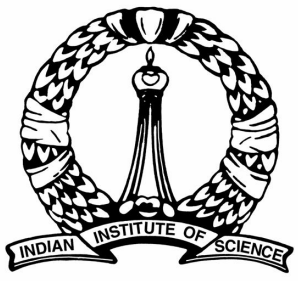
On the Consistency of Output Code Based Learning Algorithms for Multiclass Learning Problems

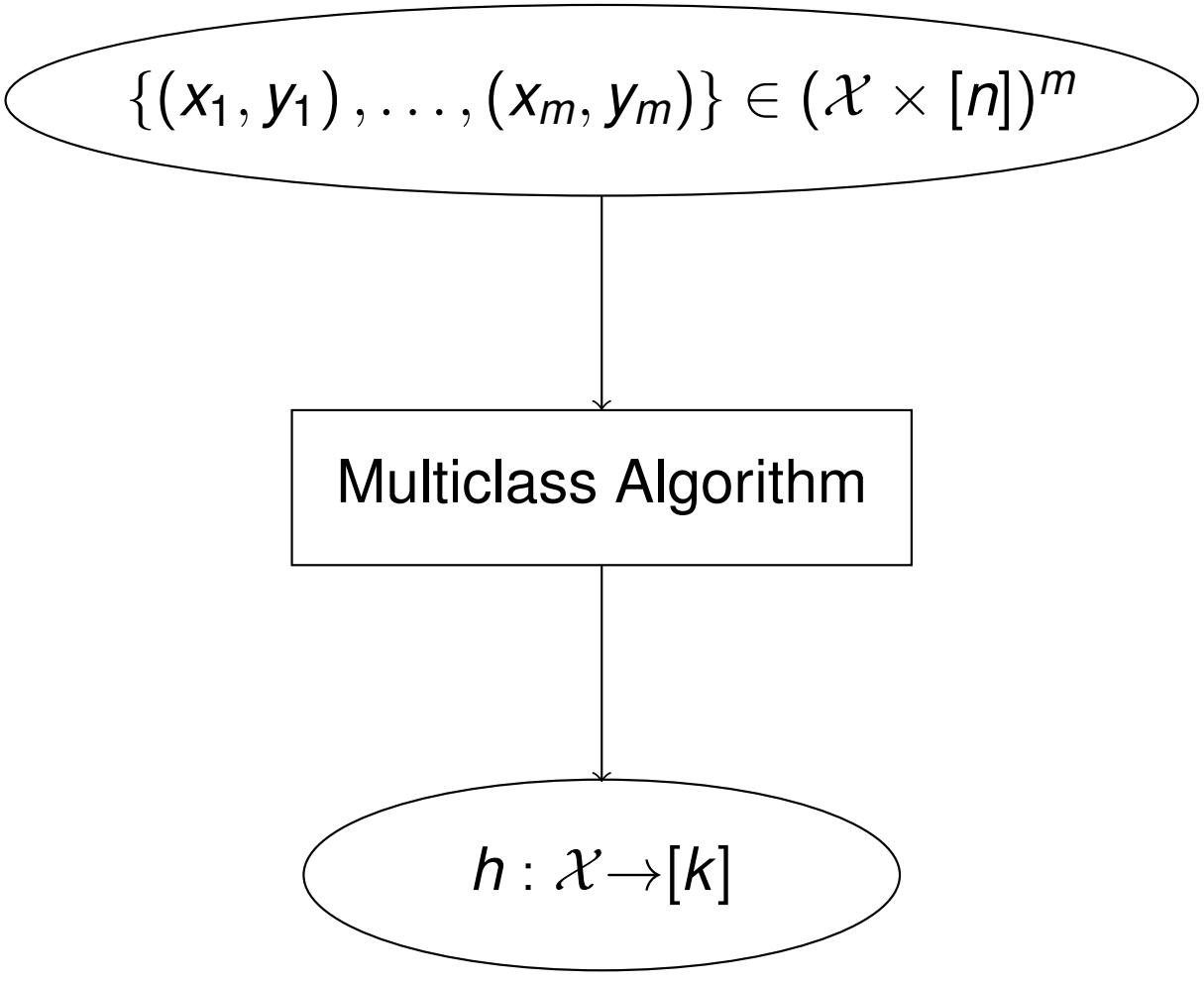
Harish G. Ramaswamy[1], S.B.Balaji[1],
Shivani Agarwal[1] and Robert Williamson [2]



