Adversarial Message Passing For Graphical Models

Theofanis Karaletsos
Geometric Intelligence

Abstract

A currently popular technique for learning generative models is generative adversarial networks (GANs). They form a basis to learning generative models by learning to discriminate true samples versus fake ones to guide a model towards good solutions that can fool a strong discriminator into assigning high probability of being true to model samples. It has been shown that GANs minimize a well-defined f-divergence, the Jensen-Shannon divergence, between the model distribution and the data distribution. However, current best practices have a number of shortcomings. Typically, GANs are considered to be models and are not understood in the context of inference. In addition, current techniques rely on global discrimination of joint distributions to perform learning, which is ineffective. We propose to alleviate this limitation by showing how to relate adversarial learning to distributed approximate Bayesian inference on factor graphs. We propose local learning rules based on message passing which minimize a global variational criterion based on adversaries used to score ratios of distributions instead of explicit likelihood evaluations. This yields an inference and learning framework that faciliates treating model specification and inference seperately by combining ideas from message passing with adversarial inference and can be used on arbitrary computational structures within the family of Directed Acyclic Graphs and models, including intractable likelihoods, non-differentiable models and generally cumbersome models. We thus present adversarial learning under the viewpoint of approximate inference and modeling. We combine adversarial learning with nonparametric variational families to yield a learning framework which performs implicit Bayesian inference on graph structures by sampling particles, without the need to evaluate densities. These approaches hold promise to be useful in the toolbox of probabilistic modelers and have the potential to enrich the gamut of flexible probabilistic programming applications beyond current practice.

1 Introduction & Related Work

We discuss adversarial learning from the perspective of distributed Bayesian inference on generative models. We generalize adversarial learning to arbitrary structured models by introducing a local message passing algorithm based on adversaries and show that is is performing a clean approximation to a posterior defined by an explicit model. We thus present novel work that explains and clarifies the separation of modeling and inference in the context of adversarial learning and opens the door to building flexible probabilistic programs using adversarial inference.

In recent work it has been shown that neural networks can be used as samplers for divergence minimization in a general class of divergences [1]. Furthermore, it was clarified in concurrent work very much in the same spirit with our paper such as [2] and [3] that Generative Adversarial Networks can be seen as a form of inference on ratios of partition functions, with early links towards training generative models. First steps towards GANs on structured models were taken in recent papers like the SeqGAN [4], Professor Forcing [5] and [6]. We highlight that a side-result of [7] is a derivation

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

of a KL-divergence loss for standard GANs and the introduction of instance noise, both of which are related to results we discuss in our Appendix. Finally, inference for a narrow class of specific fixed instances of models was introduced in similar fashion in [8], [9] and [10]) using global adversaries, but not generalized to more flexible models.

2 Generative Adversarial Networks

Basic GANs have been postulated to follow a value function playing an adversarial game between a discriminator D with parameters ξ and a generator G with parameters θ.

$$\begin{aligned}\min_{\theta}\max_{\xi} V(\xi,\theta) &= \mathbb{E}_{x\sim p^*(x)}\log D(x;\xi) + \mathbb{E}_{x\sim Q(x)}\log(1-D(x;\xi)) \\ &= \mathbb{E}_{x\sim p^*(x)}\log D(x;\xi) + \mathbb{E}_{z\sim P(z)}\log(1-D(G(x;\theta);\xi))\end{aligned} \tag{1}$$

For $m(x) = \frac{1}{2}p(x) + \frac{1}{2}q(x)$ an analogy can be shown between the value function and the following probabilistic formulation.

$$\begin{aligned}\textbf{JSD}(q(x)||p(x)) &= \frac{1}{2}\int_{x^*} q(x^*)\log\frac{q(x^*)}{m(x)}dx + \frac{1}{2}\int_{x^*} p(x^*)\log\frac{p(x^*)}{m(x)}dx \\ &= \frac{1}{2}\int_{x^*} q(x^*)\log\frac{q(x^*)}{m(x)}dx + \frac{1}{2}\int_{z} p(z)\log\frac{p(x|z)}{m(x)}dz\end{aligned} \tag{2}$$

3 Approximate Inference in Graphical Models through Adversarial Learning

We show, that instead of one large GAN discriminating between the joint distribution of all variables in graphical models (as done in [8], [9] and [10]), we can perform distributed adversarial inference by discriminating locally for each variable whether it is a valid sample or not. We can maximize these local discriminators to yield a globally convergent distributed learning procedure, **adversarial message passing**.

