CN=CPCN

Liva Ralaivola, François Denis, Christophe N. Magnan

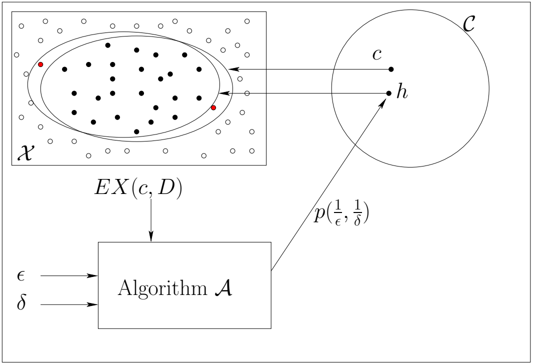
Laboratoire d'Informatique Fondamentale de Marseille
Université de Provence

ICML 2006

Outline

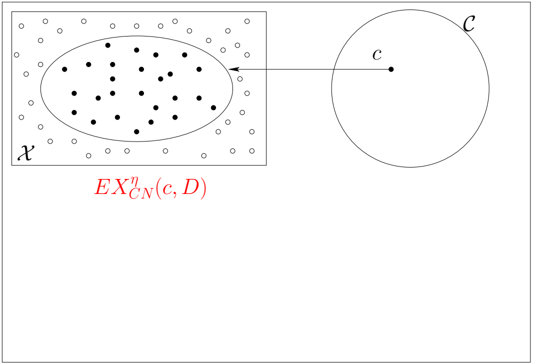
1. PAC learning
2. $CN = CCCN$
3. $CCCN = CPCN$
4. Conclusion

PAC learning



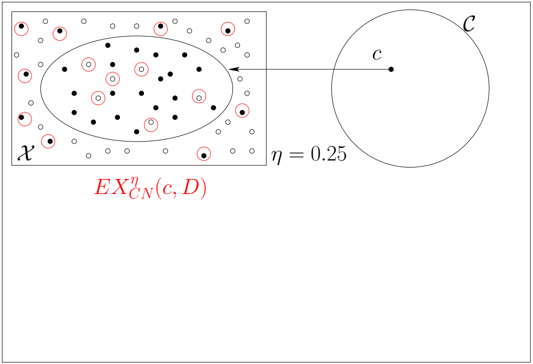
$\mathcal{C}$ is PAC learnable iff: $\exists \mathcal{A}, \forall c, \ \forall D, \ \forall \epsilon, \delta,$
$\mathcal{A}(EX(c, D), \epsilon, \delta) \rightarrow h$ s.t. $P(err_D(h) > \epsilon) < \delta$
where $err_D(h) = P_{x \sim D}(h(x) \neq c(x))$.

PAC Learning under Classification Noise [Angluin-Laird, 88]



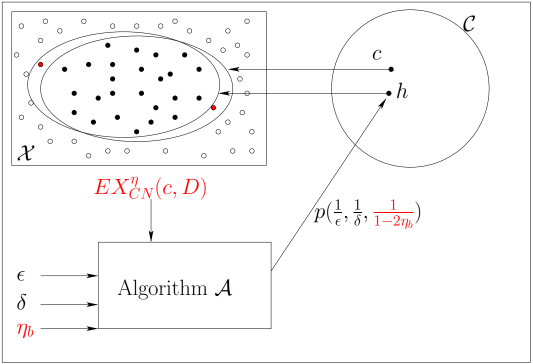
$$EX^{\eta}_{CN}(c, D) : \langle x, c^{\eta}(x)\rangle \text{ s.t. } c^{\eta}(x) = \begin{cases} c(x) \text{ with prob. } 1-\eta \\ 1-c(x) \text{ with prob. } \eta \end{cases}$$

PAC Learning under Classification Noise [Angluin-Laird, 88]



