Calibrated Surrogate Losses for Adversarially Robust Classification



Han Bao[1,2] Clayton Scott[3] Masashi Sugiyama[2,1]

1 The University of Tokyo
2 RIKEN AIP
3 University of Michigan

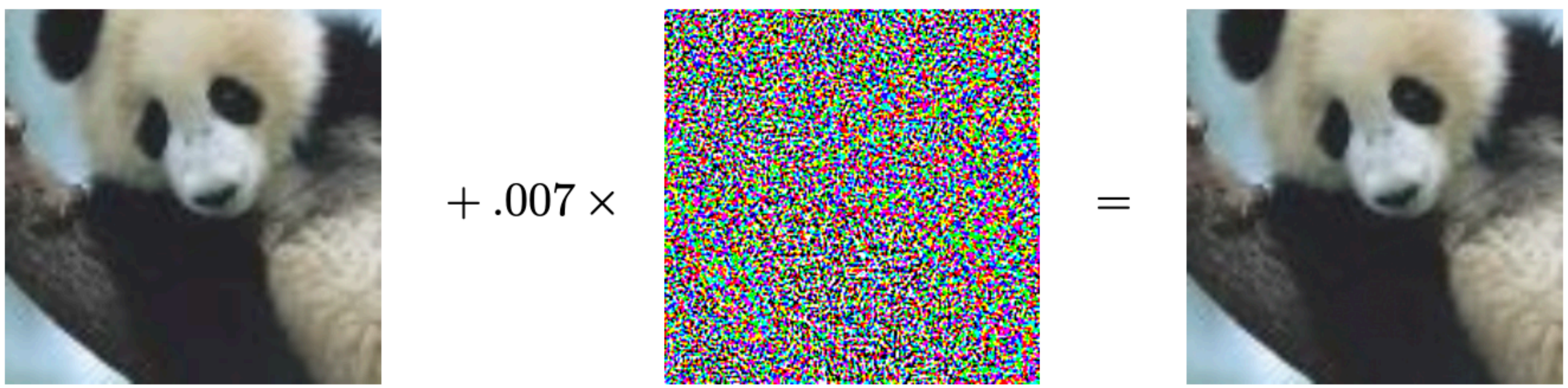
Jul. 9th - 12th @ COLT 2020

Adversarial Attacks

[Goodfellow+ 2015]

Adding inperceptible small noise can fool classifiers!

original data | perturbed data



$\boldsymbol{x}$ — "panda" 57.7% confidence

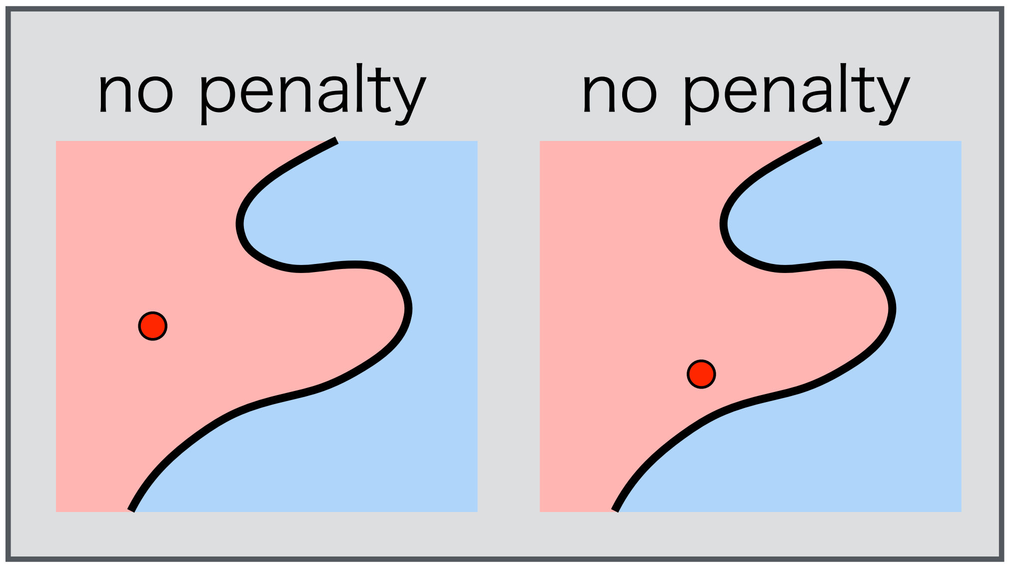
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$ — "nematode" 8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$ — "gibbon" 99.3 % confidence

