FactorNet: Learning to Factorize Intractable and Multi-Modal Posterior Distributions for Energy Disaggregation

Henning Lange
Carnegie Mellon University
Pittsburgh, USA

Mario Bergés
Carnegie Mellon University
Pittsburgh, USA

***Abstract*—Factorial Hidden Markov Models (FHMM) have emerged as a prominent modeling approach for energy disaggregation. However, because latent variables become dependent conditioned on the observation, reasoning about the posterior is usually intractable which is required for inference as well as learning. Recent approaches try to deal with these intractable posterior distributions by applying Variational Inference with an auxiliary distribution that assumes independence between latent states of the posterior. However, because posterior distributions in the context of energy disaggregation are often multi-modal, independent auxiliary distributions fail to capture *either-or* relationships between appliance states. In this paper, we introduce an auxiliary distribution over posterior states that, in principle, can approximate any multivariate Bernoulli distribution arbitrarily well, while at the same time offering a functional form that allows obtaining independent samples as well as the mode required for inference in $\mathcal{O}(N)$ where N is the number of parallel Hidden Markov chains. On top of that, training the distribution requires solely samples of the joint distribution which are typically easy to acquire. We conduct experiments in the context of waveform disaggregation illustrating the superior capacity of the proposed distribution in comparison to independent auxiliary distributions trained on minimizing the forward or backward KL-divergence.**

I. Introduction

Factorial Hidden Markov Model [1] (FHMM) are a natural choice for modeling the generative process of energy disaggregation [2], [3], [4], [5]. FHMM are a generalization of Hidden Markov Models were multiple hidden chains evolve independently in parallel. Usually, the state of a single appliance is modeled by a single HMM chain, whereas the aggregate power measured at the main distribution panel is modeled by the aggregate observation. Let $z \in \mathcal{Z} = \{0,1\}^{N \times T}$ be the latent variable and $x \in \mathbb{R}^{S \times T}$ be the aggregate observation with T number of time steps, N number of parallel HMM chains and S being the observation dimensionality. The joint distribution is defined as:

$$p(x_{1:T}, z_{1:T}) = \prod_t^T p(x_t|z_t) \prod_i^N p(z_{t,i}|z_{t-1,i}) p(z_{0,i})$$

However, reasoning about the posterior of P is usually difficult because the latent variables become conditionally dependent given the observation, specifically for the forward and filtering distribution hold respectively:

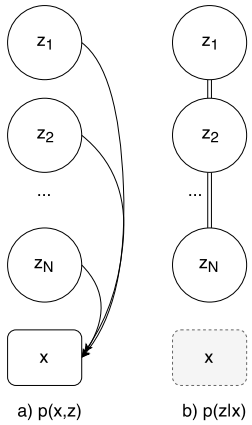

Fig. 1. a) latent variables of the proposed distribution are marginally independent, however, become b) conditionally dependent given the observation

$$p(x_{1:t}, z_t) = p(x_t|z_t) \sum_{z' \in \mathcal{Z}} p(z_t|z') p(z_{t-1}, x_{1:t-1}) \quad (1)$$

$$p(z_t|x_{1:t}) = \frac{p(x_{1:t}, z_t)}{\sum_{z' \in \mathcal{Z}} p(x_{1:t}, z')} \quad (2)$$

Note that (1) and (2) both contain summations over $\mathcal{Z}$ and that the cardinality of $\mathcal{Z}$ grows exponentially with N.

Throughout this paper, for illustration, we will consider a slightly simpler distribution that nevertheless faces the same difficulty but for which exact solutions can be obtained and visualized for small N. Consider the graphical model that arises from removing the temporal dependencies between latent variables (Figure 1a). Similar to FHMMs, latent variables become dependent conditioned on the observation x (Figure 1b). For the density holds:

$$p(x_{1:T}, z_{1:T}) = \prod_t^T p(x_t, z_t) \quad (3)$$

As for FHMMs, the posterior of (3) is intractable, i.e. the number of states grows exponentially with the latent dimensionality rendering the denominator of the posterior intractable. However, previously, statistical tools such as Variational Inference

[6] have been applied to reason about intractable posterior distributions in the context of energy disaggregation [4], [3]. The main idea of Variational Inference is to introduce a tractable auxiliary distribution Q_ψ parameterized by the variational parameters ψ. Inference is then turned into an optimization problem, i.e. ψ is optimized in such a way that Q best approximates P as measured by the KL-divergence. Then, in order to perform inference on the intractable posterior P inference can be carried out on Q_ψ instead. Since Q_ψ is required to be tractable, usually additional independence assumption are made and specifically, in the context of energy disaggregation, in order to deal with the difficulty of dependent latent variables, independence between latent states in the posterior is assumed. Note that Q_ψ is usually required to be simpler than P, i.e. to have less capacity than P.

