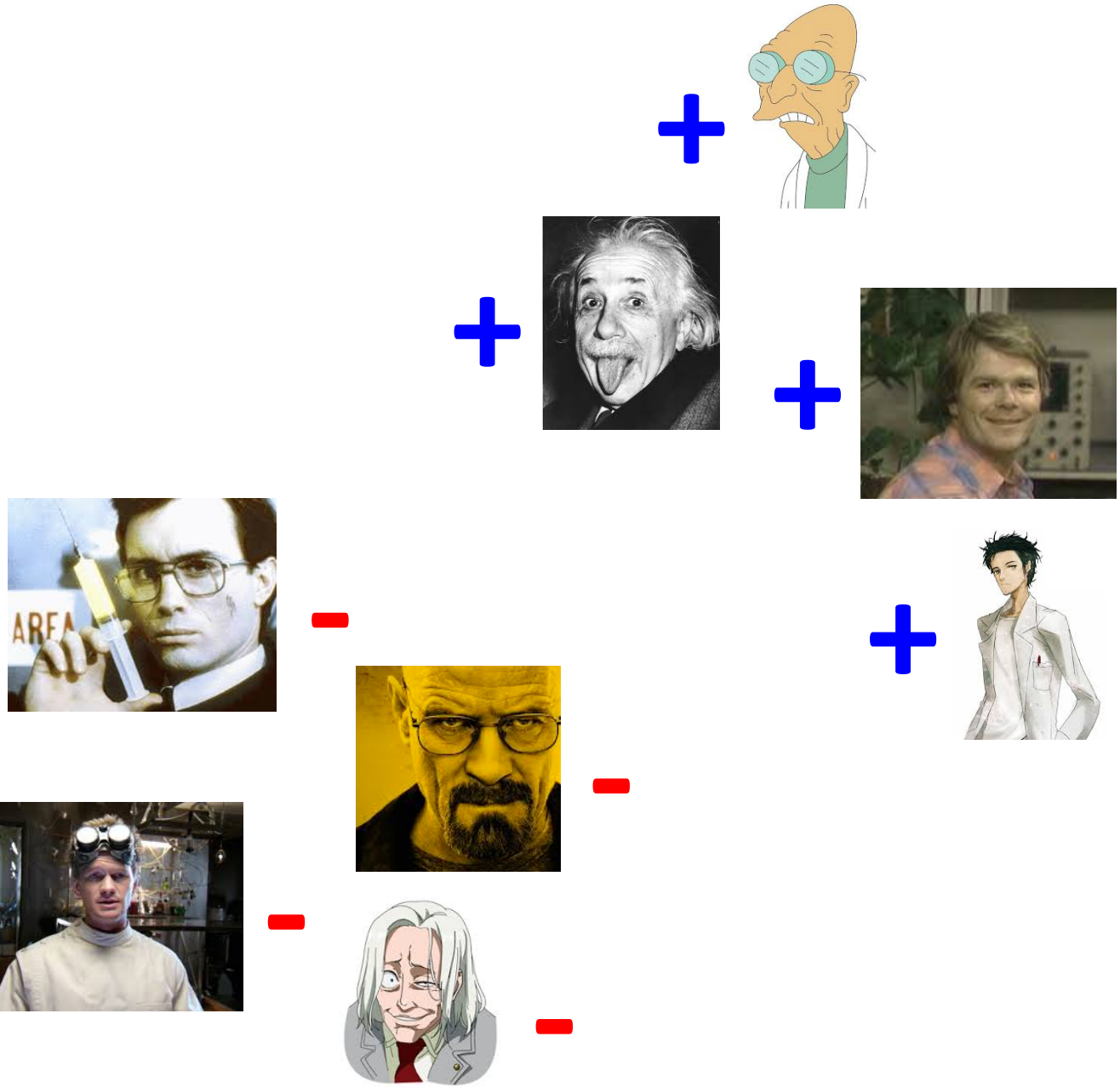
Learning from Corrupted Binary Labels via Class-Probability Estimation