We are given a joint distribution over I-many variables $p(\mathbf{X}) = p(x_0, .., x_I)$ with a graph structure $\mathcal{G}$ and a factorization given by the computational graph $p(\mathbf{X}) = \prod_i p(x_i|\text{pa}(x_i))$, where $\text{pa}(x_i)$ denote the parents of variable x_i in $\mathcal{G}$. We can derive an inverse factorization $q(\mathbf{X}) = \prod_i q(x_i|\tilde{\text{pa}}(x_i))$ which preserves the variable dependence structure. In the inverse factorization, we consider $\tilde{\text{pa}}(x_i)$ to be the part of the Markov blanket for the variable x_i needed in order to *d-seperate* it given observations. These factorizations have been explained at length in the context of stochastic inversion [11] and form a structured inverse factorization as used in variational inference [12], while also being widely used in the message passing literature [13],[14].

$$P(\mathbf{X}) = P(x_1, x_2, ..., x_D) = \prod_{i=1}^{D} P(x_i|\text{pa}(x_i)). \tag{3}$$

We use factorizations of dependencies as the basis to derive schemes for Bayesian Learning and inference which take advantage of adversarial learning.

3.1 Adversarial Message Passing For JS-Divergence Minimization

In this section, we match the local Jensen-Shannon divergence (**JSD**) of variables to perform approximate inference locally.

We use the intuition that we wish to match the local statistics of approximations to the posterior by minimizing a divergence Div at each factor indexed by i, $\text{Div}\Big(q(x_i^*, \tilde{\text{pa}}(x_i))||p(x_i, \text{pa}(x_i))\Big)$. This is a typical assumption in divergence based message passing [15].

Given a definition of $m(x_i, \text{pa}(x_i)) = \Big[0.5q(x_i, \tilde{\text{pa}}(x_i)) + 0.5p(x_i, \text{pa}(x_i))\Big]$, we can express local minimization of the **JSD** as a sum of divergences, compactly written as follows:

$$\begin{aligned}\text{Div}_{loc}\Big(q(\mathbf{X})||p(\mathbf{X})\Big) =& \frac{1}{2}\int_{x_0} p^*(x_0)...\int_{x_I} q(x_I|\tilde{\text{pa}}(x_I))\log\frac{\prod_{i=1}^{I} q(x_{i-1}, \tilde{\text{pa}}(x_{i-1}))}{\prod_{i=1}^{I} m(x_{i-1}, \text{pa}(x_{i-1}))}dx_{0...I} \\ &+ \frac{1}{2}\int_{x_I} p(x_I)...\int_{x_0} p(x_0|\text{pa}(x_0))\log\frac{\prod_{i=0}^{I-1} p(x_i, \text{pa}(x_i))}{\prod_{i=0}^{I-1} m(x_i, \text{pa}(x_i))}dx_{0...I}\end{aligned} \tag{4}$$

We rephrase the above divergence in terms of a sum of the local adversaries by noting that each factor can be expressed as an expectation over the score of the class the discriminator will assign to the bottom-up and top-down samples.

We can use an optimal discriminator D_i^* as an adversary at each local factor i to express ratios of distributions $D_i^*(x_i, \text{pa}(x_i)) = \frac{p(x_i, \text{pa}(x_i))}{m(x_i, \text{pa}(x_i))}$ and $1 - D_i^*(x_i, \text{pa}(x_i)) = \frac{q(x_i, \tilde{\text{pa}}(x_i))}{m(x_i, \text{pa}(x_i))}$. In order to calibrate these adversaries, we can derive a loss function $\mathcal{L}_{locD}$ and train models to discriminate between inference and model samples generated during training.

Combining these adversaries with Equation 4 yields a reparametrized form of the divergence term:

$$\begin{aligned}\text{Div}_{loc}\Big(q(\mathbf{X})||p(\mathbf{X})\Big) =& \frac{1}{2}\int_{x_0} p^*(x_0)...\int_{x_I} q(x_I|\tilde{\text{pa}}(x_I))\log\Big[\prod_{i=1}^{I}\Big(1 - D_i^*(x_{i-1}, \tilde{\text{pa}}(x_{i-1}))\Big)\Big]dx_{0...I} \\ &+ \frac{1}{2}\int_{x_I} p(x_I)...\int_{x_0} p(x_0|\text{pa}(x_0))\log\Big[\prod_{i=0}^{I-1}\Big(D_i^*(x_i, \text{pa}(x_i))\Big)\Big]dx_{0...I}\end{aligned} \tag{5}$$

This joint term can be approximated efficiently across each local term by performing bottom-up sampling of L particles through inference models and K top down samples from the prior. This procedure yields two Markov chains transitioning from evidence to prior and from prior to evidence in a setting similar to that used for the Bennet acceptance ratio estimator [16] and related newer work [17, 18, 19, 20, 21].