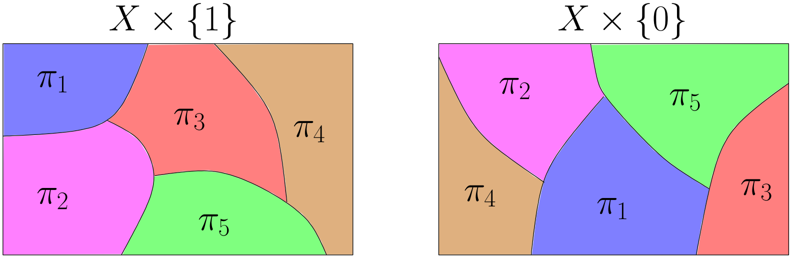
$$EX^{\eta}_{CN}(c, D) : \langle x, c^{\eta}(x) \rangle \text{ s.t. } c^{\eta}(x) = \begin{cases} c(x) \text{ with prob. } 1 - \eta \\ 1 - c(x) \text{ with prob. } \eta \end{cases}$$

PAC Learning under Classification Noise [Angluin-Laird, 88]



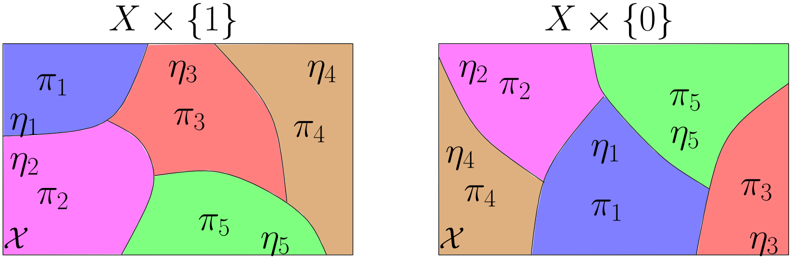
$\mathcal{C}$ is PAC learnable under CN iff: $\exists \mathcal{A}, \forall c, \ \forall D,$ $\forall \epsilon, \delta, \forall \eta \leq \eta_b < 1/2, \ \mathcal{A}(EX^{\eta}_{CN}(c, D), \epsilon, \delta, \eta_b) \rightarrow$ h s.t. $P(err_D(h) > \epsilon) < \delta$.

Constant Partition Classification Noise [Decatur, 1997]



$\Pi = \{\pi_1, ..., \pi_k\}$: partition of $\mathcal{X}$x$\{0, 1\}$

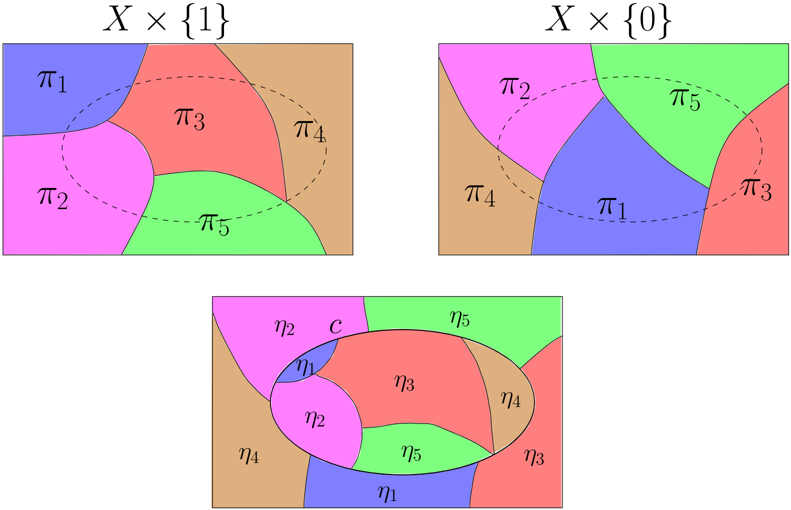
Constant Partition Classification Noise [Decatur, 1997]



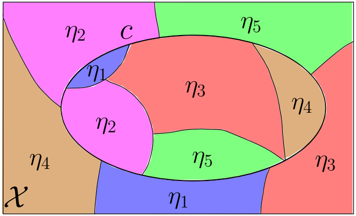
$\Pi = \{\pi_1, ..., \pi_k\}$: partition of $\mathcal{X}$x$\{0,1\}$

$\overrightarrow{\eta} = [\eta_1, ..., \eta_k]$, $\eta_i \in [0, 1/2]$

Constant Partition Classification Noise [Decatur, 1997]