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

Penalize Vulnerable Prediction

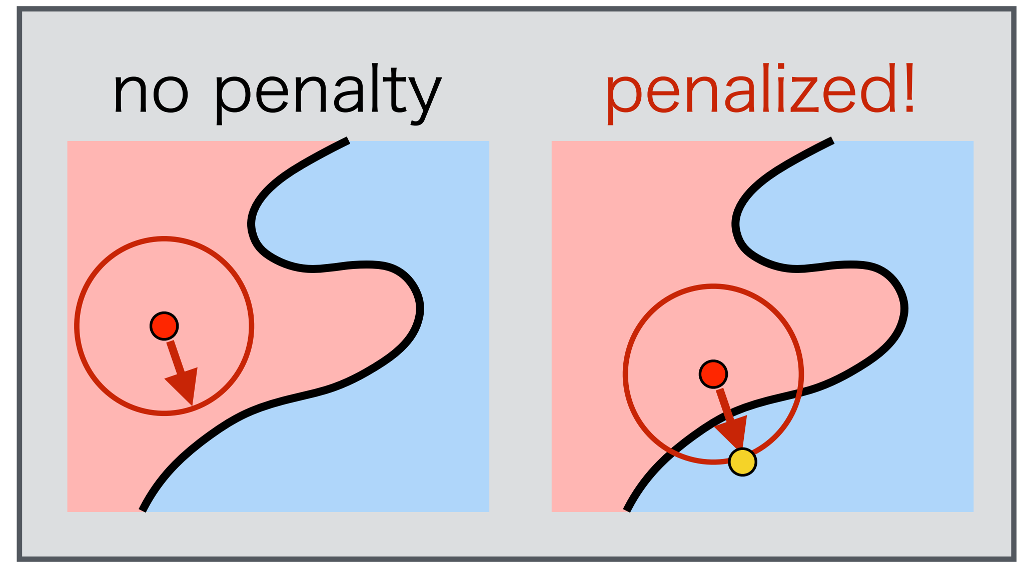
Usual Classification



usual 0-1 loss

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

Robust Classification



robust 0-1 loss

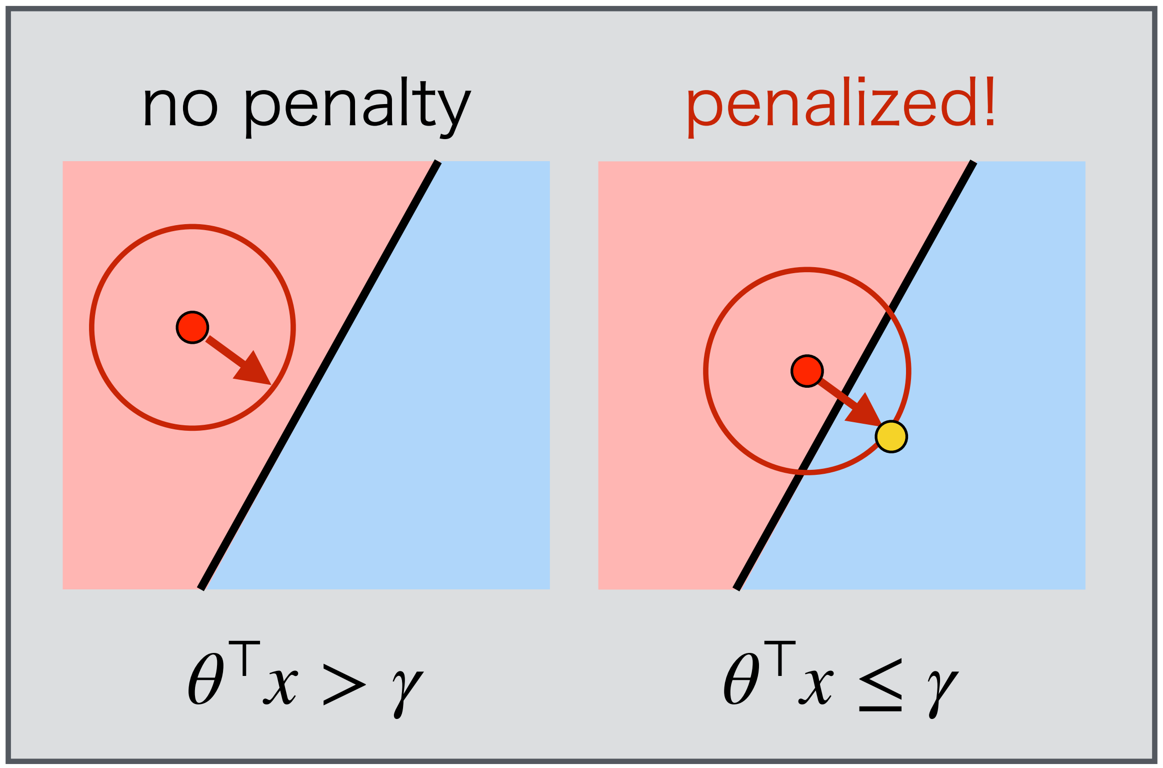
$$\ell_{\gamma}(x, y, f) = \begin{cases} 1 & \text{if } \exists \Delta \in \mathbb{B}_2(\gamma) \,.\, yf(x + \Delta) \leq 0 \\ 0 & \text{otherwise} \end{cases}$$

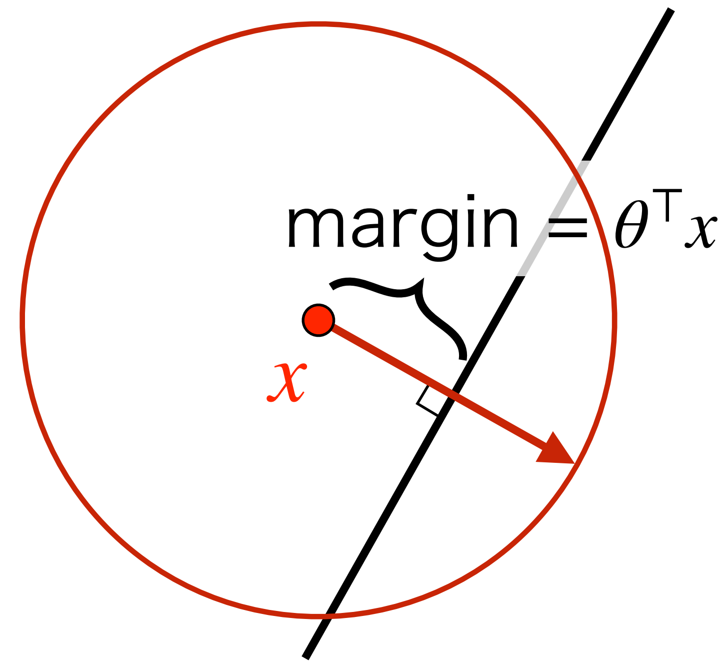
prediction too close to boundary should be penalized

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

In Case of Linear Predictors

linear predictors $\mathscr{F}_{\text{lin}} = \{x \mapsto \theta^\top x \mid \|\theta\|_2 = 1\}$





