Modified Dirichlet Distribution: Allowing Negative Parameters to Induce Stronger Sparsity

Kewei Tu (ShanghaiTech University, China)

Dirichlet Distribution

The Dirichlet distribution (Dir) is defined over probability vectors $\boldsymbol{x} = \langle x_1, \ldots, x_n \rangle$ with positive parameter vector $\boldsymbol{\alpha} = \langle \alpha_1, \ldots, \alpha_n \rangle$:

$$\mathrm{Dir}(\boldsymbol{x}; \boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{n} x_i^{\alpha_i - 1}$$

When the elements in $\boldsymbol{\alpha}$ are less than one, Dir is a sparsity prior that prefers sparse probability vectors, with smaller $\boldsymbol{\alpha}$ values inducing stronger sparsity.

However, α_i is required to be positive in Dir because otherwise the normalization factor becomes divergent. Consequently, the strength of the sparsity preference is upper bounded. This becomes problematic when a strong prior is needed.

Modified Dirichlet Distribution

Modified Dirichlet distribution (mDir) allows the parameters in $\boldsymbol{\alpha}$ to become negative. To handle the divergent normalization factor, we require that each x_i must be lower bounded by a small positive constant ϵ.

$$\mathrm{mDir}(\boldsymbol{x}; \boldsymbol{\alpha}, \epsilon) = \begin{cases} 0 & \text{if } \exists i, x_i < \epsilon \\ \frac{1}{Z(\boldsymbol{\alpha}, \epsilon)} \prod_{i=1}^{n} x_i^{\alpha_i - 1} & \text{otherwise} \end{cases}$$

where we require $0 < \epsilon \leq \frac{1}{n}$ and do *not* require α_i to be positive.

Properties of mDir:

- The normalization factor $Z(\boldsymbol{\alpha}, \epsilon)$ is guaranteed to be finite.
- mDir is still conjugate to the multinomial distribution.
- mDir achieves very strong sparsity preference when $\boldsymbol{\alpha}$ is highly negative.
- ϵ can be seen as a smoothing factor that prevents any element in $\boldsymbol{x}$ from becoming too small.

Two algorithms for finding the mode of mDir:

Algorithm 1 Mode-finding of $\mathrm{mDir}(\boldsymbol{x}; \boldsymbol{\alpha}, \epsilon)$

$S \leftarrow \{i | \alpha_i \leq 1\}$
$T \leftarrow \emptyset$
repeat
  $T \leftarrow T \bigcup S$
  for $i \in T$ **do**
    $x_i \leftarrow \epsilon$
  end for
  $z \leftarrow \sum_{i \notin T} (\alpha_i - 1)$
  for $i \notin T$ **do**
    $x_i \leftarrow \frac{\alpha_i - 1}{z} \times (1 - \epsilon |T|)$
  end for
  $S \leftarrow \{i | x_i < \epsilon\}$
until $S = \emptyset$
return $\langle x_1, \ldots, x_n \rangle$

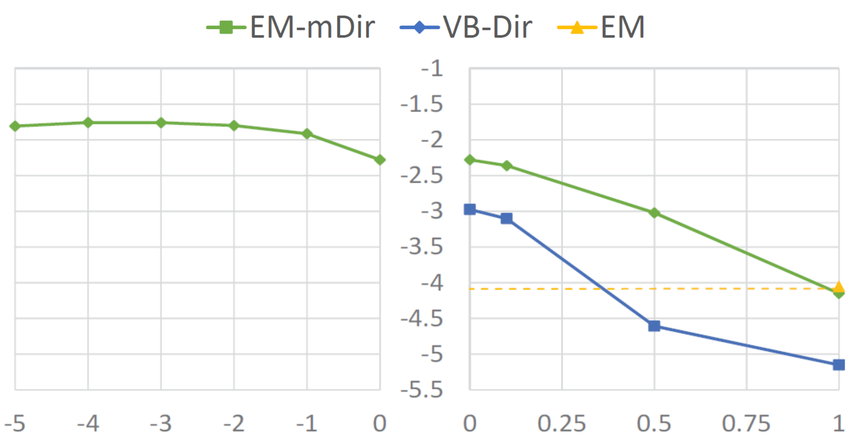
Algorithm 2 Fast mode-finding of $\mathrm{mDir}(\boldsymbol{x}; \boldsymbol{\alpha}, \epsilon)$