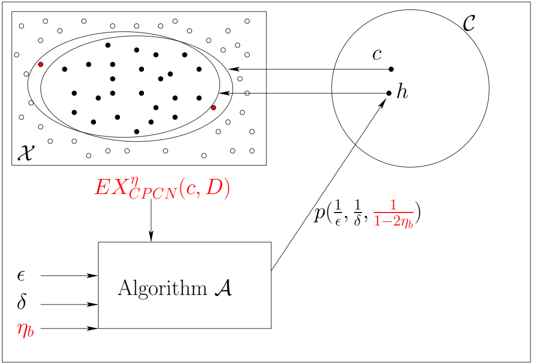
Constant Partition Classification Noise [Decatur, 1997]



EX^{η}_{CPCN} : $\langle x, c^{\eta}(x)\rangle$ s.t.

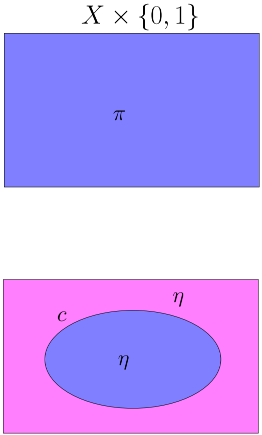
$$c^{\eta}(x) = \begin{cases} c(x) \text{ with prob. } 1-\eta_i \\ 1-c(x) \text{ with prob. } \eta_i \\ \text{where } \pi_i(\langle x, c(x)\rangle) = 1 \end{cases}$$

PAC Learning under CPCN [Decatur, 1997]

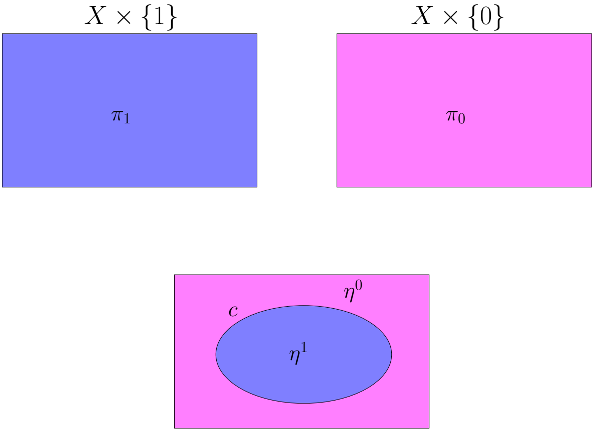


$\mathcal{C}$ is PAC learnable under CPCN iff: $\exists \mathcal{A}, \forall c, \forall D, \forall \epsilon, \delta,$
$\forall \Pi = \{\pi_1, ..., \pi_k\}, \forall \eta_b, \forall \eta = [\eta_1 ... \eta_k]$ s.t. $\eta_i \leq \eta_b < 1/2$,
$\mathcal{A}(EX^{\eta}_{CPCN}(c, D), \epsilon, \delta, \eta_b) \rightarrow h$ s.t. $P(err_D(h) \geq \epsilon) \leq \delta$.

CN: a particular case of $CPCN$

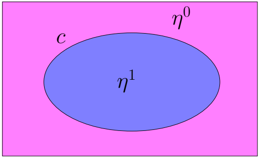


Class Conditional Classification Noise



Noise rates only depend on the class of the example.

Class Conditional Classification Noise



$$EX_{CCCN}^{[\eta^1,\eta^0]} \text{ s.t.}$$

$$c^\eta(x) = \begin{cases} 1 \text{ with prob. } 1-\eta^1 \text{ and } 0 \text{ with prob. } \eta^1 \text{ if } c(x)=1 \\ 0 \text{ with prob. } 1-\eta^0 \text{ and } 1 \text{ with prob. } \eta^0 \text{ if } c(x)=0 \end{cases}$$

Learnability Classes

- *CN*: Classes PAC learnable under Classification Noise.
- *CCCN*: Classes PAC learnable under Class Conditional Classification Noise.
- *CPCN*: Classes PAC learnable under Constant Partition Classification Noise.

