Message Passing Stein Variational Gradient Descent

Jingwei Zhuo*, Chang Liu, Jiaxin Shi, Jun Zhu, Ning Chen and Bo Zhang
Department of Computer Science and Technology, Tsinghua University.
* zjw15@mails.tsinghua.edu.cn

Motivations & Preliminaries

Variational inference: to approximate an intractable distribution $p(\mathbf{x})$ with $q(\mathbf{x})$ in some tractable family $\mathcal{Q}$ by

$$q(\mathbf{x}) = \operatorname*{argmin}_{q\in\mathcal{Q}} \mathrm{KL}(q(\mathbf{x})\|p(\mathbf{x})).$$

Stein Variational Gradient Descent (SVGD): Using a set of particles $\{\mathbf{x}^{(i)}\}_{i=1}^M$ (with the empirical distribution $\hat{q}_M(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^M \delta(\mathbf{x}-\mathbf{x}^{(i)})$) as approximation for $p(\mathbf{x})$, updated iteratively by

$$\mathbf{x}^{(i)} \leftarrow \mathbf{x}^{(i)} + \epsilon\hat{\boldsymbol{\phi}}(\mathbf{x}^{(i)}),$$

where

$$\hat{\boldsymbol{\phi}}(\mathbf{x}) = \mathbb{E}_{\mathbf{y}\sim\hat{q}_M}\Big[k(\mathbf{x},\mathbf{y})\nabla_{\mathbf{y}}\log p(\mathbf{y}) + \nabla_{\mathbf{y}}k(\mathbf{x},\mathbf{y})\Big].$$

$k(\mathbf{x},\mathbf{y})$ is a positive definite kernel, e.g., RBF kernel $k(\mathbf{x},\mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|_2^2/(2h))$.
– **Remark**: $M=1$, $\hat{\boldsymbol{\phi}}(\mathbf{x}) = \nabla_{\mathbf{x}}\log p(\mathbf{x})$, MAP.

$\hat{\boldsymbol{\phi}}$ is an unbiased estimate of $\boldsymbol{\phi}$, the steepest direction to reduce $\mathrm{KL}(q\|p)$ in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}^D$,

$$\boldsymbol{\phi}(\mathbf{x}) = \operatorname*{argmin}_{\|\boldsymbol{\phi}\|_{\mathcal{H}^D}\le 1} \nabla_\epsilon \mathrm{KL}(q_{[\mathbf{T}]}\|p)|_{\epsilon=0},$$

where $q_{[\mathbf{T}]}$ is the density of $T(\mathbf{x}) = \mathbf{x}+\epsilon\boldsymbol{\phi}(\mathbf{x})$ when the density of $\mathbf{x}$ is q.
– **Convergence Condition**: $\boldsymbol{\phi}(\mathbf{x}) \equiv \mathbf{0}$, which holds if and only if $q=p$ with a proper choice of $k(\mathbf{x},\mathbf{y})$.

$\boldsymbol{\phi}$ can be decomposed into two parts:
– **Kernel Smoothed Gradient (KSG)**:
$\mathbf{G}(\mathbf{x};p,q) = \mathbb{E}_{\mathbf{y}\sim q}[k(\mathbf{x},\mathbf{y})\nabla_{\mathbf{y}}\log p(\mathbf{y})]$.
– **Repulsive Force (RF)**:
$\mathbf{R}(\mathbf{x};q) = \mathbb{E}_{\mathbf{y}\sim q}[\nabla_{\mathbf{y}}k(\mathbf{x},\mathbf{y})]$.

Particle Degeneracy of SVGD

We observe particle degeneracy of SVGD, even for inferring $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0},\mathbf{I})$ with $M = 50, 100, 200$ particles.

