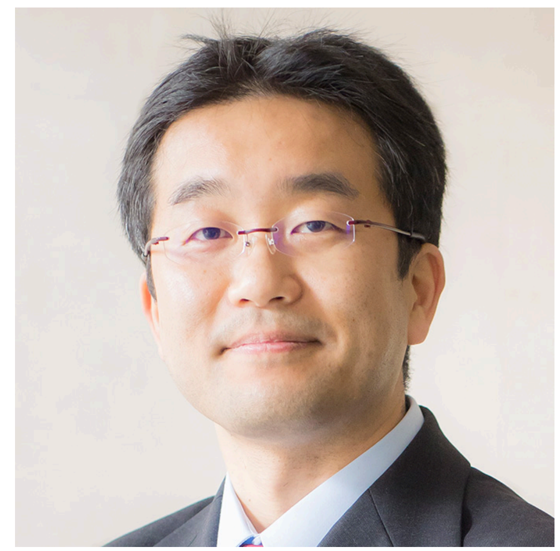
Calibrated Surrogate Maximization of Linear-fractional Utility in Binary Classification

Han Bao[1,2], Masashi Sugiyama[2,1]

1 The University of Tokyo
2 RIKEN AIP

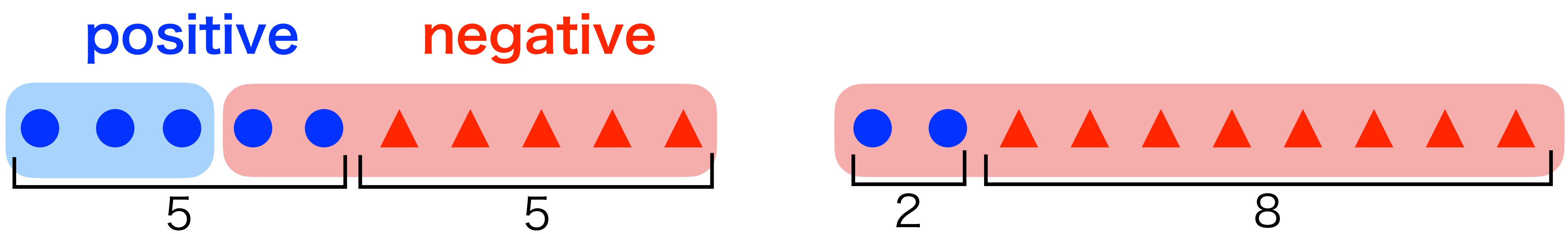
東京大学
THE UNIVERSITY OF TOKYO

RIKEN

Aug. 26th - 28th @ AISTATS 2020

Is accuracy appropriate?

- Our focus: **binary classification**



seemingly sensible classifier

unreasonable classifier

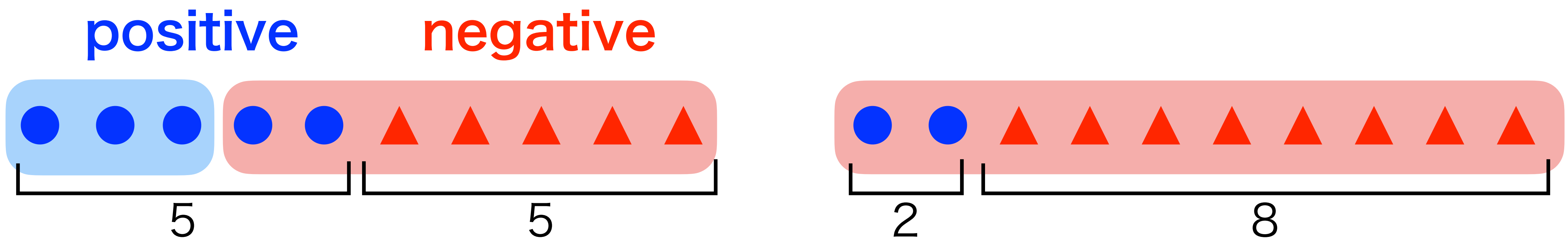
accuracy: **0.8**

accuracy: **0.8**

Accuracy can't detect unreasonable classifiers under **class imbalance**!

Is accuracy appropriate?

- F-measure is more appropriate under **class imbalance**



accuracy: **0.8**

F-measure: **0.75**

accuracy: **0.8**

F-measure: **0**

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

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

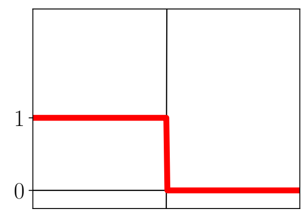
$$\mathrm{FP} = \mathbb{E}_{X,Y=-1}[1_{\{f(X)>0\}}] \qquad \mathrm{FN} = \mathbb{E}_{X,Y=+1}[1_{\{f(X)<0\}}]$$

Training and Evaluation

- Usual empirical risk minimization (ERM)

training

minimizing 0/1-error



compatible

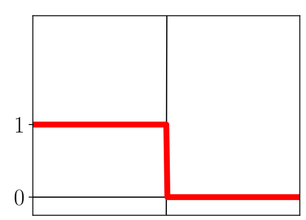
evaluation

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

- Training with accuracy but evaluating with F_1

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

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

Not only F_1, but many others

Q. Can we handle in the same way?

Accuracy

$$\mathrm{Acc} = \mathrm{TP} + \mathrm{TN}$$

Weighted Accuracy

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

F-measure

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

Balanced Error Rate

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

Gower-Legendre index

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

Jaccard index

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

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

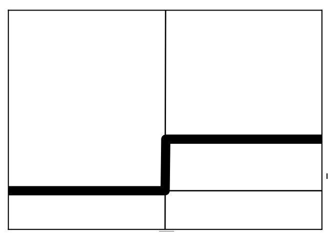
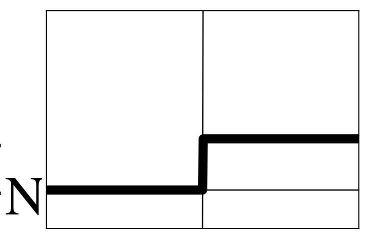
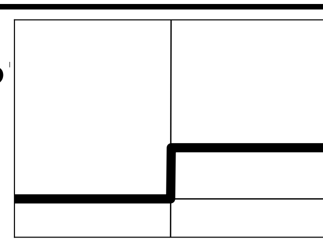
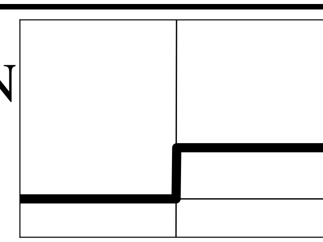
Unification of Metrics

