Calibrated Surrogate Losses and Robust Learning

Dec. 14, 2020 (in US) & Dec. 15, 2020 (in Japan)

Han Bao

東京大学
THE UNIVERSITY OF TOKYO

RIKEN

The content is based on our joint work with Clayton Scott (UMich) and Masashi Sugiyama (RIKEN / UTokyo).

Outline

- **Part 1: Calibrated surrogate losses**
  - Q. What are minimum requirements for loss functions?
- Part 2: Loss functions in robust learning
  - Q. Is it possible to design robust loss functions?

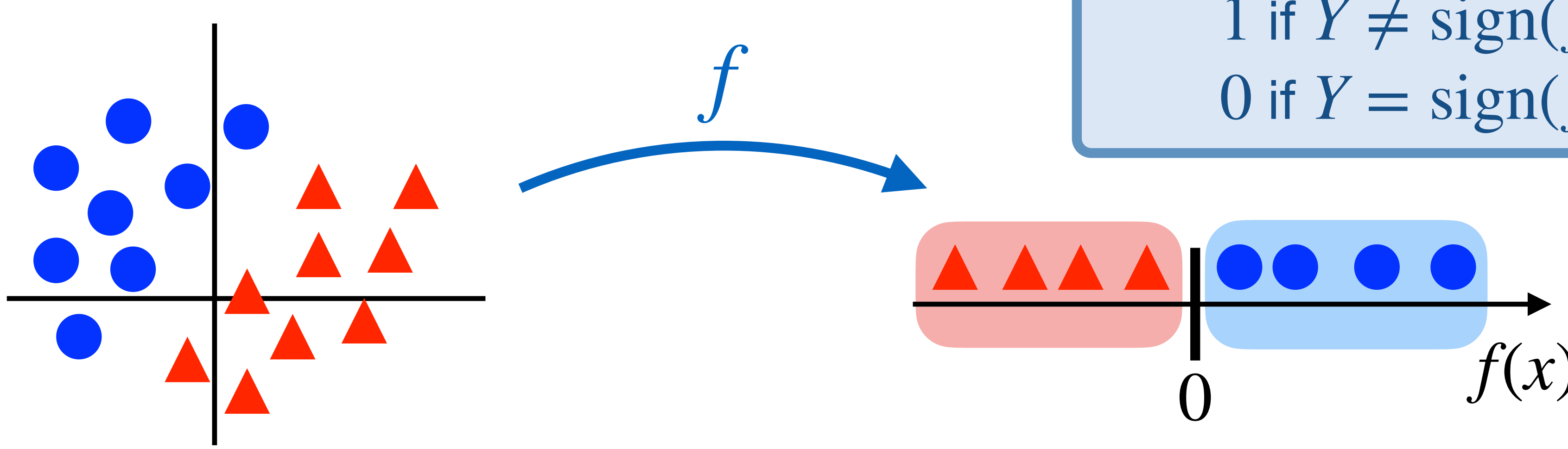
Setting: Binary Classification

- Input
  - sample $\{(x_i, y_i)\}_{i=1}^n$: feature $x_i \in \mathscr{X}$ and label $y_i \in \{\pm 1\}$
- Output: classifier $f : \mathscr{X} \to \mathbb{R}$
  - use $\mathrm{sign}(f(\cdot))$ to predict labels
  - criterion: misclassification rate $R_{01}(f) = \mathbb{E}\left[\mathbf{1}[Y \neq \mathrm{sign}(f(X))]\right]$