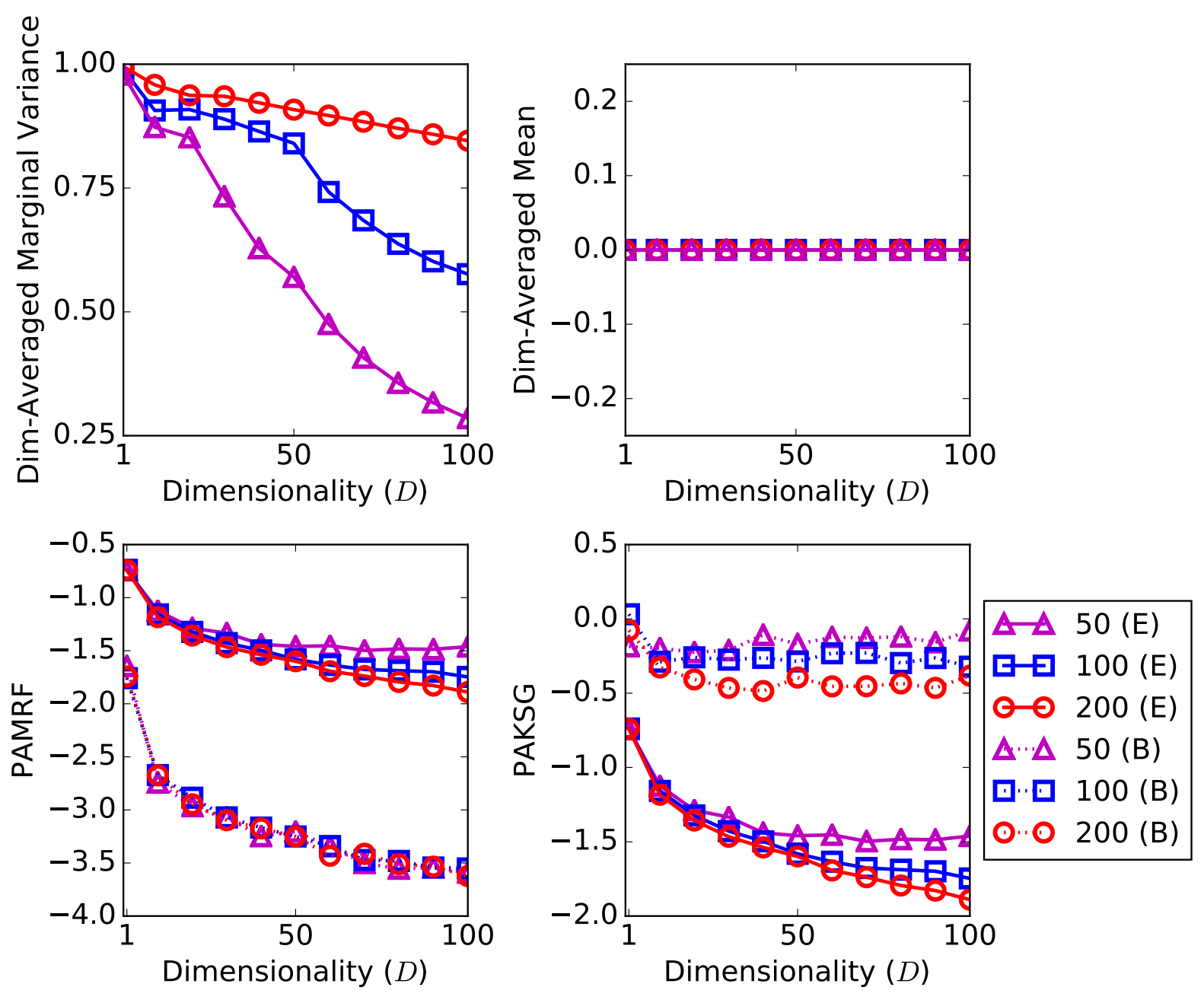


Figure 1: Top figures: Estimated variance and mean; Bottom figures: Magnitude of RF and KSG, at both the beginning (dotted;B) and the end of iterations (solid;E).

