A Concave Regularization Technique for Sparse Mixture Models

Martin Larsson, Johan Ugander

Cornell University

Motivation

A common challenge for latent variable mixture models is a desire to impose sparsity. Since mixture distributions are constrained in their L_1 norm, L_1 regularization becomes toothless, and concave regularization becomes necessary.

Concave regularization tends to involve EM algorithms that must maximize a non-concave function in their M-step. We introduce a technique for circumventing this difficulty, using the so-called Mountain Pass Theorem to provide easily verifiable conditions under which the M-step is well-behaved despite the lacking concavity.

We also develop a correspondence between logarithmic regularization and what we term the pseudo-Dirichlet distribution, a generalization of the ordinary Dirichlet distribution well-suited for inducing sparsity.

A Challenge: Sparse MAP PLSA

Probabilistic Latent Semantic Analysis (PLSA) [1] assumes the following model for each (word, document, topic) triplet:

$$P(w,d,z \mid \theta)P(\theta) = P(w \mid z)P(z \mid d)P(d)P(\theta).$$

The corresponding regularized log-likelihood is then:

$$\ell(\theta) = \underbrace{\sum_{w,d} n(w,d) \log \Big[\sum_z P(w \mid z)P(z \mid d)\Big] + \sum_d n(d) \log P(d)}_{\ell_0(\theta)} + \log P(\theta)$$

where θ consists of the model parameters $P(w \mid z)$, $P(z \mid d)$, $P(d)$, and $n(w,d)$ counts the occurrences of word w in document d, and $n(d) = \sum_w n(w,d)$.
This leads to the following **EM algorithm**:

E-step: Find $P(z \mid w,d,\theta')$, the posterior distribution of the latent variable z, given (w,d) and a current parameter estimate θ'.

M-step: Maximize $Q(\theta \mid \theta') = Q_0(\theta \mid \theta') + \log P(\theta)$ over θ, where

$$Q_0(\theta \mid \theta') = \sum_d n(d) \log P(d) + \sum_{w,d,z} n(w,d)P(z \mid w,d,\theta') \log \Big[P(w \mid z)P(z \mid d)\Big].$$

The natural sparsity-inducing prior is the Dirichlet distribution $(\alpha < 1)$. In order to infer a PLSA model with sparse priors, there are then **two challenges**:

1. M-Step maximization is **non-concave for all sparse priors.**
2. The log-likelihood function is **unbounded for Dirichlet.**

Pseudo-Dirichlet: A Sparse Prior for Regularization

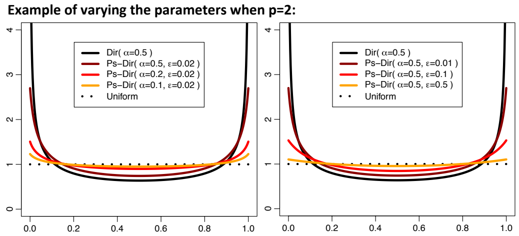
Definition:

A distribution on the simplex in $\mathbb{R}^p$ is said to follow a *pseudo-Dirichlet distribution* with concentration parameter $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_p) \in \mathbb{R}^p$ and perturbation parameter $\boldsymbol{\epsilon} = (\epsilon_1,\dots,\epsilon_p) \in \mathbb{R}^p_+$ if it has a density on the simplex given by

$$P(x_1,\dots,x_p \mid \boldsymbol{\alpha},\boldsymbol{\epsilon}) \propto \prod_{i=1}^p (\epsilon_i + x_i)^{\alpha_i - 1}$$

If $\alpha_i = \alpha$ and $\epsilon_i = \epsilon$ for all i, it is called *symmetric pseudo-Dirichlet.*

Example of varying the parameters when p=2:



Note that the Pseudo-Dirichlet density is bounded for $\epsilon > 0$ and $\alpha < 1$, while the Dirichlet density with $\alpha < 1$ is not [2]. Importantly, note also that α can here be negative. For $\epsilon = 0$ and $\alpha > 0$, it reduces to the ordinary Dirichlet density.

