損失関数をつなぐ学習理論

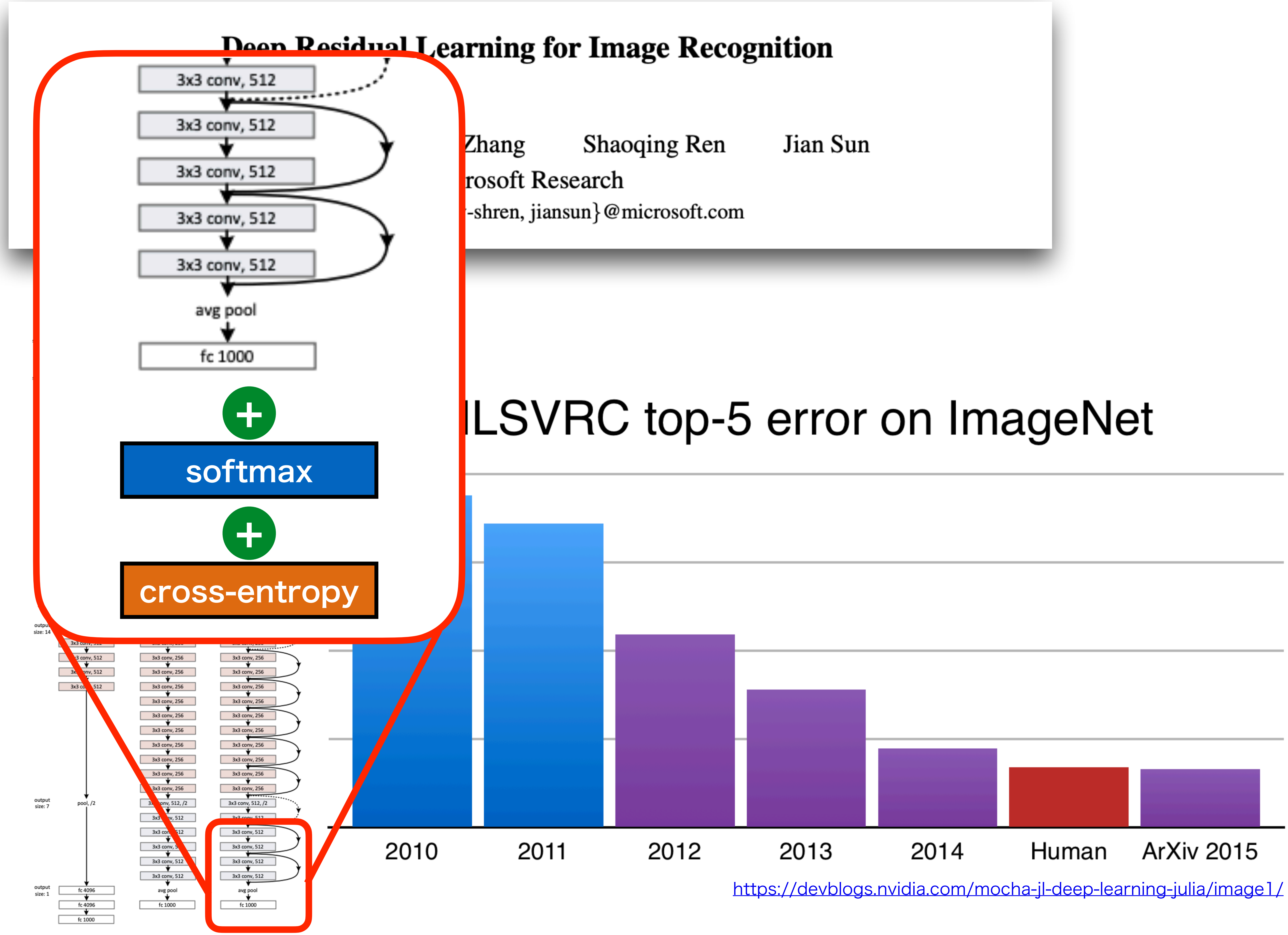
ソシオグローバル情報工学研究センター講演会 2020年9月7日

情報理工学系研究科 博士2年 包 含（つつみ ふくむ / Bao Han）

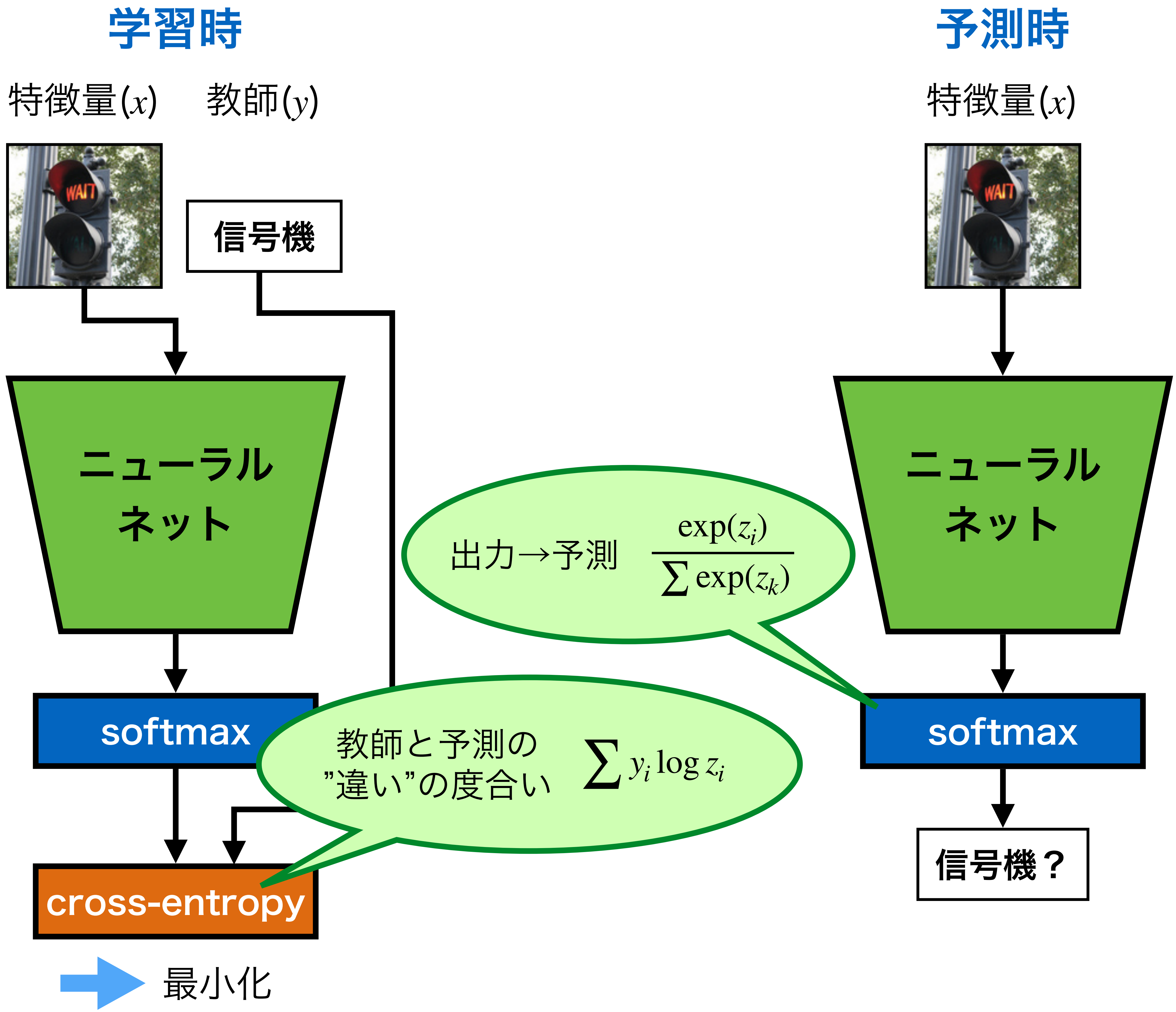
自己紹介 | 包 含（つつみ ふくむ）

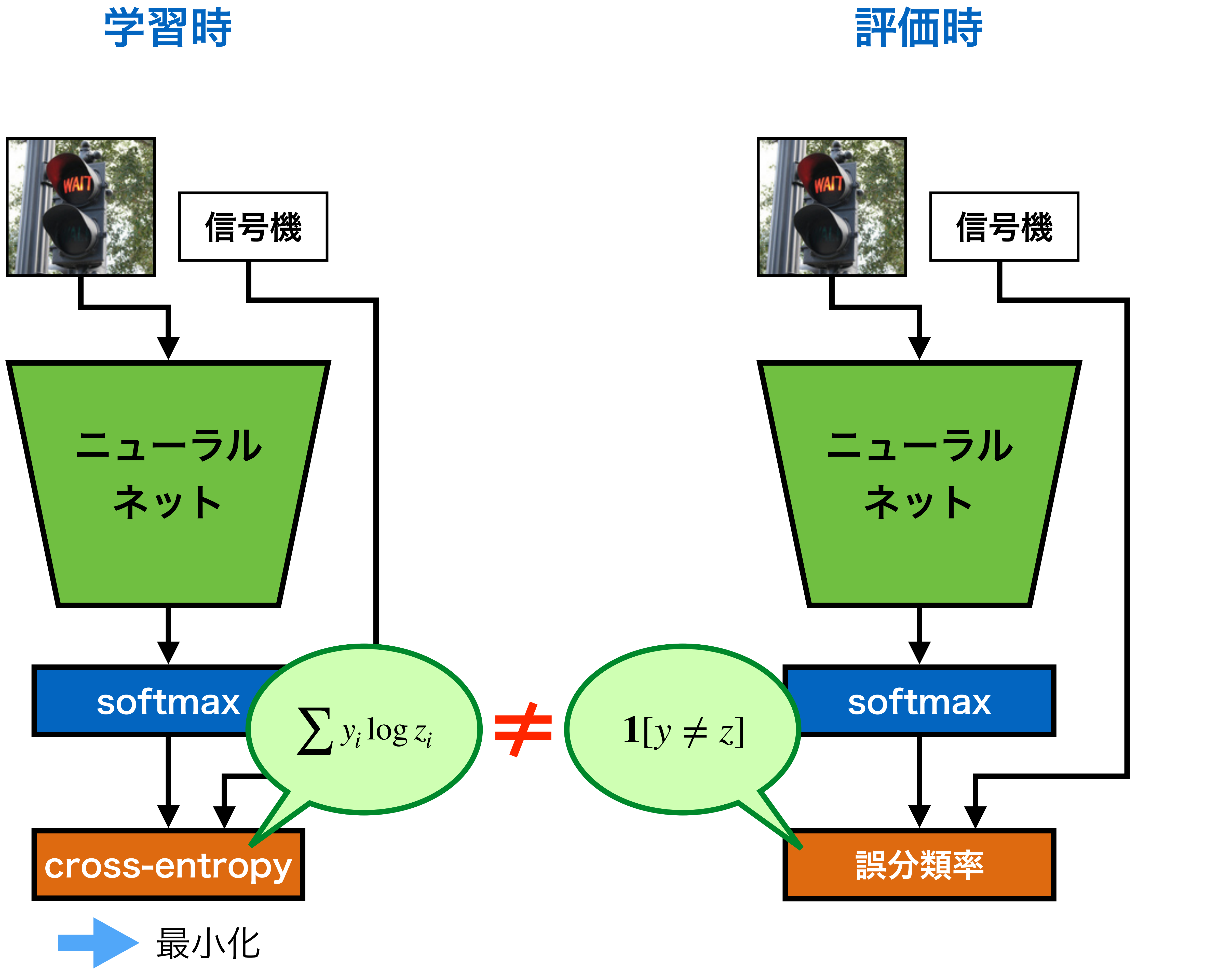
https://hermite.jp/

- 2013 - 2017 東京大学 理学部情報科学科
  - 2014/8 松浦研究室でインターン
- 2017 - 東京大学大学院 情報理工学系研究科 コンピュータ科学専攻
  - 博士2年
  - 専門: 機械学習（損失関数の理論や転移学習など）
- その他
  - 2018/10 - 2020/3 JST ACT-I 研究者
  - 2019/10 - 2020/2 米ミシガン大学にて研究滞在



https://devblogs.nvidia.com/mocha-jl-deep-learning-julia/image1/

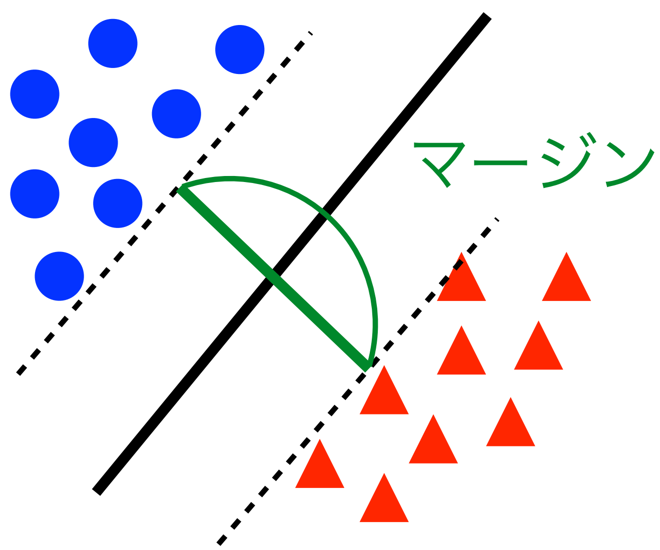

学習時
特徴量(x)
教師(y)
信号機
ニューラル
ネット
softmax
cross-entropy
最小化
予測時
特徴量(x)
ニューラル
ネット
softmax
信号機？
出力→予測
$\frac{\exp(z_i)}{\sum \exp(z_k)}$
教師と予測の
"違い"の度合い
$\sum y_i \log z_i$


学習時
評価時
信号機
WAIT
ニューラル
ネット
softmax
cross-entropy
$\sum y_i \log z_i$
≠
$\mathbf{1}[y \neq z]$
誤分類率
最小化

Machine Learning, 20, 273–297 (1995)

Support-Vector Networks

CORINNA CORTES corinna@neural.att.com
VLADIMIR VAPNIK vlad@neural.att.com
AT&T Bell Labs., Holmdel, NJ 07733, USA