We consider generative models to be parameterized by parameters θ capturing the generative factors and inverse models performing inference over unobserved variables X_u and observed variables X_o to be parametrized by ϕ denoting variational parameters or parameters of inference models. Learned adversaries have parameters ξ. We obtain the following objective function for learning graphical models using the above:

$$\mathcal{L}_{locM}(\theta, \phi|\mathbf{X}) = \text{Div}_{loc}\Big(q(\mathbf{X}|\phi))||p(\mathbf{X}|\theta)\Big) \tag{6}$$

Concurrently, since the variable-wise adversaries $D_i(\cdot|\xi)$ need to be trained to approximate optimality, we can derive a loss function for them as follows:

$$\mathcal{L}_{locD}(\xi|\mathbf{X}) = -\Big[\mathbb{E}_{x_i, \text{pa}(x_i)}\log D_i(x_i, \text{pa}(x_i)) + \mathbb{E}_{x_{i-1}, \tilde{\text{pa}}(x_{i-1})}\log(1 - D_i(x_{i-1}, \tilde{\text{pa}}(x_{i-1})))\Big] \tag{7}$$

Equality to the **JSD** holds when for each factor i we minimze the divergence between the approximation and the true distribution, obtaining $\tilde{\text{pa}}(x_i)) = p(x_i|\text{pa}(x_i))$ This also reveals that the fixed points of $\textbf{Div}_{loc}$ are the fixed points of **JSD**, which correspond to global fixed points to the true distribution. In general, $\textbf{Div}_{loc}$ provides a looser divergence than **JSD**, which intuitively makes sense since it performs a local calculation through message passing and formally can be shown by comparing the denominators in the respective divergence terms.

A divergence which is smaller or equal to **JSD** overestimates the fit to the likelihood compared to **JSD** theoretically, but we obtain the following practical benefits through distribution of our **JSD** calculation:

1. In the adversarial framework, calculating the global **JSD** requires learning and evaluation of a discriminator over the joint distribution. For larger graphical models with multiple potentially high-dimensional variables, this quickly becomes impossible or impractical.
2. As long as the discriminator is far away from the Bayes-Optimal discriminator, the assumption to reparametrize the ratio-term through the discriminator is not fulfilled. Local discriminators have a better chance of obtaining locally strong solutions for smaller tuples of variables than global discriminators of an entire graphical model state.
3. Local discriminators furthermore permit interesting learning settings, like partial observability as occuring in semi-supervised learning, time-series with irregular time-steps, multimodal data-sets with missing modalities and more.

With our framework, we perform local discrimination per factor and achieve a similar computation to that of a global discriminator needed for the global **JSD** to hold, see Algorithm 1.

Algorithm 1 Adversarial Message Passing

procedure ADVMP($X, iter$) ▷ X: a given dataset, iter: # of iterations
  $\phi_0 \sim P(\phi_{\text{init}})$
  $w_0 \sim P(w_{\text{init}})$ ▷ initialize weights of prior and model approximation
  $\epsilon_{\mathbf{0}} \sim p(\epsilon)$ ▷ Initial Noise-vector
  for $t \leq iter$ **do** ▷ Loop over iterations
    for $X_t \in X$ **do** ▷ Sample minibatch X_t
      $\forall i : x_i^l \sim q(x_i|\tilde{\text{pa}}(x_i))$ ▷ Infer parents of each variable with inference model
      $\forall i : x_i^k \sim p(x_i|\text{pa}(x_i))$ ▷ Sample from model (using θ or specified model)
      $\epsilon_{\mathbf{t}} \sim p(\epsilon)$ ▷ Sample an appropriate noise vector
      for i in factors **do** ▷ Cycle through factors and update parameters
        $\xi_{t,i} \leftarrow \xi_{t-1,i} - \frac{\partial \mathcal{L}_{locD}(\theta_{t-1},\phi_{t-1},\xi_{t-1};\epsilon_t,X_t)}{\partial \xi}$
        $\theta_{t,i} \leftarrow \theta_{t-1,i} - \frac{\partial \mathcal{L}_{locM}(\theta_{t-1},\phi_{t-1},\xi_{t-1};\epsilon_t,X_t)}{\partial \theta_{t-1}}$
        $\phi_{t,i} \leftarrow \phi_{t-1,i} - \frac{\partial \mathcal{L}_{locM}(\theta_{t-1},\phi_{t-1},\xi_{t-1};\epsilon_t,X_t)}{\partial \phi}$
  return θ_t, ϕ_t, ξ_t ▷ Parameters for the adversaries ξ, variational approximations ϕ, model θ learned from data X

