Learning from Aggregates

Giorgio Patrini
Australian National University, NICTA

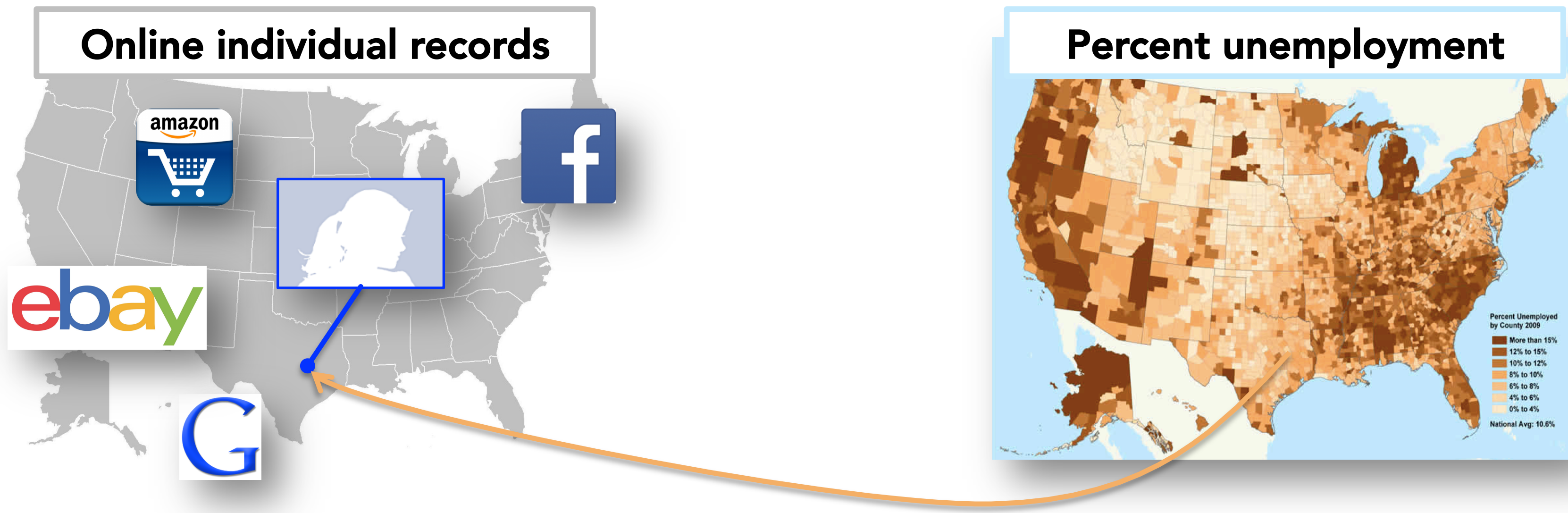
Summary

- Learning from label proportions
- Laplacian Mean Map algorithm

  G.Patrini, R.Nock, P.Rivera, T.Caetano, (Almost) no label no cry, NIPS'14
- Do we need individual feature vectors?

  R.Nock, G.Patrini, A.Friedman, Rademacher observations, private data, and boosting, ICML'15

Learning from Label Proportions (LLP)



Input: unlabeled data

Input: label proportions

Output: predictor of individual unemployment

How likely Alice is unemployed
only given her online behavior

Learning from Label Proportions (LLP)

Other applications:

- Bags of images/pixels in Computer Vision
- Classify sentences as positive/negative based on overall review score
- Data comes from physical measurements which are technically **feasible only in aggregated form**
- Potentially, applications already explored by Multiple Instance Learning (MIL)

Learning setting

- Sample $\mathcal{S} = \{(\boldsymbol{x}_i, y_i), i \in [m]\}$, on $\mathbb{R}^d \supseteq \mathcal{X} \times \{-1, +1\}$
- **No label is observed**
- Known: partition of bags $\cup_j \mathcal{S}_j = \mathcal{S}, j \in [n]$, and relative **label proportions** $\boldsymbol{\pi}_j$
- (No assumption on how the bags were made)

Goal: learn a binary (linear) classifier $\boldsymbol{\theta}$ for individual feature vectors $\boldsymbol{x}$ to predict the label as $\operatorname{sgn}\langle\boldsymbol{\theta}, \boldsymbol{x}\rangle$