Explanations:
1. KSG alone corresponds to **mode-seeking**, i.e.,

$$\frac{\mathbf{G}(\mathbf{x};p,q)}{\|\mathbf{G}(\mathbf{x};p,q)\|_{\mathcal{H}^D}} = \operatorname*{argmax}_{\|\boldsymbol{\phi}\|_{\mathcal{H}^D}\le 1} \nabla_\epsilon \mathbb{E}_{\mathbf{z}\sim q_{[\mathbf{T}]}}[\log p(\mathbf{z})]|_{\epsilon=0},$$

is the steepest direction for maximizing $\mathbb{E}_{\mathbf{x}\sim q}[\log p(\mathbf{x})]$ (instead of $\mathrm{KL}(q\|p)$), which leads $q(\mathbf{x})$ to collapse to the modes of $p(\mathbf{x})$ in convergence.

2. RF is critical for SVGD to minimize $\mathrm{KL}(q\|p)$, but its magnitude $\|\mathbf{R}(\mathbf{x};q)\|_\infty$ may **correlate negatively** with dimensionality D. E.g., for the RBF kernel with any h,

$$\|\mathbf{R}(\mathbf{x};q)\|_\infty \le \mathbb{E}_{\mathbf{y}\sim q}\left[\frac{2}{e}\cdot\frac{\|\mathbf{x}-\mathbf{y}\|_\infty}{\|\mathbf{x}-\mathbf{y}\|_2^2}\right].$$

– When $\|\mathbf{x}-\mathbf{y}\|_\infty/\|\mathbf{x}-\mathbf{y}\|_2^2 \ll 1$ for most region of q, RF would be small.

3. In high-dimensional spaces, i.e., D is large and $\|\mathbf{R}(\mathbf{x};q)\|_\infty$ is small,
– this makes SVGD dynamics greatly dependent on $\mathbf{G}(\mathbf{x};p,q)$, especially at the beginning of iterations where q does not match p and $\|\mathbf{G}(\mathbf{x};p,q)\|_\infty$ is large,
– and this **weakens** the convergence conditions between $\mathbf{G}(\mathbf{x};p,q)\equiv\mathbf{0}$ (mode-seeking) and $\boldsymbol{\phi}(\mathbf{x})\equiv\mathbf{0}$ (distribution-matching).

Theoretical Analysis

So, for which q the RF $\mathbf{R}(\mathbf{x};q)$ suffers from such a negative correlation with the dimensionality D?

Theorem 1 (Gaussian) Given the RBF kernel $k(\mathbf{x},\mathbf{y})$ and $q(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma})$, the repulsive force satisfies

$$\|\mathbf{R}(\mathbf{x};q)\|_\infty \le \frac{\sqrt{D}}{\lambda_{\min}(\boldsymbol{\Sigma})(\frac{D}{2}+1)(1+\frac{2}{D})^{\frac{D}{2}}}\|\mathbf{x}-\boldsymbol{\mu}\|_\infty,$$

where $\lambda_{\min}(\boldsymbol{\Sigma})$ is the smallest eigenvalue of $\boldsymbol{\Sigma}$. By using $\lim_{x\to 0}(1+x)^{1/x} = e$, we have $\|\mathbf{R}(\mathbf{x};q)\|_\infty \lesssim \|\mathbf{x}-\boldsymbol{\mu}\|_\infty/(\lambda_{\min}(\boldsymbol{\Sigma})\sqrt{D})$.

Theorem 2 (Bounded) Let $k(\mathbf{x},\mathbf{y})$ be an RBF kernel. Suppose $q(\mathbf{y})$ is supported on a bounded set $\mathcal{X}$ which satisfies $\|\mathbf{y}\|_\infty \le C$ for $\mathbf{y}\in\mathcal{X}$, and $\mathrm{Var}(y_d|y_1,...,y_{d-1}) \ge C_0$ almost surely for any $1\le d\le D$. Let $\{\mathbf{x}^{(i)}\}_{i=1}^M$ be a set of samples of q and $\hat{q}_M$ the corresponding empirical distribution. Then, for any $\|\mathbf{x}\|_\infty \le C$, $\alpha,\delta\in(0,1)$, there exists $D_0>0$, such that for any $D>D_0$,

$$\|\mathbf{R}(\mathbf{x};\hat{q}_M)\|_\infty \le \frac{2}{eD^\alpha}$$

holds with at least probability $1-\delta$.

