Consistency of Losses for Learning from Weak Labels

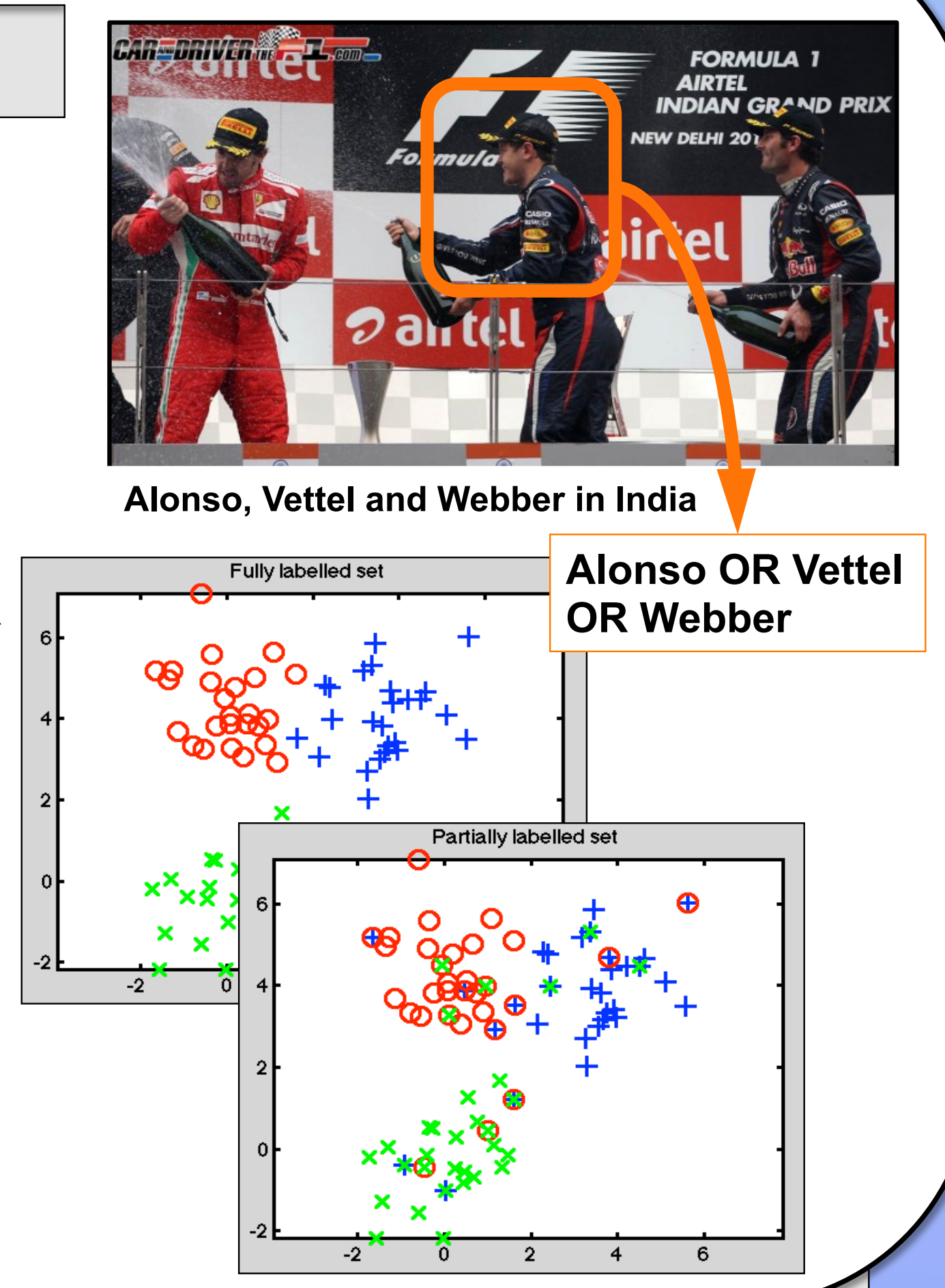
Jesús Cid Sueiro, Universidad Carlos III de Madrid, Spain
Darío García-García, Commonwealth Bank of Australia
Raúl Santos-Rodríguez, Universidad de Valencia, Spain