robust 0-1 loss

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

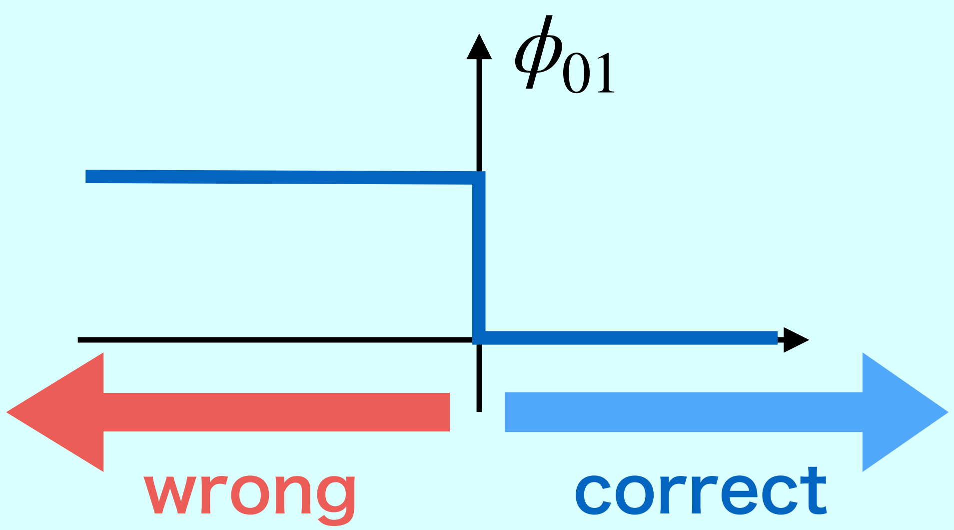
Formulation of Classification

Usual Classification

minimize 0-1 risk

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

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



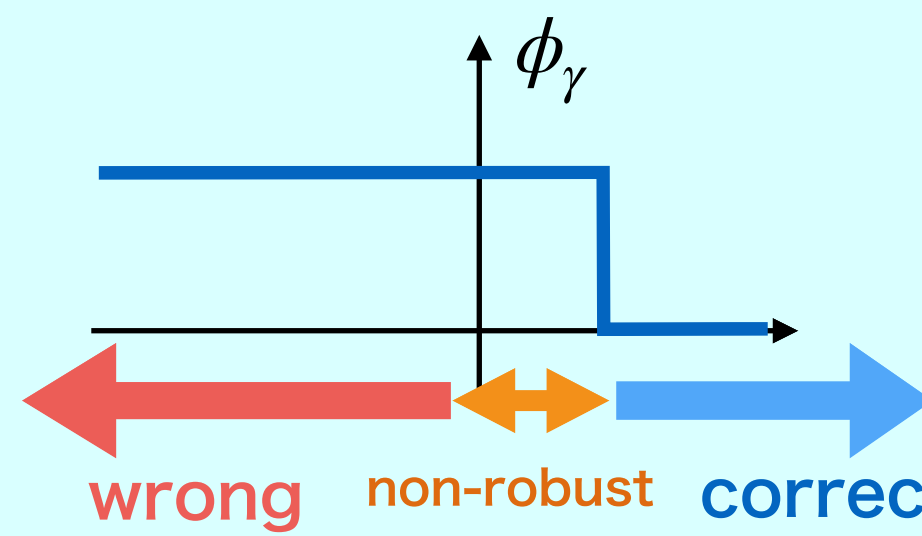
Robust Classification

minimize γ-robust 0-1 risk

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

(restricted to linear predictors)

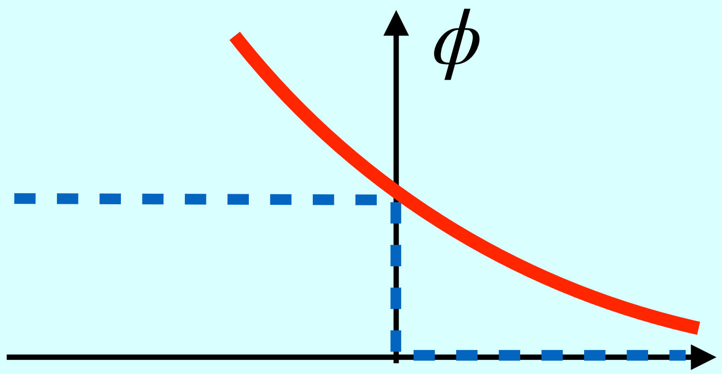
robust 0-1 loss $\phi_\gamma(\alpha) = \mathbf{1}\{\alpha \leq \gamma\}$



☹ ϕ_{01} & ϕ_γ are not easy to optimize!

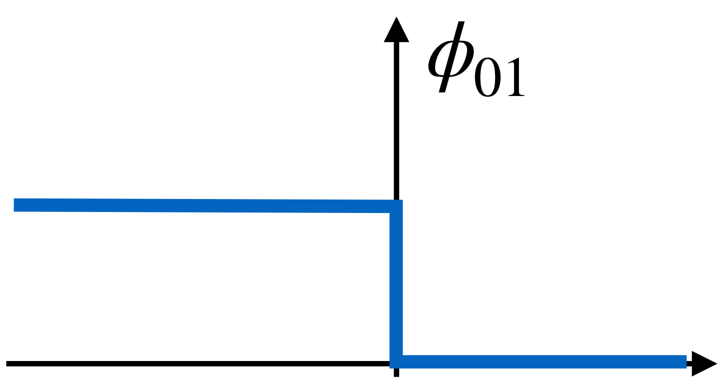
What surrogate is desirable?

Surrogate loss



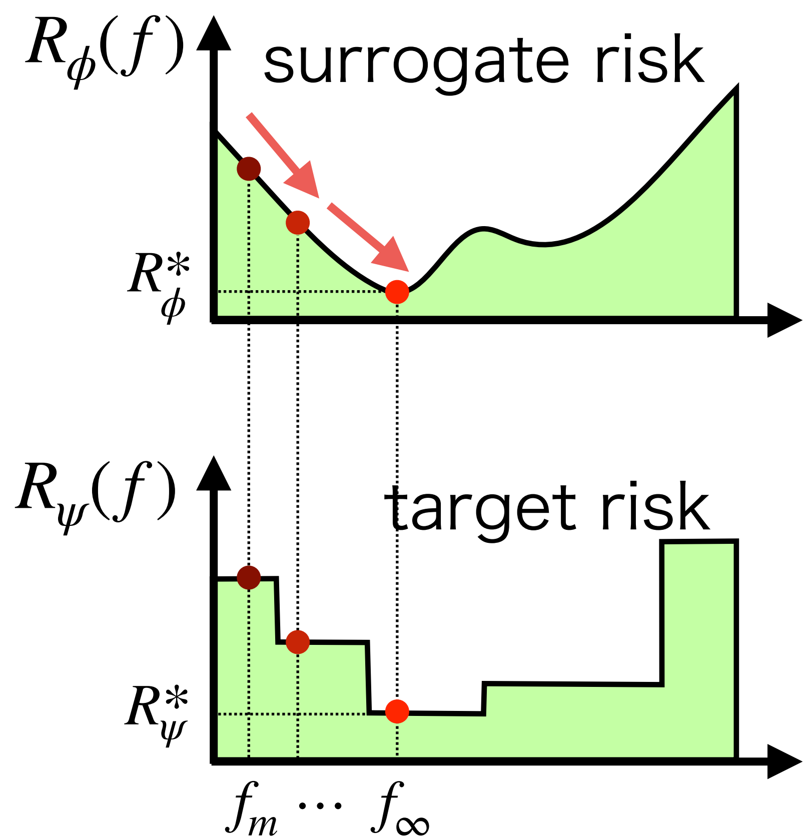
easily optimizable

Target loss (0-1 loss)



