Beyond Empirical Risk Minimization

Santiago Mazuelas

smazuelas@bcamath.org

Basque Center for Applied Mathematics-BCAM

DATAI, University of Navarra, 21 of February 2024

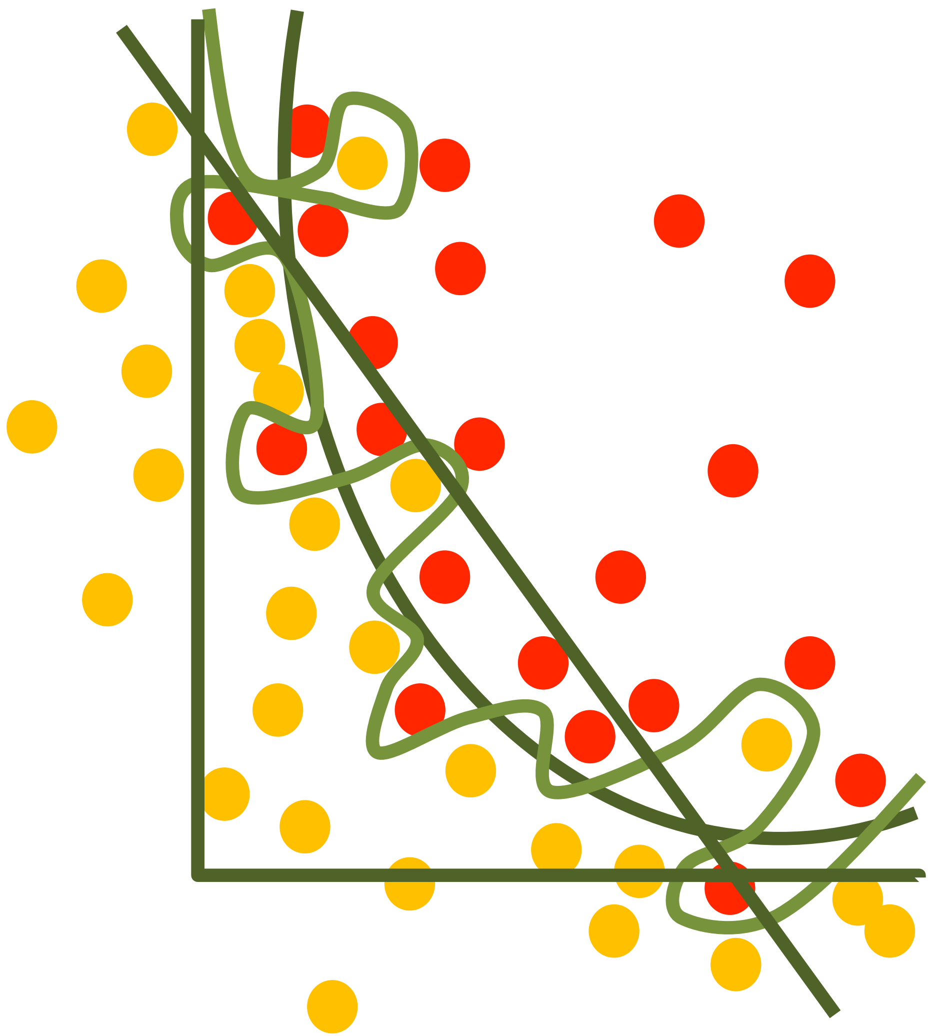
Empirical Risk Minimization

Consider multiple
“simple” hypothesis
and
choose one that fits well
the data

William of Ockham

Empirical Risk Minimization

Consider multiple
“simple” hypothesis
and
choose one that fits well
the data



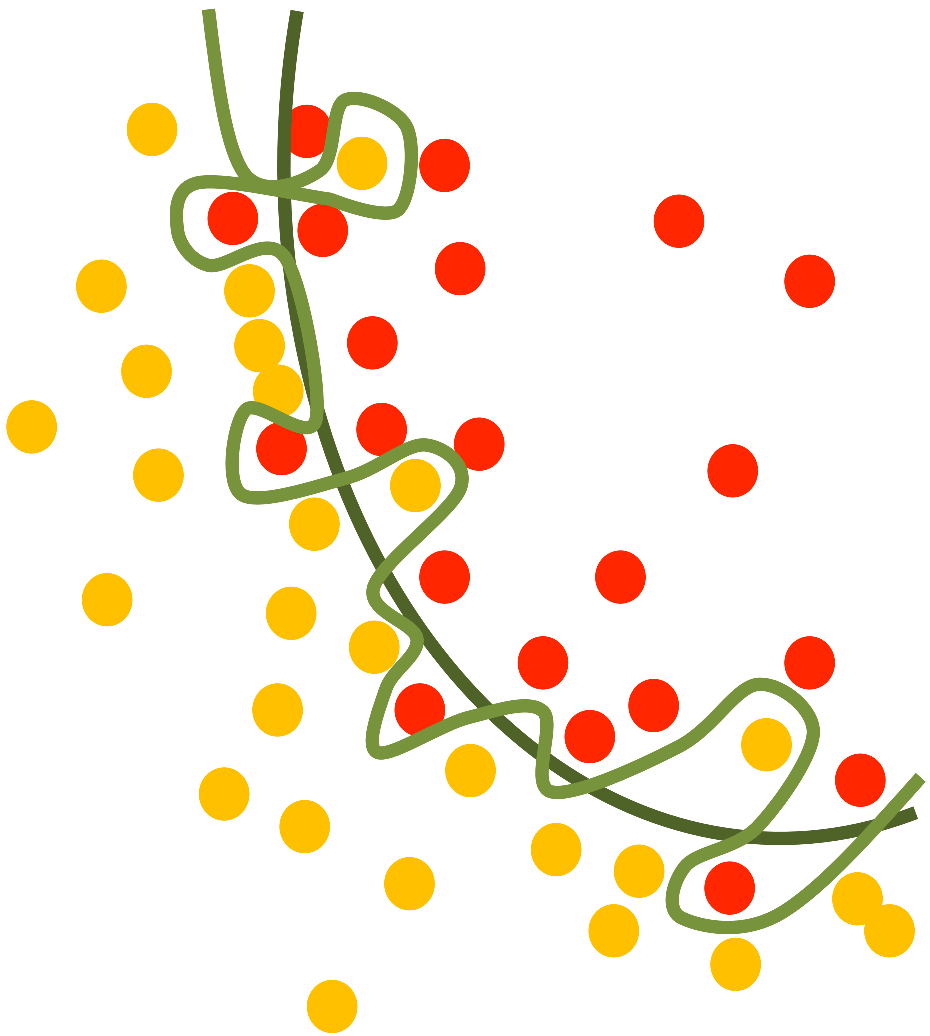
Empirical Risk Minimization

Strongly rely on the specific training samples available