Our solution, step 1: factorisation theorem

Def (*Altun&Smola COLT'06*): the **mean operator**

$$\boldsymbol{\mu} = 1/m \sum_{i=1}^{m} y_i \boldsymbol{x}_i$$

Thm (**proper losses factorisation**): $\boldsymbol{\mu}$ is **sufficient** for the label variable for most proper losses:

$$\textsc{proper-loss} = \textsc{loss w/o labels}(\boldsymbol{\theta}) - \tfrac{1}{2}\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle$$

Our solution, step 1: factorisation theorem

Def (*Altun&Smola COLT'06*): the **mean operator**

$$\boldsymbol{\mu} = 1/m \sum_{i=1}^{m} y_i \boldsymbol{x}_i$$

Thm (**proper losses factorisation**): $\boldsymbol{\mu}$ is **sufficient** for the label variable for most proper losses:

$$\text{PROPER-LOSS} = \text{LOSS W/O LABELS}(\boldsymbol{\theta}) - \tfrac{1}{2}\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle$$

***e.g.*, classic logistic loss**

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} \log(1 + e^{-y\boldsymbol{\theta}^\top x_i}) =$$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} \log \sum_{y \in \{-1,1\}} e^{-y\boldsymbol{\theta}^\top x_i} - \langle \boldsymbol{\theta}, \frac{1}{2m} \sum_{i=1}^{m} y_i \boldsymbol{x}_i \rangle$$

Our solution, step 2: estimate the mean operator

$$\boldsymbol{\mu} = \sum_{j=1}^{n} p(j)\boldsymbol{\mu}_j = \sum_{j=1}^{n} p(j) \sum_{y\in\{-1,1\}} y p(y|j)\mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j,y]$$

$$= \sum_{j=1}^{n} p(j)(\boldsymbol{\pi}_j \boldsymbol{b}_j^{+} - (1-\boldsymbol{\pi}_j)\boldsymbol{b}_j^{-})$$

$$\boldsymbol{b}_j^{y} = \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j,y]$$

Then, come up with a system of equations with $\boldsymbol{b}_j^{y}$ as only unknowns:

$$\boldsymbol{b}_j = \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j] = \sum_{y\in\{-1,1\}} p(y|j)\mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j,y] = \sum_{y\in\{-1,1\}} \boldsymbol{\pi}_j \boldsymbol{b}_j^{y}$$

$$\boldsymbol{b}_j = \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j] = \sum_{y \in \{-1,1\}} p(y|j)\mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j,y] = \sum_{y \in \{-1,1\}} \boldsymbol{\pi}_j \boldsymbol{b}_j^y$$

2 variables for each equation!

Quadrianto et al. JMLR'09

$$\boldsymbol{b}_j = \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j] = \sum_{y \in \{-1,1\}} p(y|j)\mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j,y] = \sum_{y \in \{-1,1\}} \boldsymbol{\pi}_j \boldsymbol{b}_j^y$$