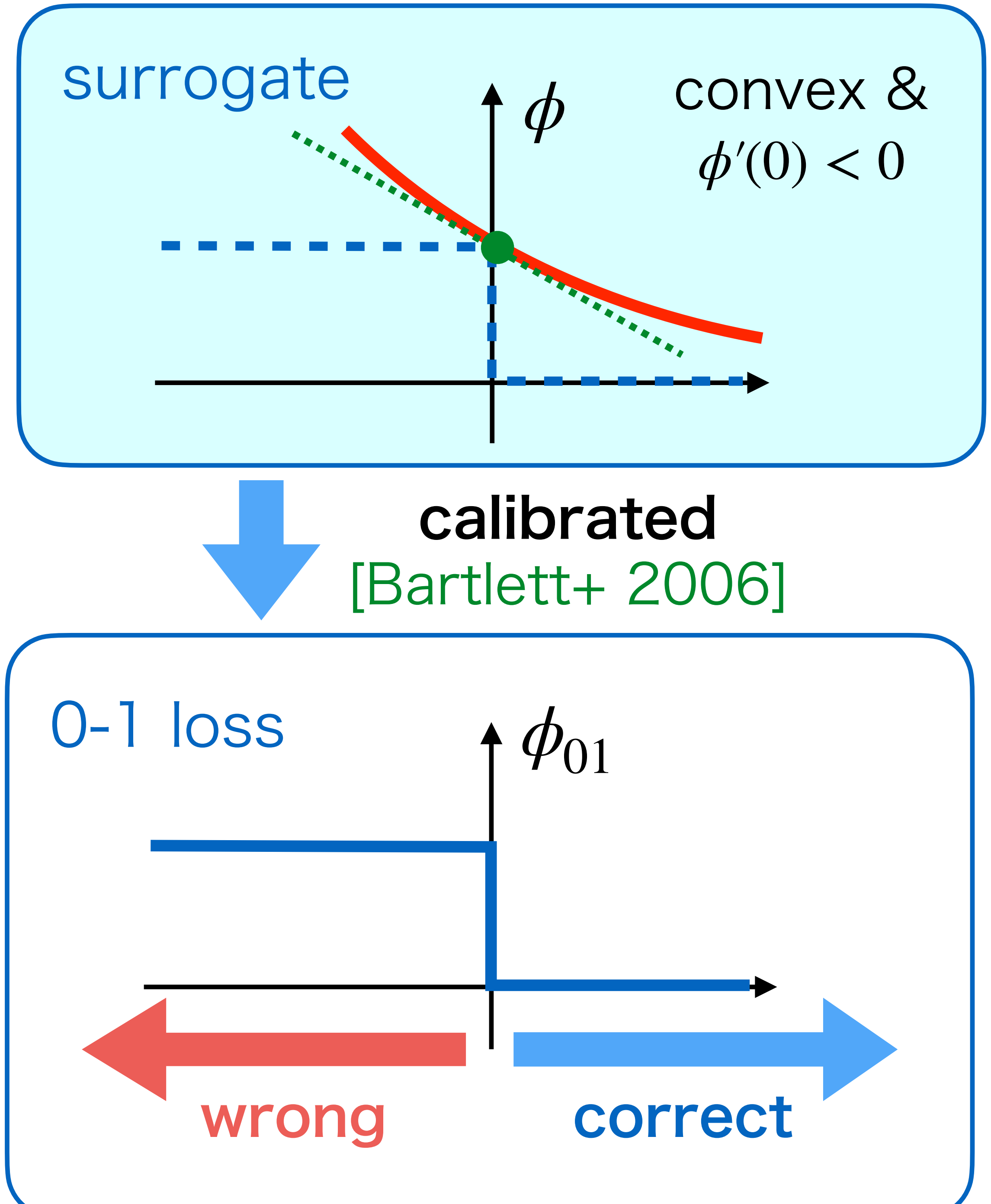
final learning criterion

Calibrated surrogate

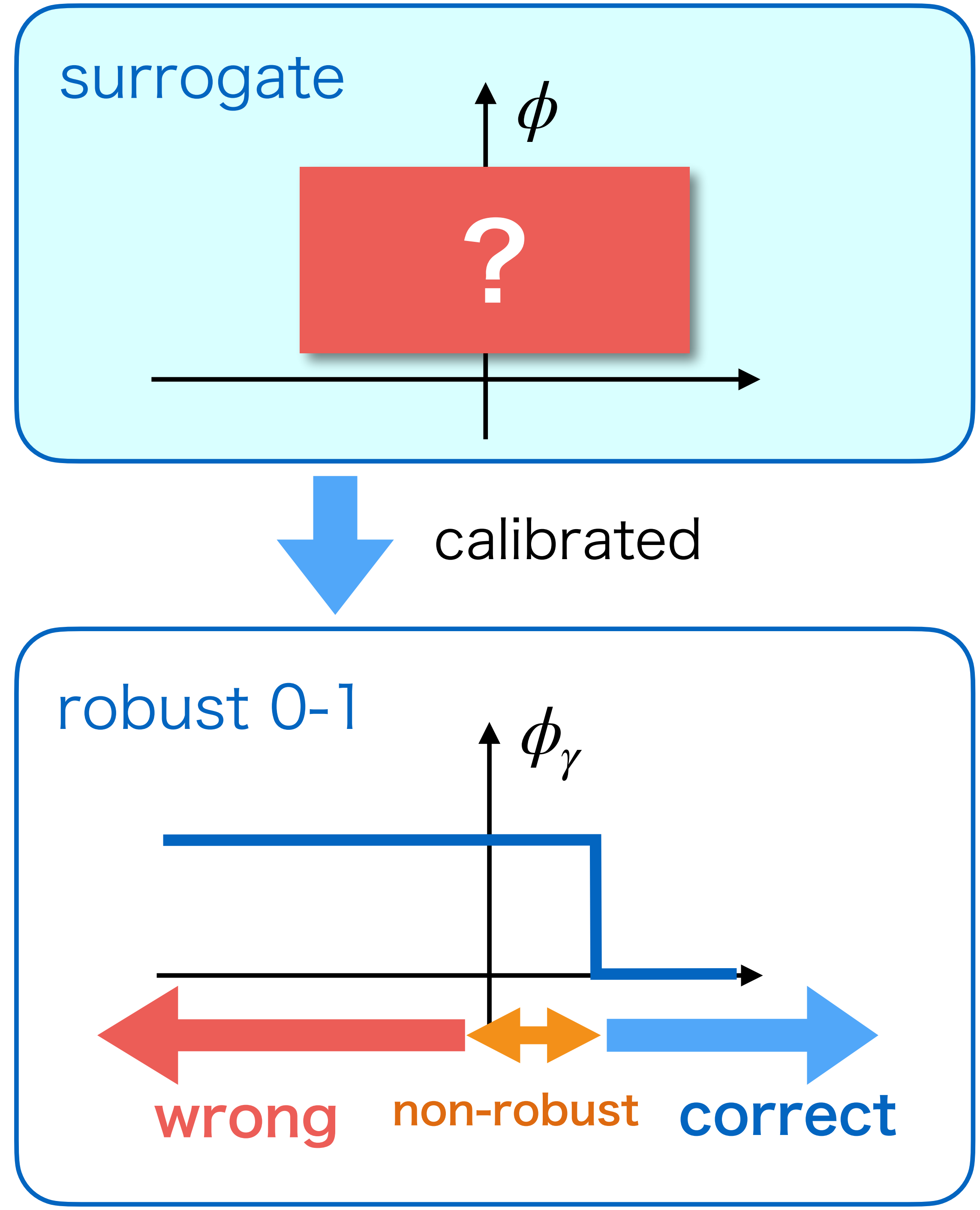


What surrogate is calibrated?