Message Passing SVGD

Key Idea: We decompose $\mathrm{KL}(q\|p)$ based on the **structural information** of $p(\mathbf{x}) \propto \prod_{F\in\mathcal{F}}\psi_F(\mathbf{x}_F)$, i.e.,

$$\mathrm{KL}(q(\mathbf{x}_{\neg d})||p(\mathbf{x}_{\neg d}))+\mathbb{E}_{q(\mathbf{x}_{\neg d})}\left[\mathrm{KL}(q(x_d|\mathbf{x}_{\neg d})||p(x_d|\mathbf{x}_{\Gamma_d}))\right],$$

where $\Gamma_d = \cup_{F\ni d}F$ is the Markov blanket (MB) of d such that $p(x_d|\mathbf{x}_{\neg d}) = p(x_d|\Gamma_d)$.

MP-SVGD: To apply SVGD with **local kernel** $k_d(\mathbf{x}_{S_d},\mathbf{y}_{S_d})$ ($S_d = \{d\}\cup\Gamma_d$) to optimize $q(x_d|\mathbf{x}_{\neg d})$ while keeping $q(\mathbf{x}_{\neg d})$ fixed, which results in,

Theorem 3 Let $\mathbf{z} = \mathbf{T}(\mathbf{x}) = [x_1,...,T_d(x_d),...,x_D]^\top$ with $T_d: x_d \to x_d + \epsilon\phi_d(\mathbf{x}_{S_d}), S_d = \{d\}\cup\Gamma_d$ where $\phi_d\in\mathcal{H}_d$ associated with the local kernel $k_d: \mathcal{X}_{S_d}\times\mathcal{X}_{S_d}\to\mathbb{R}$. Then, we have

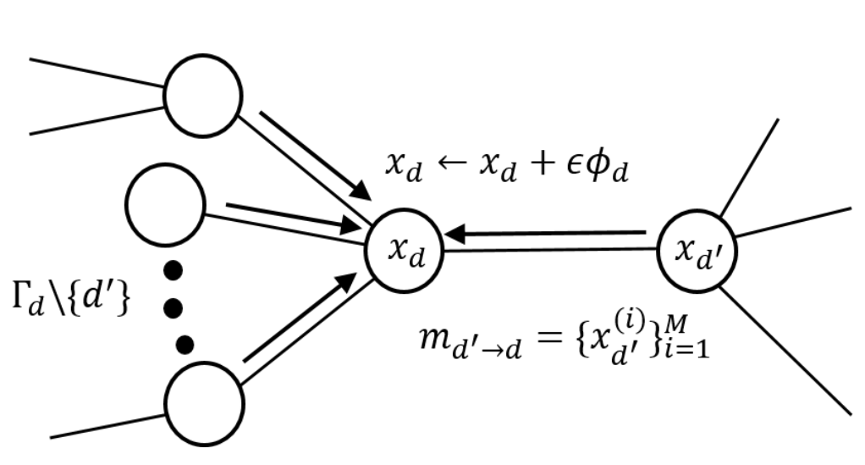
$$\nabla_\epsilon \mathrm{KL}(q_{[\mathbf{T}]}||p) = \nabla_\epsilon \mathbb{E}_{q(\mathbf{z}_{\Gamma_d})}\left[\mathrm{KL}\Big(q_{[T_d]}(z_d|\mathbf{z}_{\Gamma_d})\Big\|p(z_d|\mathbf{z}_{\Gamma_d})\Big)\right],$$

and $\phi_d(\mathbf{x}_{S_d}) = \mathrm{argmin}_{\|\phi_d\|_{\mathcal{H}_d}\le 1}\nabla_\epsilon \mathrm{KL}(q_{[\mathbf{T}]}|||p)|_{\epsilon=0} =$

$$\mathbb{E}_{\mathbf{y}_{S_d}\sim q}\left[k_d(\mathbf{x}_{S_d},\mathbf{y}_{S_d})\nabla_{y_d}\log p(y_d|\mathbf{y}_{\Gamma_d}) + \nabla_{y_d}k_d(\mathbf{x}_{S_d},\mathbf{y}_{S_d})\right].$$