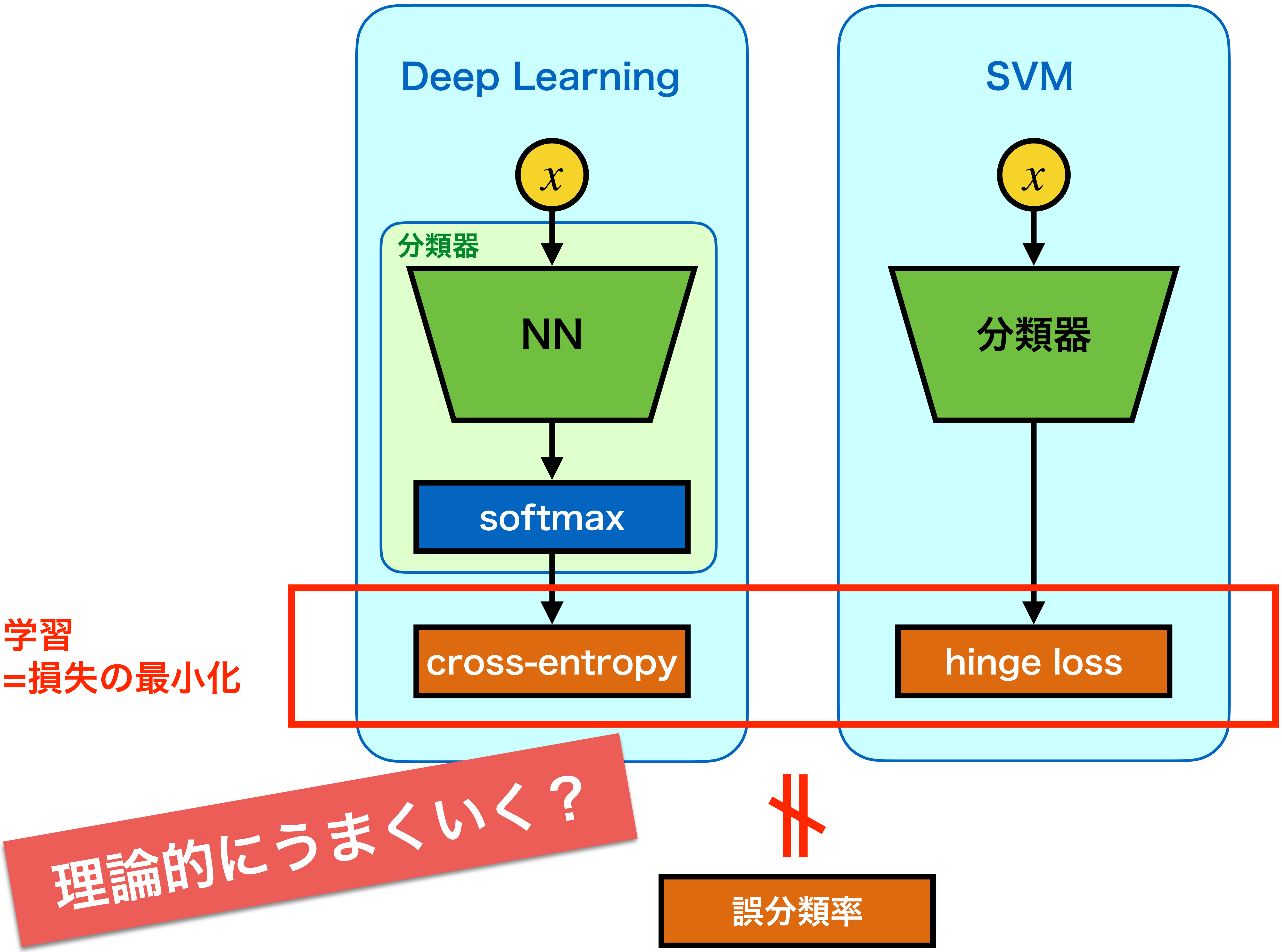
マージン最大化

$$=\ \min_{w,b} \sum_i \max\left\{0,\ 1 - y_i(w^\top x_i + b)\ \right\}$$

hinge lossの最小化

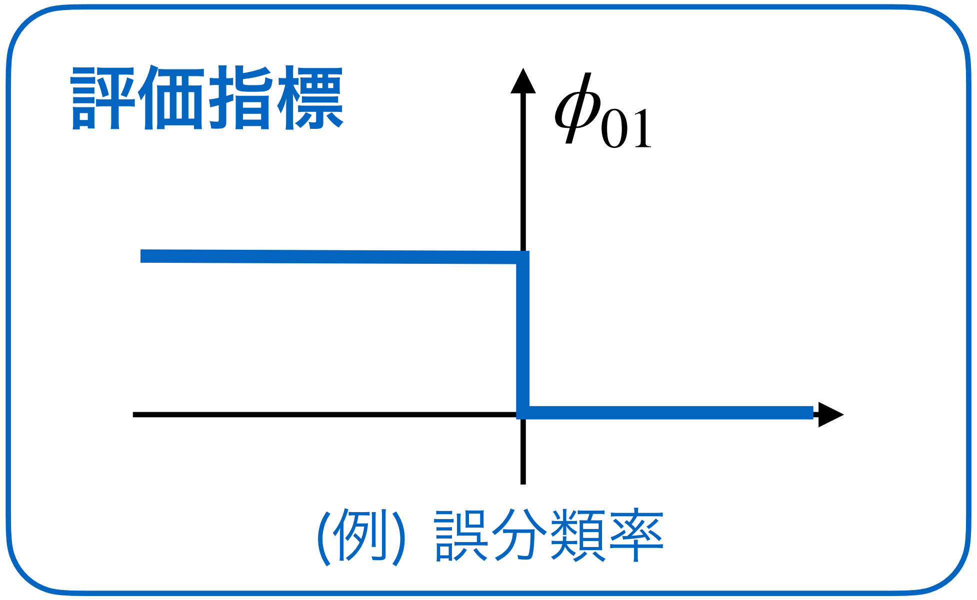
$\neq$

誤分類率の最小化

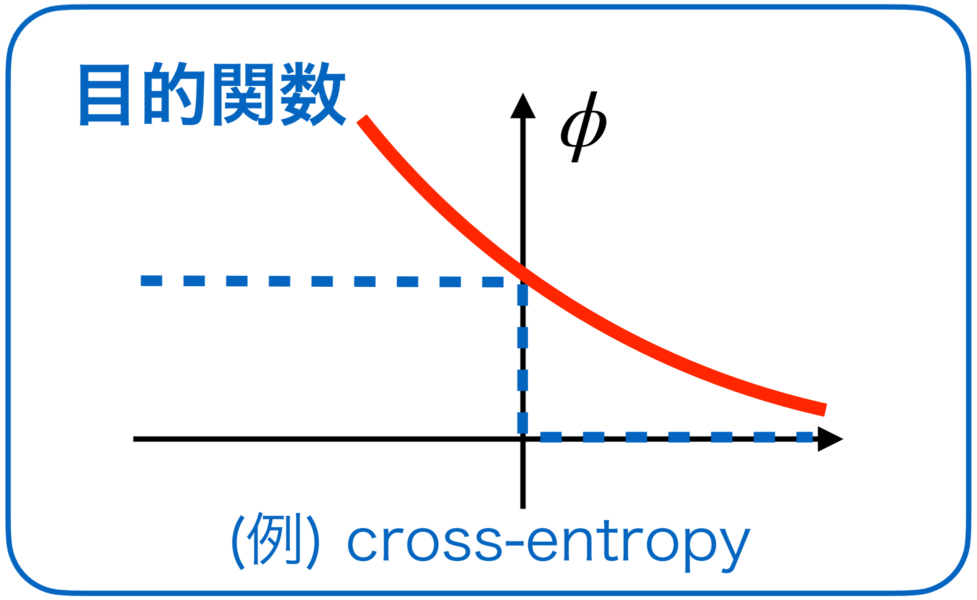

Deep Learning
SVM
x
x
分類器
NN
分類器
softmax
学習
=損失の最小化
cross-entropy
hinge loss
理論的にうまくいく？
≉
誤分類率

講演の目的

- 統計的学習理論の一端を紹介



損失関数の
関係性

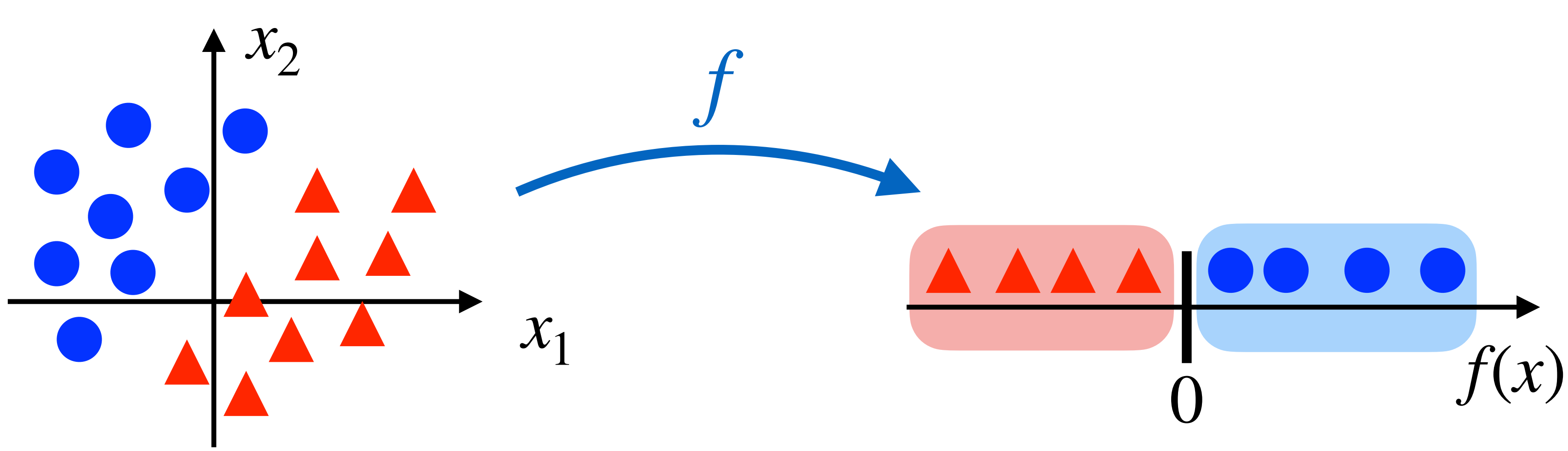


  - 機械学習を用いる際の指針として役立つように
- 応用研究の誘発
  - 応用領域の要請から 新たな評価指標 が考えられるかも？
  - 新たな評価指標 と既存の損失関数の関係性は？

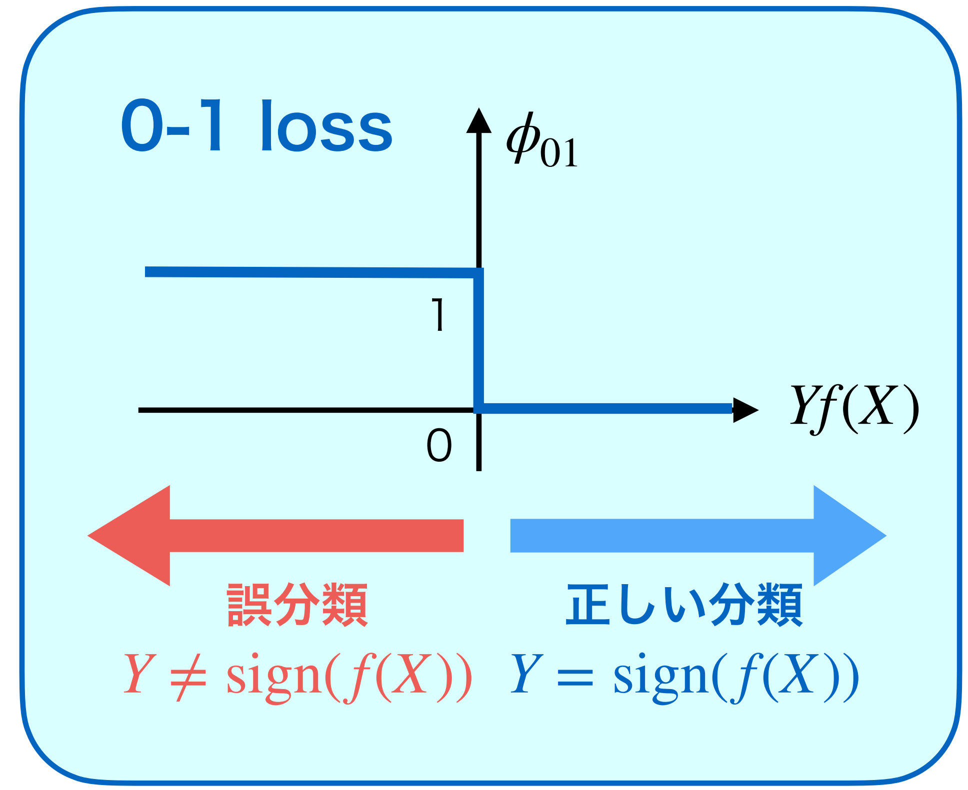
二値分類問題

- 入力
  - サンプル $\{(x_i, y_i)\}_{i=1}^n$: 特徴量 $x_i \in \mathcal{X}$ とラベル $y_i \in \{\pm 1\}$ の組
- 出力
  - 分類器 $f : \mathcal{X} \to \mathbb{R}$ の学習
  - $\mathrm{sign}(f(\cdot))$ を用いてラベルを予測
  - 基準: 誤分類率 $R_{01}(f) = \mathbb{E}\left[\mathbf{1}[Y \neq \mathrm{sign}(f(X))]\right]$

$Y \neq \mathrm{sign}(f(X))$ なら 1、
$Y = \mathrm{sign}(f(X))$ なら 0



二値分類問題



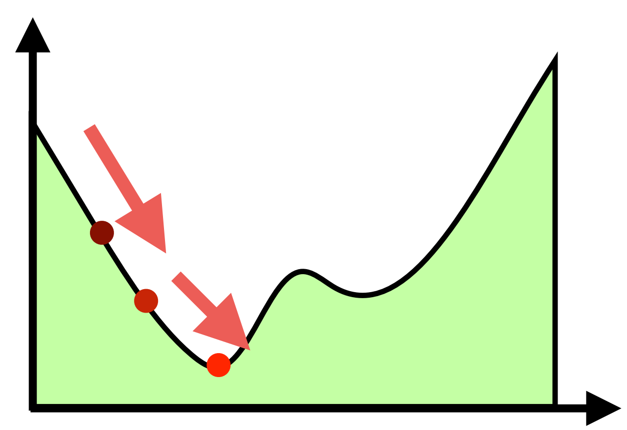
- 二値分類の真のゴール: 誤分類率の最小化

$$R_{01}(f) = \mathbb{E}\left[\mathbf{1}[Y \neq \mathrm{sign}(f(X))]\right]$$

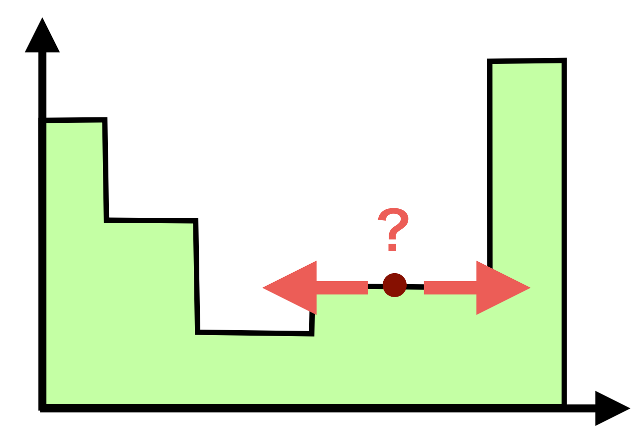
- 誤分類率 = 0-1 lossの期待値

$$\mathbf{1}[Y \neq \mathrm{sign}(f(X))] = \phi_{01}(Yf(X))$$

- 誤分類率の最小化はNP困難 [Feldman+ 2012]



最小化 = 勾配降下方向に点を更新

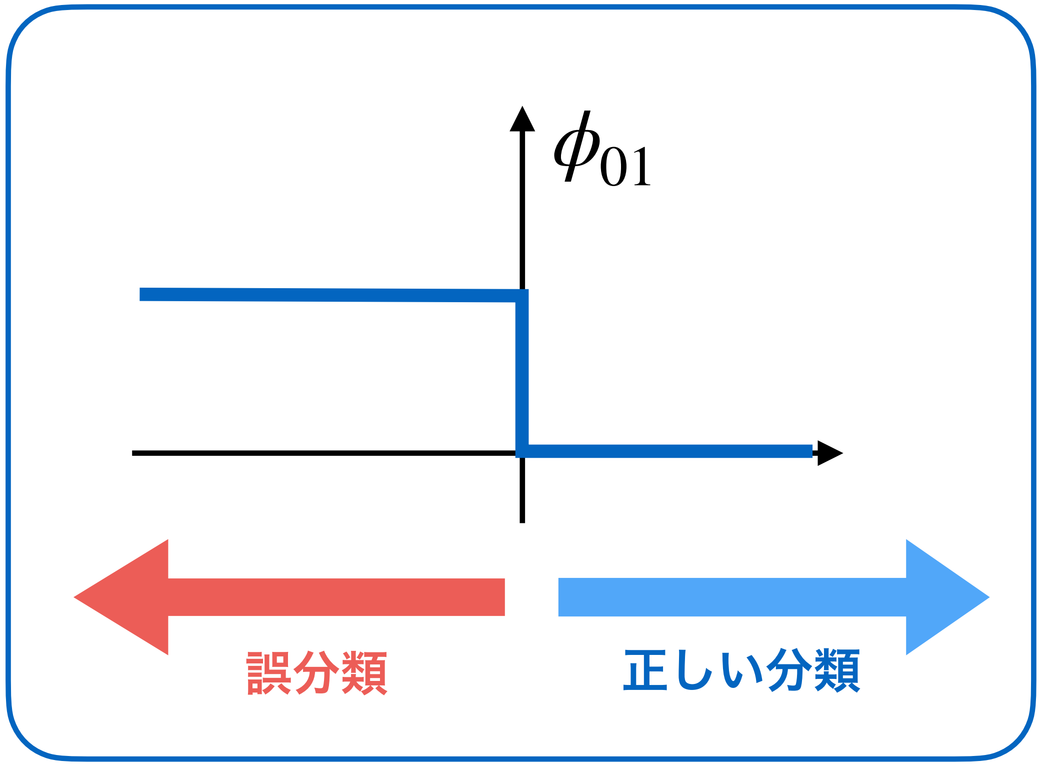


離散関数は勾配がない

Feldman, V., Guruswami, V., Raghavendra, P., & Wu, Y. (2012).
Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, *41*(6), 1558-1590.

target loss / surrogate loss
評価損失と代理損失

0-1 loss (target loss)

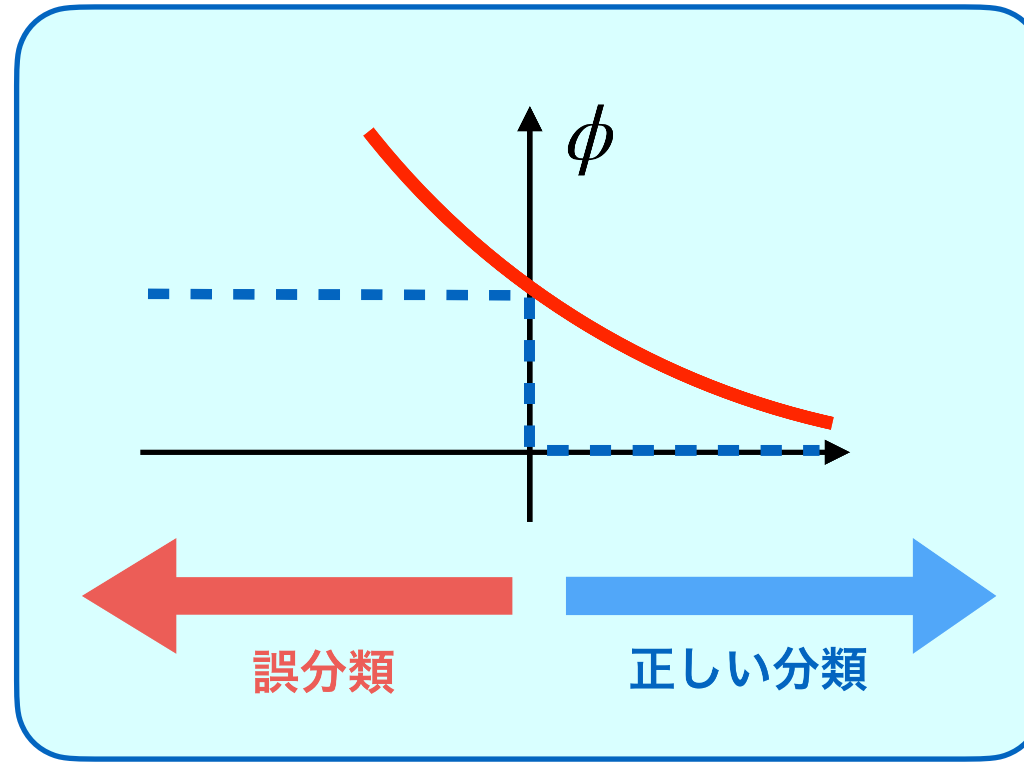


- 最終的な評価指標

$$R_{01}(f) = \mathbb{E}[\phi_{01}(Yf(X))]$$

- 最適化が困難

surrogate loss



- 最適化の容易な関数で置換

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

  - 凸上界、滑らかな関数、etc.
  - logistic loss, hinge loss, etc.

