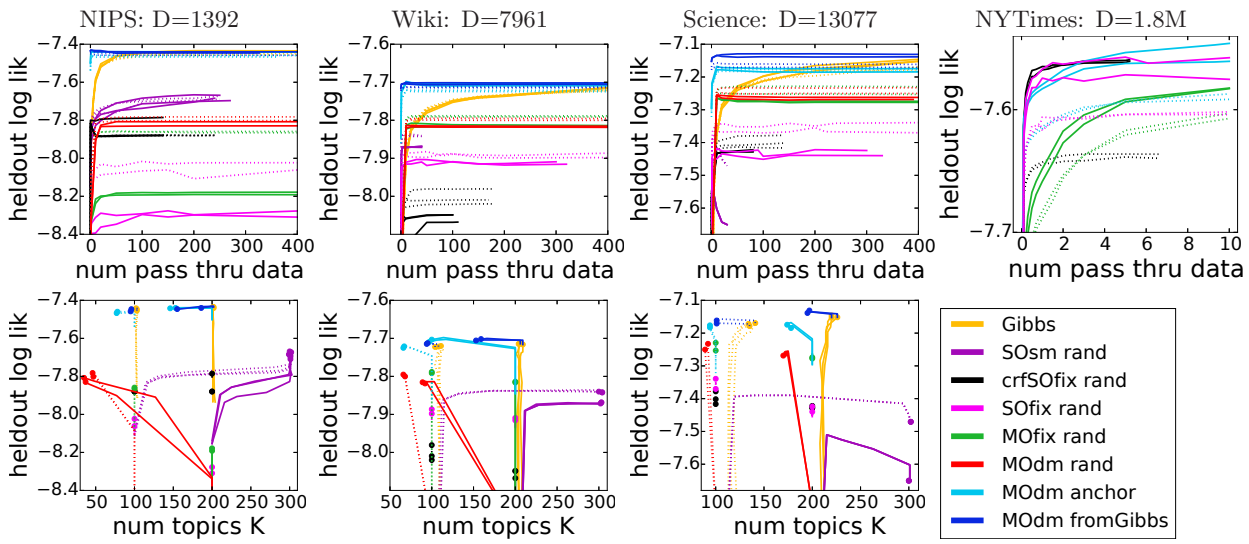

Figure 6: Comparison of inference methods on academic and news article datasets (Sec. 6.2). Line style indicates initial number of topics K: 100 is dots, 200 is solid. *Top row:* Heldout likelihood (larger is better) as more training data is seen. *Bottom row:* Trace plots of heldout likelihood and number of topics. Each solid dot marks the final result of a single run, with the trailing line its trajectory from initialization. Ideal runs move toward the upper left corner.

For this large-scale task, our direct assignment representation is more efficient than the CRF code released by Wang et al. (2011). With $K = 200$ topics, our memoized algorithm with merge and delete moves (MOdm) completes 8 laps through the 1.8 million documents in the amount of time the CRF code completes a single lap. No deletes or merges are accepted from any MOdm run, likely because 1.8M documents require more than a few hundred topics. However, the acceptance rate of sparsity-promoting restarts is 75%. With a more efficient, parallelized implementation, we believe our variational approach will enable reliable large-scale learning of topic models with larger K.

6.3 Image patch modeling.

Finally, we study 8×8 patches from grayscale natural images as in Zoran and Weiss (2012). We train on 3.5 million patches from 400 images, comparing HDP admixtures to Dirichlet process (DP) mixtures using a zero-mean Gaussian likelihood. The HDP model captures within-image patch similarity via image-specific mixture component frequencies. Both methods are evaluated on 50 heldout images scored via Eq. (22).

Fig. 7 shows merges and deletes removing junk topics while improving predictions, justifying the generality of these moves. Further, the HDP earns better prediction scores than the DP mixture. We illustrate this success by plotting sample patches from the top 4 topics (ranked by topic weight π) for several heldout images. The HDP adapts topic weights to each image, favoring smooth patches for some images (d) and textured patches for others (e-f). The less-flexible DP must use the same weights for all images (c).

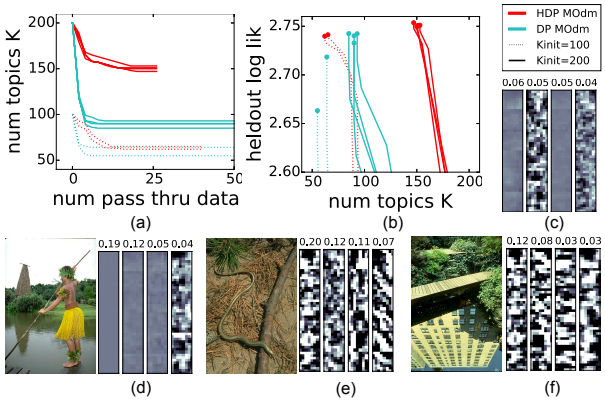

Figure 7: Comparison of DP mixtures and HDP admixtures on 3.5M image patches (Sec. 6.3). (a-b) Trace plots of number of topics and heldout likelihood, as in Fig 6. (c) Patches from the top 4 estimated DP clusters. Each column shows 6 stacked 8×8 patches sampled from one cluster. (d-f) Patches from 4 top-ranked HDP clusters for select test images from BSDS500 (Arbelaez et al., 2011).

7 CONCLUSION

We have developed a scalable variational algorithm for learning compact, interpretable HDP models from millions of examples. Our novel objective applies to any exponential family likelihood and could prove useful for sequential or relational models based on the HDP.

Acknowledgments This research supported in part by NSF CAREER Award No. IIS-1349774. M. Hughes supported in part by an NSF Graduate Research Fellowship under Grant No. DGE0228243.