Transformation Process Priors

Nicholas Andrews **Jason Eisner**
Department of Computer Science, Johns Hopkins University
3400 N. Charles St., Baltimore, MD 21218 USA
{noa,eisner}@jhu.edu

1 Introduction

Key idea. As a prior over discrete distributions over $\mathcal{X}$, it is common to use a Dirichlet or Dirichlet process [9]. However, because of the neutrality property, these priors cannot capture correlations among the probabilities of "similar" events. We propose obtaining the discrete distribution from a random walk model or *transformation model* [7], in which each observed event has evolved via a latent sequence of transformations. The transformation model is specified by a collection δ of conditional distributions (the transformations), so placing any prior Δ over δ yields the desired prior over discrete distributions over $\mathcal{X}$. Two events $x_1, x_2 \in \mathcal{X}$ have correlated probabilities in this *transformation process prior* if they tend to emerge from similar transformation sequences.

We are exploring transformation models in which the conditional distributions δ have infinite support and the prior Δ is a nonparametric distribution, such as a hierarchical Dirichlet process. This allows learning idiosyncratic transformations when they explain the data. Although for density estimation the latent structure is just a means of relating different events, there are many applications where the inferred sequences of transformations have a meaningful interpretation (examples below).

Relation to prior work. Similar to Dirichlet diffusion trees [14] (or more generally Pitman-Yor diffusion trees [12]), transformation processes model correlations via latent structure. (A certain nonparametric transformation process prior on $\mathcal{X} = \mathbb{R}$ gives a discrete analogue of Dirichlet diffusion trees.) Methods based on the Gaussian process, such as the discrete infinite logistic normal distribution, model correlations explicitly via the parameters of a covariance function, but do not reconstruct any latent structure [15]. The transformation process is especially appropriate as a prior for discrete distributions where latent structure plausibly exists in the form of derivational history. For instance, a suitable transformation process can be used to infer phylogenetic trees.

2 Transformation Models and their Priors

Prior. In a transformation model, each observation x is derived from a latent trajectory $(\diamond, x_1, x_2, ..., x_t, \square)$ through $\mathcal{X}$ space, distributed according to a stationary Markov process. We set $x_0 = \diamond$ and draw each x_i $(i > 0)$ from the transition distribution $\delta(\cdot \mid x_{t-1})$. Upon drawing the distinguished "stop" event $\square$, we set x to equal the immediately preceding draw, x_t. Likely transformations are given by the transition probabilities $\delta(x|x')$, where $x \in \mathcal{X} \cup \{\square\}$ and $x' \in \mathcal{X} \cup \{\diamond\}$. The transition graph with edges $x' \to x$ for $\delta(x|x') \neq 0$ may be a tree, a DAG, or a cyclic graph.

The expected number of visits to x on a random walk, denoted $a(x)$, emerges from the linear system

$$a(x) = \mathbf{1}(x = \diamond) + \sum_{x'} a(x')\delta(x|x') \tag{1}$$

and then $\mathrm{P}(x) = a(x)\delta(\square|x)$, which describes how often a random walk reaches x and stops. Thus, each $\mathrm{P}(x)$ is a linear combination of the $\mathrm{P}(x')$ values for x' that satisfy $\delta(x|x') > 0$, where the coefficients are determined by the conditional probabilities $\delta(x|x')$, $\delta(\square \mid x)$, and $\delta(\square \mid x')$. This is similar to the way that in Gaussian process density estimation [1], with covariance matrix Σ, each $\log \mathrm{P}(x)$ is a linear combination of the $\log \mathrm{P}(x')$ values for x' that satisfy $(\Sigma^{-1})_{xx'} \neq 0$.[1]

[1] Although Σ^{-1} is symmetric, unlike our conditional probabilities, it is possible in the case of finite $\mathcal{X}$ to convert the Gaussian process model of $\log \mathrm{P}(x)$ to a Bayesian network in which each $\log \mathrm{P}(x)$ is a linear combination of the $\log \mathrm{P}(x')$ values for the parents x' of x.

We place stick-breaking priors [11] over the transition distributions, and the transformation probabilities are therefore given by

$$\delta(x'|x) = \sum_{j=1}^{\infty} p_j \mathbf{1}(y_j = x') \tag{2}$$

where y_j are drawn IID from a (usually sparse) base distribution $\delta_0(x'|x)$, and $p_j \sim \text{GEM}(\alpha_x)$ [8]. The base distributions may themselves be drawn from a hierarchical prior or share parameters.

Posterior inference. Given a dataset of observations $x \sim P$, posterior inference consists of imputing the latent paths $\boldsymbol{x}$ and the collection δ of conditional distributions. (We may also optimize hyperparameters.) Inference via MCMC sampling is possible using a slice sampler that works directly with stick-breaking representations rather than collapsing them out [16]. To sample latent paths from the posterior, auxiliary variables are also sampled at each vertex. Conditioned on these auxiliary variables, the sampler state consists of a finite portion of the graph, from which paths (including cyclic paths and paths through unobserved events) may be sampled efficiently and without bias. In the cyclic case, exact computation of the flow $a(x)$ into each vertex requires solving a finite system of equations; alternatively, the flow may be approximated using a relaxation algorithm [7].

3 Examples of Transformation Processes

String variation. Strings, such as genetic sequences [17], bibliographic entries [10], and proper names, may undergo mutation when they are copied. Given a collection of strings, we hope to infer their evolutionary history (and thereby cluster strings with a common ancestor), by fitting a distribution over the infinite set of all strings, $\mathcal{X} = \Sigma^*$. This involves learning a language model $\delta(x \mid \diamond)$ of the archetypal strings from which other string are derived, as well a stochatic edit model $\delta(x \mid x')$ of string mutation. These stochastic edit models may be constructed using domain knowledge, such as an atomic "acronym" edit that abbreviates "Neural Information Processing Systems" to "NIPS" [10]. Names exhibit even more variation: Михаил Сергеевич Горбачёв, Mikhail Sergeyevich Gorbachev, Pres. Gorbachev, Mikhail Sergeyevich, M. S. Gorbachev, Gorbachev M., Michael Gorbachev, Mike, and many misspellings.