linear-fraction

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

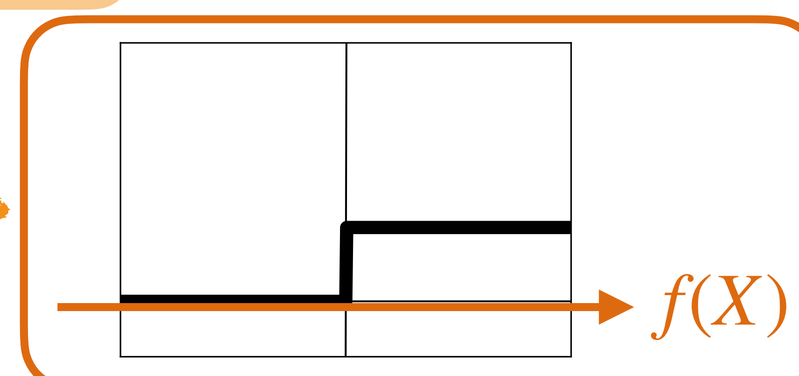
expectation divided by expecation

$= \dfrac{a_0 \mathbb{E}_{\mathrm{P}}[\ \cdot\ ] + b_0 \mathbb{E}_{\mathrm{N}}[\ \cdot\ ] + c_0}{a_1 \mathbb{E}_{\mathrm{P}}[\ \cdot\ ] + b_1 \mathbb{E}_{\mathrm{N}}[\ \cdot\ ] + c_1}$

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

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



Goal of This Talk

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

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

→

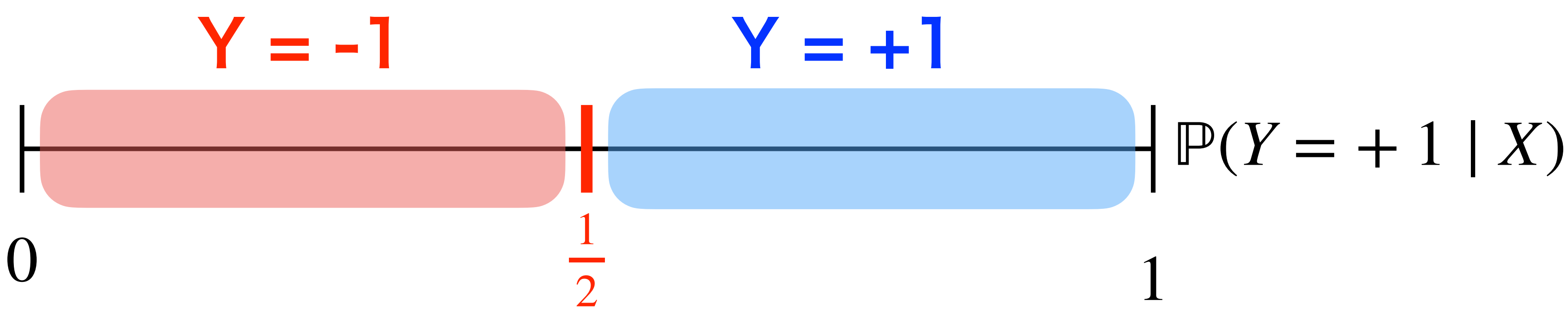
classifier $f : \mathcal{X} \to \mathbb{R}$
s.t. $U(f) = \sup_{f'} U(f')$

Related: Plug-in Classifier

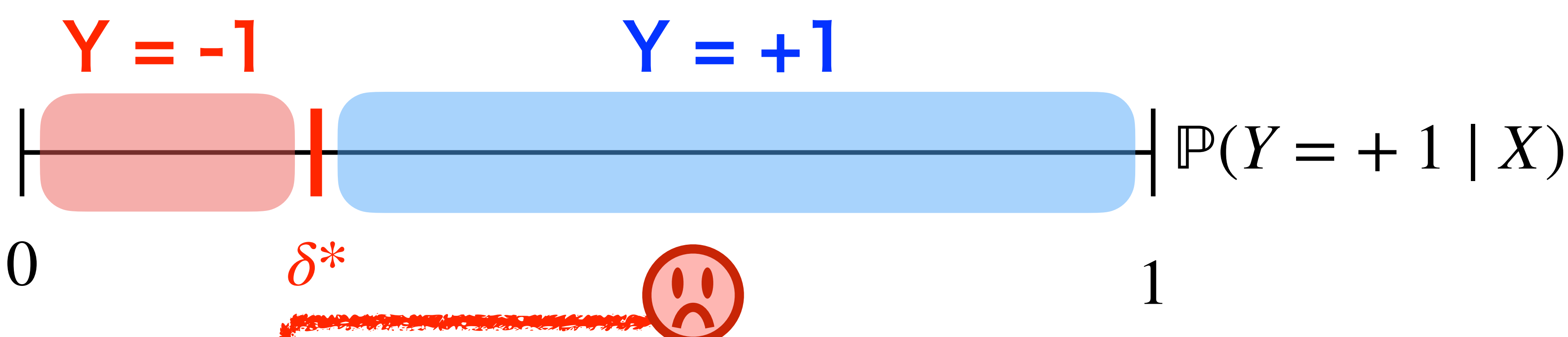
[Koyejo+ NIPS2014; Yan+ ICML2018]

- Estimating class-posterior probability is costly!

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



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



$\Rightarrow$ estimate $\mathbb{P}(Y = +1 \,|\, x)$ 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.