Not clear how to address corrupted samples or distribution shifts

The quantity optimized at learning is not very meaningful

Fitting training samples is not directly related with the prediction process



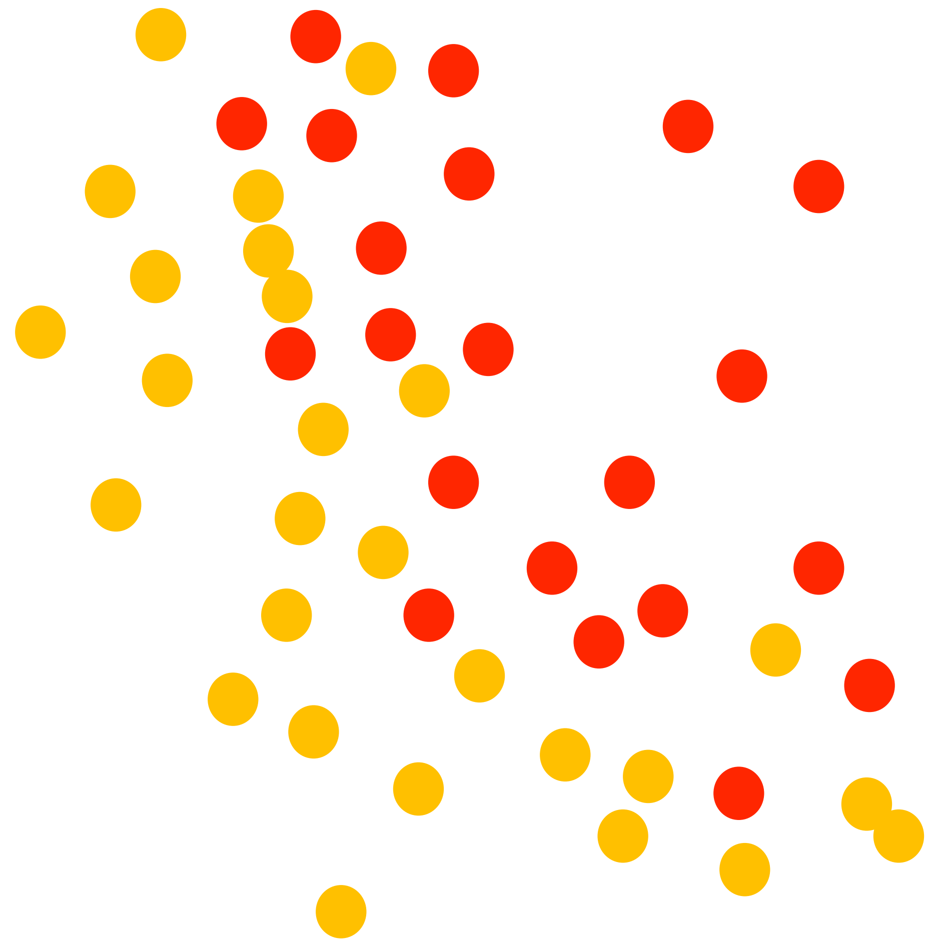
Robust Risk Minimization

Consider multiple scenarios
consistent with the data
and
choose a rule with small error
in the worst scenario

John von Neumann

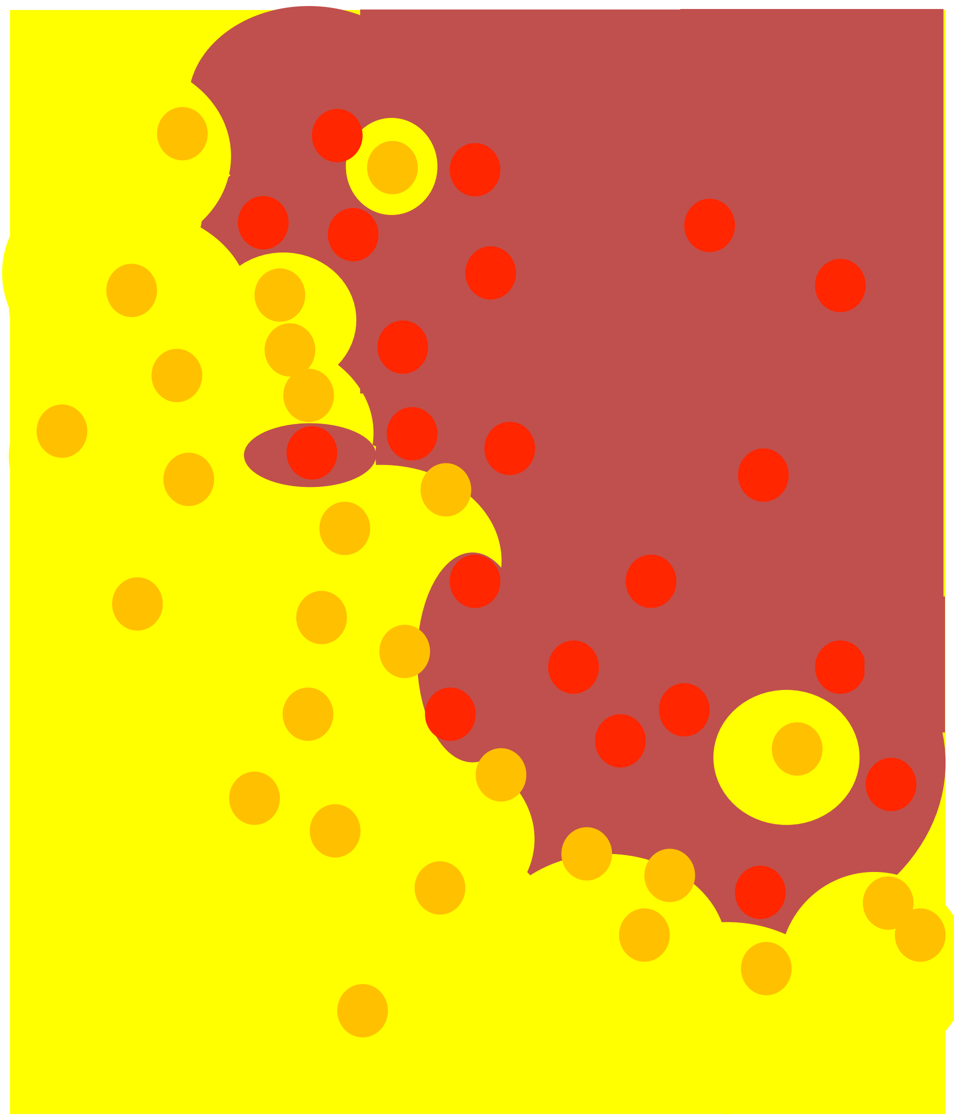
Robust Risk Minimization

Consider multiple scenarios
consistent with the data
and
choose a rule with small error
in the worst scenario



