Consistency of Losses for Learning from Weak Labels

J. Cid-Sueiro[1] D. García-García[2] R. Santos-Rodríguez[3]

[1]Universidad Carlos III de Madrid, Spain

[2]Commonwealth Bank of Australia

[3]Universidad de Valencia, Spain

European Conf. on Machine Learning, Sept. 2014

Outline

Outline

Weak labels

- On some multiclass classification problems, we may have a training dataset with partial information about the class of the data.
- Weak label: a subset of labels, probably including the true class
  - Photo captions
  - Multiple specialized binary annotators.
  - Hierarchical classes and superclass labels

Alonso, Vettel and Webber in India

Data model

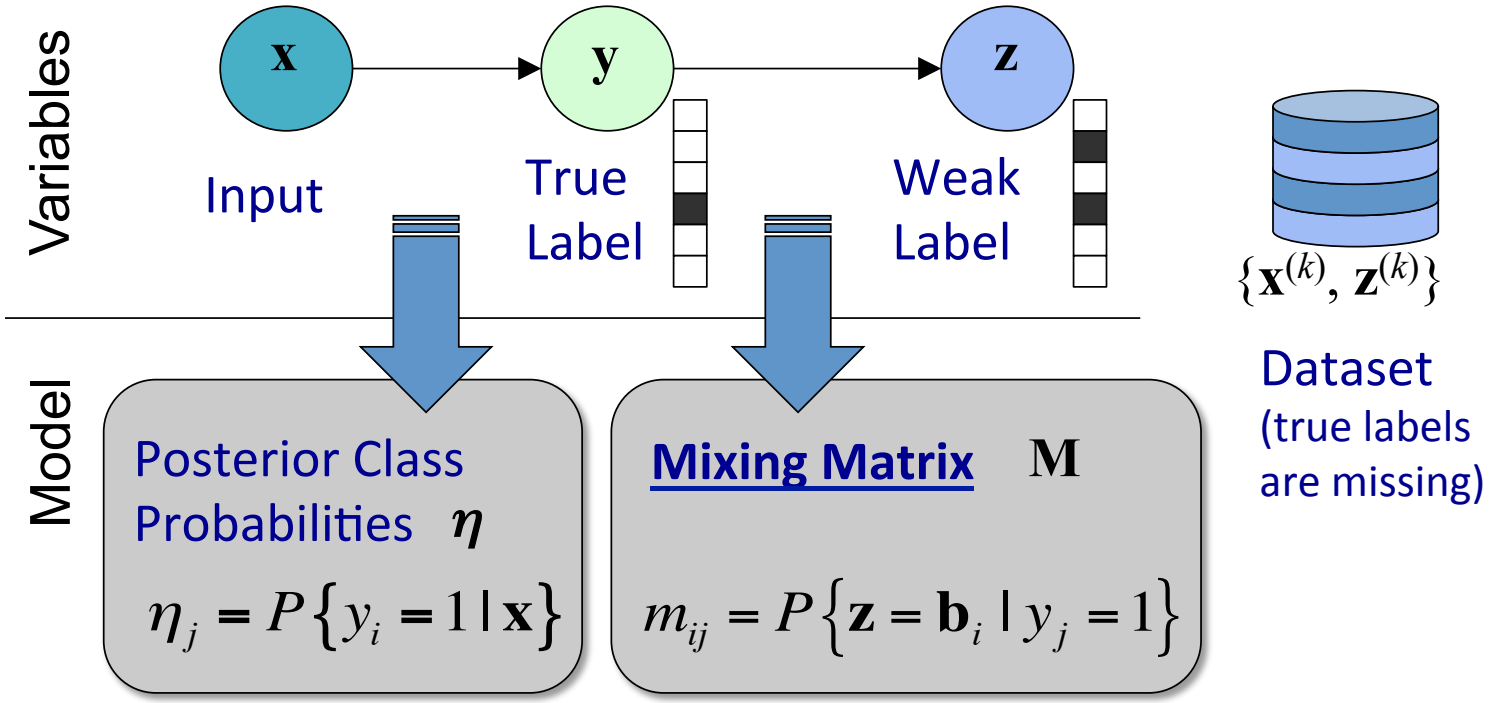


Figure: Data model

Outline

Mixing matrices.