0-1 loss $\phi_{01}(Yf(X))$

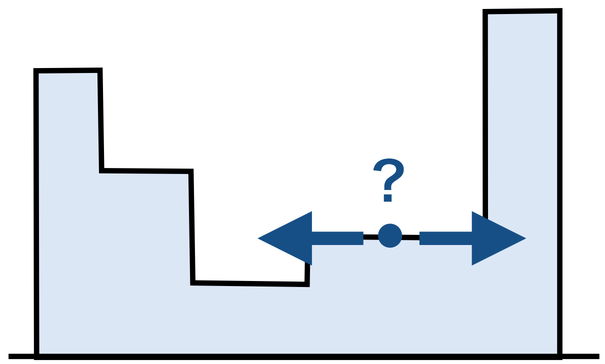
1 if $Y \neq \mathrm{sign}(f(X))$

0 if $Y = \mathrm{sign}(f(X))$

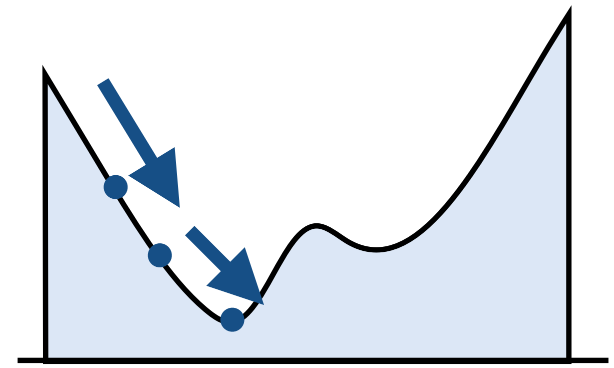


Surrogate Losses

- Motivation: minimizing 0-1 loss is NP-hard

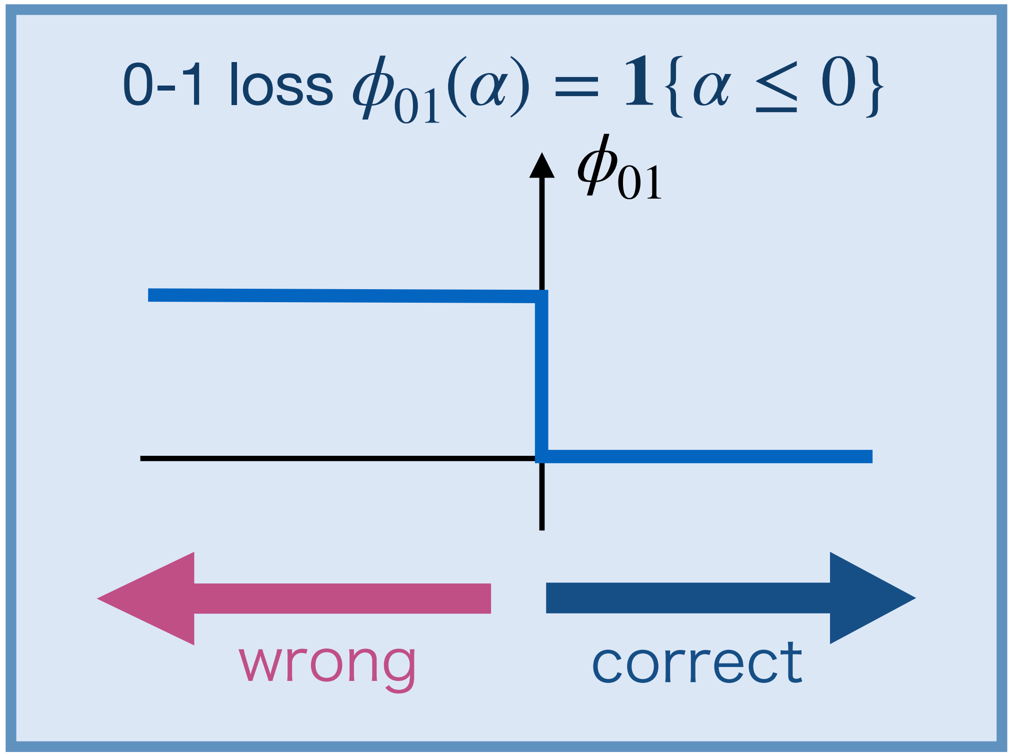


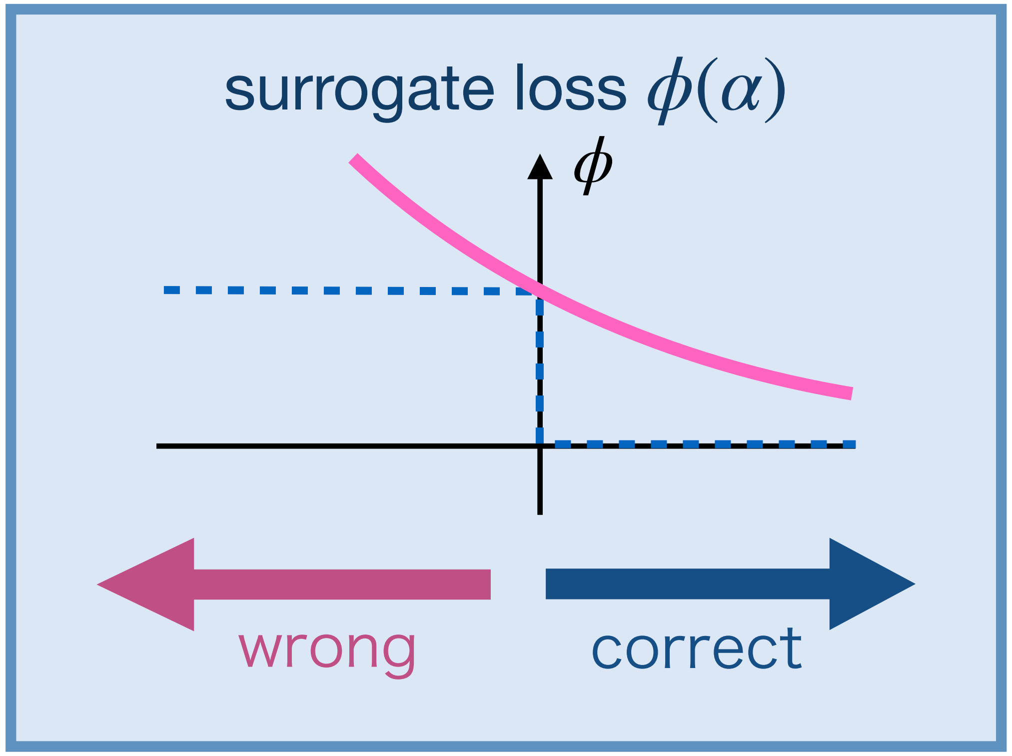
no gradient for discrete function



easily optimizable if convex and smooth

- Replace 0-1 loss with surrogate loss





hinge loss,
logistic loss, etc.

Elements of Learning Theory

(empirical) surrogate risk

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

Generalization theory:
If model is not too complicated, then justified (roughly speaking)

(population) surrogate risk

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

Our interests: **Calibration theory**

target risk

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

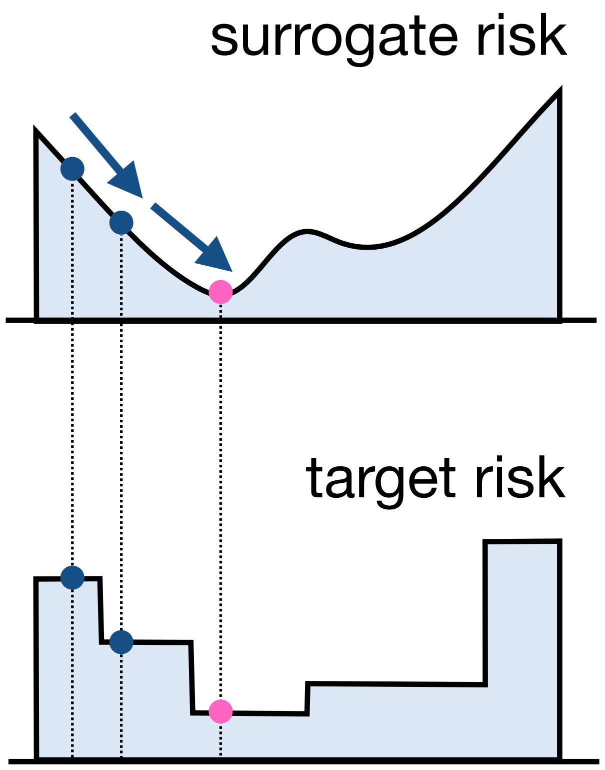
Q. What is a desirable surrogate?

- A. surrogate risk minimizer should be target risk minimizer

For two losses ψ (target) and ϕ (surrogate),

Definition. Surrogate ϕ is **calibrated** to target ψ if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all f,

$$R_\phi(f) - R_\phi^* < \delta \implies R_\psi(f) - R_\psi^* < \varepsilon.$$



How to check calibration?

