Topic Modeling for Scientific Documents

Jethro Kuan
WING-NUS

Abstract

Topic models are statistical models, used to discover abstract topics that occur in a collection of documents. Topic modeling can be applied to scientific documents for data discovery and navigation, among others. In this paper, we discuss the predominant technique for topic modeling, the Latent Dirichlet Allocation, and its extensions. We train these models on a dataset of scientific documents to observe their differences, and discuss their strengths and shortcomings.

1 Introduction

In this information age, the availability of knowledge is insufficiently met with the tools to navigate it. Archives such as Arxiv see an exponential growth in the number of documents being hosted. Developing new tools for browsing, and searching these documents is a technological challenge that calls for research in statistical modeling.

Rather than relying on keyword matching, tools like Semantic Scholar use machine learning to process scientific documents and discover meaningful structure, empowering researchers to discover papers more relevant to their work. One such statistical modeling technique is topic modeling.

Topic modeling attempts to discover abstract topics within documents. Topic models capture the intuition that if a document is of a particular topic, then words from the topic should appear more frequently.

A topic model trained on a corpora of scientific documents could learn topics such as "Neural Networks", "Biology" and "Medicine". Topics attribute a high probability to words that relate to the topic. For example, a "Neural Networks" topic would give attribute a high probability to "classifiers", and a low probability to "bacteria". The quotations around the topics are to make explicit that the labels are human interpretations of what these topics may be. Topic modeling is an unsupervised problem, and is often trained on unlabeled data. Automatic labeling of these topics is an active area of research [13, 12].

We can also view these abstract topics as a form of clustering. Researchers are better able to navigate the large corpora of scientific knowledge, by exploring documents that have similar topic proportions.

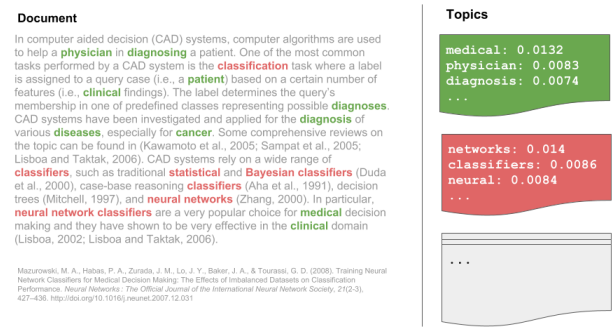

Figure 1: According to the trained topic model, topics like "Medicine" and "Neural Networks" generate the colored words of the document

Suppose we train a topic model on a corpora of scientific documents. The trained model would discover that the document in Figure 1 relates to "Medicine" and "Neural Networks", because of the frequent appearance of highly probable words in these topics.

The earliest known technique for topic modeling is Latent Semantic Indexing (LSI) [6]. The first probabilistic approach, Probabilistic Latent Semantic Analysis (pLSA), was proposed in 1999 [10]. It was only in 2003 when Blei et al. introduced Latent Dirichlet Allocation (LDA), which remains to this date the predominant technique for topic modeling.

Most of the state-of-the-art topic models are generative models. In generative models, latent variables govern the generative process of a document. A document is produced from a distribution of topics, and the topics are distributions over the vocabulary of the corpora.

In this paper, we survey the landscape of probabilistic topic models. LDA and two of its variants are discussed in detail. We train these models on a dataset of scientific documents, and present our results.

2 LDA

Latent Dirichlet Allocation is widely considered to be the simplest topic model. LDA models each document as a mixture of topics, where a topic β_k is a probability distribution over a fixed vocabulary of terms. Training LDA requires fixing the number of topics, K. We can draw the graphical model for LDA as in Figure 4.

At the risk of being pedantic, we explain the graphical model. η is the topic hyperparameter, which produces a topic distribution β_k of the Dirichlet family. There are a total of K topics. Similarly, α is a hyperparameter that produces the per-document topic proportions θ_d. These Dirichlet distributions are of dimension $K-1$, because there are a total of K topics. There are D such topic proportions, where D is the total number of documents. $Z_{d,n}$ is the per-word topic assignment, drawn from the particular θ_d. Finally, $W_{d,n}$ is the nth word in the dth document, an observed variable. It is simple to see that $P(W_{d,n}|Z_{d,n},\beta_k) = \beta_{z_{d,n},w_{d,n}}$.