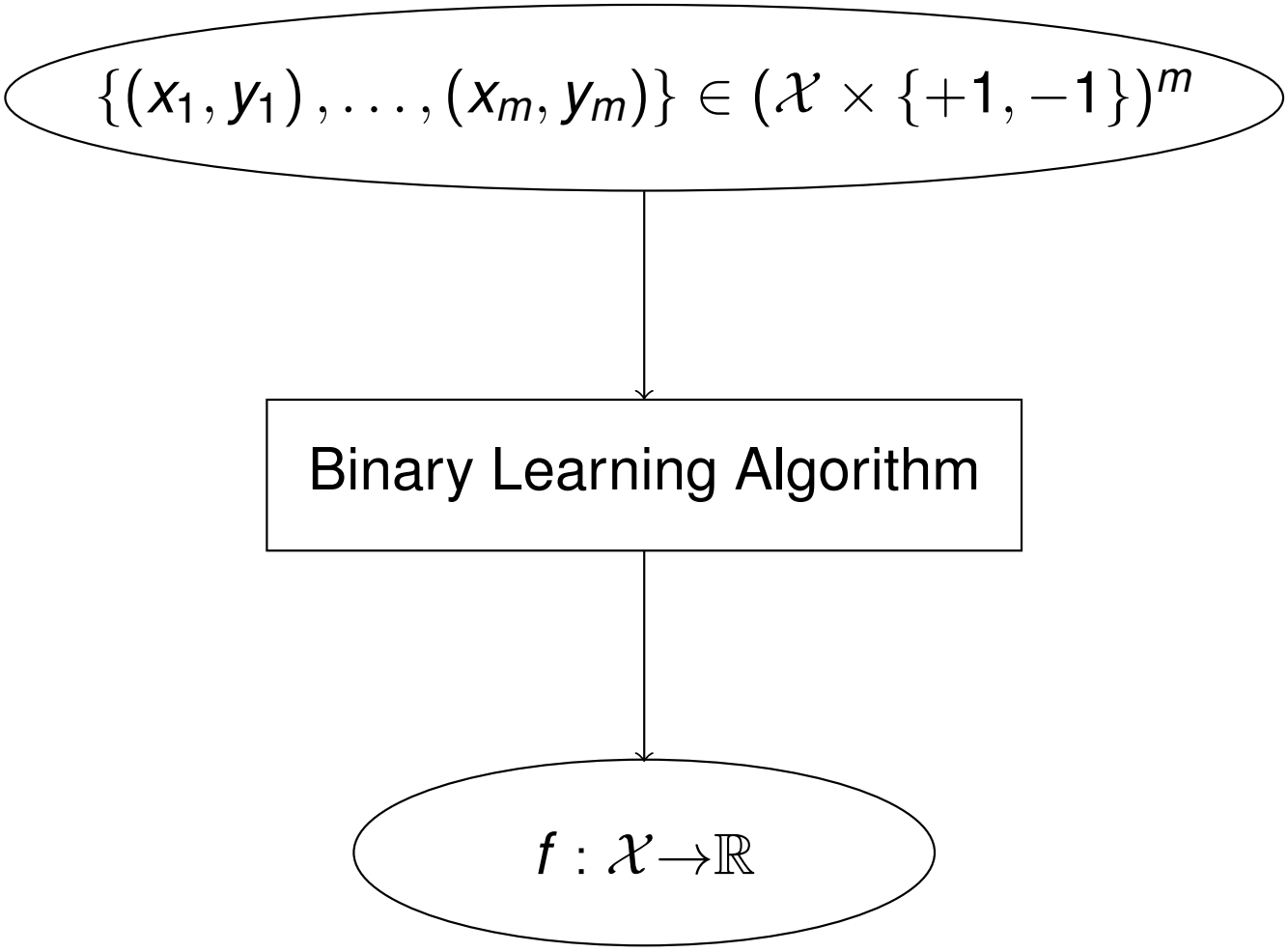
[1]Indian Institute of Science, India.