Formulation of Classification

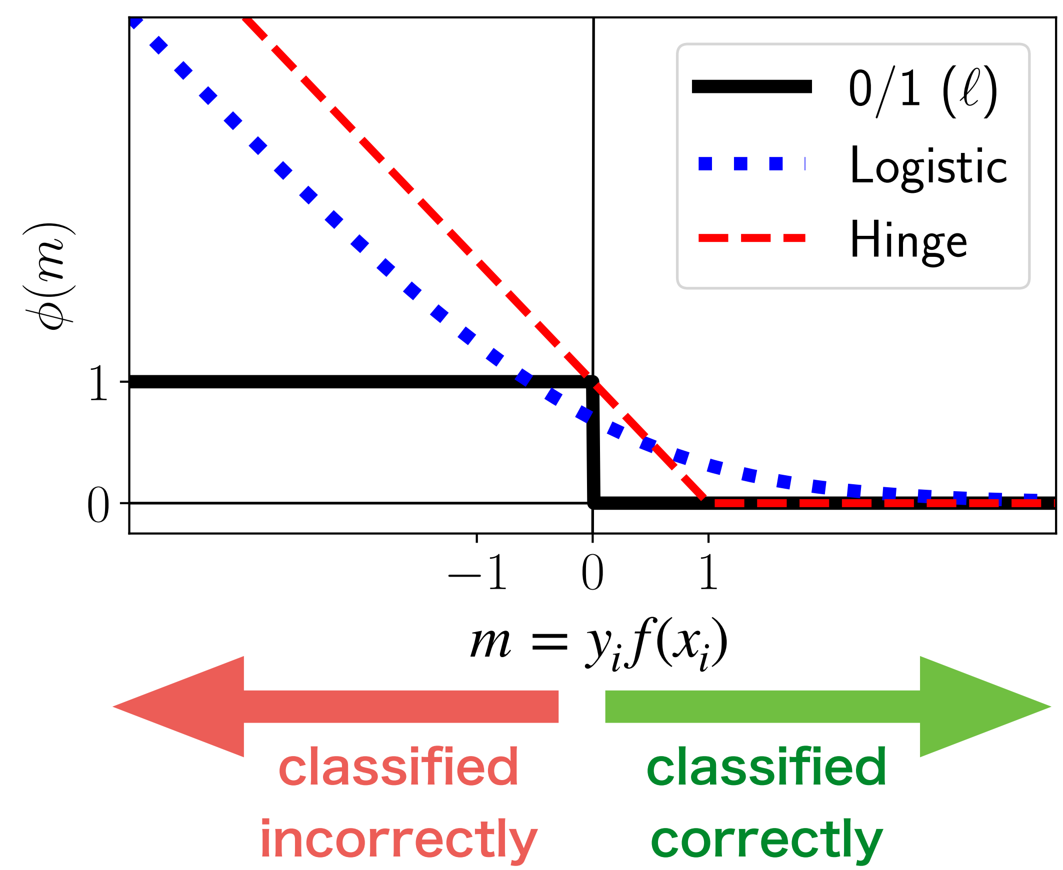
- Goal of classification: maximize accuracy
  = minimize mis-classification rate

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

↓ convexify 0/1 loss

(Empirical) Surrogate Risk

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

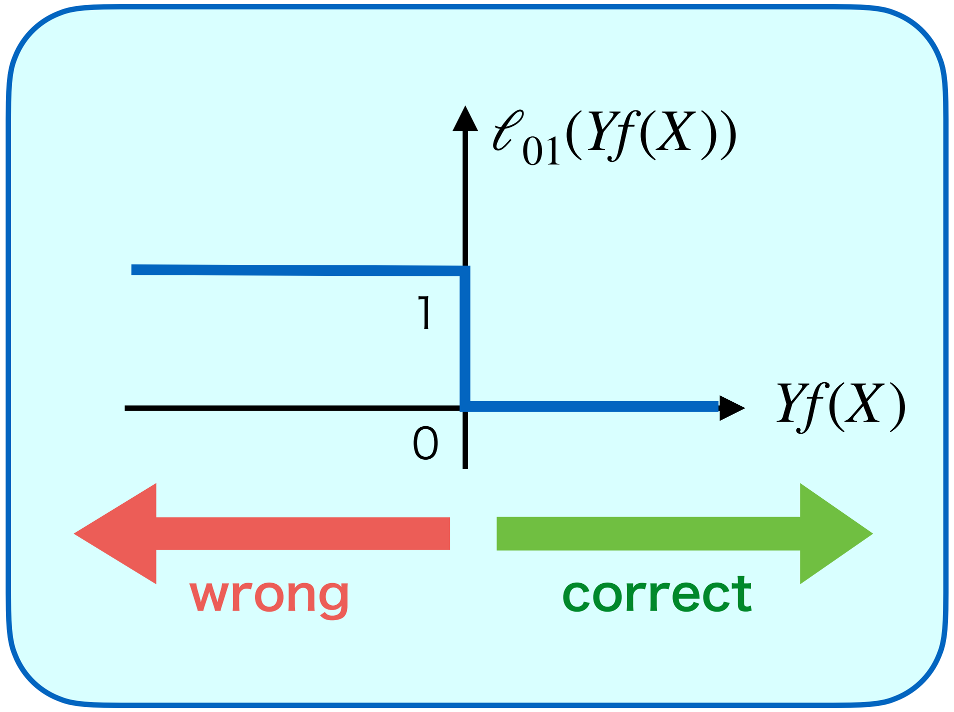


Example of ϕ

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

Target Loss and Surrogate Loss

0/1 loss (target loss)

