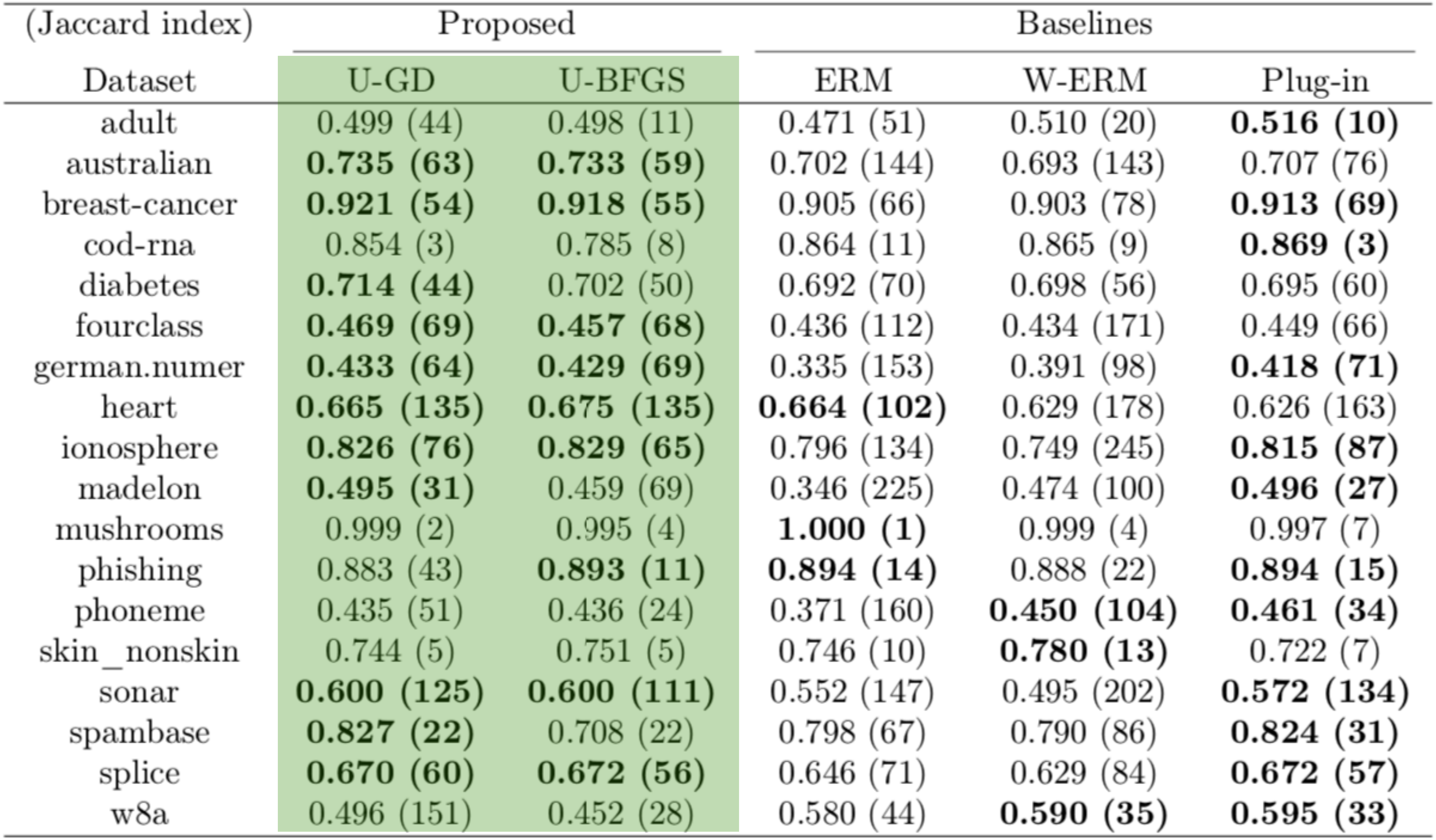
(F_1-measure)	Proposed		Baselines		
Dataset	U-GD	U-BFGS	ERM	W-ERM	Plug-in
adult	0.617 (101)	0.660 (11)	0.639 (51)	0.676 (18)	**0.681 (9)**
australian	**0.843 (41)**	**0.844 (45)**	0.820 (123)	0.814 (116)	0.827 (51)
breast-cancer	**0.963 (31)**	**0.960 (32)**	0.950 (37)	0.948 (44)	0.953 (40)
cod-rna	0.802 (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)**

