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 are powerful. 

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

Continuous probabilistic graphical models are powerful.

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

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 $Q$ 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 $Q$
  - and then optimize a (usually parametric) $q$ distribution in $Q$ to approximate the exact posterior

$$q^*(x) = \arg \min_{q \in Q} KL(q || p)$$
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 \( \phi \) is a perturbation direction chosen to maximumly decrease the KL divergence with \( p \), that is,

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

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 $\phi$ via

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

   \[
   x_i \leftarrow x_i + \epsilon \cdot \frac{1}{n} \sum_{j=1}^{n} \left[ \nabla x_j \log p(x_j) k(x_j, x_i) + \nabla x_j k(x_j, x_i) \right], \quad \forall i = 1 \cdots n
   \]

   - gradient \( G \)
   - repulsive force \( R \)
Applying SVGD to Graphical Models

**SVGD updates**  
\[ x_i \leftarrow x_i + \epsilon \cdot \frac{1}{n} \sum_{j=1}^{n} \left[ \nabla_{x_j} \log p(x_j) k(x_j, x_i) + \nabla_{x_j} k(x_j, x_i) \right] \]

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 \( k(x, x') = \exp\left(-\frac{||x-x'||^2}{2h}\right) \), where \( h \) is the bandwidth.
**SVGD for Graphical Models**

1. **Goal**: leverage Markov structures of probabilistic graphical models

2. **Idea**: construct local kernel function $k_i(x, x') = k_i(x_{C_i}, x'_{C_i})$ that depends only on the closed neighborhood $C_i$ for each node $i$, where

   \[
   \text{Markov blanket } \mathcal{N}_i := \bigcup\{s: s \in S, s \ni i\} \setminus \{i\}, \quad C_i := \mathcal{N}_i \cup \{i\}.
   \]

---

**Vanilla SVGD**

**Graphical SVGD**
### Algorithm

**Vanilla SVGD**

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

**Graphical Stein Variational Gradient Descent**

```
for node i do

x_{i,t+1}^{\ell} \leftarrow x_{i,t}^{\ell} + \epsilon \cdot \frac{1}{n} \sum_{\ell=1}^{n} \left[ \nabla_{x_{i}^{\ell}} \log p(x^{\ell}) k_{i}(x_{C_{i}}^{\ell}, x_{C_{i}}) + \partial_{x_{i}^{\ell}} k_{i}(x_{C_{i}}^{\ell}, x_{C_{i}}) \right].

end for
```
Theoretical 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 divergence 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

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

- 4-neighborhood 2D grid of size $10 \times 10$

(a) Estimating $E[x_i]$

(b) Estimating $E[x_i^2]$

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

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

References & Acknowledgment


This work is supported in part by NSF CRII 1565796.

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