4 Discussion

Adversarial Message Passing provides a framework to perform likelihood-free learning for explicit graphical models. It furthermore enriches the family of message passing algorithms by a previously intractable divergence class and faciliates the usage of nonparametric variational families for likelihood free learning and inference in graphical models. We note that more general classes of divergences such as f-divergences fall under this framework, since adversaries serve as function approximations to score ratios of distributions and can be composed locally to infer larger models. In the appendix we exhibit similar treatments for KL-divergence as an example. Interestingly, this allows us to cleanly derive combinations of adversarial loss functions with explicit parametric losses mapping to likelihood maximization, as empirically used by various previous papers without formal justification. It is also easy to mix different divergences locally depending on suitability. Furthermore, a generalization of the work presented here can use MMD [22] to perform local approximations in computational graphs. Finally, we suggest that the introduced message passing scheme can be generalized to undirected graphical models in future work.

Acknowledgements

We thank Eli Bingham, John Chodera, Noah Goodman and Zoubin Ghahramani for helpful and inspiring discussions. We furthermore acknowledge Anh Nguyen and Jason Yosinski for demonstrating the benefits of adversarial learning.

References

[1] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *arXiv preprint arXiv:1606.00709*, 2016.

[2] Masatoshi Uehara, Issei Sato, Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Generative adversarial nets from a density ratio estimation perspective. *arXiv preprint arXiv:1610.02920*, 2016.

[3] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

[4] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. *arXiv preprint arXiv:1609.05473*, 2016.

[5] A. Lamb, A. Goyal, Y. Zhang, S. Zhang, A. Courville, and Y. Bengio. Professor Forcing: A New Algorithm for Training Recurrent Networks. *ArXiv e-prints*, October 2016.

[6] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. *arXiv preprint arXiv:1606.07536*, 2016.

[7] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, 2016.

[8] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[9] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

[10] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[11] Andreas Stuhlmüller, Jacob Taylor, and Noah Goodman. Learning stochastic inverses. In *Advances in neural information processing systems*, pages 3048–3056, 2013.

[12] Matthew D Hoffman and David M Blei. Structured stochastic variational inference. In *Artificial Intelligence and Statistics*, 2015.

[13] John Winn and Christopher M Bishop. Variational message passing. *Journal of Machine Learning Research*, 6(Apr):661–694, 2005.

[14] Thomas P Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

[15] Tom Minka et al. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005.

[16] Charles H Bennett. Efficient estimation of free energy differences from monte carlo data. *Journal of Computational Physics*, 22(2):245–268, 1976.

[17] Charles J Geyer. *Reweighting monte carlo mixtures*. Citeseer, 1991.

[18] Michael R Shirts and John D Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of chemical physics*, 129(12):124105, 2008.

[19] Qiang Liu, Jian Peng, Alexander T Ihler, and John W Fisher III. Estimating the partition function by discriminance sampling. In *UAI*, pages 514–522, 2015.

[20] David Carlson, Patrick Stinson, Ari Pakman, and Liam Paninski. Partition functions from rao-blackwellized tempered sampling. *arXiv preprint arXiv:1603.01912*, 2016.

[21] Roger B Grosse, Zoubin Ghahramani, and Ryan P Adams. Sandwiching the marginal likelihood using bidirectional monte carlo. *arXiv preprint arXiv:1511.02543*, 2015.

[22] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[23] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *arXiv preprint arXiv:1602.02644*, 2016.

[24] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *arXiv preprint arXiv:1605.09304*, 2016.

[25] Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.

[26] Diane Bouchacourt, M Pawan Kumar, and Sebastian Nowozin. Disco nets: Dissimilarity coefficient networks. *arXiv preprint arXiv:1606.02556*, 2016.