EM under Pseudo-Dirichlet

We wish to utilize the natural assumption that each document contains only a few topics. We formalize this sparsity assumption by placing Pseudo-Dirichlet priors on each vector $(P(z \mid d) : z \in \mathcal{Z})$ of topic probabilities. The resulting M-step maximization of $Q(\theta \mid \theta')$ is then additively separable with decoupled constraints:

$$Q(\theta \mid \theta') = \sum_z F_z(\theta \mid \theta') + \sum_d G_d(\theta \mid \theta') + H(\theta \mid \theta').$$

Here our prior only effects the maximization of $G_d(\theta \mid \theta')$. Denoting the parameters of the d[th] prior by α_d, ϵ_d, where $\alpha_d < 1$, the Lagrangian for this constrained optimization problem is:

$$\mathcal{L}_d(\boldsymbol{x};\lambda) = \sum_z \Big[(\alpha_d - 1)\log(\epsilon_d + x_z) + c_z \log x_z\Big] + \lambda\Big[1 - \sum_z x_z\Big]$$

where $x_z = P(z \mid d)$ and $c_z = \sum_w P(w \mid w,d,\theta')n(w,d)$.

Observe that **the Lagrangian is non-concave**.

Global maximum without concavity

To address the non-concavity of the Lagrangian, we provide the following theorem:

Theorem:

Assume that:
(i) every word w is observed in at least one document d,
(ii) $P(z \mid w,d,\theta') > 0$ for all (w,d,z), and
(iii) $n(d) > (1-\alpha_d)|\mathcal{Z}|$ for each d.
Then each Lagrangian $\mathcal{L}_d$ has a unique stationary point, which is the global maximum of the corresponding optimization problem.

Sketch of Proof

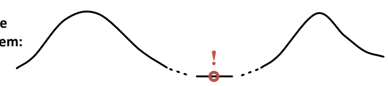
For each Lagrangian $\mathcal{L}_d$:

- Prove existence of a stationary point.
- Prove that the Hessian is negative definite at every stationary point.
- In particular, every stationary point is a strict local maximum.
- Apply the **Mountain Pass Theorem (Courant 1950) [3]**:

  If $\phi : \mathcal{O} \to \mathbb{R}$ is C^1, tends to $-\infty$ near $\partial\mathcal{O}$, and has two distinct strict local maxima, then it has a third stationary point that is not a strict local maximum.

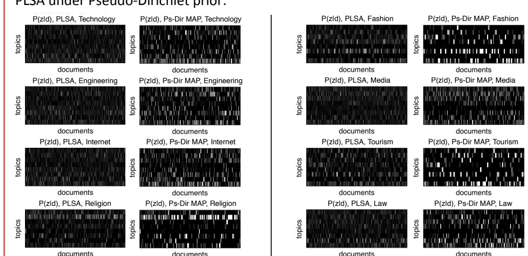
- There can only be one stationary point.

Proof by picture of the Mountain Pass Theorem:



Demonstration

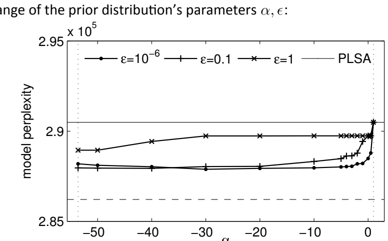
Training a topic model for a corpus of 2,406 blogs, showing ordinary PLSA vs. MAP PLSA under Pseudo-Dirichlet prior:



Using the 8 inferred topic vocabularies above, we generated 2,406 sparse topic distributions, one for each document. We used this to construct a new word-document distribution Q(w,d), from which we sampled N word-document pairs, producing a synthetic corpus. From this corpus we inferred a Pseudo-Dirichlet MAP model P(w,d), and evaluated the **model perplexity**,

$$\mathcal{P}(P(w,d)) = 2^{-\sum_{w,d} Q(w,d) \log_2 P(w,d)},$$

over a range of the prior distribution's parameters α, ϵ:



The dashed line indicates the perplexity $\mathcal{P}(Q(w,d))$ of the ground-truth distribution, which is a lower bound.

Future directions

- Can our regularization technique be applied successfully to other inference tasks analyzing mixture distributions with a fixed L_1 norm? Possible examples include portfolio optimization in finance and variable reduction in statistics.
- Similar sum-log regularization is used for sparse signal recovery ('compressed sensing') in [4]. Is there a unifying framework for these two approaches?
- We observe that PLSA is incapable of achieving true sparsity (in the L_0 sense). Can this or other methods be adapted to achieve true sparsity?

References

[1] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 42:177-196, 2001.
[2] A. Asuncion, M. Welling, P. Smyth, Y.W. Teh. On smoothing and inference for topic models. In Proc. UAI, 27-34, 2009.
[3] R. Courant. Dirichlet's principle, conformal mapping, and minimal surfaces. Interscience, New York, 1950.
[4] E.J. Candès, M.B. Wakin, S.P. Boyd. Enhancing sparsity by reweighted l_1 minimization. J Fourier Analysis and Applications, 14:877-905, 2008.

Paper: http://bit.ly/concavereg Poster: http://bit.ly/concavereg_poster