[Steinwart 2007]

Definition. Surrogate ϕ is calibrated to target ψ if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all f,

$$R_\phi(f) - R_\phi^* < \delta \implies R_\psi(f) - R_\psi^* < \varepsilon.$$

ε-δ definition of limit

$$\Updownarrow$$

$$R_\psi(f) - R_\psi^* \geq \varepsilon \implies R_\phi(f) - R_\phi^* \geq \delta$$

contraposition

Definition. (calibration function)

$$\delta(\varepsilon) = \underbrace{\inf_f R_\phi(f) - R_\phi^*}_{\text{smallest possible } \delta} \quad \text{s.t.} \quad \underbrace{R_\psi(f) - R_\psi^* \geq \varepsilon}_{\text{given lower bound of target}}$$

easy to ask existence of $\delta > 0$ given ε

Disclaimer: calibration function is defined over class-conditional risk to be precise

Main Tool: Calibration Function

[Steinwart 2007]

Definition. (calibration function)

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

smallest possible surrogate given lower bound of target

- Provides iff condition
  - ❖ calibrated to $\psi \iff \delta(\varepsilon) > 0$ for all $\varepsilon > 0$
- Provides excess risk bound
  - ❖ calibrated to $\psi \iff R_\psi(f) - R_\psi^* \leq (\delta^{**})^{-1}\left(R_\phi(f) - R_\phi^* \right)$

target risk monotone surrogate risk

minimizing surrogate risk = minimizing target risk
(we know convergence rate in addition)

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

Case: Binary Classification

[Bartlett *et al.* 2006]

Calibration function for 0-1 loss

$$\delta(\varepsilon) = \underbrace{\inf_f R_\phi(f) - R_\phi^*}_{\text{smallest possible surrogate}} \quad \text{s.t.} \quad \underbrace{R_{\phi_{01}}(f) - R_{\phi_{01}}^* \geq \varepsilon}_{\text{given lower bound of 0-1}}$$

iff condition: $\delta(\varepsilon) > 0 \quad \forall \varepsilon > 0$

calibrated iff

$$\inf_f \left\{ R_\phi(f) \,\middle|\, \underbrace{R_{\phi_{01}}(f) > R_{\phi_{01}}^*}_{f\text{ is non-optimal wrt 0-1 loss}} \right\} > \inf_f R_\phi(f)$$

minimum risk of non-optimal classifiers
V
minimum risk of all classifiers

- Check the latter condition to see if calibrated
- More simple equivalent conditions available (next slide)

Disclaimer: several literature defines calibration by the latter condition

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.

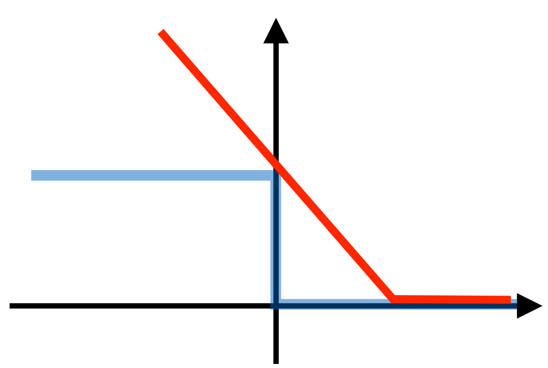
Case: Binary Classification

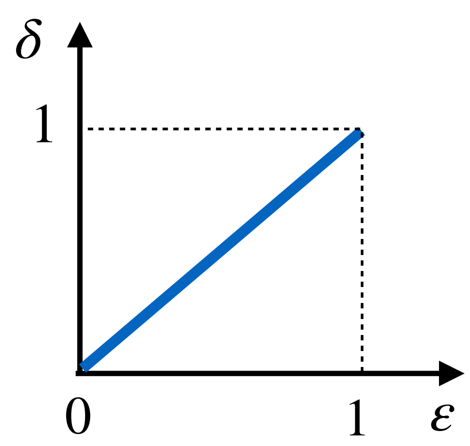
[Bartlett *et al.* 2006]

Theorem. If surrogate ϕ is convex, it is calibrated to ϕ_{01} iff

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

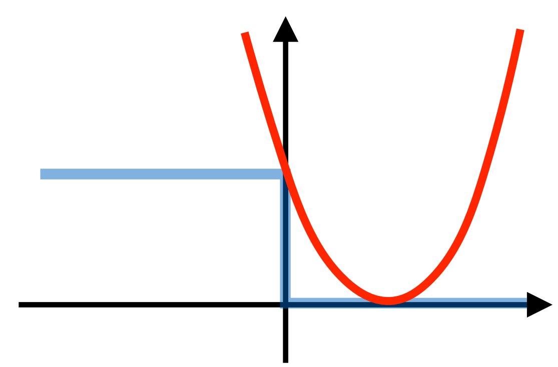
hinge loss

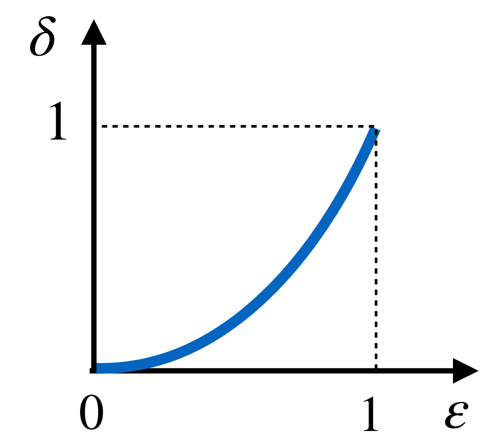




$\phi(\alpha) = [1 - \alpha]_+$ $\delta(\varepsilon) = \varepsilon$

squared loss





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

- Most of well-known losses are calibrated
  - perceptron loss $\phi(\alpha) = [-\alpha]_+$ is not

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.

Outline

- Part 1: Calibrated surrogate losses
  - ❖ Q. What are minimum requirements for loss functions?