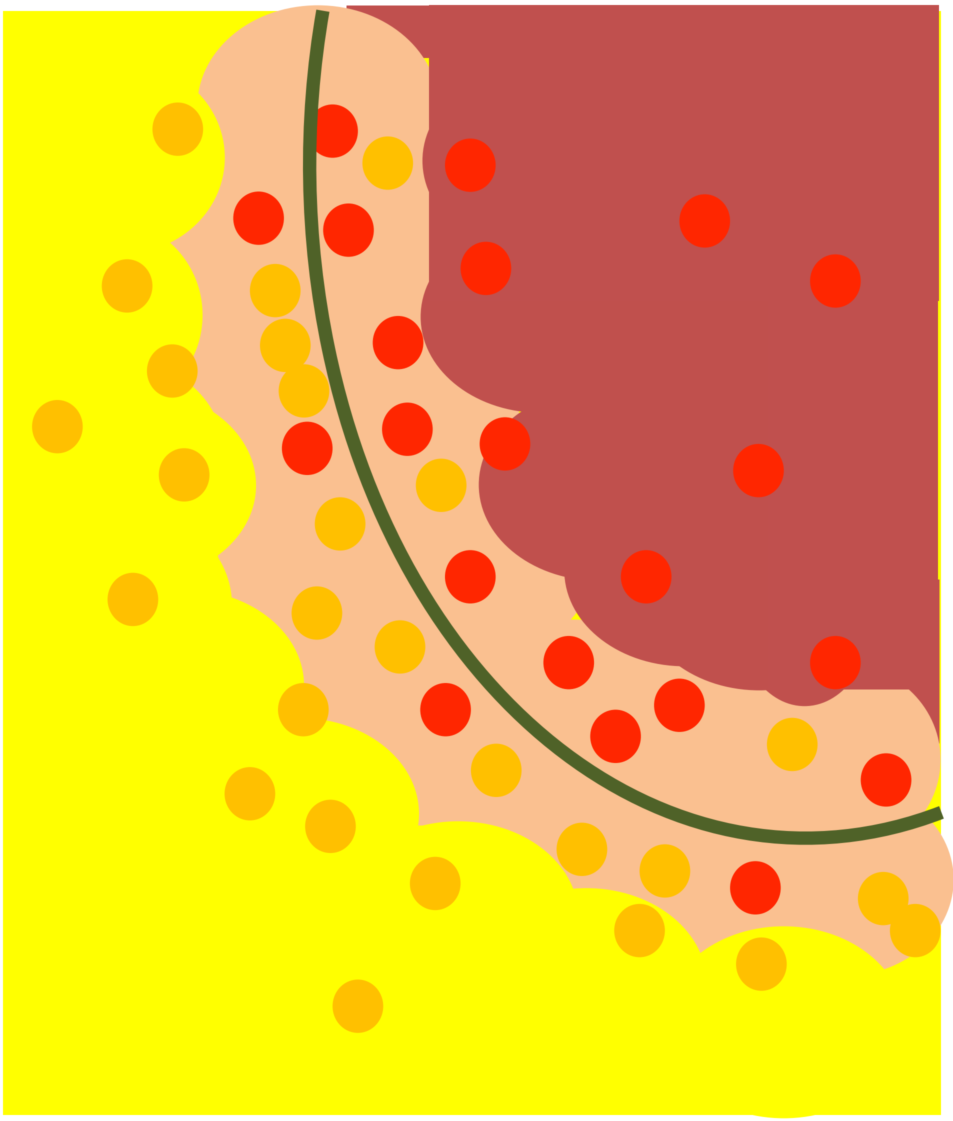
Robust Risk Minimization

Consider multiple scenarios
consistent with the data
and
choose a rule with small error
in the worst scenario



Robust Risk Minimization

Consider multiple scenarios
consistent with the data
and
choose a rule with small error
in the worst scenario



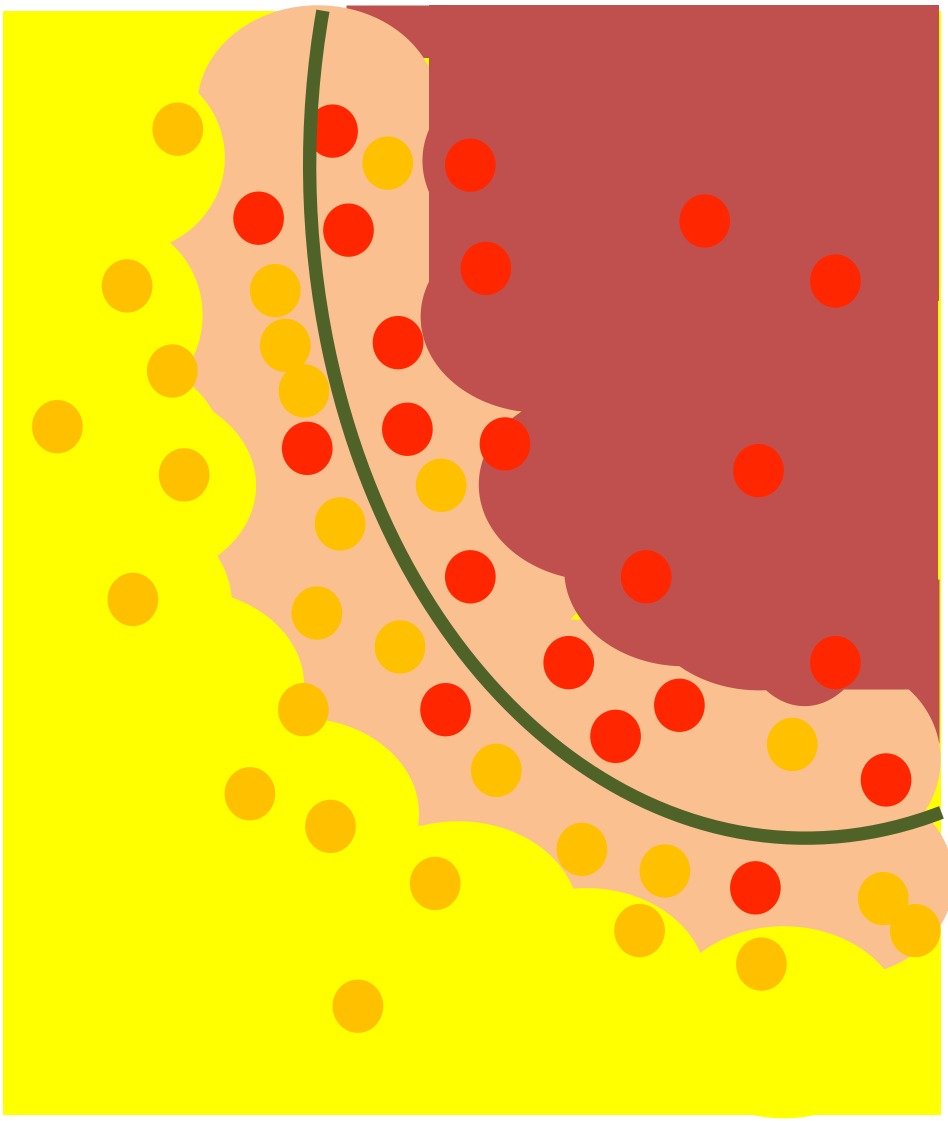
Robust Risk Minimization

Presence of corrupted samples or distribution shifts can be included on the considered scenarios

The quantity optimized at learning can provide an upper bound for the prediction error

Learning process analyzes the prediction performance of different rules

Can be too pessimistic with too broad feasible scenarios



Outline

- Introduction
- Generalized maximum entropy for supervised classification
- Minimax risk classifiers (MRCs)
- Performance guarantees 0-1 MRCs
- Multidimensional adaptation to concept drift

Maximum entropy principle

$$\begin{aligned} c: \quad \text{Words} &\rightarrow \{0,1\}^* \\ \text{“cat”} &\mapsto c(\text{“cat”}) = 1,0,1,1 \end{aligned}$$

$$\text{Words} = \{w\} \qquad \text{Codes} = \{c\} \qquad \ell(c,w) = [c(w)]$$

$$w \sim \mathrm{p}(w) \in \Delta(\text{Words})$$

$$\min_{c \in \text{Codes}} \mathbb{E}_{\mathrm{p}} \ell(c,w) = \min_{c \in \text{Codes}} \ell(c,p) = -\mathbb{E}_{\mathrm{p}} \log_2 \mathrm{p} = H(\mathrm{p})$$