[2]NICTA and ANU, Australia.

Multiclass Learning Algorithm

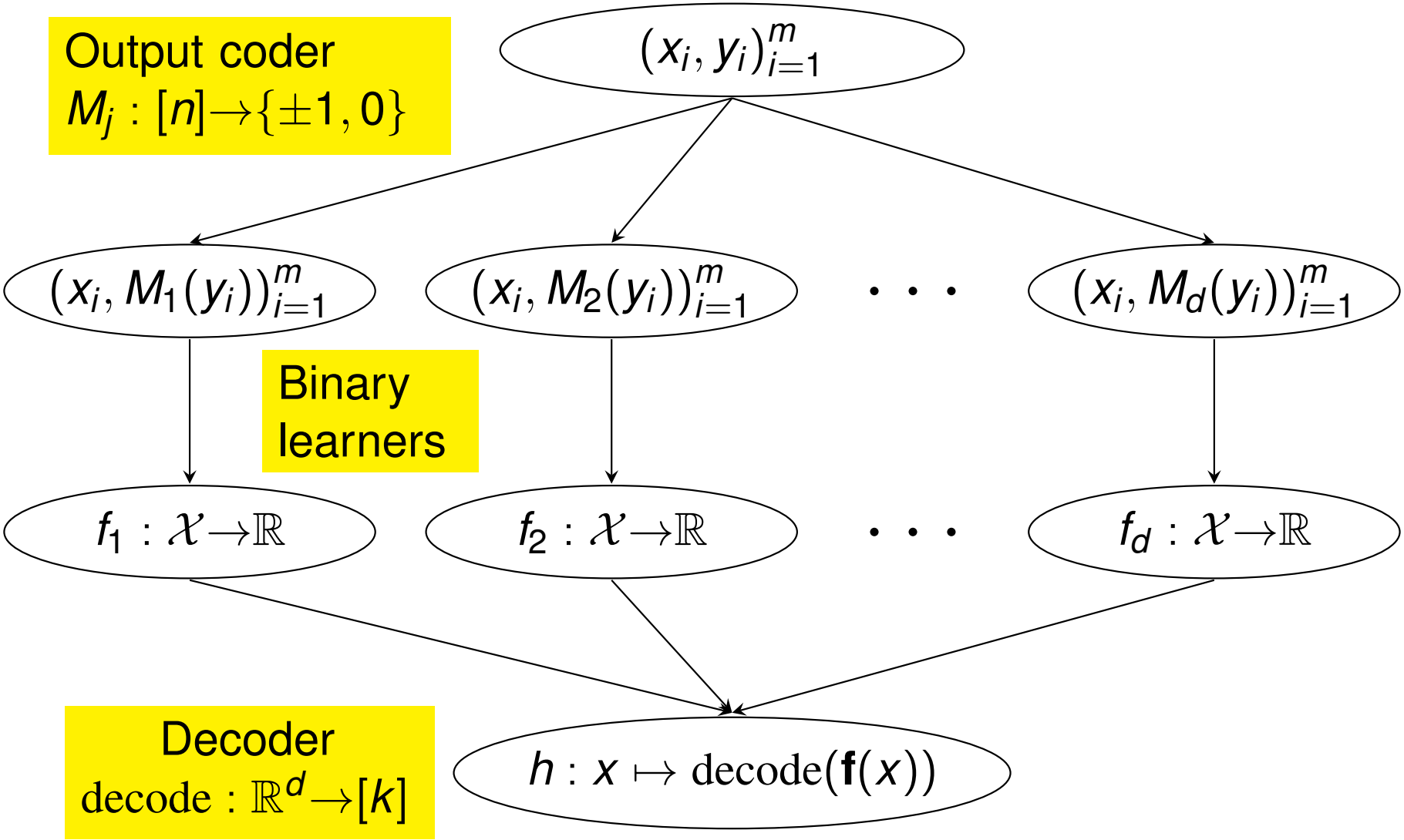


Binary Learning Algorithm



Output Coding Methods

Dietterich & Bakiri (1995), Allwein et al. (2000)



Problem

- Multiclass learning problems often solved using binary methods via output coding methods.
- Question: What are the *properties of the 3 components* of output coding methods that guarantee *'consistency' of the overall method* ?

The Coding Matrix : $\mathbf{M} \in \{-1, 0, +1\}^{n \times d}$

$$\mathbf{M} = \begin{bmatrix} M_1(1) & M_2(1) & \dots & M_d(1) \\ M_1(2) & M_2(2) & \dots & M_d(2) \\ \vdots & \vdots & \ddots & \vdots \\ M_1(n) & M_2(n) & \dots & M_d(n) \end{bmatrix}$$

One vs All and All Pairs code matrices for $n = 4$

$$\mathbf{M}^{\text{OvA}} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 \end{bmatrix} \qquad (d = n)$$

$$\mathbf{M}^{\text{all-pairs}} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & -1 & -1 \end{bmatrix} \qquad \left(d = \binom{n}{2}\right)$$

The Binary Learner

Surrogate: $\phi_1, \phi_{-1} : \mathbb{R} \to \mathbb{R}_+$

$$S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \{\pm 1\})^m$$