Aditya Krishna Menon Brendan van Rooyen
Cheng Soon Ong Robert C. Williamson

National ICT Australia and The Australian National University

Australian National University

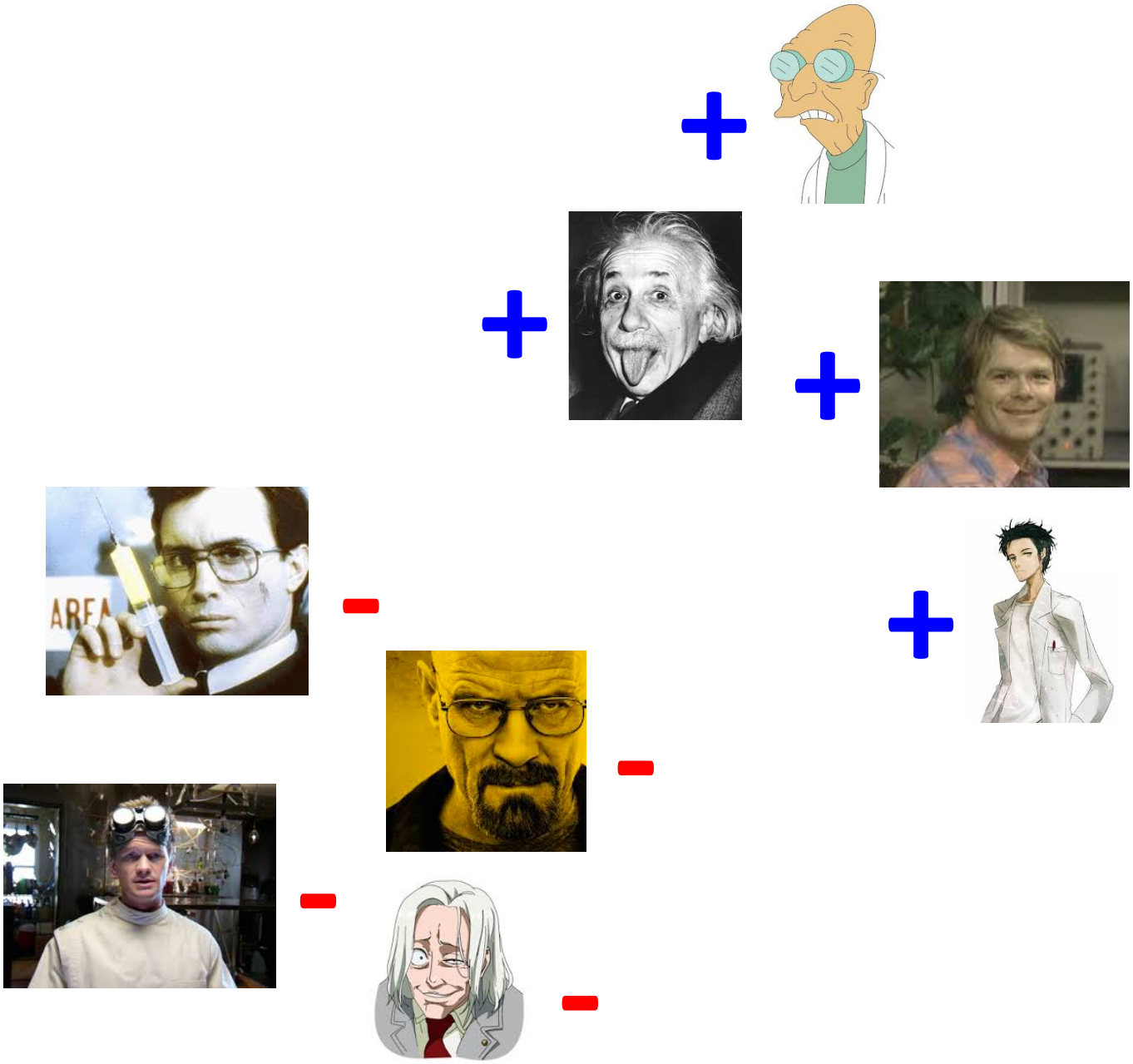
Learning from binary labels



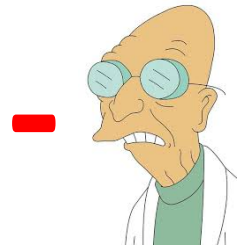
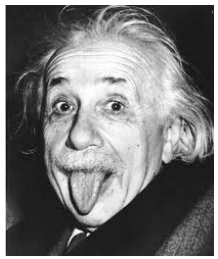
Learning from binary labels

?

Learning from binary labels



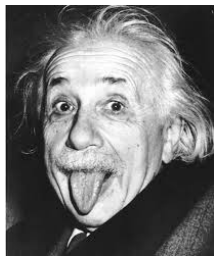
Learning from noisy labels



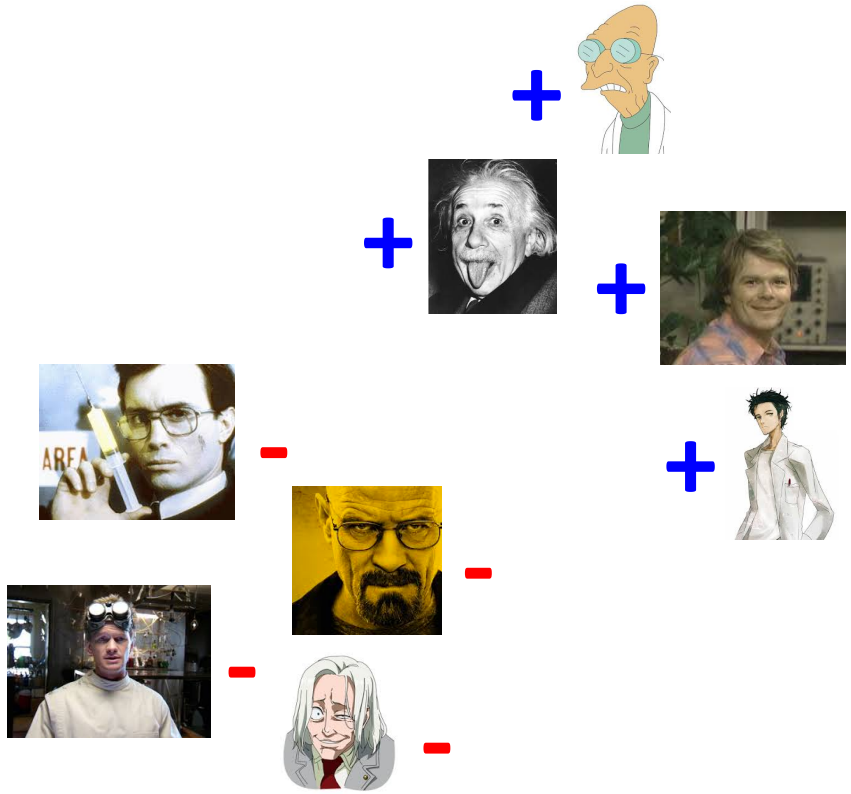


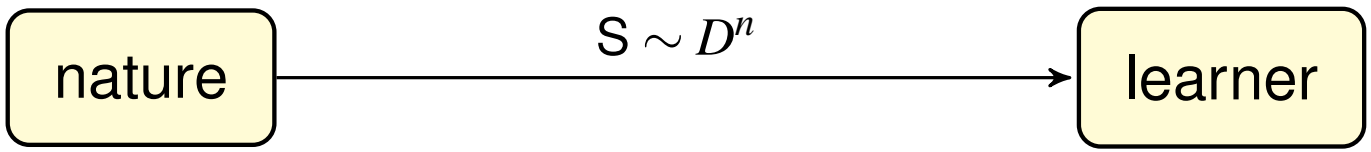
Learning from positive and unlabelled data



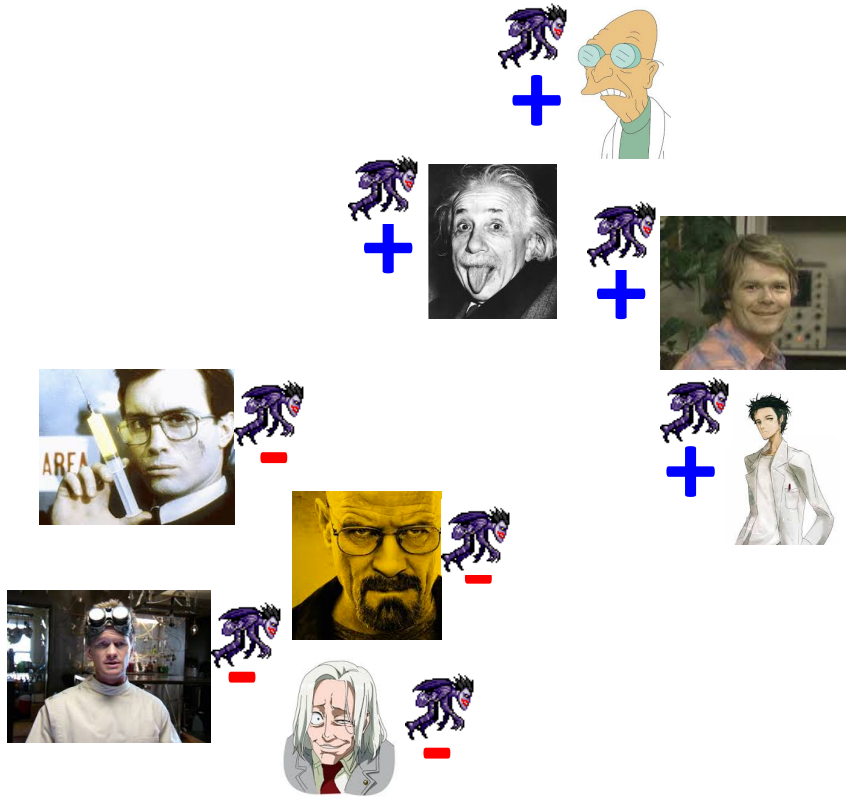

Learning from binary labels





Goal: good classification wrt distribution D

Learning from corrupted labels



Goal: good classification wrt (unobserved) distribution D

Paper summary

Can we learn a good classifier from corrupted samples?

Paper summary

Can we learn a good classifier from corrupted samples?

Prior work: in special cases (with a rich enough model), yes!

Paper summary

Can we learn a good classifier from corrupted samples?

Prior work: in special cases (with a rich enough model), yes!