A high α value encodes the belief that documents contain a mixture of many topics, rather than being largely represented by a few topics. Similarly, a high η value encodes the belief that topics has high probability for a large number of words in the vocabulary.

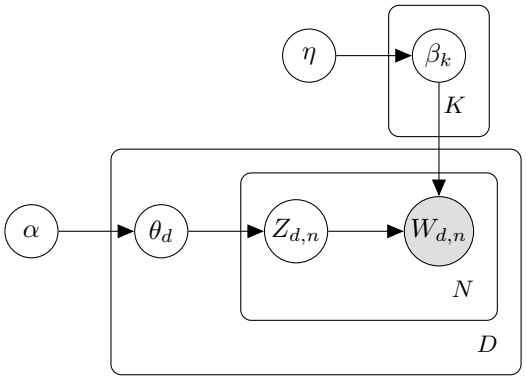

Figure 2: Plate notation for LDA.

We can fully specify the model by looking at the joint distribution $\mathcal{J}$ of all the latent and observed variables.

$$\mathcal{J} = \left(\sum_{k=1}^{K} P(\beta_k|\eta)\right)\left(\sum_{d=1}^{D} P(\theta_d|\alpha)\right) \left(\sum_{n=1}^{N} \theta_{d,Z_{d,n}}\beta_{Z_{d,n},W_{d,n}}\right) \quad (1)$$

The generative process of a document is as follows:

If we fix the topic distributions $\beta_{1:K}$, we can compute the per-document posterior θ given the document.

$$P(\theta|w_{1:n},\alpha,\beta_{1:K}) = \frac{P(\theta|\alpha)\prod_{n=1}^{N} P(z_n|\theta)P(w_n|z_n,\beta_{1:K})}{\int_\theta P(\theta|\alpha)\prod_{n=1}^{N}\sum_{z=1}^{K} P(z_n|\theta)P(w_n|z_n,\beta_{1:K})} \quad (2)$$

Algorithm 1 Generative Process of LDA

1: **for** each document d in D **do**
2: Draw topic distribution $\theta_d \sim Dir(\alpha)$
3: **for** each word n_d in N_d **do**
4: Sample topic $z_{d,n} \sim Multinomial(\theta)$
5: Sample word $w_{d,n} \sim Multinomial(\beta_{z_{d,n}})$

The denominator is intractable to compute, due to the coupling between θ and β under the multinomial assumption. [4]. Hence, we rely on techniques for approximate inference of the posterior. We discuss these techniques in section 6.

Why does LDA work? The Dirichlet distribution encourages sparsity, encoding the belief that the document-topic distribution has few topics per document, and the topic-word distribution has few words per topic. These two beliefs work against each other, and LDA discovers this sparsity balance, which gives rise to the structure of the textual data.

2.1 Statistical Assumptions

In Bayesian statistics, we cannot make inference without statistical assumptions. LDA makes several statistical assumptions, which we discuss below.

First, LDA models documents as "bag-of-words": words within the document are exchangeable. [4] I think this is a reasonable assumption to make, if the task is to discover themes within the document.

Second, LDA assumes that the order of documents do not matter. Exchangeability of both words and documents allows LDA to model the joint distribution as a mixture model. I believe this assumption to be invalid in the domain of scientific documents. The meaning of keyphrases used in scientific literature change over time. For example, the landscape of research neural networks is vastly different now, as compared to the 1990s, and LDA will fail to capture these differences.

Third, the use of the Dirichlet distribution also encodes statistical assumptions about the correlation between topics. Under the Dirichlet, components of θ_d are nearly independent. This leads to the modeling assumption that the presence of one topic is not correlated with the presence of another [3]. As explained by Blei and Lafferty, this assumption is strong and unrealistic in the domain of scientific documents. An article about genetics is also highly likely to be about health and disease, and unlikely to be about astronomy [3].