ECML PKDD 2014

Learning from Weak Labels

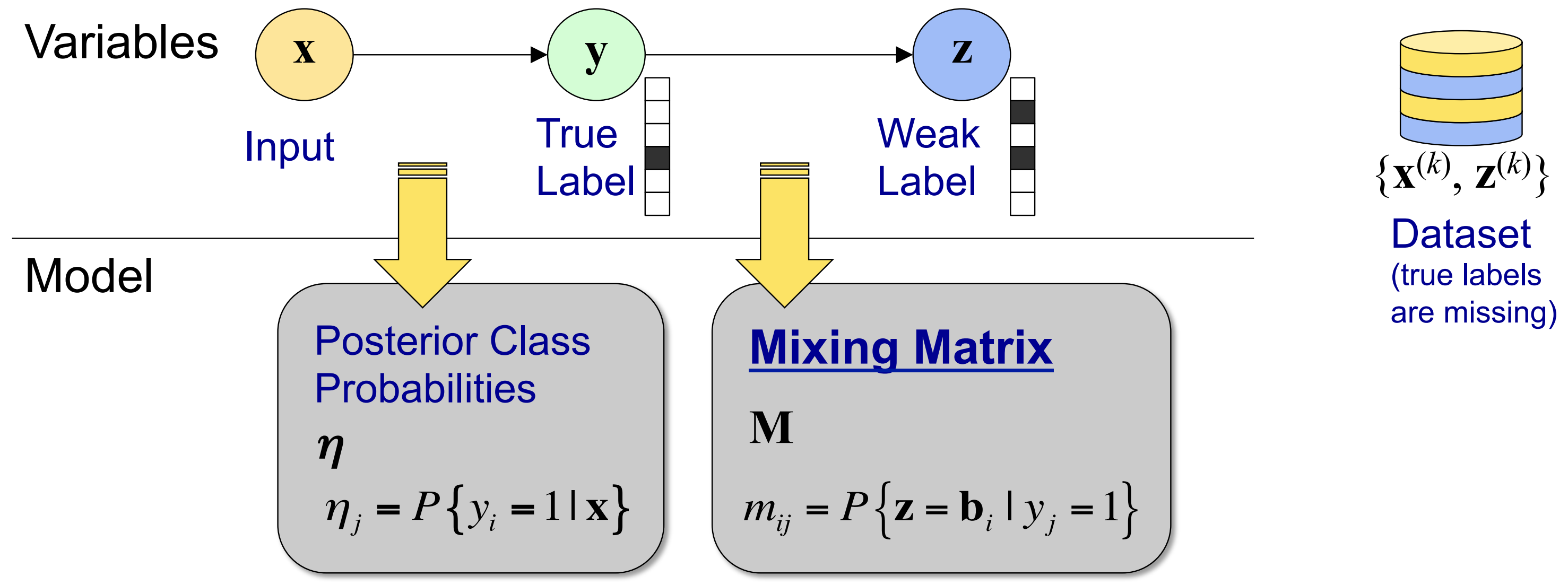
Weak Labels

- Weak labels: provide partial information about the class of the data.
  - Photo captions
  - Hierarchical classes and *superclass labels*
  - Multiple specialized annotators.
- Limitations of learning from weak labels: any loss function is effective for some weak labeling models only.
- Our goals:
  - To analyze comparatively loss functions for probability estimation, class ranking and classification from weakly labeled datasets.
  - To propose general methods to design loss functions for learning.


Alonso, Vettel and Webber in India

Alonso OR Vettel OR Webber

Model


Variables

Model

Input

True Label

Weak Label

Posterior Class Probabilities

$\boldsymbol{\eta}$

$\eta_j = P\{y_i = 1 \mid \mathbf{x}\}$

Mixing Matrix

$\mathbf{M}$

$m_{ij} = P\{\mathbf{z} = \mathbf{b}_i \mid y_j = 1\}$

$\{\mathbf{x}^{(k)}, \mathbf{z}^{(k)}\}$

Dataset (true labels are missing)

Mixing Matrix: Examples

A. True class and a noisy class

(3 classes)

True Label

Weak Label

	■□□	□■□	□□■
□□□	0	0	0
■□□	1−α	0	0
□■□	0	1−β	0
□□■	0	0	1−γ
■■□	α/2	β/2	0
■□■	α/2	0	γ/2
□■■	0	β/2	γ/2
■■■	0	0	0

