Generalized Strategic Classification and the Case of Aligned Incentives

Sagi Levanon and **Nir Rosenfeld**
Technion CS

@ICML 2022

Outline

- Strategic classification (SC)
- Generalized strategic classification (GSC)
- Incentive alignment
- Learning in GSC – not what you think!

standard classification:

learned model

train: $\underset{h}{\operatorname{argmin}}\ \mathbb{E}[\mathbb{1}\{y \neq h(x)\}]$

input features

test: $h(x) = \hat{y} \approx y$

prediction *ground truth*

strategic

classification:

user features

train: $\underset{h}{\operatorname{argmin}}\ \mathbb{E}[\mathbb{1}\{y \neq h(x)\}]$

test: $h(\Delta_h(x)) = \hat{y} \approx y$

modified features

in response to learned model

strategic

classification:

user features

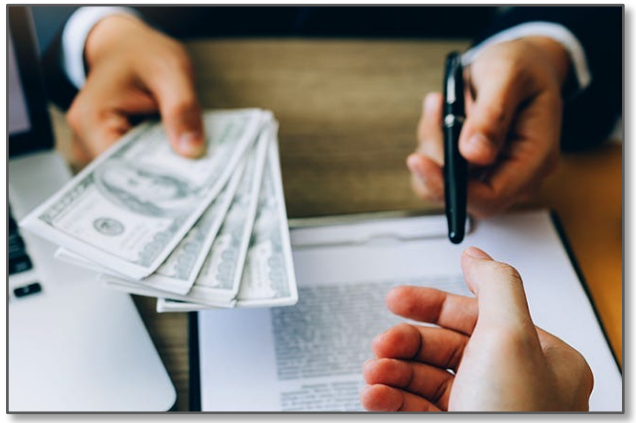
train: $\underset{h}{\operatorname{argmin}}\ \mathbb{E}[\mathbb{1}\{y \neq h(x)\}]$

test: $h(\Delta_h(x)) = \hat{y} \approx y$

modified features — $\Delta_h(x)$

in response to learned model — h

want positive predictions (*e.g., loan approved*) — $\hat{y}$

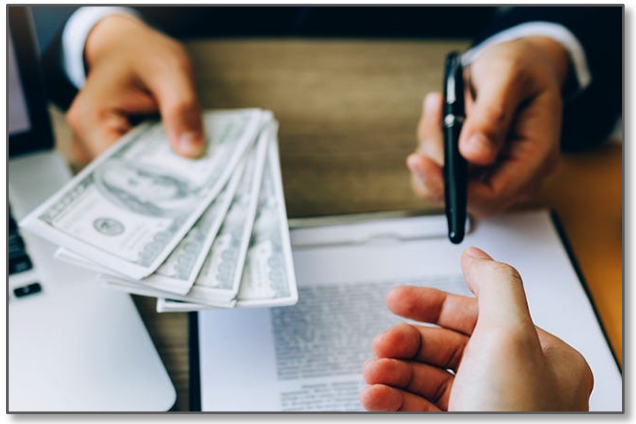


strategic

classification:

train: $\operatorname{argmin}_h \mathbb{E}[\mathbb{1}\{y \neq h(x)\}]$

test: $h(\Delta_h(x)) = \hat{y} \approx y$



response: $\Delta_h(x) = \operatorname{argmax}_{x'} h(x') - c(x, x')$

utility = prediction *(want $\hat{y} = 1$)*

cost of modifying *(assumed known)*

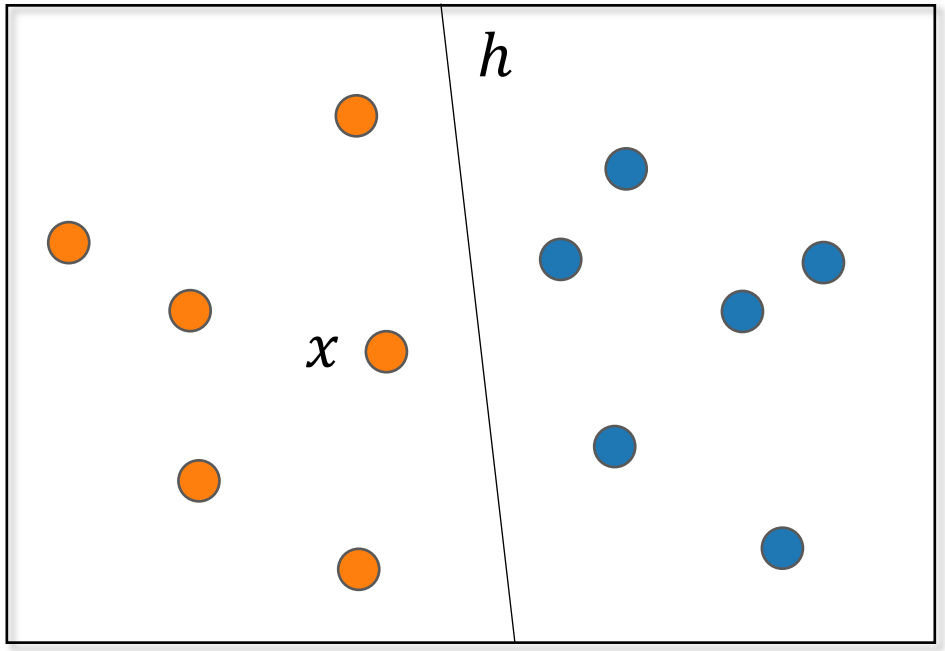
strategic

classification:

train: $\operatorname{argmin}_h \mathbb{E}[\mathbb{1}\{y \neq h(x)\}]$

test: $h(\Delta_h(x)) = \hat{y} \approx y$

response: $\Delta_h(x) = \operatorname{argmax}_{x'} h(x') - c(x, x')$



optimal non-strategic classifier

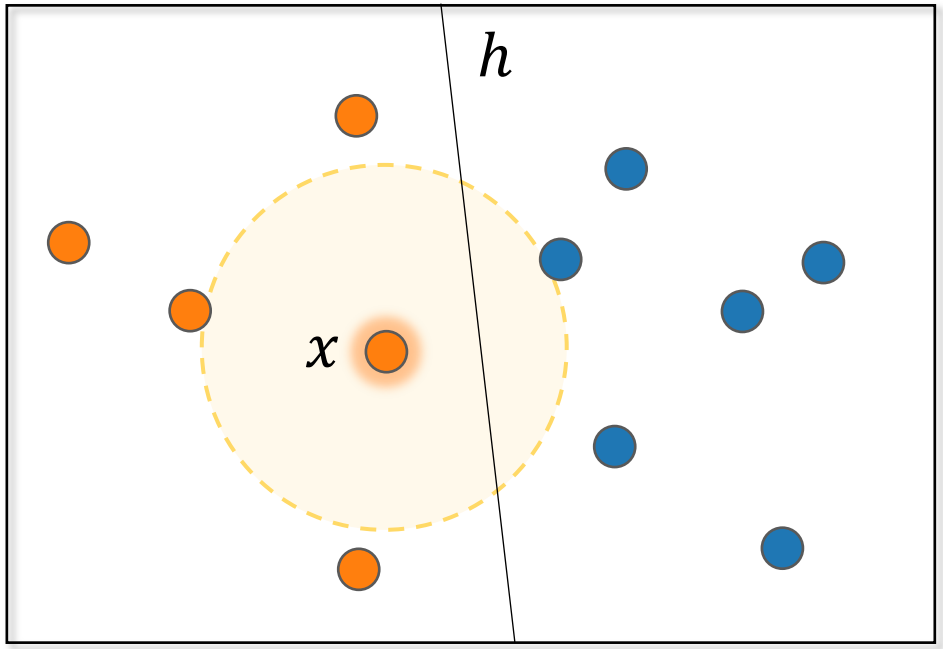
strategic

classification:

train: $\operatorname{argmin}_h \mathbb{E}[\mathbb{1}\{y \neq h(x)\}]$

test: $h(\Delta_h(x)) = \hat{y} \approx y$

response: $\Delta_h(x) = \operatorname{argmax}_{x'} h(x') - c(x, x')$



optimal non-strategic classifier

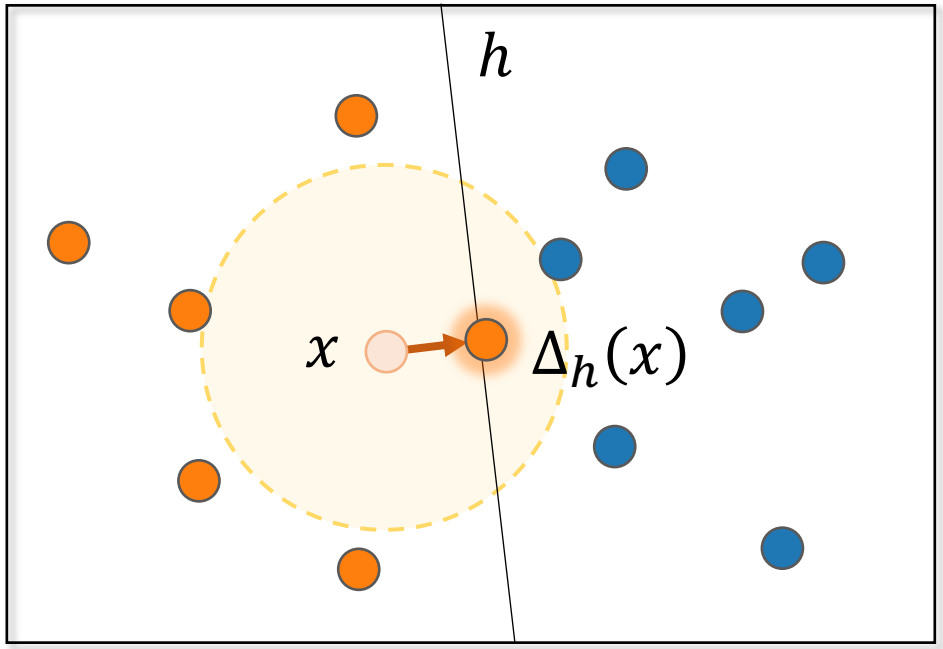
strategic

classification:

train: $\operatorname{argmin}_h \mathbb{E}[\mathbb{1}\{y \neq h(x)\}]$

test: $h(\Delta_h(x)) = \hat{y} \approx y$

response: $\Delta_h(x) = \operatorname{argmax}_{x'} h(x') - c(x, x')$



optimal non-strategic classifier

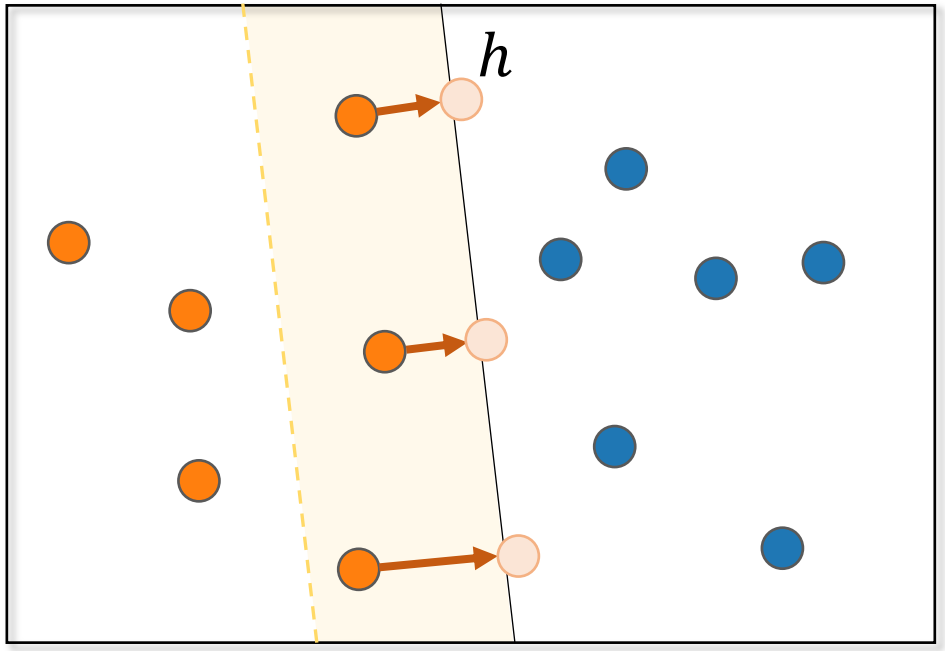
strategic

classification:

train: $\operatorname{argmin}_h \mathbb{E}[\mathbb{1}\{y \neq h(x)\}]$

test: $h(\Delta_h(x)) = \hat{y} \approx y$

response: $\Delta_h(x) = \operatorname{argmax}_{x'} h(x') - c(x, x')$



optimal non-strategic classifier

not optimal under strategic behavior

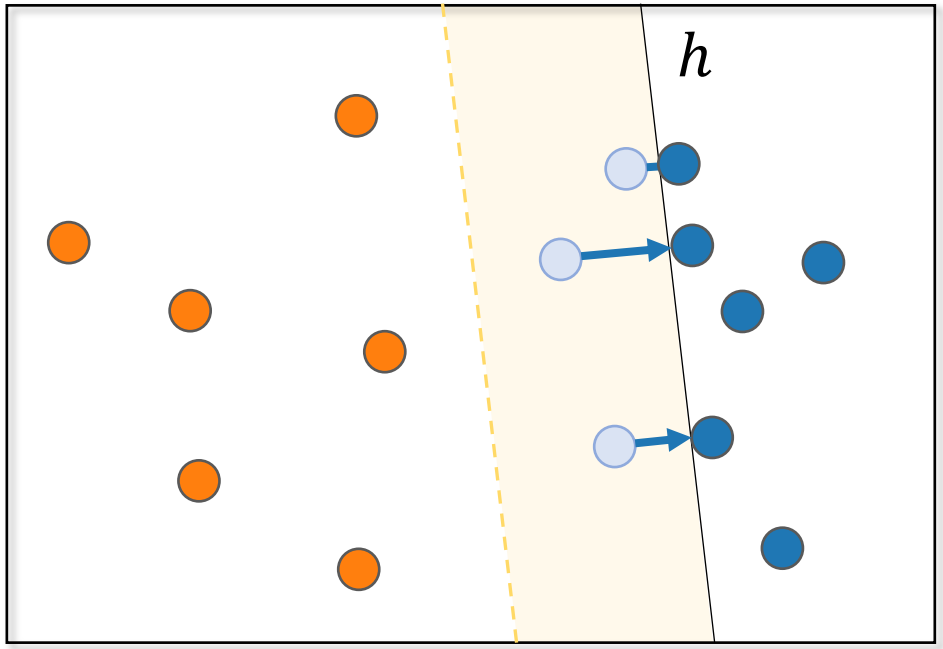
strategic

classification:

train: $\operatorname{argmin}_h \mathbb{E}[\mathbb{1}\{y \neq h(x)\}]$

test: $h(\Delta_h(x)) = \hat{y} \approx y$

response: $\Delta_h(x) = \operatorname{argmax}_{x'} h(x') - c(x, x')$



optimal classifier under strategic behavior

goal: learning that is robust to strategic "gaming" behavior

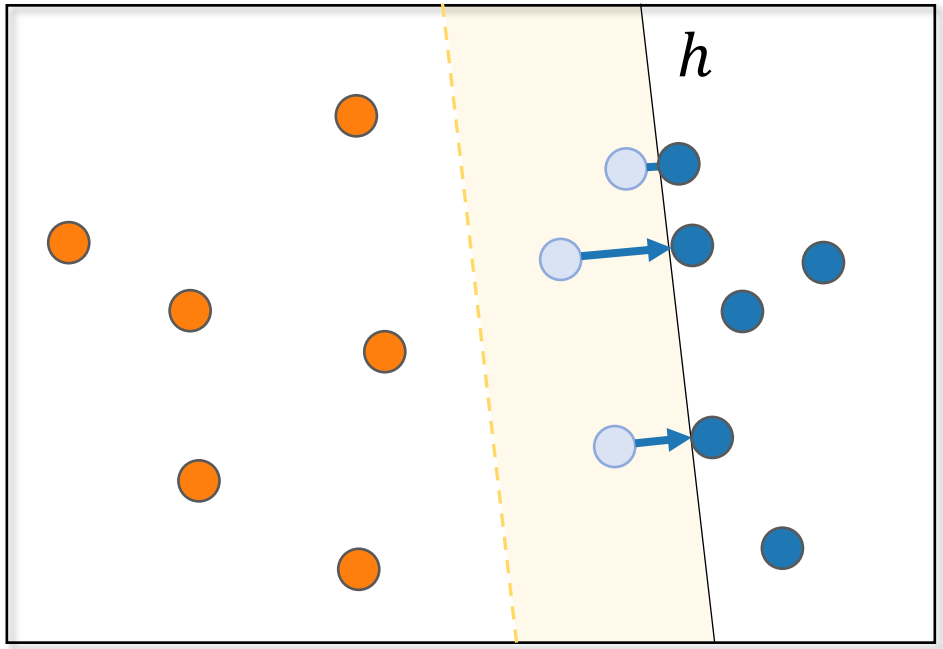
strategic

classification:

train: $\operatorname{argmin}_h \mathbb{E}[\mathbb{1}\{y \neq h(x)\}]$ **?**

test: $h(\Delta_h(x)) = \hat{y} \approx y$

response: $\Delta_h(x) = \operatorname{argmax}_{x'} h(x') - c(x, x')$



optimal classifier under strategic behavior

goal: learning that is robust to strategic "gaming" behavior

strategic

classification:

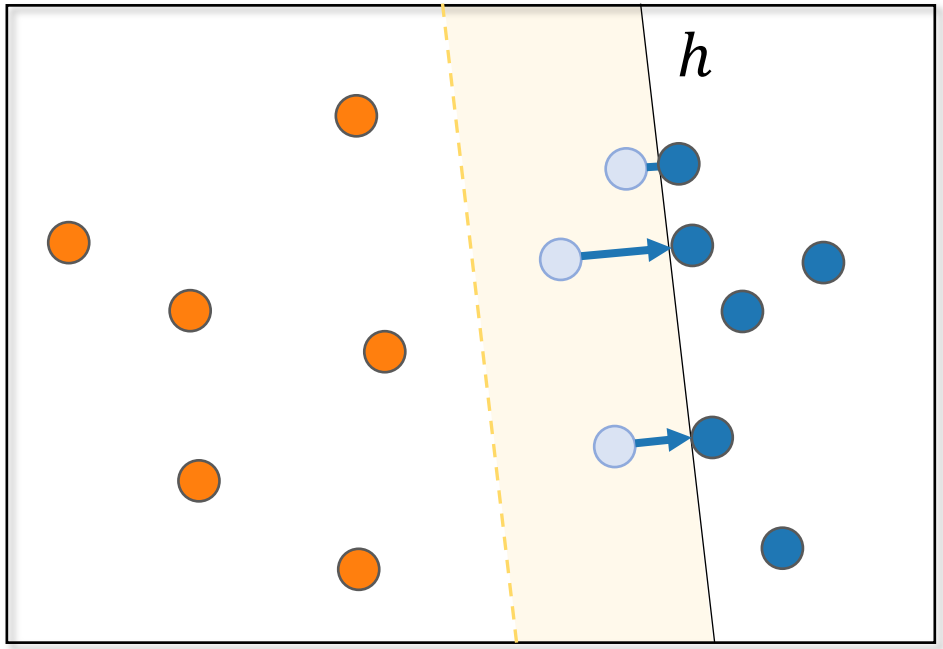
natural solution:

train: $\operatorname{argmin}_h \mathbb{E}[\mathbb{1}\{y \neq h(\Delta_h(x))\}]$

test: $h(\Delta_h(x)) = \hat{y}$

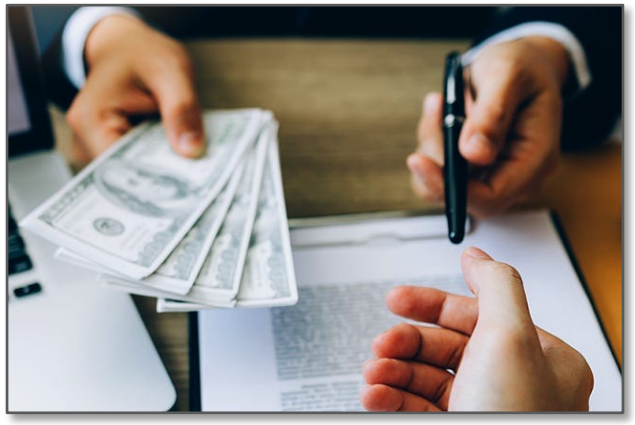
main takeaway:
not so simple!

response: $\Delta_h(x) = \operatorname{argmax}_{x'} h(x') - c(x, x')$



optimal classifier under strategic behavior

goal: learning that is robust to strategic "gaming" behavior

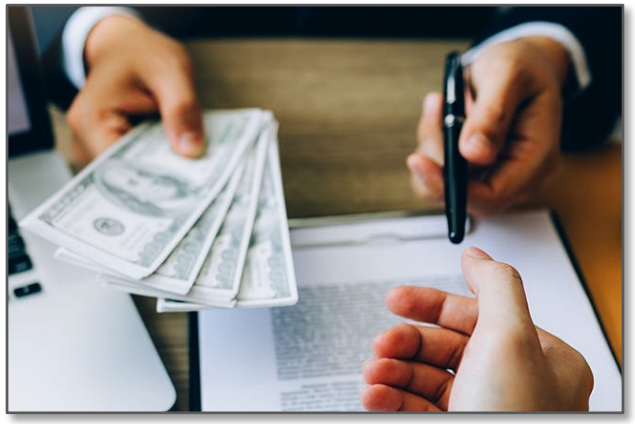


- **pro**: cleanly captures natural, prevalent tension
- **con**: narrow perspective, limited modeling power

?

?

?

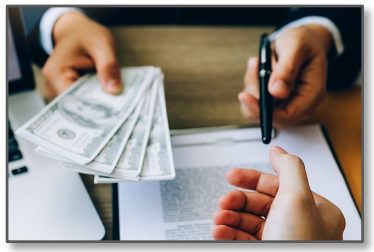


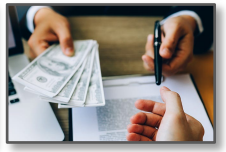

INSURANCE

?

?

?

?	?	?	?	?
?	?	?	?	?
?			 INSURANCE	?
?	?	?	?	?
?	?	?	?	?

?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?
?	?	?			 INSURANCE	?	?	?
?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?

generalized

strategic

classification:

response: $\Delta_h(x) = \underset{x'}{\operatorname{argmax}}\ h(x') - c(x, x')$

*users are **rational:*** *maximize* *utility* - *cost*

generalized

strategic

classification:

general utility

response: $\Delta_h(x) = \underset{x'}{\operatorname{argmax}}\ u_h(x') - c(x, x')$

Generaliing SC:

1. allow arbitrary utility functions u (that depend on h)

generalized

strategic

classification:

perceived utility

response: $\Delta_h(x) = \underset{x'}{\operatorname{argmax}} \tilde{u}_h(x') - c(x, x')$

$\neq u_h(x')$ ← *actual utility*

Generaliing SC:

1. allow arbitrary utility functions u (that depend on h)
2. let users act on perceived utility $\tilde{u}$ (can differ from *true* utility)

generalized

strategic

classification:

perceived utility *private information*

response: $\Delta_h(x, z) = \underset{x'}{\operatorname{argmax}} \tilde{u}_h(x', z) - c(x, x')$

$\neq u_h(x')$ ← *actual utility*

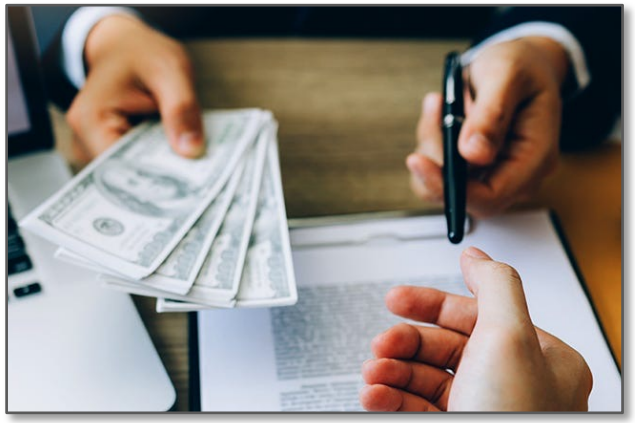
Generaliing SC:

1. allow arbitrary utility functions u (that depend on h)
2. let users act on perceived utility $\tilde{u}$ (can differ from *true* utility)
3. permit users to hold private information z (on which $\tilde{u}$ relies)

generalizes:
Hardt et al. (2016),
Sundaram et al. (2021),
Jagadeesan et al. (2021),
Ghalme et al. (2021), …

main questions: can we learn? when? how?

strategic = “gaming”

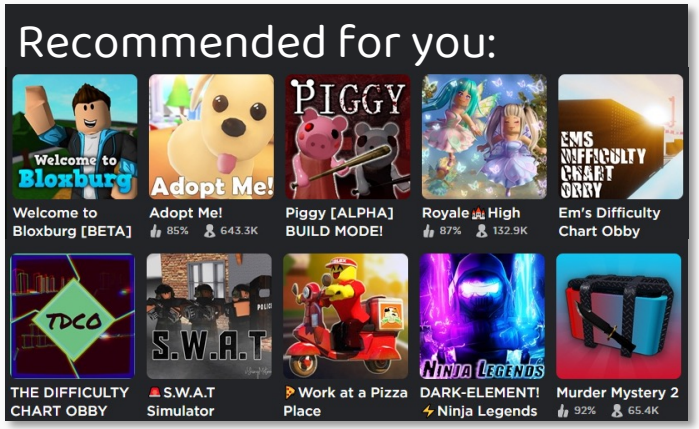


classification *about* humans

system wants: *correct* predictions

users want: *positive* predictions

incentives align



classification *for* humans

system wants: *correct* predictions

users want: *correct* predictions

- **Incentive Alignment (IA):** [definition]

$$\exists h \text{ s.t. } \mathbb{E}[\mathbb{1}\{y \neq h(\Delta_h(x, z))\}] < \min_{h'} \mathbb{E}[\mathbb{1}\{y \neq h'(x)\}]$$

- **Begin simple**: users act on noisy label beliefs:

$$z = \tilde{y} = y \text{ w.p. } 1 - \epsilon$$

$$\Rightarrow \Delta_h(x, \tilde{y}) = \operatorname*{argmax}_{x'} \overbrace{\mathbb{1}\{h(x') = \tilde{y}\}}^{\tilde{u}} - c(x, x')$$

Expect learning to imporve if:

- private (noisy) labels informative of true labels
- it can utilize users' willingness to invest effort

- **Incentive Alignment (IA):** [definition]

$$\exists h \text{ s.t. } \mathbb{E}[\mathbb{1}\{y \neq h(\Delta_h(x, z))\}] < \min_{h'} \mathbb{E}[\mathbb{1}\{y \neq h'(x)\}]$$

- **Begin simple**: users act on noisy label beliefs:

$$z = \tilde{y} = y \text{ w.p. } 1 - \epsilon$$

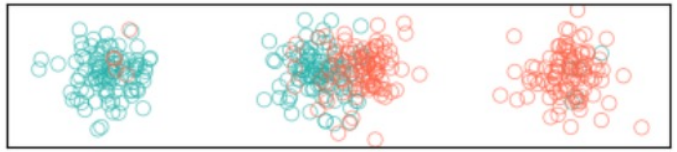
$$\Rightarrow \Delta_h(x, \tilde{y}) = \operatorname*{argmax}_{x'} \overbrace{\mathbb{1}\{h(x') = \tilde{y}\}}^{\tilde{u}} - c(x, x')$$

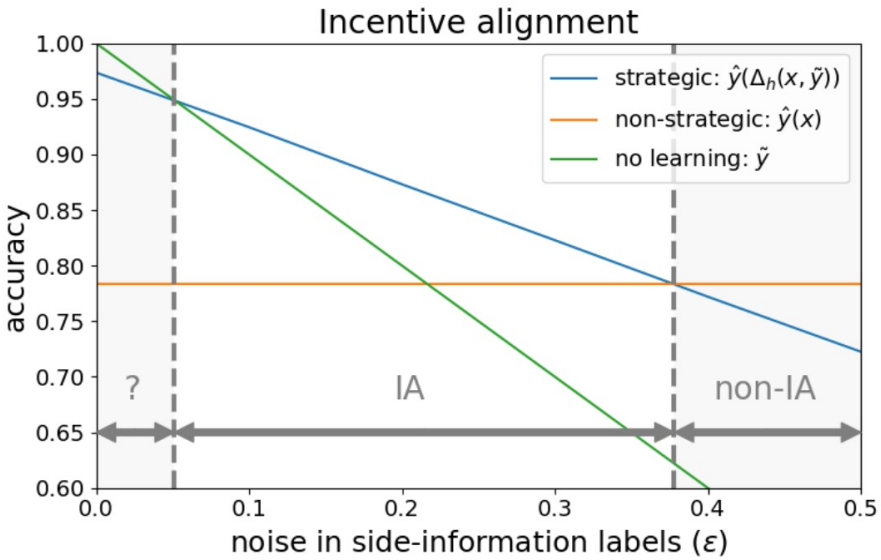
Expect learning to imporve if:

- private (noisy) labels informative of true labels
- it can utilize users' willingness to invest effort

- **Empirically** – learning aligns incentives for many ϵ:
- **Theoretically** – full characterization (see paper)

Data:





strategic behavior can be helpful!

learning:

problem: *train-test discrepancy*

train: $\underset{h}{\operatorname{argmin}}\, \mathbb{E}[\mathbb{1}\{y \neq h(x)\}]$

test: $h(\Delta_h(x, \tilde{y})) = \hat{y} \approx y$

response: $h(\Delta_h(x, \tilde{y})) = \underset{x'}{\operatorname{argmax}}\, \mathbb{1}\{h(x') = \tilde{y}\} - c(x, x')$

learning:

***natural solution**: anticipate user response*

train: $\underset{h}{\operatorname{argmin}}\, \mathbb{E}[\mathbb{1}\{y \neq h(\Delta_h(x, \tilde{y}))\}]$

test: $h(\Delta_h(x, \tilde{y})) = \hat{y} \approx y$

response: $h(\Delta_h(x, \tilde{y})) = \underset{x'}{\operatorname{argmax}}\, \mathbb{1}\{h(x') = \tilde{y}\} - c(x, x')$

learning:

0/1 loss:

train: $\underset{h}{\operatorname{argmin}}\, \mathbb{E}[\mathbb{1}\{y \neq h(\Delta_h(x, \tilde{y}))\}]$ ✓

test: $h(\Delta_h(x, \tilde{y})) = \hat{y} \approx y$

response: $h(\Delta_h(x, \tilde{y})) = \underset{x'}{\operatorname{argmax}}\, \mathbb{1}\{h(x') = \tilde{y}\} - c(x, x')$

learning:

hinge loss:

train: $\underset{h}{\operatorname{argmin}}\ \mathbb{E}[\max\{0,1 - yw^{\top}\Delta_h(x,\tilde{y})\}]$ **?**

test: $h(\Delta_h(x,\tilde{y})) = \hat{y} \approx y$

response: $h(\Delta_h(x,\tilde{y})) = \underset{x'}{\operatorname{argmax}}\ \mathbb{1}\{h(x') = \tilde{y}\} - c(x,x')$

learning:

hinge loss:

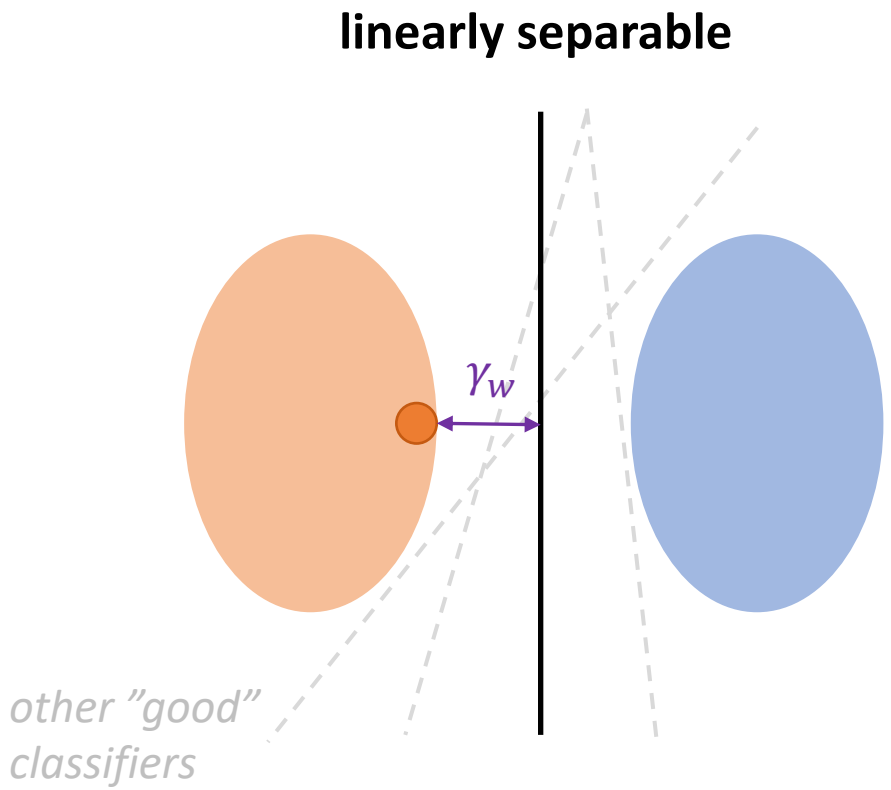
train: $\underset{h}{\operatorname{argmin}}\ \mathbb{E}[\max\{0,1 - yw^{\top}\Delta_h(x,\tilde{y})\}]$ 

test: $h(\Delta_h(x,\tilde{y})) = \hat{y} \approx y$

response: $h(\Delta_h(x,\tilde{y})) = \underset{x'}{\operatorname{argmax}}\ \mathbb{1}\{h(x') = \tilde{y}\} - c(x,x')$

standard hinge loss:

$$\max\{0,1 - yw^\top x\}$$



standard hinge loss:

$$\max\{0,1 - yw^\top x\}$$

not linearly separable

max-margin classifier

- selection criterion

- good generalization

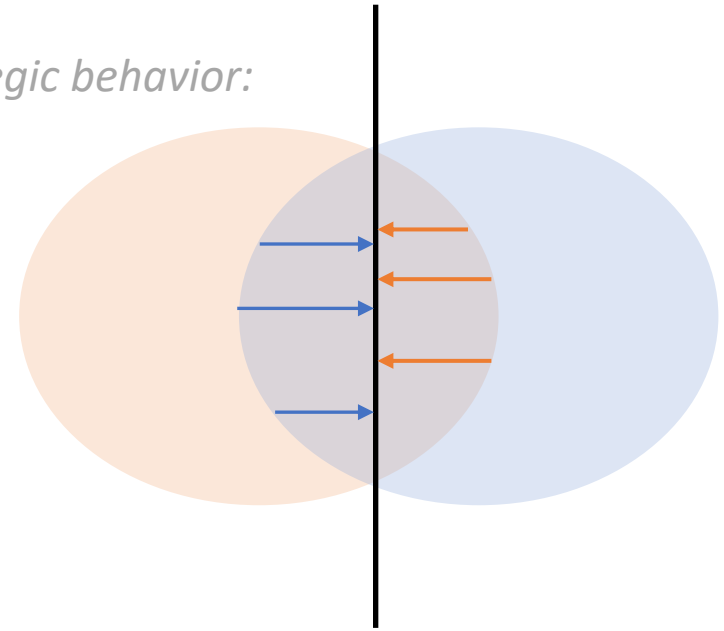
naïve hinge loss:

$$\max\{0,1 - yw^\top \Delta_h(x,y)\}$$

easiest to imagine!

not linearly separable

strategic behavior:

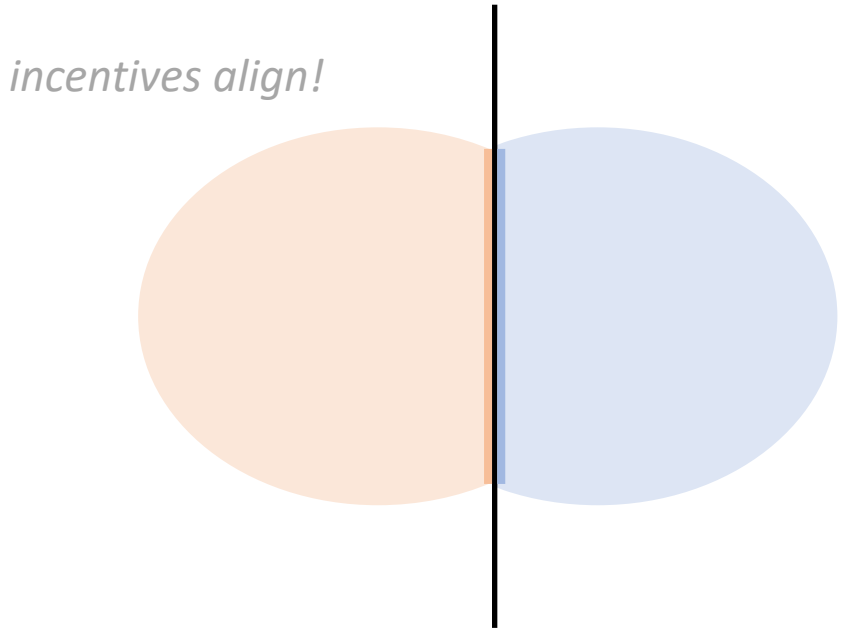


(naïve) **max-margin classifier**

naïve hinge loss:

$$\max\{0,1 - yw^{\top}\Delta_h(x,y)\}$$

strategically linearly separable



(naïve) **max-margin classifier**

perfect accuracy!

naïve hinge loss:

$$\max\{0,1 - yw^\top \Delta_h(x,y)\}$$

strategically linearly separable

(naïve) **max-margin = 0**

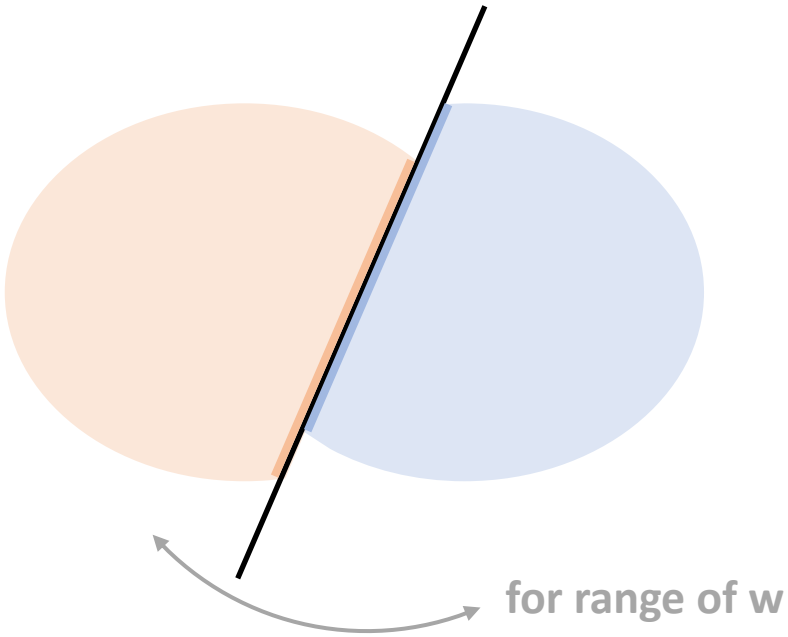
perfect accuracy?

