Notes on Poulis and Dasgupta (2017)

Karl Stratos

The paper's content can be greatly simplified as follows.

1 Setting

- We have a fixed topic model $\theta : \mathcal{X} \rightarrow \Delta^{T-1}$.
- There is some ground truth mapping $l : [T] \rightarrow [k] \cup \{?\}$. Let $P := \{t \in [T] : \; l(t) \neq ?\}$.
- Each labeled document (x, y) is generated by
  1. Drawing x from some distribution over $\mathcal{X}$, and
  2. Drawing $t \sim \text{Categorical}(\theta(x))$ and setting $y = l(t)$.

It's a graphical model. Capital letters denote random variables (so T is overloaded). The model defines

$$\Pr(X = x, Y = y) = \Pr(Y = y | X = x) \times \Pr(X = x)$$

where the conditional label distribution is given by

$$\begin{aligned}
\Pr(Y = y | X = x) &= \sum_{t=1}^{T} \Pr(T = t, Y = y | X = x) \\
&= \sum_{t=1}^{T} \Pr(T = t | X = x) \times \Pr(Y = y | X = x, T = t) \\
&= \sum_{t:\; y = l(t)} \theta_t(x)
\end{aligned}$$

2 Problem

The goal is to estimate l from labeled documents $(x_1, y_1) \ldots (x_n, y_n)$. Poulis and Dasgupta (2017) suggest finding the maximum-likelihood estimator

$$l^* = \arg\max_{l} \prod_{i=1}^{n} \sum_{t:\; l(t) = y_i} \theta_t(x_i)$$

They go on to show that finding any l that assigns nonzero probability to given data is NP-complete (Lemma A.1).

But the data is arbitrary. The documents are labeled adversarially, *not* by the model.

3 Solution

Now suppose that we receive n documents actually labeled by the model. Let

$$n_{ty} := \sum_{i=1}^{n} \mathbb{1}\left(t \sim \text{Categorical}(\theta(x_i)) \wedge y = y_i\right)$$

and $n_t := \sum_{y \in [k]} n_{ty}$. The estimator

$$\hat{l}(t) := \underset{y \in [k] \cup \{?\}}{\arg\max}\ n_{ty}$$

is then Bayes optimal and consistent in expectation. That is,

$$\mathbf{E}\left[\hat{l}\right](t) = \arg\max_{y}\ \Pr(Y = y | X = x, T = t) = l(t)$$

3.1 Finite Samples

Consider the expected value of n_{ty},

$$\begin{aligned}
\mathbf{E}\left[n_{ty}\right] &= \sum_{i=1}^{n} \Pr\left(T = t, Y = y | X = x_i\right) \\
&= \sum_{i=1}^{n} \theta_t(x_i) \times \Pr\left(Y = y | X = x_i, T = t\right)
\end{aligned}$$

Under the model, clearly we have

$$\Pr(Y = y | X = x, T = t) \geq 2\lambda \qquad \forall t \in P,\ y = l(t) \tag{1}$$

$$\Pr(Y = y | X = x, T = t) \leq \frac{\lambda}{2} \qquad \forall t \in [T],\ y \neq l(t) \tag{2}$$

for some $\lambda \leq 1/2$. (In particular, we can use $\lambda = 1/2$.) Let

$$n_0 := \frac{6}{\lambda} \log \frac{(k+1)T}{\delta}$$

Lemma A.3 shows that w.p. $\geq 1 - \delta$, for all t such that $n_t \geq n_0$ and for all y,

- $t \in P$: $n_{ty} > \lambda n_t$ if $y = l(t)$, $n_{ty} < \lambda n_t$ if $y \neq l(t)$
- $t \notin P$: $n_{ty} < \lambda n_t$ if $y \neq ?$

Thus w.p. $\geq 1 - \delta$, the following estimator

$$\hat{l}(t) = \begin{cases} y & \text{if } n_t \geq n_0 \text{ and } n_{ty} \geq \lambda n_t \\ ? & \text{otherwise} \end{cases}$$

is consistent for all t with $n_t \geq n_0$. For $t \in P$, we ensure $n_t \geq n_0$ in expectation if

$$n \geq \frac{n_0}{\min_{i=1}^{n} \theta_t(x_i)}$$

3.2 So What's the Paper Doing?

Essentially a bunch of unnecessary steps.

We see above that

1. Estimating l is hard if documents are allowed to be labeled adversarially.
2. Estimating l is *not* hard if documents are labeled by the model.