$CN = CCCN = CPCN$

Trivially,

$$CPCN \subseteq CCCN \subseteq CN$$

We prove that

$$CN \subseteq CCCN \text{ and } CCNN \subseteq CPCN.$$

Outline

1. PAC learning
2. $CN = CCCN$
3. $CCCN = CPCN$
4. Conclusion

$CN \subseteq CCCN$

- Given $\mathcal{A}$ that PAC-learns $\mathcal{C}$ under CN,
- let us design $\mathcal{A}'$ that PAC-learns $\mathcal{C}$ under CCCN.

Idea:

- Add class conditional classification noise to the examples drawn from $EX_{CCCN}^{[\eta^1,\eta^0]}$, according to a tuning parameter γ, in order to approach uniform classification noise: S_γ.
- For each value of γ, let $\mathcal{A}(S_\gamma) \rightarrow h_\gamma$;
- Select a hypothesis from $\mathcal{H} = \{h_{\gamma_1}, \ldots, h_{\gamma_l}\}$.

Adding noise

Let $\gamma = [\rho, s]$ where $\rho \in [0, 1]$ and $s \in \{0, 1\}$.
Add CCC noise according to $[\rho s, \rho(1 - s)]$.
The resulting noise is

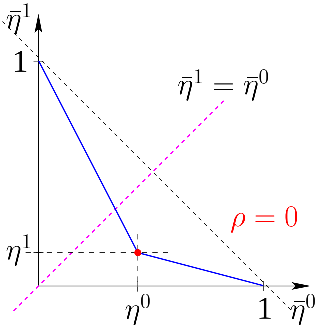
$$\begin{cases} \bar{\eta}^1 = (1-\rho)\eta^1 + (1-s)\rho \\ \bar{\eta}^0 = (1-\rho)\eta^0 + s\rho \end{cases}$$

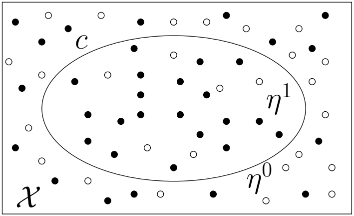
Let $\rho_{opt} = \frac{|\eta^1 - \eta^0|}{1+|\eta^1 - \eta^0|}$ and $s_{opt} = 1$ if $\eta^1 > \eta^0$ and 0 otherwise.

$$EX_{CCCN}^{[\bar{\eta}^1 \bar{\eta}^0]} \equiv EX_{CN}^{\eta_{opt}}$$

where $\eta_{opt} = \frac{max(\eta^1, \eta^0)}{1+|\eta^1 - \eta^0|}$. Remark that $\eta^0, \eta^1 < \eta_b \Rightarrow \eta_{opt} < \eta_b$.