[27] Francisco JR Ruiz, Michalis K Titsias, and David M Blei. The generalized reparameterization gradient. *arXiv preprint arXiv:1610.02287*, 2016.

[28] Christian A Naesseth, Francisco JR Ruiz, Scott W Linderman, and David M Blei. Rejection sampling variational inference. *arXiv preprint arXiv:1610.05683*, 2016.

[29] R. Ranganath, J. Altosaar, D. Tran, and D. M. Blei. Operator Variational Inference. *ArXiv e-prints*, October 2016.

[30] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

5 Appendix

5.1 Learning Deep Generative Models

We exemplify how to use the introduced framework at the example of a deep generative model with two stochastic layers, applied to modeling MNIST digits.

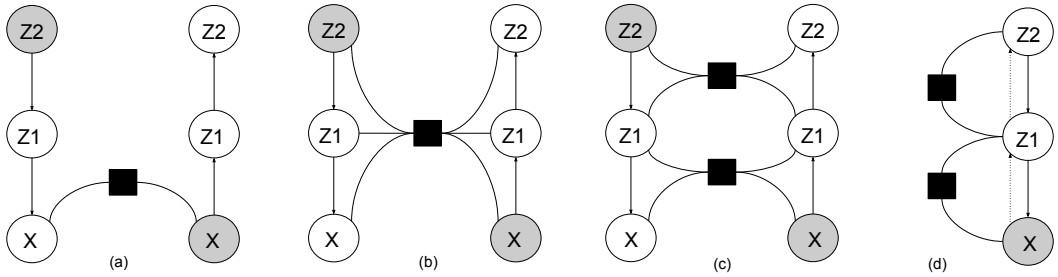

Figure 1: We show the four different learning variants. Black boxes indicate adversaries connected to their input variables. (**a**) A standard adversarial network which only has to generate observable X (**b**) A deep variant of a global bidirectional adversarial network (**c**) A model using adversarial message passing with JSD minimization using local adversaries (**d**) A model using adversarial message passing with KL minimization using local adversaries.

The generative model is defined as follows:

1. $z_2 \sim P(z_2)$

2. $z_1 \sim P(z_1|z_2)$

3. $x \sim P(x|z_1)$

We use two adversaries $D_1(x, z_1)$ and $D_2(z_1, z_2)$ to drive learning. The inverse factorization here is trivial since Markov blankets on chain-graphs form unique tupels of variables. We show the different inferential strategies in Figure 5.1.

We note that compared to the usual application of GANs, we explicitly define the model here. For instance, $P(z_2) = \mathcal{N}(0, 1)$, $P(z_1|z_2) = \mathcal{N}(\mu_{z_2}, \Sigma_{z_2})$, $P(x|z_1) = Ber(\mu_{z_1})$.Interestingly, when we generate from the priors we also sample observation noise from the Bernoulli likelihood. This yields similar results to what is defined as instance noise in [7], since a layer of noise is added to all generated images before they are passed into adversaries.

5.2 Derivations for Variational Inference

For a model $P(x, z)$ with variable z we can derive the following:

$$\begin{aligned}
\mathrm{KL}(q(z|x)||p(z|x)) &= \int_z q(z|x)\log\frac{q(z|x)}{p(z|x)}dz \\
&= \int_z q(z|x)\log\frac{q(z|x)p(x)}{p(x,z)}dz \\
&= \int_z q(z|x)\log\frac{q(z|x)p(x)}{p(z)p(x|z)}dz \\
&= \int_z q(z|x)\log\frac{q(z|x)}{p(z)p(x|z)}dz + \log p(x) \\
&= \int_z q(z|x)\log\frac{q(z|x)}{p(z)}dz - \int_z q(z|x)\log p(x|z)dz + \log p(x) \\
\log p(x) &= \int_z q(z|x)\log p(x|z)dz - \int_z q(z|x)\log\frac{q(z|x)}{p(z)}dz + \mathrm{KL}(q(z|x)||p(z|x)) \\
\log p(x) &\geq \int_z q(z|x)\log p(x|z)dz - \int_z q(z|x)\log\frac{q(z|x)}{p(z)}dz \\
\log p(x) &\geq \int_z q(z|x)\log p(x|z)dz - \mathrm{KL}(q(z|x)||p(z))
\end{aligned} \tag{8}$$

5.3 Generative Adversarial Networks For KL-divergence minimization

Assuming $D(x) = \frac{p(x)}{q(x)+p(x)}$ and $(1 - D(x)) = \frac{q(x)}{q(x)+p(x)}$ and $D(x)$ being a Bayes-optimal discriminator , we can derive the following divergence:

$$\begin{aligned}
\mathrm{KL}(q(x)||p(x)) &= \int_x q(x)\log\frac{q(x)}{p(x)}dx \\
&= \int_x q(x)\log\frac{1 - D(x)}{D(x)}dx
\end{aligned} \tag{9}$$

This has also been considered as a loss function for adversarial learning in recent work on image super-resolution [7].