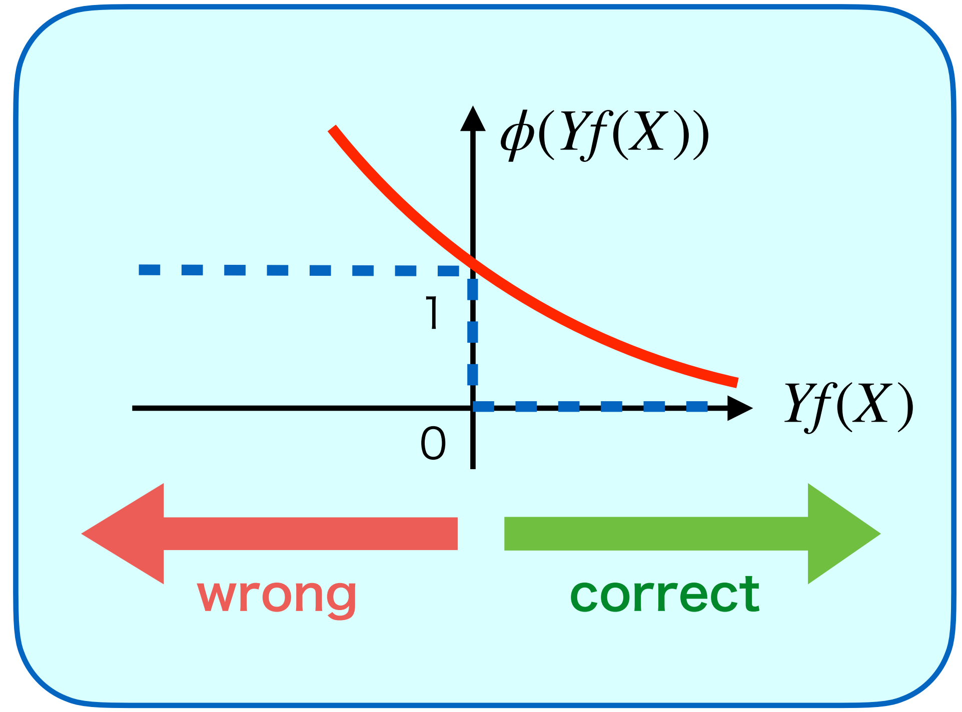


- Final learning criteria

$$R(f) = \mathbb{E}[\ell_{01}(Yf(X))]$$

- (Usually) hard to optimize

surrogate loss



- Easily-optimizable criteria

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

  - usually convex, smooth

Convexity & Statistical Property

tractable (convex)

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

intractable

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

Q. argmin R_ϕ = argmin R ?

A. Yes, w/ calibrated surrogate

Theorem. [Bartlett+ 2006]

Assume ϕ: convex.

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

Convexity & Statistical Property

Q. How to make tractable surrogate?

Accuracy

tractable (convex)

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

calibrated

intractable

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

Linear-fractional Metrics

① tractable?

? ? ?

② calibrated?

intractable

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

Non-concave, but quasi-concave

Idea: $\dfrac{\text{concave}}{\text{convex}}$ = 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 $\{x \mid f/g \geq \alpha\}$ is convex.

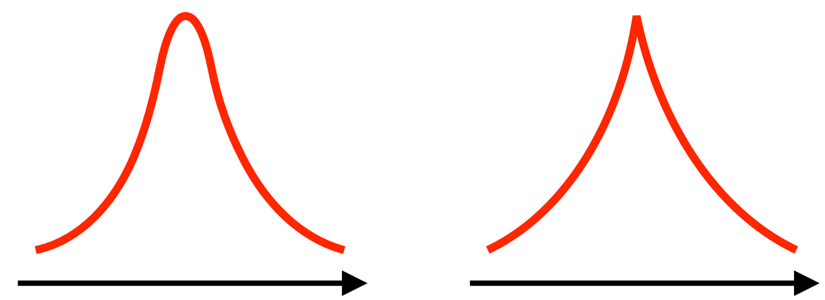
$$\frac{f(x)}{g(x)} \geq \alpha \iff \underbrace{f(x) - \alpha g(x)}_{\text{concave}} \geq 0$$

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

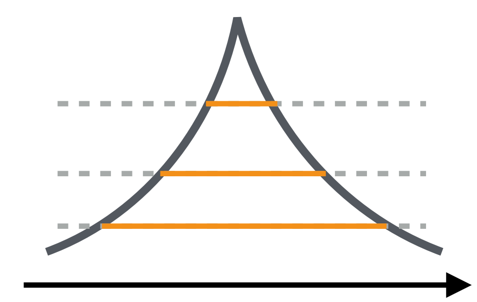
$\therefore \{x \mid f/g \geq \alpha\}$ is convex for $\forall \alpha \geq 0$