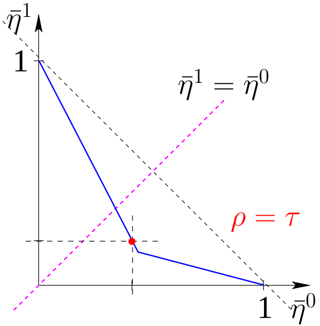
Adding noise

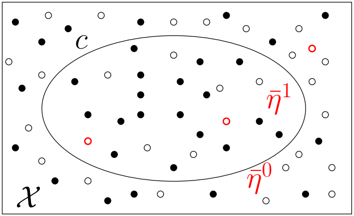




$$\mathcal{H} = \{h_0\}$$

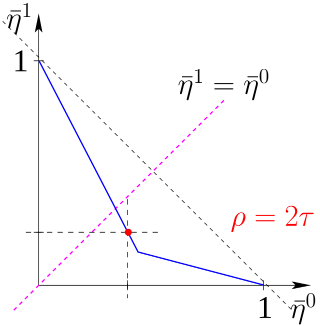
Adding noise

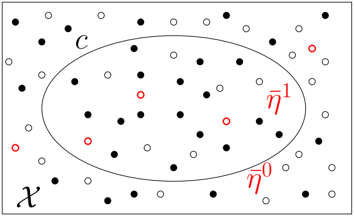




$$\mathcal{H} = \{h_0, h_1\}$$

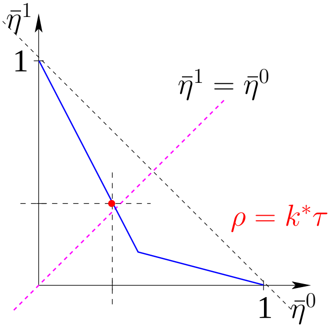
Adding noise

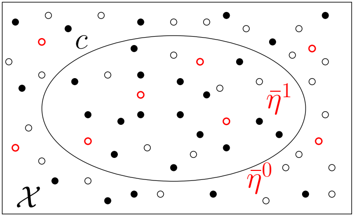




$$\mathcal{H} = \{h_0, h_1, h_2\}$$

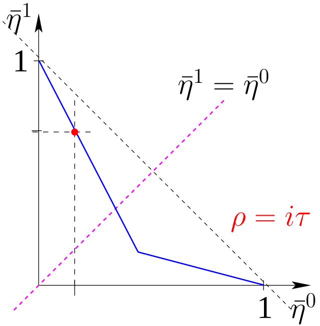
Adding noise

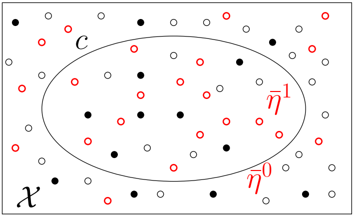




$$\mathcal{H} = \{h_0, h_1, h_2, ..., h_{k^*}\}$$

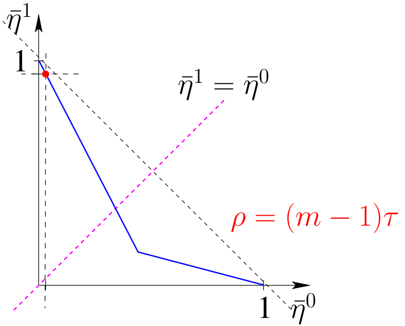
Adding noise

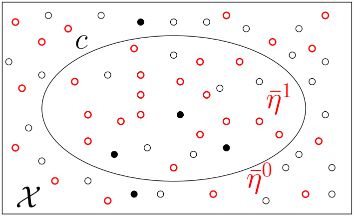




$$\mathcal{H} = \{h_0, h_1, h_2, ..., h_{k^*}, ..., h_i\}$$

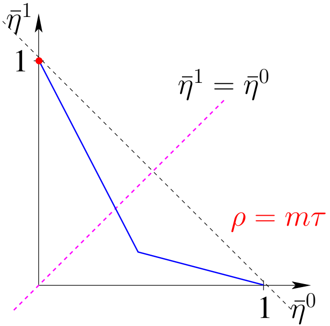
Adding noise

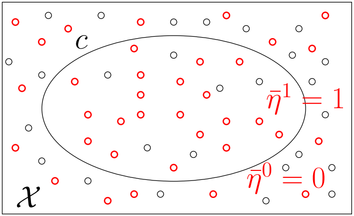




$$\mathcal{H} = \{h_0, h_1, h_2, \ldots, h_{k^*}, \ldots, h_i, \ldots, h_{m-1}\}$$

Adding noise





$$\mathcal{H} = \{h_0, h_1, h_2, ..., h_{k^*}, ..., h_i, ..., h_{m-1}, h_m\}$$

Approaching η_{opt}

$l = p(1/\epsilon, 1/\delta, 1/(1-2\eta_b))$: number of examples required by $\mathcal{A}$ to learn $\mathcal{C}$ under CN with learning parameters ϵ, δ.

$\mathcal{S} = \{(x_1, c_1), \ldots, (x_l, c_l)\}$: sample drawn according to $EX(c, D)$.
Add to $\mathcal{S}$

- *uniform classification noise*: $\mathcal{S}^{\eta}$,
- *CC classification noise*: $\mathcal{S}^{\eta^0,\eta^1}$

in such a way that

- $\mathcal{S}^{\eta}$ follows the distribution EX^{η}_{CN},
- $\mathcal{S}^{\eta^0,\eta^1}$ follows the distribution $EX^{\eta^0,\eta^1}_{CCCN}$,
- With probability $\geq 1 - Max_i|\eta^i - \eta|$, a given example of $\mathcal{S}$ has the same label in $\mathcal{S}^{\eta}$ and $\mathcal{S}^{\eta^0,\eta^1}$.