naïve hinge loss:

$$\max\{0,1 - yw^\top \Delta_h(x,y)\}$$

conclusion: adapting hinge to strategic settings requires **rethinking the concept of margin**

strategically linearly separable



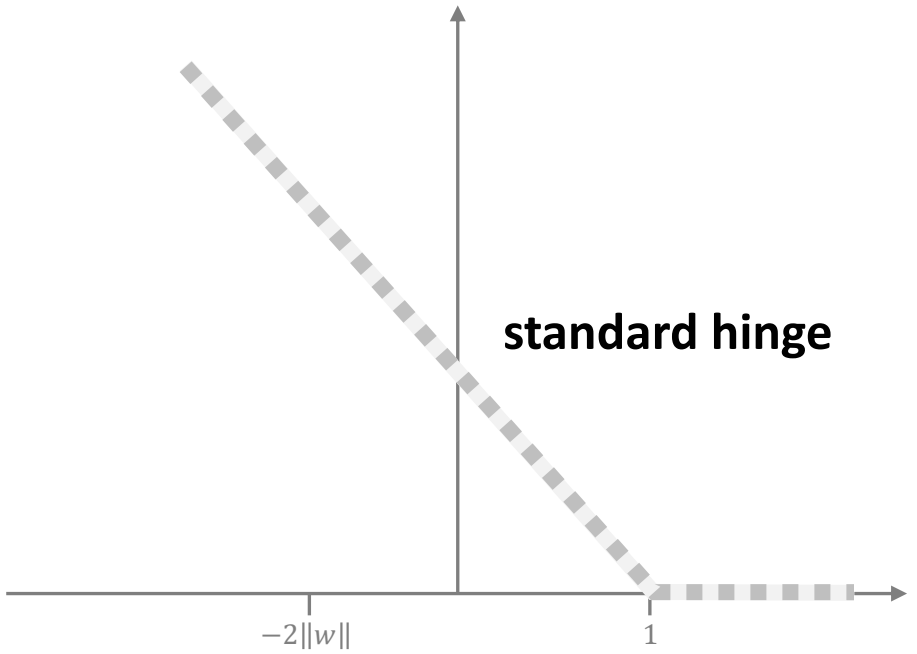
(naïve) **max-margin = 0**

⇒ **vocous selection criterion**

⇒ **won't help generalization!**

- **standard hinge:**

$$\max\{0, 1 - yw^\top x\}$$



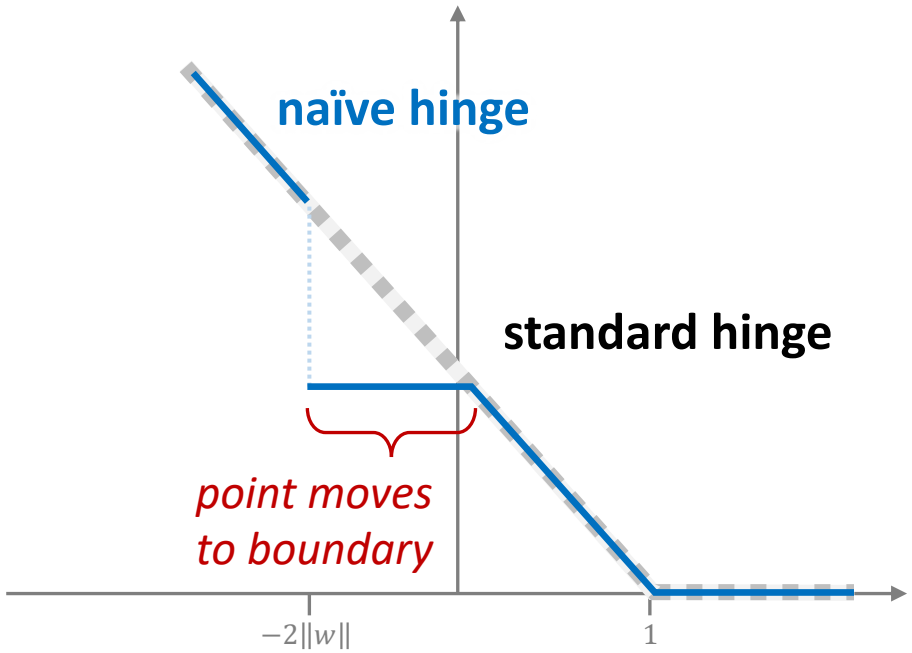
- **standard hinge:**

$$\max\{0, 1 - yw^\top x\}$$

- **naïve hinge:**

$$\max\{0, 1 - yw^\top \Delta_h(x, \tilde{y})\}$$

suspicious: does not depend on x!



- **standard hinge:**

$$\max\{0, 1 - yw^\top x\}$$

- **naïve hinge:**

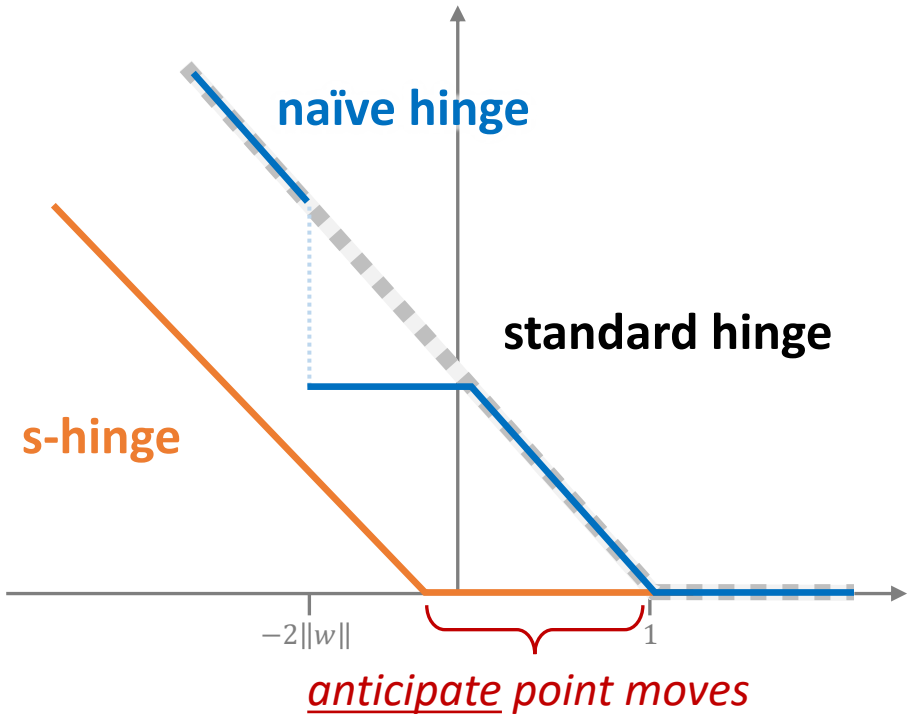
$$\max\{0, 1 - yw^\top \Delta_h(x, \tilde{y})\}$$

- **our strategic hinge:** (s-hinge)

$$\max\{0, 1 - yw^\top x - 2y\tilde{y}\|w\|\}$$

prediction before movement…

… but relative to "shifted", personalized decision boundary



- **standard hinge:**

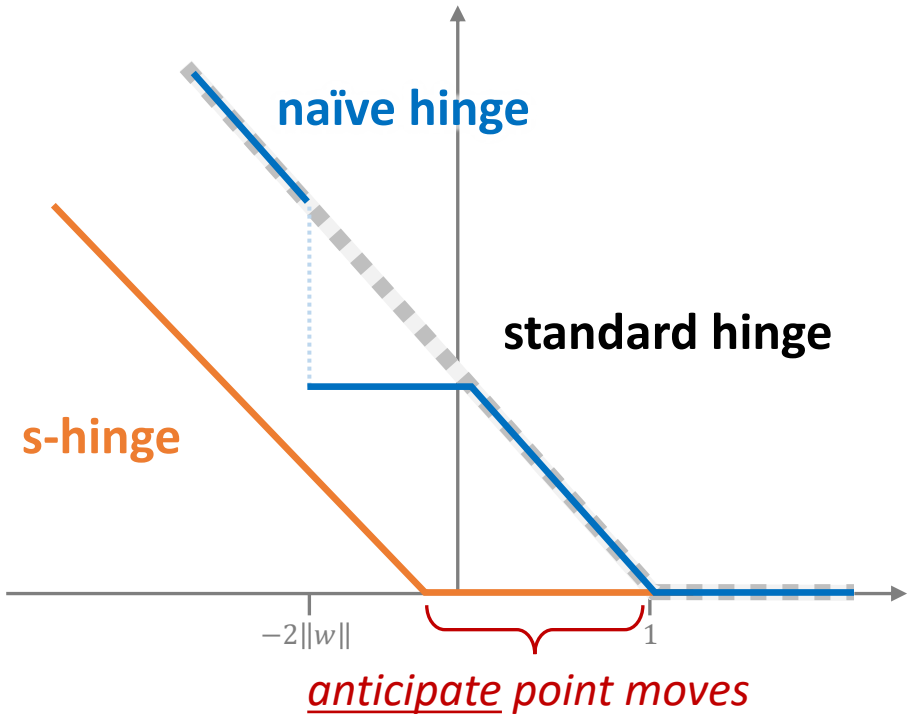
$$\max\{0, 1 - yw^\top x\}$$

- **naïve hinge:**

$$\max\{0, 1 - yw^\top \Delta_h(x, \tilde{y})\}$$

- **our strategic hinge:** (s-hinge)