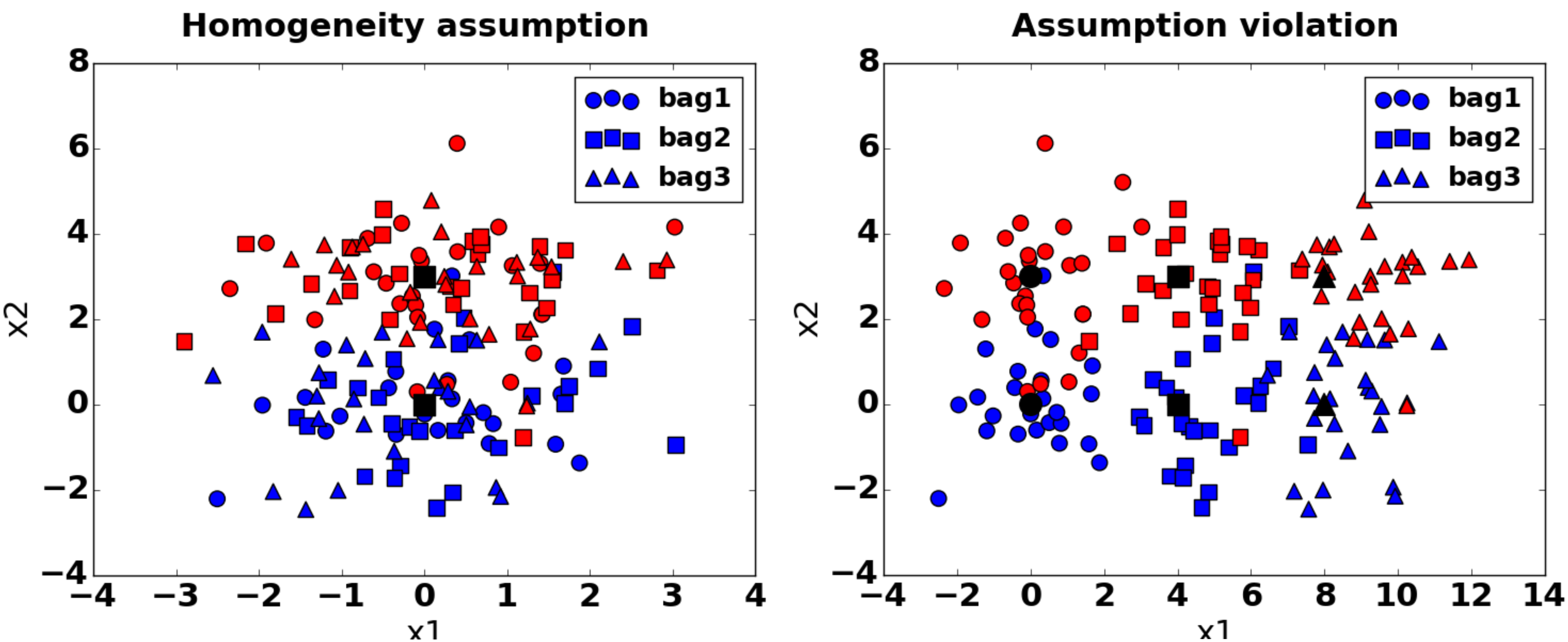
2 variables for each equation!

Solution of *Quadrianto et al. JMRL'09* with *Mean Map*, **homogeneity** assumption:

$$\forall_j \ \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j,y] = \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|y]$$

"Unemployed people in all the counties behave online in the same way, in average"

Homogeneity assumption: $\forall_j \; \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j,y] = \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|y]$



We relax it

$$\boldsymbol{b}_j = \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j] = \sum_{y \in \{-1,1\}} p(y|j)\mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j, y] = \sum_{y \in \{-1,1\}} \boldsymbol{\pi}_j \boldsymbol{b}_j^y$$

We only asks **smoothness** on "similar" bags:

$$\forall_{j,j'}\ \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j] \approx \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j'] \implies \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j, y] \approx \mathbb{E}_{\mathcal{S}}[\boldsymbol{x}|j', y]$$

"The more similar the counties, the more similar the online behaviour of the people unemployed there"

Our solution, step 3: Laplacian regularization

Let $v_{j,j'}$ be the similarity between bags. Then we can encode our assumption in a **regularized least square**:

$$\operatorname*{argmin}_{\boldsymbol{b}_j^y} \sum_j (\boldsymbol{b}_j - \sum_{y\in\{-1,1\}} \pi_j \boldsymbol{b}_j^y)^2 + \gamma \sum_{j,j'} v_{j,j'} [(\boldsymbol{b}_j^+ - \boldsymbol{b}_{j'}^+)^2 + (\boldsymbol{b}_j^- - \boldsymbol{b}_{j'}^-)^2]$$

Then, in matrix form:

$$\begin{aligned} \mathrm{B} &= [\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_n]^\top, \ \mathrm{B}^\pm = [\boldsymbol{b}_1^+, \boldsymbol{b}_2^+, ..., \boldsymbol{b}_n^+, \boldsymbol{b}_1^-, \boldsymbol{b}_2^-, ..., \boldsymbol{b}_n^-]^\top, \\ \Pi &= [\textsc{Diag}(\boldsymbol{\pi}) | \textsc{Diag}(\mathbf{1} - \boldsymbol{\pi})] \end{aligned}$$

$$\operatorname*{argmin}_{\mathrm{B}^\pm} \operatorname{tr}\left((\mathrm{B} - \Pi\mathrm{B}^\pm)^\top (\mathrm{B} - \Pi\mathrm{B}^\pm)\right) + \gamma \operatorname{tr}\left((\mathrm{B}^\pm)^\top \mathrm{L}\mathrm{B}^\pm\right)$$

Laplacian matrix on $v_{j,j'}$

Our solution: Laplacian Mean Map algorithm

(steps in reverse order)

Laplacian Mean Map (LMM)