The "feature feedback" component of the paper confusingly mashes with the model. We exploit (1) and (2) exactly as before,

$$\begin{aligned} \Pr(Y = y|X = x, T = t) &\geq 2\lambda & \forall t \in P,\ y = l(t) \\ \Pr(Y = y|X = x, T = t) &\leq \frac{\lambda}{2} & \forall t \in [T],\ y \neq l(t) \end{aligned}$$

which are trivially true under the model. But we assume that *humans* generate topics T that satisfy these conditions.

4 A Method-of-Moments Estimator

Define $L \in \{0, 1\}^{(k+1)\times T}$ by

$$L_{y,t} = \begin{cases} 1 & \text{if } y = l(t) \\ 0 & \text{otherwise} \end{cases}$$

(With appropriate ordering, L is block diagonal.) Conditioning on documents x, each sample can be regarded as $y \sim \text{Categorical}(h(x))$ where

$$h(x) = L\theta(x)$$

Given n documents, let $H \in \mathbb{R}^{(k+1)\times n}$ be a matrix with columns $h(x) \in \Delta^{k+1}$, and let $\Theta \in \mathbb{R}^{T\times n}$ be a matrix with columns $\theta(x) \in \Delta^T$. Then

$$H = L\Theta$$

so if $n \geq \max\{k + 1, T\}$ and Θ is full-rank, we can recover the labeling by $L = H\Theta^+$ if we observe H.

A Lemmas

Lemma A.1. *The problem: given any topic model $\theta : \mathcal{X} \to [T]$ and labeled documents $(x_1, y_1) \ldots (x_n, y_n) \in \mathcal{X} \cup \{1, 2, ?\}$, find a topic-label mapping $l : [T] \to \{1, 2, ?\}$ such that for every $i = 1 \ldots n$ there is $t \in [T]$ with $\theta_t(x_i) > 0$ and $l(t) = y_i$. This problem is NP-complete.*

Proof. Let $\phi(z_1 \ldots z_q) = C_1 \wedge \cdots \wedge C_p$ be a 3-SAT instance with q Boolean variables $z_1 \ldots z_q \in \{0, 1\}$ and p clauses $C_1 \ldots C_p$ (e.g., $C_j = \bar{z}_3 \vee z_{10} \vee \bar{z}_1$). We construct a one-to-one correspondence between z_i values and topics by having $2q$ topics.

- Topics $1 \ldots q$ are associated with $z_1 \ldots z_q$.
- Topics $(q+1) \ldots 2q$ are associated with $\bar{z}_1 \ldots \bar{z}_q$.

Construct $2q$ labeled documents as follows. For each $i = 1 \ldots q$, let x be a document such that $\theta_i(x) = \theta_{q+i}(x) = 1/2$, then add $(x, 1)$ as the i-th labeled document and $(x, 2)$ as the $(q+i)$-th labeled document. If l is a valid topic-label mapping, then for the first q labeled documents it must assign label 1 to some $t \in \{i, q+i\}$ and for the next q labeled documents it must assign label 2 to some $t \in \{i, q+i\}$. This means for each $i \in [q]$, either

- $l(i) = 1$ and $l(q+i) = 2$, or
- $l(i) = 2$ and $l(q+i) = 1$.

Note that at this point, $l(i)$ is either 1 or 2 and can be treated like a Boolean variable. Construct p additional labeled documents as follows. For each $j = 1 \ldots p$, denote the three topics corresponding to the three literals in C_j by $j_1, j_2, j_3 \in [2q]$ and let x be a document such that $\theta_{j_1}(x) = \theta_{j_2}(x) = \theta_{j_3}(x) = 1/3$. Add $(x, 2)$ as the $(2q+j)$-th labeled document. To handle these last p labeled documents, a valid mapping l must assign $l(t) = 2$ for some $t \in \{j_1, j_2, j_3\}$ for every $j = 1 \ldots p$. A satisfying assignment to ϕ is now given by

$$z_i = \begin{cases} 1 & \text{if } l(i) = 2 \\ 0 & \text{if } l(i) = 1 \end{cases} \qquad\qquad \forall i = 1 \ldots q$$

Conversely, if we have a satisfying assignment to ϕ, a valid mapping for this topic model and dataset is given by setting $l(i) = 2$ and $l(q+i) = 1$ if $z_i = 1$ and $l(i) = 1$ and $l(q+i) = 2$ if $z_i = 0$. Thus 3-SAT and the considered problem are equivalent (the construction takes polynomial time). The problem is in NP since given l we can check its validity in polynomial time. $\square$

Lemma A.2. *Let $X = \sum_{i=1}^{n} X_i$ where $X_i \in \{0, 1\}$ are independent. Suppose $\boldsymbol{E}[X] \leq U$ and $\boldsymbol{E}[X] \geq L$. Then*

$$\Pr(X \geq 2U) \leq \exp\left(-\frac{U}{3}\right)$$
$$\Pr\left(X \leq \frac{L}{2}\right) \leq \exp\left(-\frac{L}{8}\right)$$