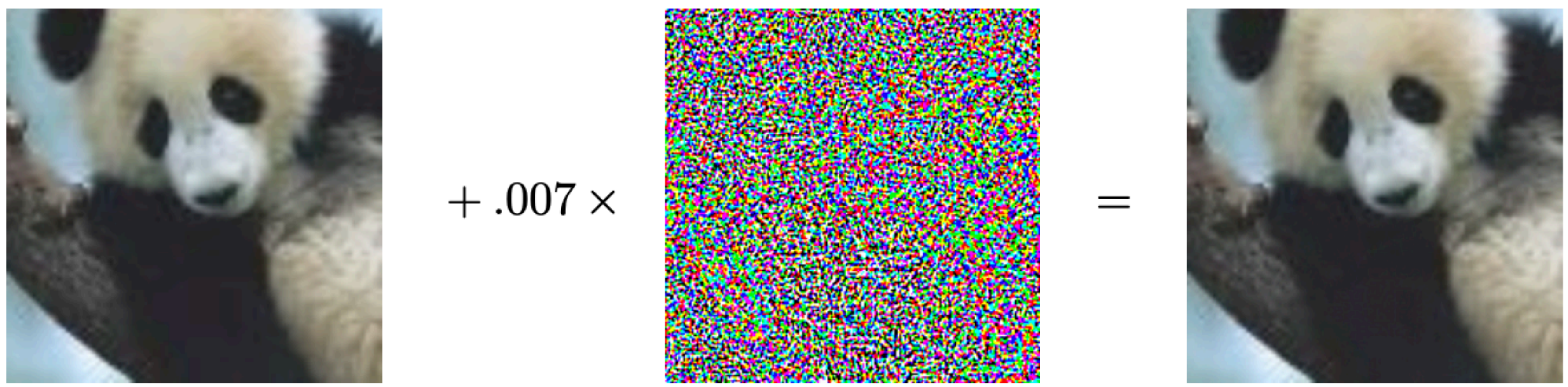
A. calibration: surrogate minimizer = target minimizer
- confirmed via calibration function
- simple iff conditions for binary classification

- **Part 2: Loss functions in robust learning**
  - ❖ Q. Is it possible to design robust loss functions?

Classifier is vulnerable to "attacks"

[Goodfellow *et al.* 2015]

- Adversarial attacks:
  manipulate predictions by adding imperceptible small noise



$+ .007 \times$ $=$

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

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

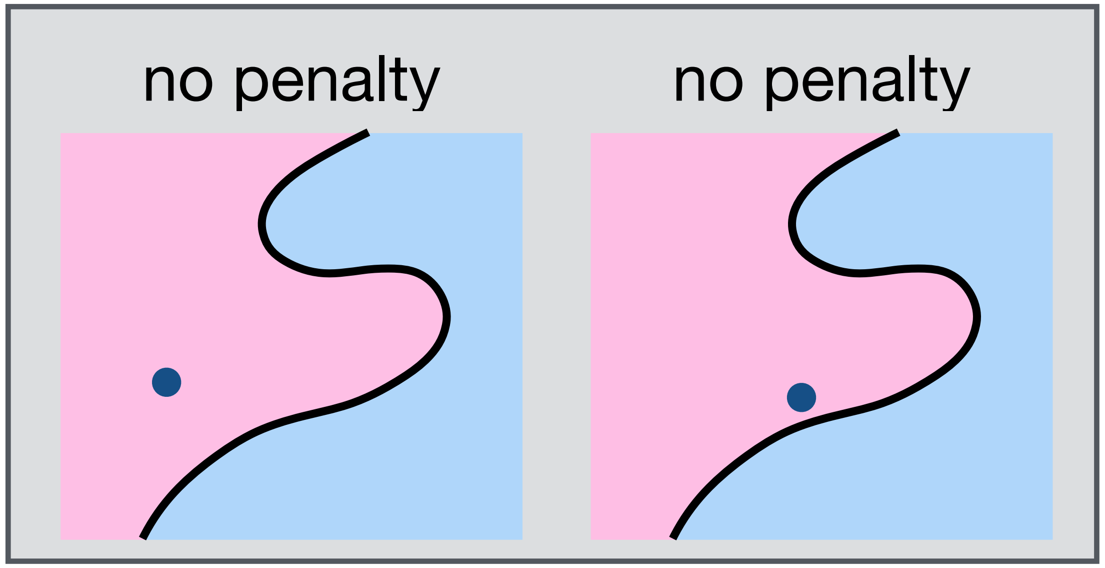
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$ "gibbon" 99.3 % confidence

- More interests in whether our learning method are robust
  - important in applications such as autonomous driving, medical diagnosis

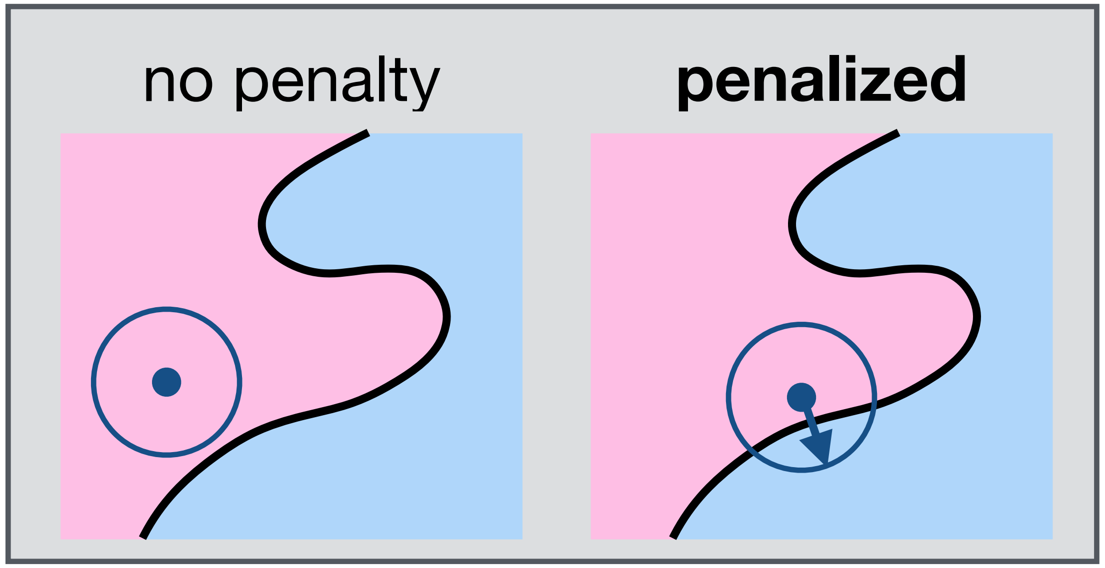
Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *ICLR*, 2015.