- Supervised learning:

True Label

Weak Label

	0	0	0
	1	0	0
	0	1	0
	0	0	1
	0	0	0
	0	0	0
	0	0	0
	0	0	0

Probability of observing weak label (classes 1 and 3) when the true class is (class 3)

Mixing matrices.

- Unsupervised learning:

True Label

Weak Label

	α	β	γ
	$1-\alpha$	0	0
	0	$1-\beta$	0
	0	0	$1-\gamma$
	0	0	0
	0	0	0
	0	0	0
	0	0	0

Mixing matrices.

- Noisy labels:

True Label

Weak Label			
	■□□	□■□	□□■
□□□	0	0	0
■□□	$1-\alpha$	$\beta/2$	$\gamma/2$
□■□	$\alpha/2$	$1-\beta$	$\gamma/2$
□□■	$\alpha/2$	$\beta/2$	$1-\gamma$
■■□	0	0	0
■□■	0	0	0
□■■	0	0	0
■■■	0	0	0

Mixing matrices.

- True label and one noisy label:

True Label

Weak Label			
	■□□	□■□	□□■
□□□	0	0	0
■□□	$1-\alpha$	0	0
□■□	0	$1-\beta$	0
□□■	0	0	$1-\gamma$
■■□	$\alpha/2$	$\beta/2$	0
■□■	$\alpha/2$	0	$\gamma/2$
□■■	0	$\beta/2$	$\gamma/2$
■■■	0	0	0

Mixing matrices.

- True label and independent noisy labels:

True Label

Weak Label	■□□	□■□	□□■
□□□	0	0	0
■□□	$(1-\beta)^2$	0	0
□■□	0	$(1-\beta)^2$	0
□□■	0	0	$(1-\beta)^2$
■■□	$\beta(1-\beta)$	$\beta(1-\beta)$	0
■□■	$\beta(1-\beta)$	0	$\beta(1-\beta)$
□■■	0	$\beta\ (1-\beta)$	$\beta(1-\beta)$
■■■	β^2	β^2	β^2

Outline

Inference from partial labels.

- If $\mathbf{M}$ is known, since

$$\mathbf{p} = \mathbf{M}\eta$$

we can estimate η from $\mathbf{p}$ given $\mathbf{M}$ as

$$\eta = \mathbf{M}^{+}\mathbf{p}$$

- Two major problems:
  1. Even knowing $\mathbf{M}$, estimating (a high dimensional) $\mathbf{p}$ from data may be infeasible
  2. $\mathbf{M}$ is generally unknown, and cannot be estimated from $\mathcal{S}$.

Inference from partial labels.

- If $\mathbf{M}$ is known, since

$$\mathbf{p} = \mathbf{M}\eta$$

we can estimate η from $\mathbf{p}$ given $\mathbf{M}$ as

$$\eta = \mathbf{M}^{+}\mathbf{p}$$

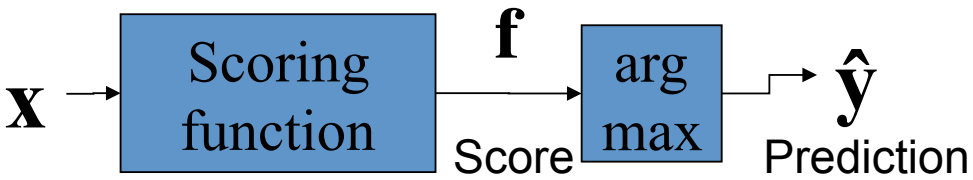
- Two major problems:
  1. Even knowing $\mathbf{M}$, estimating (a high dimensional) $\mathbf{p}$ from data may be infeasible
  2. $\mathbf{M}$ is generally unknown, and cannot be estimated from $\mathcal{S}$.

Limitations of Learning from Weak Labels

- Any learning algorithm is effective for some weak labelling models only.
- Our goals:
  - Determine conditions or requirements for learnability
  - To analyze comparatively the impact of these requirements in probability estimation, class ranking and classification.
  - To propose general methods to design loss functions for learning.