Proof. Define $Y_U := X - \mathbf{E}[X] + U$. Note that $\mathbf{E}[Y_U] = U$ and $Y_U \geq X$. Thus

$$\Pr(X \geq 2U) \leq \Pr(Y_U \geq 2U) \leq \exp\left(-\frac{U}{3}\right)$$

where we use the multiplicative Chernoff $\Pr(Z \geq 2\mathbf{E}[Z]) \leq \exp\left(-\frac{\mathbf{E}[Z]}{3}\right)$. Define $Y_L := X - \mathbf{E}[X] + L$. Note that $\mathbf{E}[Y_L] = L$ and $Y_L \leq X$. Thus

$$\Pr\left(X \leq \frac{L}{2}\right) \leq \Pr\left(Y_L \leq \frac{L}{2}\right) \leq \exp\left(-\frac{L}{8}\right)$$

where we use the multiplicative Chernoff $\Pr\left(Z \leq \frac{\mathbf{E}[Z]}{2}\right) \leq \exp\left(-\frac{\mathbf{E}[Z]}{8}\right)$. ☐

Lemma A.3. *With probability at least $1-\delta$ the following holds. For all $t \in [T]$ and $y \in [k] \cup \{?\}$, either $n_t < n_0$ or*

- *$t \in P$: $n_{ty} > \lambda n_t$ if $y = l(t)$, $n_{ty} < \lambda n_t$ if $y \neq l(t)$*
- *$t \notin P$: $n_{ty} < \lambda n_t$ if $y \neq ?$*

Proof. Using (1) and (2), conditioning on the value of n_t,

$$\mathbf{E}[n_{ty}] \geq 2\lambda n_t \qquad \forall t \in P,\ y = l(t)$$
$$\mathbf{E}[n_{ty}] \leq \frac{\lambda}{2} n_t \qquad \forall t \in [T],\ y \neq l(t)$$

Then by Lemma A.2,

$$\Pr(n_{ty} \leq \lambda n_t) \leq \exp\left(-\frac{\lambda n_t}{4}\right) \qquad \forall t \in P,\ y = l(t)$$
$$\Pr(n_{ty} \geq \lambda n_t) \leq \exp\left(-\frac{\lambda n_t}{6}\right) \qquad \forall t \in [T],\ y \neq l(t)$$

Let

$$E_{ty} := (t \in P \wedge y = l(t) \wedge n_{ty} \leq \lambda n_t) \vee (t \in [T] \wedge y \neq l(t) \wedge n_{ty} \geq \lambda n_t)$$

Note that $\Pr(E_{ty} | n_t \geq n_0) \leq \frac{\delta}{(k+1)T}$. Apply the union bound as follows:

$$\Pr(\exists (t,y):\ n_t \geq n_0 \wedge E_{ty}) \leq \sum_{(t,y)} \Pr(E_{ty} | n_t \geq n_0) \leq \delta$$

☐

Lemma A.4. *Let $A \succ 0$. The dual of the norm $||\cdot||_A$ is $||\cdot||_{A^{-1}}$.*

Proof. Let $||\cdot||_*$ denote the dual of $||\cdot||_A$. Then

$$\begin{aligned} ||x||_*^2 := \max_{u:\ ||u||_A = 1} (x^\top u)^2 &= \max_{u:\ u^\top A u = 1} u^\top x x^\top u \\ &= \max_{v:\ ||v||_2 = 1} v^\top A^{-1/2} x x^\top A^{-1/2} v \\ &= \left|\left|A^{-1/2} x\right|\right|_2^2 \end{aligned}$$

where the last step uses the fact that the only positive eigenvalue of a rank-1 matrix $zz^\top$ is given by $||z||_2^2$ (with z as the eigenvector). ☐

Lemma A.5. *Let $A \succ 0$. The squared norm $||\cdot||_A^2$ is 2-strongly convex wrt. itself.*

Proof. Since $\nabla ||x||_A^2 = 2Ax$, we have

$$\langle \nabla ||x||_A^2 - \nabla ||y||_A^2,\ x - y \rangle = 2\langle Ax - Ay,\ x - y \rangle \geq 2\, ||x - y||_A$$

☐

B External Theorems

Theorem B.1 (Theorem 1, Kakade et al. (2009))**.** *The class of bounded linear models* $\mathcal{F} = \{w : ||w|| \leq W\}$ *where* $||\cdot||^2$ *is* σ*-strongly convex wrt. itself has the Rademacher complexity bounded as follows:*

$$\mathcal{R}_n(\mathcal{F}) \leq W \max_x ||x||_* \sqrt{\frac{2}{\sigma n}}$$