$\langle \alpha_{k_1}, \ldots, \alpha_{k_n} \rangle \leftarrow \langle \alpha_1, \ldots, \alpha_n \rangle$ in ascending order
$s_n \leftarrow \alpha_{k_n} - 1$
for $i = n-1, \ldots, 1$ **do**
  $s_i = s_{i+1} + \alpha_{k_i} - 1$ ▷ So $s_i = \sum_{j \geq i} (\alpha_{k_j} - 1)$
end for
$t \leftarrow 0$
for $i = 1, \ldots, n$ **do**
  $x_{k_i} \leftarrow \frac{\alpha_{k_i} - 1}{s_i} \times (1 - \epsilon t)$
  if $x_{k_i} < \epsilon$ **then**
    $x_{k_i} \leftarrow \epsilon, \quad t \leftarrow t + 1$
  end if
end for
return $\langle x_1, \ldots, x_n \rangle$

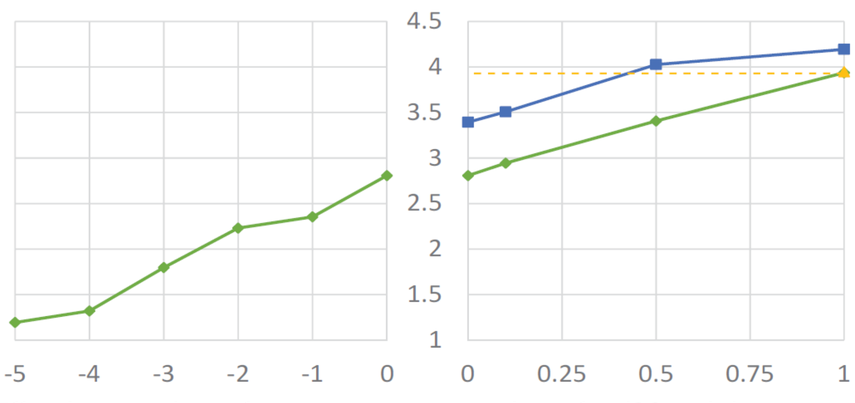
The time complexity of Algorithm 1 is $O(n^2)$ in the worst case, but it decreases to $O(n)$ when ϵ is small.

Algorithm 2 has time complexity $\Theta(n \log n)$ and can be more efficient than Algorithm 1 when both n and ϵ are large.

