Appendix for "Stein Variational Message Passing for Continuous Graphical Models"

A. Proof of Theorem 1

Proof. Consider $\boldsymbol{f} = [f_1, \ldots, f_d]^\top \in \mathcal{H}$, where $\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_d$. Using the reproducing property of $\mathcal{H}_i$, we have for any $f_i \in \mathcal{H}_i$

$$\mathbb{E}_{x\sim q}[\mathcal{P}_{x_i} f_i(x)] = \langle f_i,\ \mathbb{E}_{x\sim q}[\mathcal{P}_{x_i} k_i(x, \cdot)]\rangle_{\mathcal{H}_i}.$$

Recall that $\phi_i^*(\cdot) = \mathbb{E}_{x\sim q}[\mathcal{P}_{x_i} k_i(x,\cdot)]$, and $\boldsymbol{\phi}^* = [\phi_1^*, \ldots, \phi_d^*]^\top$. The optimization of the Stein Discrepancy is framed into

$$\begin{aligned}
\mathbb{D}(q \,||\, p) &= \max_{\boldsymbol{f}\in\mathcal{H}, ||\boldsymbol{f}||_{\mathcal{H}}\leq 1} \mathbb{E}_{x\sim q}[\mathcal{P}_x{}^\top \boldsymbol{f}(x)] \\
&= \max_{\boldsymbol{f}\in\mathcal{H}, ||\boldsymbol{f}||_{\mathcal{H}}\leq 1} \sum_{i=1}^d \mathbb{E}_{x\sim q}[\mathcal{P}_{x_i} f_i(x)] \\
&= \max_{\boldsymbol{f}\in\mathcal{H}, ||\boldsymbol{f}||_{\mathcal{H}}\leq 1} \sum_{i=1}^d \langle f_i, \phi_i^*\rangle_{\mathcal{H}_i} \\
&= \max_{\boldsymbol{f}\in\mathcal{H}, ||\boldsymbol{f}||_{\mathcal{H}}\leq 1} \langle \boldsymbol{f}, \boldsymbol{\phi}^*\rangle_{\mathcal{H}}.
\end{aligned}$$

This shows that the optimal $\boldsymbol{f}$ should equal $\boldsymbol{\phi}^*/||\boldsymbol{\phi}^*||_{\mathcal{H}}$, and $\mathbb{D}(q \,||\, p) = \langle \boldsymbol{\phi}^*/||\boldsymbol{\phi}^*||_{\mathcal{H}}, \boldsymbol{\phi}^*\rangle_{\mathcal{H}} = ||\boldsymbol{\phi}^*||_{\mathcal{H}}$. □

B. Proof of Theorem 2

Proof. Plugging the optimal solution in Theorem 1 into the definition of Stein discrepancy (2), we get

$$\begin{aligned}
\mathbb{D}(q \,||\, p) &= \frac{1}{||\boldsymbol{\phi}^*||_{\mathcal{H}}} \mathbb{E}_{x\sim q}[\mathcal{P}_x{}^\top \boldsymbol{\phi}^*(x)] \\
&= \frac{1}{||\boldsymbol{\phi}^*||_{\mathcal{H}}} \sum_{i=1}^d \mathbb{E}_{x\sim q}[\mathcal{P}_{x_i} \phi_i^*(x)] \\
&= \frac{1}{||\boldsymbol{\phi}^*||_{\mathcal{H}}} \sum_{i=1}^d \mathbb{E}_{x,x'\sim q}[\mathcal{P}_{x_i}\mathcal{P}_{x_i'} k_i(x,x')].
\end{aligned}$$

On other hand, because $\mathbb{D}(q \,||\, p) = ||\boldsymbol{\phi}^*||_{\mathcal{H}}$, we have

$$\mathbb{D}(q \,||\, p)^2 = \sum_{i=1}^d \mathbb{E}_{x,x'\sim q}[\mathcal{P}_{x_i}\mathcal{P}_{x_i'} k_i(x,x')]. \tag{B.1}$$

To prove (10), note that

$$\begin{aligned}
&\mathbb{E}_{x\sim q}[\mathcal{P}_{x_i} f(x)] \\
&= \mathbb{E}_{x\sim q}[\mathcal{P}_{x_i} f] - \mathbb{E}_q[\mathcal{Q}_{x_i} f] \\
&= \mathbb{E}_{x\sim q}[(\nabla_{x_i}\log p(x) - \nabla_{x_i}\log q(x)) f(x)] \\
&= \mathbb{E}_{x\sim q}[(\nabla_{x_i}\log p(x_i|x_{\neg i}) - \nabla_{x_i}\log q(x_i|x_{\neg i})) f(x)] \\
&= \mathbb{E}_{x\sim q}[\delta_i(x) f(x)].
\end{aligned}$$

Applying this equation twice to (B.1) gives

$$\mathbb{D}(q \,||\, p)^2 = \sum_{i=1}^d \mathbb{E}_{x,x'\sim q}[\delta_i(x) k_i(x,x') \delta_i(x')]. \tag{B.2}$$