5.4 Adversarial Message Passing For KL-Divergence Minimization

In the following we will derive two distinct learning rules which will enable us to perform implicit divergence minimization using adversarial learning as a deterministic posterior approximation technique using the KL divergence. This is a similar procedure to the one considered in the main paper, but minimizes a different divergence and matches reconstructive statistics over marginal ones as performed with **JSD**.

5.4.1 Adversarial Inference With Tractable Likelihoods

The first learning rule is appropriate when we have explicitly stated models using the log-likelihood. Good-looking samples have been obtained in previous literature by blending adversarial losses and reconstruction losses and here we derive a principled explanation for some instances of them.

We assume $D(z,x) = \frac{p(z)}{q(z|x)+p(z)}$ and $(1 - D(z,x)) = \frac{q(z|x)}{q(z|x)+p(z)}$.

$$
\begin{aligned}
\log p(x) &= \int_z q(z|x) \log p(x|z) dz - \int_z q(z|x) \log \frac{q(z|x)}{p(z)} dz + \mathrm{KL}(q(z|x)||p(z|x)) \\
\log p(x) &\geq \int_z q(z|x) \log p(x|z) dz - \int_z q(z|x) \log \frac{q(z|x)}{p(z)} dz \\
&= \int_z q(z|x) \log p(x|z) dz - \int_z q(z|x) \log \frac{1 - D_z(z,x)}{D_z(z,x)} dz \\
&= \mathcal{L}_{rec}(x|\theta,\phi) - \int_z q(z|x) \log \frac{1 - D_z(z,x)}{D_z(z,x)} dz
\end{aligned}
\tag{10}
$$

We can easily draw samples for $p(z)$ and $q(z|x)$ from the prior and inference model, respectively, and can thus easily train a powerful classifier D_z to perform the required discrimination.

This setting is particularly useful when combining adversarial training with tractable likelihoods and intractable posteriors and matches the model used for Adversarial Autoencoders [10].

5.4.2 Adversarial Variational Inference With Intractable Likelihoods

For $q(x)$ being the true data distribution represented by samples of a dataset and $p(z)$ a prior, we assume $D_z(z,x) = \frac{p(z)}{q(z|x)+p(z)}$ and $(1 - D_z(z,x)) = \frac{q(z|x)}{q(z|x)+p(z)}$. We furthermore similarly assume that $1 - D_x(x,z) = \frac{q(x)}{q(x)+p(x|z)}$. Then, we can express the results from Section 5.4.1 such as to avoid having to calculate an explicit reconstruction and can express that term through the adversarial score assigned to a reconstruction.

$$
\begin{aligned}
\log p(x) &= \int_z q(z|x) \log p(x|z) dz - \int_z q(z|x) \log \frac{q(z|x)}{p(z)} dz + \mathrm{KL}(q(z|x)||p(z|x)) \\
\log p(x) &\geq \int_z q(z|x) \log p(x|z) dz - \int_z q(z|x) \log \frac{q(z|x)}{p(z)} dz \\
&= \int_z q(z|x) \log p(x|z) dz - \int_z q(z|x) \log \frac{1 - D_z(z,x)}{D_z(z,x)} dz \\
&= \int_z q(z|x) \log(1 - D_x(p(x|z))) dz - \int_z q(z|x) \log \frac{1 - D_z(z,x)}{D_z(z,x)} dz
\end{aligned}
\tag{11}
$$

This framework reveals how a carefully chosen adversarial cost and an explicit likelihood represent the same terms and can be combined. This is intuitively performed in various papers in previous literature [23, 24] and explained formally here.

