# End-to-End Statistical Learning, with or without Labels

Corinne Jones[*] Vincent Roulet[*] Zaid Harchaoui[*]

**Abstract**
We introduce an approach that allows one to learn a feature representation and perform clustering of unlabeled data. The approach can also leverage any amount of additional labeled data in order to boost the statistical performance. The proposed method is based on a semi-implicit stochastic optimization algorithm and an entropy-regularized optimal transport algorithm. A numerical illustration on a real dataset shows the promise of the proposed approach.

**Key Words:** discriminative clustering, unsupervised learning, semi-supervised learning, representation learning

## 1. Introduction

In many domains, ranging from healthcare to astronomy, collecting large quantities of labeled data can be time-consuming and even prohibitively expensive. A paucity of such labeled data can pose a problem when trying to perform classification tasks using large models such as deep networks. However, collecting unlabeled data can be relatively inexpensive. Recent work leverages unlabeled data in a variety of ways to learn feature representations using deep networks, either with unsupervised or semi-supervised methods (Chapelle et al., 2010; Oliver et al., 2018).

While these unsupervised and semi-supervised methods are often effective, we would ideally like to have a single approach that works regardless of the quantity of labeled data. The crux of learning with unlabeled data is to ensure that it does not result in degenerate solutions. Specifically, the algorithm must avoid assigning all points to the same class and moreover avoid mapping all of the raw features to the same embedded feature vector. Previously these issues were dealt with using heuristic techniques, such as randomly creating a new cluster when a cluster becomes empty.

In this work we propose an approach that naturally reduces to an unsupervised clustering method when no labeled data is available and to a supervised classification method when no unlabeled data is available. Our approach learns a feature representation as well, and hence learns all components of the statistical modeling end to end. To optimize our objective function we develop an effective algorithm that simultaneously learns the parameters of a network and predicts the labels of the unlabeled data. We guard ourselves against degenerate solutions by enforcing cluster balancing constraints and penalizing the norm of the learned features. We present results showing the promise of our algorithm on a real dataset.

## 2. Related Work

We refer the reader to the surveys of Chapelle et al. (2010) and Oliver et al. (2018) for overviews of semi-supervised learning methods. One line of research has explored discriminative clustering methods that incorporate labeled data (Bach and Harchaoui, 2007; Joulin and Bach, 2012; Xu et al., 2009; White and Schuurmans, 2012). In this work we extend DIFFRAC (Bach and Harchaoui, 2007) by equipping it with an ability to learn a feature representation as well. The approach we outline here may be interpreted as learning

[*] Department of Statistics, University of Washington, Seattle, WA 98195, USA

a similarity measure between input data points. This paper is an extended abstract of our working paper (Jones et al., 2019).

## 3. Learning Regardless of the Level of Supervision

In this section we present the end-to-end learning framework allowing us to take advantage of any amount of labeled and unlabeled data.

### 3.1 Problem formulation

Let $x_1, \dots, x_n \in \mathbb{R}^d$ be a sequence of observations, and assume each observation belongs to one of $k$ classes. The class label corresponding to each $x_i$, denoted by $y_i^\star \in \{0,1\}^k$, may or may not be observed. We denote by $\mathcal{S}$ the set of indices corresponding to the labeled data. We aim to use both the labeled and unlabeled data to learn both the parameters $V$ of a network $\phi(\cdot; V) : \mathbb{R}^d \to \mathbb{R}^D$ and the parameters $W \in \mathbb{R}^{D\times k}$ and $b \in \mathbb{R}^k$ of a classifier on the outputs $\phi(x_i; V)$, $i = 1, \dots, n$.

To do so, we denote by $y_i, i = 1, \dots, n$ the rows of a matrix $Y \in \mathbb{R}^{n\times k}$ and consider solving the problem

$$\min_{Y\in\mathcal{C},V,W,b} \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i, W^T\phi(x_i;V)+b\right) + \Omega(V,W)\ ,$$

where $\mathcal{C} = \{Y \in \{0,1\}^{n\times k} : Y\mathbb{1}_k = \mathbb{1}_n, y_i = y_i^\star \text{ for } i \in \mathcal{S}\}$ is the constraint set on the labels, $\ell(y,\hat{y}) = \|y-\hat{y}\|^2$ is the square loss, and $\Omega(V,W) := \alpha\|V\|_F^2 + \lambda\|W\|_F^2$ includes the regularization terms. The scalars $\alpha \geq 0$ and $\lambda \geq 0$ are regularization parameters. Note that the label matrix $Y$ is constrained so each $x_i$ is assigned to a unique class.