学習理論ことはじめ

(empirical)
surrogate risk

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

汎化誤差の理論: (大雑把に言えば)
モデルが複雑過ぎなければ収束
よくある学習理論の話

(population)
surrogate risk

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

今回の鍵:
損失関数の **適合性** 理論
calibration

target risk

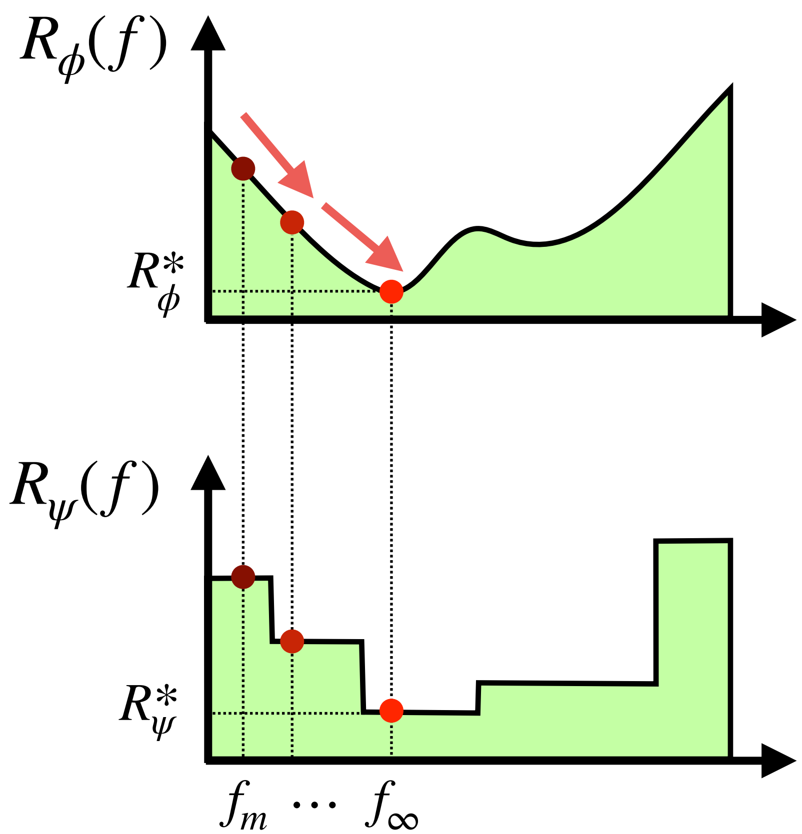
$$R_{01}(f) = \mathbb{E}[\phi_{01}(Yf(X))]$$

妥当な代理損失 ϕ とは？

[Steinwart 2007]

A. R_ϕ の最小化が R_ψ の最小化を誘導する ϕ

(surrogate: R_ϕ, target: R_ψ)



$R_\phi(f_m) \overset{m\to\infty}{\to} R_\phi^*$ となる列 $\{f_m\}_{m\geq 1}$ について

$R_\psi(f_m) \overset{m\to\infty}{\to} R_\psi^*$ が成立

定義. 適合的損失 (calibrated surrogate loss)

任意の f と $\varepsilon > 0$ についてある $\delta > 0$ が存在して以下が成り立つとき、surrogate ϕ は target ψ に対して**適合的**という。

$$R_\phi(f) < R_\phi^* + \delta \quad \Longrightarrow \quad R_\psi(f) < R_\psi^* + \varepsilon.$$

limの定義を
ε-δ で言い換え

Steinwart, I. (2007). How to compare different loss functions and their risks. *Constructive Approximation*, *26*(2), 225-287.

どうやって適合性を確認する？

Idea: 条件を満たす δ を ε の関数として書く

定義. 適合的損失 (calibrated surrogate loss)

任意の f と $\varepsilon > 0$ についてある $\delta > 0$ が存在して以下が成り立つとき、surrogate ϕ は target ψ に対して**適合的**という。

$$R_\phi(f) < R_\phi^* + \delta \quad \Longrightarrow \quad R_\psi(f) < R_\psi^* + \varepsilon.$$

定義. 適合関数 (calibration function)

$$\delta(\varepsilon) = \inf_f \ R_\phi(f) - R_\phi^* \quad \text{s.t.} \quad R_\psi(f) - R_\psi^* \geq \varepsilon$$

制約付き変分問題として解ける / 詳細略
(cf. [Steinwart 2007; Osokin+ 2017] など)

- ϕ は適合的 $\Leftrightarrow$ すべての $\varepsilon > 0$ について $\delta(\varepsilon) > 0$

Steinwart, I. (2007). How to compare different loss functions and their risks. *Constructive Approximation*, *26*(2), 225-287.
Osokin, A., Bach, F., & Lacoste-Julien, S. (2017).
On structured prediction theory with calibrated convex surrogate losses. In *NeurIPS*.

適合性理論

任意の f と $\varepsilon > 0$ についてある $\delta > 0$ が存在して以下が成り立つとき、surrogate ϕ は target ψ に対して**適合的**という。

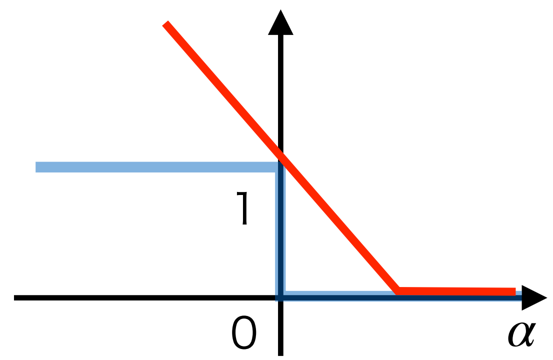
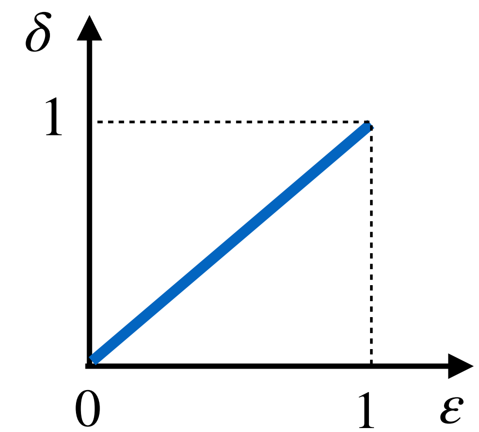
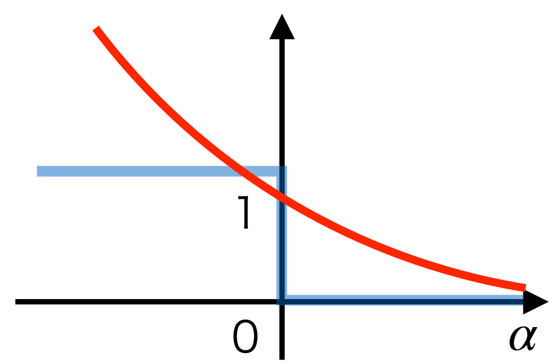
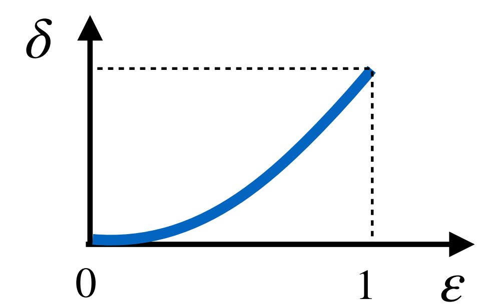
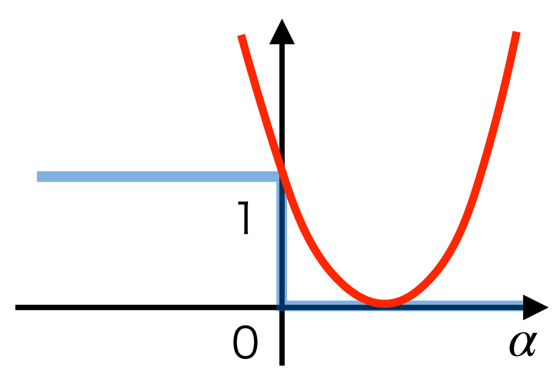
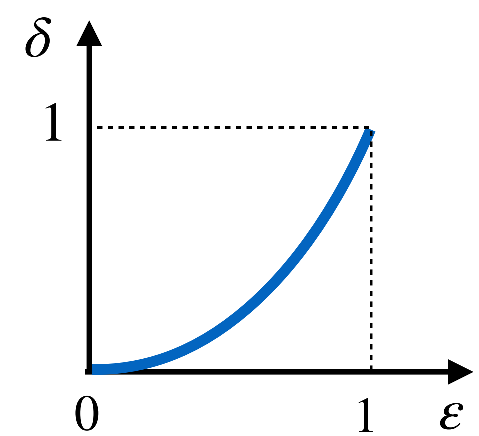
$$R_\phi(f) < R_\phi^* + \delta \quad \Longrightarrow \quad R_\psi(f) < R_\psi^* + \varepsilon.$$

適合関数 $\delta(\varepsilon) = \inf_f R_\phi(f) - R_\phi^*$ s.t. $R_\psi(f) - R_\psi^* \geq \varepsilon$

- 損失関数を**「定性的」**につなぐ
  - すべての $\varepsilon > 0$ について $\delta(\varepsilon) > 0 \Rightarrow$ 適合的
- 損失関数を**「定量的」**につなぐ
  - (適合関数の定義から) 任意の f について $\delta(R_\psi(f) - R_\psi^*) \leq R_\phi(f) - R_\phi^*$
  - δ が可逆なら $R_\psi(f) - R_\psi^* \leq \delta^{-1}(R_\phi(f) - R_\phi^*)$

surrogate riskが減少したら
target riskがどれくらい減少するか

代理損失の例

	損失の形	適合関数
hinge loss $\phi_{\mathrm{hinge}}(\alpha) = \max\{0,1-\alpha\}$		 $\delta(\varepsilon) = \varepsilon$
logistic loss $\phi_{\log}(\alpha) = \ln(1+e^{-\alpha})$		 $\delta(\varepsilon) = \dfrac{(1+\varepsilon)\ln(1+\varepsilon)+(1-\varepsilon)\ln(1-\varepsilon)}{2}$
squared loss $\phi_{\mathrm{sq}}(\alpha) = (1-\alpha)^2$		 $\delta(\varepsilon) = \varepsilon^2$

凸損失の場合

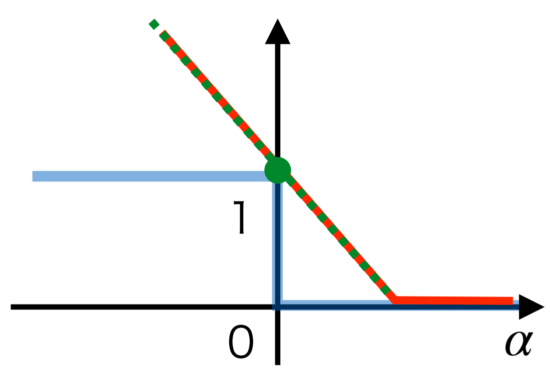
[Bartlett+ 2006]

■ 適合性の必要十分条件が簡潔に記述可能

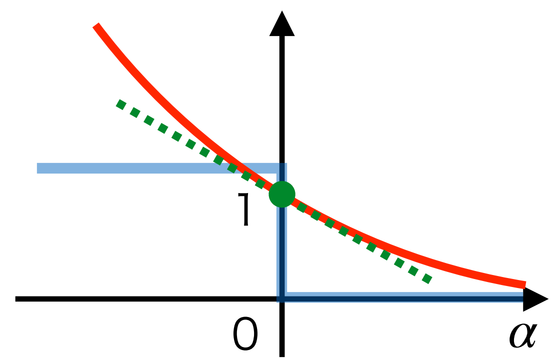
定理. Surrogate ϕ が凸関数なら以下の場合に限り0-1 lossに対して適合的

- 原点で微分可能
- $\phi'(0) < 0$

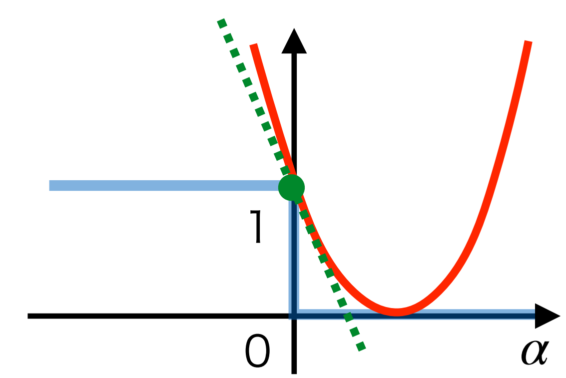
hinge loss



logistic loss



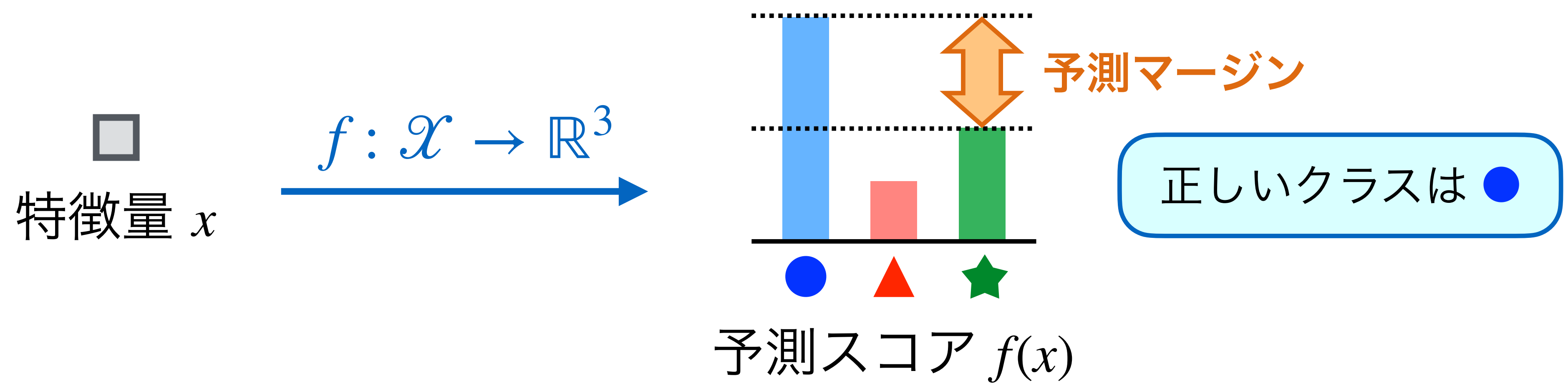
squared loss



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.

適合性は必ずしも直感的ではない

- 例: 多値分類
  - 予測マージン最大化として定式化



Crammer-Singer loss [Crammer & Singer 2001]

$\max\{0, 1 - \text{予測マージン}\}$

hinge lossの多値拡張のひとつ

Crammer-Singer lossは0-1 lossに対して適合的でない！

logistic lossの同様な拡張なら適合的

[Zhang 2004]

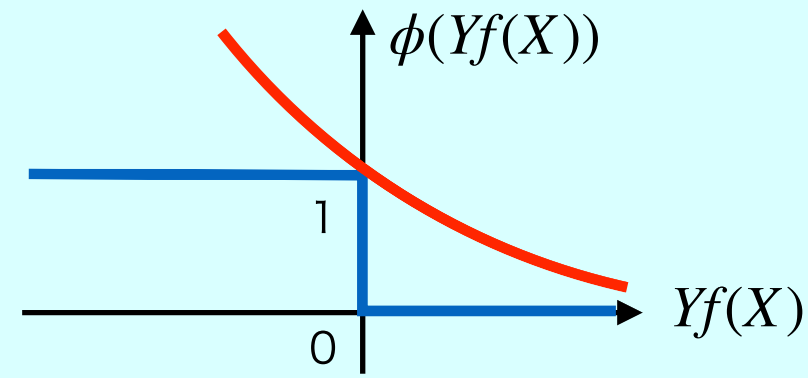
Crammer, K., & Singer, Y. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, *2*(Dec), 265-292

Zhang, T. (2004). Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, *5*(Oct), 1225-1251.

