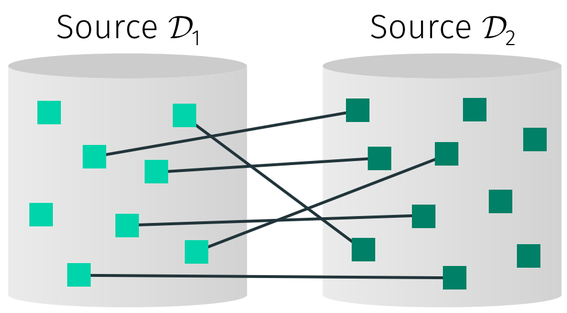
In Search of an Entity Resolution OASIS:
Optimal Asymptotic Sequential Importance Sampling

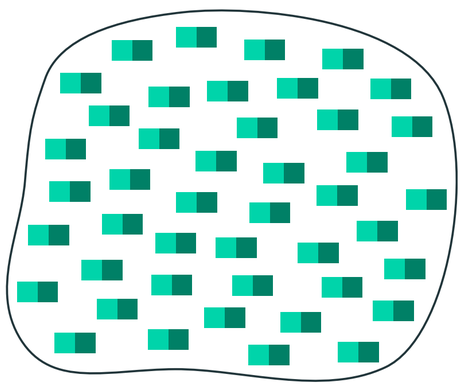
Neil Marchant and Ben Rubinstein
School of Computing and Information Systems, University of Melbourne, Australia

THE UNIVERSITY OF MELBOURNE

1. Evaluation of Entity Resolution

Entity resolution (ER) is the task of identifying records across data sources $\{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$ that refer to the *same entities*. It may be cast as a binary classification problem on the product space $\mathcal{Z} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_m$.





Product space: $\mathcal{Z} = \mathcal{D}_1 \times \mathcal{D}_2$

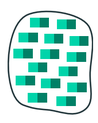
Evaluation is important since ER is an inherently ambiguous task. However sound evaluation is made difficult due to extreme class imbalance. Finding a matching record when labelling is like finding an oasis in a desert! For every "match" there are at least $\max(|\mathcal{D}_1|, |\mathcal{D}_2|)$ "non-matches". This makes standard approaches based on uniform sampling infeasible.

2. Problem formulation

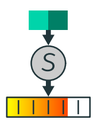
Motivated by the inefficiency/inaccuracy of standard evaluation methods, we seek to develop a new method of estimating F-measure that is:

- *statistically consistent*: converges in probability to the true value
- *statistically efficient*: requires minimal labels

In evaluating a predicted ER, we assume access to:



Pool of record pairs: ideally the pool P is a subset of $\mathcal{Z}$ drawn randomly. However P could also be selected based on blocking.



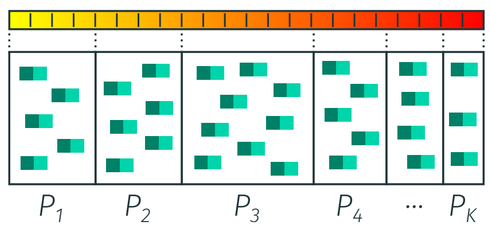
Similarity scores: quantify the degree of similarity between records. Most ER methods produce such scores.

Oracle: returns ground truth labels (match/non-match) for record pairs in the pool—e.g. implemented via crowdsourcing.

3. Key ingredients of OASIS

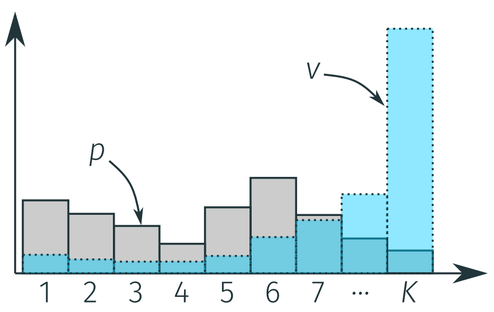
Stratification

The pool is partitioned into K strata based on the similarity scores. By grouping "similar" items together, the number of parameters in the subsequent model may be reduced.



Sequential Importance Sampling (SIS)

SIS can effectively achieve variance reduction. Rather than sampling from the strata proportionately (p), they are sampled according to a biased instrumental distribution (v), chosen to minimise the estimator variance. In SIS v is updated *sequentially* to approach optimality.



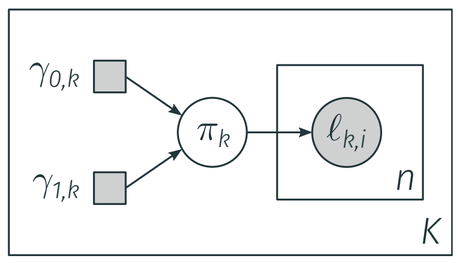
Simple Bayesian model for the Oracle

Estimates the likelihood that the Oracle returns a "match" label in each stratum:

$$\pi_k \sim \text{Beta}(\gamma_{0,k}^{(0)}, \gamma_{1,k}^{(0)})$$

$$\ell_{k,i} \sim \text{Bern}(\pi_k)$$

Needed to estimate the optimal v.

