Amortized Population Gibbs Samplers with Neural Sufficient Statistics

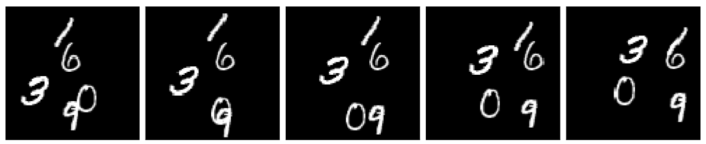
Hao Wu[1], Heiko Zimmermann[1], Eli Sennesh[1], Tuan Anh Le[2], Jan-Willem van de Meent[1]
Northeastern University[1], Massachusetts Institute of Technology[2]

Summary

We develop amortized population Gibbs (**APG**) Samplers, a class of scalable methods in structured variational inference. APG samplers construct high-dimensional proposals by iterating over updates to lower-dimensional blocks of variables.We train each conditional proposal by minimizing the inclusive KL divergence with respect to the conditional posterior. To appropriately account for the size of the input data, we develop a new parameterization in terms of neural sufficient statistics.

Structured Deep Probabilistic Models

Example: Unsupervised Tracking



- Corpus level (many videos): Digit shapes, Transition dynamics
- Instances level (single videos): Object features
- Data-points level (single frames): Object positions

Reweighted Wake-Sleep (RWS) Style Methods

(Standard Variational Methods)

$x_{1:T}$: data, η: features, $c_{1:T}$: positions.

Generative Model

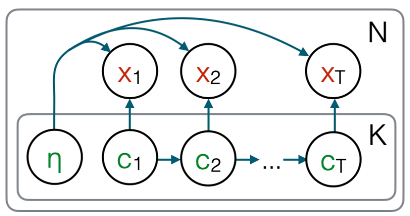
$$\max_{\theta} \ \log p_\theta(x)$$

objective: marginal likelihood

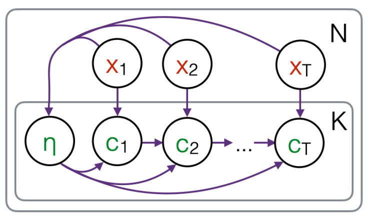
Inference Model

$$\min_{\phi} \ \mathrm{KL}\big(p_\theta(\eta, c \mid x) \,||\, q_\phi(\eta, c \mid x)\big)$$

objective: inclusive KL divergence



θ: network weights



ϕ: network weights

Self-normalized Gradient Estimates

$$\nabla_\theta \log p_\theta(x) = \mathbb{E}_{p_\theta(\eta, c \mid x)}\big[\nabla_\theta \log p_\theta(x, \eta, c)\big] \simeq \sum_l \frac{w^l}{\sum_{l'} w^{l'}} \nabla_\theta \log p_\theta(x, \eta^l, c^l)$$

$$-\nabla_\phi \mathrm{KL}\big(p_\theta(\eta, c \mid x) \,||\, q_\phi(\eta, c \mid x)\big) = \mathbb{E}_{p_\theta(\eta, c \mid x)}\big[\nabla_\phi \log q_\phi(\eta, c \mid x)\big] \simeq \sum_l \frac{w^l}{\sum_{l'} w^{l'}} \nabla_\phi \log q_\phi(\eta^l, c^l \mid x)$$

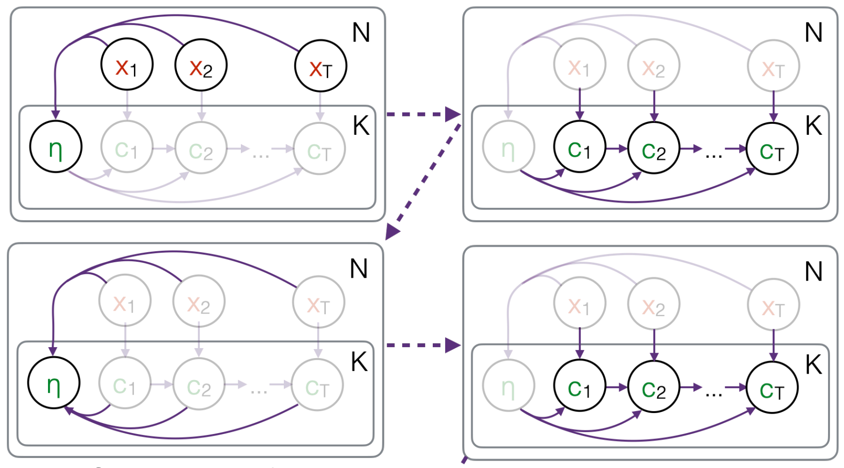
$$w^l = \frac{p_\theta(x, \eta^l, c^l)}{q_\phi(\eta^l, c^l \mid x)} \qquad \eta^l, c^l \sim q_\phi(\eta, c \mid x)$$

Amortized Population Gibbs Samplers

APG samplers iterate between conditional proposals to blocks of variables to construct high-quality

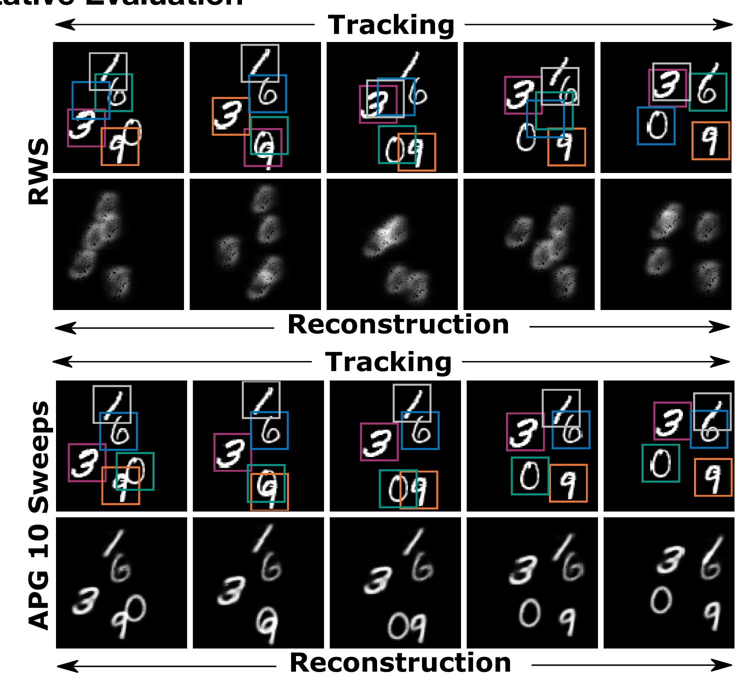
Hard: Guessing consistent features for K digits in T frames

Easier: Guessing positions of digits in each frame given digit features.



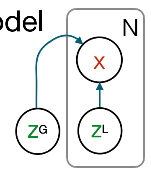
Easier: Guessing digit features given digits positions in each frame

Qualitative Evaluation



Neural Sufficient Statistics

Structured Model



$z = \{z^L, z^G\}$

Datapoint-level (local) — Instance-level (global)

Conditional Independencies $z^L_n, z^L_{-n} \mid x, z^G$

$$q_\phi(z^L \mid x, z^G) := p(z^L; \tilde{\lambda}^L) = p(z^L; \lambda^L + T^L_\phi(x_n, z^G))$$

Priors — Neural Sufficient Statistics

$$q_\phi(z^G \mid x, z^L) := p(z^G; \tilde{\lambda}^G) = p(z^G; \lambda^G + \sum_{n=1}^{N} T^G_\phi(x_n, z^L_n))$$

Example: Clustering Task