Variants of LDA relax these statistical assumptions, or make other assumptions in place. We discuss Dynamic Topic Modeling (DTM) in section 4 and Correlated Topic Modeling(CTM), and briefly mention the rest in the appendix.

3 Correlated Topic Modeling

The Correlated Topic Model (CTM) addresses the model assumption that topic proportions are not cor-

related. In scientific documents, this assumption is unlikely to hold true.

Instead of drawing from a Dirichlet distribution, the CTM uses a logistic normal distribution. CTM draws a real valued random vector from a multivariate Gaussian, and maps it to the simplex to obtain a multinomial parameter [3]. The $K \times K$ covariance matrix Σ models dependencies between the topics. This tweak is evident in the generative process shown in Algorithm 2, contrasting it with Algorithm 1.

Algorithm 2 Generative Process of CTM

1: **for** each document d in D **do**
2: Draw topic distribution $\beta_d \sim \mathcal{N}(\mu, \Sigma)$
3: **for** each word n_d in N_d **do**
4: Sample topic $z_{d,n} \sim Multinomial(\pi(\beta_d))$
5: Sample word $w_{d,n} \sim Multinomial(\beta_{z_{d,n}})$

π maps the multinomial parameters to the mean parameters, $\pi(x) = \frac{e^x}{\sum_i e^{x_i}}$

We can also visualize CTM with a graphical model.

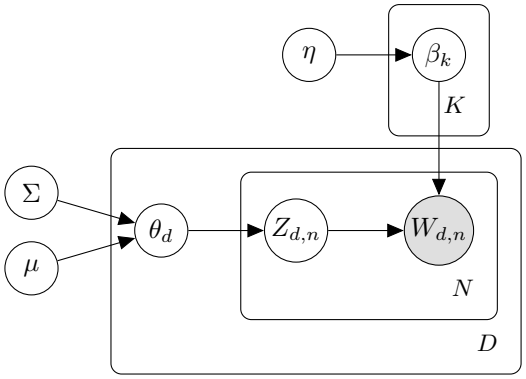

Figure 3: Plate notation for CTM.

The Multinomial and Gaussian distributions are not conjugate distributions. This causes difficulty in inference via Gibbs sampling, and variational inference is used instead.

4 Dynamic Topic Modeling

Dynamic Topic Modeling (DTM) was proposed to remove the assumption that documents are *exchangeable*. [2]

The order of documents are important for scientific documents, since both the content, and the meaning of words evolve over time.

In DTM, data is divided by discrete time slices. The topics associated with time slice t evolve from the time slice $t-1$. Because the Dirichlet distribution is not amenable to sequential modeling, we use the Gaussian distribution to model the sequence of random variables.

The generative process for time slice t is as follows:

Algorithm 3 Generative Process of DTM

1: Draw topic distribution $\beta_t|\beta_{t-1} \sim N(\beta_{t-1}, \sigma^2 I)$
2: Draw $\alpha_t|\alpha_{t-1} \sim N\left(\alpha_{t-1}, \delta^2 I\right)$
3: **for** each document w **do**
4: Draw $\eta_{w,t} \sim N\left(\alpha_t, a^2 I\right)$
5: **for** each word at position n **do**
6: Sample topic $z_{t,n} \sim Multinomial(\pi(\eta_{w,t}))$
7: Sample word $w_{t,d,n} \sim Multinomial(\beta_{t,z,n})$

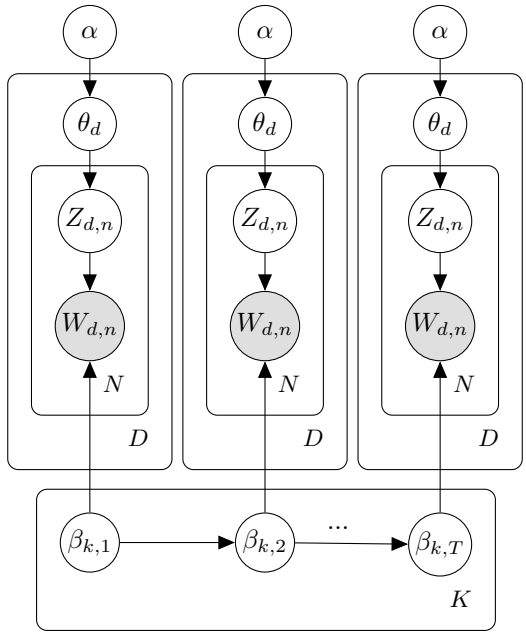

Figure 4: Plate notation for DTM.