Input $\mathcal{S}_j, \pi_j, j \in [n]; \lambda, \gamma > 0; \text{V};$

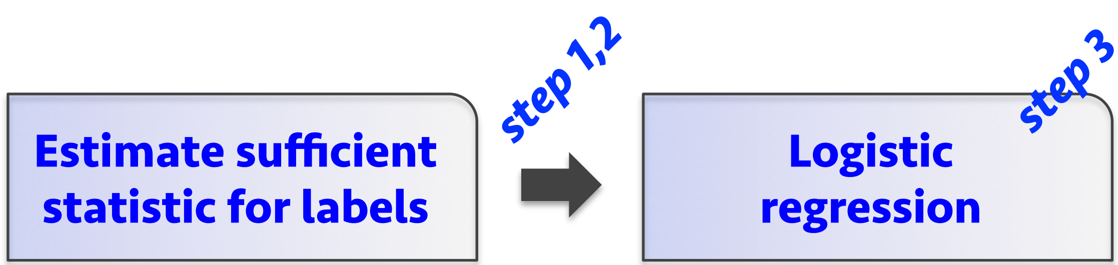
Step 1 : let $\text{B}^{\pm} \leftarrow (\Pi\Pi^T + \gamma L)^{-1}\Pi B$

Step 2 : let $\boldsymbol{\mu} \leftarrow \sum_j p_j(\pi_j \boldsymbol{b}_j^+ - (1 - \pi_j)\boldsymbol{b}_j^-)$

Step 3 : let $\boldsymbol{\theta}_* \leftarrow \arg\min_{\boldsymbol{\theta}} \text{LOSS W/O LABEL}(\boldsymbol{\theta}) + \frac{1}{2}\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle + \lambda \|\boldsymbol{\theta}\|_2^2;$

Return $\boldsymbol{\theta}^*$

Scalability: Step 1 is only $O(n^3) \ll O(m^3)$



Approximation of the mean operator

Theorem 1 *Suppose that* γ *satisfies* $\gamma\sqrt{2} \leq \max_{j \neq j'} v_{jj'}$. *Let* $\mathrm{M} \doteq [\boldsymbol{\mu}_1 | \boldsymbol{\mu}_2 | ... | \boldsymbol{\mu}_n]^\top \in \mathbb{R}^{n \times d}$, $\tilde{\mathrm{M}} \doteq [\tilde{\boldsymbol{\mu}}_1 | \tilde{\boldsymbol{\mu}}_2 | ... | \tilde{\boldsymbol{\mu}}_n]^\top \in \mathbb{R}^{n \times d}$ *and* $\psi(\mathrm{V}, \mathrm{B}^{\pm}) \doteq (\max_{j \neq j'} v_{jj'})^2 \|\mathrm{B}^{\pm}\|_F$. *The following holds:*

$$\|\mathrm{M} - \tilde{\mathrm{M}}\|_F \quad \leq \quad \sqrt{n/2} \times \psi(\mathrm{V}, \mathrm{B}^{\pm}) \ .$$

(Assuming homogeneity with Mean Map, the norm is unbounded.)

Choose the similarity $v^G_{jj'} \doteq \exp(-\|\boldsymbol{b}_j - \boldsymbol{b}_{j'}\|_2^2)$

Under mild conditions, it holds, *w.r.t.* the max norm of $\boldsymbol{b}^y_j = \mathbb{E}_{\mathcal{S}}[x | j, y]$:

$$\psi(\mathrm{V}^G, \mathrm{B}^{\pm}) = o(1)$$

Approximation of the model

Theorem 1 *Let $\boldsymbol{\theta}_*$ be the model computed with the true mean operator $\boldsymbol{\mu}$. Let $\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\theta}}_*$ be the respective estimates. For any proper loss L_2-regularizated with parameter $\lambda > 0$, there exists $q > 0$ such that:*

$$\|\tilde{\boldsymbol{\theta}}_* - \boldsymbol{\theta}_*\|_2^2 \leq 1/(2\lambda + q)\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2$$