However, because inference is carried out on a simpler distribution, Variational Inference maximizes a lower bound on the data likelihood $p(x)$, i.e. it performs inference up to a constant and it can be shown that this constant is the KL divergence between P and Q_ψ. Note also that because Q is required to be simpler than P, the KL divergence usually never becomes 0.

Furthermore, if independence between latent states is assumed in Q_ψ, i.e. the posterior is factored as:

$$q_\psi(z_t|x_t) = \prod_i f_\psi(x_t)_i^{z_i}(1 - f_\psi(x_t)_i)^{1-z_i} \tag{4}$$

with f being bounded by $[0, 1]$, Q_ψ is often overly simple. It is easy to show that depending on whether the forward or backward KL divergence is employed as a divergence measure, the Q introduced in (4) either learns the mean or the mode of P. Specifically, for energy disaggregation, such a unimodal Q is unable to learn *either* this appliance *or* the other.

Consider a scenario with 2 two-state appliances with comparable power draw and an aggregate observation x' that is similar to the power consumption of each appliances. Thus we can assume that for the posterior the following holds:

$$p(z|x') = \begin{pmatrix} 0 & 0.5 & 0.5 & 0 \end{pmatrix}$$
$$\text{with } z = \begin{pmatrix} 0,0 & 0,1 & 1,0 & 1,1 \end{pmatrix}$$

Note that approaches that assume independence between latent states of the auxiliary distribution fail at capturing the *either-or* relationship between appliance states. Let ψ_f^* and ψ_b^* be optimal variational parameters that minimize the forward and backward KL-divergence between P and Q respectively. It can be shown that:

$$q_{\psi_f^*}(z|x') = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}$$
$$q_{\psi_b^*}(z|x') = \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} \text{ or } q_{\psi_b^*}(z|x') = \begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix}$$

It is easy to see that independent of the choice of divergence measurement, Q cannot capture a significant proportion of the information present in P, specifically the fact that one of the appliances is active but not both or none.

That is why we argue that previous approaches based on Variational Inference can be improved by a better choice of the auxiliary distribution. Thus, in this paper, we introduce a tractable auxiliary distribution g that despite being tractable can approximate any discrete distribution arbitrarily well. To sum up, we propose an auxiliary distribution that has the following characteristics:

1) No independence assumptions and therefore unlimited capacity, i.e. in general, any multivariate Bernoulli distribution can be approximated arbitrarily well
2) The posterior can be trained efficiently based on samples of the joint $p(x, z)$
3) Computing the mode and drawing independent samples can be achieved in $\mathcal{O}(N)$

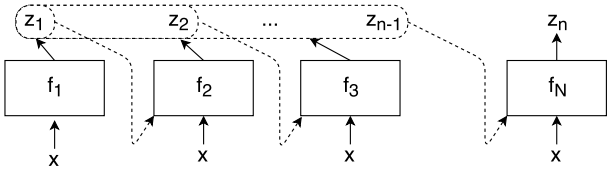

Fig. 2. A graphical depiction of the cascaded neural networks that factorize the joint probability distribution.

In the next section we will provide a brief introduction into Variational Inference and introduce FactorNet, the proposed auxiliary distribution. We then conduct experiments in section 3 and conclude our findings.

II. Variational Inference and FactorNet

Variational Inference (VI) has experienced a recent surge in attention from various academic communities [7], [8]. One of the key advantages of VI over its alternatives such as Markov Chain Monte Carlo [9] (MCMC) is speed. Since, as stated earlier, VI translates statistical inference into an optimization problem that produces a tractable distribution that best approximates the true posterior, inference can be amortized, i.e. time training the auxiliary distribution is spend once and after training, inference can be carried out extremely fast. This characteristic has direct implications in the context of energy disaggregation: VI-based approaches allow for inference on cheap hardware such as an electricity meter located in the premises whereas MCMC would require remotely collecting, storing and processing data. However, even in the asymptotic regime, VI is an approximate inference technique whereas (albeit slowly) MCMC is known to converge to the true posterior. The quality of the VI-based approximation crucially depends on the choice of the auxiliary distribution which can be seen when investigating the commonly used Expectational Lower Bound as the variational objective:

$$\log p(x) = \sum_z \log p(x|z)p(z) \tag{5}$$
$$= \sum_z \frac{q(z|x)}{q(z|x)} \log p(x|z)p(z) \tag{6}$$
$$\geq D_{KL}[q(z|x)||p(z)] + \mathbb{E}_{q(z|x)}[\log p(x|z)] \tag{7}$$