– **Convergence Condition**: $\phi_d(\mathbf{x}_{S_d})\equiv 0$, $\forall d$, which holds if and only if $q(x_d|\mathbf{x}_{\Gamma_d}) = p(x_d|\mathbf{x}_{\Gamma_d})$ with a proper choice of $k_d(\mathbf{x}_{S_d},\mathbf{y}_{S_d})$.

Final Algorithms:
To approximate $p(\mathbf{x}) \propto \psi_F(\mathbf{x}_F)$ with a set of particles $\{\mathbf{x}^{(i)}\}_{i=1}^M$ (with the empirical distribution $\hat{q}_M(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^M\delta(\mathbf{x}-\mathbf{x}^{(i)})$), updated iteratively by


$x_d \leftarrow x_d + \epsilon\phi_d$
x_d
$x_{d'}$
$\Gamma_d\backslash\{d'\}$
$m_{d'\to d} = \{x_{d'}^{(i)}\}_{i=1}^M$

$$\mathbf{x}_d^{(i)} \leftarrow \mathbf{x}_d^{(i)} + \epsilon\hat{\phi}_d(\mathbf{x}_{S_d}^{(i)}),$$

where $\hat{\phi}_d(\mathbf{x}_{S_d}) =$

$$\mathbb{E}_{\mathbf{y}_{S_d}\sim\hat{q}_M}\left[k_d(\mathbf{x}_{S_d},\mathbf{y}_{S_d})\nabla_{y_d}\log p(y_d|\mathbf{y}_{\Gamma_d})+\nabla_{y_d}k(\mathbf{x}_{S_d},\mathbf{y}_{S_d})\right].$$

Experiments: Synthetic Results

Targets: $p(\mathbf{x}) \propto \prod_{d\in V}\psi_d(x_d)\prod_{(d,t)\in E}\psi_{dt}(x_d,x_t)$, where
$\psi_d(x_d) = \alpha_1\mathcal{N}(x_d-y_d|-2,1)+\alpha_2\mathcal{G}(x_d-y_d|2,1.3)$,
$\psi_{dt}(x_d,x_t) = \mathcal{L}(x_d-x_t|0,2)$,
with a 10 × 10 grid.
Methods:
We compare SVGD, MP-SVGD with EP, HMC (slow, asymptotically exact algorithm) and EPBP (original state-of-the-art method on this example).
Ground Truth:
4 million HMC samples.