損失関数をつなぐ理論

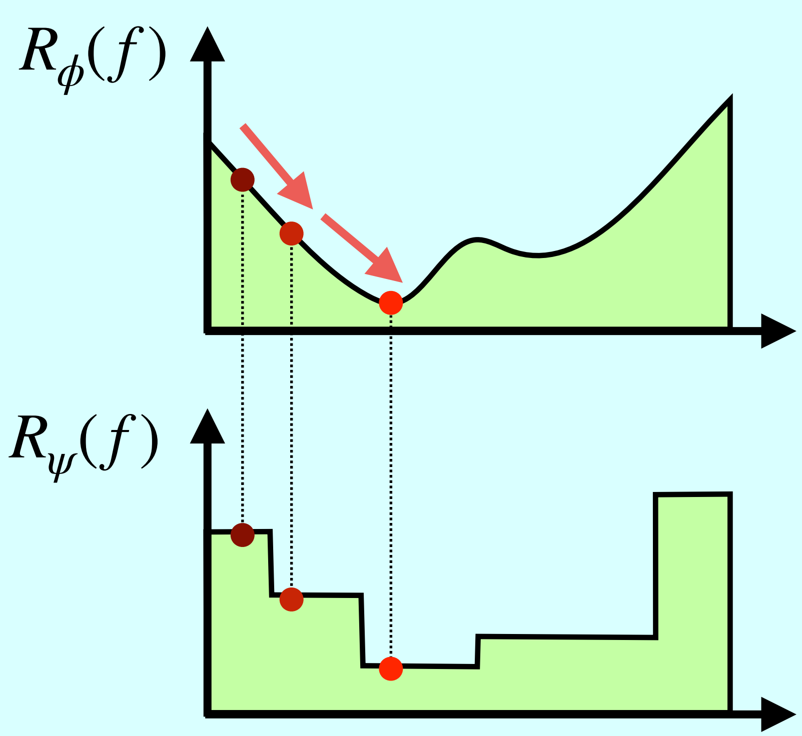
Surrogate vs. Target loss

評価損失 (target) は最適化がしばしば困難
⇒ 代理損失 (surrogate) で置換



適合的損失

targetの最小化が
保証されるsurrogate



二値分類

Hinge, logisticなどが適合的
$\phi'(0) < 0$ が適合性に必要十分

多値分類

CS-loss (多値hinge loss) は
適合的でない！

cross-entropyは適合的 (詳細略)

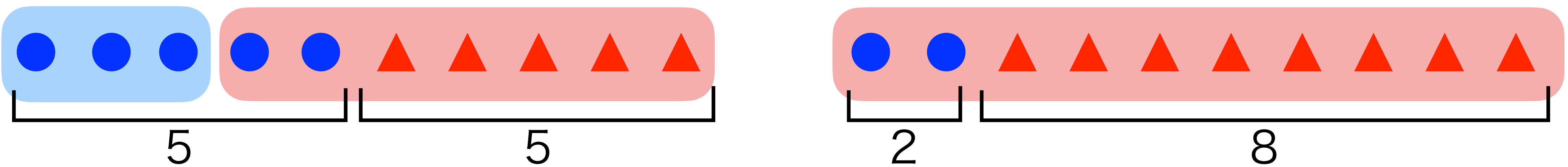
代理損失の妥当性が厳密に議論可能に！

評価損失が0-1 lossでないとき

H. Bao and M. Sugiyama.
Calibrated Surrogate Maximization of Linear-fractional Utility in Binary Classification. In *AISTATS*, 2020.

分類正答率は適切？

- 例: 二値分類

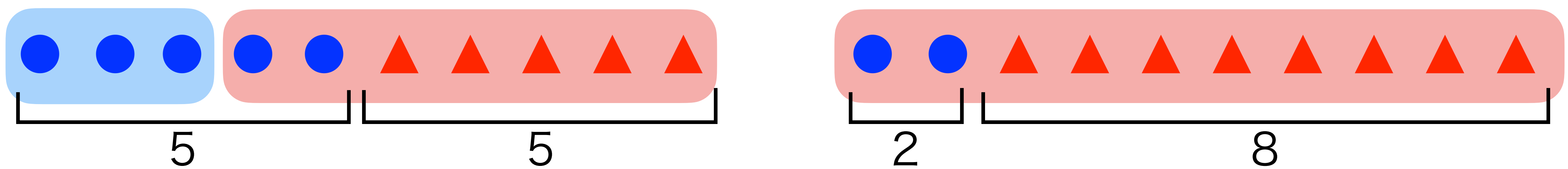


正答率: **0.8**

正答率: **0.8**

医療診断などでは重大な問題に！

分類正答率は適切？



正答率: **0.8**

F値: **0.75**

正答率: **0.8**

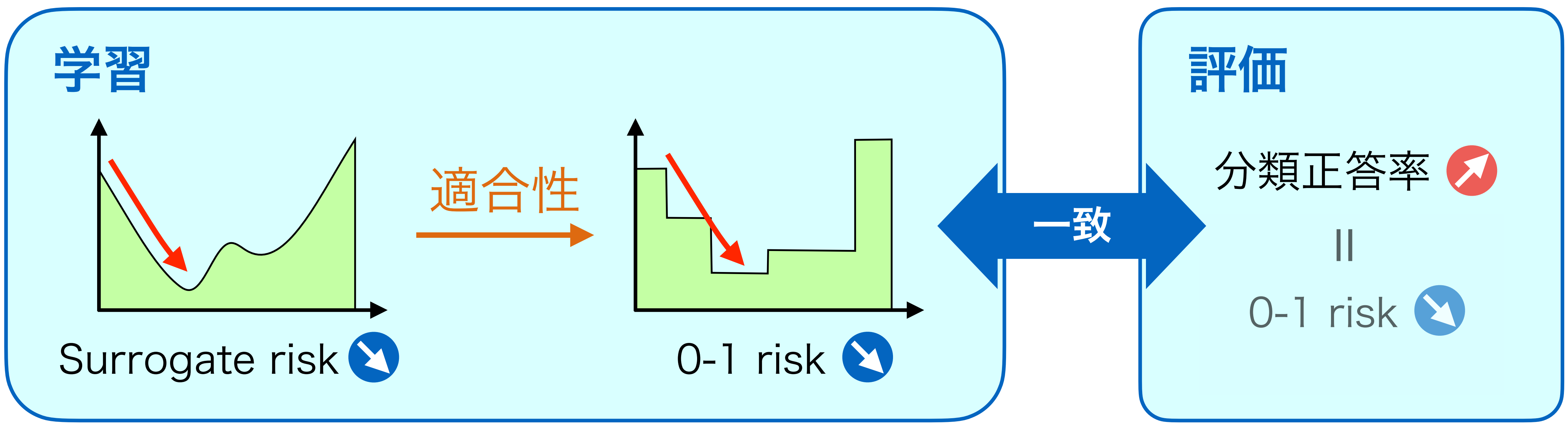
F値: **0**

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

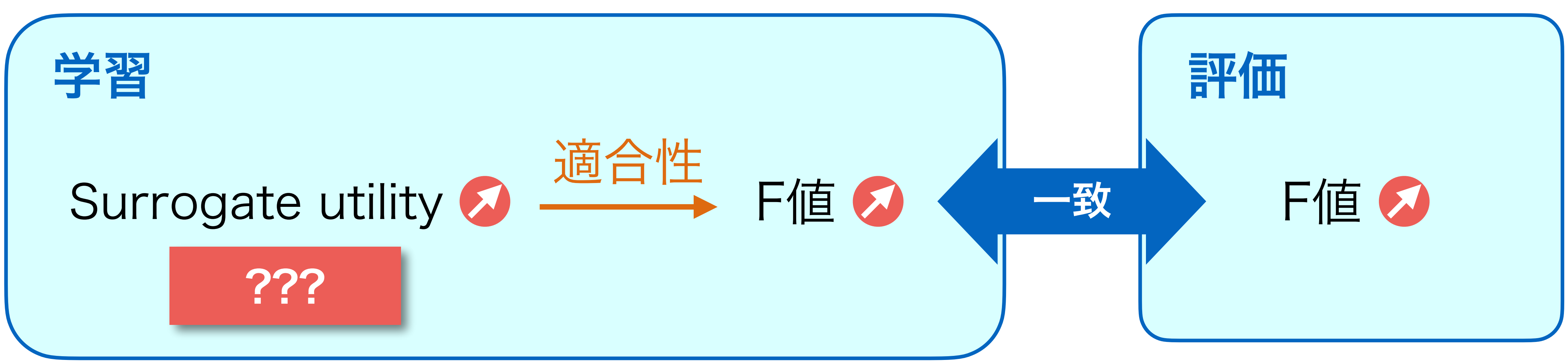
TP: True Positive	TN: True Negative
FP: False Positive	FN: False Negative

学習 vs. 評価

- 通常の二値分類



- F値で評価したい場合



Fowlkes-Mallows index

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

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

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

Ac

統一したい！

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

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

評価指標の統一

Note:

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

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

評価指標の例

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

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

分数線形型

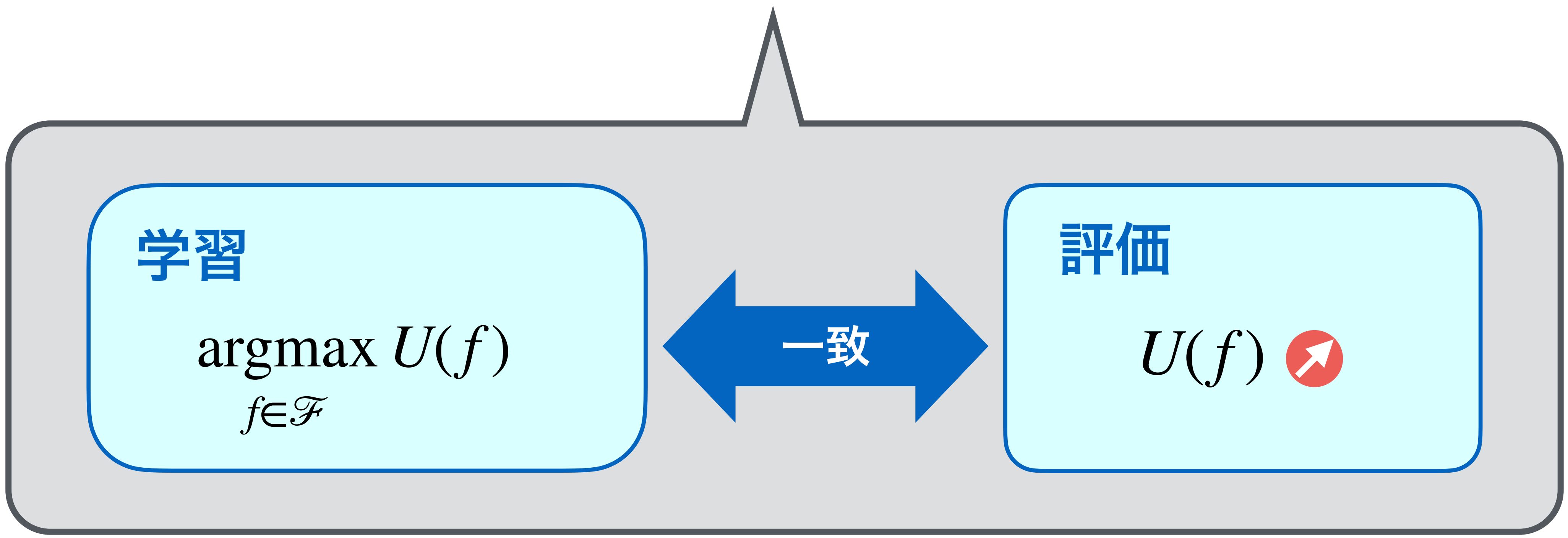
$$U(f) = \frac{a_0\mathrm{TP} + b_0\mathrm{FP} + c_0}{a_1\mathrm{TP} + b_1\mathrm{FP} + c_1}$$

a_k, b_k, c_k : 評価指標に依存する定数

分数線形型の評価指標の下で学習するには？

識別器の評価指標 $U(f) = \dfrac{a_0\mathrm{TP} + b_0\mathrm{FP} + c_0}{a_1\mathrm{TP} + b_1\mathrm{FP} + c_1}$ を1つ決めたとき、

Q. $U(f)$ **の下でどのように性能を <u>直接</u> 最大化する？**

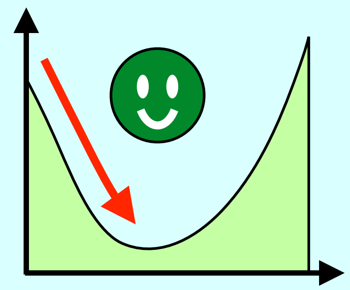


適合性 & 最適化の容易さ

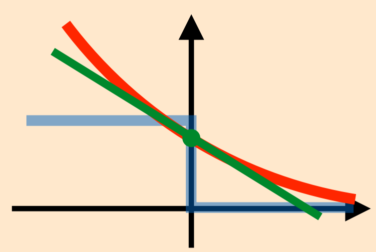
分類正答率の場合

surrogate risk

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



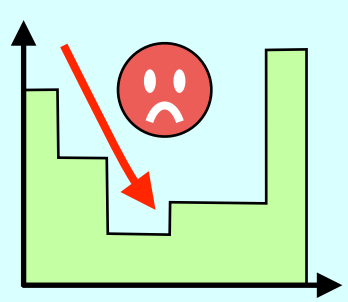
最適化が容易（例: convex）



適合性

target risk

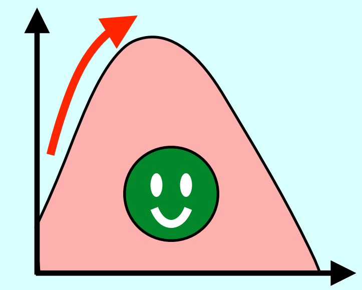
$$R_{01}(f) = \mathbb{E}[\phi_{01}(Yf(X))]$$



分数線形型の場合

surrogate utility

???