Approaching η_{opt}

Let α s.t. $|\eta^i - \eta| < \alpha$ for $i = 0, 1 \Rightarrow Pr(\mathcal{S}^\eta \neq \mathcal{S}^{\eta^0,\eta^1}) < \delta$.

Suppose that $|\overline{\eta}^i - \eta_{opt}| < \alpha$ for $i = 0, 1$.

The probability that $\mathcal{A}(\mathcal{S}^{\overline{\eta}^0,\overline{\eta}^1})$ provides a *bad* hypothesis is smaller than

- the probability that $\mathcal{S}^{\eta_{opt}} \neq \mathcal{S}^{\overline{\eta}^0,\overline{\eta}^1}$ +
- the probability that $\mathcal{A}(\mathcal{S}^{\eta_{opt}})$ provides a *bad* hypothesis

Lemma: The probability that $\mathcal{A}(\mathcal{S}^{\overline{\eta}^0,\overline{\eta}^1})$ provides a *bad* hypothesis is smaller than 2δ.

Tuning the increment τ

Given ϵ, δ, η_b and $l = p(1/\epsilon, 1/\delta, 1/(1 - 2\eta_b))$, the number of examples required by $\mathcal{A}$ to learn under CN.

Lemma: If we set $\tau = \delta/l$, with probability $\geq 1 - 2\delta$, $\mathcal{H}$ will contain a hypothesis h^* s.t.

$$err_D(h^*) \leq \epsilon.$$

Minimizing the empirical error on noisy data

Let $p = P(c(x) = 1)$. For any h:

$$\begin{aligned} P(h(x) \neq c^{\eta^0,\eta^1}(x)) &= p\eta^1 + (1-p)\eta^0 \\ &+ (1-2\eta^1)P(h(x)=1, c(x)=0) \\ &+ (1-2\eta^0)P(h(x)=0, c(x)=1) \end{aligned}$$

Therefore,

$$err_D(h) \leq \frac{P(h(x) \neq c^{\eta^0,\eta^1}(x)) - (p\eta^1 + (1-p)\eta^0)}{1-2\eta_b}.$$

Minimizing the empirical error on noisy data is a good strategy.

Selecting a correct hypothesis.

- Draw a new test set S containing

$$O\left(\frac{1}{\varepsilon^2(1-2\eta_b)^2}\ln\frac{16l}{\delta^2}\right)$$

  examples according to EX^{η}_{CCCN}
- Select the hypothesis h_{min} from $\mathcal{H}$ that minimizes the empirical error on S.

With probability $\geq 1-\delta$, h_{min} has true error lower than ϵ.

CCCN=CN

Proposition: Any concept class that is efficiently CN-learnable is also efficiently CCCN-learnable:

$$CN \subseteq CCCN.$$

Outline

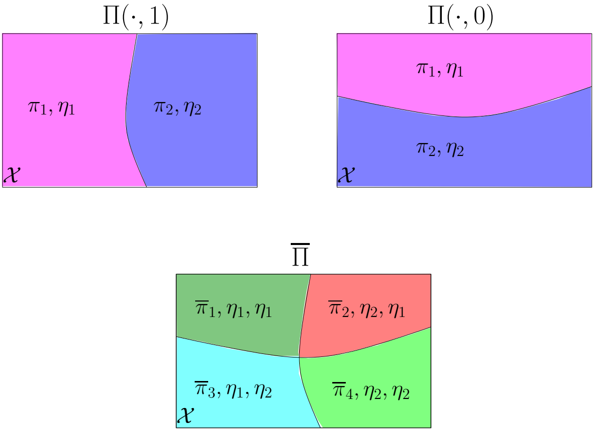
$CCCN \subseteq CPCN$

- Given $\mathcal{A}$ that PAC-learns $\mathcal{C}$ under CCCN,
- let us design $\mathcal{A}'$ that PAC-learns $\mathcal{C}$ under CPCN.