This holds for **any estimator of** μ, even outside the LLP setting

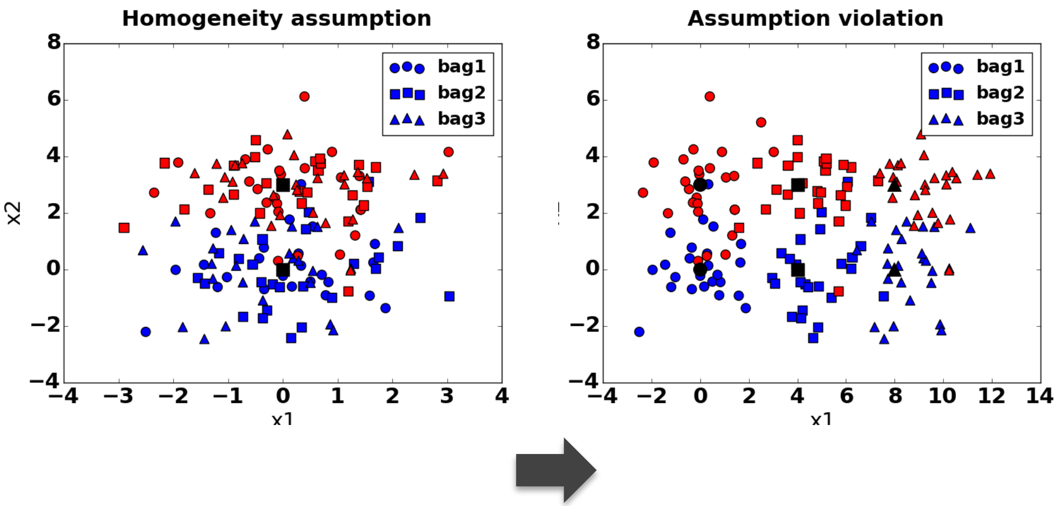
Estimate sufficient statistic for labels → (**bound holds**) → **Logistic regression**

And more in the paper

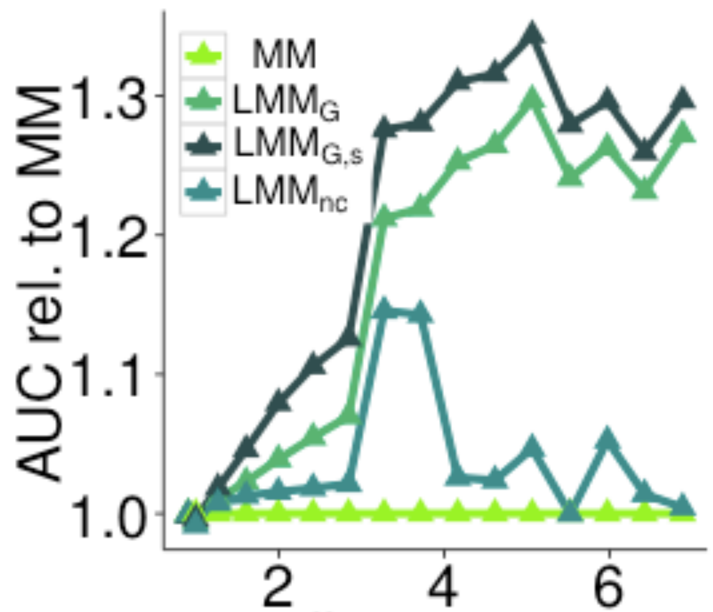
- **Alternating Mean Map**: use LMM as initialization and optimizes further, inferring labels as latent variables (similar to Expectation Maximization)

- We also provide **generalization bounds** based on **Rademacher Complexity**.

Experiments: homogeneity assumption



gradual violation of homogeneity



Experiments: comparative tests

14 UCI datasets **converted to LLP** (up to ~300K examples)

- Select a categorical feature, use its value to assign bags and proportions; then remove the feature.
- Compare with SVMs (*Yu et al. ICML'13*) and InvCal (*Rueping ICML'10*)

Table 1: 10 small domains results. #win/#lose for row vs column on 50 tests; ties not reported. Bold faces when p-val $<$ .001 for Wilcoxon signed-rank tests.