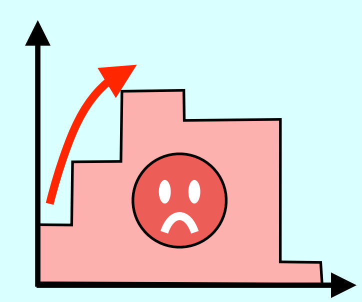


① 最適化が容易（例: concave）

② 適合性

target utility

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

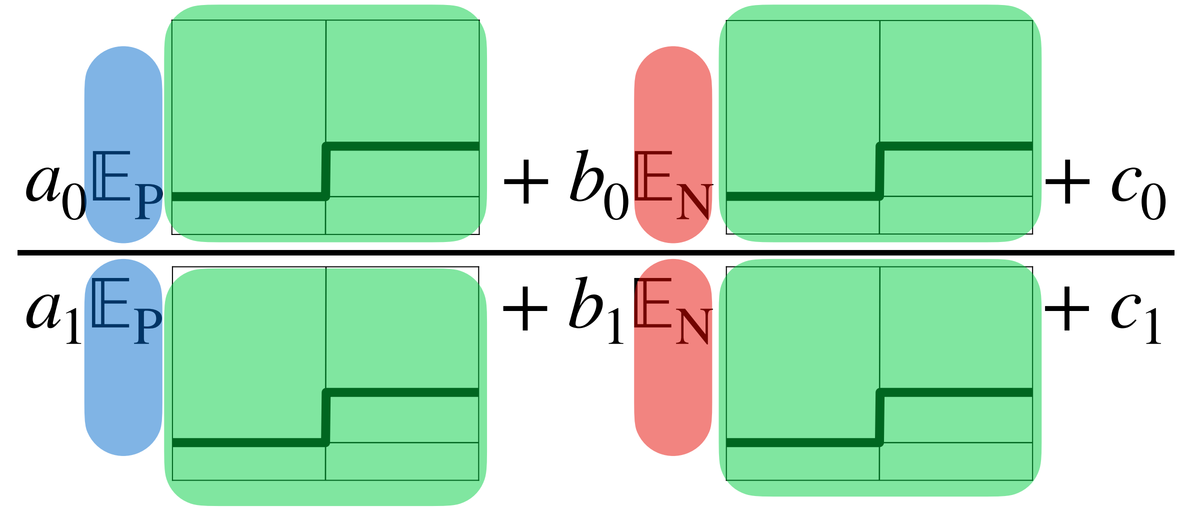


Surrogate Utility

分数線形型

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

=



- TP / FP = 0/1 lossの期待値

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

ラベルが正 && 予測が正

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

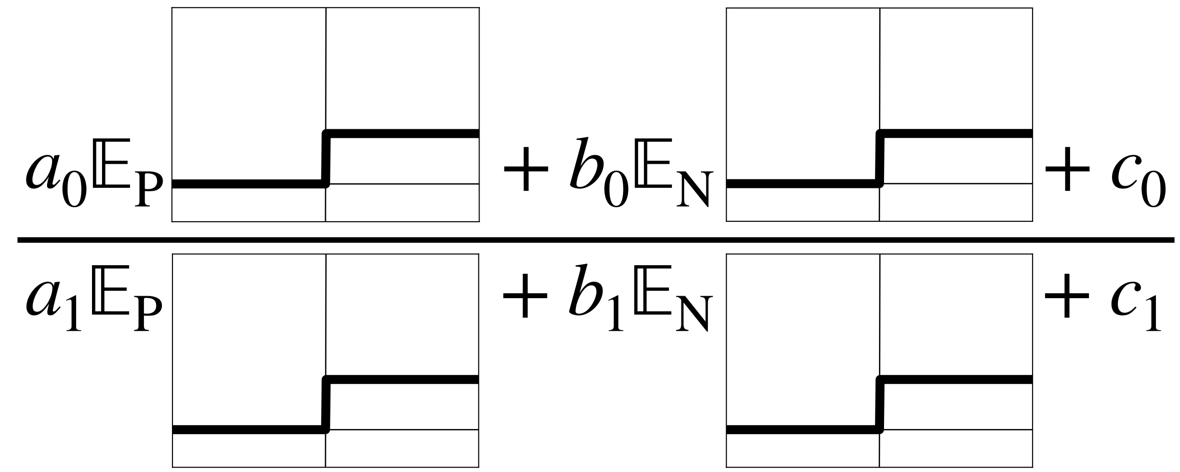
ラベルが負 && 予測が正

Surrogate Utility

分数線形型

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

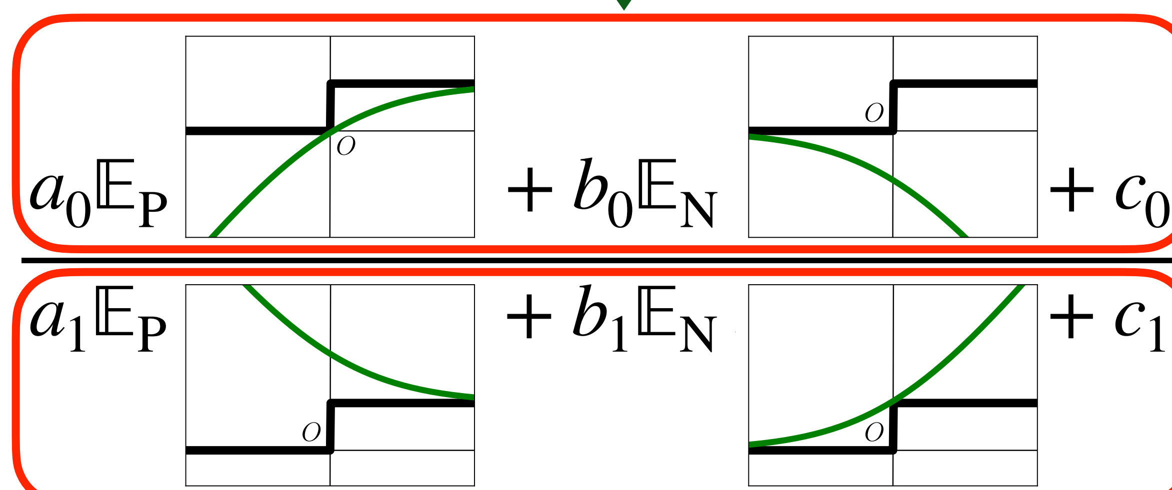
=



分子は下から抑える

concaveの非負和
⇒ concave

≥



convexの非負和
⇒ convex

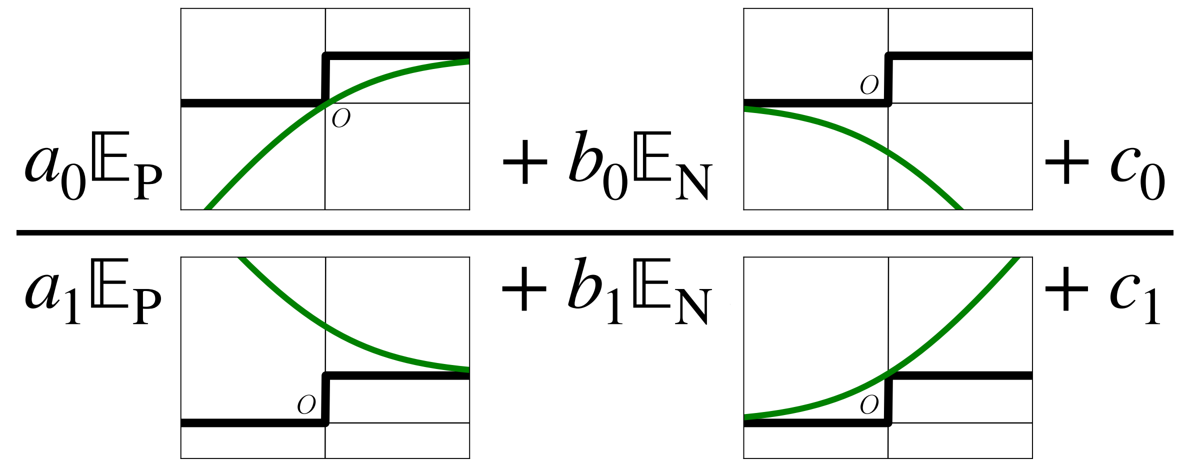
分母は上から抑える

Surrogate Utility

分数線形型

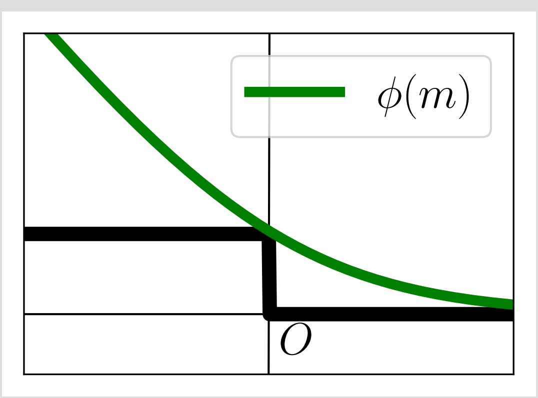
$$U(f) = \frac{a_0 \mathrm{TP} + b_0 \mathrm{FP} + c_0}{a_1 \mathrm{TP} + b_1 \mathrm{FP} + c_1}$$

$\geq$



$\parallel$

surrogate loss

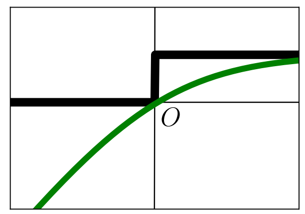


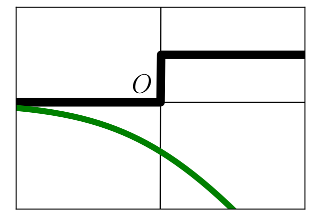
Surrogate Utility

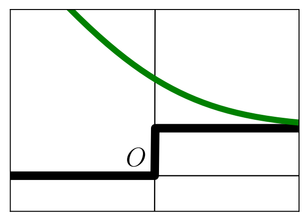
$$U_\phi(f) = \frac{a_0 \mathbb{E}_\mathrm{P}[1 - \phi(f(X))] + b_0 \mathbb{E}_\mathrm{N}[-\phi(-f(X))] + c_0}{a_1 \mathbb{E}_\mathrm{P}[1 + \phi(f(X))] + b_1 \mathbb{E}_\mathrm{N}[\ \phi(-f(X))\ ] + c_1}$$

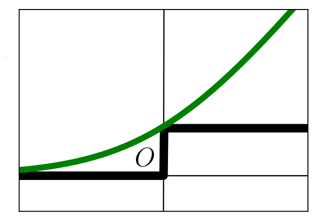
①Surrogate Utilityの最適化

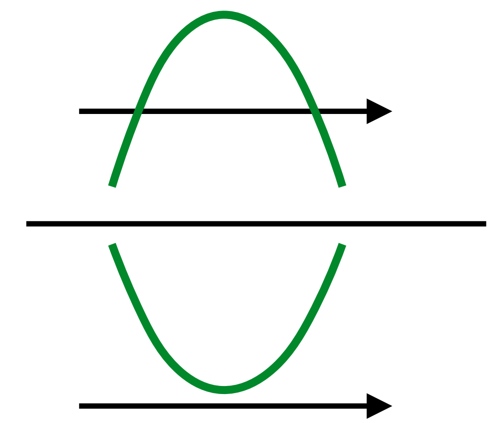
$$U_\phi(f) = \frac{a_0\mathbb{E}_P + b_0\mathbb{E}_N + c_0}{a_1\mathbb{E}_P + b_1\mathbb{E}_N + c_1} =$$











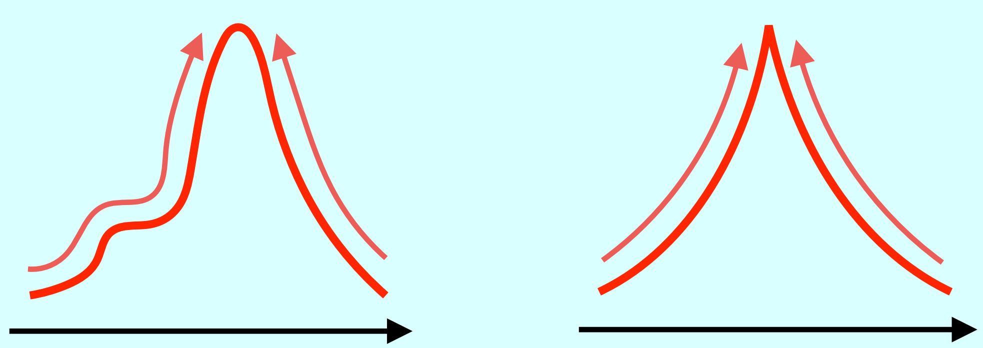
ポイント: concave / convex = quasi-concave

quasi-concave: (直感的には) 山がひとつの関数

⇒ 勾配上昇方向に更新すると値が増加（勾配法）

concaveとは限らない

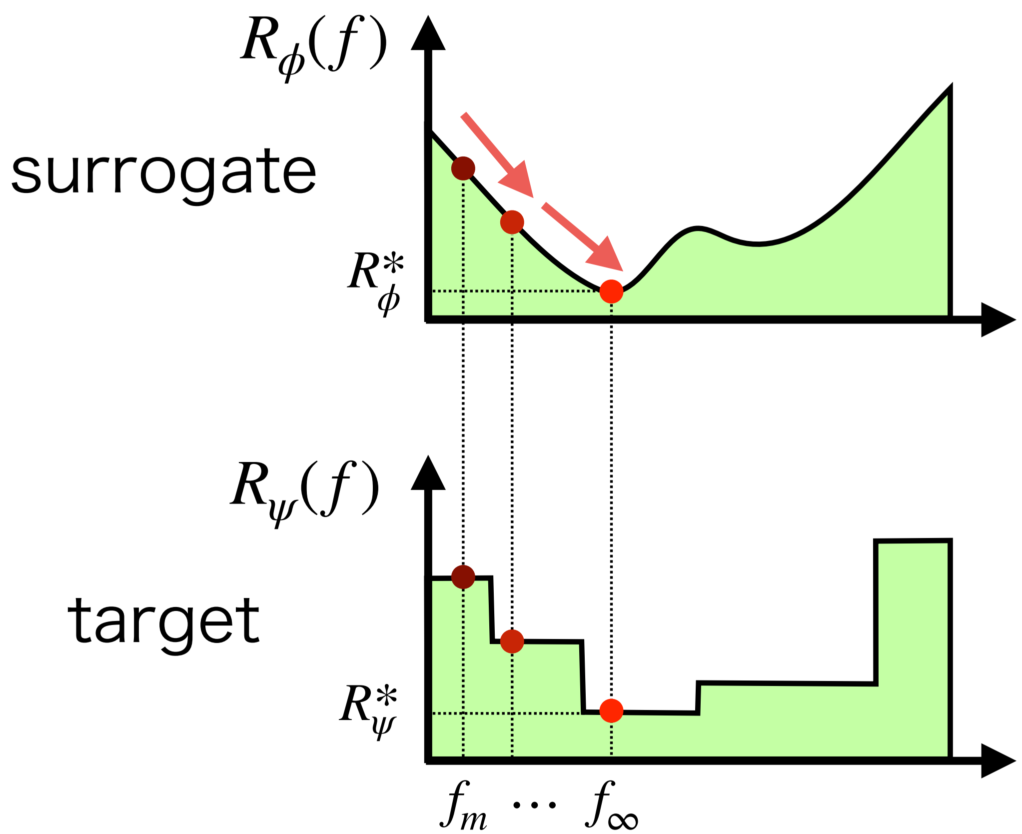
[Hazan+ NeurIPS2015]



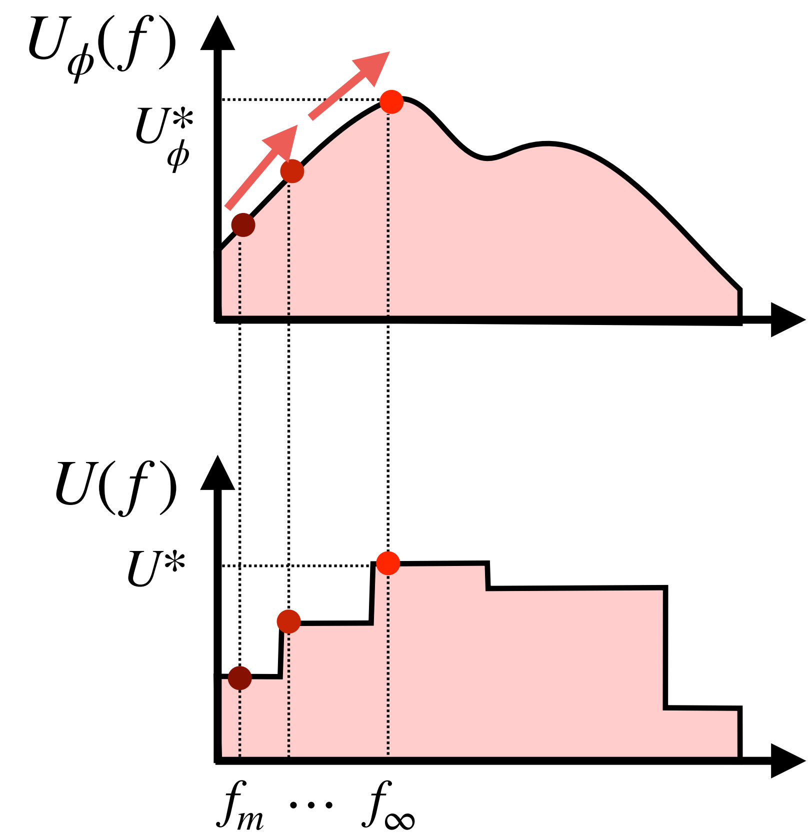
Hazan, E., Levy, K., & Shalev-Shwartz, S. (2015). Beyond convexity: Stochastic quasi-convex optimization. In *Advances in Neural Information Processing Systems* (pp. 1594-1602).

②Surrogate Utilityの適合性

分類正答率の場合



分数線形型の場合



ϕ はどのような性質を満たす必要がある？

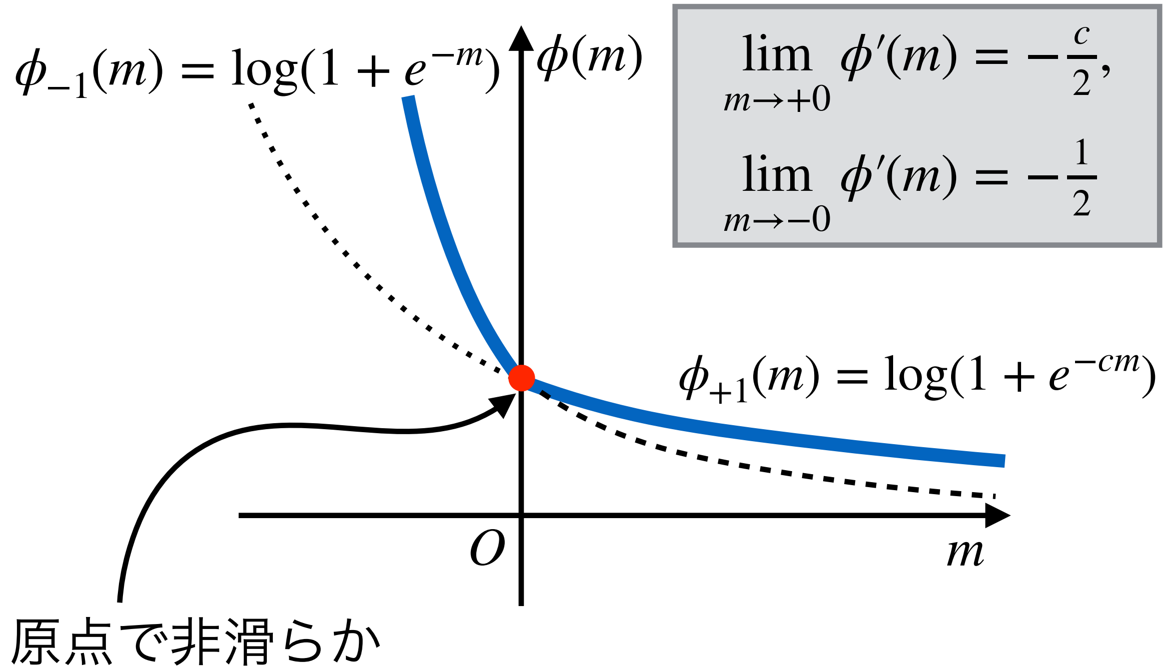
②Surrogate Utilityの適合性

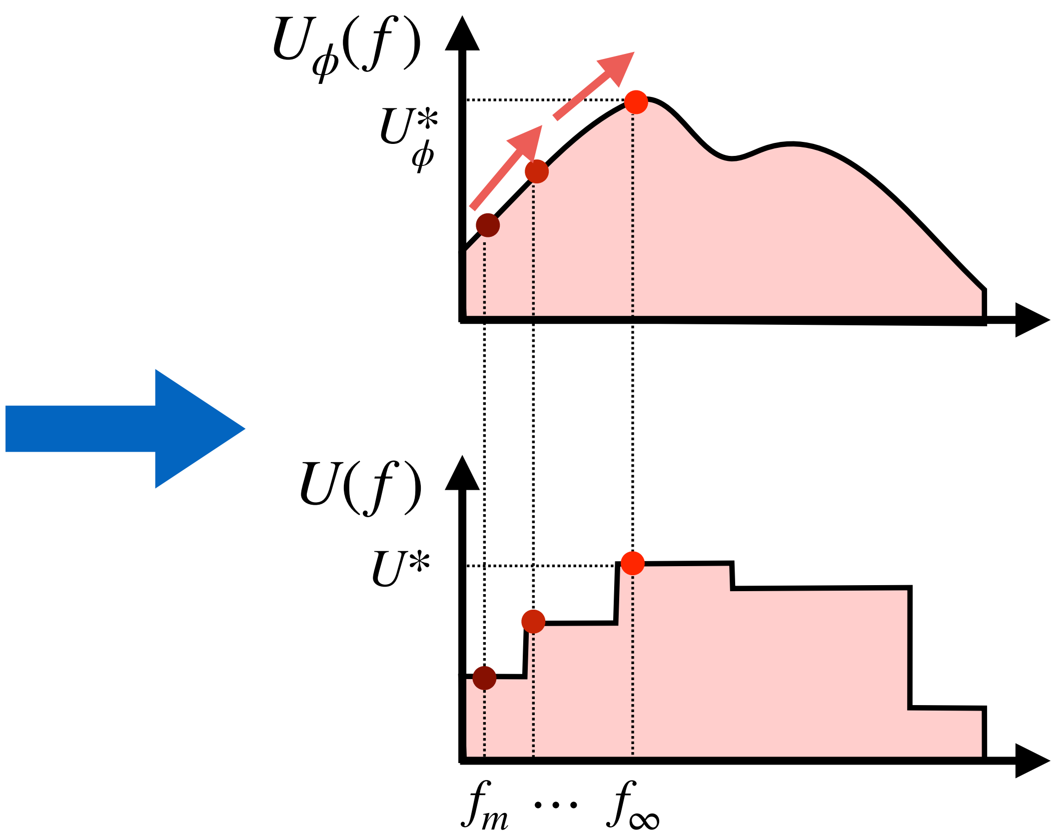
Special Case: F値の場合

定理. ϕ が以下を満たすとき U_ϕ は適合的

- ϕ: 単調非増加
- ϕ: convex
- $\exists c \in (0,1)$ s.t. $\sup_f U_\phi(f) \geq \frac{2c}{1-c}$, $\lim_{m \to +0} \phi'(m) \geq c \lim_{m \to -0} \phi'(m)$

