Stein Variational Message Passing for Continuous Graphical Models

Dilin Wang*+ Zhe Zeng*# Qiang Liu+
* (equal contribution)

+ Department of Computer Science, The University of Texas at Austin
# School of Mathematical Sciences, Zhejiang University

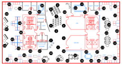
Continuous Probabilistic Graphical Models

1. Continuous probabilistic graphical models are powerful.

$$p(x) \propto \exp[\sum_{s \in \mathcal{S}} \psi(x_s)], \quad \mathcal{S} \text{ denotes the clique set.}$$

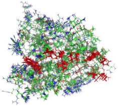
Crowdsourcing

Pose Estimation [Pacheco et al., 2014]



Sensor Localization

Optical Flow [Pacheco et al., 2015]



Proteins [Pacheco et al., 2015]

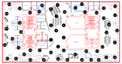
Continuous Probabilistic Graphical Models

1. Continuous probabilistic graphical models are powerful.

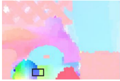
$$p(x) \propto \exp[\sum_{s \in \mathcal{S}} \psi(x_s)], \quad \mathcal{S} \text{ denotes the clique set.}$$

Crowdsourcing

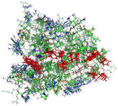
Pose Estimation [Pacheco et al., 2014]



Sensor Localization



Optical Flow [Pacheco et al., 2015]



Proteins [Pacheco et al., 2015]

Challenges for efficient inference

1. Standard belief propagation (BP) is best applicable only to discrete or Gaussian variable models.
2. Variants of particle message passing (PMP): 1) are sensitive to choices of re-sampling proposals; 2) don't use gradient information.

Recap: Approximate Inference

- given intractable $p(x)$
- find $q(x)$ in some family $\mathcal{Q}$ s.t. $q(x) \approx p(x)$
- and in inference time, approximate

$$\mathbb{E}_{p(x)}[f(x)] \approx \mathbb{E}_{q(x)}[f(x)]$$

Recap: Approximate Inference

- Monte Carlo sampling methods, e.g. Markov chain Monte Carlo (MCMC), gibbs sampling.
- Variational inference
  - pick a family of tractable distributions $\mathcal{Q}$
  - and then optimize a (usually parametric) q distribution in $\mathcal{Q}$ to approximate the exact posterior

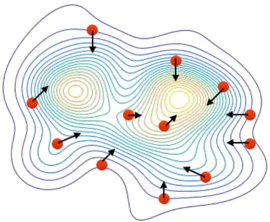
$$q^*(x) = \arg\min_{q \in \mathcal{Q}} KL(q||p)$$

Stein Variational Gradient Descent (SVGD) [Liu et al., 2016, a]

Idea: Iteratively move $\{x_i\}_{i=1}^n$ towards the target p by updates of form

$$x_i' \leftarrow x_i + \epsilon\phi(x_i),$$

where ϕ is a perturbation direction chosen to maximumly decrease the KL divergence with p, that is,



$$\phi^* = \arg\max_{\phi \in \mathcal{F}} \left\{ -\frac{\partial}{\partial \epsilon} \mathrm{KL}(q_{[\epsilon\phi]} \mid\mid p)\big|_{\epsilon=0} \right\}$$

where $q_{[\epsilon\phi]}$ is the density of $x' = x + \epsilon\phi(x)$. $\mathcal{F}$ is a function set that includes the possible velocity fields.

Stein Variational Gradient Descent (SVGD)

The optimization problem can be solved by the following basic observation shown in Liu & Wang [2016]:

- assume $x \sim q$ and $q_{[\epsilon\phi]}$ is the distribution of $x' = x + \epsilon\phi(x)$,
- then we have

$$KL(q_{[\epsilon\phi]} \parallel p) = KL(q \parallel p) - \epsilon\, \mathbb{E}_{x\sim q}[\mathcal{T}_x^\top \phi(x)] + O(\epsilon^2),$$

where $\mathcal{T}$ is a linear operator, called Stein operator, that acts on function ϕ via

$$\mathcal{T}_x^\top \phi(x) \quad \overset{\text{def}}{=} \quad \nabla_x \log p(x)^\top \phi(x) + \nabla_x^\top \phi(x).$$