Formulation of Adversary

- Standard learning:
  no penalty if classified to
  the correct side of the boundary



- Robust learning:
  prediction close to the boundary
  will be penalized even if
  correctly classified
  - the boundary will be crossed over by attacks
  - assume L_2-ball attack

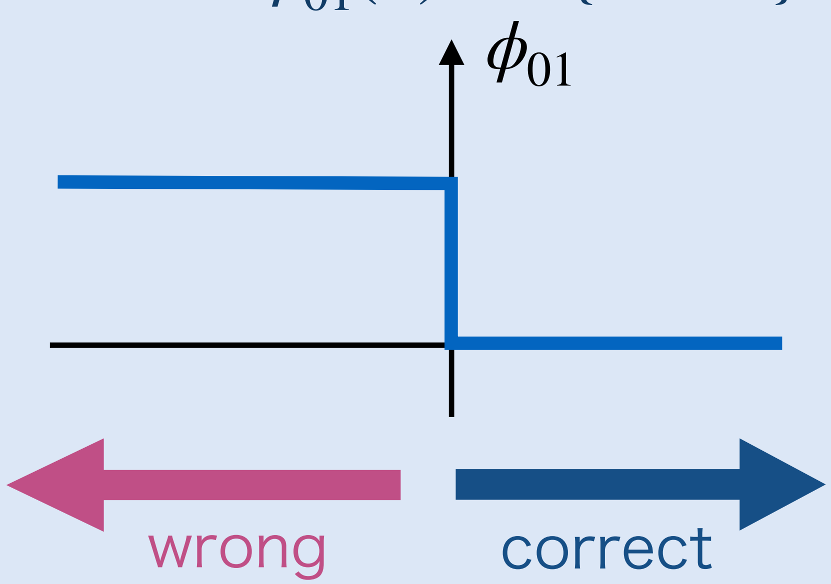


Standard vs. Robust Learning

- Standard learning: minimize 0-1 loss

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

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



- Robust learning: minimize robust 0-1 loss

$$R_{\phi_\gamma}(f) = \mathbb{E}\left[\underbrace{\max_{\Delta \in B_2(\gamma)}}_{\text{worst } L_2\text{-attack}} \phi_{01}(Yf(X+\Delta))\right]$$

learn best (min) classifier
under worst-case (max) attack
= **robust optimization**

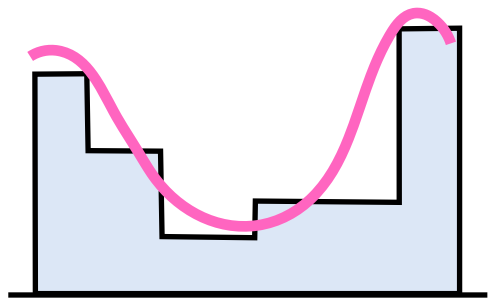
Relaxation of Robust Optimization

- Direct optimization of robust 0-1 loss is hard
- Existing relaxation

Not necessarily calibrated to robust 0-1 loss!

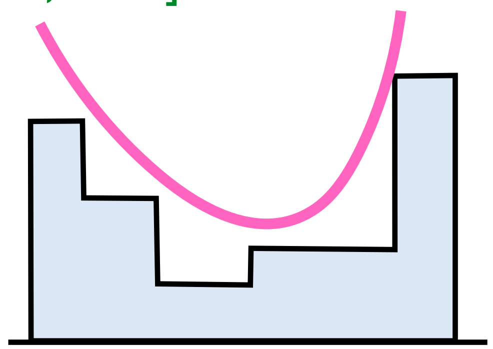
- ❖ Taylor approximation [Shaham *et al.* 2018; etc.]

  local approximation of original objective
  does not necessarily lead to global minimum



- ❖ Minimize convex upper bound [Wong & Kolter 2018; etc.]

  global minimum of upper bound
  does not necessarily equal to minima
  of original objective



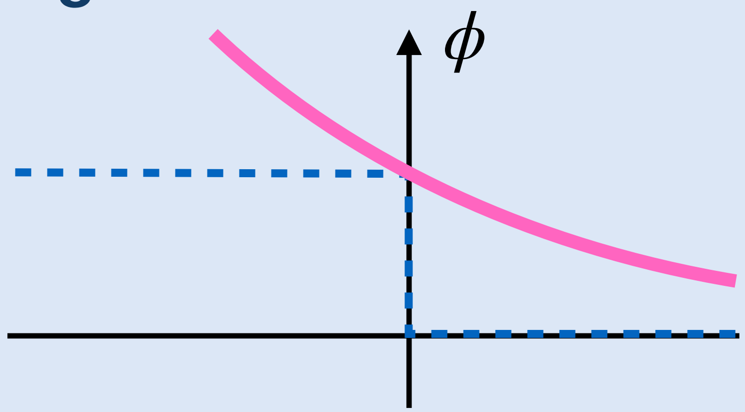
Shaham, U., Yamada, Y., & Negahban, S. (2018). Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 195-204.

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).

What is calibrated surrogates?