$$f_m = \operatorname{argmin}_{f \in \mathcal{F}_m} \sum_{i=1}^{m} \phi_{y_i}(f(x_i))$$

Strictly Proper Composite Surrogate Losses

Reid & Williamson (2010)

There exists an invertible *link function* $\lambda : [0,1] \to \mathbb{R}$ s.t.

$$\lambda(\eta) = \text{argmin}_{v \in \mathcal{V}} \; \eta\, \phi_1(v) + (1-\eta)\, \phi_{-1}(v)$$

Loss	$\mathcal{V}$	$\phi_1(v)$	$\phi_{-1}(v)$	$\lambda(\eta)$	$\lambda^{-1}(v)$
Logistic	$\overline{\mathbb{R}}$	$\ln(1+e^{-v})$	$\ln(1+e^{v})$	$\ln\left(\frac{\eta}{1-\eta}\right)$	$\frac{1}{1+e^{-v}}$
Exponential	$\overline{\mathbb{R}}$	e^{-v}	e^{v}	$\frac{1}{2}\ln\left(\frac{\eta}{1-\eta}\right)$	$\frac{1}{1+e^{-2v}}$
Least-squares	$[-1,1]$	$(v-1)^2$	$(v+1)^2$	$2\eta - 1$	$\frac{v+1}{2}$

The Decoder

$$\text{decode} : \mathbb{R}^d \to [k]$$

One vs All ($d = n$) and All Pairs ($d = \binom{n}{2}$) decoders:

$$\begin{aligned} \text{decode}^{\text{OvA}}(\mathbf{u}) &= \text{argmax}_i \, u_i \\ \text{decode}^{\text{all-pairs}}(\mathbf{u}) &= \text{argmax}_i \sum_{j=1}^{n} \mathbf{1}\left(u_{i,j} > \frac{1}{2}\right); \quad u_{i,j} = 1 - u_{j,i} \end{aligned}$$