Experiments: Synthetic Results

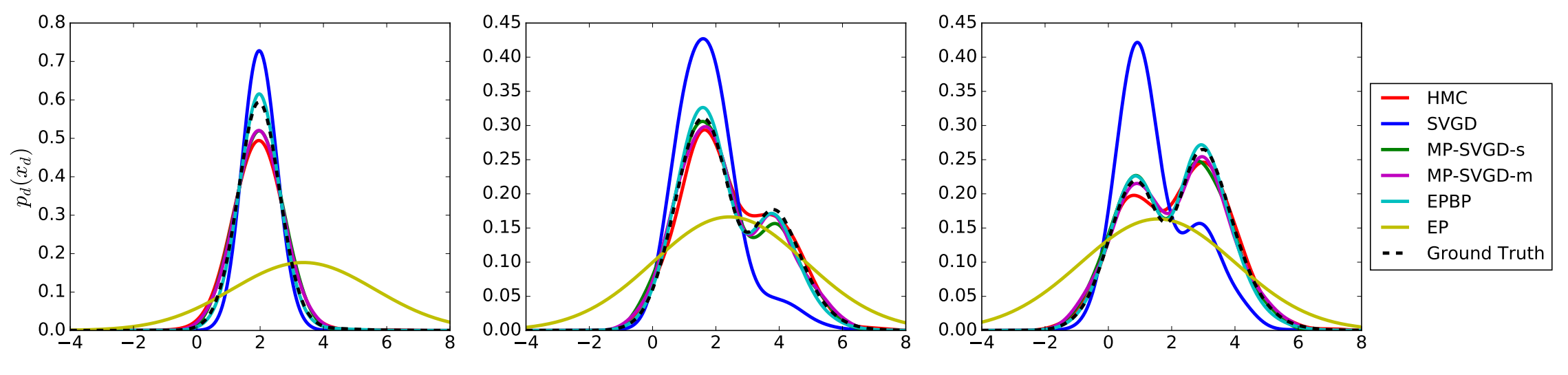


Figure 2: A qualitative comparison of inference methods with 100 particles (except EP) for estimating marginal densities of three randomly selected nodes.

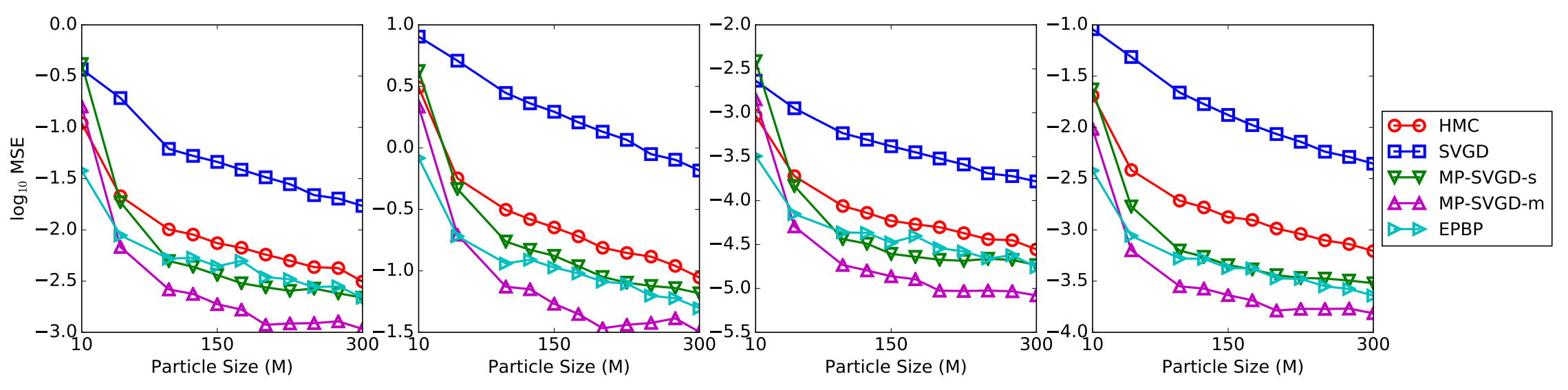


Figure 3: A quantitative comparison of inference methods with varying number of particles. Performance is measured by the MSE of the estimation of expectation $\mathbb{E}_{\mathbf{x}\sim\hat{q}_M}[\mathbf{f}(\mathbf{x})]$ for test functions $\mathbf{f}(\mathbf{x}) = \mathbf{x}$, $\mathbf{x}^2$, $1/(1+\exp(\boldsymbol{\omega}\circ\mathbf{x}+\mathbf{b}))$ and $\cos(\boldsymbol{\omega}\circ\mathbf{x}+\mathbf{b})$, arranged from left to right, where $\circ$ denotes the element-wise product and $\boldsymbol{\omega},\mathbf{b}\in\mathbb{R}^{100}$ with $\omega_d\sim\mathcal{N}(0,1)$ and $b_d\in\mathrm{Uniform}[0,2\pi]$, $\forall d\in\{1,...,100\}$.