- Standard learning

Surrogate loss



calibrated
[Bartlett *et al.* 2006]

Target loss = 0-1 loss

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

- Robust learning

Q. What kind of losses are calibrated?

calibrated

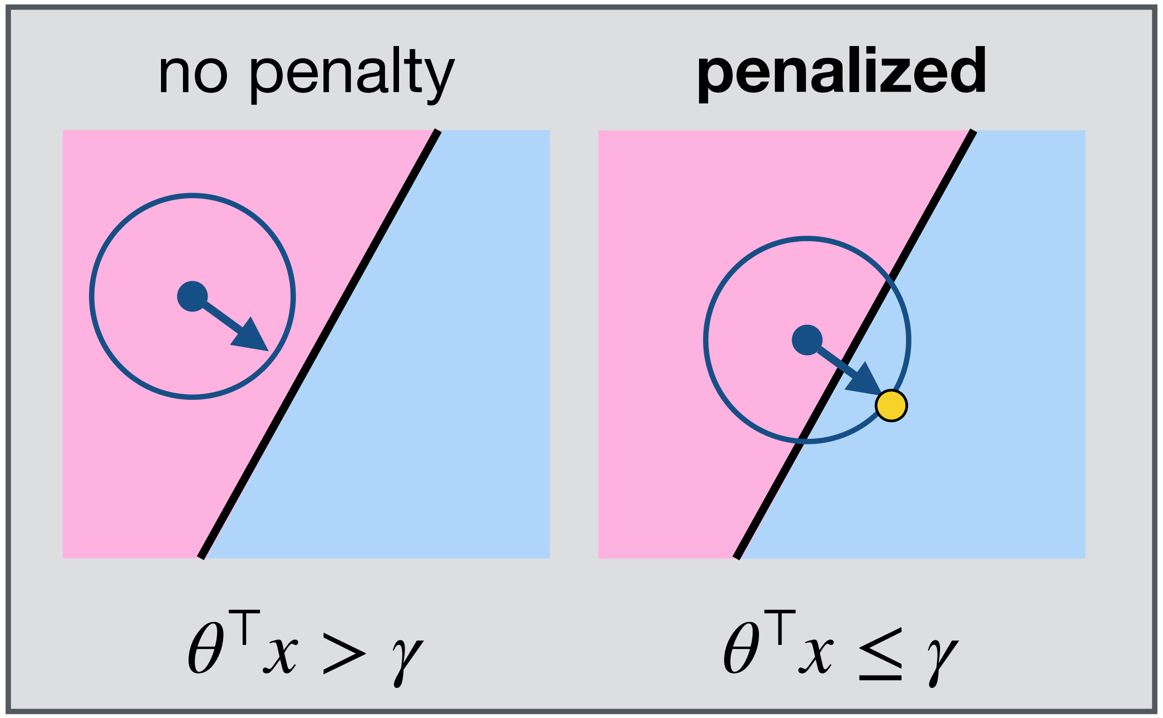
Target loss = robust 0-1 loss

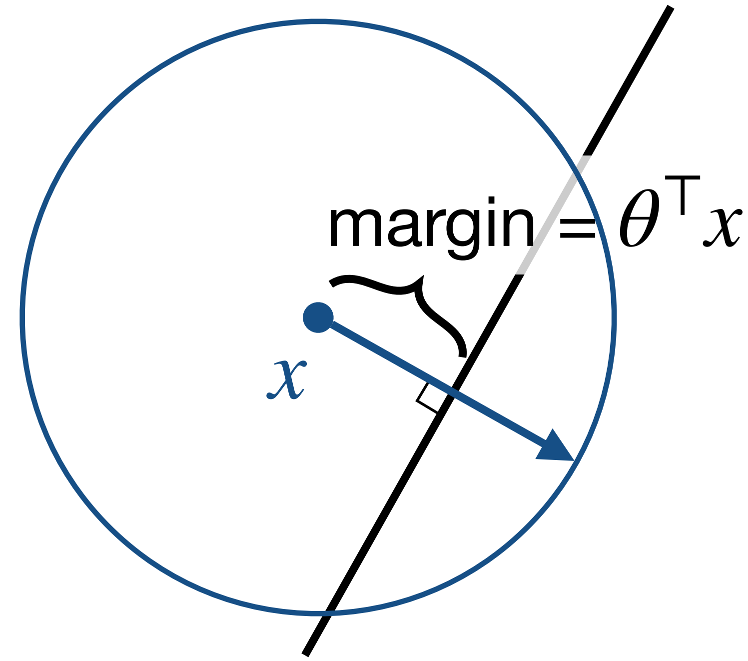
$$R_{\phi_\gamma}(f) = \mathbb{E}\left[\max_{\Delta \in B_2(\gamma)} \phi_{01}(Yf(X+\Delta))\right]$$

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: linear model + L$_2$-attack

- Linear model $f_\theta(x) = \theta^\top x$ where $\|\theta\|_2 = 1$





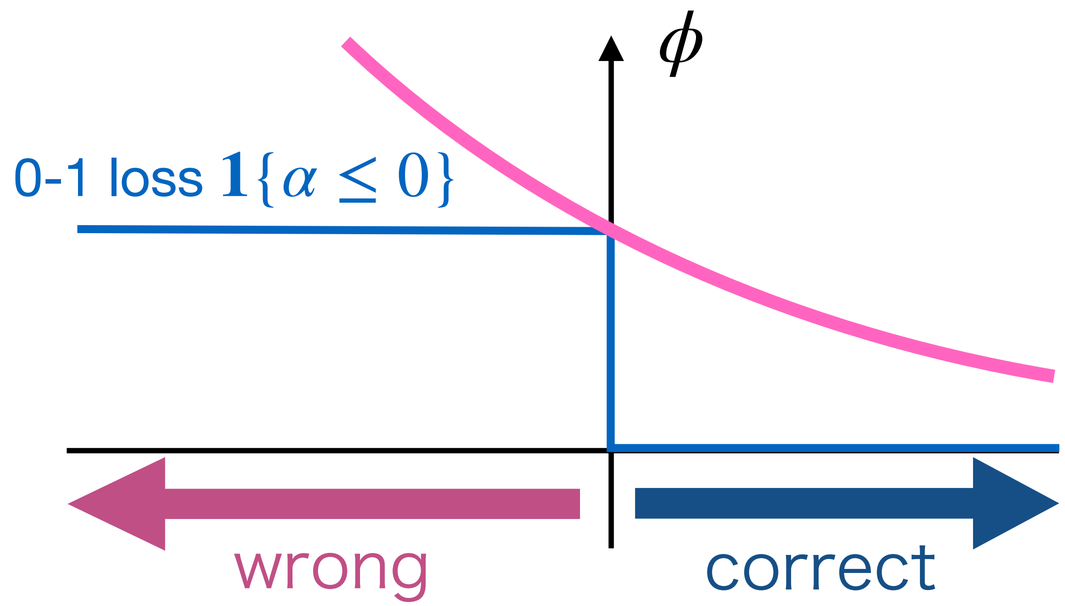
robust 0-1 loss

$$\max_{\Delta \in B_2(\gamma)} \phi_{01}(Yf(X+\Delta)) \quad = \quad \mathbf{1}\{Yf(X) \leq \gamma\} := \phi_\gamma(Yf(X))$$

- General case is hard to analyze

Isn't it a piece of cake?