Usual Classification



Robust Classification



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.

Short Course on Calibration Analysis

— how to analyze loss calibration property —

Ingo Steinwart.
How to compare different loss functions and their risks.
Constructive Approximation, 2007.

Conditional Risk and Calibration

Conditional Risk = Risk at a single x

$$R_\phi(f) = \mathbb{E}_X\left[\mathbb{P}(Y=+1\,|\,X)\phi(f(X)) + \mathbb{P}(Y=-1\,|\,X)\phi(-f(X))\right]$$

$\mathbb{P}(Y=+1\,|\,X) := \eta$ (class prob.)

$f(X) := \alpha$ (prediction)

$$C_\phi(\alpha,\eta) := \eta\phi(\alpha) + (1-\eta)\phi(-\alpha)$$

Definition. ϕ is $(\psi,\mathcal{F})$-**calibrated** for a target loss ψ if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all $\alpha \in A_{\mathcal{F}}$ and $\eta \in [0,1]$,

$$C_\phi(\alpha,\eta) - C^*_{\phi,\mathcal{F}}(\eta) < \delta \implies C_\psi(\alpha,\eta) - C^*_{\psi,\mathcal{F}}(\eta) < \varepsilon.$$

surrogate excess conditional risk — target excess conditional risk

$A_{\mathcal{F}} := \{f(x)\,|\,f \in \mathcal{F}, x \in \mathcal{X}\}$

Main Tool: Calibration Function

Definition. (calibration function)

$$\delta(\varepsilon) = \inf_{\eta \in [0,1]} \inf_{\alpha \in A_{\mathscr{F}}} C_\phi(\eta, \alpha) - C^*_{\phi,\mathscr{F}}(\eta) \quad \text{s.t.} \quad C_\psi(\eta, \alpha) - C^*_{\psi,\mathscr{F}}(\eta) \geq \varepsilon$$

surrogate excess conditional risk — target excess conditional risk

- Provides **iff condition**
  - $(\psi,\mathscr{F})$-calibrated $\iff \delta(\varepsilon) > 0$ for all $\varepsilon > 0$
- Provides **excess risk bound**
  - $(\psi,\mathscr{F})$-calibrated $\implies R_\psi(f) - R^*_\psi \leq (\delta^{**})^{-1}\left(R_\phi(f) - R^*_\phi\right)$

target excess risk: $R_\psi(f) - R^*_\psi$; monotonically increasing: $(\delta^{**})^{-1}$; surrogate excess risk: $R_\phi(f) - R^*_\phi$

$A_{\mathscr{F}} := \{f(x) \mid f \in \mathscr{F}, x \in \mathscr{X}\}$

δ^{**}: biconjugate of δ

Example: Binary Classification (ϕ_{01})

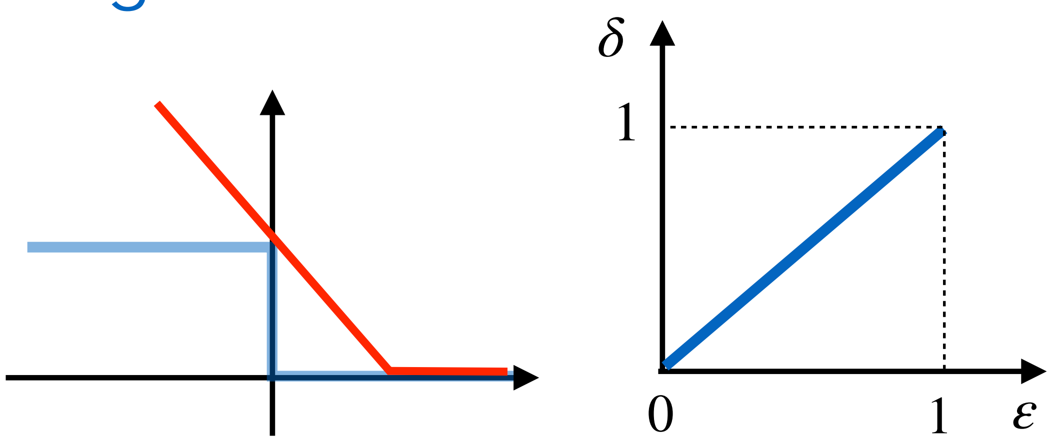
[Bartlett+ 2006]

Theorem. If surrogate ϕ is convex, it is $(\phi_{01}, \mathscr{F}_{\text{all}})$-calibrated iff

- differentiable at 0
- $\phi'(0) < 0$

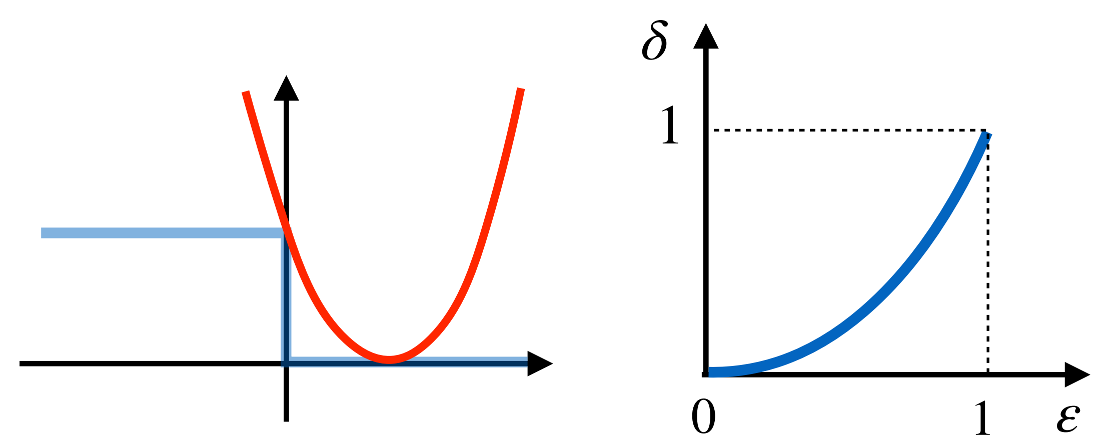
$\mathscr{F}_{\text{all}}$: all measurable functions