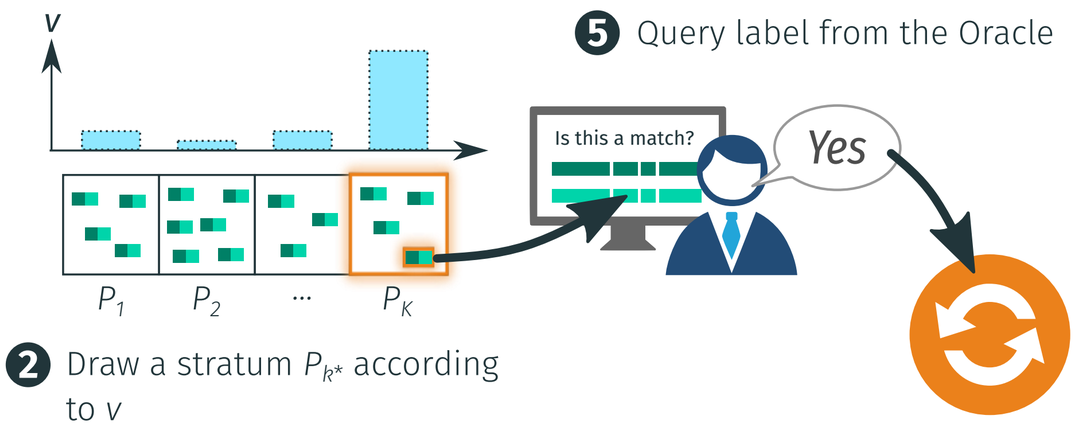


4. The OASIS algorithm

Initialise: stratify pool and generate initial estimates based on scores

Sample: at each iteration do the following:

1. Update instrumental distribution (v)
2. Draw a stratum P_{k^*} according to v
3. Draw a record pair uniformly from stratum P_{k^*}
4. Record the importance weight for bias correction
5. Query label from the Oracle
6. Update estimates and Bayesian model for the Oracle

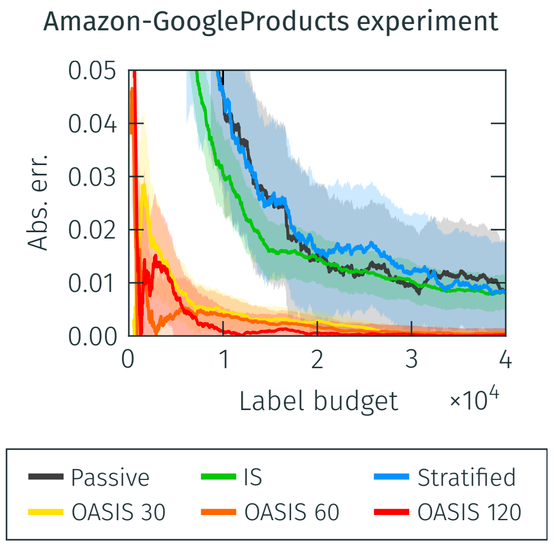


5. Theoretical guarantee

Theorem: OASIS is a consistent estimator of the α-weighted F-measure (includes precision, recall, F1 score)

Challenges for the proof: samples generated by OASIS are not i.i.d.; non-linearity of the F-measure; and ensuring that v dominates p.

6. Experimental results



- We compared OASIS with 3 baseline evaluation methods (Passive, Stratified, non-adaptive IS) on 5 ER datasets
- OASIS outperforms the baselines on all but one dataset (where it remains competitive)
- Example (left): OASIS achieves an 83% reduction in labelling requirements (for an exp. err. of 0.01) compared to the prior state-of-the-art.

7. Open-source Python package

The `oasis` package implements OASIS and the baseline evaluation methods.

To install from PyPI run: `pip3 install --user oasis`

For documentation and more info visit: `https://git.io/OASIS`

References

[1] C. Sawade, N. Landwehr, and T. Scheffer. Active Estimation of F-measures. In *NIPS*, pages 2083–2091. 2010.

[2] G. Druck and A. McCallum. Toward Interactive Training and Evaluation. In *CIKM*, pages 947–956, 2011.

[3] H. Köpcke, A. Thor, and E. Rahm. Evaluation of Entity Resolution Approaches on Real-world Match Problems. *PVLDB*, 3(1):484–493, 2010.

[4] T. Dalenius and J. L. Hodges. Minimum Variance Stratification. *JASA*, 54(285):88–101, 1959.

[5] R. Y. Rubinstein and D. P. Kroese. *Simulation and the Monte Carlo Method*. Wiley, 2007.

N. Marchant acknowledges the support of an Australian Government Research Training Program Scholarship.
B. Rubinstein acknowledges the support of the Australian Research Council (DP150103710).