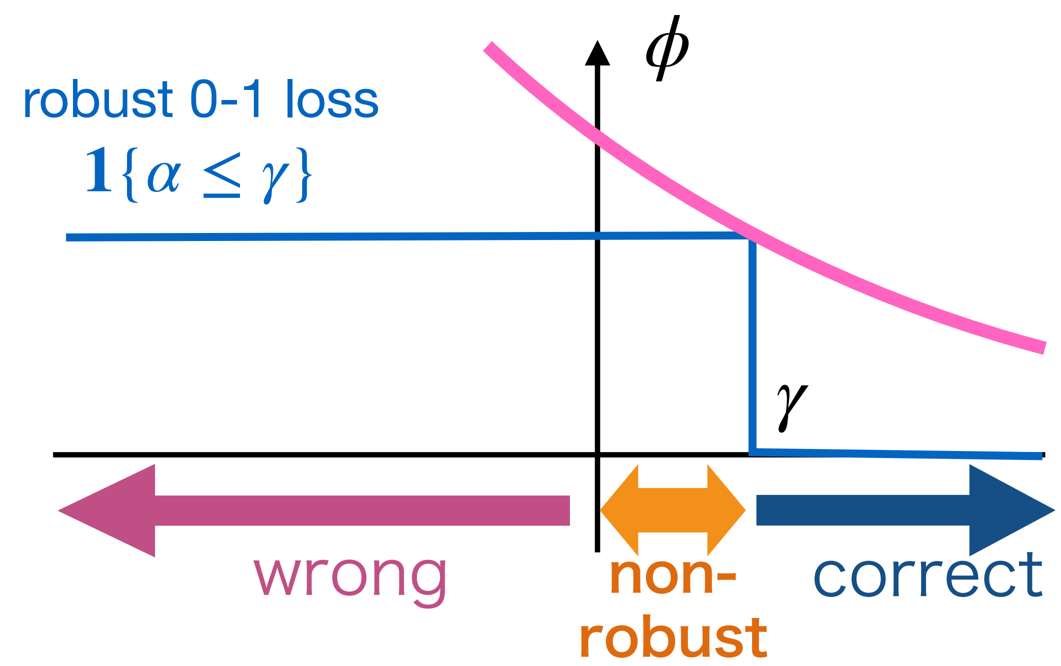
- Standard learning



Theorem [Bartlett *et al.* 2006]**.**
If surrogate ϕ is convex,
1. ϕ is differentiable at 0
2. $\phi'(0) < 0$

are necessary and sufficient for calibration.

- Robust learning



Conjecture.
If surrogate ϕ is convex,
1. ϕ is differentiable at $\alpha = \gamma$
2. $\phi'(\gamma) < 0$

are necessary and sufficient?

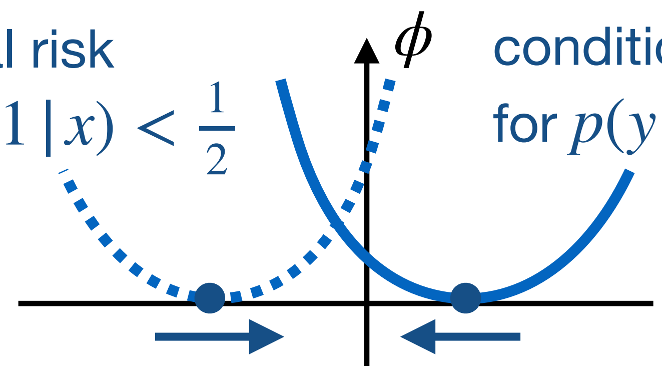
P. L. Bartlett, M. I. Jordan, & J. D. McAuliffe. (2006).
Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473), 138-156.

Main Result

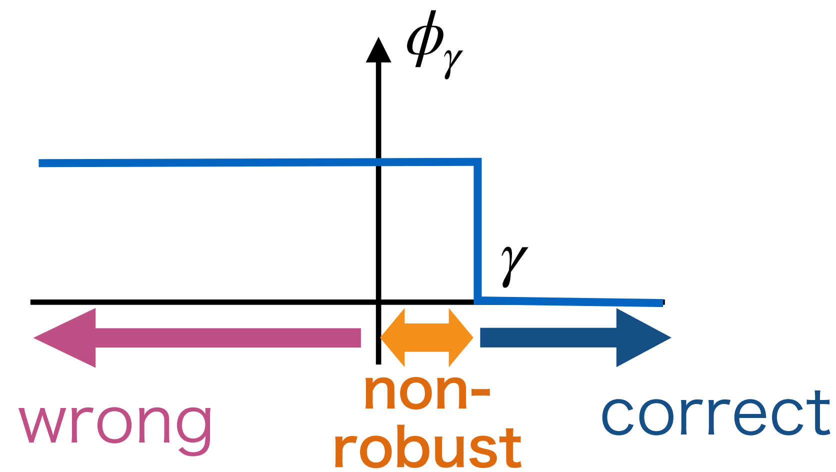
Theorem [Bao *et al.* 2020]**.**
Any convex surrogates are not calibrated to robust 0-1 loss under linear models + L_2 attack.