**Avoiding degenerate solutions.** The above objective can lead to two different types of trivial solutions: one that maps all observations to the same embedded point, i.e., $\phi(x_1;V) = \phi(x_2;V) = \cdots = \phi(x_n;V)$; and one that assigns all observations to the same cluster, i.e., $y_1 = y_2 = \cdots = y_n$. We avoid the first problem by subtracting the penalty $\rho\sum_{i=1}^n \|\phi(x_i;V) - \bar{\phi}\|_2^2$ on the squared norms of the centered embeddings, where $\bar{\phi} = 1/n\sum_{i=1}^n \phi(x_i;V)$. The second problem exists even when the network parameters $V$ are fixed, as noted by Bach and Harchaoui (2007). To avoid this behavior we add constraints enforcing that the clusters have a minimum and maximum size, i.e., $n_{\min}\mathbb{1}_k \leq Y^T\mathbb{1}_n \leq n_{\max}\mathbb{1}_k$ for some $n_{\max} \geq n_{\min}$.

Formally, we consider then the problem

$$\min_{Y\in\mathcal{C}',V,W,b} \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i, W^T\phi(x_i;V)+b\right) + \mathcal{R}(V,W) \qquad (1)$$

where

$$\mathcal{R}(V,W) = \alpha\|V\|_F^2 + \lambda\|W\|_F^2 - \rho\sum_{i=1}^{n}\|\phi(x_i;V) - \bar{\phi}\|_2^2$$

with $\alpha, \lambda, \rho \geq 0$. The constraint set for $Y$ is now

$$\mathcal{C}' = \{Y \in \{0,1\}^{n\times k} : Y\mathbb{1}_k = \mathbb{1}_n, y_i = y_i^\star \text{ for } i \in \mathcal{S}, n_{\min}\mathbb{1}_k \leq Y^T\mathbb{1}_n \leq n_{\max}\mathbb{1}_k\}\ .$$

We define $\phi_i(V) = \phi(x_i;V)$ and $\Phi(V) = (\phi_1(V), \dots, \phi_n(V))^T$.

### 3.2 Optimization

For $Y$ and $V$ fixed, the minimization over $W$ and $b$ in problem (1) can be performed analytically, leading to the objective

$$\min_{V} \min_{\substack{M=YY^\top \\ Y\in\mathcal{C}}} \frac{1}{2}\operatorname{tr}(A_\lambda(V)M) + \alpha\|V\|_F^2 - \rho\|\Phi(V) - \mathbb{1}_n\bar{\phi}^\top\|_F^2\,, \tag{2}$$

where $M$ is an equivalence matrix and $A_\lambda(V) = \lambda\Pi_n\left(\Pi_n\Phi(V)\Phi(V)^T\Pi_n + n\lambda\,\mathrm{I}_n\right)^{-1}\Pi_n$ with $\Pi_n = \mathrm{I}_n - \mathbb{1}_n\mathbb{1}_n^\top$. Note that by optimizing over the equivalence matrix $M$ instead of the assignment matrix $Y$ we avoid having many equivalent solutions that can be obtained by simply permuting the cluster labels. To optimize the objective, we consider an iterative scheme, where at each iteration: (i) an approximate solution $\hat{M}$ is computed on a mini-batch of size $n_b$ for fixed parameters $V$ by relaxing the constraints in the assignment problem; and (ii) a gradient step is taken to update $V$ for $\hat{M}$ fixed. This optimization strategy can be related to the one in profile likelihood methods, where secondary variables are profiled out and defined implicitly with respect to primary variables (Barndorff-Nielsen and Cox, 1994).

**Cluster assignment.** Consider the objective (2) for $V$ fixed. This problem is hard due to its combinatorial nature. Therefore, we relax the discrete constraints $M \in \{0,1\}^{n_b\times n_b}$ to inequality constraints on the row and column sums of $M$ and consider the following regularized problem:

$$\begin{aligned}
\min_{M} \quad & \frac{1}{2}\operatorname{tr}(MA_\lambda(V)) + \mu D_h(M; M_0)\\
\text{subject to} \quad & M_{ij} = m_{ij} \quad \forall\,(i,j)\in\mathcal{K}\\
& n_{\min}\mathbb{1}_{n_b} \le M\mathbb{1}_{n_b} \le n_{\max}\mathbb{1}_{n_b}\\
& n_{\min}\mathbb{1}_{n_b} \le M^T\mathbb{1}_{n_b} \le n_{\max}\mathbb{1}_{n_b}\,,
\end{aligned}$$

where $D_h(M; M_0) = h(M) - h(M_0) - \langle\nabla h(M_0), M - M_0\rangle$ is the Bregman divergence of the entropic regularizer $h(M) = \sum_{ij} M_{ij}\log(M_{ij})$, and the matrix $M_0$ is an initial guess for the optimizer. In addition, the scalars $m_{ij}$ are the known entries of $M$, where $i, j \in \mathcal{K} := (\mathcal{S}\times\mathcal{S}) \cup \{(1,1), \ldots, (n,n)\}$. This problem may be viewed as an entropy-regularized optimal transport problem (Sinkhorn and Knopp, 1967; Peyré and Cuturi, 2019). We optimize its dual via alternating minimization.