$\mathrm{p} \in \mathcal{U}$ Given by expectation constraints $\rightarrow \max_{\mathrm{p} \in \mathcal{U}} H(\mathrm{p})$

[Jaynes, Phys. Rev., 1957], [P. Grünwald and A. P. Dawid, Annals of Statistics, 2004]

Supervised classification

$\text{Examples} = \{(x, y)\}$ x instance or attribute y label or class

$\text{Classification rules} = \{\mathrm{h} : \mathcal{X} \to \Delta(\mathcal{Y})\} = \mathrm{T}(\mathcal{X}, \mathcal{Y})$

$$\ell_{\text{0-1}}(\mathrm{h}, (x, y)) = \begin{cases} 1 & \text{if} \quad \mathrm{h}(x) \neq y \\ 0 & \text{if} \quad \mathrm{h}(x) = y \end{cases} \qquad \ell_{\text{0-1}}(\mathrm{h}, (x, y)) = 1 - \mathrm{h}(y|x)$$

$$\ell_{\log}(\mathrm{h}, (x, y)) = -\log \mathrm{h}(y|x)$$

ℓ-entropy for classification

$$H_\ell(\mathrm{p}) = \min_{\mathrm{h} \in \mathrm{T}(\mathcal{X}, \mathcal{Y})} \mathbb{E}_\mathrm{p} \ell(\mathrm{h}, (x, y)) = \min_{\mathrm{h} \in \mathrm{T}(\mathcal{X}, \mathcal{Y})} \ell(\mathrm{h}, \mathrm{p})$$

$$H_{\text{0-1}}(\mathrm{p}) = 1 - \sum_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \mathrm{p}(x, y)$$

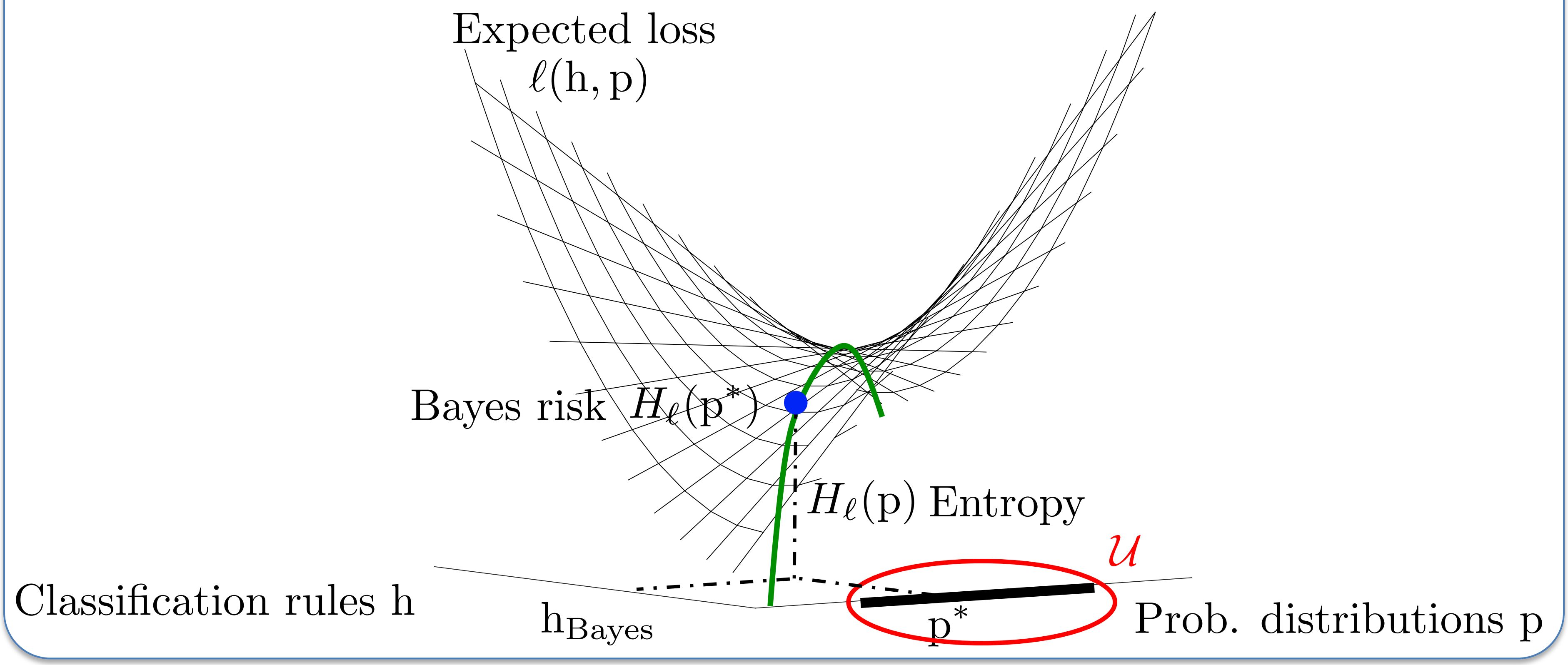
$$H_{\log}(\mathrm{p}) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathrm{p}(x, y) \log \frac{\mathrm{p}(x)}{\mathrm{p}(x, y)}$$

Maximum entropy and minimax rules

p^* underlying distribution of instance-label pairs

$$H_\ell(\mathrm{p}^*) = \min_{\mathrm{h} \in \mathrm{T}(\mathcal{X},\mathcal{Y})} \ell(\mathrm{h}, \mathrm{p}^*) \text{ Bayes risk}$$

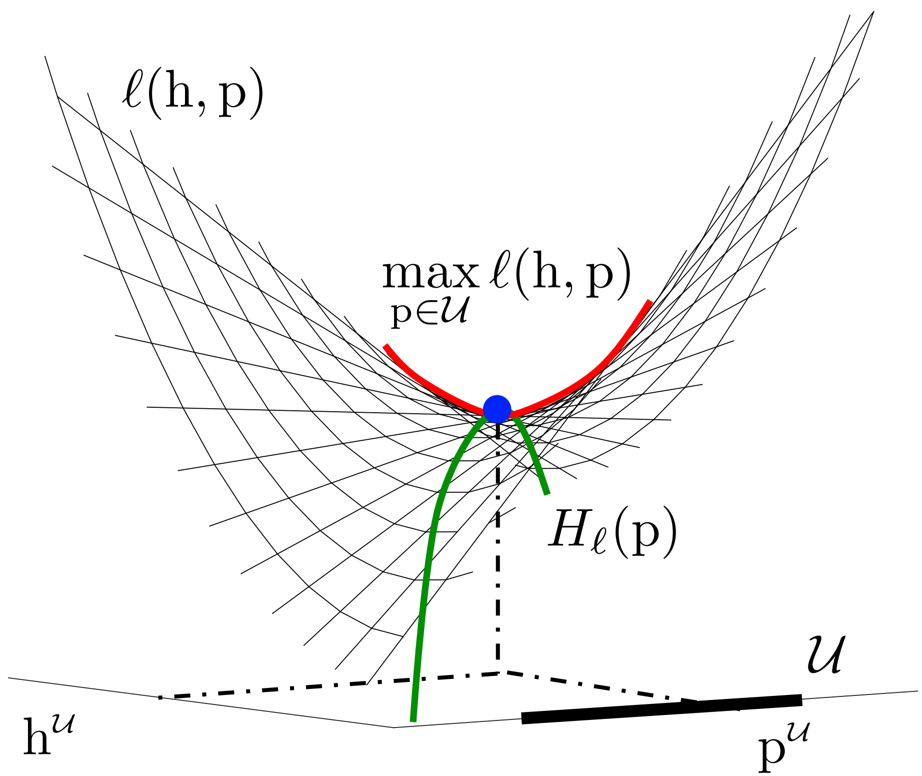
$$H_\ell(\mathrm{p}^*) = \ell(\mathrm{h}_{\mathrm{Bayes}}, \mathrm{p}^*),\ \mathrm{h}_{\mathrm{Bayes}} \text{ Bayes classifier}$$