By (B.2) and the definition of strictly integrally positive definite kernels, we can see that $\mathbb{D}(q \,||\, p) = 0$ implies $\delta_i(x) = 0$, $\forall i \in [d]$, if $k_i(x,x')$ is strictly integrally positive definite for each i. Note that $\delta_i(x) = 0$ means p and q matches the conditional probabilities:

$$p(x_i|x_{\neg i}) = q(x_i|x_{\neg i}), \quad \forall i \in [d]. \tag{B.3}$$

For positive densities, this implies that $p(x) = q(x)$ (see e.g., Brook (1964); Besag (1974)). □

C. Proof of Theorem 3

Proof. For a graphical model $p(x)$ with Markov blanket $\mathcal{N}_i$ for node i, we have

$$\nabla_{x_i}\log p(x_i|x_{\neg i}) = \nabla_{x_i}\log p(x_i|x_{\mathcal{N}_i}) \quad \forall i \in [d].$$

Moreover, by Stein's identity on q, we have

$$\mathbb{E}_{x\sim q}[\nabla_{x_i}\log q(x_i \mid x_{\mathcal{N}_i}) f(x) + \nabla_{x_i} f(x)] = 0, \quad \forall i \in [d].$$

With a similar argument as the proof of Theorem 2, we get

$$\begin{aligned}
&\mathbb{E}_{x\sim q}[\mathcal{P}_{x_i} f(x)] \\
&= \mathbb{E}_{x\sim q}[(\nabla x_i \log p(x_i|x_{\mathcal{N}_i}) - \nabla x_i \log q(x_i|x_{\mathcal{N}_i})) f(x)] \\
&= \mathbb{E}_{x\sim q}[\delta_i(x_{\mathcal{C}_i}) f(x)],
\end{aligned}$$

where $\delta_i(x_{\mathcal{C}_i}) = \nabla_{x_i}\log q(x_i|x_{\mathcal{N}_i}) - \nabla_{x_i}\log p(x_i|x_{\mathcal{N}_i})$.

Applying this equation twice to (B.1) gives

$$\mathbb{D}(q \,||\, p)^2 = \sum_{i=1}^d \mathbb{E}_{x,x'\sim q}[\delta_i(x_{\mathcal{C}_i}) k_i(x,x') \delta_i(x'_{\mathcal{C}_i})].$$

Therefore, if $k_i(x, x')$ is strictly integrally positive definite on $x_{\mathcal{C}_i}$, Stein discrepancy $\mathbb{D}(q \mid\mid p) = 0$ if and only if $q(x_i|x_{\mathcal{N}_i}) = p(x_i|x_{\mathcal{N}_i})$.

$\square$

References

Besag, J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 192–236, 1974.

Besse, F., Rother, C., Fitzgibbon, A., and Kautz, J. PMBP: Patchmatch belief propagation for correspondence field estimation. *International Journal of Computer Vision*, 110(1):2–13, 2014.

Brook, D. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, 51 (3/4):481–483, 1964.

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

Gorham, J. and Mackey, L. Measuring sample quality with kernels. *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Hoffman, M. D. and Gelman, A. The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1): 1593–1623, 2014.

Ihler, A. and McAllester, D. Particle belief propagation. In *Artificial Intelligence and Statistics*, pp. 256–263, 2009.

Ihler, A. T., Fisher, J. W., Moses, R. L., and Willsky, A. S. Nonparametric belief propagation for self-localization of sensor networks. *IEEE Journal on Selected Areas in Communications*, 23(4):809–819, 2005.

Lauritzen, S. L. *Graphical models*, volume 17. Clarendon Press, 1996.

Liu, Q. Stein variational gradient descent as gradient flow. In *Advances in neural information processing systems (NIPS)*, pp. 3118–3126, 2017.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances In Neural Information Processing Systems (NIPS)*, pp. 2370–2378, 2016.

Liu, Q., Ihler, A. T., and Steyvers, M. Scoring workers in crowdsourcing: How many control questions are enough? In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1914–1922, 2013.

Liu, Q., Lee, J. D., and Jordan, M. I. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of the International Conference on Machine Learning*, 2016.

Nelsen, R. B. *An introduction to copulas*. Springer Science & Business Media, 2007.

Oates, C. J., Cockayne, J., Briol, F.-X., and Girolami, M. Convergence rates for a class of estimators based on Stein's identity. *arXiv preprint arXiv:1603.03220*, 2016.

Pacheco, J. and Sudderth, E. Proteins, particles, and pseudo-max-marginals: A submodular approach. In *International Conference on Machine Learning (ICML)*, pp. 2200–2208, 2015.

Pacheco, J., Zuffi, S., Black, M., and Sudderth, E. Preserving modes and messages via diverse particle selection. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pp. 1152–1160, 2014.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988.

Song, L., Gretton, A., Bickson, D., Low, Y., and Guestrin, C. Kernel belief propagation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 707–715, 2011.

Sudderth, E. B., Ihler, A. T., Isard, M., Freeman, W. T., and Willsky, A. S. Nonparametric belief propagation. *Communications of the ACM*, 53(10):95–103, 2010.

Wainwright, M. and Jordan, M. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, 2008.

Wang, D., Fisher III, J. W., and Liu, Q. Efficient observation selection in probabilistic graphical models using Bayesian lower bounds. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2016.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239, 2003.

Zhuo, J., Liu, C., Shi, J., Zhu, J., Chen, N., and Zhang, B. Message passing Stein variational gradient descent. *arXiv preprint arXiv:1711.04425v2*, 2018.