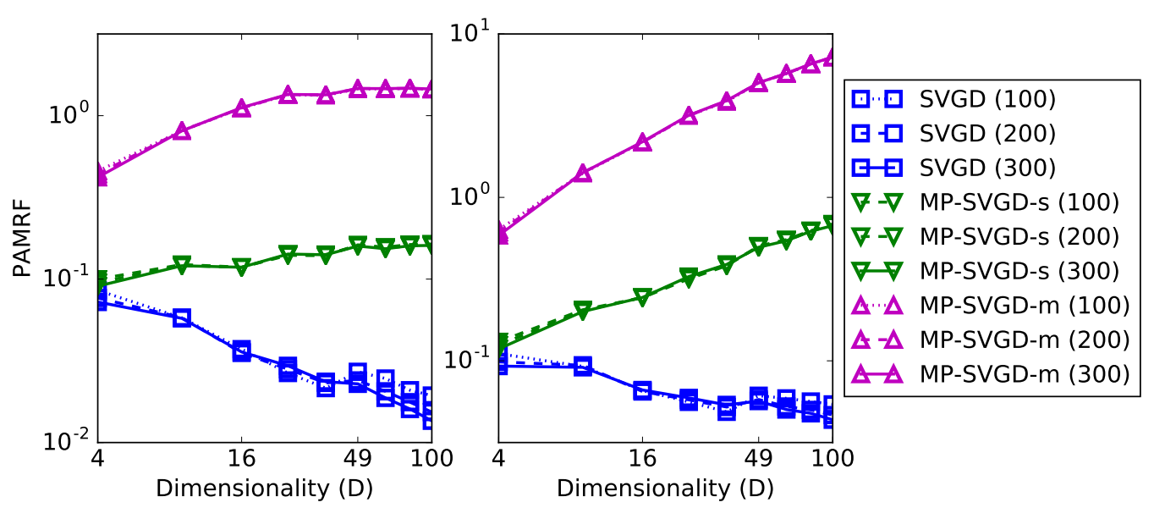


Figure 4: PAMRF $\frac{1}{M}\sum_{i=1}^M\|\mathbf{R}(\mathbf{x}^{(i)};\hat{q}_M)\|_r$ for converged $\{\mathbf{x}^{(i)}\}_{i=1}^M$ with $r=\infty$ (left) and $r=2$ (right). The grid size ranges from 2 × 2 to 10 × 10 and the number of particles is denoted in the bracket.

Experiments: Image Denoising

Targets:$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$, where
$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{x},\sigma_n^2\mathbf{I})$: Noise distribution.
$p(\mathbf{x}) \propto \exp(-\frac{\epsilon||\mathbf{x}||_2^2}{2})\prod_{C\in\mathcal{C}}\prod_{i=1}^N\phi(\mathbf{J}_i^{\mathrm{T}}\mathbf{x}_C;\boldsymbol{\alpha}_i)$: Fields of Experts (FOE) model, where $\phi(\mathbf{J}_i^{\mathrm{T}}\mathbf{x}_C;\boldsymbol{\alpha}_i) = \sum_{j=1}^J\alpha_{ij}\mathcal{N}(\mathbf{J}_i^{\mathrm{T}}\mathbf{x}_C|0,\sigma_i^2/s_j)$.
Methods:
We compare SVGD, MP-SVGD and Gibbs sampling with auxiliary variables (original inference method).
Evaluation:
peak signal-to-noise ratio (**PSNR**) and structural similarity index (**SSIM**).