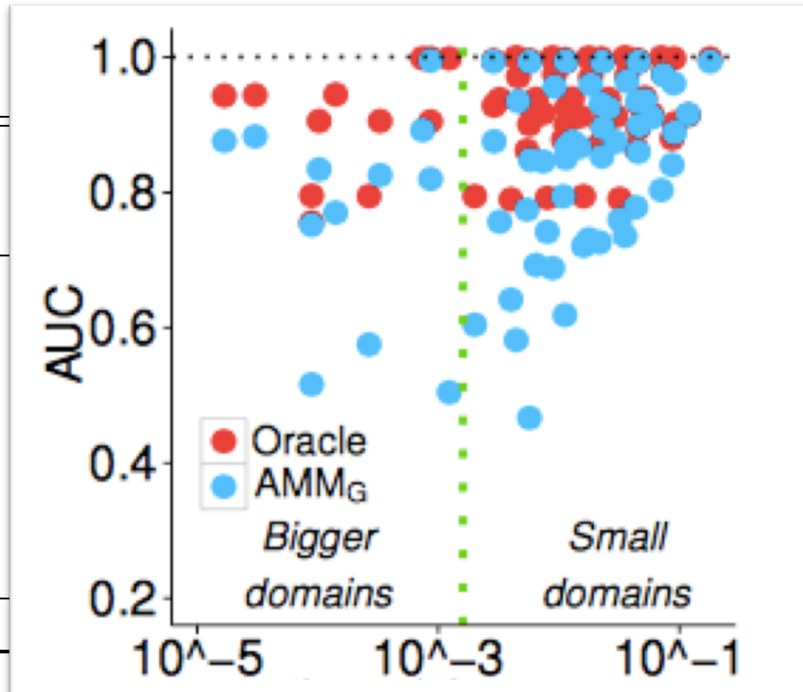
	algorithm	MM	LMM G	LMM G,s	LMM nc	InvCal	AMM$^{\text{min}}$ MM	AMM$^{\text{min}}$ G	AMM$^{\text{min}}$ G,s	AMM$^{\text{min}}$ 10ran	*conv-*$\propto SVM$
LMM	G	**36**/4									
LMM	G,s	**38**/3	**30**/6								
LMM	nc	**28**/12	3/**37**	2/**37**							
	InvCal	4/**46**	3/**47**	4/**46**	4/**46**						
AMM$^{\text{min}}$	MM	**33**/16	26/24	25/25	32/18	**46**/4					
AMM$^{\text{min}}$	G	**38**/11	**35**/14	**30**/20	**37**/13	**47**/3	**31**/7				
AMM$^{\text{min}}$	G,s	**35**/14	**33**/17	**30**/20	**35**/15	**47**/3	**24**/11	7/**15**			
AMM$^{\text{min}}$	10ran	27/22	24/26	22/28	26/24	**44**/6	20/**30**	16/**34**	19/**31**		
SVM	*conv-*$\propto$	21/**29**	2/**48**	2/**48**	2/**48**	2/**48**	4/**46**	3/**47**	3/**47**	4/**46**	
SVM	*alter-*$\propto$	0/**50**	0/**50**	0/**50**	0/**50**	20/**30**	0/**50**	0/**50**	0/**50**	3/**47**	27/23

Experiments: no label no cry

algorithm		*adult*: **48842 × 89**			*marketing*: **45211 × 41**			*census*: **299285 × 381**		
		IV(5)	V(16)	VI(42)	V(4)	VII(4)	VIII(12)	IV(5)	VIII(9)	VI(42)
	MM	80.93	76.65	74.01	54.64	50.71	49.70	75.21	**90.37**	75.52
	LMM_G	81.79	78.40	78.78	**54.66**	**51.00**	51.93	75.80	71.75	**76.31**
	$\text{LMM}_{G,s}$	**84.89**	**78.94**	**80.12**	49.27	**51.00**	**65.81**	**84.88**	60.71	69.74
$\text{AMM}^{\min}$	AMM_{MM}	**83.73**	77.39	80.67	52.85	**75.27**	58.19	89.68	84.91	68.36
	AMM_G	83.41	**82.55**	**81.96**	51.61	75.16	57.52	87.61	88.28	76.99
	$\text{AMM}_{G,s}$	81.18	78.53	**81.96**	52.03	75.16	53.98	**89.93**	83.54	52.13
	AMM_1	81.32	75.80	80.05	65.13	64.96	66.62	89.09	**88.94**	56.72
$\text{AMM}^{\max}$	AMM_{MM}	82.57	71.63	81.39	48.46	51.34	56.90	50.75	66.76	58.67
	AMM_G	82.75	72.16	81.39	50.58	47.27	34.29	48.32	67.54	**77.46**
	$\text{AMM}_{G,s}$	82.69	70.95	81.39	**66.88**	47.27	34.29	80.33	74.45	52.70
	AMM_1	75.22	67.52	77.67	66.70	61.16	**71.94**	57.97	81.07	53.42
	Oracle	90.55	90.55	90.50	79.52	75.55	79.43	94.31	94.37	94.45