Maximum entropy and minimax rules

$$\mathrm{p}^{\mathcal{U}} \in \arg\max_{\mathrm{p}\in\mathcal{U}} H_\ell(\mathrm{p}) \quad \text{Maximum entropy distribution}$$

$$\mathrm{h}^{\mathcal{U}} \in \arg\min_{\mathrm{h}\in\mathrm{T}(\mathcal{X},\mathcal{Y})} \left(\max_{\mathrm{p}\in\mathcal{U}} \ell(\mathrm{h},\mathrm{p})\right) \quad \text{Minimax risk classifier}$$

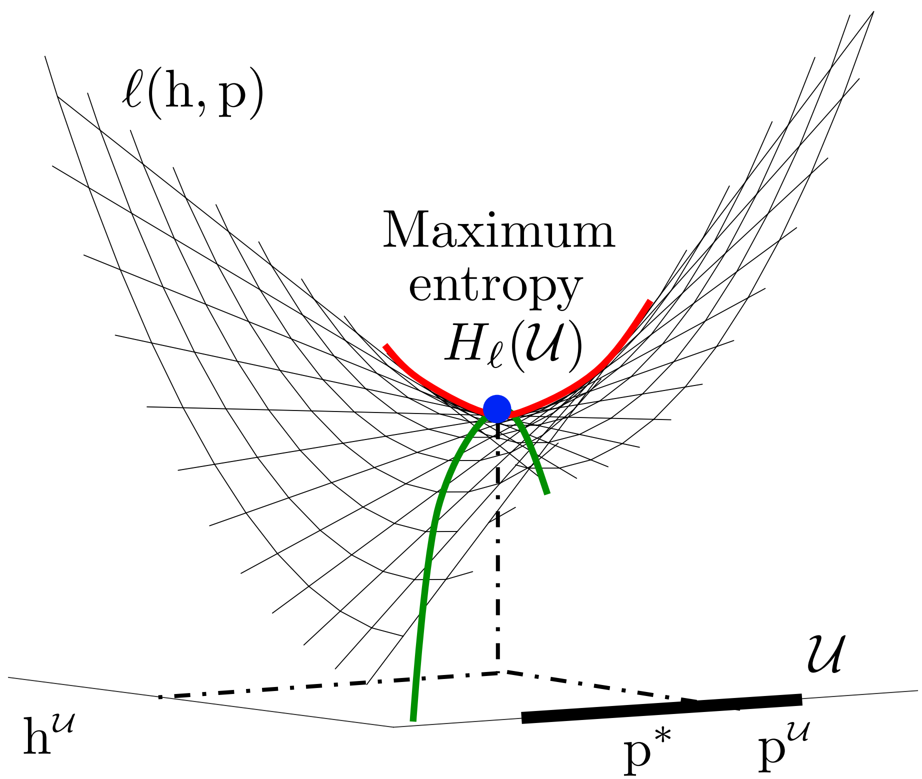


Maximum entropy and minimax rules

Theorem [S. Mazuelas, Y. Shen, A. Perez, IEEE Trans. Inf. Theory, 2022]

$$H_\ell(\mathcal{U}) = \max_{\mathrm{p}\in\mathcal{U}} H_\ell(\mathrm{p}) = \ell(\mathrm{h}^{\mathcal{U}}, \mathrm{p}^{\mathcal{U}}) = \min_{\mathrm{h}\in\mathrm{T}(\mathcal{X},\mathcal{Y})} \Big(\max_{\mathrm{p}\in\mathcal{U}} \ell(\mathrm{h},\mathrm{p}) \Big)$$

$$\mathrm{h}^{\mathcal{U}} \in \arg\min_{\mathrm{h}} \ell(\mathrm{h}, \mathrm{p}^{\mathcal{U}}) \qquad \mathrm{p}^{\mathcal{U}} \in \arg\max_{\mathrm{p}} \ell(\mathrm{h}^{\mathcal{U}}, \mathrm{p})$$

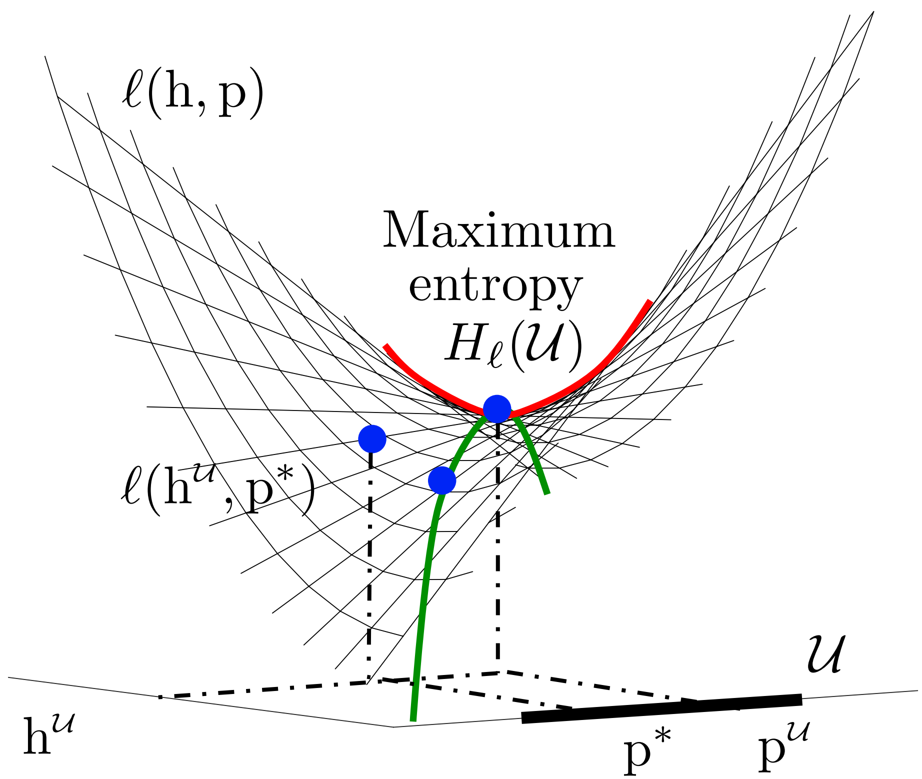


Maximum entropy and minimax rules

<u>Theorem</u> [S. Mazuelas, Y. Shen, A. Perez, IEEE Trans. Inf. Theory, 2022]

$$H_\ell(\mathcal{U}) = \max_{\mathrm{p}\in\mathcal{U}} H_\ell(\mathrm{p}) = \ell(\mathrm{h}^{\mathcal{U}}, \mathrm{p}^{\mathcal{U}}) = \min_{\mathrm{h}\in\mathrm{T}(\mathcal{X},\mathcal{Y})} \left(\max_{\mathrm{p}\in\mathcal{U}} \ell(\mathrm{h}, \mathrm{p}) \right)$$

$$\mathrm{p}^* \in \mathcal{U} \Rightarrow H_\ell(\mathrm{p}^*) \leq \ell(\mathrm{h}^{\mathcal{U}}, \mathrm{p}^*) \leq H_\ell(\mathcal{U})$$