Outline

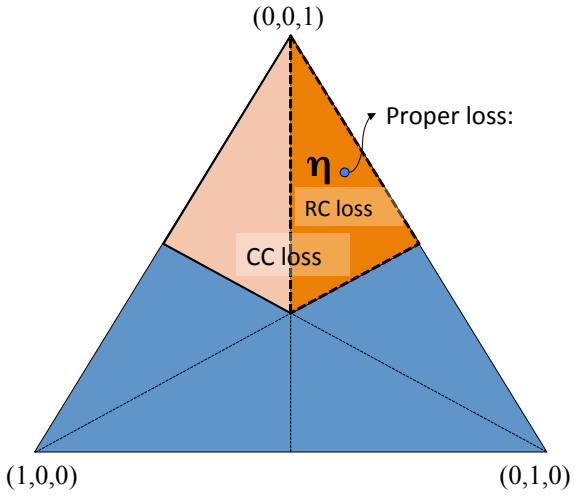
Losses for weak labels.



- Label-based losses
  - A function of the true label and the score: $\tilde{\Psi}(\mathbf{y}, \mathbf{f})$
- Weak losses
  - A function of the weak label and the score: $\Psi(\mathbf{z}, \mathbf{f})$.

Proper, RC and CC losses

- Let $\mathbf{f}^*$ be a minimizer of $\mathbb{E}_{\mathbf{z}}\{\Psi(\mathbf{z}, \mathbf{f})\}$
- Three types of losses:
  1. **M**-Proper: a loss for posterior class probability estimation
     - $\mathbf{f}^* = \boldsymbol{\eta}$
  2. **M**-Ranking Calibrated (RC): A loss for ranking classes
     - $f_i^* > f_j^* \Leftrightarrow \eta_i > \eta_j$
  3. **M**-Classification Calibrated (CC): A loss to mininimize errors
     - $f_i^* > \max_{j \neq i} f_j^* \Leftrightarrow \eta_i > \max_{j \neq i} \eta_j$



Outline

Theorem

Consider a weak loss Ψ *and a mixing matrix* $\mathbf{M}$*, and let the* equivalent *(label-based) loss* $\tilde{\Psi}$ *be given by*

$$\tilde{\boldsymbol{\Psi}}(\mathbf{f}) = \mathbf{M}^{\mathsf{T}}\boldsymbol{\Psi}(\mathbf{f}) \tag{1}$$

- Ψ *is (strictly)* $\mathbf{M}$*-proper iff* $\tilde{\Psi}$ *is (strictly) proper.*
- Ψ *is* $\mathbf{M}$*-RC iff* $\tilde{\Psi}$ *is RC.*
- Ψ *is* $\mathbf{M}$*-CC iff* $\tilde{\Psi}$ *is CC.*