適合的な ϕ の例





数値実験: F値

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

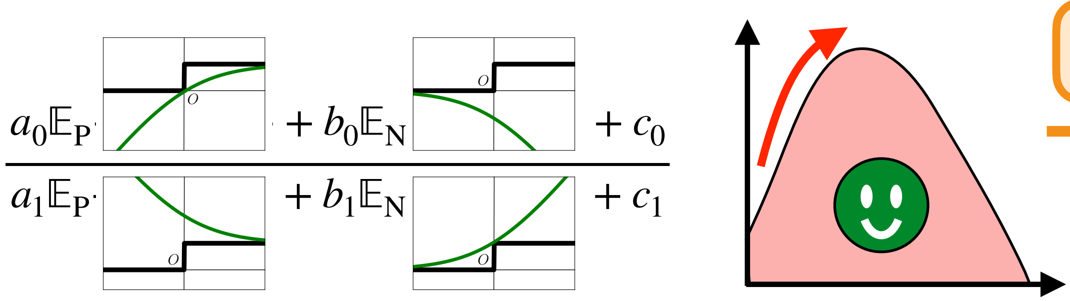
より複雑な評価指標と損失関数

分数線形型の評価指標

F値、Jaccard指標などを包摂
不均衡データを扱うときによく利用

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

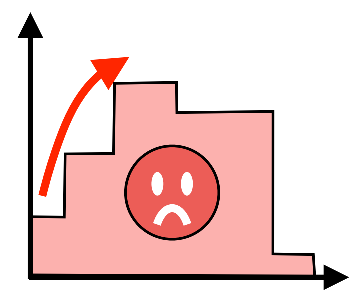
surrogate utility



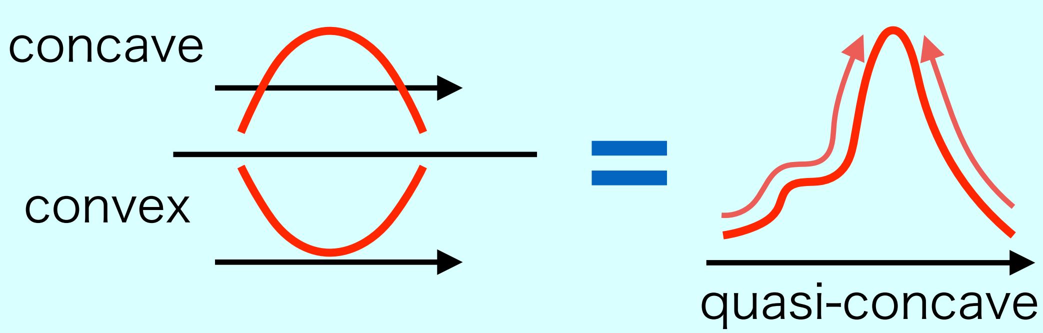
適合性

target utility

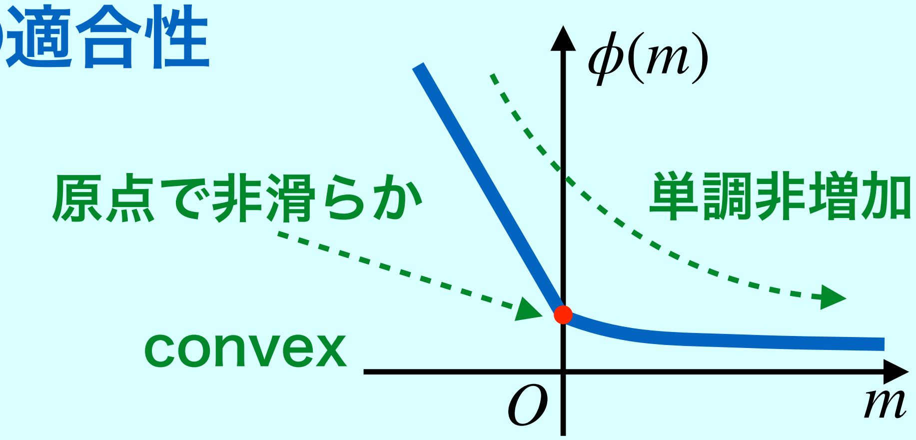
$$U(f) = \frac{a_0\mathrm{TP} + b_0\mathrm{FP} + c_0}{a_1\mathrm{TP} + b_1\mathrm{FP} + c_1}$$



①最適化が容易: quasi-concave



②適合性



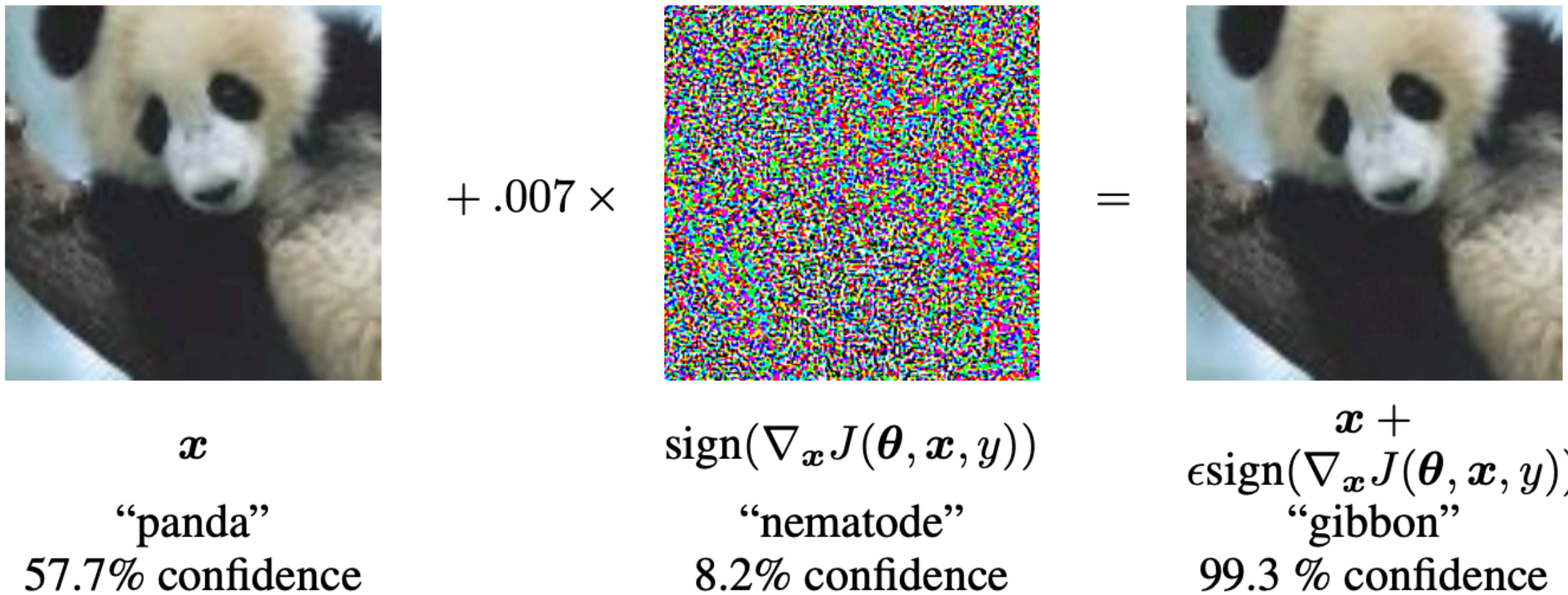
複雑な評価指標に対する代理損失の設計指針に！

ロバストな学習と損失関数

H. Bao, C. Scott, and M. Sugiyama.
Calibrated Surrogate Losses for Adversarially Robust Classification.
In *COLT*, 2020.

敵対者による分類器への攻撃

[Goodfellow+ 2015; Eykholt+ 2018]

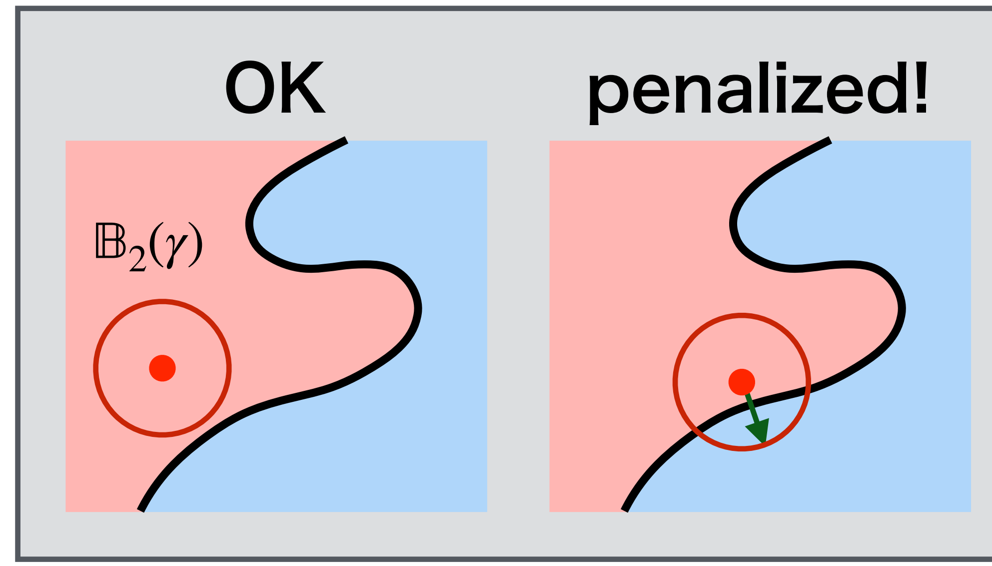


Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *ICLR*, 2015.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *CVPR*, 2018.

攻撃者の定式化

- 攻撃者:ℓ_2-ノルムが $\gamma \in (0,1)$ 以下で分類器の予測を変えるノイズ
  - 損失関数として定式化



通常の 0-1 loss　　予測が間違っていたら

$$\ell_{01}(x, y, f) = \begin{cases} 1 \text{ if } \mathrm{sign}(f(x)) \neq \mathrm{sign}(y) \\ 0 \text{ otherwise} \end{cases}$$

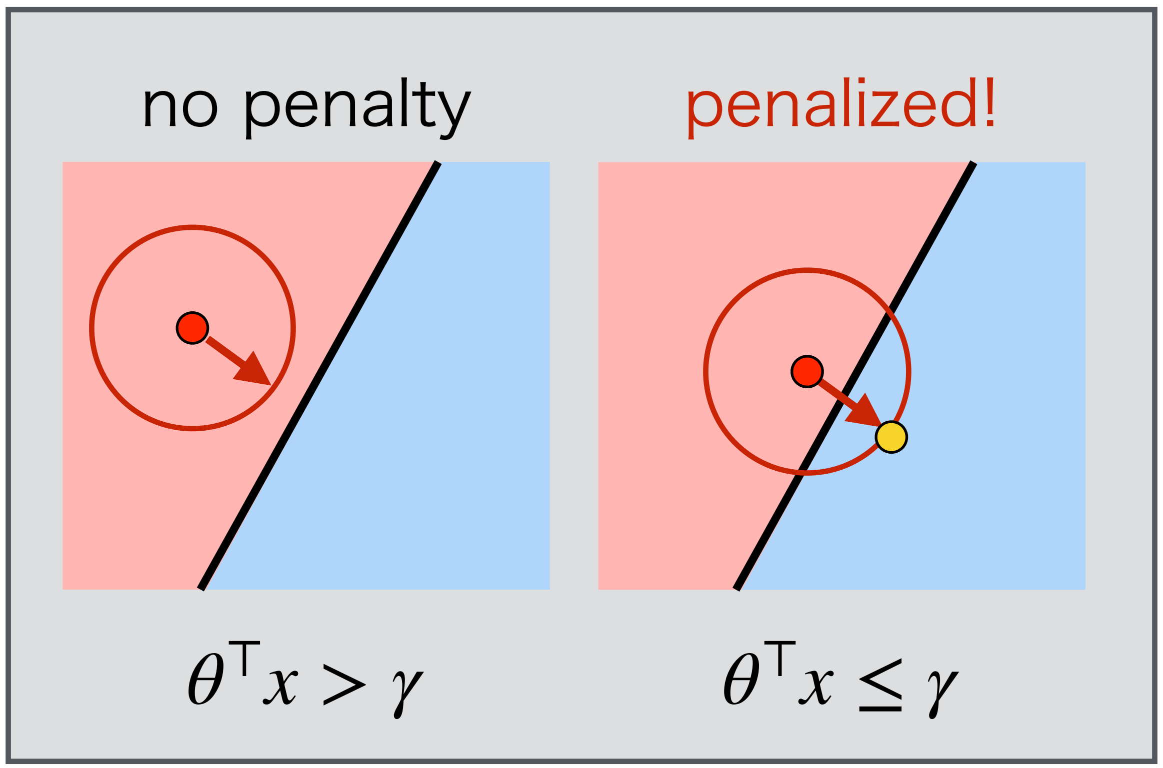
ロバストな 0-1 loss　　予測を間違うノイズが存在するなら

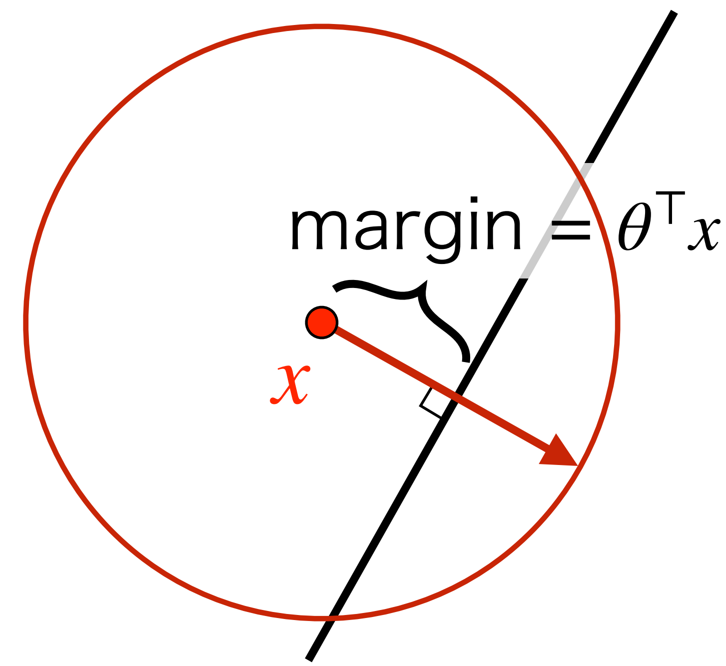
$$\ell_\gamma(x, y, f) = \begin{cases} 1 & \text{if } \exists \Delta \in \mathbb{B}_2(\gamma) \text{ s.t. } \mathrm{sign}(f(x+\Delta)) \neq \mathrm{sign}(y) \\ 0 & \text{otherwise} \end{cases}$$

$\mathbb{B}_2(\gamma) = \{x \in \mathbb{R}^d \mid \|x\|_2 \leq \gamma\}$: γ-ball

線形分類境界の場合

線形分類器 $\mathscr{F}_{\text{lin}} = \{x \mapsto \theta^\top x \mid \|\theta\|_2 = 1\}$





ロバストな 0-1 loss

$$\ell_\gamma(x, y, f) = \begin{cases} 1 & \text{if } \exists \Delta \in \mathbb{B}_2(\gamma)\,.\, yf(x+\Delta) \leq 0 \\ 0 & \text{otherwise} \end{cases} \quad = \quad \mathbf{1}\{yf(x) \leq \gamma\} := \phi_\gamma(yf(x))$$

