NEURAL VARIATIONAL RANDOM FIELD LEARNING

Volodymyr Kuleshov and Stefano Ermon
Department of Computer Science,
Stanford University
{kuleshov,ermon}@cs.stanford.edu

ABSTRACT

We propose variational bounds on the log-likelihood of an undirected probabilistic graphical model p that are parametrized by flexible approximating distributions q. These bounds are tight when $q = p$, are convex in the parameters of q for interesting classes of q, and may be further parametrized by an arbitrarily complex neural network. When optimized jointly over q and p, our bounds enable us to accurately track the partition function during learning.

1 INTRODUCTION

Probabilistic graphical modeling is one of the most fundamental techniques in artificial intelligence and representation learning. However, learning rich representational models involves major computational challenges. One of the main approximate inference techniques that deals with these challenges is variational inference. This approach seeks to find a tractable approximating distribution q to a complex model p. Ideal q's should be expressive, easy to optimize over, and admit tractable inference procedures. Recent work has shown that neural network-based models possess many of these qualities (Kingma & Welling, 2013; Rezende et al., 2014; Burda et al., 2015).

Here, we seek to extend this line of work via new variational inference techniques aimed at undirected probabilistic graphical models. We propose variational upper bounds on the log-partition function parametrized by an approximating distribution q. These bounds are tight when $q = p$ and are convex in the parameters of q for interesting classes of q; for increased expressivity, q can also be parametrized by an arbitrarily complex neural network. Most interestingly, we also give a new concave lower bound on the log-likelihood function; when optimized jointly over q and p, it enables us to accurately track the partition function during learning. Our techniques may serve as subroutines in several classes of algorithms for learning representations.

2 SETUP AND BACKGROUND

Undirected graphical models. For expository purposes, we will focus our attention on Markov random fields (MRFs), which are probabilistic models of the form $p_\theta(x) = \tilde{p}_\theta(x)/Z(\theta)$, where $\tilde{p}_\theta(x) = \exp(\theta \cdot x)$ is an unnormalized probability and $Z(\theta) = \mathbb{E}_x \tilde{p}_\theta(x)$ is the partition function. Our approach also naturally extends to conditional random field (CRF) models.

Importance sampling. The partition function of an MRF in an intractable integral over $\tilde{p}(x)$. We may, however, rewrite it as $I := \int_x \tilde{p}_\theta(x)dx = \int_x \frac{\tilde{p}_\theta(x)}{q(x)} q(x)dx = \int_x w(x)q(x)dx$, where q is a proposal distribution. Integral I can in turn be approximated by a Monte-Carlo estimate $\hat{I} := \frac{1}{n}\sum_{i=1}^n w(x_i)$, where $x_i \sim q$. The variance of this *importance sampling* estimate $\hat{I}$ has a closed-form expression: $\mathbb{E}_{q(x)}[w(x)^2] - I^2$. By Jensen's inequality, it equals 0 when $p = q$.

3 VARIATIONAL BOUNDS

The first term in the variance of the importance sampler is a natural bound on the partition function:

$$\mathbb{E}_{q(x)}\left[\frac{\tilde{p}(x)^2}{q(x)^2}\right] \geq Z(\theta)^2 \tag{1}$$

Again, this bound is tight when $q = p$. It implies a natural algorithm for computing $Z(\theta)$: minimize (1) over q in some family $\mathcal{Q}$. This can be interpreted as both minimizing the variance of $\hat{I}$, and as minimizing a tight upper bound on the partition function. A key decision concerns the choice of approximating family $\mathcal{Q}$: it needs to be expressive, easy to optimize over, and admit tractable inference procedures. Here, we propose two such families.

3.1 Choice of approximating family

Non-parametric variational inference. First, as suggested by Gershman et al. (2012), we may take q to be a uniform mixture of exponential families $\sum_{k=1}^{K} \frac{1}{K} q_k(x; \phi_k)$. In practice, the q_k may be either Gaussians or Bernoulli, depending on whether x is discrete or continuous. This choice of q lets us potentially model arbitrarily complex p given enough components; we will also see below that such q are easy to optimize.

Auxiliary-variable neural networks. Alternatively, we may further parametrize q by an arbitrarily complex neural network. This approach is complicated by the fact that unlike earlier methods that parametrized conditional distributions $q(z|x)$ over hidden variables z, our setting does not admit a natural input/output to a neural network.