Evaluation Metrics 1: Zero-One loss

$$\mathbf{L}^{0\text{-}1} \in \mathbb{R}_{+}^{n \times n}$$

			
	0	1	1
	1	0	1
	1	1	0

Evaluation Metrics 2: DCG loss

$$\mathbf{L}^{\mathrm{DCG}} \in \mathbb{R}_{+}^{2^r \times r!}$$

	σ_1	σ_2	$\ldots$	$\sigma_{r!}$
$(0,\ldots,0)$				
$(0,\ldots,1)$				
$\vdots$	$\mathbf{L}^{\mathrm{DCG}}_{y,\sigma} = C - \sum_{i=1}^{r} \frac{2^{y_i}-1}{\log(1+\sigma(i))}$			
$(1,\ldots,1)$				

General Multiclass Loss

$$\mathbf{L} \in \mathbb{R}_{+}^{n \times k}$$

Classes \ Predictions	1	2	...	k
1	L_{11}	L_{12}	...	L_{1k}
2	L_{21}	L_{22}	...	L_{2k}
⋮	⋮	⋮	⋱	⋮
n	L_{n1}	L_{n2}	...	L_{nk}

Objective of the Multiclass Learning Algorithm

Given $S \in (\mathcal{X} \times [n])^m$ drawn i.i.d. from D

Find $h : \mathcal{X} \to [k]$ minimizing

$$\mathbf{E}_{(X,Y)\sim D}\left[\mathbf{L}_{Y,h(X)}\right]$$

Consistency

Sequence of functions $h_i : \mathcal{X} \to [k]$ s.t.

$$\mathbf{E}_{(X,Y)\sim D}\left[\mathbf{L}_{Y,h_i(X)}\right] \longrightarrow \inf_h \mathbf{E}_{(X,Y)\sim D}\left[\mathbf{L}_{Y,h(X)}\right]$$

Output Coding as a Multiclass Surrogate

A multiclass surrogate $\psi : [n] \times \mathbb{R}^d \to \mathbb{R}_+$.

$$h^* = \operatorname{argmin}_{h \in \mathcal{H}} \sum_{i=1}^{m} \psi(y_i, h(x_i))$$

Surrogate corresponding to a $(\mathbf{M}, \phi_1, \phi_{-1})$ output coding method:

$$\psi(y, \mathbf{u}) = \sum_{j=1}^{d} \mathbf{1}(M_j(y) = 1)\phi_1(u_j) + \mathbf{1}(M_j(y) = -1)\phi_{-1}(u_j)$$

Output Coding as a Multiclass Surrogate

A multiclass surrogate $\psi : [n] \times \mathbb{R}^d \to \mathbb{R}_+$.

$$h^* = \text{argmin}_{h \in \mathcal{H}} \sum_{i=1}^{m} \psi(y_i, h(x_i))$$

Surrogate corresponding to a $(\mathbf{M}, \phi_1, \phi_{-1})$ output coding method:

$$\psi(y, \mathbf{u}) = \sum_{j=1}^{d} \mathbf{1}(M_j(y) = 1)\phi_1(u_j) + \mathbf{1}(M_j(y) = -1)\phi_{-1}(u_j)$$

One vs All

Theorem (Strictly proper composite surrogates)

If (ϕ_1, ϕ_{-1}) *is strictly proper composite binary, then* $(\mathbf{M}^{\mathrm{OvA}}, \phi_1, \phi_{-1}, \mathrm{decode}^{\mathrm{OvA}} \circ \lambda^{-1})$ *output coding algorithm is consistent w.r.t zero-one loss.*

Theorem (Lee et al., 2004 – Hinge loss)

$(\mathbf{M}^{\mathrm{OvA}}, \phi_1^{\mathrm{hinge}}, \phi_{-1}^{\mathrm{hinge}})$ *output coding algorithm is not consistent w.r.t zero-one loss for any decoder.*