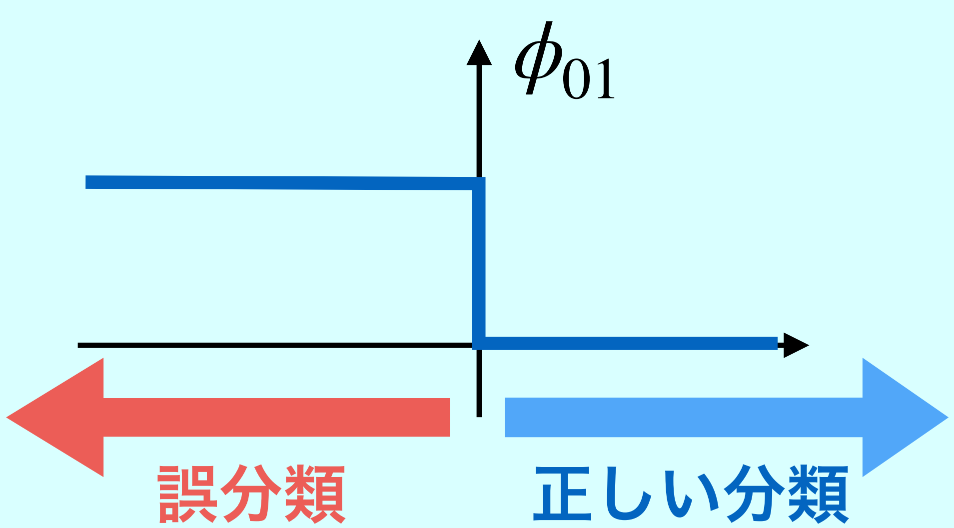
分類問題の定式化

通常の二値分類

0-1 riskの最小化

$$R_{\phi_{01}}(f) = \mathbb{E}\left[\phi_{01}(Yf(X))\right]$$

0-1 loss $\phi_{01}(\alpha) = \mathbf{1}\{\alpha \le 0\}$



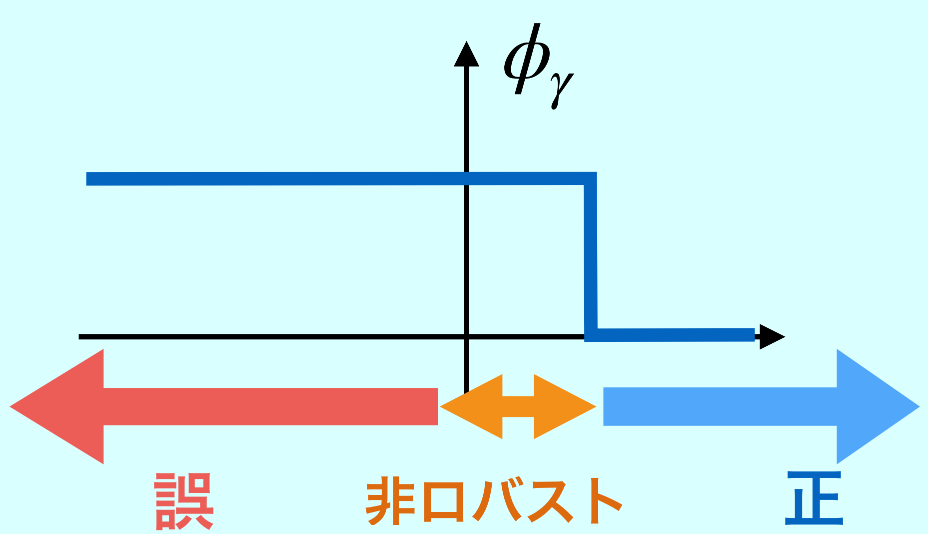
ロバストな二値分類

γ-robust 0-1 riskの最小化

$$R_{\phi_\gamma}(f) = \mathbb{E}\left[\phi_\gamma(Yf(X))\right]$$

(※: 線形分類境界の場合)

robust 0-1 loss $\phi_\gamma(\alpha) = \mathbf{1}\{\alpha \le \gamma\}$

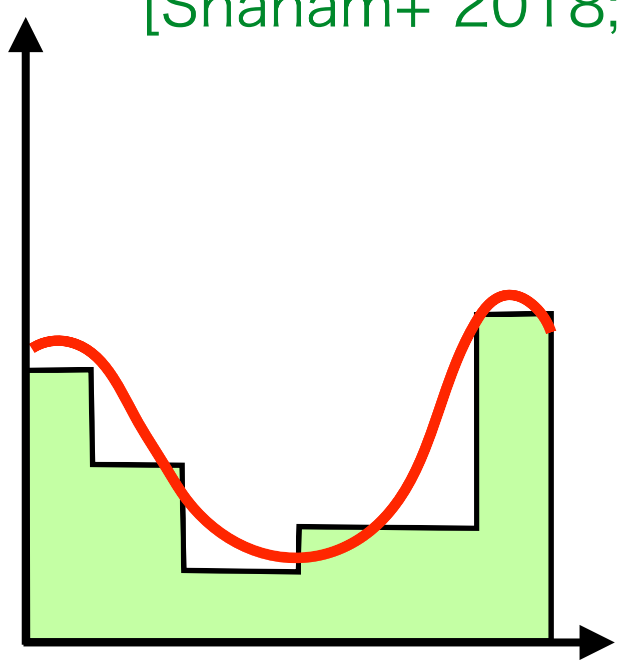


既存のロバストな学習方法

Robust risk $R_{\phi_\gamma}(w) = \mathbb{E}[\phi_\gamma(Y(w^\top X))]$ の直接最小化は困難

Taylor近似

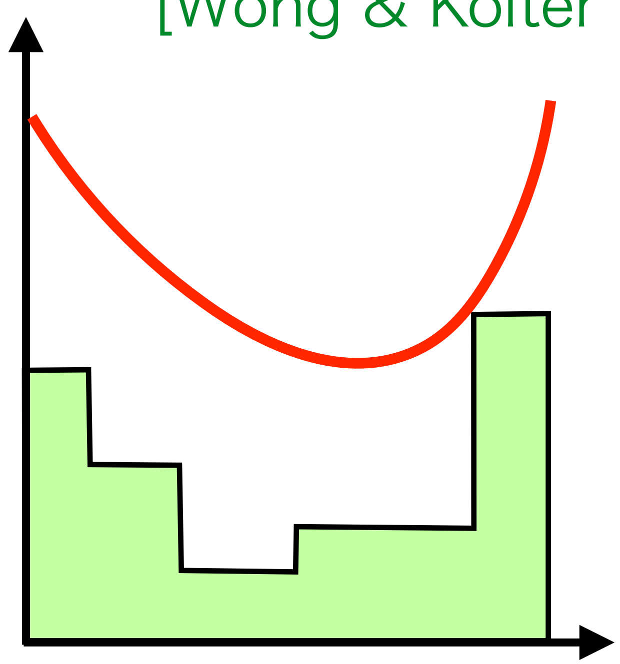
[Shaham+ 2018; etc.]



目的関数の近似の最小化が
正しい解を導くとは限らない

上界の最小化

[Wong & Kolter 2018; etc.]



上界の最小化の正しい解への
収束性は示されていない

Shaham, U., Yamada, Y., & Negahban, S. (2018).
Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 195-204.

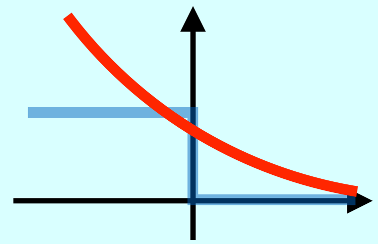
Wong, E., & Kolter, Z. (2018,). Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope.
In *International Conference on Machine Learning* (pp. 5286-5295).

適合性 & 最適化の容易さ

通常の二値分類の場合

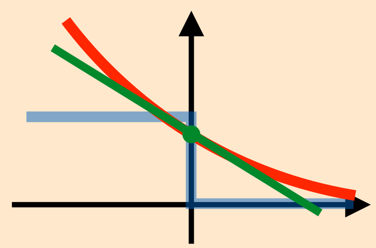
surrogate risk

$R_\phi(w) = \mathbb{E}[\phi(Y(w^\top X))]$



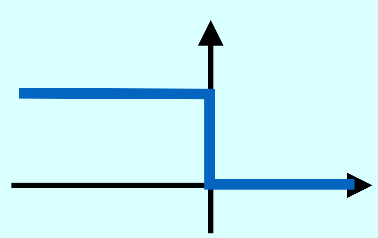
最適化が容易（例: convex）

適合性



0/1 risk

$R_{01}(w) = \mathbb{E}[\phi_{01}(Y(w^\top X))]$



ロバストな二値分類の場合

surrogate risk

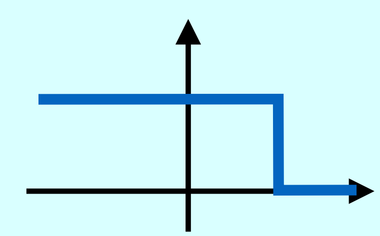
???

① **最適化が容易**（例: convex）

② **適合性**

robust 0/1 risk

$R_{\phi_\gamma}(w) = \mathbb{E}[\phi_\gamma(Y(w^\top X))]$



意外とシンプル？

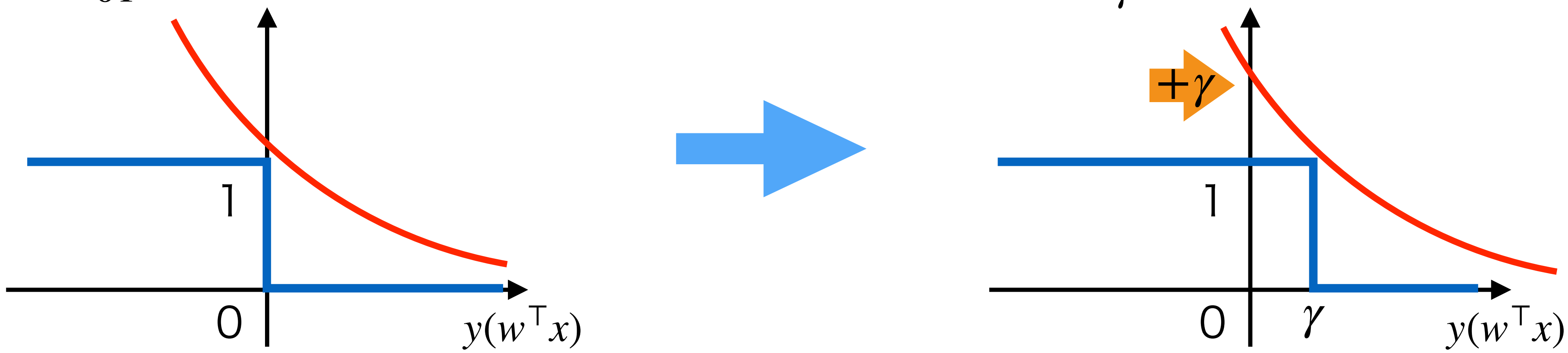
定理. Surrogate ϕ が凸関数なら以下の場合に限り0-1 lossに対して適合的
- 原点で微分可能
- $\phi'(0) < 0$

通常の 0-1 loss

$$\phi_{01}(\alpha) = \mathbf{1}\{\alpha \leq 0\}$$

ロバストな 0-1 loss

$$\phi_\gamma(\alpha) = \mathbf{1}\{\alpha \leq \gamma\}$$



$\phi'(\gamma) < 0$ であればロバストな 0-1 lossに対して適合的？

凸 & 適合的なSurrogateは存在しない！

定理. 任意のconvex surrogateは（線形分類器の中では）robust lossに対して適合的でない

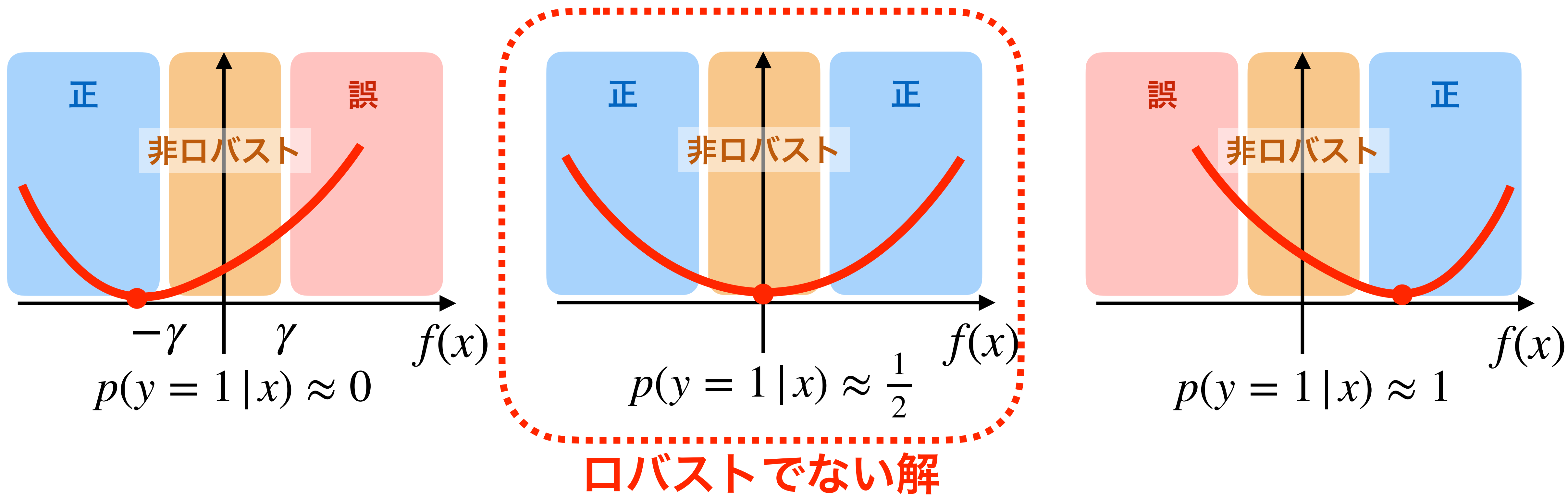
証明の概略: calibration function $\delta(\varepsilon) = 0$ となる分布の存在を示す

calibration function

f に関して凸

意味: 分類器 f がロバストでない

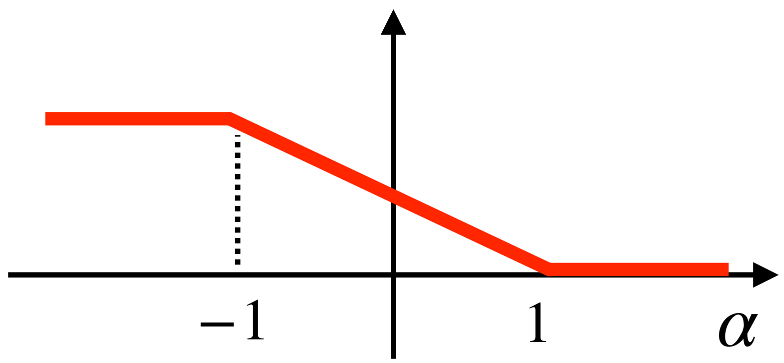
$$\delta(\varepsilon) = \inf_f R_\phi(f) - R_\phi^* \quad \text{s.t.} \quad R_{\phi_\gamma}(f) - R_{\phi_\gamma}^* \geq \varepsilon$$



適合的な代理損失の例: ramp loss

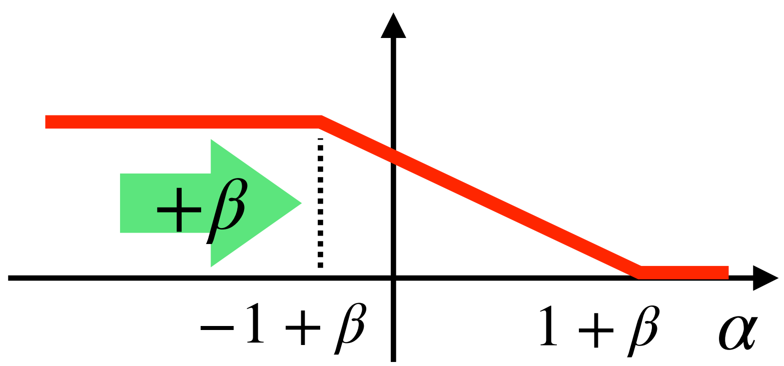
Ramp loss

$$\phi(\alpha) = \mathrm{clip}_{[0,1]}\left(\frac{1-\alpha}{2}\right)$$

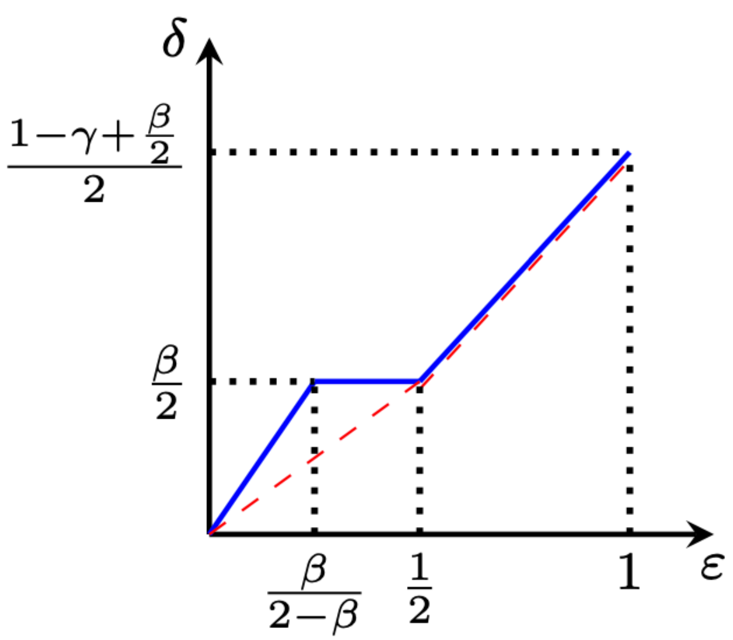


Shifted ramp loss

$$\phi_\beta(\alpha) = \mathrm{clip}_{[0,1]}\left(\frac{1-\alpha+\beta}{2}\right)$$