This inequality is tight if and only if $p(z|x) = q(z|x)$, however, this cannot be achieved when Q is simpler than

P. Furthermore note, that (7) is typically evaluated by Monte Carlo techniques, i.e. by evaluating the expectation by sampling from Q. Thus, in order for a Variational approach to be successful, Q needs to be complex enough to be fit to P tightly but simple enough to be sampled from efficiently.
For continuous distributions the problem of choosing a suitable posterior distribution has recently been addressed by introducing normalizing flows[10], i.e. a succession of invertible non-linear transformations of the random variable z. However, for discrete random variables this approach does not seem to be possible since the flow-operators are required to be differentiable but to be mapping into the same domain (in this case $\{0,1\}^N$). Or in other words, the flow-operator cannot at the same time be mapping into the discrete domain whilst being smooth and differentiable.
Furthermore, another difficulty that arises for VI-based approaches is the fact that the true posterior is usually not obtainable, thus all updates need to be made based on samples of the joint $p(x,z)$. Typically, this is circumvented by maximizing the variational objective (7), however, in the experience of the authors (7) has suboptimal convergence properties.
Thus, in this paper, we follow a different strategy. We directly learn the conditional factorization of the joint and show that once the joint is factorized, obtaining the posterior can be done efficiently. First, we note that any joint probability distribution can be factored according to the chain rule of probabilities:

$$\begin{aligned} p(z_t, x_t) &= p(z_{t,1}, x_t)p(z_{t,2}, x_t|z_{t,1})...p(z_{t,N}, x_t|z_{t,N-1}, ..., z_{t,1}) \\ &= \prod_n^N p(z_{t,n}, x_t|z_{t,1:n}) \end{aligned}$$

The goal now is to learn this factorization. This is achieved by approximating every factor of the probability distribution by a neural network that takes the respective condition as input and produces the conditional joint probability. Thus, let g be the FactorNet distribution and f_n and $\overline{f}_n$ with $1 \leq n \leq N$ be the N neural networks approximating the *on* and *off* factors of the joint distribution, i.e.:

$$\begin{aligned} f_i(x_t, z_1, ..., z_{i-1}) &\approx p(x_t, z_i = 1|z_1, ..., z_{i-1}) \\ \overline{f}_i(x_t, z_1, ..., z_{i-1}) &\approx p(x_t, z_i = 0|z_1, ..., z_{i-1}) \end{aligned}$$

therefore:

$$\begin{aligned} &p(z_i = 1|x_t, z_1, ..., z_{i-1}) \\ \approx &\frac{f_i(x_t, z_1, ..., z_{i-1})}{f_i(x_t, z_1, ..., z_{i-1}) + \overline{f}_i(x_t, z_1, ..., z_{i-1})} \\ = &f_i^*(x_t, z_1, ..., z_{i-1}) \end{aligned}$$

For the FactorNet joint distribution the following then holds:

$$\begin{aligned} g(z_t, x_t) = \prod_i^N &f_i(x_t, z_1, ..., z_{i-1})^{z_i} \\ &\overline{f}_i(x_t, z_1, ..., z_{i-1})^{(1-z_i)} \end{aligned}$$

and for its posterior:

$$\begin{aligned} g(z_t|x_t) = \prod_i^N &f_i^*(x_t, z_1, ..., z_{i-1})^{z_i} \\ &(1 - f_i^*(x_t, z_1, ..., z_{i-1}))^{(1-z_i)} \end{aligned}$$

Note that because the joint instead of the posterior probability is factorized, $f_i(x_t, z_1, ..., z_{i-1}) + \overline{f}_i(x_t, z_1, ..., z_{i-1}) \neq 1$ and that even though no independence assumption between latent variables has been made, evaluating the joint as well as the posterior probability is linear in the latent dimensionality as opposed to exponential for evaluating P. Furthermore, we can take independent samples from the posterior of G efficiently, i.e. linear time. That is, we do not have to resort to Markov Chain Monte Carlo techniques for drawing samples from g, which would, in principle, allow for an efficient Monte Carlo approximation of the expectation of (7) given the samples from Q. See Algorithm 1 for how to sample from $g(z|x)$.