hinge loss



$\phi(\alpha) = [1 - \alpha]_+$ $\delta(\varepsilon) = \varepsilon$

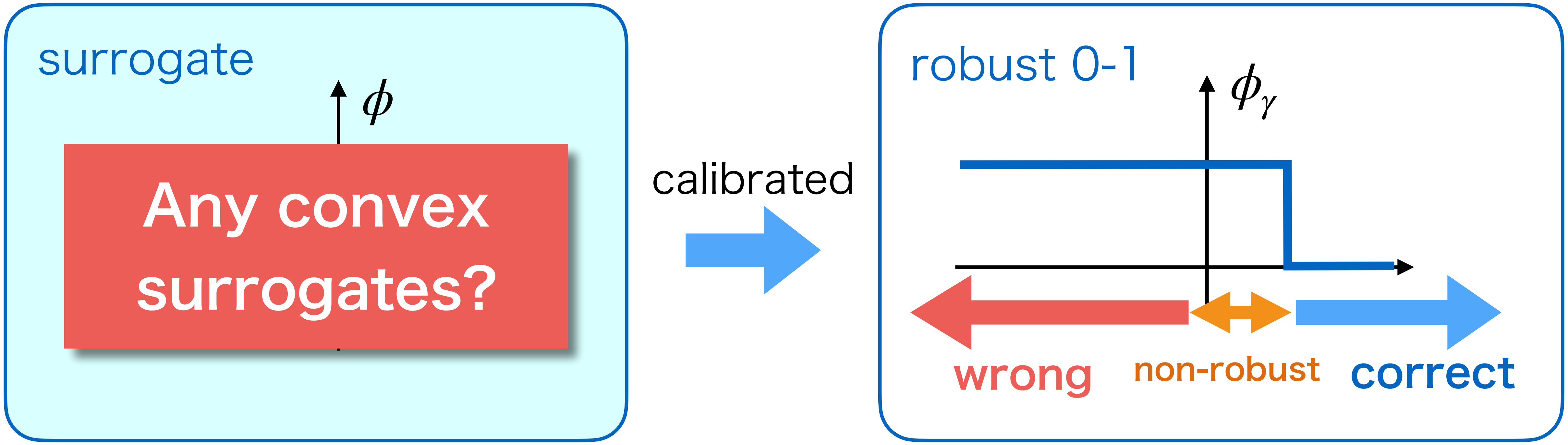
squared loss



$\phi(\alpha) = (1 - \alpha)^2$ $\delta(\varepsilon) = \varepsilon^2$

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.

Analysis of Robust Classification



restricted to linear predictors

No convex calibrated surrogate

Theorem. Any convex surrogate is not $(\phi_\gamma, \mathscr{F}_{\text{lin}})$-calibrated.

Proof Sketch

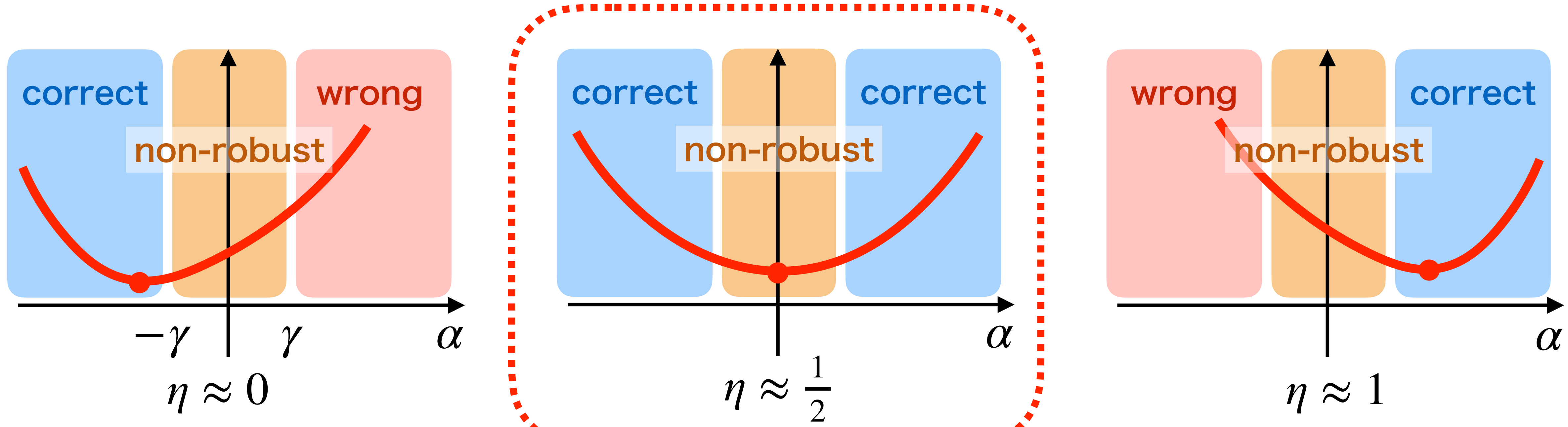
calibration function

convex in α

$|\alpha| \leq \gamma$ is non-robust

$$\delta(\varepsilon) = \inf_{\eta \in [0,1]} \inf_{\alpha \in A_{\mathscr{F}}} C_\phi(\eta, \alpha) - C^*_{\phi, \mathscr{F}}(\eta) \quad \text{s.t.} \quad C_{\phi_\gamma}(\eta, \alpha) - C^*_{\phi_\gamma, \mathscr{F}}(\eta) \geq \varepsilon$$

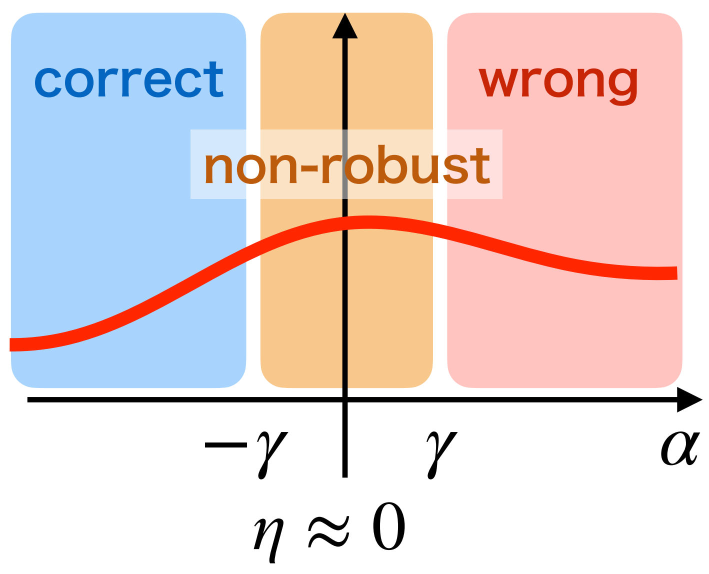
non-robust minimizer!



surrogate conditional risk is plotted

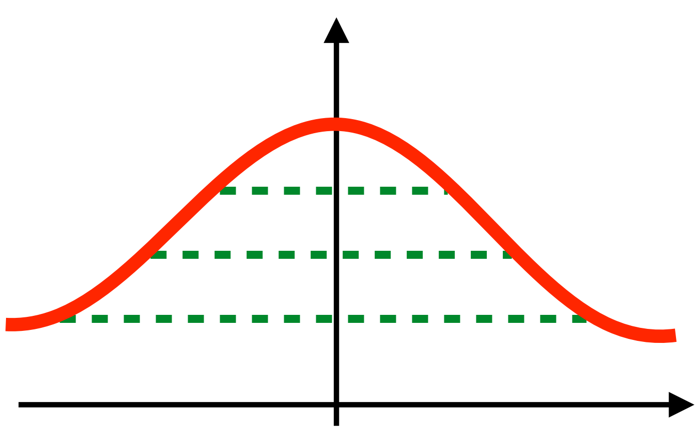
How to find calibrated surrogate?