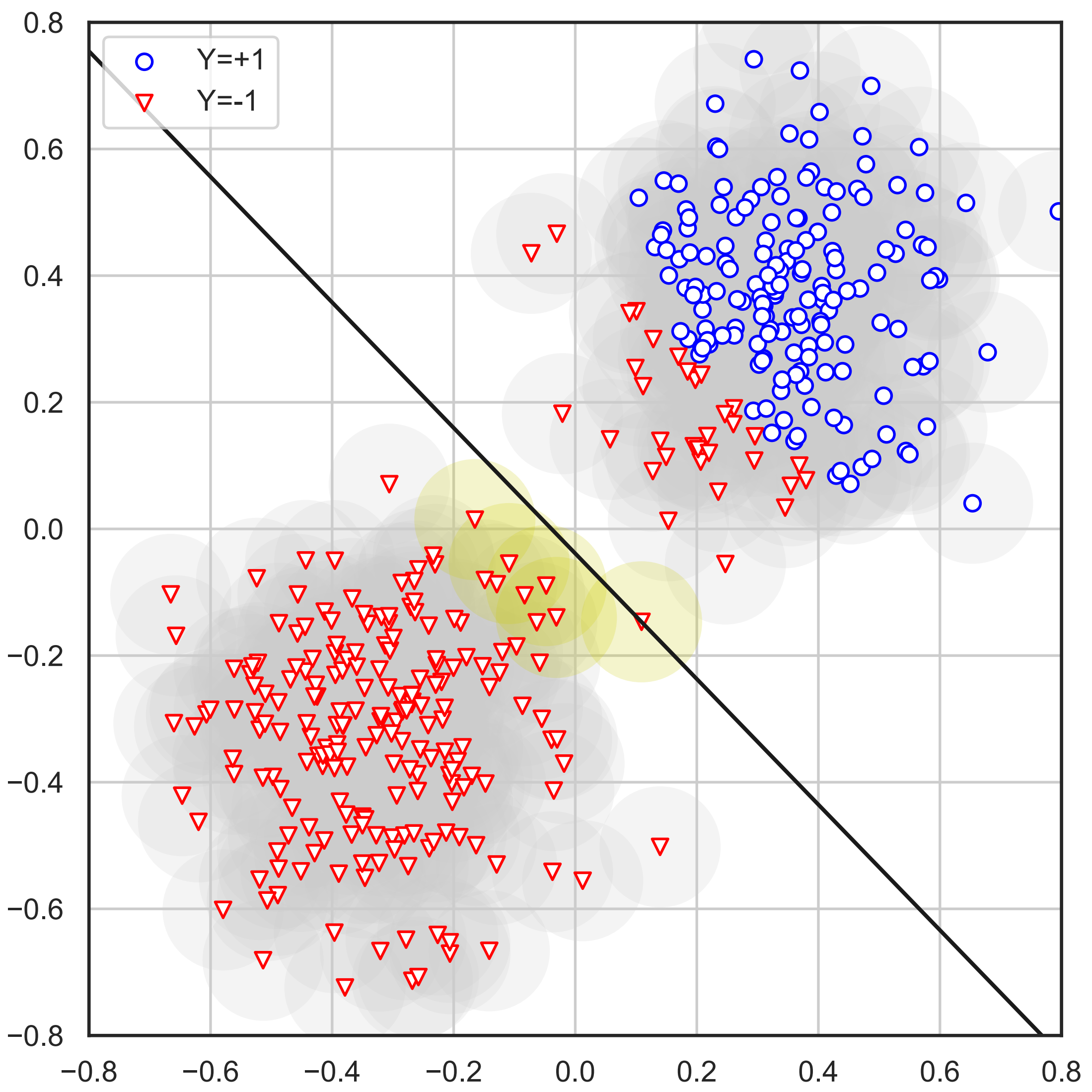
calibration function



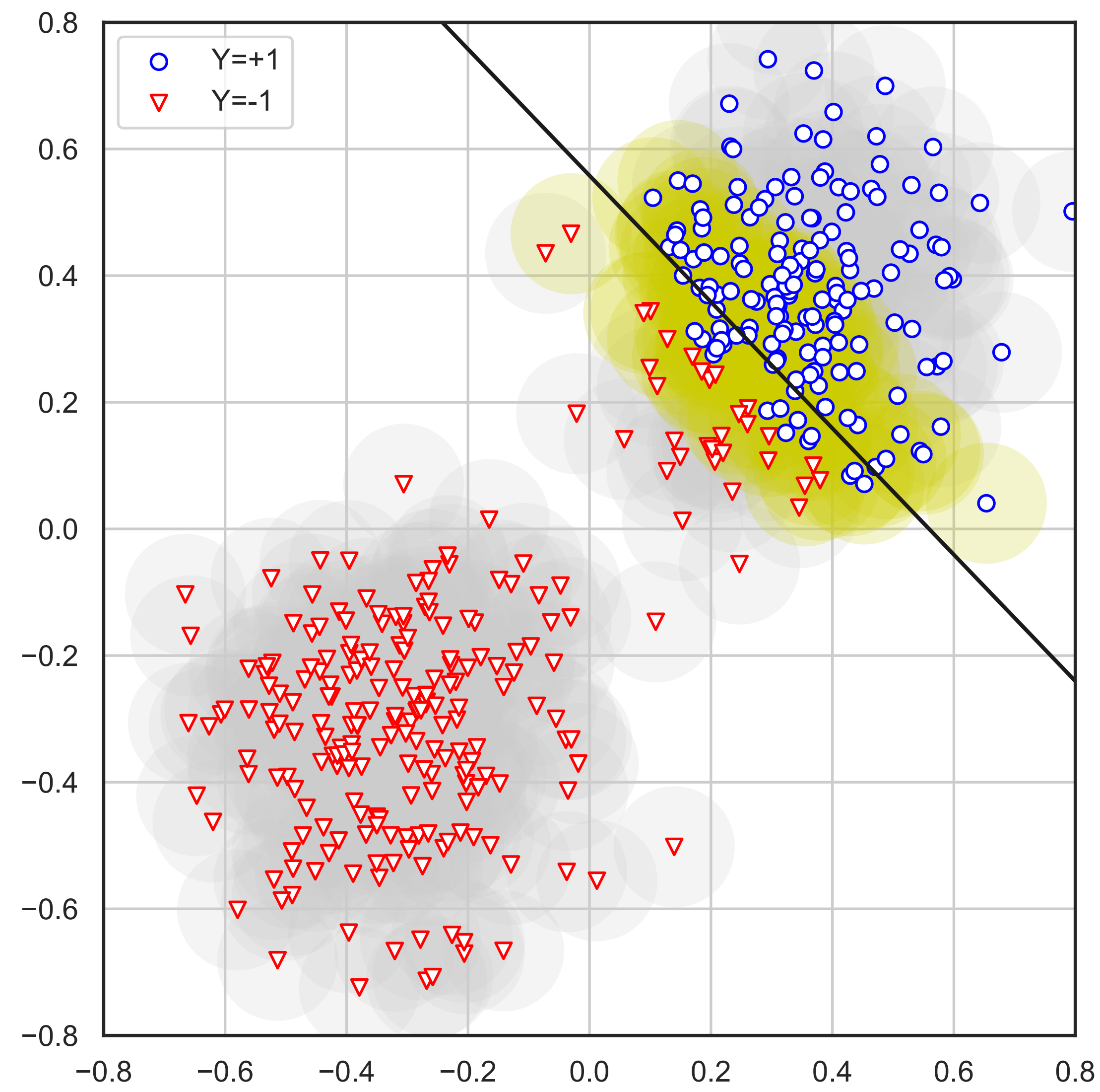
assume $0 < \beta < 1-\gamma$

数値実験

Ramp loss



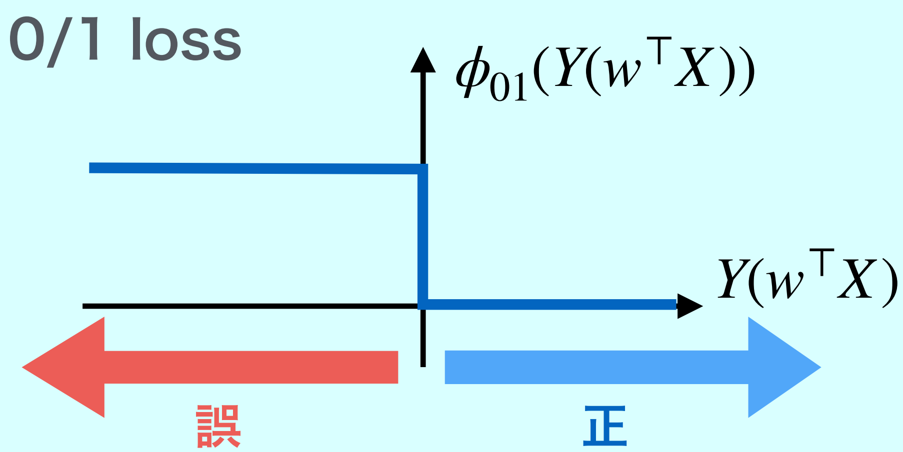
Hinge loss

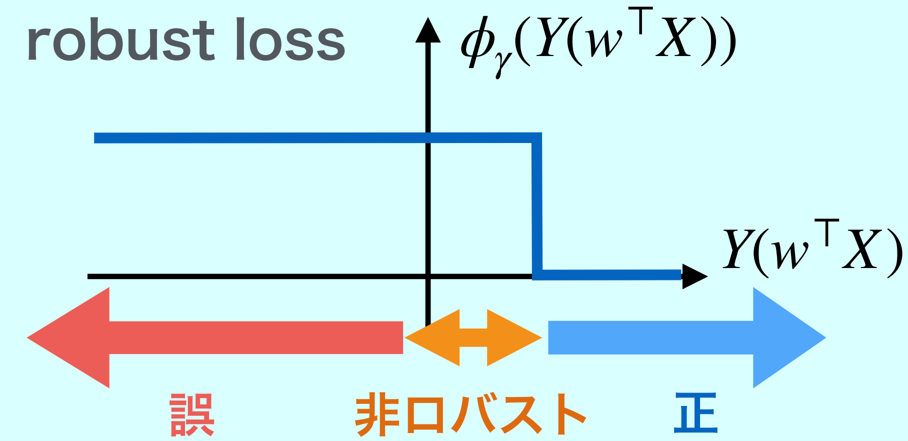


各点に付随する球は γ-ball / 黄色の球は決定境界に触れている (=非ロバストな) 点

ロバストな学習と損失関数

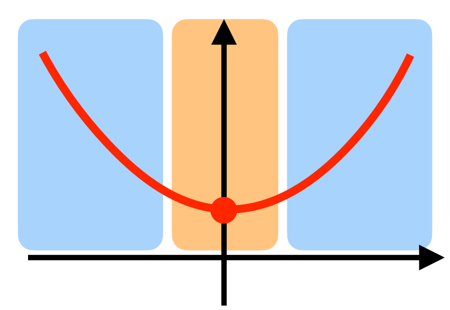
損失関数にロバスト性を「埋め込む」



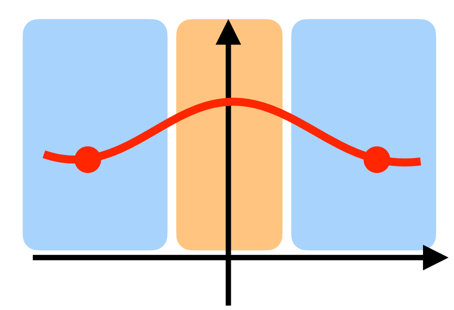


損失関数は予測の正誤だけでなく、予測のロバスト性を埋め込むことも可能

凸な代理損失では
ロバスト性が得られない



凸関数は非ロバストな
領域に解を出力



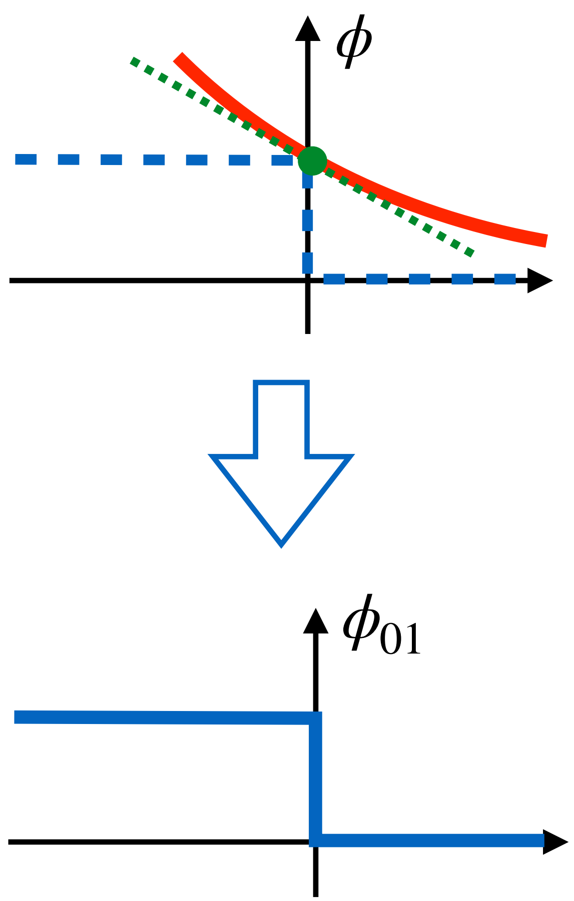
ロバストな目的関数

適合性理論は分類器の性質を調べるのにも役立つ！

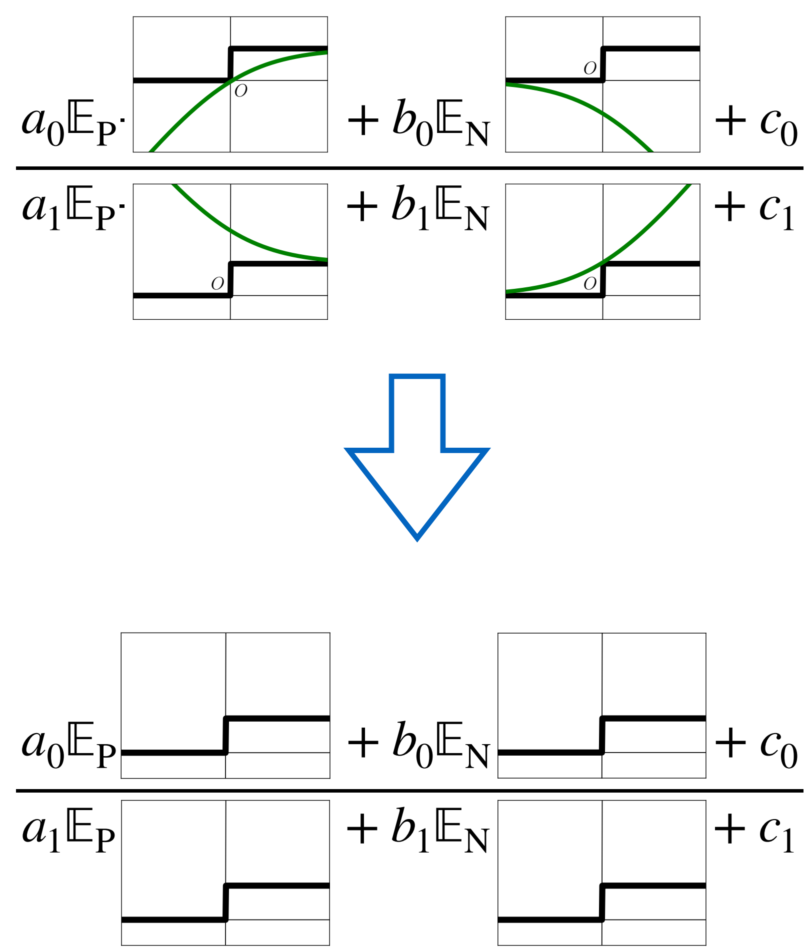
まとめ

まとめ

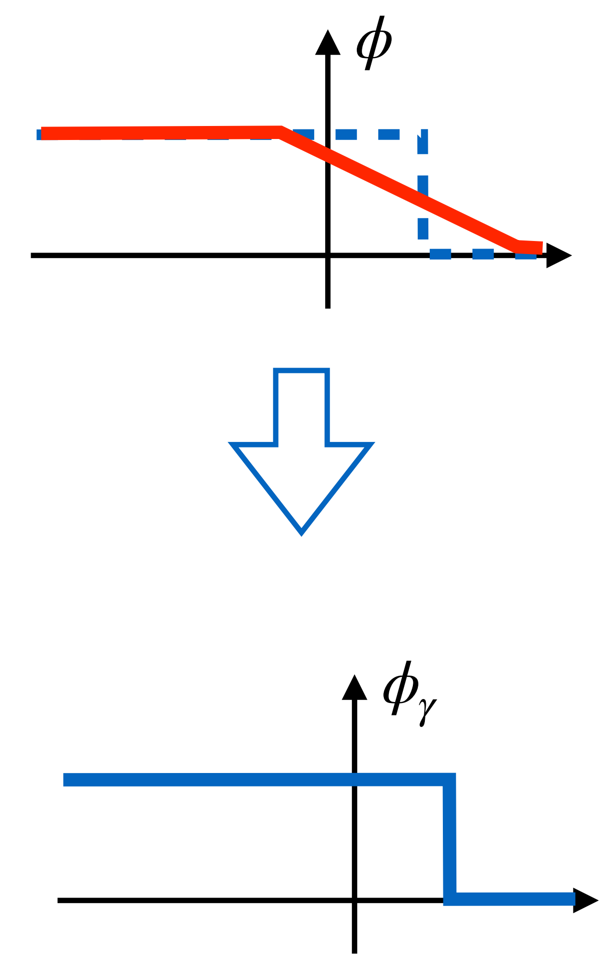
二値分類



不均衡データ



敵対的攻撃



- 損失関数の適合性解析の紹介
- ロバスト性の適用を行った最新の研究