Stein Variational Gradient Descent (SVGD)

1. Closed-form solution in RKHS [Liu et al., 2016, b],

$$\phi^*(\cdot) \propto \mathbb{E}_{x \sim q}[\mathcal{T}_x k(x, \cdot)].$$

Related, the optimal decreasing rate, which is called Stein discrepancy equals

$$\mathbb{D}(q||p) = \mathbb{E}_{x,x' \sim q}[\mathcal{T}_x^\top (\mathcal{T}_{x'} k(x, x'))].$$

2. iteratively update $\{x_i\}$ until convergence:

$$x_i \leftarrow x_i + \epsilon \cdot \frac{1}{n} \sum_{j=1}^{n} [\underbrace{\nabla_{x_j} \log p(x_j) k(x_j, x_i)}_{\text{gradient } G} + \underbrace{\nabla_{x_j} k(x_j, x_i)}_{\text{repulsive force } R}], \quad \forall i = 1 \cdots n$$

Applying SVGD to Graphical Models

SVGD updates $x_i \leftarrow x_i + \epsilon \cdot \frac{1}{n}\sum_{j=1}^{n}[\underbrace{\nabla_{x_j}\log p(x_j)k(x_j,x_i)}_{\text{gradient } G} + \underbrace{\nabla_{x_j}k(x_j,x_i)}_{\text{repulsive force } R}].$

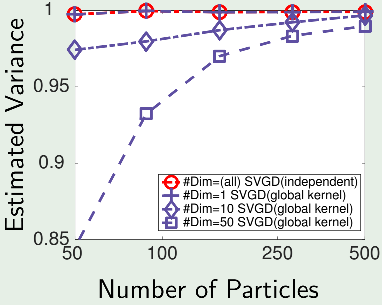
1. Problem 1: Kernel introduces global dependency; algorithms can be not distributed
2. Problem 2: The repulsive force is less effective with high dimensions

Example

1. $p(x)$ as the standard multivariate Gaussian distribution $\mathcal{N}(0, I_d)$,

$$p(x) = \prod_{i=1}^{d} p_i(x_i).$$

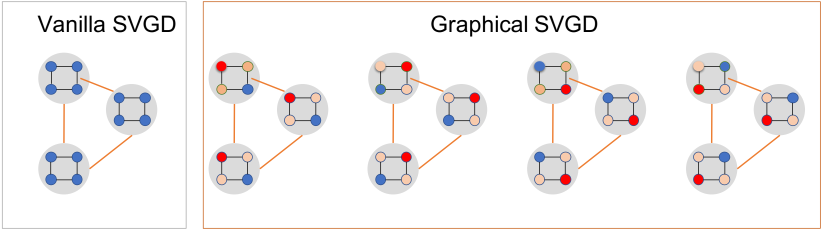
2. Taking $k(x,x') = \exp(-\frac{||x-x'||^2}{2h})$, where h is the bandwidth.



SVGD for Graphical Models

1. Goal: leverage Markov structures of probabilistic graphical models
2. Idea: construct local kernel function $k_i(x, x') = k_i(x_{\mathcal{C}_i}, x'_{\mathcal{C}_i})$ that depends only on the closed neighborhood $\mathcal{C}_i$ for each node i, where

$$\textit{Markov blanket} \;\; \mathcal{N}_i := \cup\{s\colon s \in \mathcal{S},\;\; s \ni i\} \setminus \{i\}, \;\; \mathcal{C}_i := \mathcal{N}_i \cup \{i\}.$$



Algorithm

Vanilla SVGD

$$x^{\ell,t+1} \leftarrow x^{\ell,t} + \epsilon \cdot \frac{1}{n} \sum_{\ell=1}^{n} \Big[\nabla_{x^\ell} \log p(x^\ell) k(x^\ell, x) + \nabla_{x^\ell} k(x^\ell, x) \Big].$$