non-concave, but unimodal

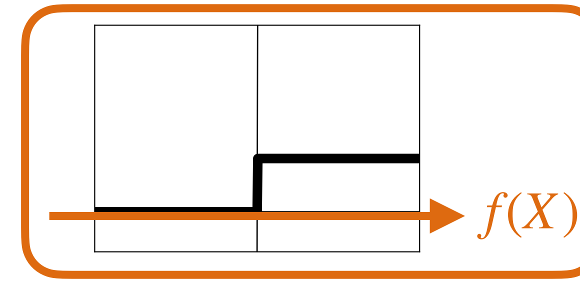
⇒ efficiently optimized



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



Surrogate Utility

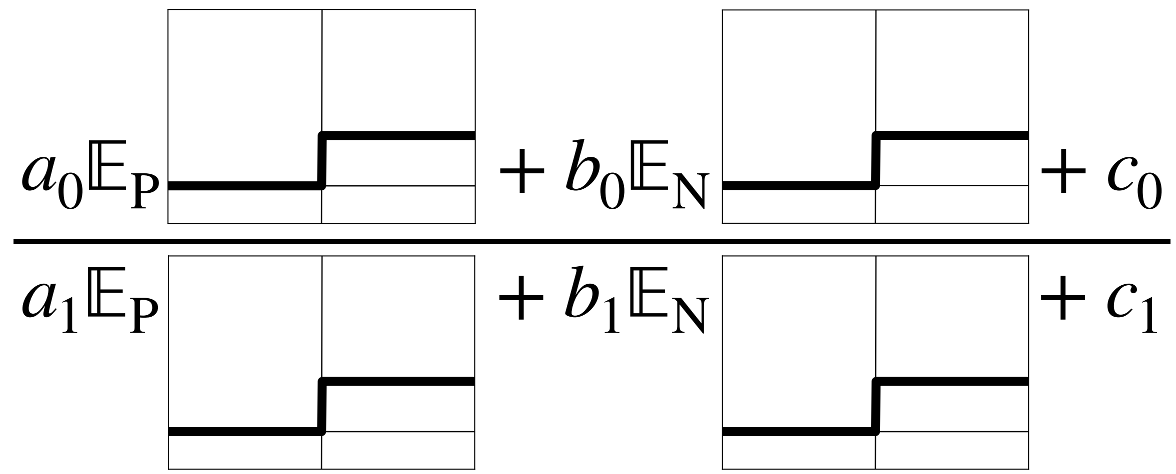


- Idea: bound true utility from below

linear-fraction

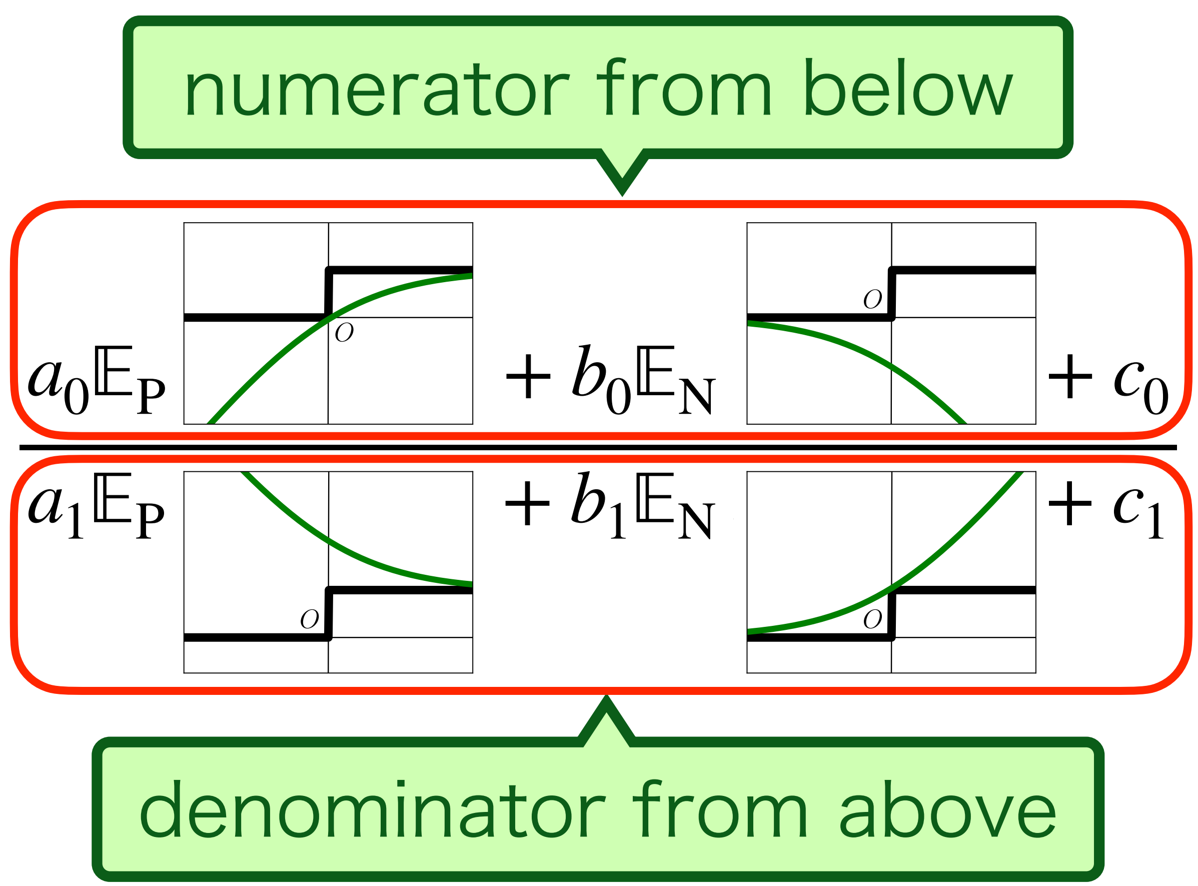
$$U(f) = \frac{a_0 \mathsf{TP} + b_0 \mathsf{FP} + c_0}{a_1 \mathsf{TP} + b_1 \mathsf{FP} + c_1}$$

=



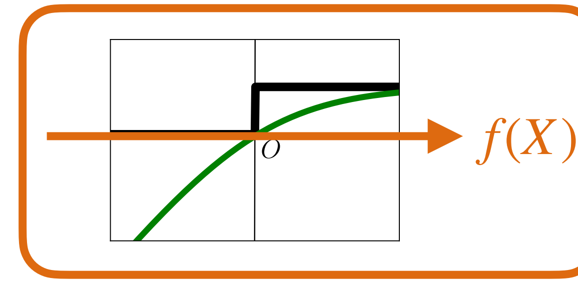
non-negative sum of concave
⇒ concave

≥



non-negative sum of convex
⇒ convex

Surrogate Utility

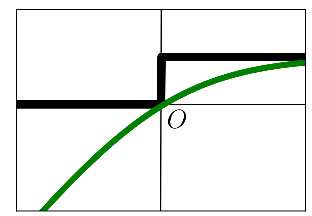


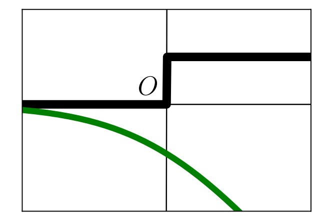
- Idea: bound true utility from below

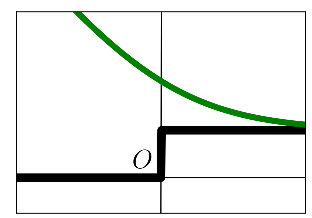
linear-fraction

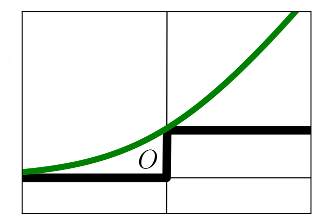
$$U(f) = \frac{a_0\mathrm{TP} + b_0\mathrm{FP} + c_0}{a_1\mathrm{TP} + b_1\mathrm{FP} + c_1}$$

$$\geq \frac{a_0\mathbb{E}_\mathrm{P}[\;\cdot\;] + b_0\mathbb{E}_\mathrm{N}[\;\cdot\;] + c_0}{a_1\mathbb{E}_\mathrm{P}[\;\cdot\;] + b_1\mathbb{E}_\mathrm{N}[\;\cdot\;] + c_1}$$



