Appendix

A. Omitted Proofs

Proposition. *If $q(x,\theta)$ is a probability density function such that $q''_x(x,\theta) = q''_\theta(x,\theta)$, then we have*

$$L''(\theta) = \int_{-\infty}^{+\infty} q''_\theta(x,\theta) f(x) dx = \int_{-\infty}^{+\infty} q(x,\theta) f''(x) dx \tag{10}$$

for sufficiently smooth f. This also holds in higher dimensions under the mean-field assumption.

Proof. Integrate by parts twice. First, you have

$$\int_{-\infty}^{+\infty} q(x,\theta) f''(x) dx = q(x,\theta) f'(x)\Big]_{x=-\infty}^{x=+\infty} - \int_{-\infty}^{+\infty} q'_x(x,\theta) f'(x) dx$$

using $u = q(x,\theta)$ and $v = f'(x)$. Then integrate by part again

$$\int_{-\infty}^{+\infty} q(x,\theta) f''(x) dx = q(x,\theta) f'(x)\Big]_{x=-\infty}^{x=+\infty} - q'_x(x,\theta) f(x)\Big]_{x=-\infty}^{x=+\infty} + \int_{-\infty}^{+\infty} q''_x(x,\theta) f(x) dx$$

using $u = q'_x(x,\theta)$ and $v = f(x)$ □

Proposition. *ELBO($\boldsymbol{\rho}$) is DR-Submodular in $\boldsymbol{\rho}$.*

Proof. We already proved that $\frac{\partial^2 F}{\partial \rho_{ij} \partial \rho_{kl}} = 0$ when $i = k$ for all j, l and $\frac{\partial^2 F}{\partial \rho_{ij} \partial \rho_{kl}} \leq 0$ when $i \neq k$ for all j, l. On the other hand, $\frac{\partial^2 H_T}{\partial \rho_{ij} \partial \rho_{kl}} \leq 0$ when $i = k$ for all j, l and $\frac{\partial^2 H_T}{\partial \rho_{ij} \partial \rho_{kl}} = 0$ when $i \neq k$ for all j, l since $H(\boldsymbol{\rho}_i) = -(\rho_{i1} \log \rho_{i1} + \rho_{i2} \log \rho_{i2} + \ldots + (1 - \rho_{i1} - \ldots - \rho_{i,k-1}) \log(1 - \rho_{i1} - \ldots - \rho_{i,k-1}))$. Therefore $\frac{\partial^2 \text{ELBO}}{\partial \rho_{ij} \partial \rho_{kl}} \leq 0$ for all i, j, k, l. □

Proposition. *Considering maximizing ELBO($\boldsymbol{\rho}$) in Equation (7), if one only optimize for $\boldsymbol{\rho}_i$ while keeping all other marginals fixed, we have the following closed form solution (let $\nabla_{ij} := \nabla_{\rho_{ij}} F(\boldsymbol{\rho})$ for notational simplicity):*

$$\rho_{ij} = \frac{\exp(\nabla_{ij})}{1 + \sum_{j'} \exp(\nabla_{ij'})}, \forall j \in \{1, ..., k-1\}. \tag{11}$$

Proof. Firstly notice that the generalized multilinear extension $F(\boldsymbol{\rho})$ is linear in terms of each ρ_{ij}, and it is separable for $\rho_{i1}, \rho_{i2}, ..., \rho_{i,k-1}$ for a fixed i. So if we fix all of the other marginals except for $\boldsymbol{\rho}_i$, ∇_{ij} will be a constant for all $j \in [k-1]$.

Secondly the entropy term $H(\boldsymbol{\rho})$ is concave in terms of ρ_{ij}, so the ELBO($\boldsymbol{\rho}$) is concave in terms of ρ_{ij}. In order to find the maximizer of this $(k-1)$ dimensional concave function, we just need to set $\nabla_{\boldsymbol{\rho}_i}$ELBO($\boldsymbol{\rho}$) to be zero. One can verify that

$$\begin{aligned} \nabla_{ij}\text{ELBO}(\boldsymbol{\rho}) &= \nabla_{ij} + \nabla_{ij} H(\boldsymbol{\rho}) \\ &= \nabla_{ij} + \log \frac{1 - \sum_{j'} \rho_{ij'}}{\rho_{ij}} \stackrel{!}{=} 0, \forall j \in \{1, ..., k-1\}. \end{aligned} \tag{12}$$

Solving the above $k-1$ equations, we get that $\rho_{ij} = \frac{\exp(\nabla_{ij})}{1 + \sum_{j'} \exp(\nabla_{ij'})}$.