(F_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



(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)**

(Jaccard index is shown)

model: linear-in-parameter

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

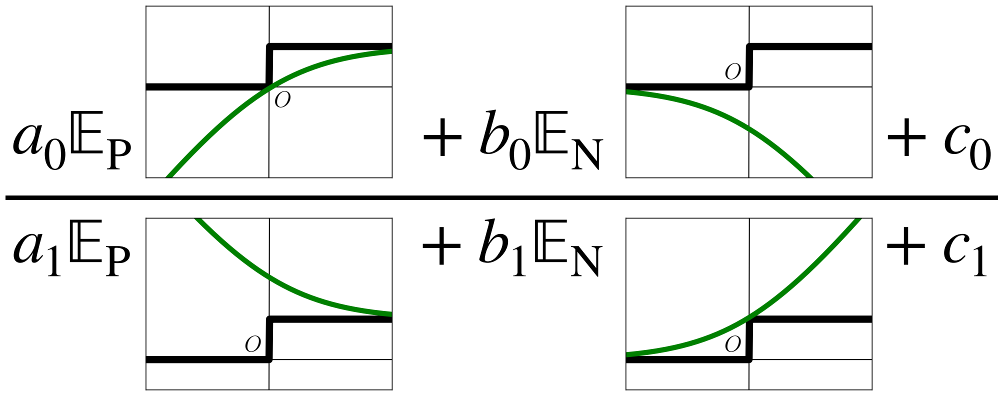
Goal

$$\boxed{U(f) = \frac{a_0\mathrm{TP} + b_0\mathrm{FP} + c_0}{a_1\mathrm{TP} + b_1\mathrm{FP} + c_1}}$$

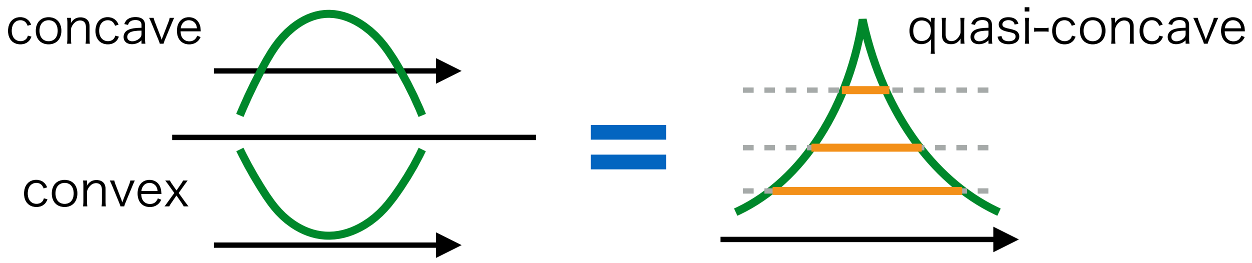
maximize linear-fractional utility

Tractable Optimization

surrogate utility

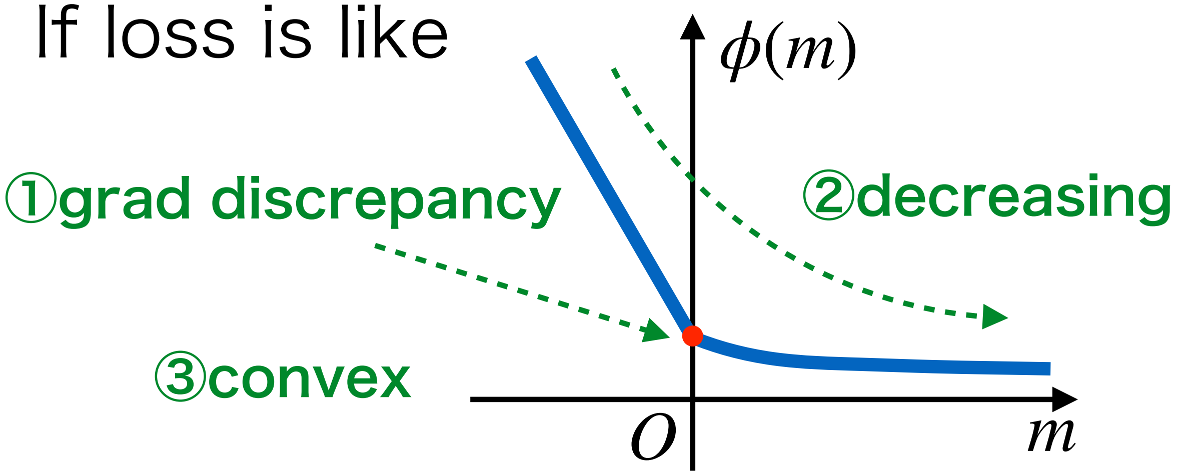


quasi-concave optimization



Calibrated Surrogate

If loss is like



then

$$\mathrm{argmax}_f\, U_\phi(f) = \mathrm{argmax}_f\, U(f)$$

Open Problems

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