- can treat samples as if uncorrupted!
- (Elkan and Noto, 2008), (Zhang and Lee, 2008), (Natarajan et al., 2013), (duPlessis and Sugiyama, 2014) ...

Paper summary

Can we learn a good classifier from corrupted samples?

Prior work: in special cases (with a rich enough model), yes!

- can treat samples as if uncorrupted!
- (Elkan and Noto, 2008), (Zhang and Lee, 2008), (Natarajan et al., 2013), (duPlessis and Sugiyama, 2014) ...

This work: unified treatment via class-probability estimation

- analysis for general class of corruptions

Assumed corruption model

Learning from binary labels: distributions

Fix instance space $\mathcal{X}$ (e.g. $\mathbb{R}^N$)

Underlying distribution D over $\mathcal{X} \times \{\pm 1\}$

Constituent components of D:

$$(P(x), Q(x), \pi) = (\mathbb{P}[\mathsf{X} = x | \mathsf{Y} = 1], \mathbb{P}[\mathsf{X} = x | \mathsf{Y} = -1], \mathbb{P}[\mathsf{Y} = 1])$$

Learning from binary labels: distributions

Fix instance space $\mathcal{X}$ (e.g. $\mathbb{R}^N$)

Underlying distribution D over $\mathcal{X} \times \{\pm 1\}$

Constituent components of D:

$$(P(x), Q(x), \pi) = (\mathbb{P}[\mathsf{X} = x | \mathsf{Y} = 1], \mathbb{P}[\mathsf{X} = x | \mathsf{Y} = -1], \mathbb{P}[\mathsf{Y} = 1])$$
$$(M(x), \eta(x)) = (\mathbb{P}[\mathsf{X} = x], \mathbb{P}[\mathsf{Y} = 1 | \mathsf{X} = x])$$

Learning from corrupted binary labels

Samples from corrupted distribution $\bar{D} = (\bar{P}, \bar{Q}, \bar{\pi})$

Goal: good classification wrt (unobserved) distribution D

Learning from corrupted binary labels

Samples from corrupted distribution $\bar{D} = (\bar{P}, \bar{Q}, \bar{\pi})$, where

$$\bar{P} = (1 - \alpha) \cdot P + \alpha \cdot Q$$

$$\bar{Q} = \beta \cdot P + (1 - \beta) \cdot Q$$

and $\bar{\pi}$ is arbitrary

- α, β are noise rates
- mutually contaminated distributions (Scott et al., 2013)

Goal: good classification wrt (unobserved) distribution D

Special cases

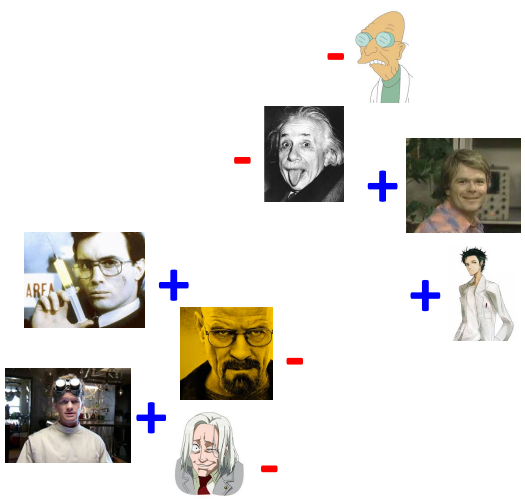
Label noise

Labels flipped w.p. ρ

$$\bar{\pi} = (1 - 2\rho) \cdot \pi + \rho$$

$$\alpha = \bar{\pi}^{-1} \cdot (1 - \pi) \cdot \rho$$

$$\beta = (1 - \bar{\pi})^{-1} \cdot \pi \cdot \rho$$

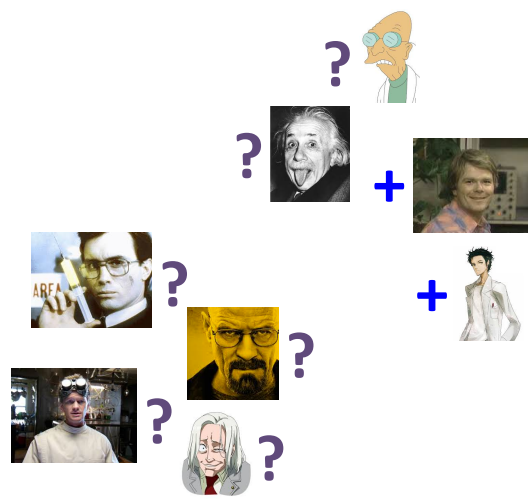


PU learning

Observe M instead of Q

$$\bar{\pi} = \text{ arbitrary}$$

$$\bar{P} = 1 \cdot P + 0 \cdot Q$$
$$\bar{Q} = M$$
$$= \pi \cdot P + (1 - \pi) \cdot Q$$



Corrupted class-probabilities

Structure of corrupted class-probabilities underpins analysis

Corrupted class-probabilities

Structure of corrupted class-probabilities underpins analysis

Proposition

For any $D, \bar{D}$,

$$\bar{\eta}(x) = \phi_{\alpha,\beta,\pi}(\eta(x))$$

where $\phi_{\alpha,\beta,\pi}$ *is strictly monotone for fixed* α, β, π.

Corrupted class-probabilities

Structure of corrupted class-probabilities underpins analysis

Proposition

For any $D, \bar{D}$,

$$\bar{\eta}(x) = \phi_{\alpha,\beta,\pi}(\eta(x))$$

where $\phi_{\alpha,\beta,\pi}$ is strictly monotone for fixed α, β, π.

Follows from Bayes' rule:

$$\frac{\bar{\eta}(x)}{1-\bar{\eta}(x)} = \frac{\bar{\pi}}{1-\bar{\pi}} \cdot \frac{\bar{P}(x)}{\bar{Q}(x)}$$

Corrupted class-probabilities

Structure of corrupted class-probabilities underpins analysis

Proposition

For any $D, \bar{D}$*,*

$$\bar{\eta}(x) = \phi_{\alpha,\beta,\pi}(\eta(x))$$

where $\phi_{\alpha,\beta,\pi}$ *is strictly monotone for fixed* α, β, π*.*

Follows from Bayes' rule:

$$\frac{\bar{\eta}(x)}{1-\bar{\eta}(x)} = \frac{\bar{\pi}}{1-\bar{\pi}} \cdot \frac{\bar{P}(x)}{\bar{Q}(x)} = \frac{\bar{\pi}}{1-\bar{\pi}} \cdot \frac{(1-\alpha) \cdot \frac{P(x)}{Q(x)} + \alpha}{\beta \cdot \frac{P(x)}{Q(x)} + (1-\beta)}.$$

Corrupted class-probabilities: special cases

Label noise

$$\bar{\eta}(x) = (1 - 2\rho) \cdot \eta(x) + \rho$$

ρ unknown

(Natarajan et al., 2013)