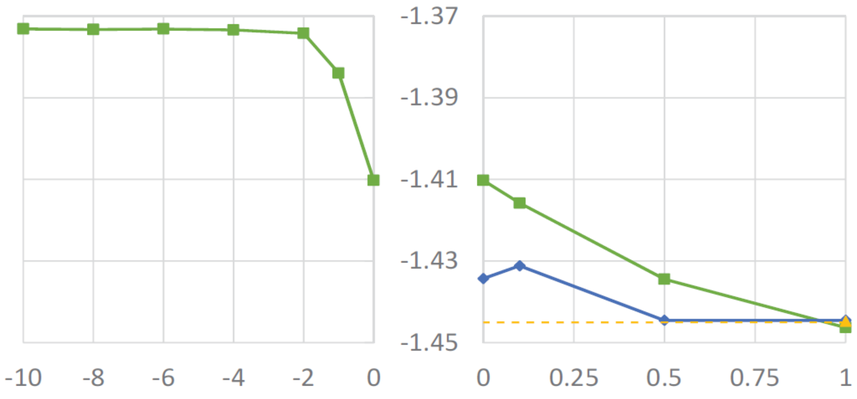
Exp 1: Learning Mixtures of Gaussians



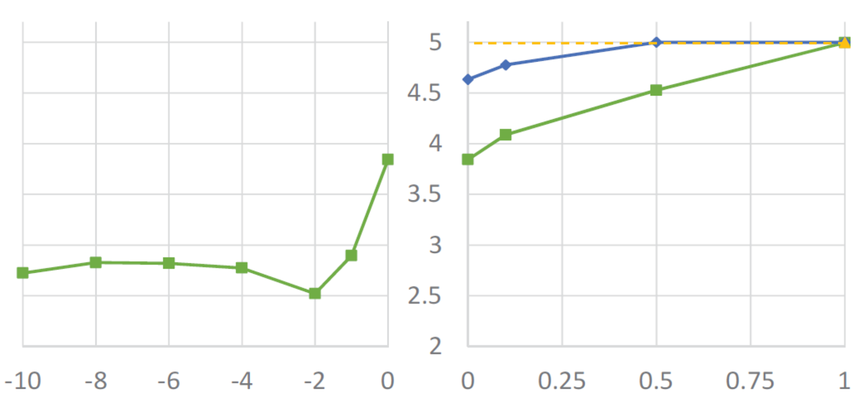
Test set log likelihood vs. value of α (20 training samples)



Effective number of components vs. value of α (20 training samples)

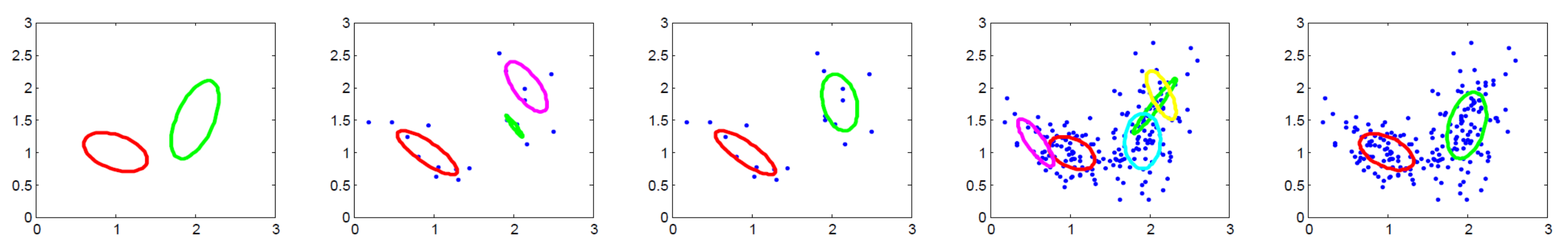


Test set log likelihood vs. value of α (200 training samples)



Effective number of components vs. value of α (200 training samples)

- **EM**: maximum likelihood estimation using expectation-maximization, which has no sparsity preference;
- **VB-Dir**: mean-field variational Bayesian inference with a Dir prior over the mixing probabilities, which is the most frequently used inference approach for Dir with $\alpha < 1$;
- **EM-mDir**: maximum a posteriori estimation using expectation-maximization with a mDir prior over the mixing probabilities.



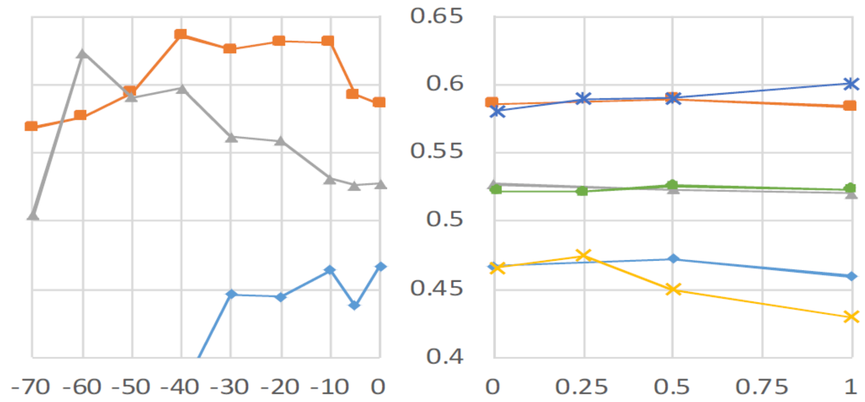
(a) Ground-truth (b) VB-Dir, $\alpha = 10^{-5}$ (c) EM-mDir, $\alpha = -2$ (d) VB-Dir, $\alpha = 10^{-5}$ (e) EM-mDir, $\alpha = -30$

The ground-truth model and four typical models learned by VB-Dir and EM-mDir from 20 and 200 training samples.

