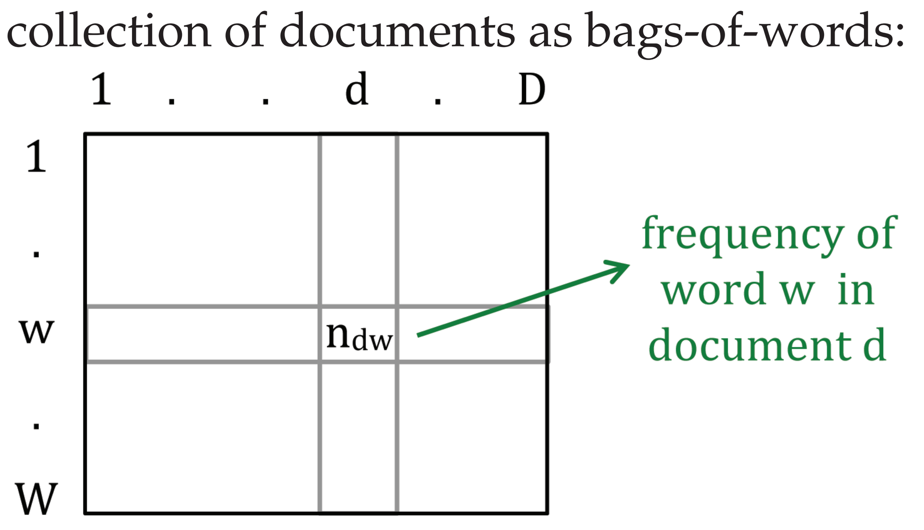
Additive Regularization for Learning Interpretable Topic Models

Anna Potapenko, anya_potapenko@mail.ru

Task of Topic Modeling

Given:
A collection of documents as bags-of-words:



Model:
Assume that each observable word w in document d refers to latent topic t and

$$p(w|d) = \sum_{t \in T} p(w|t)\, p(t|d) = \sum_{t \in T} \phi_{wt}\, \theta_{td}$$

Find:
- $\phi_{wt} \equiv p(w|t)$ — words for each topic,
- $\theta_{td} \equiv p(t|d)$ — topics for each document,

resulting in $p(w|d)$ close to $\hat{p}(w|d) \propto n_{dw}$.

Problems

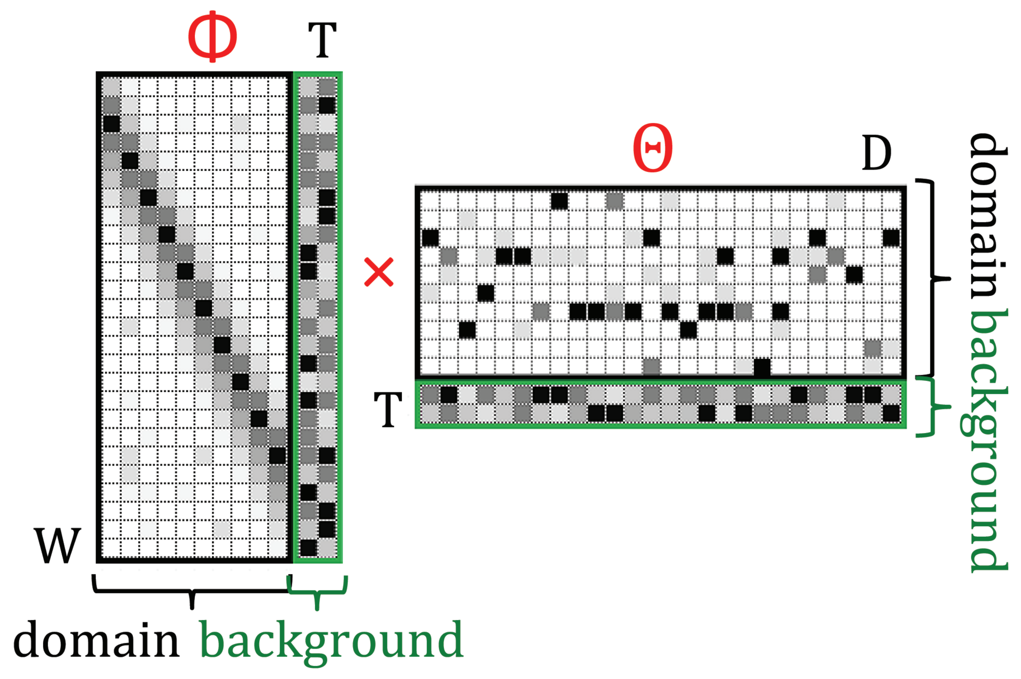
- The task is ill-posed and solution is non-unique: $\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$
- The choice depends on random initialization rather than on required properties of the solution such as interpretability or sparsity.