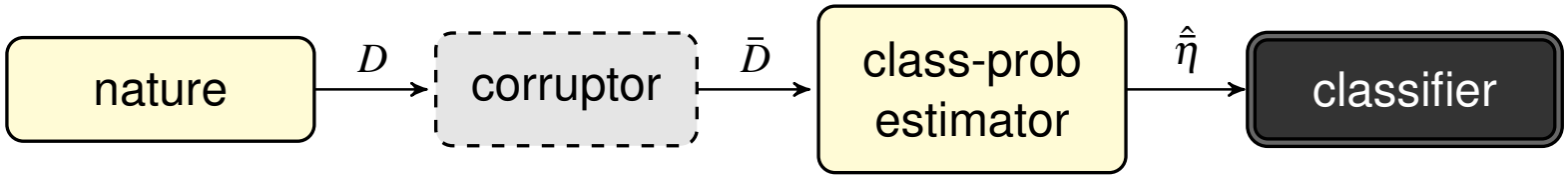
PU learning

$$\bar{\eta}(x) = \frac{\pi \cdot \eta(x)}{\pi \cdot \eta(x) + (1 - \pi) \cdot \bar{\pi}}$$

π unknown

(Ward et al., 2009)

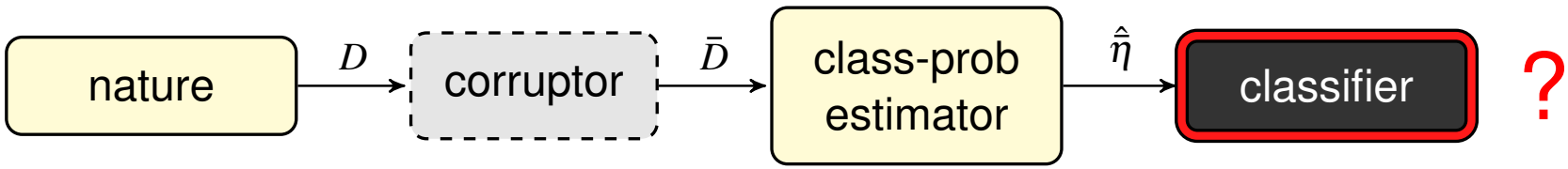
Roadmap



Kernel logistic regression

Roadmap

Exploit monotone relationship between η and $\bar{\eta}$



Kernel logistic regression

Classification with noise rates

Class-probabilities and classification

Many classification measures optimised by $\operatorname{sign}(\eta(x) - t)$

- 0-1 error $\rightarrow t = \frac{1}{2}$
- Balanced error $\rightarrow t = \pi$
- F-score $\rightarrow$ optimal t depends on D
  - (Lipton et al., 2014, Koyejo et al., 2014)

Class-probabilities and classification

Many classification measures optimised by $\operatorname{sign}(\eta(x) - t)$

- 0-1 error $\rightarrow t = \frac{1}{2}$
- Balanced error $\rightarrow t = \pi$
- F-score $\rightarrow$ optimal t depends on D
  - (Lipton et al., 2014, Koyejo et al., 2014)

We can relate this to thresholding of $\bar{\eta}$!

Corrupted class-probabilities and classification

By monotone relationship,

$$\eta(x) > t \iff \bar{\eta}(x) > \phi_{\alpha,\beta,\pi}(t).$$

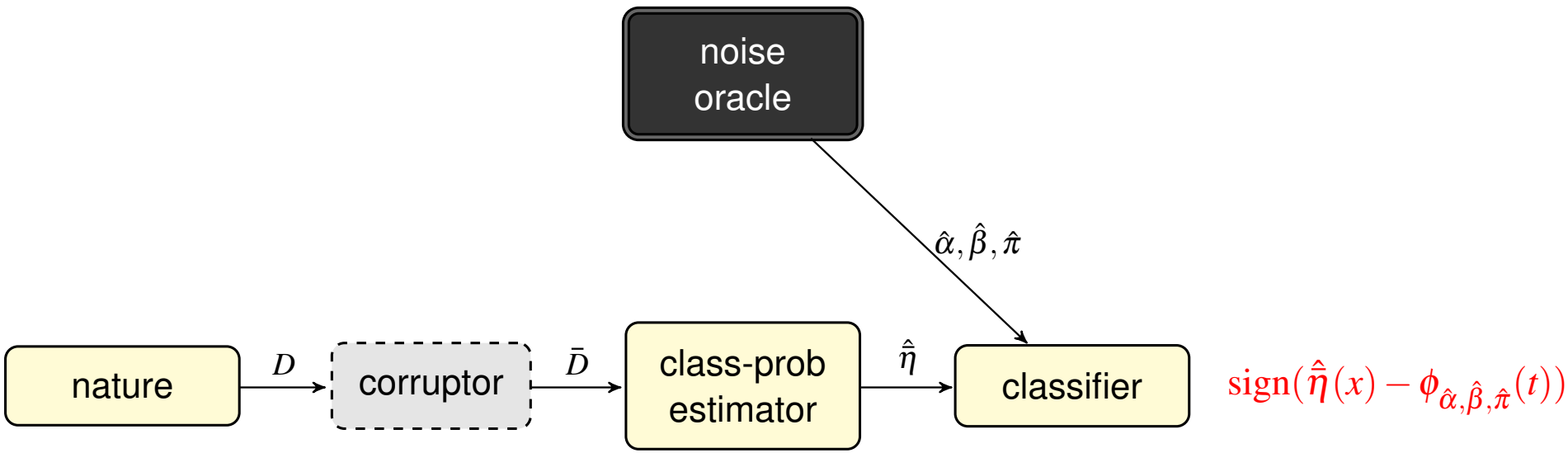
Threshold $\bar{\eta}$ at $\phi_{\alpha,\beta,\pi}(t) \to$ optimal classification on D

Can translate into regret bound e.g. for 0-1 loss

Story so far

Classification scheme requires:

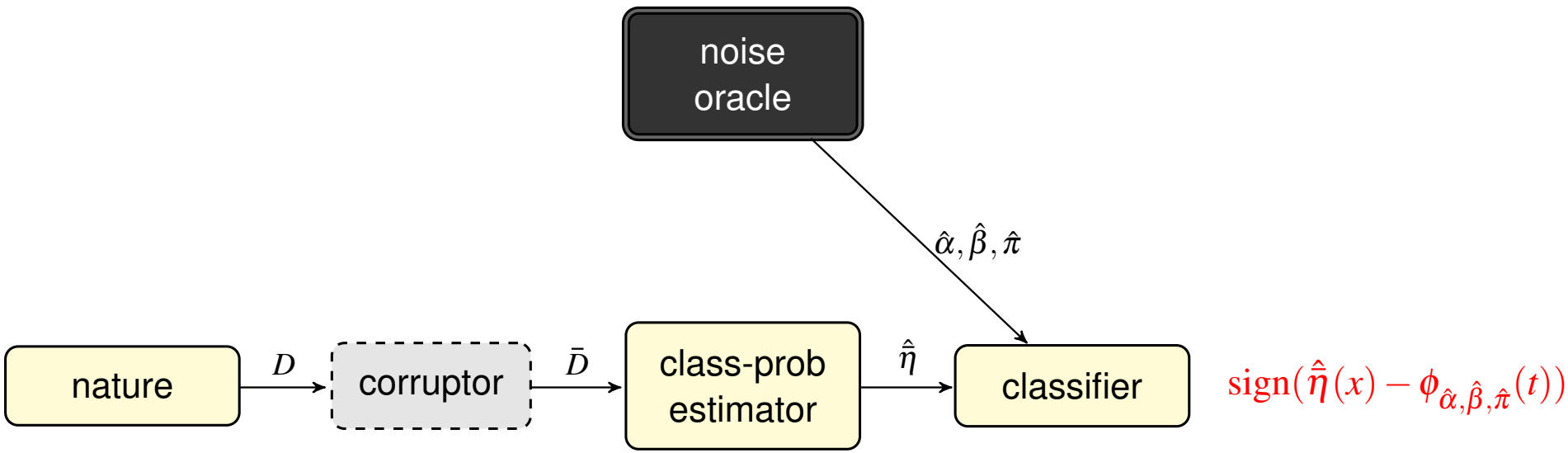
- $\bar{\eta}$
- t
- α, β, π



Story so far

Classification scheme requires:

- $\bar{\eta} \rightarrow$ class-probability estimation
- t
- α, β, π

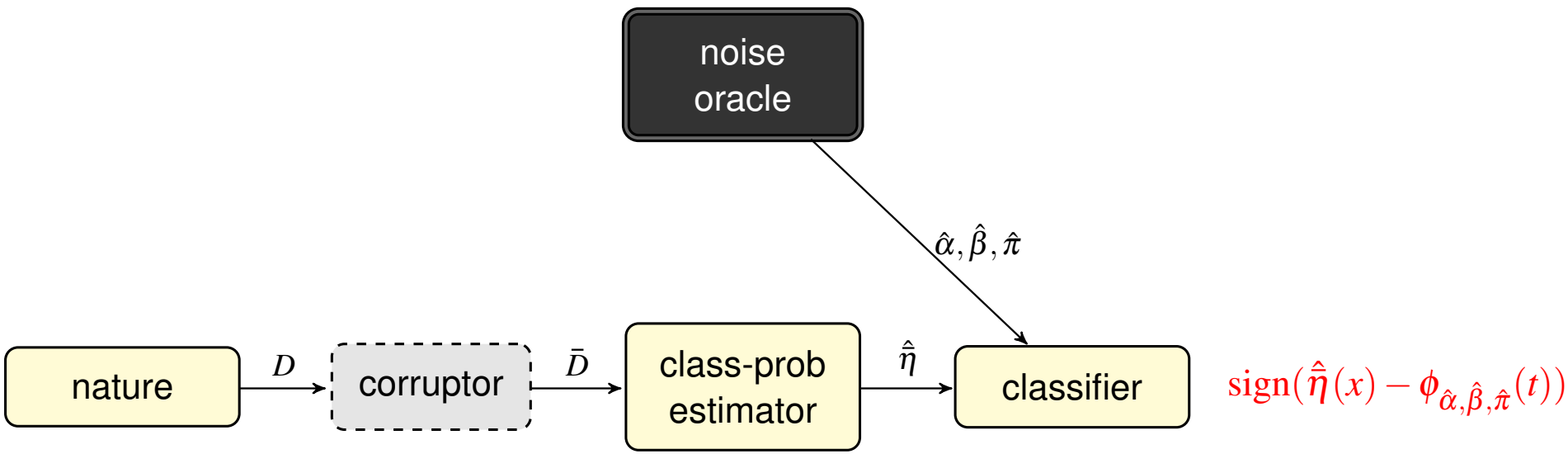


Kernel logistic regression

Story so far

Classification scheme requires:

- $\bar{\eta} \rightarrow$ class-probability estimation
- $t \rightarrow$ if unknown, alternate approach (see poster)
- α, β, π

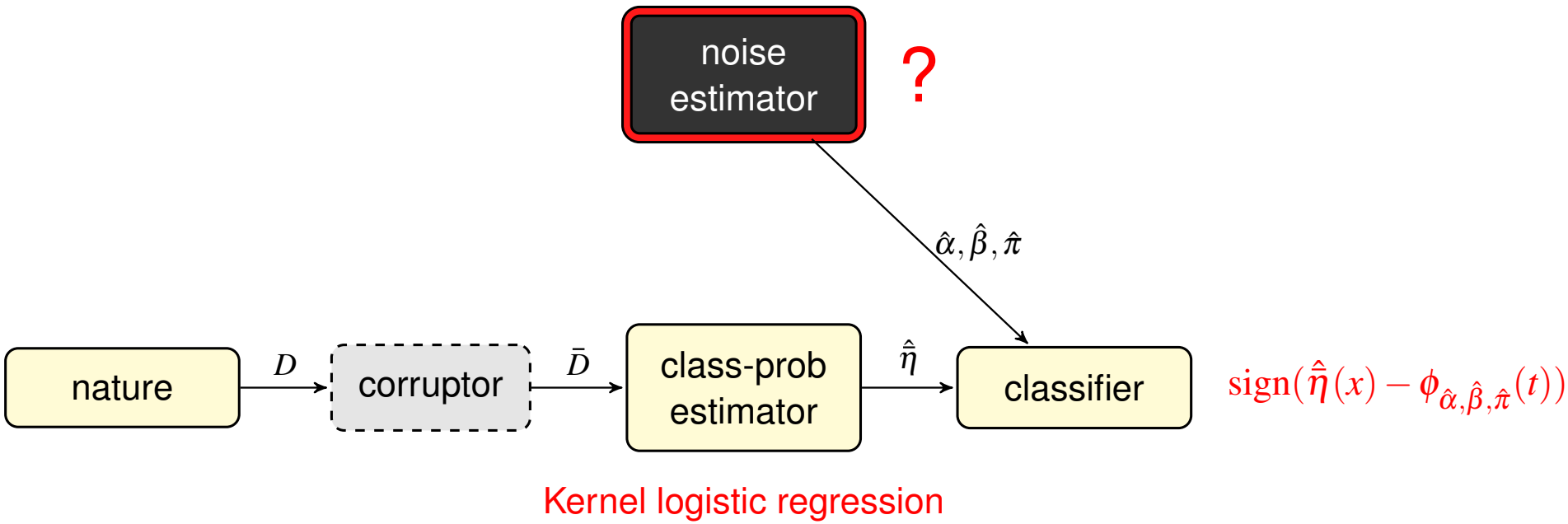


Kernel logistic regression

Story so far

Classification scheme requires:

- $\bar{\eta} \rightarrow$ class-probability estimation
- $t \rightarrow$ if unknown, alternate approach (see poster)
- $\alpha, \beta, \pi \rightarrow$ **can we estimate these?**



Kernel logistic regression

Estimating noise rates: some bad news

π strongly non-identifiable!

- $\bar{\pi}$ allowed to be arbitrary (e.g. PU learning)

α, β non-identifiable without assumptions (Scott et al., 2013)

Can we estimate α, β under assumptions?

Weak separability assumption

Assume that D is "weakly separable":

$$\min_{x \in \mathcal{X}} \eta(x) = 0$$

$$\max_{x \in \mathcal{X}} \eta(x) = 1$$

- i.e. $\exists$ deterministically +'ve and -'ve instances
- weaker than full separability

Weak separability assumption

Assume that D is "weakly separable":

$$\min_{x \in \mathcal{X}} \eta(x) = 0$$

$$\max_{x \in \mathcal{X}} \eta(x) = 1$$

- i.e. $\exists$ deterministically +'ve and -'ve instances
- weaker than full separability

Assumed range of η constrains observed range of $\bar{\eta}$!