Similar to CTM, the use of the logistic normal distribution causes variational inference to be the preferred approximate inference technique.

Further extensions of this approach include the continuous Dynamic Topic Models (cDTM), which removes the discretization of the time slices [17]. This model has been used to predict the timestamp of documents.

5 Bayesian Non-parametric Models

One important component we have yet to discuss is the choice of K, the number of topics used to model the corpora. Notice that the value of K affects model complexity. LDA, CTM and DTM are Bayesian parametric models: documents are modeled as mixtures of K distributions. The choice of K is therefore also an important parameter that requires tuning. This is a pitfall of parametric models. A misfit between the complexity of the model and the amount and quality of data available can lead to severe underfitting or overfitting [15].

Bayesian non-parametric models provide an elegant solution to this issue. The non-parametric approach allows the model to grow in complexity as more data is observed. These non-parametric models have also been

extended to hierarchies of topics, moving from more general to more concrete [5].

Bayesian non-parametric models has seen some success in recommendation systems [7], a popular modeling technique in production systems because of its ability to adapt with growing data size.

6 Inference Methods

Approximating intractable probability densities is a well-studied problem in modern statistics. This problem arises often in Bayesian statistics, where computing posterior probablity densities in requires inference over latent variables. Many learning algorithms have been developed, including collapsed Gibbs Sampling, Variational Inference, Collapsed Variational Inference, and MAP estimation. Each of these approximation techniques have their own strength and shortcomings.

The two inference methods are briefly discussed below, and a comparison between them relegated to the appendix in subsection 6.3.

6.1 MCMC Sampling

Historically, Markov Chain Monte Carlo (MCMC) sampling has been the dominant technique for approximating posterior densities. In MCMC, we construct an ergodic Markov chain on the latent variable z, whose stationary distribution is the posterior $P(z|x)$. Samples are drawn from the stationary distribution, and used to approximate the posterior empirically.

In Gibbs sampling, the space of the Markov Chain is the space of the configurations of the hidden variables. In Gibbs sampling, the next state is reached by sequentially sampling all variables from the distribution, conditioned on all the current sampled values. After a "burn-in" period, the samples would be drawn from the posterior distribution.

Algorithm 4 Gibbs Sampling

1: $x^0 \leftarrow q(x)$
2: **for** $i = 1, 2, 3, \ldots$ **do**
3: **for** $d = 1, 2, 3, \ldots, D$ **do**
4: $x_d^i \sim P(X_1 = x_1 | X_k = x_k^{i-1}$ for $k = \{1..n \setminus d\})$

A full treatment of Gibbs Sampling applied to LDA can be found in [8].

6.2 Variational Inference

In variational inference, the posterior distribution over a set of unobserved variables $p(Z|H)$ is approximated by a distribution $q(Z)$, selected to be in a family that can approximately model the true posterior. Inference is performed by minimizing the distance between $p(Z|H)$ and $q(Z)$. One common metric used is the KL-divergence. Here, we discuss variational inference for LDA.

Mean field variational inference (MFVI) breaks the coupling between θ and z by introducing free variational parameters γ over θ and ϕ over z and dropping the edges between them. This results in an approximate posterior $q(\theta, z|\gamma, \phi) = q_\gamma(\theta) \prod_n q_\phi(z_n)$. This is illustrated in the graphical model in Figure 5.

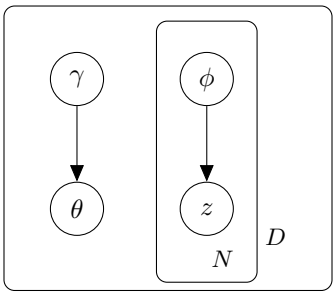

Figure 5: Graphical model for the approximate posterior for variational inference.