Contributions

- We impose additional requirements on Φ and Θ in form of regularization penalty terms for more reasonable choice of the solution.
- Within the framework of additive regularization we work out the topic model that outperforms classic PLSA model in interpretability.

Hypothesis and Regularizers

We propose a set of regularizers to meet the hypothetical structure of well-interpreted topics:



A set of topics is split into two categories: *domain* and *background* topics.

1. **Sparsing.** Each *domain* topic contains a small number of domain-related words, each document relates to a few topics:

$$R_1(\Phi, \Theta) = \sum_t KL(\phi_t, \beta) + \sum_d KL(\theta_d, \alpha),$$

where α, β are uniform (u), or β is a background distribution of words in language/collection (b).

2. **Decorrelating.** *Domain* topics are significantly different, in other words correlation is low.

$$R_2(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt}\phi_{ws}$$

3. **Smoothing.** *Background* topics accumulate general lexis and neutral words. They are inherent in all documents and contain all words of the vocabulary with nonzero probability.

$$R_3(\Phi, \Theta) = -\sum_t KL(\phi_t, \beta) - \sum_d KL(\theta_d, \alpha)$$

Optimization Technique

Regularized Likelihood Maximization:

$$\underbrace{\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td}}_{L(\Phi,\Theta)} + \underbrace{\sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta}$$

EM-algorithm – iterative process, alternating:
E-step (Bayes' Rule):

$$p(t|d,w) \propto \phi_{wt}\theta_{td}$$

M-step (maximization):

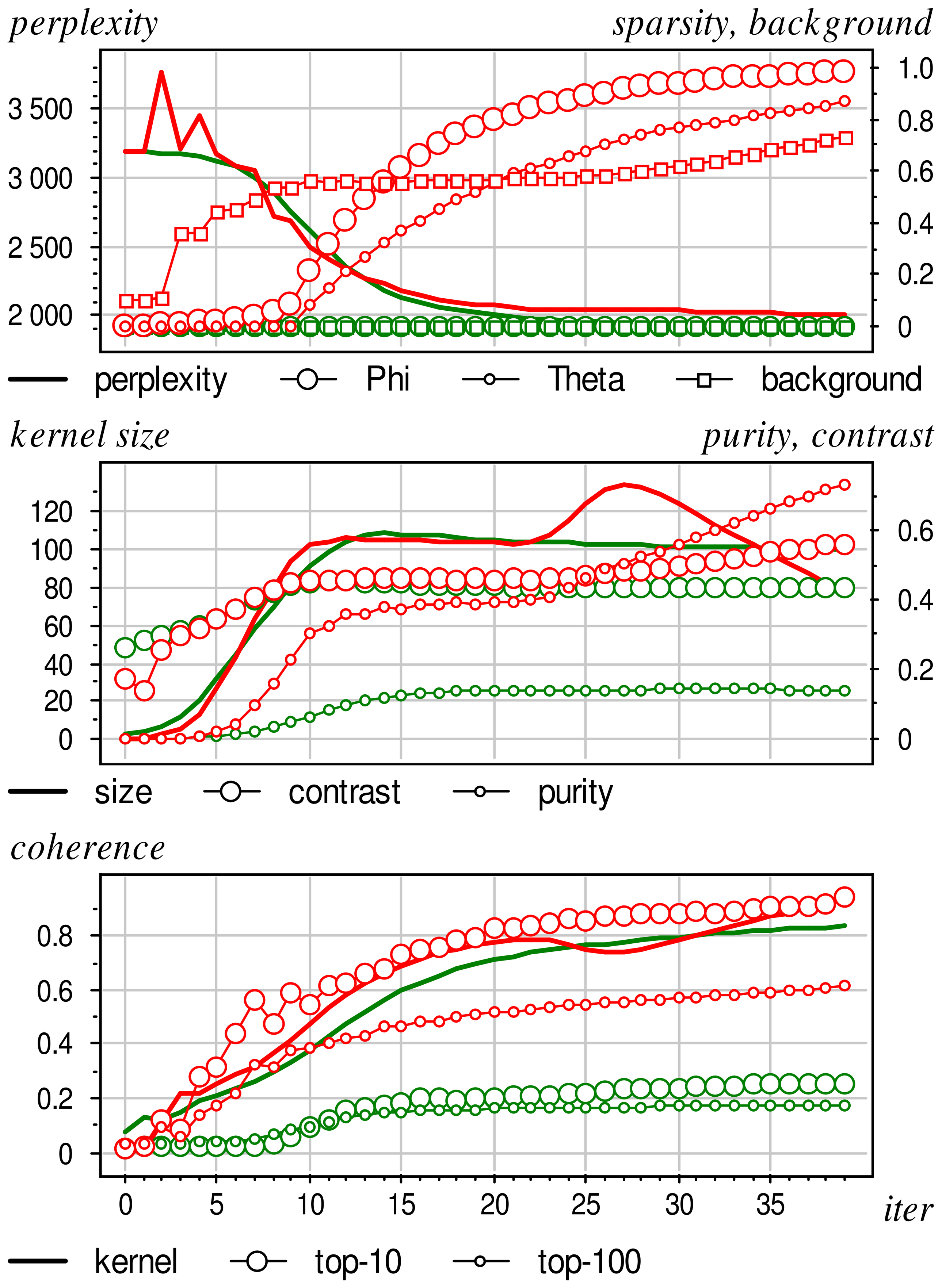
$$\phi_{wt} \propto \left(\sum_{d \in D} n_{dw} p(t|d,w) + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)_+$$