Experiments: no label no cry

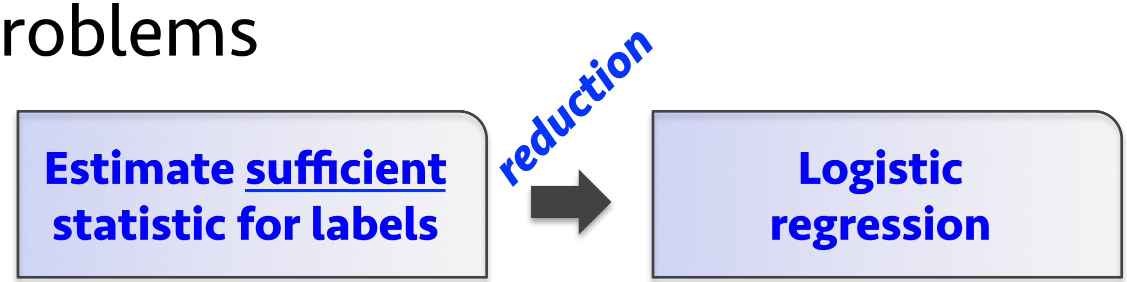
algorithm		*adult*: IV(5)	*census*: 299285 × 381 IV(5)	VIII(9)	VI(42)
	MM	80.93	75.21	**90.37**	75.52
	$\mathrm{LMM}_{\mathrm{G}}$	81.79	75.80	71.75	**76.31**
	$\mathrm{LMM}_{\mathrm{G,s}}$	**84.89**	**84.88**	60.71	69.74
$\mathrm{AMM}^{\min}$	$\mathrm{AMM}_{\mathrm{MM}}$	**83.73**	89.68	84.91	68.36
	$\mathrm{AMM}_{\mathrm{G}}$	83.41	87.61	88.28	76.99
	$\mathrm{AMM}_{\mathrm{G,s}}$	81.18	**89.93**	83.54	52.13
	AMM_{1}	81.32	89.09	**88.94**	56.72
$\mathrm{AMM}^{\max}$	$\mathrm{AMM}_{\mathrm{MM}}$	82.57	50.75	66.76	58.67
	$\mathrm{AMM}_{\mathrm{G}}$	82.75	48.32	67.54	**77.46**
	$\mathrm{AMM}_{\mathrm{G,s}}$	82.69	80.33	74.45	52.70
	AMM_{1}	75.22	57.97	81.07	53.42
	Oracle	90.55	94.31	94.37	94.45



more supervised
#bags/#instances

Take-home messages (until here)

- **(Almost) no label no cry**: few proportions can suffice to learn. Privacy threat?
- **Sufficiency of mean operator**: any "*weakly-supervised*" learner can exploit the same trick, *e.g.* semi-supervised, MIL, noisy labels. Bound for the classifier holds.
- **Do not reinvent the wheel**: *reduction* between ML problems



But what about individual *feature vectors*?

- Is there an analogue of the mean operator that allows us to learn with *aggregate feature vectors*?

- YES. Define a **Rademacher observation** as a (non-normalized) mean operator restricted to a subsample $s \in \mathcal{S}$:

$$\boldsymbol{\mu}_s = \sum_{i:(\boldsymbol{x}_i, y_i) \in s} y_i \boldsymbol{x}_i$$

Rademacher observations and logistic loss

$$\operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{m} \sum_{i=1}^{m} \log(1 + e^{-y\boldsymbol{\theta}^\top x_i}) =$$

$$\operatorname*{argmin}_{\boldsymbol{\theta}} \log(2) + \frac{1}{m} \log \left(\frac{1}{2^m} \sum_{s \subseteq \mathcal{S}} e^{-\boldsymbol{\theta}^\top \boldsymbol{\mu}_s} \right)$$

they are all aggregated here

The number of $\boldsymbol{\mu}_s$ is exponential in m, but we can still learn on a small subset of Rademacher observations. See our *ICML'15* for details.

Yes, but why?

- When we have all the data but do not want to share it entirely with the learner, but still want to learn good models. **Privacy** constraints.
  - Can prove differential privacy
  - Properties of non-reconstruct-ability of the original data (NP-harness and algebraic impossibility)

Conclusion

Learning from aggregate data is possible, with unexpected applications on

- weakly-supervised learning
- privacy
- distributed learning - one μ_s per cluster?
- and social sciences, *e.g.* the ecological inference

➢ **NIPS'15 workshop** on "*Learning and privacy with incomplete data and weak supervision*"