Probability of observing weak label ■□■ when the true class is □□■

B. Independent noisy labels

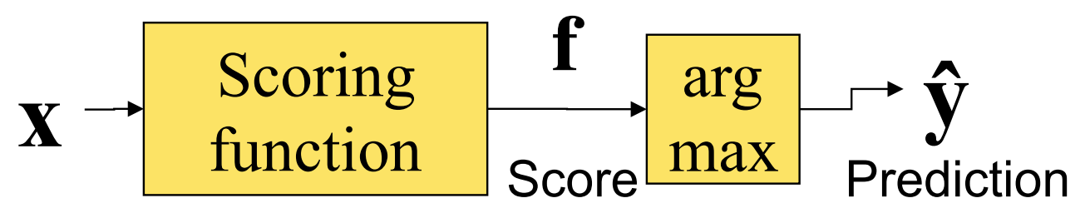
True Label

Weak Label

	■□□	□■□	□□■
□□□	0	0	0
■□□	(1−β)²	0	0
□■□	0	(1−β)²	0
□□■	0	0	(1−β)²
■■□	β (1−β)	β (1−β)	0
■□■	β (1−β)	0	β (1−β)
□■■	0	β (1−β)	β (1−β)
■■■	β²	β²	β²

Loss functions

Label-based loss vs weak loss



- Label-based loss: A function of the true label and the score

$\boxed{\tilde{\Psi}(\mathbf{y},\mathbf{f})}$ (Scalar form) $\Leftrightarrow$ $\boxed{\tilde{\boldsymbol{\Psi}}(\mathbf{f})}$ (Vector form)

- Weak loss: A function of the weak label and the score:

$\boxed{\Psi(\mathbf{z},\mathbf{f})}$ (Scalar form) $\Leftrightarrow$ $\boxed{\boldsymbol{\Psi}(\mathbf{f})}$ (Vector form)

Three types of losses

- Let $\mathbf{f}^*$ be a minimizer of the expected loss,
  - ($\mathbf{f}^*$ depends on the mixing matrix, $\mathbf{M}$)

$$\mathbf{f}^* = \arg\min_{\mathbf{f}} E_{\mathbf{z}}\left\{\hat{\Psi}(\mathbf{z},\mathbf{f}) \middle| \mathbf{x}\right\}$$

- Three types of weak losses:
  - $\mathbf{M}$-Proper loss:
    - A loss for estimating posterior class probabilities

$$\mathbf{f}^* = \boldsymbol{\eta}$$

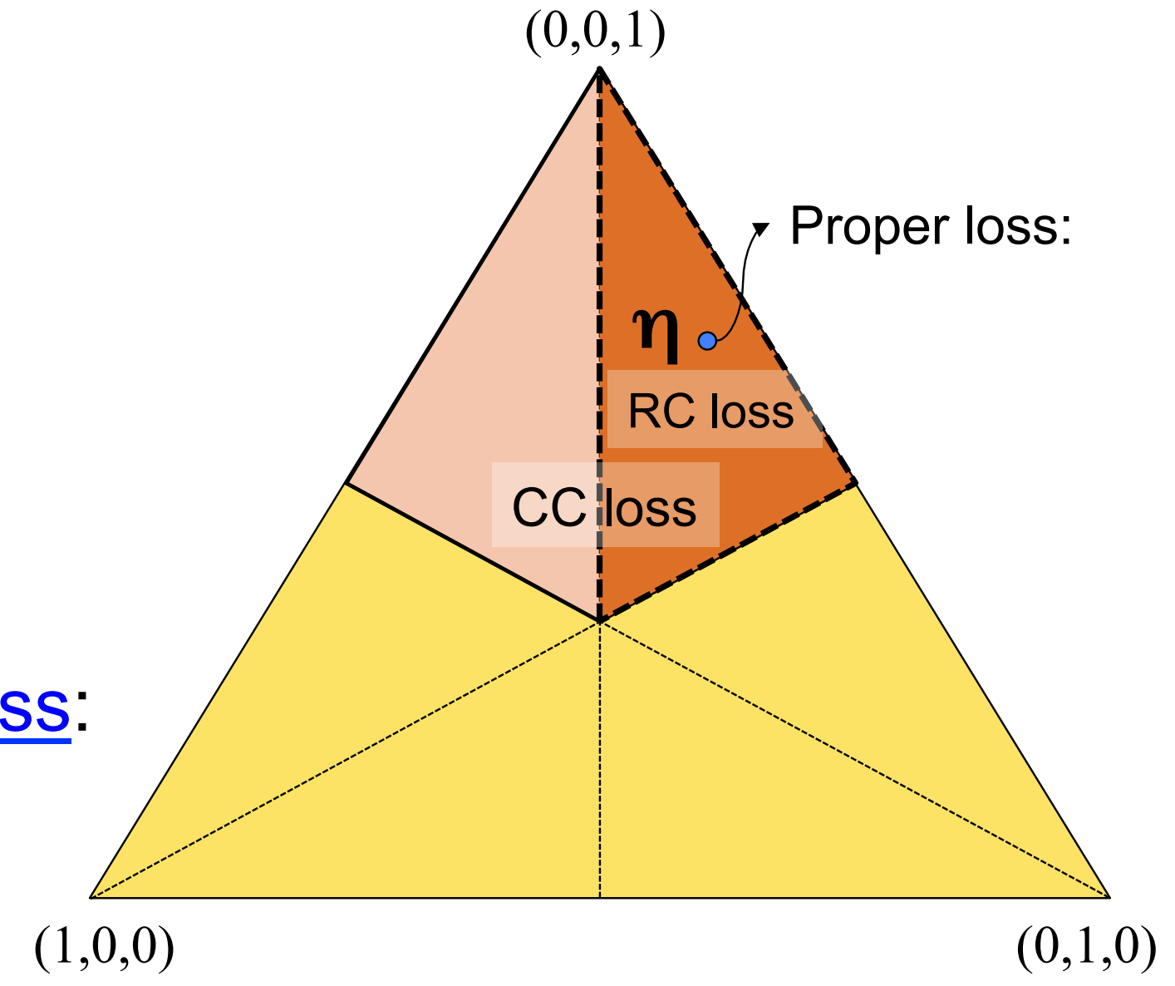
  - $\mathbf{M}$-Ranking Calibrated ($\mathbf{M}$-RC) loss:
    - A loss for ranking classes