Idea. To make conditional risk not minimized in **non-robust area**



surrogate conditional risk is plotted

consider a surrogate ϕ such that conditional risk is **quasiconcave**

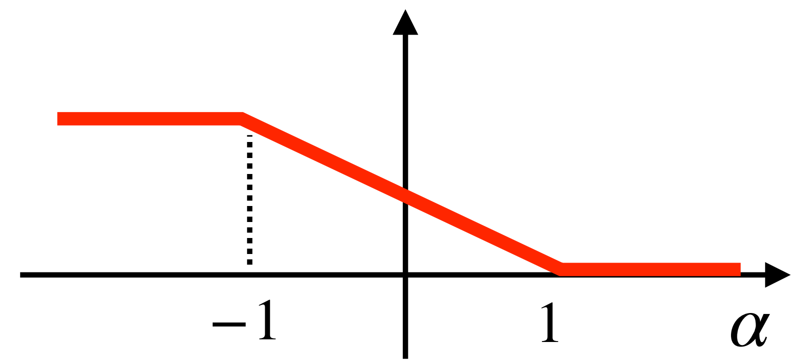


all superlevels are convex

Example: Shifted Ramp Loss

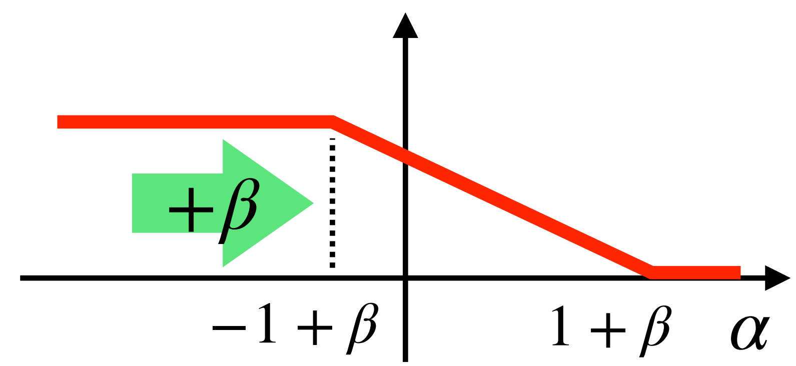
Ramp loss

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

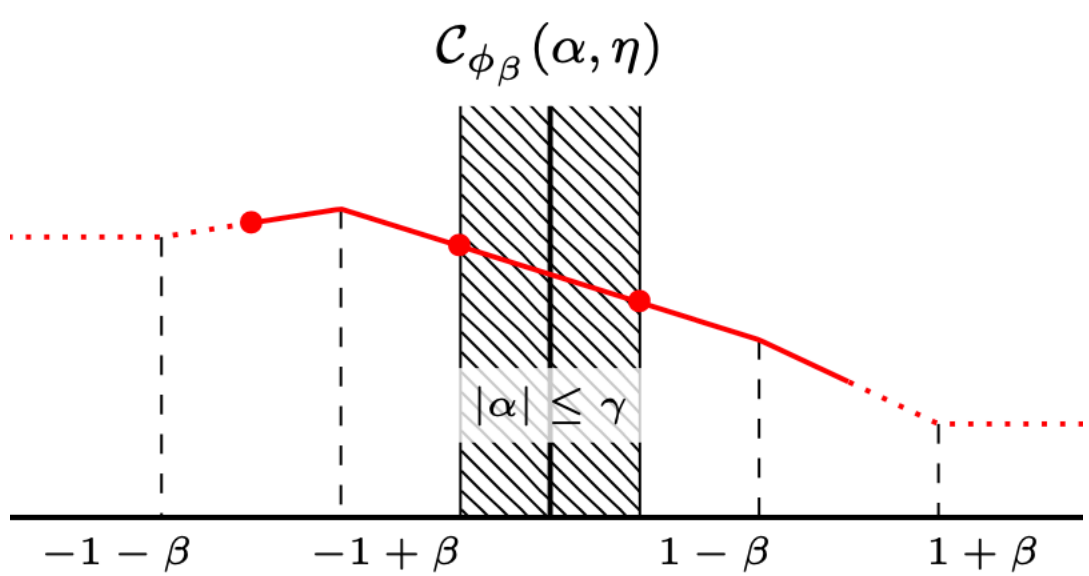


Shifted ramp loss

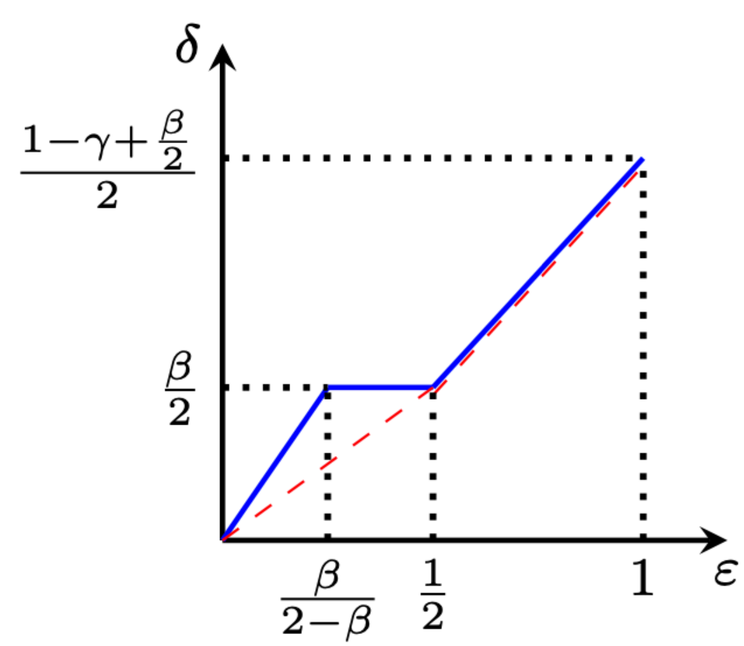
$$\phi_\beta(\alpha) = \mathrm{clip}_{[0,1]}\left(\frac{1-\alpha+\beta}{2}\right)$$