Result: Sample or Mode of $g(z|x_t)$
$z = \{\}$;
for $n = 1, ..., N$ **do**
  $p_n = f_n(x_t, z)/(f_n(x_t, z) + \overline{f}_n(x_t, z))$;
  if $p_n > threshold$ **then**
    Append 1 to z
  else
    Append 0 to z;
  end
end

Algorithm 1: Outputs either an independent sample or the mode of $g(z|x_t)$. If the mode is desired, set $threshold = 0.5$ and to a sample from $g(z|x_t)$ set $threshold \sim U[0,1]$, i.e. to a sample from a uniform distribution.

However, as stated above, (7) has suboptimal convergence properties that can be circumvented by exploiting the fact that G allows to efficiently obtain the joint as well as the posterior. That is why we propose a learning objective that directly minimizes the KL-divergence between the joint distributions, i.e.:

$$\mathcal{L} = -g(z_t, x_t) \log \frac{p(z_t, x_t)}{g(z_t, x_t)}$$

Note that we do not allow the gradients to flow into the fraction, i.e. we treat $g(z_t, x_t)$ in the denominator as a constant.

III. Experiments

The efficacy of FactorNet is evaluated on a synthetic experiment in the context of supervised waveform disaggregation. Specifically, we choose 8 appliances from the PLAID dataset[11] and extract a single steady-state current waveform for every appliance aligned by zero-crossing of the voltage line. PLAID is a publicly available dataset containing high-frequency current and voltage measurements of single appliances. Since PLAID is collected at 30kHz, approximately 500 samples are collected per voltage cycle. Thus a matrix $W \in \mathbb{R}^{500 \times 8}$ was extracted from PLAID and Figure 3 shows

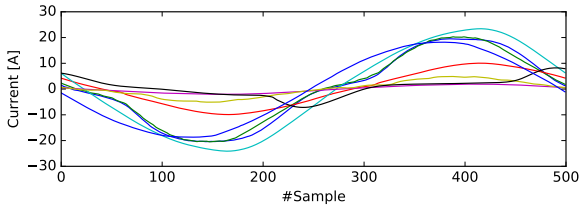

Fig. 3. The current waveforms used in the synthetic experiment taken from PLAID datasets. Current waveforms were extracted by alignment to zero-crossings in the voltage line.

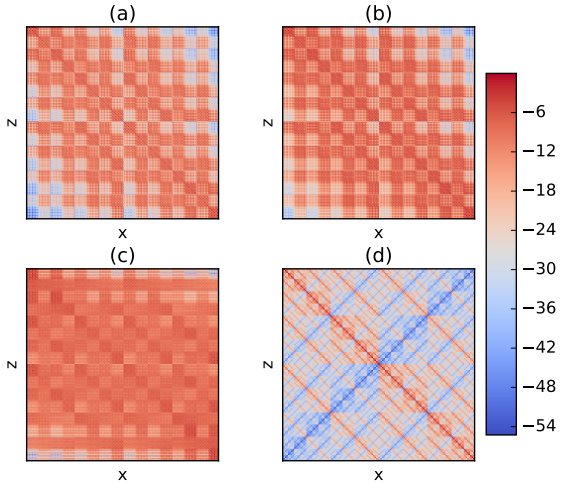

Fig. 4. (a) The true posterior $\log p(z|x)$ (b) The FactorNet posterior $\log g(z|x)$ (c) The posterior $\log q(z|x)$ minimizing the forward KL-divergence (d) The posterior $\log q(z|x)$ minimizing the backward KL-divergence. Note that all probabilities were clipped between 0.001 and 0.999 to avoid $\log(0)$

the waveforms used in the experiments. The 8 appliance waveforms were then mixed up, i.e. all 256 possible combinations of waveforms were created and corrupted by Gaussian noise: $X = \{Wz + \mathcal{N}(0, 0.1I)|z \in \{0,1\}^8\}$. The probability of the aggregate observation was defined as:

$$p(x_t|z_t) = \mathcal{N}(x_t|Wz_t, 0.1I)$$

with W being a matrix containing the appliance waveforms and I being the identity matrix. For the posterior thus holds:

$$p(z_t|x_t) = \frac{\mathcal{N}(x_t|Wz_t, 0.1I)}{\sum_z \mathcal{N}(x|Wz, 0.1I)}$$

For every combination of $z \in \{0,1\}^8$ and $x \in X$, $\log p(z|x)$ was computed and stored. See Figure 4(a) for a plot of the resulting 256×256 matrix.