Outline

- Introduction
- Generalized maximum entropy for supervised classification
- Minimax risk classifiers (MRCs)

Minimax Risk Classifiers (MRCs)

$$\mathscr{P}_{\text{MRC}}: \quad \min_{\text{h}\in T(\mathcal{X},\mathcal{Y})} \max_{\text{p}\in\mathcal{U}} \ell(\text{h},\text{p})$$

$$\mathcal{U} = \left\{\text{p} \in \Delta(\mathcal{X}\times\mathcal{Y}) : \ |\mathbb{E}_{\text{p}}\{\Phi\} - \boldsymbol{\tau}| \preceq \boldsymbol{\lambda}\right\}$$

$\Phi : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^m$ feature mapping

$$\Phi(x,y) = \mathbf{e}_y \otimes \boldsymbol{\Psi}(x), \quad \boldsymbol{\Psi}(x) = [\psi_{v_1}(x), \psi_{v_2}(x), \ldots, \psi_{v_D}(x)]^{\text{T}}$$

e.g., random features $\mathbb{E}_{p(v)}\{\psi_v(x)\psi_v(x')\} = k(x,x')$

$$\boldsymbol{\tau}_n = \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i,y_i)$$

$\boldsymbol{\lambda}$ confidence vector at level $1-\delta \ \Rightarrow \ \text{p}^* \in \mathcal{U}$ w.p. $\geq 1-\delta$

MRCs Learning. Representer Theorem

<u>Theorem (0-1 loss)</u> [S. Mazuelas, A. Zanoni, A. Perez, NeurIPS 2021]

Let $\boldsymbol{\mu}^* \in \mathbb{R}^m$ be a solution of

$$\min_{\boldsymbol{\mu} \in \mathbb{R}^m} 1 - \boldsymbol{\tau}^{\mathrm{T}}\boldsymbol{\mu} + \varphi(\boldsymbol{\mu}) + \boldsymbol{\lambda}^{\mathrm{T}}|\boldsymbol{\mu}|$$

where $\varphi(\boldsymbol{\mu}) = \max_{x \in \mathcal{X}, \mathcal{C} \subseteq \mathcal{Y}} \frac{\sum_{y \in \mathcal{C}} \Phi(x,y)^{\mathrm{T}}\boldsymbol{\mu} - 1}{|\mathcal{C}|}$.

If $\mathrm{h}^{\mathcal{U}} \in T(\mathcal{X}, \mathcal{Y})$ satisfies

$$\mathrm{h}^{\mathcal{U}}(y|x) \geq \Phi(x,y)^{\mathrm{T}}\boldsymbol{\mu}^* - \varphi(\boldsymbol{\mu}^*), \ \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

Then

$$\mathrm{h}^{\mathcal{U}} \in \arg \min_{\mathrm{h} \in T(\mathcal{X}, \mathcal{Y})} \max_{\mathrm{p} \in \mathcal{U}} \ell(\mathrm{h}, \mathrm{p})$$

$$H_{0\text{-}1}(\mathcal{U}) = \overline{R}(\mathcal{U}) = 1 - \boldsymbol{\tau}^{\mathrm{T}}\boldsymbol{\mu}^* + \varphi(\boldsymbol{\mu}^*) + \boldsymbol{\lambda}^{\mathrm{T}}|\boldsymbol{\mu}^*|$$

Outline

- Introduction
- Generalized maximum entropy for supervised classification
- Minimax risk classifiers (MRCs)
- Performance guarantees 0-1 MRCs

MRCs Generalization Bounds

$$R_\Phi = \min_{\boldsymbol{\mu}} 1 - \mathbb{E}_{\mathrm{p}^*}\Phi(x,y)^{\mathrm{T}}\boldsymbol{\mu} + \varphi(\boldsymbol{\mu}) = 1 - \mathbb{E}_{\mathrm{p}^*}\Phi(x,y)^{\mathrm{T}}\boldsymbol{\mu}_\infty + \varphi(\boldsymbol{\mu}_\infty)$$

The MRC given by $\boldsymbol{\mu}_\infty$ is the optimal minimax rule for Φ:

$$R_\Phi \leq \overline{R}(U) \quad \forall\ \mathcal{U} = \{\mathrm{p} \in \Delta(\mathcal{X},\mathcal{Y}) :\ |\mathbb{E}_{\mathrm{p}}\{\Phi\} - \boldsymbol{\tau}| \preceq \boldsymbol{\lambda}\} \text{ s.t. } \mathrm{p}^* \in \mathcal{U}$$

<u>Theorem</u> [S. Mazuelas, M. Romero, P. Grünwald, JMLR 2023]

$$R(\mathrm{h}^{\mathcal{U}}) \leq \overline{R}(\mathcal{U}) + (|\mathbb{E}_{\mathrm{p}^*}\Phi - \boldsymbol{\tau}| - \boldsymbol{\lambda})^{\mathrm{T}}|\boldsymbol{\mu}^*|$$

$$R(\mathrm{h}^{\mathcal{U}}) \leq R_\Phi + |\mathbb{E}_{\mathrm{p}^*}\Phi - \boldsymbol{\tau}|^{\mathrm{T}}|\boldsymbol{\mu}_\infty - \boldsymbol{\mu}^*| + \boldsymbol{\lambda}^{\mathrm{T}}(|\boldsymbol{\mu}_\infty| - |\boldsymbol{\mu}^*|)$$

In particular, if $\mathbb{P}\{|\mathbb{E}_{\mathrm{p}^*}\Phi - \boldsymbol{\tau}| \preceq \boldsymbol{\lambda}\} \geq 1 - \delta$

$$R(\mathrm{h}^{\mathcal{U}}) \leq \overline{R}(\mathcal{U})$$

$$R(\mathrm{h}^{\mathcal{U}}) \leq R_\Phi + 2\boldsymbol{\lambda}^{\mathrm{T}}|\boldsymbol{\mu}_\infty|$$

w. p. at least $1 - \delta$.

MRCs Universal Consistency

<u>Theorem</u> [S. Mazuelas, M. Romero, P. Grünwald, JMLR 2023]

Let Φ_n be a feature mapping given by D_n random features corresponding with a characteristic kernel.

If $\lim_{n\to\infty} D_n = \infty$, for any underlying distribution p^* we have that

$$\lim_{n\to\infty} R_{\Phi_n} = R_{\text{Bayes}}$$

with probability 1.

If h_n is the MRC given by n samples and

$$\lim_{n\to\infty} \boldsymbol{\lambda}_n = \mathbf{0} \qquad \lim_{n\to\infty} \boldsymbol{\lambda}_n \sqrt{n/\log n} = \infty$$

for any underlying distribution p^* we have that

$$\lim_{n\to\infty} R(\mathrm{h}_n) = R_{\text{Bayes}}$$

with probability 1.