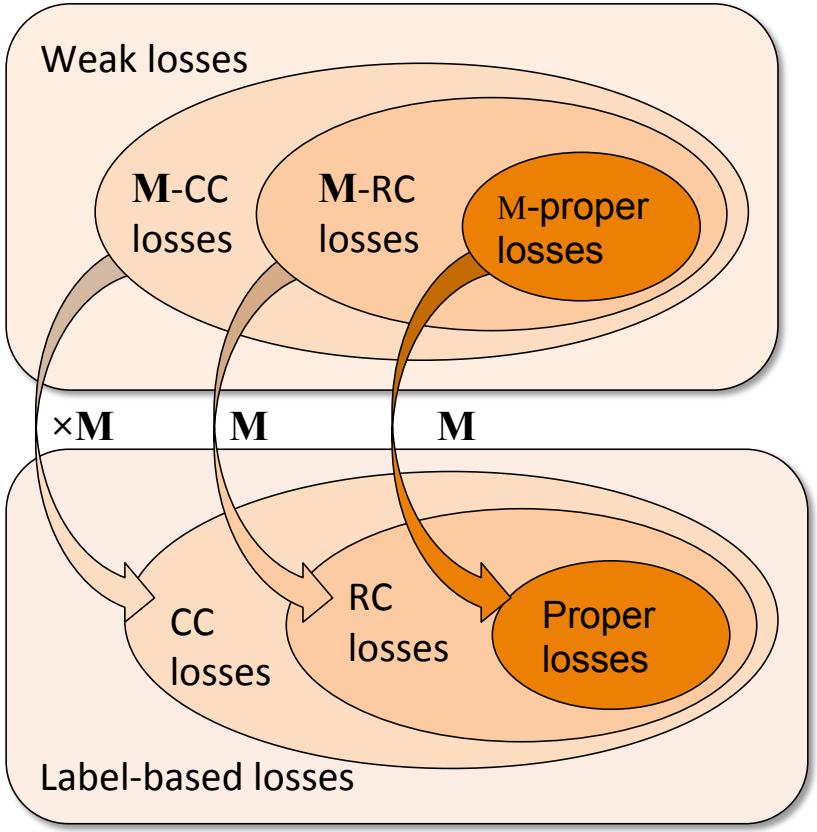


Figure: From Weak to labeled

- Let us proceed in a constructive way: we can design weak losses as a linear transformation of a label based loss $\boldsymbol{\Psi} = \tilde{\mathbf{Y}}^{\intercal}\tilde{\boldsymbol{\psi}}$
- Maximal sets:
  1. $\mathcal{Q}_{\text{proper}}(\tilde{\mathbf{Y}})$ = the set of all mixing matrices, $\mathbf{M}$ s.t. $\boldsymbol{\Psi}$ is $\mathbf{M}$-proper.
  2. $\mathcal{Q}_{rc}(\tilde{\mathbf{Y}})$ = the set of all mixing matrices, $\mathbf{M}$ s.t. $\boldsymbol{\Psi}$ is $\mathbf{M}$-RC.
  3. $\mathcal{Q}_{cc}(\tilde{\mathbf{Y}})$ = the set of all mixing matrices, $\mathbf{M}$ s.t. $\boldsymbol{\Psi}$ is $\mathbf{M}$-C.

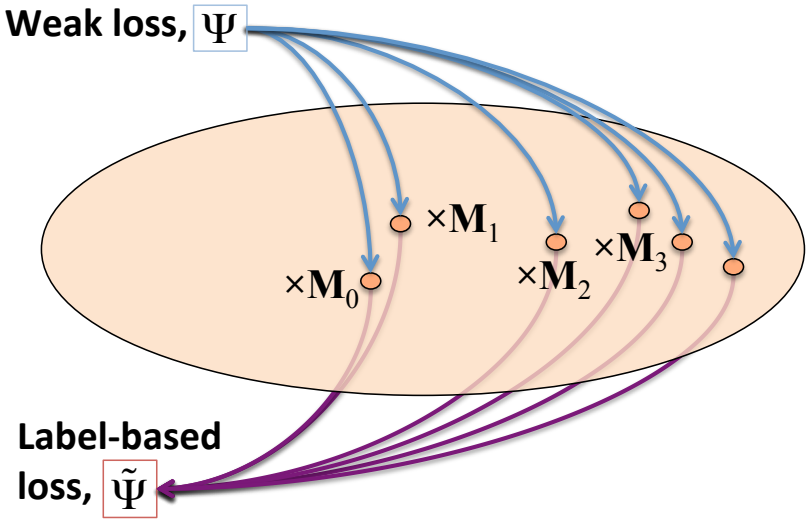
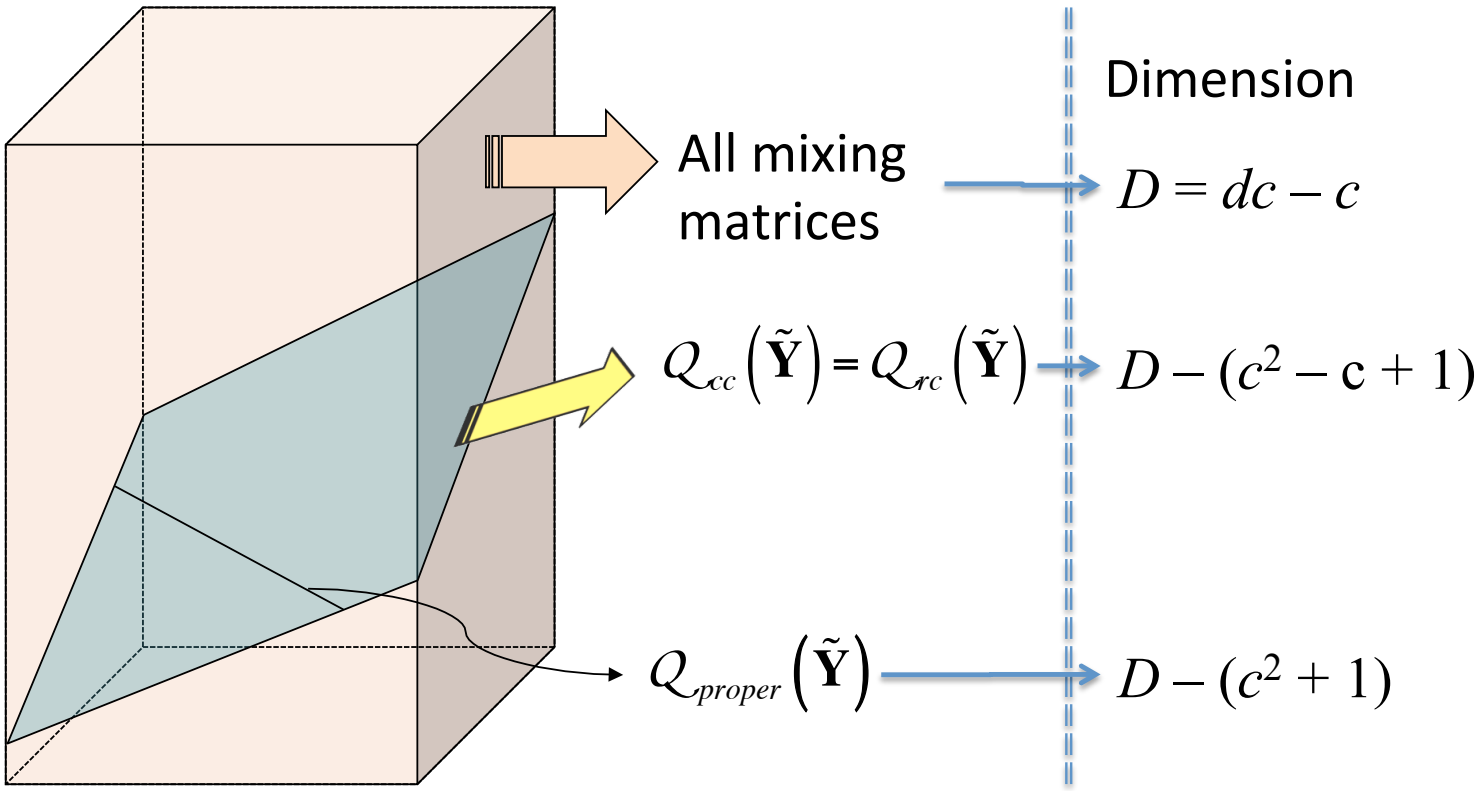