$$\max\{0, 1 - yw^\top x - 2y\tilde{y}\|w\|\}$$



- **Benefits:**
  1. Optimization – differentiable; no nasty Δ!
  2. Generalization guarantees (up next)
  3. Extends to broader strategic settings

general-preference strategic classification:

noisy labels:

$$\Delta_h(x;\tilde{y}) = \operatorname*{argmax}_{x'} \mathbb{1}\{h(x') = \tilde{y}\} - c(x,x')$$

standard SC: $\mathbb{1}\{h(x') = 1\}$

"adversarial": $\mathbb{1}\{h(x') \neq y\}$

$\vdots$

general preferences: $\mathbb{1}\{h(x') = y'\}, \quad y' \in \{\pm 1\}$

also Sundaram et al. (2021)

general-preference strategic classification:

general preferences:

$$\Delta_h(x; y') = \underset{x'}{\operatorname{argmax}}\, \mathbb{1}\{h(x') = y'\} - c(x, x')$$

1) s-hinge extends to GP:

$$\max\{0, 1 - yw^\top x - 2yy'\|w\|\}$$

differentiable – no Δ!

2) generalization bound:

$$\mathcal{L}_{0/1} \le \mathcal{L}_{0/1}^{\mathrm{NL}} + \frac{4r\|\hat{w}\|}{\sqrt{m}} + \left(1 + 2\rho^{\mathrm{GP}}\|w\|\right)\sqrt{\frac{2\ln(4\|\hat{w}\|/\delta)}{m}}$$

- extends and closely matches non-strategic case
- relations: $\rho^{\mathrm{NL}}_{\epsilon<0.5} \le \rho \le \rho^{\mathrm{NL}}_{\epsilon>0.5} \le \rho^{\mathrm{SC}} = \rho^{\mathrm{ADV}} = \rho^{\mathrm{GP}}$

general-preference strategic classification:

general preferences:

$$\Delta_h(x; y') = \underset{x'}{\operatorname{argmax}}\, \mathbb{1}\{h(x') = y'\} - c(x, x')$$

1) s-hinge extends to GP:

$$\max\{0, 1 - yw^\top x - 2yy'\|w\|\}$$

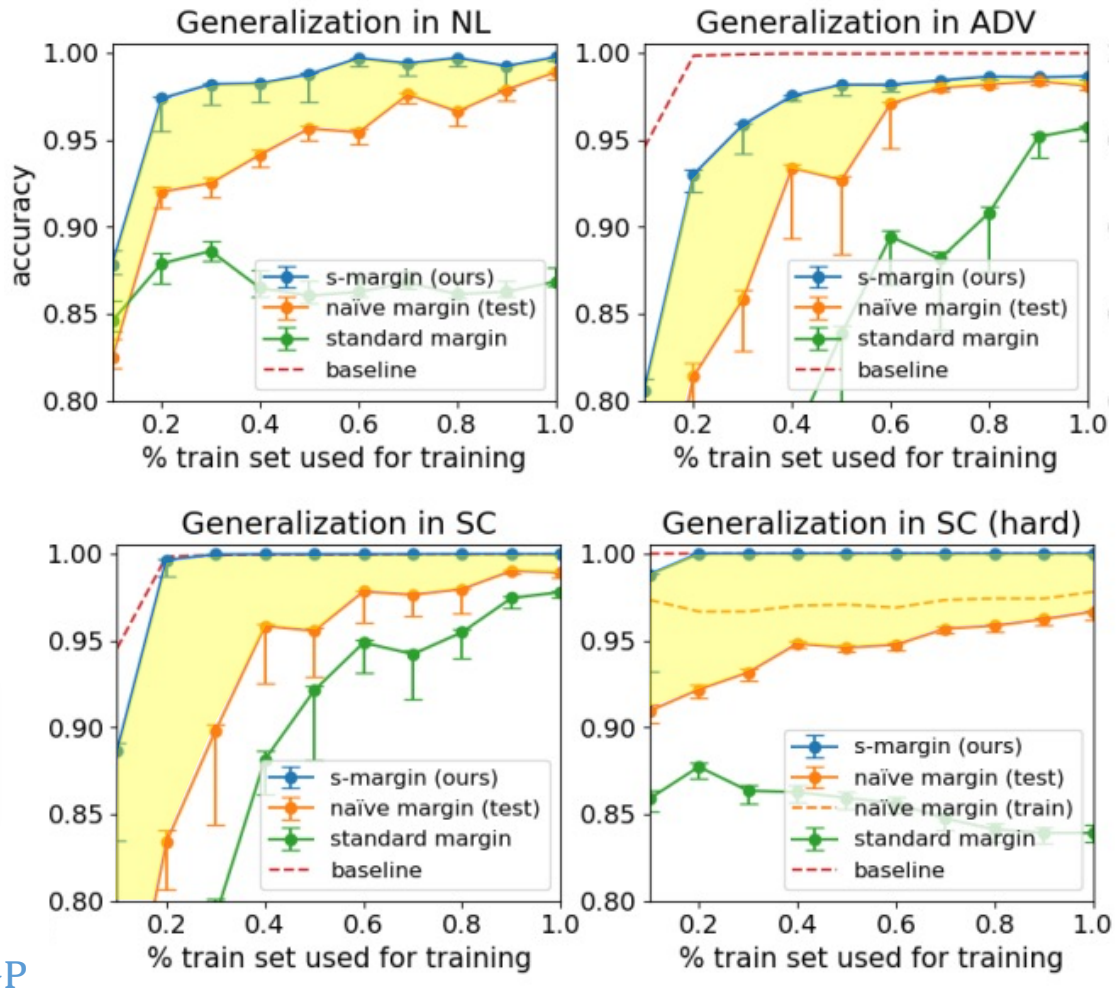
differentiable – no Δ!