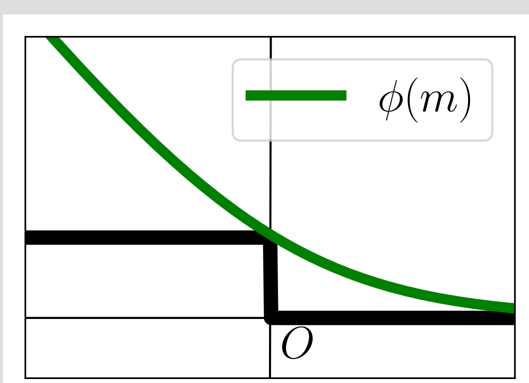






||

surrogate loss



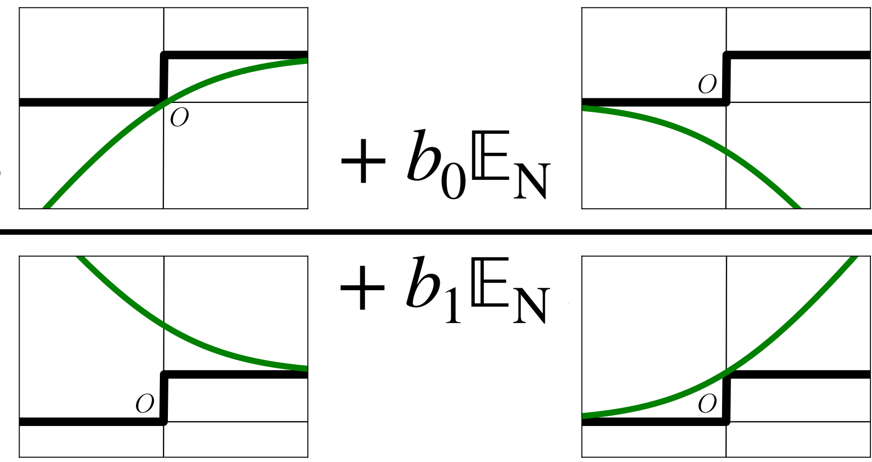
$$U_\phi(f) = \frac{a_0\mathbb{E}_\mathrm{P}[1-\phi(f(X))] + b_0\mathbb{E}_\mathrm{N}[-\phi(-f(X))] + c_0}{a_1\mathbb{E}_\mathrm{P}[1+\phi(f(X))] + b_1\mathbb{E}_\mathrm{N}[\;\phi(-f(X))\;] + c_1}$$

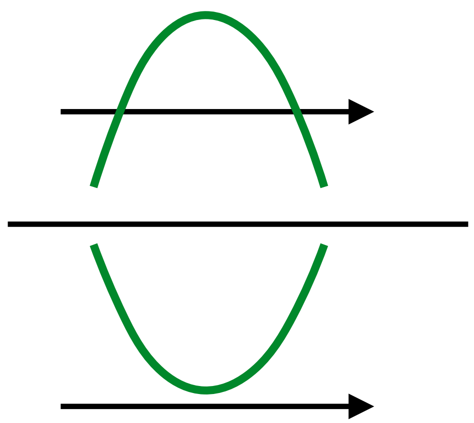
$$:= \frac{\mathbb{E}[W_{0,\phi}]}{\mathbb{E}[W_{1,\phi}]}$$