Estimating noise rates

Proposition

Pick any weakly separable D. Then, for any $\bar{D}$,

$$\alpha = \frac{\eta_{\min} \cdot (\eta_{\max} - \bar{\pi})}{\bar{\pi} \cdot (\eta_{\max} - \eta_{\min})} \text{ and } \beta = \frac{(1 - \eta_{\max}) \cdot (\bar{\pi} - \eta_{\min})}{(1 - \bar{\pi}) \cdot (\eta_{\max} - \eta_{\min})}$$

where

$$\eta_{\min} = \min_{x \in \mathcal{X}} \bar{\eta}(x)$$

$$\eta_{\max} = \max_{x \in \mathcal{X}} \bar{\eta}(x)$$

α, β can be estimated from corrupted data alone

Estimating noise rates: special cases

Label noise

$$\rho = 1 - \eta_{\max}$$
$$= \eta_{\min}$$
$$\pi = \frac{\bar{\pi} - \eta_{\min}}{\eta_{\max} - \eta_{\min}}$$

(Elkan and Noto, 2008),
(Liu and Tao, 2014)

PU learning

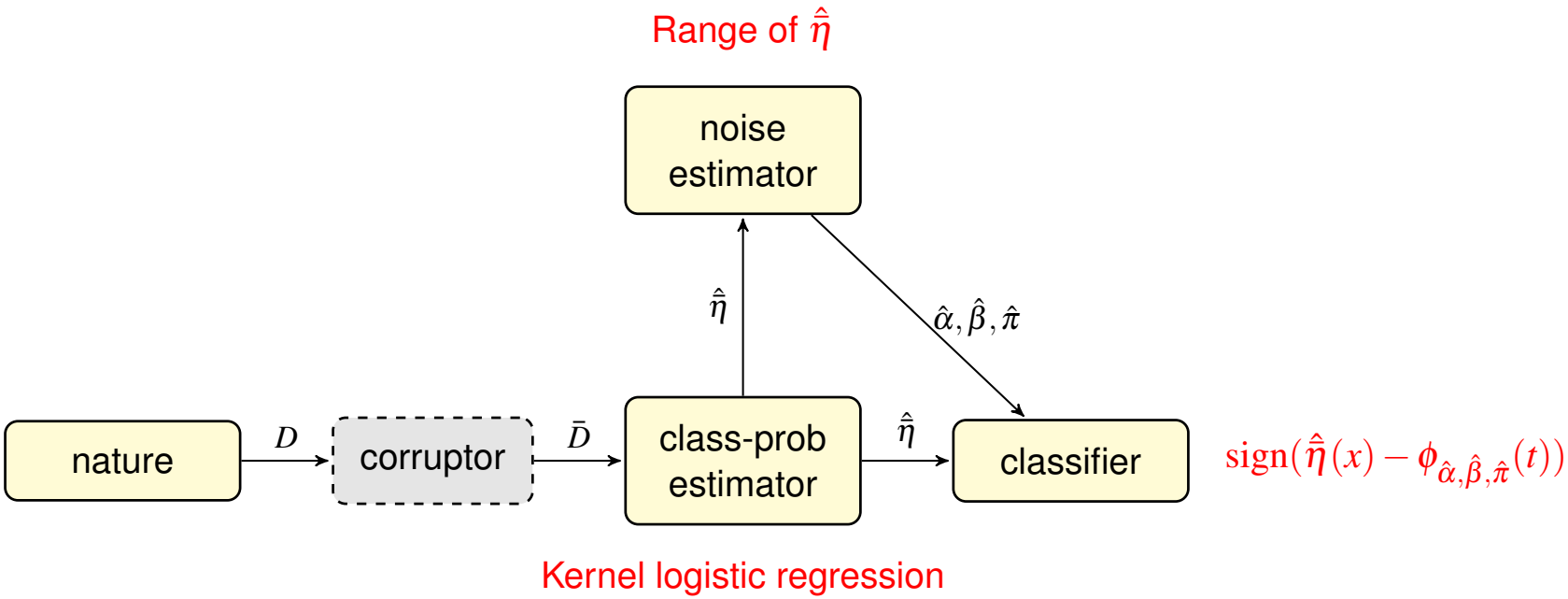
$$\alpha = 0$$
$$\beta = \pi$$
$$= \frac{1 - \eta_{\max}}{\eta_{\max}} \cdot \frac{\bar{\pi}}{1 - \bar{\pi}}$$

c.f. mixture proportion estimate of (Scott et al., 2013)

In these cases, π can be estimated as well

Story so far

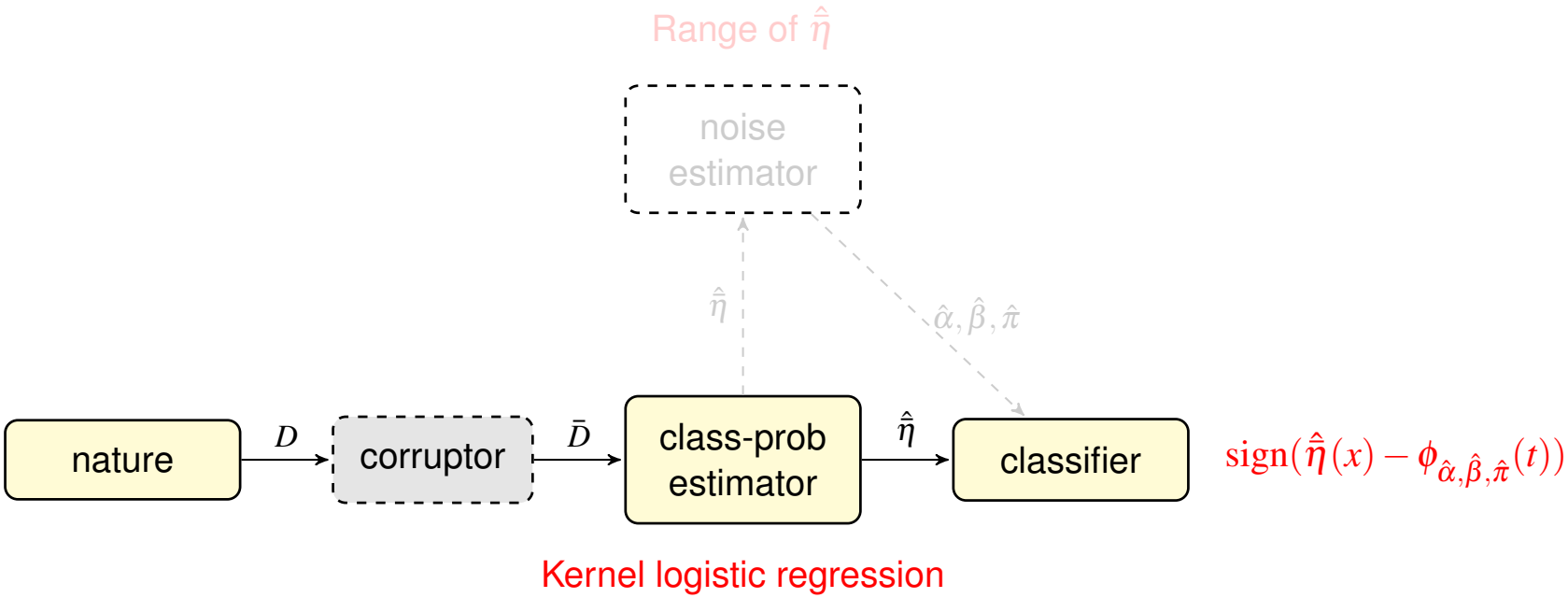
Optimal classification in general requires α, β, π