$$f_i^* > f_j^* \Leftrightarrow \eta_i > \eta_j$$

  - $\mathbf{M}$-Classification Calibrated ($\mathbf{M}$-CC) loss:
    - A loss for minimizing errors

$$f_i^* > \max_{j \neq i} f_j^* \Leftrightarrow \eta_i > \max_{j \neq i} \eta_j$$

Example: probability triangle, 3 classes.
Possible values of $\mathbf{f}^*$:



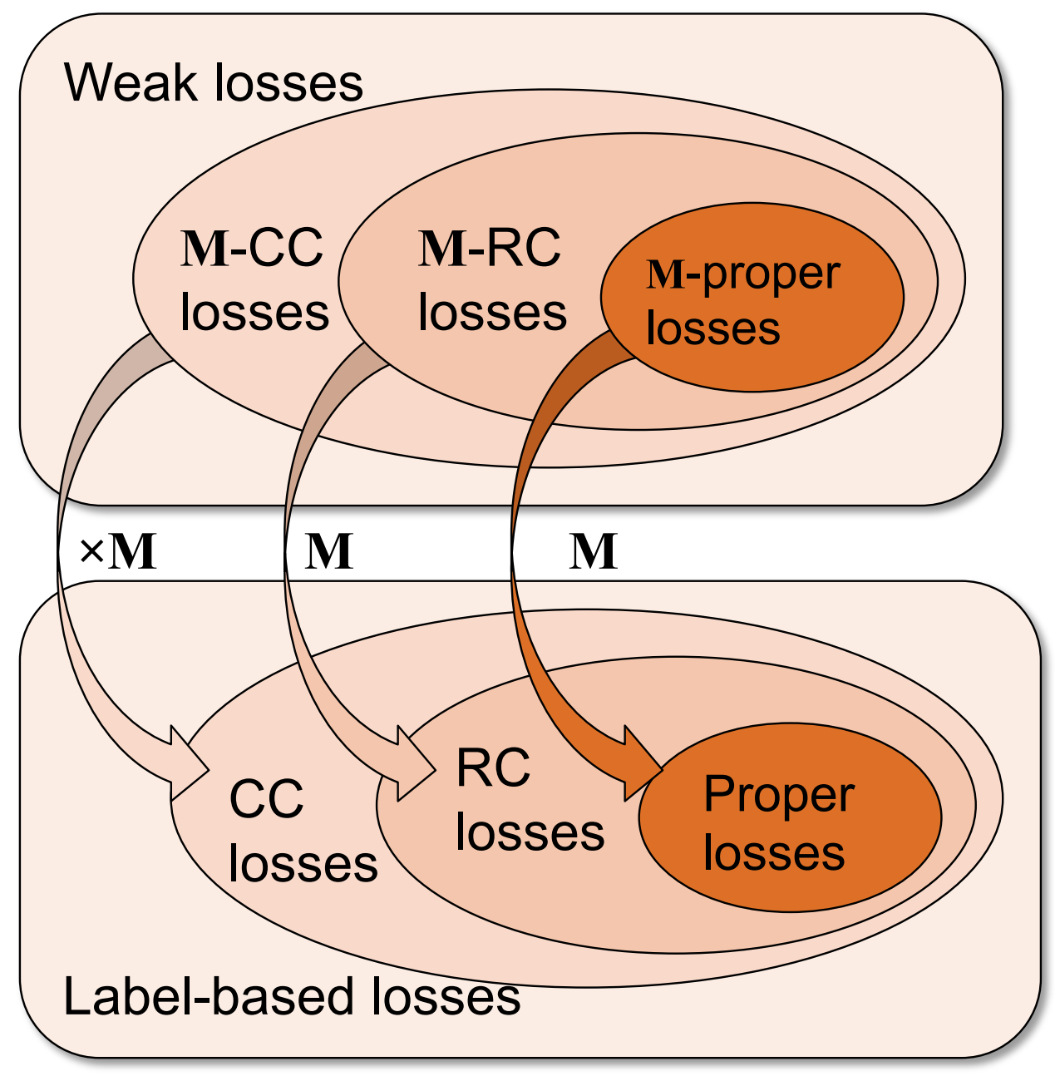
Proper/RC/CC Losses for partial labels

- A (theoretical) test to verify if a weak loss is proper, RC or CC using the label-based loss

$$\tilde{\boldsymbol{\Psi}} = \mathbf{M}^T \boldsymbol{\Psi}$$

Theorem 1:

$\boldsymbol{\Psi}$ is $\mathbf{M}$-proper iff $\tilde{\boldsymbol{\Psi}}$ is proper
$\boldsymbol{\Psi}$ is $\mathbf{M}$-RC iff $\tilde{\boldsymbol{\Psi}}$ is RC
$\boldsymbol{\Psi}$ is $\mathbf{M}$-CC iff $\tilde{\boldsymbol{\Psi}}$ is CC



Constructing Weak Losses

From label-based losses to Weak Losses

- We can construct weak losses as a linear transformation of a label-based loss

$$\boldsymbol{\Psi} = \tilde{\mathbf{Y}}^T \tilde{\boldsymbol{\Psi}}$$

- For which mixing matrices, $\mathbf{M}$, is $\boldsymbol{\Psi}$ $\mathbf{M}$-proper, $\mathbf{M}$-RC or $\mathbf{M}$-CC?

Definitions: Maximal sets for a given conventional loss

$Q_{\text{proper}}(\tilde{\mathbf{Y}}) =$ The set of all mixing matrices, M, for which $\boldsymbol{\Psi} = \tilde{\mathbf{Y}}^T \tilde{\boldsymbol{\Psi}}$ is proper

$Q_{rc}(\tilde{\mathbf{Y}})$ = The set of all mixing matrices, for which $\boldsymbol{\Psi} = \tilde{\mathbf{Y}}^T \tilde{\boldsymbol{\Psi}}$ is RC

$Q_{cc}(\tilde{\mathbf{Y}})$ = The set of all mixing matrices, for which $\boldsymbol{\Psi} = \tilde{\mathbf{Y}}^T \tilde{\boldsymbol{\Psi}}$ is CC

Characterizing maximal sets

- Main result:

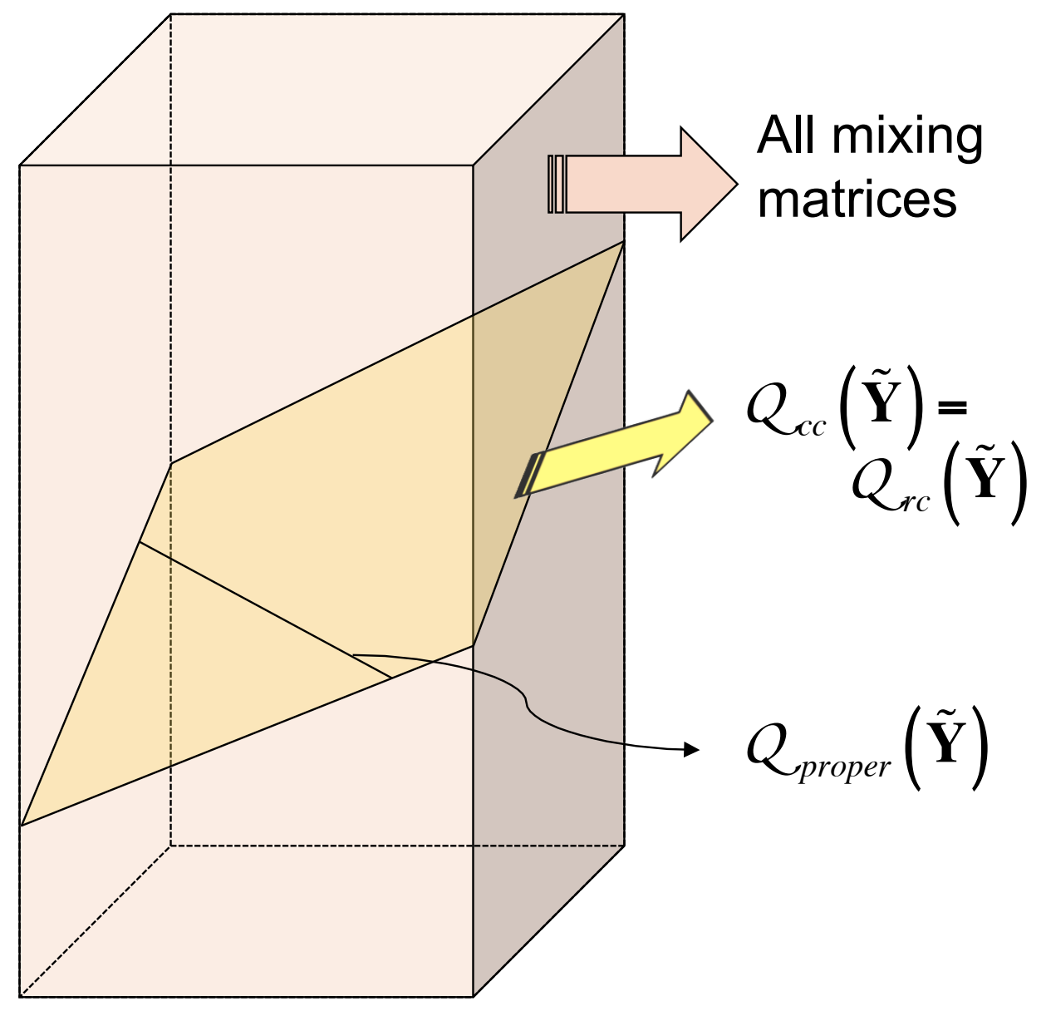
Ths. 2-4: Maximal set characterization

$$Q_{\text{proper}}(\tilde{\mathbf{Y}}) = \left\{\mathbf{M} \;\middle|\; \tilde{\mathbf{Y}}\mathbf{M} = \alpha\mathbf{I}\right\}$$

$$Q_{rc}(\tilde{\mathbf{Y}}) = Q_{cc}(\tilde{\mathbf{Y}}) = \left\{\mathbf{M} \;\middle|\; \tilde{\mathbf{Y}}\mathbf{M} = \alpha\mathbf{I} + \mathbf{1}\mathbf{v}^T\right\}$$

$$\dim\left\{Q_{proper}(\tilde{\mathbf{Y}})\right\} \leq dc - c^2 - c + 1$$

$$\dim\left\{Q_{rc}(\tilde{\mathbf{Y}})\right\} = \dim\left\{Q_{cc}(\tilde{\mathbf{Y}})\right\} \leq dc - c^2 + 1$$



- For a given loss, classification imposes much less constraints on the mixing matrices.

Contributions:

- A simple linear transformation can be used to construct a proper, RC or CC weak loss from a label-based loss.
- Classification and class ranking impose much less constraints (on the mixing matrix) than probability estimation
- In the paper, we derive and propose losses for specific weak labeling models:
  - Proper losses for quasi-independent labels (generalizing example A)
  - Convex RC or CC weak losses for independent labels (generalizing the noisy label example B): take $\mathbf{Y}$ equal to a matrix of weak labels.