We address this difficulty via extra *auxiliary variables* z in the approximating model q. First, we define $\tilde{p}(x, z) = \tilde{p}(x)$ for all x, z, and let $q(x, z) = q(x|z)q(z)$, where $q(z)$ is some simple prior (e.g. normal or uniform), and $q_\phi(x|z)$ is an exponential family distribution whose natural parameters are parametrized by a neural net, e.g. $q(x|z) = N(\mu_\phi(z), \sigma_\phi(z)I)$ for continuous x. We may perform importance sampling as follows: $\int \tilde{p}(x)dx = \int \tilde{p}(x, z)dxdz \approx \frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{p}(x,z)}{q(x|z)q(z)}$, where $x_i, z_i \sim q(x, z)$. Note that this reduces to the previous case with an appropriate choice of q.

3.2 Convexity properties

A key property of our bound is that it is jointly log-convex in θ and $\{\phi_k\}_{k=1}^{K}$, which is the set of natural parameters in a mixture of exponential families $q = \sum_{k=1}^{K} \frac{1}{K} q_k(x; \phi_k)$. Note that this immediately implies that our non-parametric inference approach leads to a convex optimization problem. If we choose to further parametrize ϕ by a neural net, the resulting non-convexity will originate solely from the neural network, and not from our choice of loss function.

To establish log-convexity, it suffices to look at $\tilde{p}_\theta(x)^2/q(x)$ for one x, since the sum of log-convex functions is log-convex. Note that $\log \frac{\tilde{p}_\theta(x)^2}{q(x)} = 2\theta^T x - \log \sum_k \pi_k q_{\phi_k}(x)$. One can easily check that a non-negative concave function is also log-concave. If the q_k are in the exponential family, it follows that $\sum_k \pi_k q_{\phi_k}(x)$ is log-concave, and hence the above expression is convex.

3.3 Optimization

Assuming a non-parametric variational approximation $\sum_{k=1}^{K} \frac{1}{K} q_k(x; \phi_k)$, it is easy to show that the gradient w.r.t. ϕ_k is $\nabla_{\phi_k} \mathbb{E}_q \frac{\tilde{p}(x)^2}{q(x)^2} = \mathbb{E}_{q_k}\left[\frac{\tilde{p}(x)^2}{q(x)^2} d_k(x)\right]$, where $d_k(x)$ is the difference between x and its expectation under q_k. Thus, we may optimize the bound (1) using stochastic gradient descent by taking samples from q_k. Note also that if our goal is to compute the partition function, we may collect all intermediary samples for computing the gradient and use them as regular importance samples. This may be interpreted as a form of adaptive sampling.

4 Variational random field learning

Next, we turn our attention to the problem of learning the parameters of an MRF. Given data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{n}$, our training objective is the log-likelihood $\log p(\mathcal{D}|\theta) := \sum_{i=1}^{n} \log p_\theta(x^{(i)})$. We can use our earlier bound to upper bound the log-partition function by $\log\left(\mathbb{E}_{x\sim q} \frac{\tilde{p}_\theta(x)^2}{q(x)^2}\right)$. By our previous discussion, this expression is convex; however, unlike Equation 1, we may no longer approximate the expectation with Monte-Carlo estimates due to the non-linearity introduced by the log.

To deal with this issue we further linearize the log using the identity $\log(x) \leq ax - \log(a) - 1$, which is tight for $a = 1/x$. Together with our bound on the log-partition function, this yields

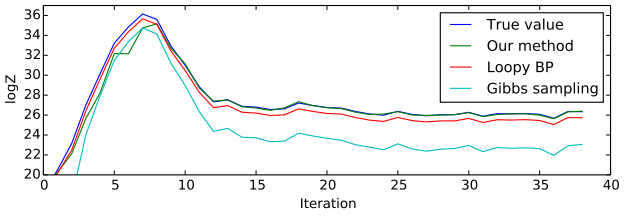
$$\log p(\mathcal{D}|\theta) \geq \max_{\theta,q} \frac{1}{n} \sum_{i=1}^{n} \theta^T x^{(i)} - \frac{1}{2} \left(a \mathbb{E}_{x\sim q} \frac{\tilde{p}_\theta(x)^2}{q(x)^2} - \log(a) - 1 \right). \tag{2}$$

This expression is convex in each of (θ, ϕ) and a, but is not jointly convex. However, it is straightforward to show that equation (2) and its unlinearized version have a unique point satisfying first-order stationarity conditions. This may be done by writing out the KKT conditions of both problems and using the fact that $a^* = (\mathbb{E}_{x\sim q} \frac{\tilde{p}_\theta(x)^2}{q(x)^2})^{-1}$ at the optimum. See Gopal & Yang (2013) for more details.