To best approximate the true posterior, we frame it as an optimization problem, minimizing L where:

$$L(\gamma, \phi|\alpha, \beta) = D_{KL}\left[q(\theta, z|\gamma, \phi)||p(\theta, z|\alpha, \beta)\right] - \log p(w|\alpha, \beta) \quad (3)$$

This optimization has closed form coordinate descent equations for LDA, because the Dirichlet is conjugate to the Multinomial distribution. This computational convenience comes at the expense of robustness, making it difficult to apply to other more complicated topic models [4].

Both DTM and CTM show that the simplicity of LDA makes it promising as a base model that can be adapted to the problem domain. CTM and DTM both relax assumptions that should lead to improvements in the topic distributions generated. The evaluation of topic models is discussed in section 8.

6.3 Choosing an Inference Method

How do we know which technique to use to approximate the posterior density? MCMC methods are computationally more intensive, but provide samples that are approximately exact from the target posterior density. In contrast, VI methods view the problem as an optimization problem, which allows it to utilize efficient learning algorithms such as stochastic optimization. This is much quicker to compute, and is suited for larger datasets.

However, MCMC methods, cover a large family of sampling methods. Gibbs sampling requires that the prior and posterior are conjugate distributions. When this is not possible, such as in DTM, VI methods can perform better than other methods in the MCMC family.

A closer look at the different inference approximation algorithms, however, shows that the performance differences can be explained away by setting certain smoothing hyperparameters [1].

7 Example Topic Model Training

To see topic modeling in action, we train them on a dataset of scientific papers. This dataset was obtained from Kaggle, and contains all papers from the NIPS conference, along with the paper metadata[1]. Jupyter notebooks for data exploration and model training can be found here [2].

We choose 2 models to train, LDA and DTM, because their implementations are readily available in Gensim, a popular topic modeling library.

To clean the raw text, we lemmatize, remove stop words, and keep only the nouns. After removing tokens that rarely occur, or occur too often to be meaningful, the corpus of 7241 documents produced a dictionary of 54254 unique tokens. We set $K = 30$ topics.

LDA trained in 65 seconds, and the results are presented in the appendix (Figure 7).

1	mixture, covariance, likelihood, component, density, estimation
2	policy, action, reward, game, agent, regret
3	theorem, bound, cluster, lemma, complexity, proof

Figure 6: Short selection of LDA topic results.

Figure 6 shows a selection of the topics generated by LDA. We can see that LDA learns several meaningful topics: topic 1 is likely to be of high proportion in documents about topic modeling, and similarly topic 2 in documents about reinforcement learning, and topic 3 in documents about algorithmic complexity.

8 Evaluating Topic Models

During the analysis of our experimental results, we used human judgment to evaluate the quality of the abstract topics generated. In addition, we judged the quality of the topic by how amenable it is to human interpretation, which may not be suitable for the task of some topic models. This evaluation metric is subjective and not reproducible, and leaves much to be desired.

The evaluation of topic models is also an area of research that is relatively unexplored. The unsupervised nature of topic modeling makes model selection difficult: there are no gold labels for topics, and document topic proportions. The most common evaluation metric is the probability of held-out documents given a trained model [16], and we refer you to Wallach et al. for a full treatment.

9 Conclusion

In this paper, we have discussed the importance of topic modeling, and the opportunities it presents in the domain of scientific documents. LDA is widely regarded as the simplest topic model, which operates on the intuition that a document has few topics, and a topic has few words. LDA has inspired other topic models, such as CTM and DTM, this in no small part due to its simplicity. LDA makes several statistical assumptions, and in the context of scientific documents some of these assumptions are not appropriate. CTM and DTM were introduced in response to these assumptions.

To demonstrate what topic models are capable of, we trained LDA and DTM on the NIPS dataset, and provided the Jupyter notebooks for reference. We also discussed the difficulties faced in evaluating the models, that arise from the unsupervised nature of the problem.

10 Future Work

There are many topic models that have not covered in this survey. For example, the topic models we have covered train only on document text. Yet, scientific documents often come with useful metadata. The author-topic model exploits this [14].