: Surrogate Utility

Hybrid Optimization Strategy

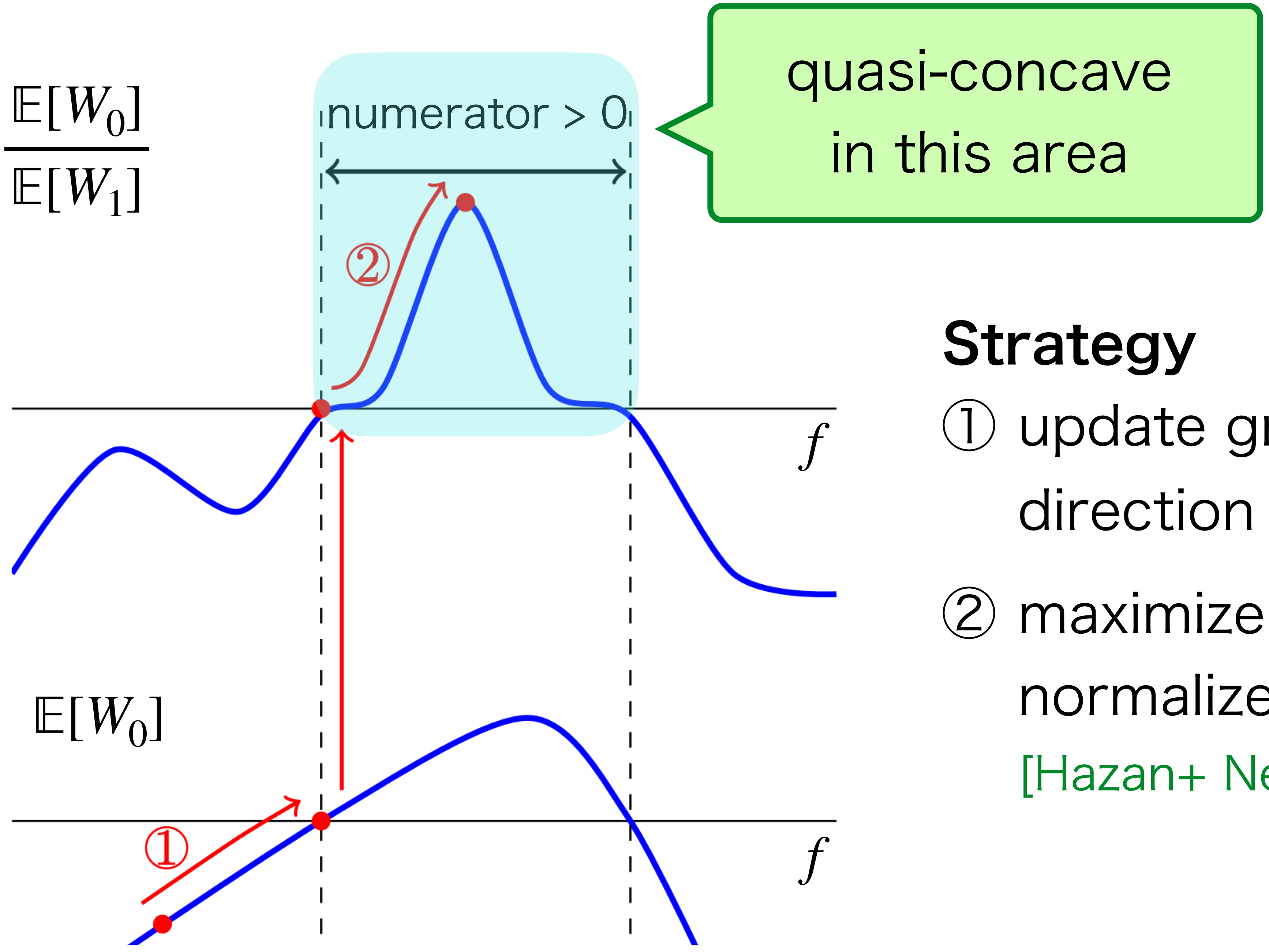
$$U_\phi(f) = \frac{a_0 \mathbb{E}_{\mathrm{P}} + b_0 \mathbb{E}_{\mathrm{N}} + c_0}{a_1 \mathbb{E}_{\mathrm{P}} + b_1 \mathbb{E}_{\mathrm{N}} + c_1} =$$





- Note: numerator can be negative
  - U_ϕ isn't quasi-concave only if numerator < 0
  - make numerator positive first (concave), then maximize fractional form (quasi-concave)

Hybrid Optimization Strategy



numerator is always concave

Strategy

① update gradient-ascent direction while $\mathbb{E}[W_0] < 0$

② maximize fraction by normalized-gradient ascent [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).

Convexity & Statistical Property

Q. How to make surrogate calibrated?

Accuracy

tractable (convex)

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

↓ calibrated

intractable

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

Linear-fractional Metrics

① tractable?

? ? ?

↓ ② calibrated?

intractable

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

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: F1-measure

merely sufficient!

Theorem

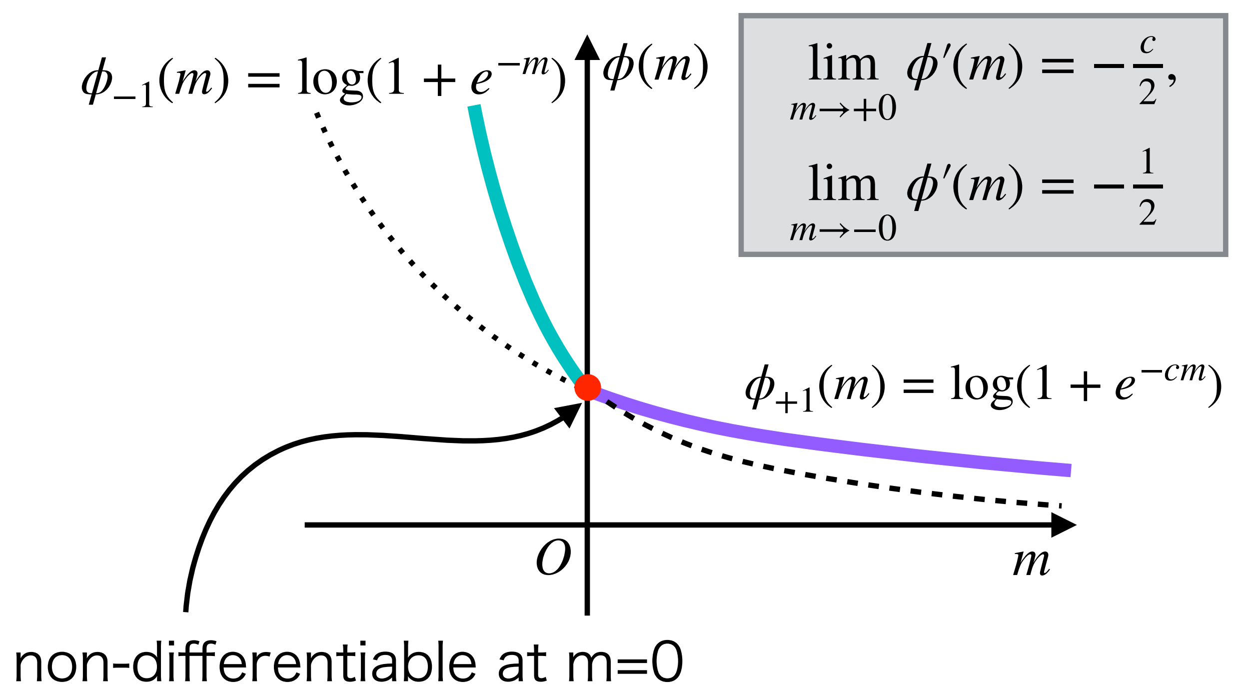
$$U_\phi(f_n) \overset{n\to\infty}{\to} 1 \implies U(f_n) \overset{n\to\infty}{\to} 1 \quad \forall \{f_n\}$$

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



Intuition:
trade off TP and FP
by gradient steepness

Experiment: F_1-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: $f_\theta(x) = \theta^\top x$