Outline

- Introduction
- Generalized maximum entropy for supervised classification
- Minimax risk classifiers (MRCs)
- Performance guarantees 0-1 MRCs
- Multidimensional adaptation to concept drift

Supervised classification under concept drift

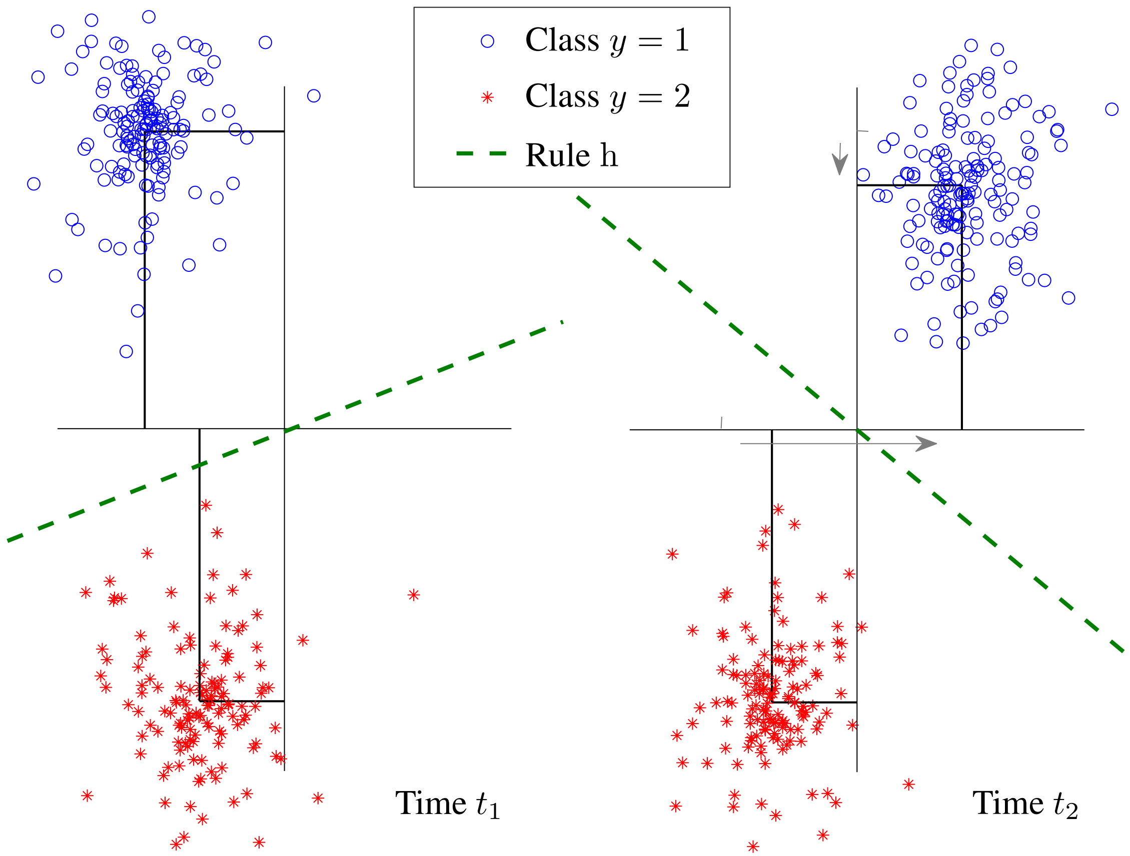
$$
\begin{array}{ccccccccc}
\mathrm{p}_1 & & \mathrm{p}_2 & \cdots & \mathrm{p}_t & & \mathrm{p}_{t+1} & \cdots \\
\wr & & \wr & & \wr & & \wr & \\
(x_1, y_1), & & (x_2, y_2), & \ldots, & (x_t, y_t), & & (x_{t+1}, y_{t+1}), & \ldots \\
\downarrow & & \downarrow & & \downarrow & & \downarrow & \\
\mathrm{h}_1 & \rightarrow & \mathrm{h}_2 & \cdots & \mathrm{h}_t & \rightarrow & \mathrm{h}_{t+1} & \cdots
\end{array}
$$

For instance,

$$\mathrm{h}_{t+1} = \mathrm{h}_t - k \nabla \ell(\mathrm{h}_t, (x_t, y_t))$$

$k \in \mathbb{R}$ accounts for a global rate of change

Supervised classification under concept drift



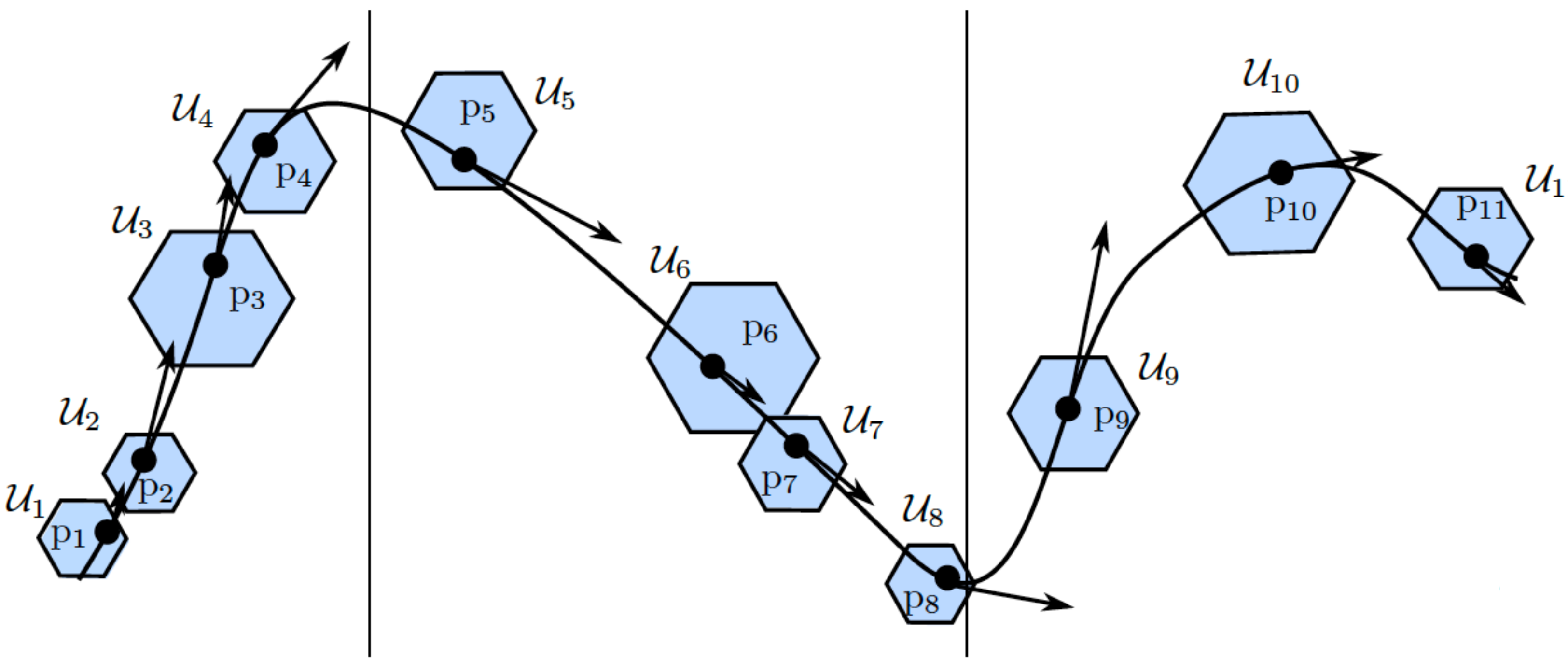
$$\mathrm{h}_{t+1} = \mathrm{h}_t - \boxed{k} \nabla \ell(\mathrm{h}_t, (x_t, y_t))$$