- Intuition (Note: proven by checking $\delta(\varepsilon) = 0$ for some ε to be precise)
  1. predictions becomes close to 0 as $p(y \mid x) \to \frac{1}{2}$
  2. predictions close to 0 are regarded as non-robust

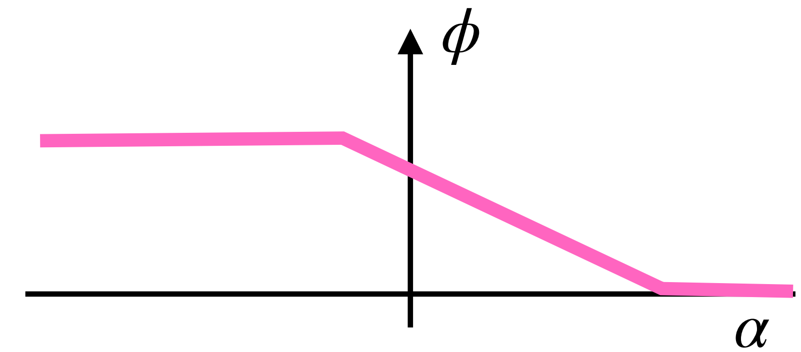
conditional risk for $p(y = 1 \mid x) < \frac{1}{2}$



conditional risk for $p(y = 1 \mid x) > \frac{1}{2}$



- Nonconvex calibrated surrogates exist
  - e.g. ramp loss

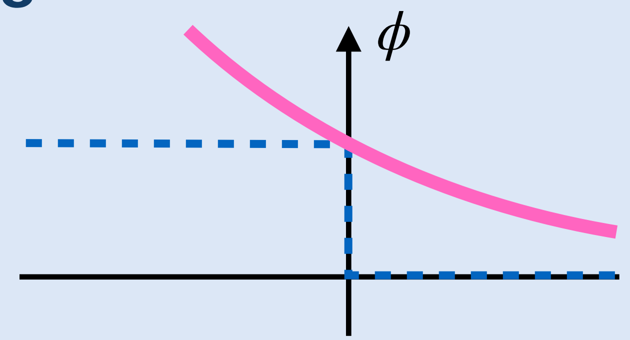


Bao, H., Scott, C., & Sugiyama, M. (2020).
Calibrated Surrogate Losses for Adversarially Robust Classification. In *COLT*, 2020.

Summary | Calibrated Surrogates and Robust Learning

- Standard learning

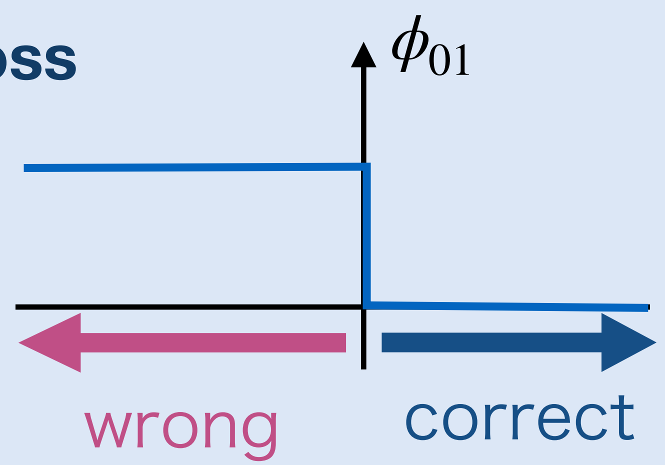
Surrogate loss



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

calibrated
[Bartlett *et al.* 2006]

0-1 loss



- Robust learning

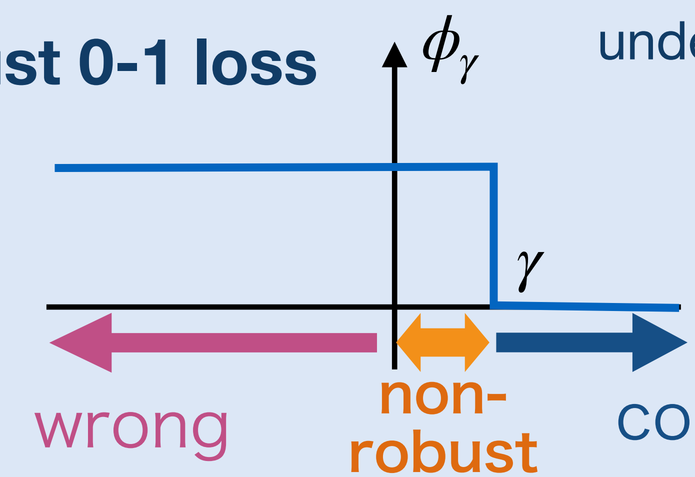
Result

- no convex calibrated surrogates
- nonconvex calibrated surrogates exist (e.g. ramp loss)

calibrated
[Bao *et al.* 2020]

robust 0-1 loss

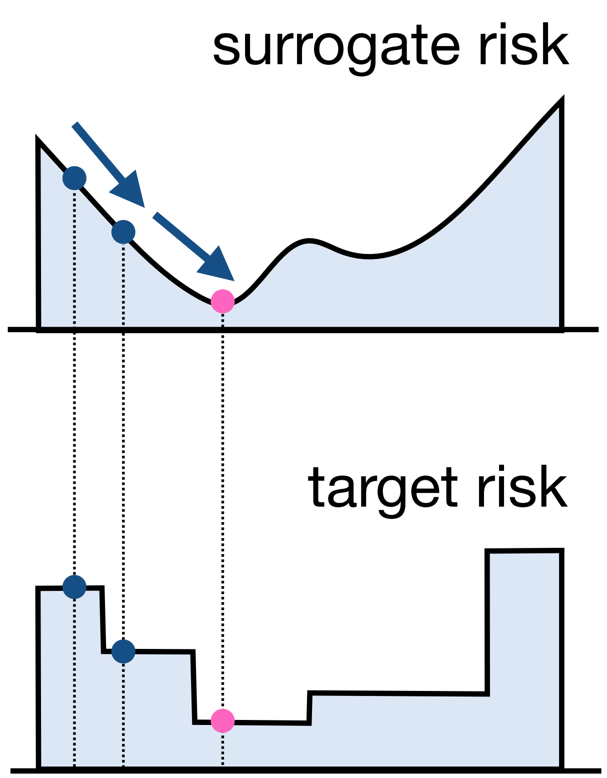
under linear model + L_2-attack



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

Summary

- **Calibrated surrogate losses**:
  surrogate risk minimizer = target risk minimizer
  - ❖ can be confirmed via calibration function
- **Robust learning** from calibration perspective:
  ⇒ no convex calibrated losses
  - ❖ future: how about minimax surrogates?
- Take home:



calibration is interesting not only for minimizer consistency
but also for robust loss design!

- ❖ similar idea adopted to analyze robustness
  to symmetric label noise [Reid & Williamson 2010]

Reid, M. D., & Williamson, R. C. (2010). Composite binary losses. *Journal of Machine Learning Research*, *11*, 2387-2422.