## 4. Numerical Illustrations

In the numerical illustrations we focus on the case when either none or only a few labels are known. The goal is to demonstrate that the proposed approach can make use of unlabeled data in order to decrease the classification error relative to a model trained with only labeled data.

### 4.1 Setup

In the numerical illustrations we focus on the dataset MAGIC (Bock et al., 2004). MAGIC contains measurements related to 19,020 simulated particles observed by a gamma telescope. The goal is to distinguish between gamma particles and hadrons. Since this dataset does not have a train/test split, we randomly split the data 75%/25% into train/test sets, and further split the training dataset 80%/20% into training and validation sets. In the illustrations we will focus on the case where the classes are balanced, i.e., where $n_{\min} = n_{\max}$. Therefore,

**Table 1**: Classification error of the kernel network on the test set of the MAGIC dataset.

| | # labeled observations | | |
|---|---|---|---|
| | 0 | 50 | 100 |
| Random initialization | 0.34 | 0.25 | 0.23 |
| Supervised initialization | N/A | 0.25 | 0.24 |
| Our method | 0.26 | 0.24 | 0.22 |

we randomly deleted 5,644 observations from the dataset that had label 1. We standardized each of the ten features prior to training. The network we use is a single-layer kernel network. This network approximates a Gaussian RBF kernel using the Nyström method with 32 landmarks (Williams and Seeger, 2000). The bandwidth was set using the median pairwise distance rule of thumb.

The initialization, training, and evaluation proceed as follows. The parameters $V$ of the Nyström approximation are initialized by randomly sampling from the inputs. During training we use all of the available labeled data at each iteration. However, we set the batch size for the unlabeled data to 4096. We train on only the labeled data for the first 100 iterations. Afterward, we train on both the labeled and the unlabeled data for another 400 iterations. In order to evaluate the performance of the learned features, we first either run spectral clustering (in the case of no labeled observations) or 1-nearest neighbor (in the case of some labeled observations) in order to estimate the labels of the unlabeled observations. We then train a regularized least squares classifier after combining the labeled data and the unlabeled data with the estimated labels.

### 4.2 Results

Table 1 presents the average classification error across 10 trials when varying the number of labeled observations. We compare the error of our method to (a) the error when the network parameters are fixed to their random values at initialization ("random initialization"); and (b) the error when the network is trained using only the available labeled data ("supervised initialization"). From the results we can see that, on average, our method outperforms both the random initialization and the supervised initialization in each setting. Moreover, this difference in performance is larger when the number of labeled observations is smaller. For example, our method performs 24% better than the random initialization when there are no labeled observations. However, this difference drops to 4% when there are 100 labeled observations.

## 5. Conclusion

In this work we proposed a framework for jointly learning feature representations and performing clustering regardless of the level of supervision. The objective function gracefully interpolates between unsupervised clustering and supervised classification objectives, depending on the ratio of labeled to unlabeled data. It recovers discriminative clustering when no labeled data exists and supervised classification when no unlabeled data exists. The numerical illustrations demonstrated that improvements can be obtained relative to a purely supervised alternative.

## Acknowledgements

The authors would like to gratefully acknowledge support from NSF CCF-1740551, NSF DMS-1810975, the program "Learning in Machines and Brains" of CIFAR, and faculty research awards. This work was first presented at the Women in Machine Learning Workshop in December 2019, for which the first author received travel funding from NSF IIS-1833154.

## References

F. R. Bach and Z. Harchaoui. DIFFRAC: a discriminative and flexible framework for clustering. In *Advances in Neural Information Processing Systems*, pages 49–56, 2007.

O. E. Barndorff-Nielsen and D. R. Cox. *Inference and asymptotics.*, volume 52. London: Chapman and Hall, 1994.

R. Bock, A. Chilingarian, M. Gaug, F. Hakl, T. Hengstebeck, M. Jirina, J. Klaschka, E. Kotrc, P. Savicky, S. Towers, A. Vaicilius, and W. Wittek. Methods for multidimensional event classification: A case study using images from a Cherenkov gamma-ray telescope. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 516(2):511–528, 2004.

O. Chapelle, B. Schlkopf, and A. Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.

C. Jones, V. Roulet, and Z. Harchaoui. End-to-end learning, with or without labels. *arXiv preprint arXiv:1912.12979*, 2019.

A. Joulin and F. R. Bach. A convex relaxation for weakly supervised classifiers. In *International Conference on Machine Learning*, 2012.

A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. J. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, pages 3239–3250, 2018.

G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21:343–348, 1967.

M. White and D. Schuurmans. Generalized optimal reverse prediction. In *International Conference on Artificial Intelligence and Statistics*, pages 1305–1313, 2012.

C. K. I. Williams and M. W. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, pages 682–688, 2000.

L. Xu, M. White, and D. Schuurmans. Optimal reverse prediction: a unified perspective on supervised, unsupervised and semi-supervised learning. In *International Conference on Machine Learning*, pages 1137–1144, 2009.