Equaiton 2 may be optimized jointly over θ, ϕ, with periodical updates for a. By training p and q jointly, the two distributions may help each other. In particular, we may start learning at an easy θ (where p is not too peaked) and use slowly q to track p, thus controlling the variance in the gradient.

5 Experiments

We evaluated empirically our learning strategy on a 5×5 Ising MRF with coupling factor J and unaries chosen randomly in $\{10^{-2}, -10^{-2}\}$. We set $J = -0.6$, sampled 1000 examples from the model, and fit another MRF to this data. We followed a non-parametric inference approach with a mixture of $K = 8$ Bernoullis. We optimized (2) using SGD with fixed stepsizes chosen by cross-validation; we alternated between ten steps over the ϕ_k and one step over θ, a. We drew 100 Monte Carlo samples per q_k. Our method converged in about 25 steps over θ. At each iteration we computed $\log Z$ via importance sampling.



Our Figure shows the evolution of $\log Z$ during learning. It also plots $\log Z$ computed by brute force enumeration, loopy BP, and Gibbs sampling (using the same number of samples). Our method accurately tracks the partition function after about 10 iterations. In particular, our method fares better than the others when $J \approx -0.6$, which is when the Ising model is entering its phase transition.

6 Discussion and related work

Our work is inspired by variational autoencoders (Kingma & Welling, 2013), which involve tightening variational lower bounds using neural networks. Our work provides analogous upper bounds that also hold for undirected and discrete variable models; interestingly, they may be interpreted as an inclusive α-divergence (Minka, 2005). Alternative rich proposal distribution families include normalizing flows (Rezende & Mohamed, 2015) and Variational Gaussian Processes (Tran et al., 2015). Finally, the unpublished manuscript of Ryu & Boyd (2014) proposed similar adaptive importance sampling methods, but did not discuss tightness or applications to MRF learning.

Limitations. Our technique's main shortcoming is high variance in the Monte Carlo gradient estimates if q is initially far from p, and the latter is "peaked"; in such cases, we may never sample from the modes of p. Thus, our techniques are more suitable for learning, where p is initially "easy", and q tracks p during the learning procedure.

Future work. Our next steps are to validate the method in more complex models, such as restricted Boltzmann machines and CRFs, to use more complex neural-network reparametrizations, and to compare with additional methods such as annealed importance sampling.

Our methods may also augment existing inference methods, for example by bounding the log-partition function within classical variational lower bounds. Our bound may also serve as a loss for training variational autoencoders: since it corresponds to an inclusive divergence, it may help avoid overfitting distributions to specific modes, a problem that has recently received research attention (Burda et al., 2015).

REFERENCES

Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *CoRR*, abs/1509.00519, 2015. URL http://arxiv.org/abs/1509.00519.

Samuel Gershman, Matthew D. Hoffman, and David M. Blei. Nonparametric variational inference. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012. URL http://icml.cc/discuss/2012/360.html.

Siddharth Gopal and Yiming Yang. Distributed training of large-scale logistic models. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp. 289–297, 2013. URL http://jmlr.org/proceedings/papers/v28/gopal13.html.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL http://arxiv.org/abs/1312.6114.

Thomas Minka. Divergence measures and message passing. Technical report, 2005. URL https://www.seas.harvard.edu/courses/cs281/papers/minka-divergence.pdf.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pp. 1530–1538, 2015. URL http://jmlr.org/proceedings/papers/v37/rezende15.html.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pp. 1278–1286, 2014. URL http://jmlr.org/proceedings/papers/v32/rezende14.html.

Ernest K. Ryu and Stephen P. Boyd. Adaptive importance sampling via stochastic convex programming. Unpublished manuscript, November 2014.

Dustin Tran, Rajesh Ranganath, and David M. Blei. Variational gaussian process. *CoRR*, abs/1511.06499, 2015. URL http://arxiv.org/abs/1511.06499.