Eight neural networks with a similar topology were created with an input dimensionality of $500+(n-1)$, two intermediate *relu*-layers with 512 hidden units and two-unit *sigmoid* output-layer for f and $\overline{f}$ respectively. The network was trained by minimizing $\mathcal{L}$ introduced earlier. The objective was minimized by drawing mini-batches of 144 samples uniformly from

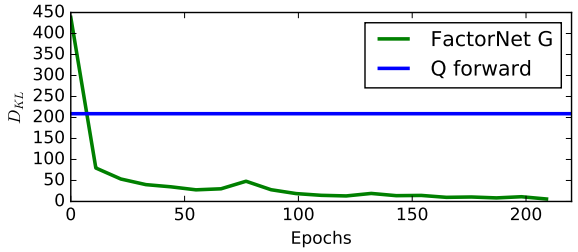

Fig. 5. The KL-divergence $D_{KL}(p(z|x)||f(z|x))$ summed over all x. In this case f is either the FactorNet distribution g or the $q_{\psi_f^*}$ minimizing the forward KL divergence. Note that $q_{\psi_b^*}$ minimizing the backward KL divergence did not fit onto the plot with a divergence of approximately 3800.

the joint distribution $p(z, x)$. The training procedure did not assume knowledge of the posterior $p(z|x)$ and was solely presented with sampled of the joint. The performance of the algorithm is compared to distributions $q_{\psi_f^*}$ and $q_{\psi_b^*}$ introduced earlier, i.e. distributions that assume independence between latent states in the posterior and minimize the forward and backward KL-divergence respectively. The parameters ψ_f^* and ψ_b^* were obtained with the knowledge of the true posterior that usually is not available, thus we compare to distributions in their globally optimal configuration.

Figure 4 shows a visual comparison of the different resulting posterior distributions. One can see that FactorNet G captures much more information present in P compared to Q in both settings. Figure 5 emphasizes this fact as it shows the KL-divergence over time. One can see that FactorNet reaches a KL-divergence of practically 0 after approximately 100 iterations.

IV. Conclusion

We introduced an auxiliary distribution capable of approximating any multivariate Bernoulli distribution arbitrarily well whilst at the same time having a functional form that is simple enough to allow for drawing samples as well as computing the mode of the posterior efficiently. The joint as well as posterior distribution can be obtained in linear time by approximating the chain rule factorization through a succession of neural networks, which allows for using a training objective that minimizes the divergence between the joint distributions directly circumventing the need for ELBO minimization. Positive experimental results of the performance were obtained in the setting of supervised waveform disaggregation.

However, experiments in which FactorNet incorporates temporal dependencies have not yet been conducted. Note that FactorNet was conceived out of the realization that auxiliary distributions that assume independence in the posterior are detrimental when modeling temporal dependencies, i.e. the posterior collapses onto a single state and most of the uncertainty is falsely explained away. This prohibits temporal models from reversing previous decisions like e.g. the Viterbi [12] algorithm would. FactorNets performance with temporal dependencies needs yet to be determined.

REFERENCES

[1] Z. Ghahramani and M. I. Jordan, "Factorial hidden markov models," *Machine learning*, vol. 29, no. 2-3, pp. 245–273, 1997.

[2] J. Z. Kolter and T. Jaakkola, "Approximate inference in additive factorial hmms with application to energy disaggregation," in *International conference on artificial intelligence and statistics*, 2012, pp. 1472–1482.

[3] H. Lange and M. Bergés, "Variational bolt: Approximate learning in factorial hidden markov models with application to energy disaggregation," in *to appear: AAAI 2018*, 2018.

[4] Y. C. Ng, P. M. Chilinski, and R. Silva, "Scaling factorial hidden markov models: Stochastic variational inference without messages," in *Advances in Neural Information Processing Systems*, 2016, pp. 4044–4052.

[5] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[6] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[7] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, 2013.

[8] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[9] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 6, pp. 721–741, 1984.

[10] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," *arXiv preprint arXiv:1505.05770*, 2015.

[11] J. Gao, S. Giri, E. C. Kara, and M. Bergés, "Plaid: a public dataset of high-resoultion electrical appliance measurements for load identification research: demo abstract," in *proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*. ACM, 2014, pp. 198–199.

[12] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *Information Theory, IEEE Transactions on*, vol. 13, no. 2, pp. 260–269, 1967.