All Pairs

Theorem (Strictly proper composite surrogates)

If (ϕ_1, ϕ_{-1}) *is strictly proper composite binary, then* $(\mathbf{M}^{\text{all-pairs}}, \phi_1, \phi_{-1}, \text{decode}^{\text{all-pairs}} \circ \lambda^{-1})$ *output coding algorithm is consistent w.r.t zero-one loss.*

Theorem (Hinge loss)

$(\mathbf{M}^{\text{all-pairs}}, \phi_1^{\text{hinge}}, \phi_{-1}^{\text{hinge}}, \text{decode}^{\text{all-pairs}} \circ \lambda^{-1})$ *output coding algorithm is consistent w.r.t zero-one loss.*

Column Space Condition

A coding matrix $\mathbf{M} \in \{\pm 1\}^{n \times d}$ satisfies the *column space condition* with $\mathbf{L} \in \mathbb{R}_{+}^{n \times k}$ if

$$\mathrm{Span}([\mathbf{M}, \mathbf{e}]) \supseteq \mathrm{Span}(\mathbf{L}).$$

Consistency w.r.t. General Losses

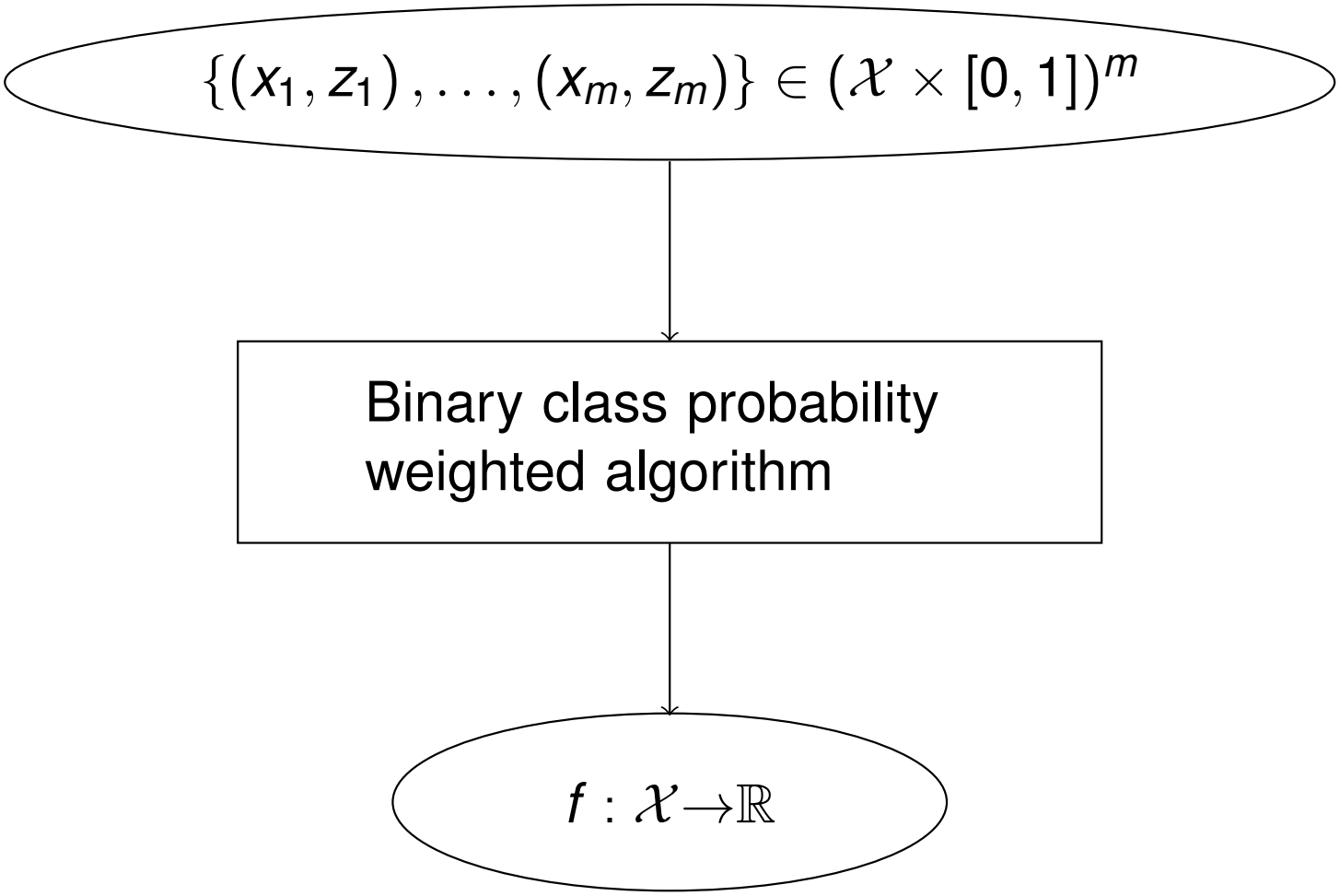
Theorem

If

- $\mathbf{M} \in \{\pm 1\}^{n \times d}$ *satisfies the column space condition with* $\mathbf{L} \in \mathbb{R}_+^{n \times k}$*, and*
- (ϕ_1, ϕ_{-1}) *is strictly proper composite*

then $(\mathbf{M}, \phi_1, \phi_{-1}, \text{decode}^{\mathbf{M},\mathbf{L},\lambda})$ *output coding algorithm is consistent w.r.t.* $\mathbf{L}$*.*

Binary Class Probability Weighted Learning Algorithm



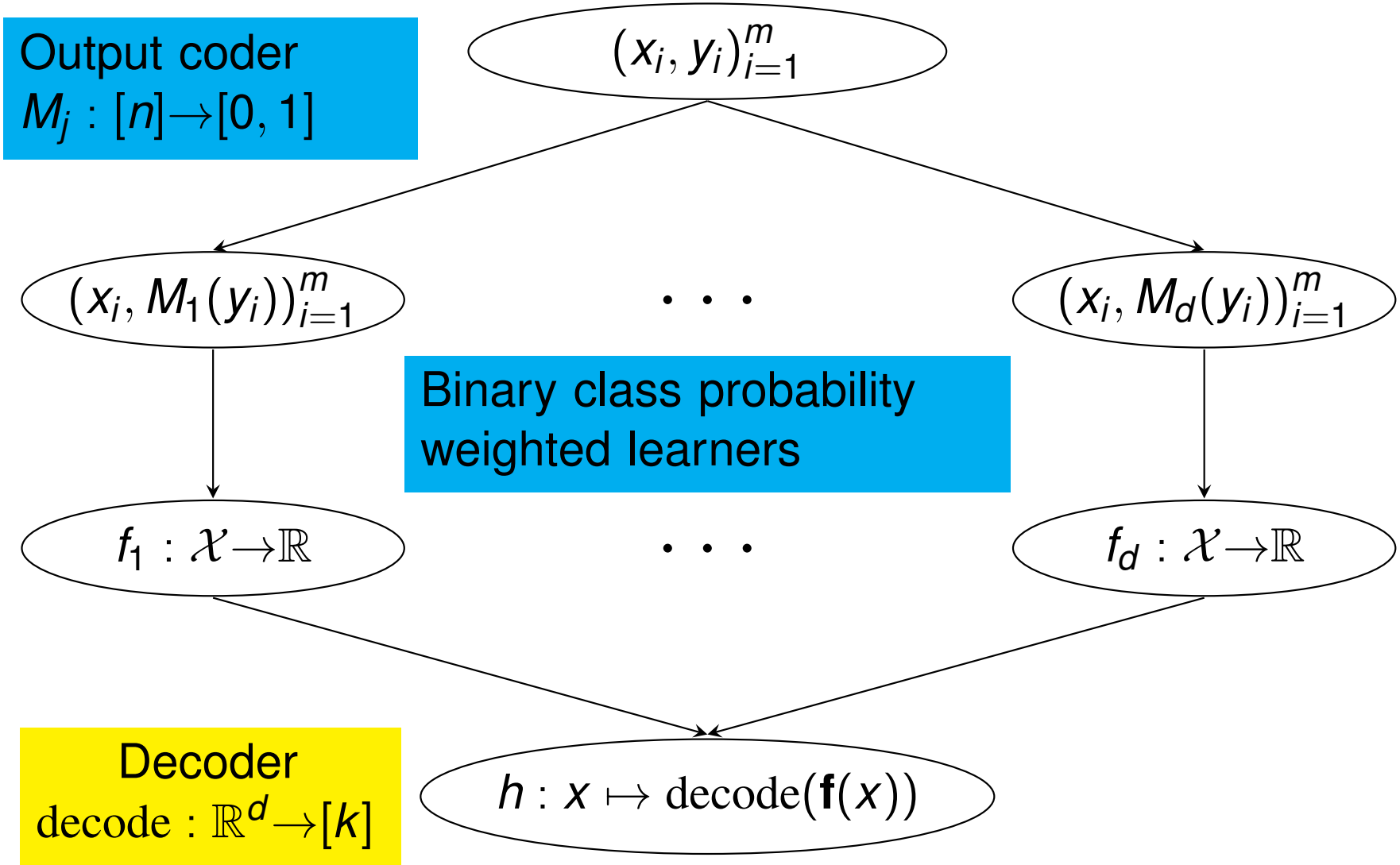
The Binary Class Probability Weighted Learner