Story so far

Optimal classification in general requires α, β, π

- **when does $\phi_{\alpha,\beta,\pi}(t)$ not depend on α, β, π?**



Classification without noise rates

Balanced error (BER) of classifier

Balanced error (BER) of a classifier $f : \mathcal{X} \to \{\pm 1\}$ is:

$$\mathrm{BER}^D(f) = \frac{\mathrm{FPR}^D(f) + \mathrm{FNR}^D(f)}{2}$$

for false positive and negative rates $\mathrm{FPR}^D(f), \mathrm{FNR}^D(f)$

- average classification performance on each class
- optimal classifier is $\mathrm{sign}(\eta(x) - \pi)$

BER “immunity” under corruption

Proposition (c.f. (Zhang and Lee, 2008))

For any $D, \bar{D}$, and classifier $f : \mathcal{X} \to \{\pm 1\}$,

$$\mathrm{BER}^{\bar{D}}(f) = (1 - \alpha - \beta) \cdot \mathrm{BER}^{D}(f) + \frac{\alpha + \beta}{2}$$

BER "immunity" under corruption

Proposition (c.f. (Zhang and Lee, 2008))

For any $D, \bar{D}$, and classifier $f : \mathcal{X} \to \{\pm 1\}$,

$$\mathrm{BER}^{\bar{D}}(f) = (1 - \alpha - \beta) \cdot \mathrm{BER}^{D}(f) + \frac{\alpha + \beta}{2}$$

BER-optimal classifiers on clean and corrupted coincide

- $\mathrm{sign}(\eta(x) - \pi) = \mathrm{sign}(\bar{\eta}(x) - \bar{\pi})$

BER "immunity" under corruption

Proposition (c.f. (Zhang and Lee, 2008))

For any $D, \bar{D}$, and classifier $f : \mathcal{X} \to \{\pm 1\}$,

$$\mathrm{BER}^{\bar{D}}(f) = (1 - \alpha - \beta) \cdot \mathrm{BER}^{D}(f) + \frac{\alpha + \beta}{2}$$

BER-optimal classifiers on clean and corrupted coincide

- $\mathrm{sign}(\eta(x) - \pi) = \mathrm{sign}(\bar{\eta}(x) - \bar{\pi})$

Minimise clean BER $\to$ don't need to know corruption rates!

- threshold on $\bar{\eta}$ does not need α, β, π

BER "immunity" & class-probability estimation

Trivially, we also have

$$\text{regret}_{\text{BER}}^{D}(f) = (1 - \alpha - \beta)^{-1} \cdot \text{regret}_{\text{BER}}^{\bar{D}}(f).$$

i.e. **good corrupted BER $\implies$ good clean BER**

- can make $\text{regret}_{\text{BER}}^{\bar{D}}(f) \to 0$ by class-probability estimation

Similar result for AUC (see poster)

BER “immunity” under corruption: proof

From (Scott et al., 2013),

$$\begin{bmatrix} \mathrm{FPR}^{\bar{D}}(f) & \mathrm{FNR}^{\bar{D}}(f) \end{bmatrix}^T = \begin{bmatrix} \mathrm{FPR}^{D}(f) & \mathrm{FNR}^{D}(f) \end{bmatrix}^T \cdot \begin{bmatrix} 1-\beta & -\alpha \\ -\beta & 1-\alpha \end{bmatrix} + \begin{bmatrix} \beta & \alpha \end{bmatrix}^T,$$

BER "immunity" under corruption: proof

From (Scott et al., 2013),

$$\begin{bmatrix} \mathrm{FPR}^{\bar{D}}(f) & \mathrm{FNR}^{\bar{D}}(f) \end{bmatrix}^T = \begin{bmatrix} \mathrm{FPR}^{D}(f) & \mathrm{FNR}^{D}(f) \end{bmatrix}^T \cdot \begin{bmatrix} 1-\beta & -\alpha \\ -\beta & 1-\alpha \end{bmatrix} + \begin{bmatrix} \beta & \alpha \end{bmatrix}^T,$$

and $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is an eigenvector of $\begin{bmatrix} 1-\beta & -\alpha \\ -\beta & 1-\alpha \end{bmatrix}$

Are other measures "immune"?

BER is only (non-trivial) performance measure for which:

- corrupted risk = affine transform of clean risk
  - because of eigenvector interpretation
- corrupted threshold is independent of α, β, π
  - because of nature of $\phi_{\alpha,\beta,\pi}$

(see poster)

Other performance measures $\rightarrow$ need (one of) α, β, π

Experiments

Experimental setup

Injected label noise on UCI datasets

Estimate corrupted class-probabilities via neural network

- well-specified if D linearly separable:

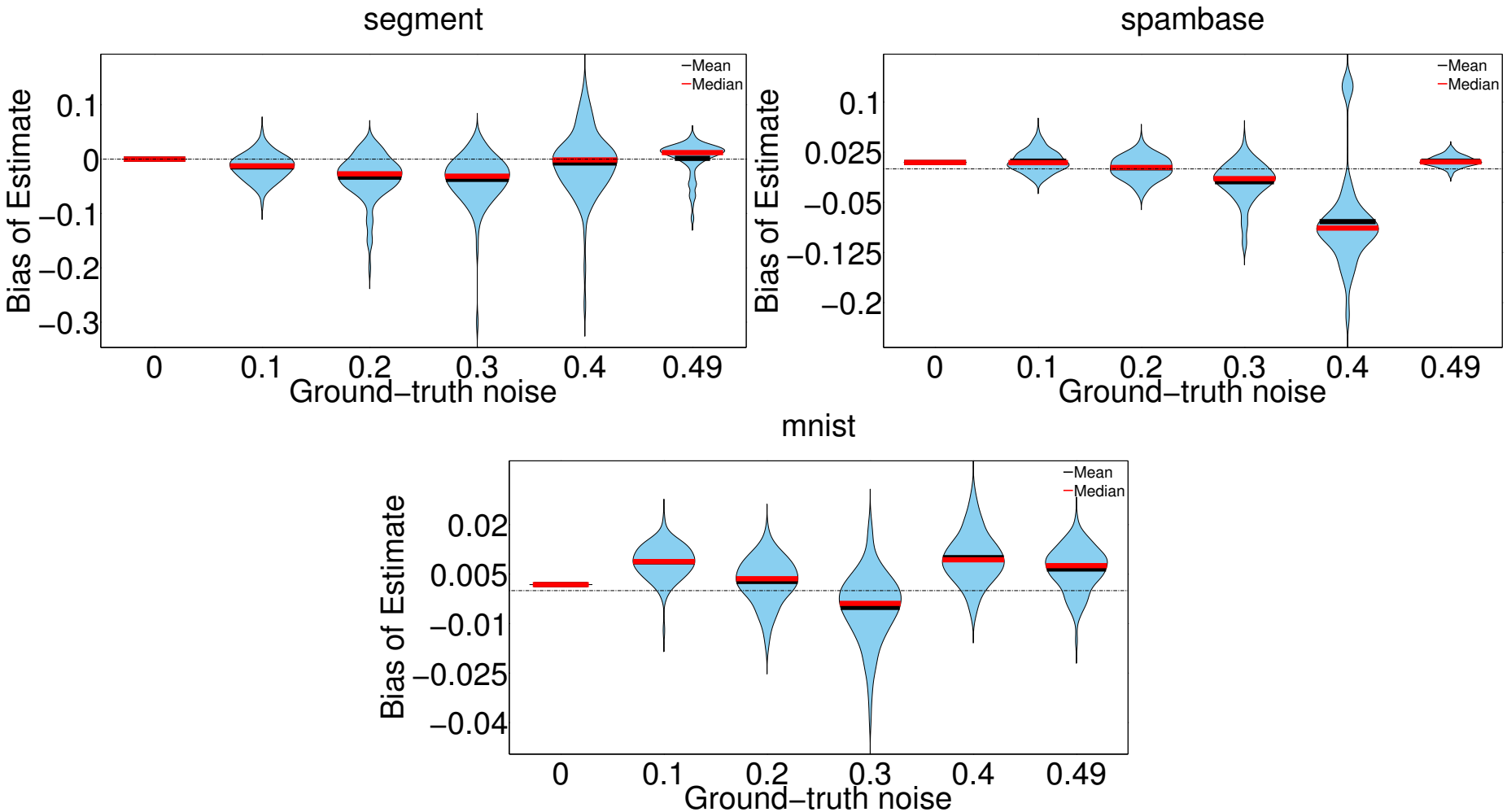
$$\eta(x) = \sigma(\langle w, x \rangle) \implies \bar{\eta}(x) = a \cdot \sigma(\langle w, x \rangle) + b$$

Evaluate:

- reliability of noise estimates
- BER performance on clean test set
  - corrupted data used for training and validation
- 0-1 performance on clean test set (see poster)

Experimental results: noise rates

Estimated noise rates are generally reliable



Experimental results: BER immunity

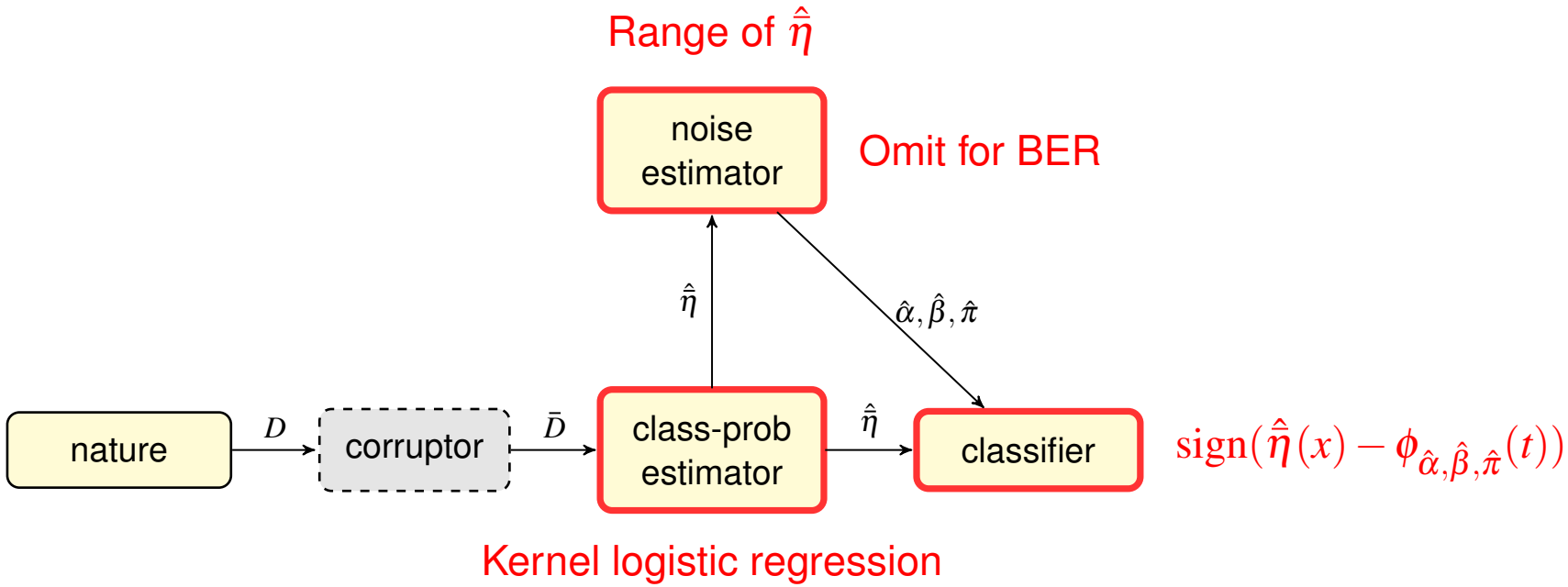
Generally, low observed degradation in BER

Dataset	**Noise**	1 - **AUC** (%)	**BER** (%)
segment	None	0.00 ± 0.00	0.00 ± 0.00
	$(\rho_+, \rho_-) = (0.1, 0.0)$	0.00 ± 0.00	0.01 ± 0.00
	$(\rho_+, \rho_-) = (0.1, 0.2)$	0.02 ± 0.01	0.90 ± 0.08
	$(\rho_+, \rho_-) = (0.2, 0.4)$	0.03 ± 0.01	3.24 ± 0.20
spambase	None	2.49 ± 0.00	6.93 ± 0.00
	$(\rho_+, \rho_-) = (0.1, 0.0)$	2.67 ± 0.02	7.10 ± 0.03
	$(\rho_+, \rho_-) = (0.1, 0.2)$	3.01 ± 0.03	7.66 ± 0.05
	$(\rho_+, \rho_-) = (0.2, 0.4)$	4.91 ± 0.09	10.52 ± 0.13
mnist	None	0.92 ± 0.00	3.63 ± 0.00
	$(\rho_+, \rho_-) = (0.1, 0.0)$	0.95 ± 0.01	3.56 ± 0.01
	$(\rho_+, \rho_-) = (0.1, 0.2)$	0.97 ± 0.01	3.63 ± 0.02
	$(\rho_+, \rho_-) = (0.2, 0.4)$	1.17 ± 0.02	4.06 ± 0.03

Conclusion

Learning from corrupted binary labels

Monotone relationship $\bar{\eta}(x) = \phi_{\alpha,\beta,\pi}(\eta(x))$ facilitates:



Future work

Better noise estimators in special cases?

- c.f. (Elkan and Noto, 2008) when D separable

Fusion with “loss transfer” (Natarajan et al., 2013) approach

- assumes noise rates known
- better for misspecified models?
  - c.f. non-robustness of convex surrogate minimisation

Thanks![1]

[1] Drop by the poster for more (Paper ID 69)