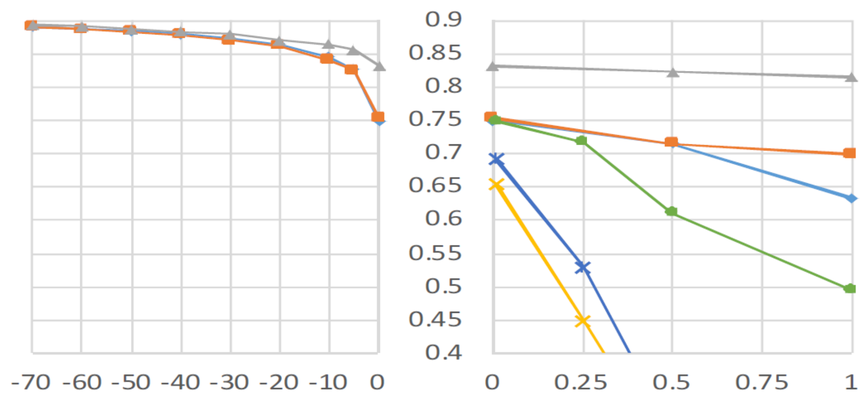
Exp 2: Unsupervised Dependency Parsing

We tried to learn a dependency model with valence (DMV) from the Wall Street Journal corpus. The number of dependency rules in DMV is small relative to the training corpus size.

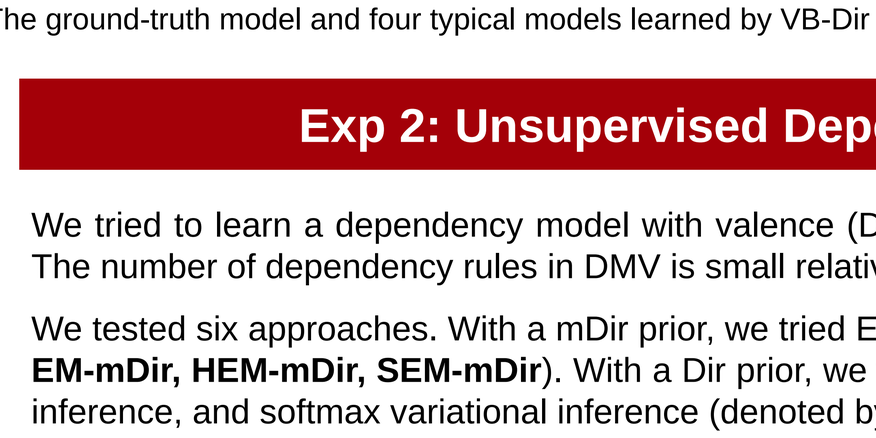
We tested six approaches. With a mDir prior, we tried EM, hard EM, and softmax-EM (denoted by **EM-mDir, HEM-mDir, SEM-mDir**). With a Dir prior, we tried variational inference, hard variational inference, and softmax variational inference (denoted by **VB-Dir, HVB-Dir, SVB-Dir**).



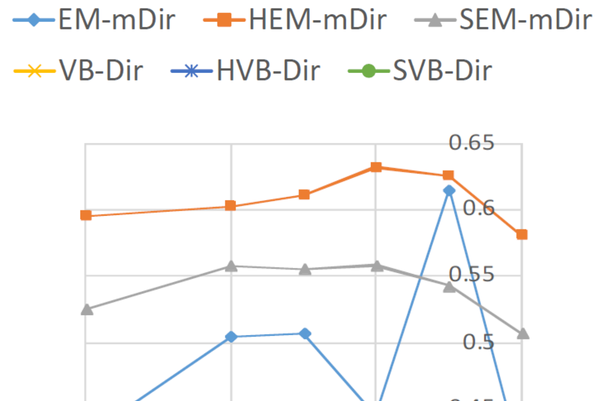
Test set log likelihood vs. value of α (200 training samples)



Sparsity of the learned grammars vs. value of α



Parsing accuracy vs. value of α



Parsing accuracy vs. value of ϵ

Scan for code & paper link