In addition, I have only trained the topic model, and have made no attempt on intrinsic and extrinsic evaluation. Producing benchmarks for topic models on the same corpus may provide value to the research community.

The topic models we have surveyed are directed graphical models. Despite the recent discoveries in deep learning, LDA, which is over a decade old, has not yet been replaced as the mainstay for topic modeling. Deep generative models in particular hold much promise, and is seeing some recent development [9, 11]. In future we hope to explore this area.

[1] https://www.kaggle.com/benhamner/nips-papers

[2] https://github.com/jethrokuan/data-science-notebooks/tree/master/04%20-%20Topic%20Modeling%20(UROP)/nips-papers

References

[1] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. *CoRR*, 2012. URL http://arxiv.org/abs/1205.2662v1.

[2] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[3] David M Blei and John D Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35, 2007.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[5] David M Blei, Thomas L Griffiths, and Michael I Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):7, 2010.

[6] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of*

the American society for information science, 41 (6):391, 1990.

[7] Prem Gopalan, Francisco J Ruiz, Rajesh Ranganath, and David Blei. Bayesian nonparametric poisson factorization for recommendation systems. In *Artificial Intelligence and Statistics*, pages 275–283, 2014.

[8] Tom Griffiths. Gibbs sampling in the generative model of latent dirichlet allocation. *Stanford University*, 2002.

[9] Geoffrey E Hinton and Ruslan R Salakhutdinov. Replicated softmax: an undirected topic model. In *Advances in neural information processing systems*, pages 1607–1614, 2009.

[10] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

[11] Hugo Larochelle and Stanislas Lauly. A neural autoregressive topic model. In *Advances in Neural Information Processing Systems*, pages 2708–2716, 2012.

[12] Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1536–1545. Association for Computational Linguistics, 2011.

[13] Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499. ACM, 2007.

[14] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

[15] Yee Whye Teh. Dirichlet process. In *Encyclopedia of machine learning*, pages 280–287. Springer, 2011.

[16] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112. ACM, 2009.

[17] Chong Wang, David Blei, and David Heckerman. Continuous time dynamic topic models. *CoRR*, 2012. URL `http://arxiv.org/abs/1206.3298v2`.

11 Appendix

	word0	word1	word2	word3	word4	word5
topic0	online	batch	rnn	learner	eigenvalue	stability
topic1	effect	adaptation	gain	control	line	response
topic2	word	layer	sequence	recognition	image	context
topic3	loss	estimator	regression	estimation	prediction	noise
topic4	image	map	object	detection	attention	motion
topic5	layer	latent	code	user	architecture	unit
topic6	graph	tensor	vertex	edge	decomposition	ranking
topic7	group	optimization	constraint	iteration	marginals	solver
topic8	cvpr	influence	rating	relu	effect	prediction
topic9	mixture	covariance	likelihood	component	density	estimation
topic10	oracle	preprint	memory	cnn	phase	capacity
topic11	policy	action	reward	game	agent	regret
topic12	inference	arxiv	source	markov	chain	factorization
topic13	gradient	convergence	iteration	policy	update	equation
topic14	image	filter	patch	video	pixel	scene
topic15	arxiv	event	dropout	speech	sequence	recognition
topic16	classification	datasets	classifier	kernel	query	tree
topic17	topic	rule	round	program	word	atom
topic18	theorem	bound	cluster	lemma	complexity	proof
topic19	image	distance	subspace	dataset	component	face
topic20	bound	log	regularization	variance	hyperparameters	complexity
topic21	brain	cell	population	correlation	activity	response
topic22	gene	design	power	cnns	circuit	calibration
topic23	neuron	spike	response	activity	stimulus	population
topic24	edge	segmentation	image	target	label	field
topic25	document	price	market	consensus	bregman	day
topic26	kernel	projection	recovery	operator	embeddings	column
topic27	node	message	language	communication	runtime	parent
topic28	optimization	gradient	norm	loss	convex	descent
topic29	arm	trajectory	feedback	control	song	movement

Figure 7: Top 6 words of the 30 topic distributions from the LDA model trained on the NIPS paper dataset.