Surrogate: $\phi_1, \phi_{-1} : \mathbb{R} \to \mathbb{R}_+$

$$S = \{(x_1, z_1), \ldots, (x_m, z_m)\} \in (\mathcal{X} \times [0, 1])^m$$

$$f_m = \text{argmin}_{f \in \mathcal{F}_m} \sum_{i=1}^{m} z_i \phi_1(f(x_i)) + (1 - z_i)\phi_{-1}(f(x_i))$$

Probabilistic Output Coding



Consistency of Probabilistic Output Coding

Probabilistic code matrix $\mathbf{M} \in [0,1]^{n \times d}$

$$\mathbf{M} = \begin{bmatrix} M_1(1) & M_2(1) & \dots & M_d(1) \\ M_1(2) & M_2(2) & \dots & M_d(2) \\ \vdots & \vdots & \ddots & \vdots \\ M_1(n) & M_2(n) & \dots & M_d(n) \end{bmatrix}$$

Theorem

If

- $\mathbf{M} \in [0,1]^{n \times d}$ *satisfies the column space condition with* $\mathbf{L} \in \mathbb{R}_+^{n \times k}$*, and*
- (ϕ_1, ϕ_{-1}) *is strictly proper composite*

then $(\mathbf{M}, \phi_1, \phi_{-1}, \text{decode}^{\mathbf{M},\mathbf{L},\lambda})$ *probabilistic output coding algorithm is consistent w.r.t.* $\mathbf{L}$*.*

Application to DCG loss

$$\mathbf{L}_{y,\sigma}^{\mathrm{DCG}} = C - \sum_{i=1}^{r} \frac{2^{y_i} - 1}{\log(1 + \sigma(i))}$$

- Probabilistic output coding : r binary tasks.
- One vs All output coding : 2^r binary tasks.

Summary

- Gave general conditions for coding methods with the one-vs-all & all-pairs code matrices to be consistent w.r.t. 0-1 loss.
- Gave a column space condition on the code matrix for consistency w.r.t. general multiclass losses.
- Introduced probabilistic output coding methods which can require fewer binary problems to achieve consistency w.r.t. a general multiclass loss.