Nonparametric priors over δ, with low concentration parameters, are crucial here to fit the fact that only a few of the plausible edits have actually gained purchase. Mikhail Gorbachev's first name is much more often replaced by "Pres." than by "Mike," even though the base distributions may suggest that "Михаил $\rightarrow$ Mikhail $\rightarrow$ Michael $\rightarrow$ Mike $\rightarrow \square$" is an *a priori* plausible way to transform his first name. In turn, a nonparametric model of the family of base distributions is what lets us learn that "Mikhail $\rightarrow$ Michael" is a common Anglicization (across all Russian names) and more generally that "k $\rightarrow$ c" is common before "h".[2] Sharing statistical strength among the base distributions allows the sampler to propose plausible latent forms in the paths explaining the observed strings.

Syntactic transformations. Each word of a natural language sentence is analogous to a prefix, postfix, or infix operator that takes several arguments. A word type may be overloaded with several calling conventions, known as subcategorization frames. Linguistic tradition suggests that its frames are derived from one another by transformation [4, 3, etc]. Estimating the distribution of frames for each word is useful for syntactic parsing. While we have attempted to do so by MAP estimation of a parametric transformation process with a log-linear parameterization of δ [7], a nonparametric version would allow us to take a proper Bayesian approach to the infinite set of parameters.

Word sense disambiguation. WordNet [13] is a directed acyclic graph whose leaves are words and whose internal nodes are meanings. A path through the graph successively specializes a meaning until it arrives at a word: ENTITY $\rightarrow$ PHYSICAL ENTITY $\rightarrow \cdots \rightarrow$ LEADER $\rightarrow$ POLITICIAN $\rightarrow$ STATESMAN $\rightarrow$ Gorbachev $\rightarrow \square$. A given word may be reachable by multiple paths because it has multiple senses or because an internal node inherits from multiple parents. Suppose we wish to use WordNet to help estimate a joint distribution $P(\text{context}, \text{word})$. We may use a transformation process whose structure is drawn from the WordNet graph [5, 2]—or more precisely, a separate copy of that graph for each context c. A hierarchical prior allows us to learn which sub-concepts of (e.g.) PHYSICAL ENTITY are common overall, and which are unusually popular in a given context. Raising the probability of any sub-concept also raises the probability of all words that descend from it. As a result, words that are close in WordNet will have correlated probabilities, as desired.

[2] Nonparametric models for edits in arbitrary amounts of context may be defined using graphical Pitman-Yor priors [18]. More simply, one could use any of the several parametric stochastic edit models, such as [6].

References

[1] Ryan Prescott Adams, Iain Murray, and David J. C. MacKay. The Gaussian process density sampler. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21 (NIPS 2008)*. 2009.

[2] Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. A topic model for word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1024–1033.

[3] Joan Bresnan. A realistic transformational grammar. In Morris Halle, Joan Bresnan, and George A. Miller, editors, *Linguistic Theory and Psychological Reality*. MIT Press, Cambridge, MA, 1978.

[4] Noam Chomsky. *Syntactic Structures*. Mouton & Co., The Hague, 1957.

[5] Jia Cui and Jason Eisner. Finite-state Dirichlet allocation: Learned priors on finite-state models. Technical Report 53, Center for Language and Speech Processing, Johns Hopkins University, April 2006. 18 pages.

[6] Markus Dreyer, Jason Smith, and Jason Eisner. Latent-variable modeling of string transductions with finite-state methods. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1080–1089, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

[7] Jason Eisner. *Smoothing a Probabilistic Lexicon via Syntactic Transformations*. PhD thesis, University of Pennsylvania, July 2001. 318 pages.

[8] W. Ewens. Population Genetics Theory - The Past and the Future. In S. Lessard, editor, *Mathematical and Statistical Developments of Evolutionary Theory*, pages 177–227. Kluwer Academic Publishers, 1990.

[9] T. S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[10] Rob Hall, Charles Sutton, and Andrew McCallum. Unsupervised deduplication using cross-field dependencies. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 310–317, New York, NY, USA, 2008. ACM.

[11] H. Ishwaran and James. Gibbs Sampling Methods for Stick Breaking Priors. *Journal of the American Statistical Association*, pages 161–173, March 2001.

[12] David Knowles and Zoubin Ghahramani. Pitman-yor diffusion trees. In *27th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 63–70, 2011.

[13] George A. Miller. Nouns in WordNet: A Lexical Inheritance System. *International Journal of Lexicography*, 3(4):245–264, December 1990.

[14] R. M. Neal. Density modeling and clustering using dirichlet diffusion trees. In J. M. Bernardo, editor, *Bayesian Statistics 7*, pages 619–629. 2003.

[15] J. Paisley, C. Wang, and D. Blei. The discrete infinite logistic normal distribution for mixed-membership modeling. In *Artificial Intelligence and Statistics*, 2011.

[16] Stephen G. Walker. Sampling the dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation*, 36(1):45–54, 2007.

[17] Fabian L. Wauthier, Michael I. Jordan, and Nebojsa Jojic. Nonparametric Combinatorial Sequence Models. In *RECOMB Proceedings*. 2011.

[18] F. Wood and Y. W. Teh. A hierarchical nonparametric Bayesian approach to statistical language model domain adaptation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 12, 2009.