$k \in \mathbb{R}$ accounts for a global rate of change

MRCs for concept drift adaptation

$$\mathcal{U}_t = \left\{ \mathrm{p} \in \Delta(\mathcal{X} \times \mathcal{Y}) : \ |\mathbb{E}_{\mathrm{p}}\{\Phi\} - \widehat{\boldsymbol{\tau}}_t| \preceq \boldsymbol{\lambda}_t \right\}$$

$$\widehat{\boldsymbol{\tau}}_t \approx \widehat{\boldsymbol{\tau}}_{t-1} \in \mathbb{R}^m$$



Multidimensional Adaptation

[V. Alvarez, S. Mazuelas, J. A. Lozano, ICML 2022]
[V. Alvarez, S. Mazuelas, J. A. Lozano, NeurIPS 2023]

Tracking underlying distribution: obtain $\widehat{\boldsymbol{\tau}}_t, \boldsymbol{\lambda}_t$

Dynamical system: model the evolution of each component of $\boldsymbol{\tau}_t$

$$\boldsymbol{\eta}_{t,i} = \mathbf{H}_t \boldsymbol{\eta}_{t-1,i} + \mathbf{w}_{t,i}$$
$$\Phi_i(x_t, y_t) = \tau_{t,i} + v_{t,i}$$

with $\boldsymbol{\eta}_{t,i} = \left[\tau_{t,i}, \tau'_{t,i}, \tau''_{t,i}, \ldots, \tau^{k)}_{t,i}\right]^{\mathrm{T}}$

Unbiased linear estimator of $\boldsymbol{\eta}_{t,i}$ with minimum MSE

$$\widehat{\boldsymbol{\eta}}_{t,i} = \mathbf{H}_t \widehat{\boldsymbol{\eta}}_{t-1,i} - \mathbf{k}_{t,i}(\widehat{\tau}_{t-1,i} - \Phi_i(x_{t-1}, y_{t-1}))$$

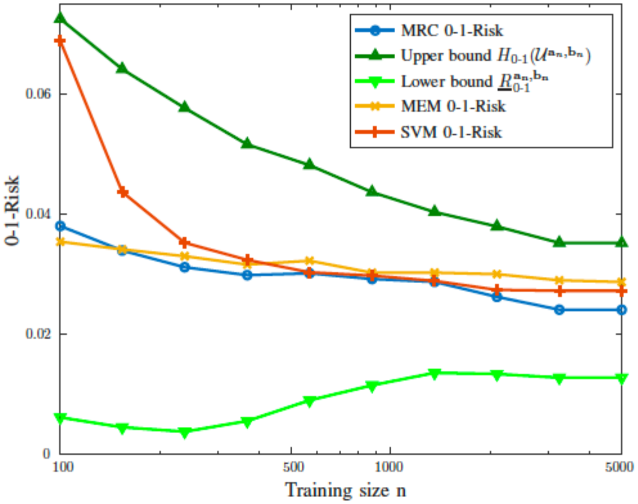
$\mathbf{k}_{t,1}, \mathbf{k}_{t,2}, \ldots, \mathbf{k}_{t,m}$ account for multidimensional time changes

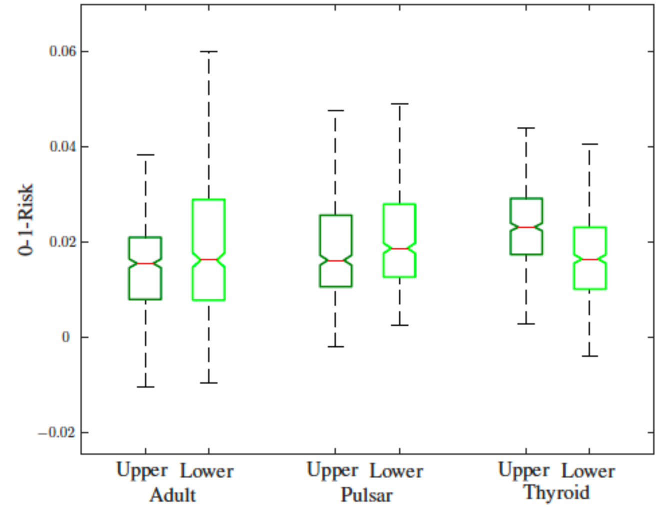
Performance guarantees

Error probability $R(\mathrm{h}_t) \leq R(\mathcal{U}_t)$

Accumulated mistakes $\sum_{t=1}^{T} \mathbb{I}\left\{\hat{y}_t \neq y_t\right\} \leq \sum_{t=1}^{T} R(\mathcal{U}_t) + \sqrt{2T \log \frac{1}{\delta}}$

Some numerical results





https://github.com/MachineLearningBCAM/MRCpy https://machinelearningbcam.github.io/MRCpy/

Algorithm	Weather	Elec2	German	Chess	Usenet1	Email	Poker
DWM	**30.0**	36.3	41.2	35.2	36.3	39.5	22.0
Projectron	30.7	36.8	40.0	35.1	49.1	48.2	22.6
Unid. AMRC	31.3	40.1	30.3	38.3	46.3	48.4	39.4
AMRC	32.3	35.8	30.3	**27.7**	35.7	43.7	**21.9**
Det. AMRC	**30.0**	**33.9**	**30.0**	33.4	**32.0**	**33.9**	**21.9**

Conclusions

- Learning methodology that relies on expectation estimates and not on specific training samples
- MRCs can minimize the worst-case expected 0-1 loss over general classification rules, and provide performance guarantees during learning
- Described MRCs' generalization bounds and universal consistency
- The proposed methods can lead to algorithms that seamlessly address distribution shifts

References

- S. Mazuelas, A. Zanoni, and A. Perez, “Minimax classification with 0-1 loss and performance guarantees,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 302–312.
- S. Mazuelas, Y. Shen, and A. Perez, “Generalized maximum entropy for supervised classification,” *IEEE Transactions on Information Theory*, vol. 68, no. 4, pp. 2530–2550, Apr. 2022.
- V. Alvarez, S. Mazuelas, and J. A. Lozano, “Minimax classification under concept drift with multidimensional adaptation and performance guarantees,” *International Conference on Machine Learning (ICML)* 2022.
- S. Mazuelas, M. Romero, and P. Grünwald, “Minimax classification with 0-1 loss,” *Journal of Machine Learning Research (JMLR*), vol. 24, no. 208, pp. 1-48, 2023.
- V. Alvarez, S. Mazuelas, and J. A. Lozano, “Minimax Forward and Backward Learning of Evolving Tasks with Performance Guarantees,” *Advances in Neural Information Processing Systems (NeurIPS).* 2023.
- J. Segovia, S. Mazuelas, and A. Liu, “Double-weighting for covariate shift adaptation,” *International Conference on Machine Learning (ICML)* 2023.