conditional risk $(\eta > 1/2)$



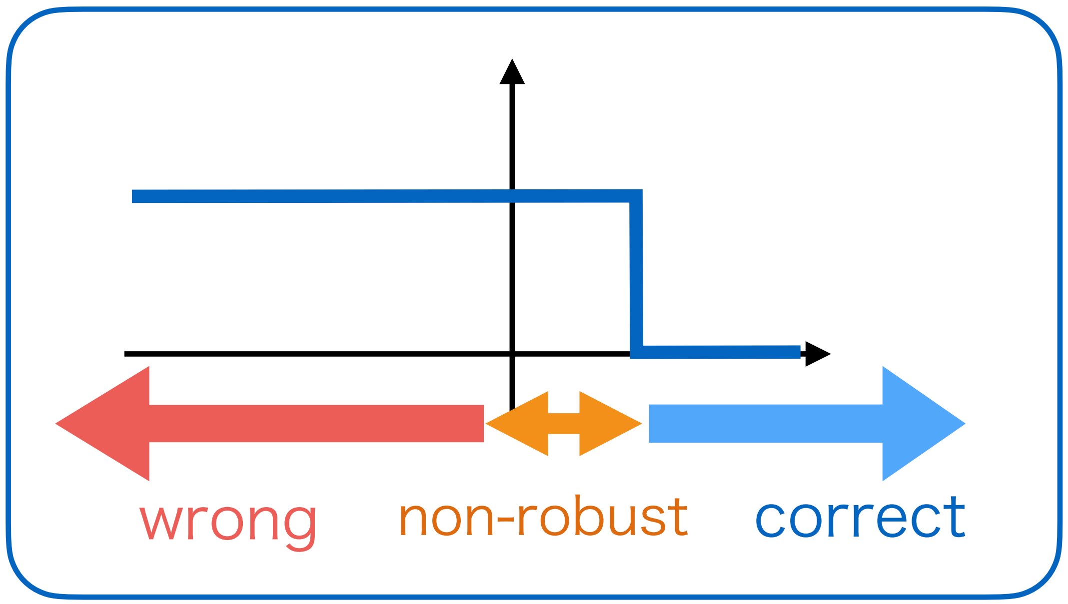
calibration function



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

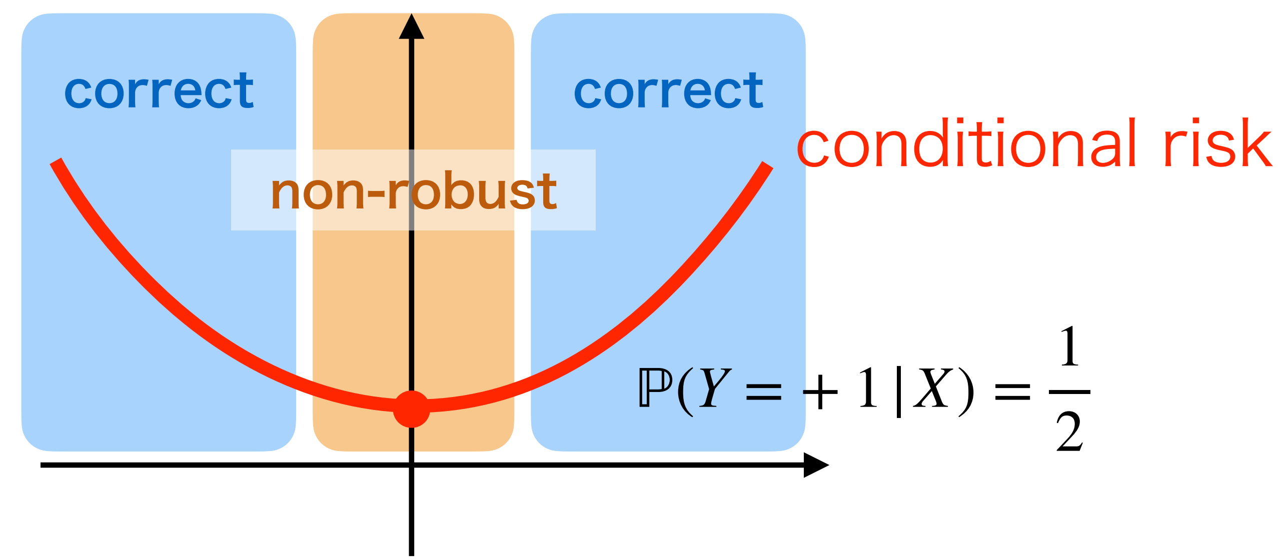
Calibrated Surrogate Losses for Adversarially Robust Classification

Robust classification
= minimize **robust 0-1 loss**



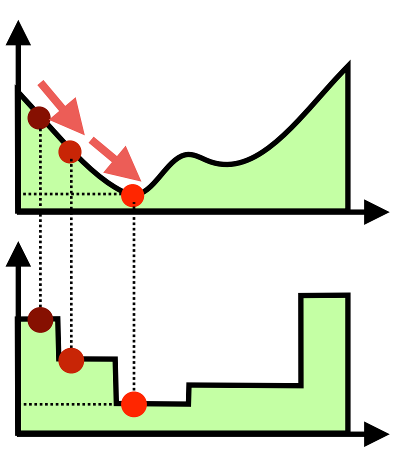
under restriction to linear predictors

No convex calibrated surrogate
under linear predictors



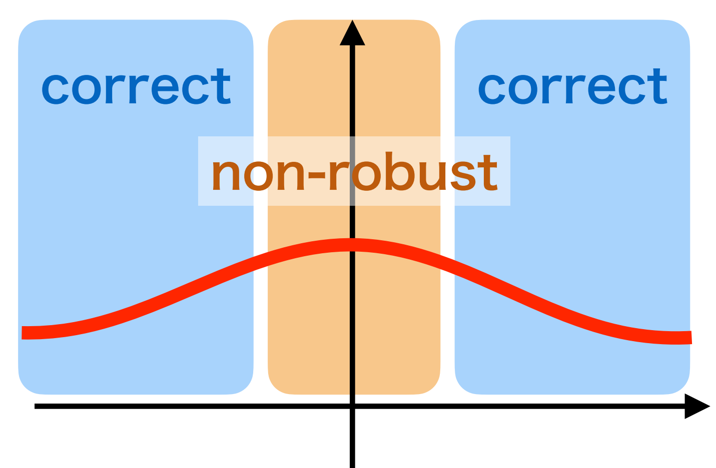
because minimizer lies in non-robust area

Calibrated surrogate loss



minimizing surrogate
⇓
minimizing target

Quasiconcavity is important



Example:
shifted ramp loss