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

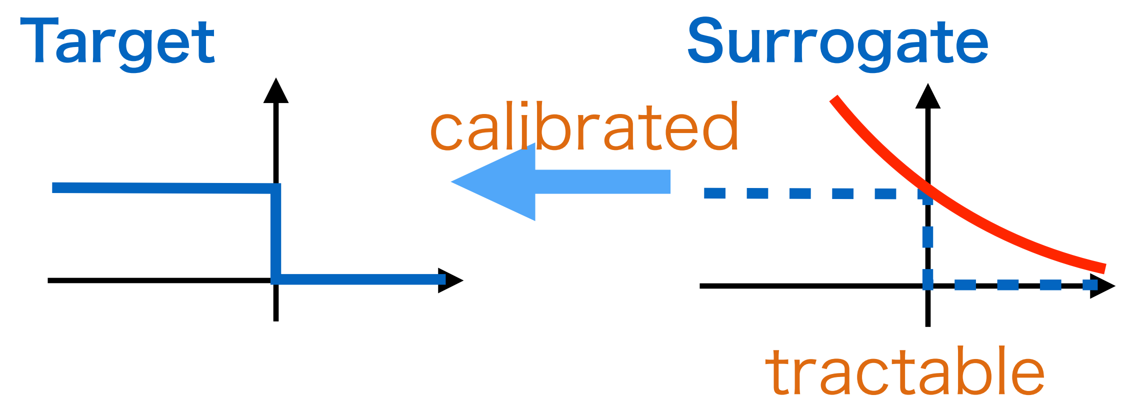
Summary: Calibrated and Tractable Surrogate for Class-imbalance

Goal

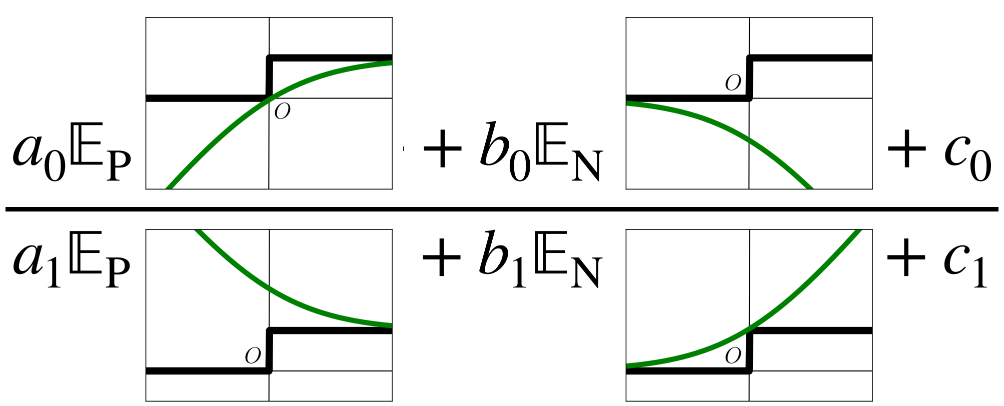
maximize linear-fractional utility

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

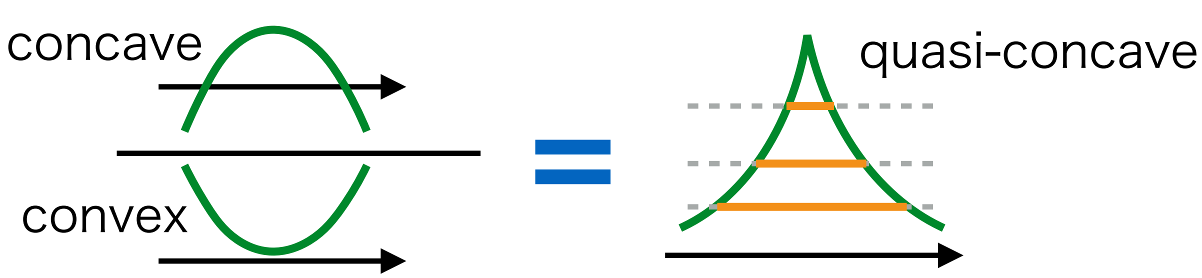
In usual binary classification...



Tractable Optimization

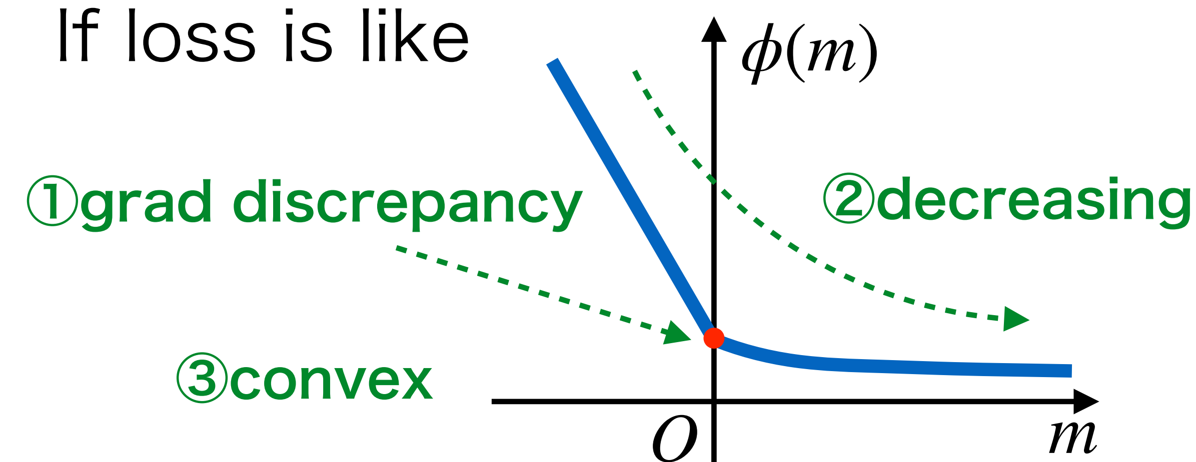


quasi-concave optimization



Calibrated Surrogate

If loss is like



then

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