5.4.3 Mixed Adversarial Variational Inference

In Sections 5.4.1 and 5.4.2 we show how KL divergence can lead to adversarial objective functions for tractable and intractable likelihoods. It is easy to see, that the two objectives shown are precisely the same for optimal discriminators and known likelihoods, since the regularizer involving the latent variable is the same. This clearly explains the ability to learn strong generative models when combining both approaches, since they correspond to the same criterion but are calculated in different ways. Optimization-wise, it may confer benefits for the learning of the discriminator to blend its cost with an explicit likelihood or regularizer on the latent variable, if such an explicit parametric form is known. Similarly, this can be chosen at any factor in a graph: applying the trick of replacing ratios with adversaries can be used at will at every factor, since the objective is not affected.

As such, we have shown that for generative modeling it is still a separate task to determine a model from its explicit learning and inference algorithm. Additionally, the choice of divergence and overall learning procedure is unrelated to picking adversarial or likelihood-based learning. Both stem from the same objective and should be used where appropriate to facilitate robust approximate inference in graphical models. Adversarial learning can better cope with intractable distributions at the cost of potential saddle points during optimization while explicit likelihood-based learning is stable at the cost of complexity in the variational approximation it induces.

5.5 Feature-based Message Passing

An alternative representation stems from a feature view on density ratios. The introduction of maximum mean discrepancy [22] provides the theoretical underpinnings to understand any distribution as a point in an adequately complicated vector space and a two-sample test to depend on the statistics on the distances betweeen different distributions represented by points in that space. The basis of many divergences is the evaluation and minimization of expectations of ratios or, in the case of the **JSD**, a softmax ratio between two distributions. In the context of MMD, this corresponds to minimizing distances in appropriate spaces between the approximate and the true distributions.

MMD-networks [25] use this methodology as a means to learn generative models and our framework fits this as well.

5.6 Divergence Minimization and Generation With Nonparametric Observation Models

Currently, sampling from $q(x|\tilde{\text{pa}}(x))$ is typically implemented using the reparametrization trick and generalizations thereof and takes the form:

$$q(x|\tilde{\text{pa}}(x)) = \int_{\epsilon} g_{rt}(f_{pm}(\tilde{\text{pa}}(x)), \epsilon). \tag{12}$$

where f_{pm} is a mapping (for instance a neural inference network) from an input to a parametric variational family.

We propose to free variational families from their parametric corsets and parametrize a more flexible variational family through a nonlinear function f_{vf}. We directly sample from the approximate posterior by injecting the noise vectors as additional inputs into the nonlinear transformation of the parents, $x^l = f_{vf}(\tilde{\text{pa}}(x), \epsilon_l)$. A (not necessarily normalized) variational family is thus modeled by:

$$q^*(x|\tilde{\text{pa}}(x)) = \int_{\epsilon} p(\epsilon) f_{vf}(\tilde{\text{pa}}(x), \epsilon) d\epsilon. \tag{13}$$

The subtle but powerful difference is that now the samples x^l can represent an arbitrary distribution, constrained only by the capacity of the nonlinear function f_{vf} and the dimensionality of the noise vector ϵ_l. This trick also forms the basis of DISCO networks [26] and was mentioned in the context of adversarial autoencoders [10]. However, we re-introduce this trick as a general tool to represent rich variational families, which are a good fit with our flexible adversarial message passing framework, thereby generalizing from the specific cases mentioned ahead to a general approximate inference framework. Specifically, previous variational inference techniques require a parametric form of the approximate posterior, such as obtained when using the reparametrization trick, in order to evaluate the divergence term needed to regularize learning. Within our framework, this divergence term is implicitly represented through samples which are scored within the adversarial framework, relieving the probabilistic modeler of the need to choose an explicit parametric form for approximate posterior families. Together with other recent powerful advances in variational inference, such as the generalized reparametrization gradient [27] and a rejection sampling generalization [28] which both learn explicit transformations $h(g(\cdot))$ to represent complex parametric variational families, this enables practical use of complicated modeling assumptions which are not limited by tractability of the typically occuring ratios within many divergence terms. We also note the concurrently published work [29], which focuses on a related idea irrespective of the link to adversarial inference, but gives deeper theoretical insights into the applicability of the same trick and provides further justification for our application thereof. Finally, we note that the same approach can also be used to specify implicit observation noise models in generative models, such as done in generative neural samplers

as introduced in the original GAN paper [30]. While this is not explicitly mentioned in [30], it is plausible that generative adversarial networks can learn arbitrary noise models that may be hard to represent analytically and the typically high-dimensional inputs to the networks can be interpreted to factorize into noise contributions and actual latent variables.