Graphical Stein Variational Gradient Descent

for node i **do**

$$x_i^{\ell,t+1} \leftarrow x_i^{\ell,t} + \epsilon \cdot \frac{1}{n} \sum_{\ell=1}^{n} \Big[\nabla_{x_i^\ell} \log p(x^\ell) k_i(x_{\mathcal{C}_i}^\ell, x_{\mathcal{C}_i}) + \partial_{x_i^\ell} k_i(x_{\mathcal{C}_i}^\ell, x_{\mathcal{C}_i}) \Big].$$

end for

Theoritical Results

1. Stein discrepancy with each variable i equipped with a local kernel k_i,

$$\mathbb{D}(q||p)^2 = \sum_{i=1}^{d} \mathbb{E}_{x,x'\sim q}[\mathcal{T}_{x_i}^{\top}(\mathcal{T}_{x_i'} k_i(x,x'))].$$

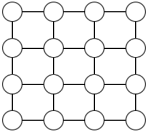
2. Similar to SVGD, we can show that as the particle size increases, the KL divrergence decreasing rate equals a generalized Stein discrepancy, in which each coordinate uses a separate kernel.

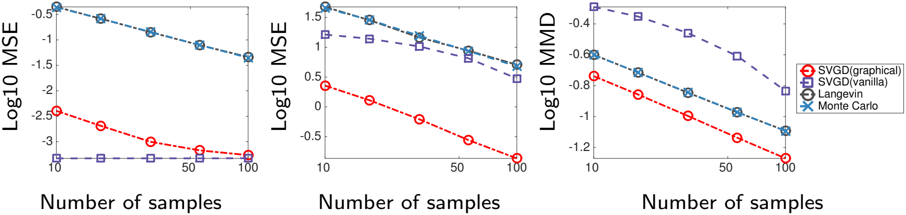
3. If all local kernels $k_i(x,x')$ are strictly integrally positive definite and under mild assumptions, we show

$$\mathbb{D}(q \mid\mid p) = 0 \quad \text{iff} \quad q(x_i|x_{\mathcal{N}_i}) = p(x_i|x_{\mathcal{N}_i}), \;\; \forall i \in [d].$$

Experiments: Gaussian MRFs

1. 4-neighborhood 2D grid of size 10×10

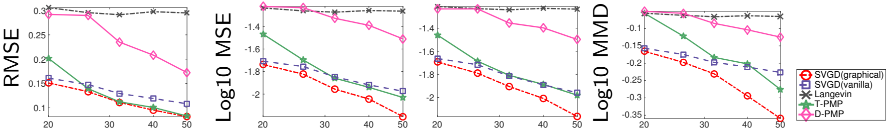




(a) Estimating $\mathbb{E}[x_i]$ (b) Estimating $\mathbb{E}[x_i^2]$ (c) MMD vs. n

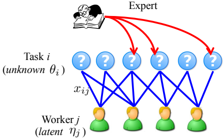
Experiments

Sensor Network Localization

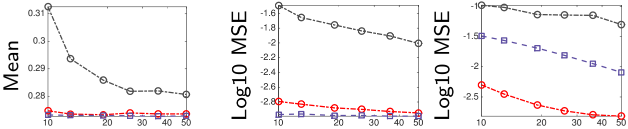


(a) Localization Error (b) $\mathbb{E}[x_i]$ (c) $\mathbb{E}[x_i^2]$ (d) MMD vs. n

Crowdsourcing: $x_{ij} \sim \mathcal{N}(\theta_i + b_j, v_j), \eta_j = [b_j, v_j]$



(a) MSE w.r.t. true labels (b) Mean (c) Variance



Thank You

References & Acknowledgment

[1] Qiang Liu and Dilin Wang. Stein variational gradient descent: a general purpose bayesian inference algorithm. NIPS. 2016.

[2] Qiang Liu and Jason Lee and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. ICML. 2016.

[3] Jason Pacheco and Erik Sudderth. Proteins, particles, and pseudo-max-marginals: a submodular approach. ICML. 2015.

[4] Jason Pacheco and Silvia Zuffi and Michael Black and Erik Sudderth. Preserving modes and messages via diverse particle selection. ICML. 2014.

This work is supported in part by NSF CRII 1565796.

Poster #61. 18:15 - 21:00 on 2018-07-12 in Hall B