Figure: From Weak to labeled

Theorem

- $\mathcal{Q}_{proper}(\tilde{\mathbf{Y}}) = \{\mathbf{M} | \tilde{\mathbf{Y}}\mathbf{M} = \lambda \mathbf{I}, \lambda > 0\}$
- $\mathcal{Q}_{cc}(\tilde{\mathbf{Y}}) = \mathcal{Q}_{rc}(\tilde{\mathbf{Y}}) = \{\mathbf{M} | \tilde{\mathbf{Y}}\mathbf{M} = \lambda \mathbf{I} + \mathbb{1}_c \mathbf{v}^\intercal, \lambda \in \mathbb{R}^+, \mathbf{v} \in \mathbb{R}^c\}$



- Two be more practical: a proper loss recommendation:
  1. Take any label-based proper loss, e.g. the cross entropy:

$$\tilde{\boldsymbol{\Psi}}(\mathbf{f}) = \log(\mathbf{f}) \tag{2}$$

  2. Take $\tilde{\mathbf{y}}$ given by

$$\tilde{y}_j = \begin{bmatrix} 1 & z_j = 1 \\ -\frac{|\mathbf{z}|-1}{c-|\mathbf{z}|} & z_j = 0 \end{bmatrix} \tag{3}$$

  3. Take $\Psi(\mathbf{z}, \mathbf{f}) = \tilde{\mathbf{y}}^\intercal \tilde{\boldsymbol{\Psi}}(\mathbf{f})$
- $\boldsymbol{\Psi}$ is proper for any quasi-independent mixing models.

- ... and a RC or CC loss recommendation:
  1. Take any label-based RC or CC loss (e.g. the cross entropy)
  2. Take $\Psi(\mathbf{z}, \mathbf{f}) = \mathbf{z}^\intercal \tilde{\boldsymbol{\Psi}}(\mathbf{f})$
- $\boldsymbol{\Psi}$ is RC or CC for any mixing model given by

$$P(\mathbf{z}|y_c = 1) = \alpha^{z_m}(1-\alpha)^{1-z_m}\beta^{|\mathbf{z}|-1}(1-\beta)^{c-|\mathbf{z}|} \tag{4}$$

  for any $\alpha > \beta$.
- $\boldsymbol{\Psi}$ is convex