2) generalization bound:

$$\mathcal{L}_{0/1} \le \mathcal{L}_{0/1}^{\mathrm{NL}} + \frac{4r\|\hat{w}\|}{\sqrt{m}} + \left(1 + 2\rho^{\mathrm{GP}}\|w\|\right)\sqrt{\frac{2\ln(4\|\hat{w}\|/\delta)}{m}}$$

- extends and closely matches non-strategic case
- relations: $\rho^{\mathrm{NL}}_{\epsilon<0.5} \le \rho \le \rho^{\mathrm{NL}}_{\epsilon>0.5} \le \rho^{\mathrm{SC}} = \rho^{\mathrm{ADV}} = \rho^{\mathrm{GP}}$

3) empirical generalization gaps:



generalized strategic classification:

$$\Delta_h(x; z) = \underset{x'}{\operatorname{argmax}}\ \tilde{u}_h(x', z) - c(x, x')$$

perceived utility — *side information*

- **standard hinge:**

$$\max\{0, 1 - yw^\top x\}$$

$$= \max\{0, 1 - \operatorname{sign}(yw^\top x)|w^\top x|\}$$

correctness *distance*

generalized strategic classification:

$$\Delta_h(x; z) = \underset{x'}{\operatorname{argmax}} \; \tilde{u}_h(x', z) - c(x, x')$$

perceived utility — *side information*

- **standard hinge:**

$$\max\{0, 1 - yw^\top x\}$$

$$= \max\{0, 1 - \operatorname{sign}(yw^\top x)|w^\top x|\}$$

correctness *distance*

- **naïve hinge:**

$$\max\{0, 1 - \operatorname{sign}(yw^\top \Delta_h(x, z))|w^\top \Delta_h(x, z)|\}$$

- no original x
- distance <u>after</u> move
- distance to hyperplane

- **generalized strategic hinge:** (gs-hinge)

$$\max\{0, 1 - \operatorname{sign}(yw^\top \Delta_h(x, z)) d_\Delta(x, z; w)\}$$

<u>anticipated</u> distance

generalized strategic classification:

$$\Delta_h(x;z) = \underset{x'}{\operatorname{argmax}}\ \tilde{u}_h(x',z) - c(x,x')$$

- **standard hinge:**

$$\max\{0, 1 - yw^\top x\}$$

$$= \max\{0, 1 - \operatorname{sign}(yw^\top x)|w^\top x|\}$$

- **naïve hinge:**

$$\max\{0, 1 - \operatorname{sign}(yw^\top \Delta_h(x,z))|w^\top \Delta_h(x,z)|\}$$

- **generalized strategic hinge:** (gs-hinge)

$$\max\{0, 1 - \operatorname{sign}(yw^\top \Delta_h(x,z))\, d_\Delta(x,z;w)\}$$

reinterpretation of "margin":

distance to nearest x' who's movement *flips label*:
subsumes non-strategic case

$$d_\Delta(x,z;w) = \min_{x'} \frac{\|x - x'\|}{\|w\|}$$

minimal distance between points (normalized)

flip label (after movement) s.t. $h\big(\Delta_h(x,z)\big) \neq h(\Delta_h(x',z))$

- *subsumes standard hinge for non-strat.*
- *for GP, get s-hinge as special case*
- *for general GSC, may be hard to compute*

extended generalization bound:

$$\mathcal{L}_{0/1} \leq \mathcal{L}_{0/1}^{\mathrm{NL}} + \frac{8r\|\widehat{w}\|}{\sqrt{m}} + \left(1 + 2\rho^{\mathrm{GSC}}\|w\|\right)\sqrt{\frac{2\ln(4\|\widehat{w}\|/\delta)}{m}}$$

strategic content reccomendation:

perceived utility

$$\Delta_h(x;z) = \underset{x'}{\operatorname{argmax}}\ \tilde{u}_h(x',z) - c(x,x')$$

personal history $= \{(a_j, y_j)\}_{j=1}^{n}$

modify features to have previous experiences classified correctly

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{h(x', a_j) = y_j\}$$

want correct predictions on future items:

$$u_h(x) = \mathbb{1}\{h(x,a) = y\}$$

true utility — *like/dislike* — *items (e.g., movies)*

act in hope of improving (future) accuracy:

$$u_h\big(\Delta_h(x,z)\big) = \mathbb{1}\{h(\Delta_h(x,z), a) = y\}$$

- **Challenge**: in essence, users *also* aim to minmize 0-1 loss
- Get *coupling* of learning problems (system+users) – implicit cooperation!

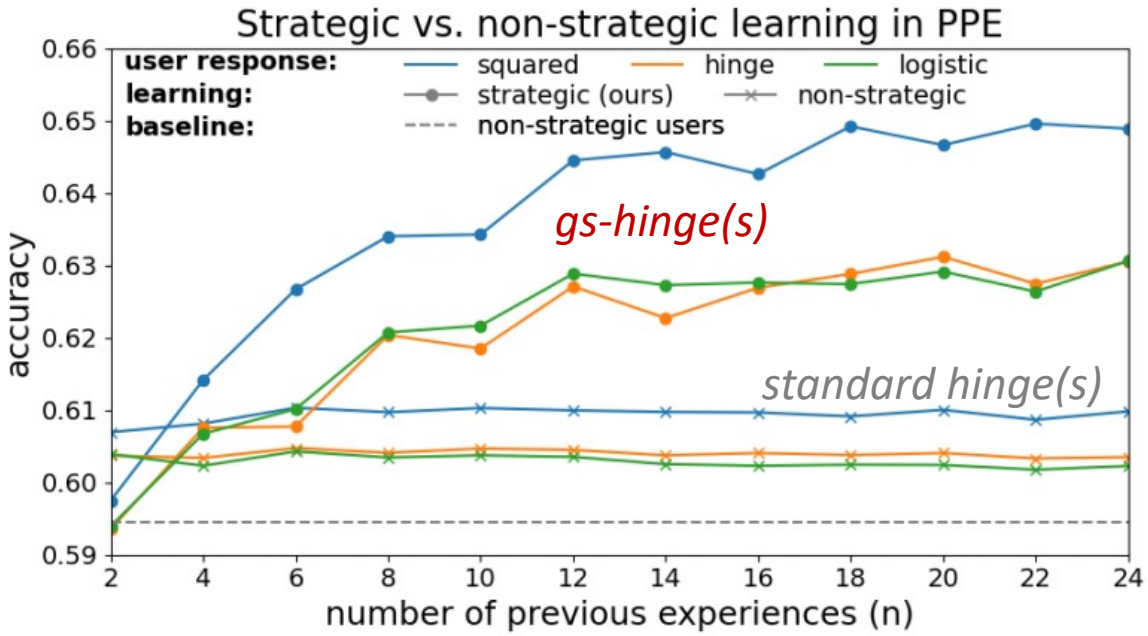
strategic content reccomendation:

$$\Delta_h(x;z) = \underset{x'}{\operatorname{argmax}}\ \tilde{u}_h(x', z) - c(x, x')$$

personal history $= \{(a_j, y_j)\}_{j=1}^{n}$

modify features to have previous experiences classified correctly

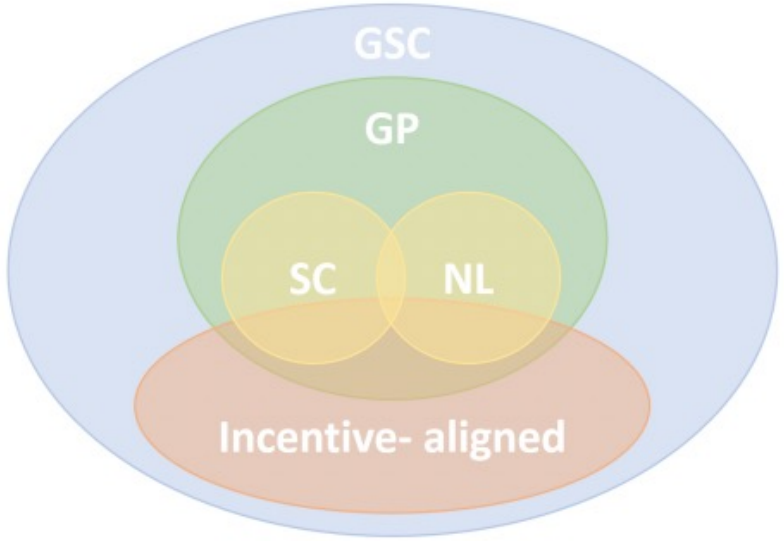
$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{h(x', a_j) = y_j\}$$



- **Challenge**: in essence, users *also* aim to minmize 0-1 loss
- Get *coupling* of learning problems (system+users) – implicit cooperation!
- We propose differentiable proxy
- **Results**: highly incentive-aligned task!

Conclusions

1. Strategic behavior comes in **many flavors**
   - Heirarchy of problem classes
   - Mild variations in user response
     ⇒ very different learning problems
   - We highlight incentive alignment
   - Others interesting classes?
     (e.g., bounded-rational, behavioral, Bayesian)



2. Strategic behavior can **break** proxy losses
   - Choose with care!
   - We propose **gs-hinge**:
     differentible (sometimes) + theoretical guarantees + empirical results
   - Other losses? (e.g., strategic cross entropy)

Thanks!

(come to our poster)

check out our other paper on

strategic representation

(also @ICML2022)