$$\theta_{td} \propto \left(\sum_{w \in W} n_{dw} p(t|d,w) + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+$$

Quality Measures

1. **Perplexity** (is based on likelihood):
   $P = \exp\left(-\frac{1}{n} L(\Theta, \Phi)\right)$
2. Weight of **background** topics (B)
3. **Sparsity**: proportion of zeros in Φ (S_Φ), Θ (S_Θ)
4. **Size of topic kernel** (distinguishing words):
   size = $|W_t|$, $\quad W_t = \{w \colon p(t|w) > 0.25\}$
5. **Topic contrast**: con = $\frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$
6. **Topic purity**: pur = $\sum_{w \in W_t} p(w|t)$
7. **Coherence**: $C = \frac{2}{k(k-1)} \sum_{j=2}^{k} \sum_{i=1}^{j} \mathrm{PMI}(w_i, w_j)$
   $C^{\text{ker}}, C^{10}, C^{100}$: kernel, top-10, top-100 words

Experiments and Results



green: PLSA (standard), red: ARTM (introduced)

Dataset: standard NIPS collection (1700 papers).

- **Figure.** A combination of regularizers improves all measures of sparsity and interpretability and doesn't virtually affect the perplexity.

  To adjust regularization coefficients we observe the state of the model by a set of measures during the iteration process.

- **Table.** In ARTM approach any combination of regularizers is valid. Comparison of all possible combinations by a set of measures shows that the combination of sparsing (Sp) + decorrelating (Dc) + smoothing (Sm) gives the best result.

Sp	Dc	Sm	P	B	S_Φ	S_Θ	size	con	pur	C^{ker}	C^{10}	C^{100}
–	–	–	1923	0.00	0.000	0.000	100	0.43	0.14	0.84	0.25	0.17
u	–	–	2114	0.24	0.957	0.867	71	0.53	0.20	0.91	0.25	0.18
b	–	–	2507	0.51	0.957	0.867	151	0.46	0.56	0.71	0.60	0.58
–	+	–	2025	0.57	0.561	0.000	109	0.46	0.38	0.82	0.94	0.56
u	–	+	1961	0.25	0.957	0.867	64	0.51	0.20	**0.97**	0.26	0.18
b	–	+	2025	0.49	0.957	0.867	128	0.45	0.52	0.77	0.55	0.55
–	+	+	1985	0.59	0.582	0.000	97	0.46	0.39	0.87	0.93	0.57
u	+	+	2010	0.73	**0.980**	0.867	78	**0.56**	0.73	**0.94**	0.94	0.62
b	+	+	2026	**0.80**	**0.979**	0.867	111	0.52	**0.89**	0.81	**0.96**	**0.83**

- **Word lists.** General lexis words are grouped into background topics. Domain topics are free of them and contain domain-related words with high probabilities (kernel words are red).

PLSA:	face, images, faces, recognition, set, image, based, hme, facial, representation, view, figure, model, experts, network, human, expert, space, examples, system
ARTM:	face, faces, facial, Cottrell, Pentland, gesture, lane, emotion, person, steering, appearance, Baluja, setpoint, camera, tracking, pose, Pomerleau, mouth, Darrell, lighting
Background:	model, data, models, parameters, noise, neural, mixture, prediction, set, gaussian, likelihood, networks, test, figure, training, performance, network, number, input, results

References: *Vorontsov K. V., Potapenko A. A.* Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. — Analysis of Images, Social Networks, and Texts (AIST-2014). — LNCS, Springer.