Idea:

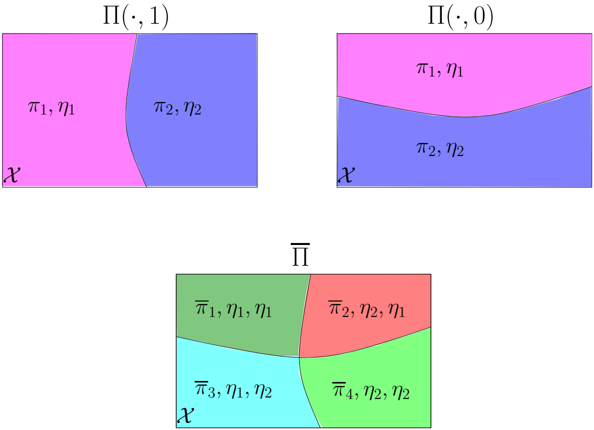
- Given a partition Π of $\mathcal{X} \times \{0, 1\}$, refine it to build a partition of $\mathcal{X}$,
- Transform the original problem in k learning problem under CCCN.

Partitioning $\mathcal{X}$



$$x \sim x' \text{ iff}$$
$$(x, 0) \sim_\pi (x', 0) \text{ and } (x, 1) \sim_\pi (x', 1)$$

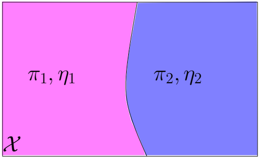
Partitioning $\mathcal{X}$



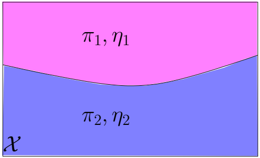
In each part of $\overline{\Pi}$, the noise rates are constant within each class label.

Partitioning $\mathcal{X}$

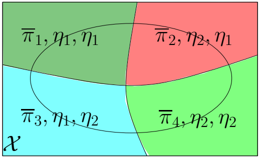
$\Pi(\cdot, 1)$



$\Pi(\cdot, 0)$

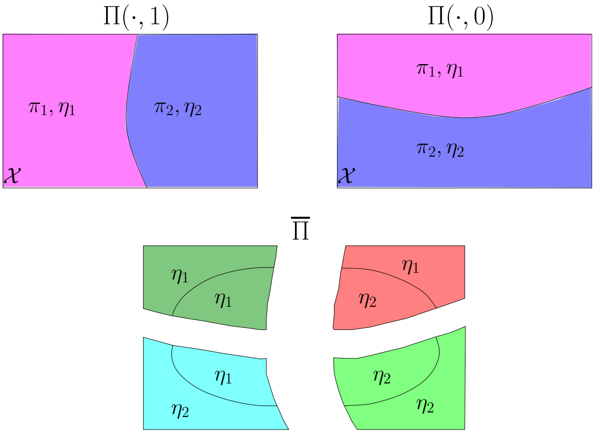


$\overline{\Pi}$



In each part of $\overline{\Pi}$, the noise rates are constant within each class label.

Partitioning $\mathcal{X}$



The original CPCN learning problem yields $|\overline{\Pi}|$ CCCN learning problems.

Learning under *CPCN*

Given access to $EX^{\eta}_{CPCN}(c, D)$,

- Compute $\bar{\Pi} = \{\bar{\pi}_1, \ldots, \bar{\pi}_k\}$,
- Eliminate parts $\bar{\pi}_i$ such that $D(\bar{\pi}_i)$ too small,
- For every remaining part $\bar{\pi}$, learn c using $EX^{\eta}_{CCCN}(c, D_{|\bar{\pi}})$.

Let $h_1, \ldots, h_k$ the output hypotheses: they are sufficient to learn c.

In order to get proper learning,

- Use the learned classifiers $h_1, \ldots, h_k$ to relabel the examples drawn from $EX^{\eta}_{CPCN}(c, D)$,
- Use these new examples to learn a classifier $h \in \mathcal{C}$.

$CCCN = CPCN$

Proposition: Any concept class that is efficiently CCCN-learnable is also efficiently CPCN-learnable:

$$CCCN \subseteq CPCN.$$

Outline

Conclusion and prospects

- We have shown that $CN = CCCN = CPCN$ given a bound η_b of the noise rates.
- Can this condition be ruled out?
- We have supposed that the noise rates are $< 1/2$. What results hold when the noise rates $\in [0, 1]$?