Lastly, with this update rule, the simplex constraints are always satisfied. Because after this update, we have

$$\sum_{j'} \rho_{ij'} = \frac{\sum_{j'} \exp(\nabla_{ij})}{1 + \sum_{j'} \exp(\nabla_{ij})} \tag{13}$$

$$< 1. \tag{14}$$

□

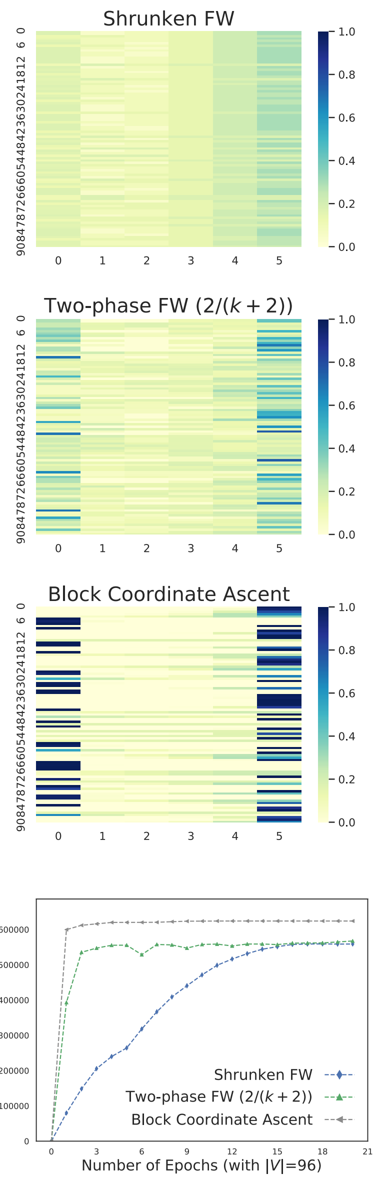
B. More on Experiments

The graph datasets and corresponding experimental parameters are documented in the following table:

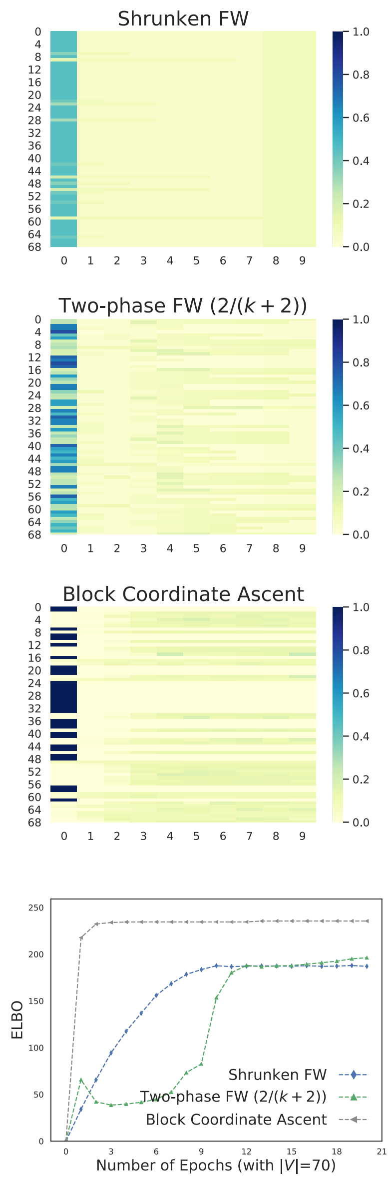
Dataset	n	#edges	q	#categories
"Seventh graders"	29	376	0.7	6
"Highschool"	70	366	0.2	10
"Reality Mining"	96	1,086,404 (multiedge)	0.75	6
"Residence hall"	217	2,672	0.75	10
"Infectious"	410	17,298	0.7	6

Table 1. Graph datasets and corresponding experimental parameters

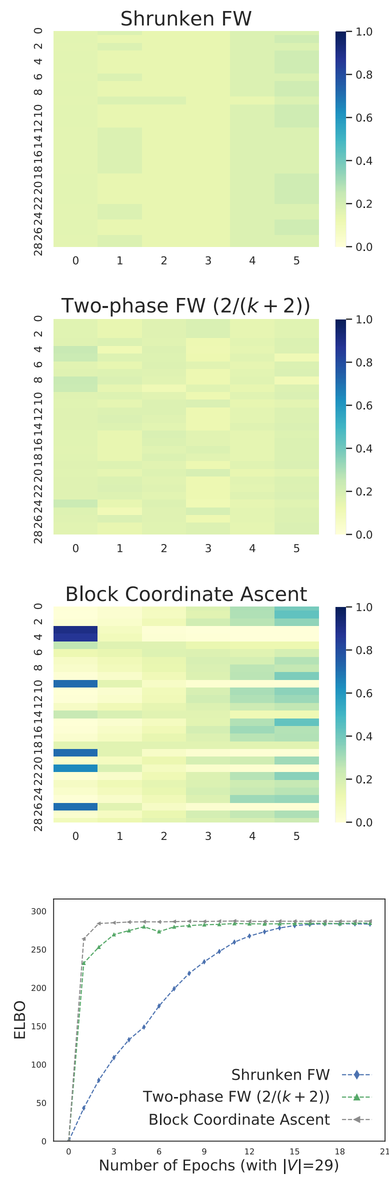
In the plots for the marginals, we always observe the same behaviour: Shrunken FW gives smoother marginals than the Block CA. When we look at the trajectories, we see that Block CA always obtains the highest ELBO value and converges the fastest. Shrunken FW usually obtains slower convergence than Two Phase FW and obtains lower ELBO values.



(a) “Reality Mining” dataset $n = 96$.



(b) “Highschool” dataset $n = 70$.



(c) “Seventh Graders” dataset $n = 29$.

Figure 4. Marginals and Trajectories for Different Datasets and Functions