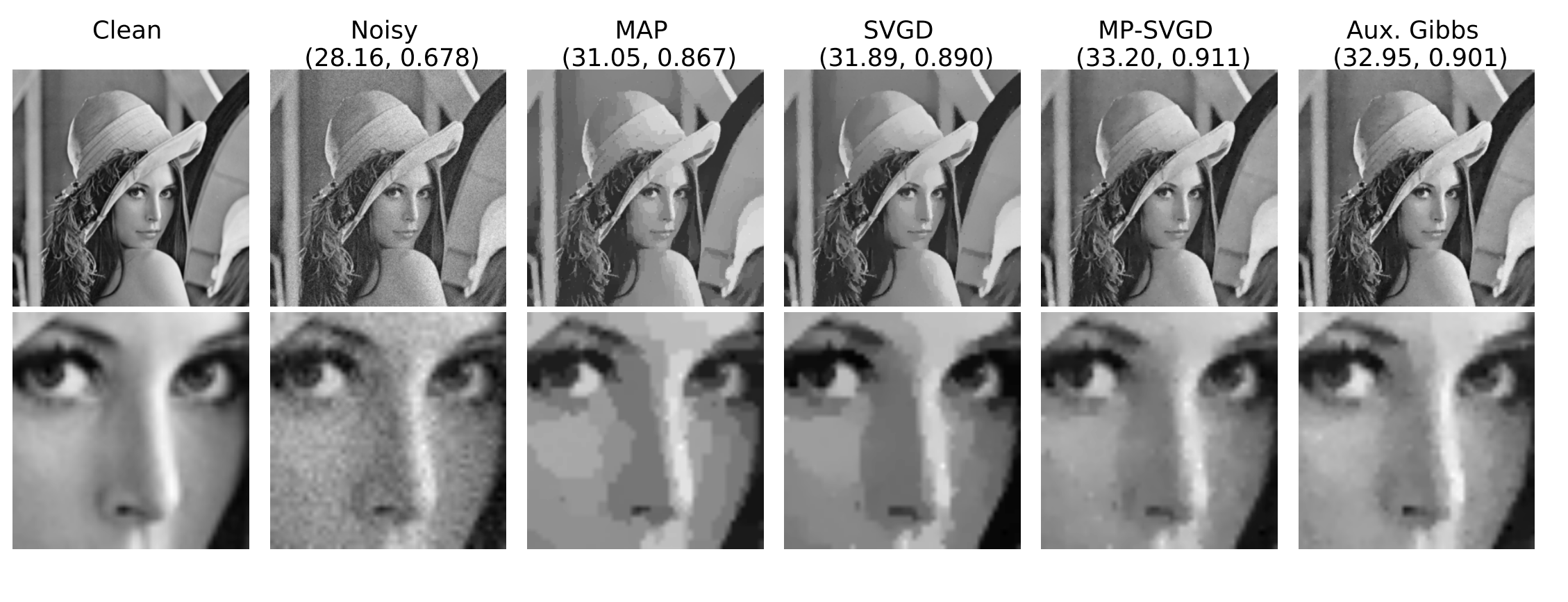


Figure 5: Denoising results for *Lena* using 50 particles, 256 × 256 pixels, $\sigma_n = 10$. The number in bracket is PSNR and SSIM. Upper Row: The full size image; Bottom Row: The 50 × 50 patches.

Inference	avg. PSNR		avg. SSIM	
	$\sigma_n=10$	$\sigma_n=20$	$\sigma_n=10$	$\sigma_n=20$
MAP	30.27	26.48	0.855	0.720
Aux. Gibbs	32.09	**28.32**	0.904	0.808
Aux. Gibbs ($M=50$)	31.87	28.05	0.898	0.795
Aux. Gibbs ($M=100$)	31.98	28.17	0.901	0.801
SVGD ($M=50$)	31.58	27.86	0.894	0.766
SVGD ($M=100$)	31.65	27.90	0.896	0.767
MP-SVGD ($M=50$)	32.09	28.21	0.905	0.808
MP-SVGD ($M=100$)	**32.12**	28.27	**0.906**	**0.809**

Table 1: Denoising results for 10 test images from BSD dataset. The first two rows are cited from the original paper while the other rows are based on our implementation.

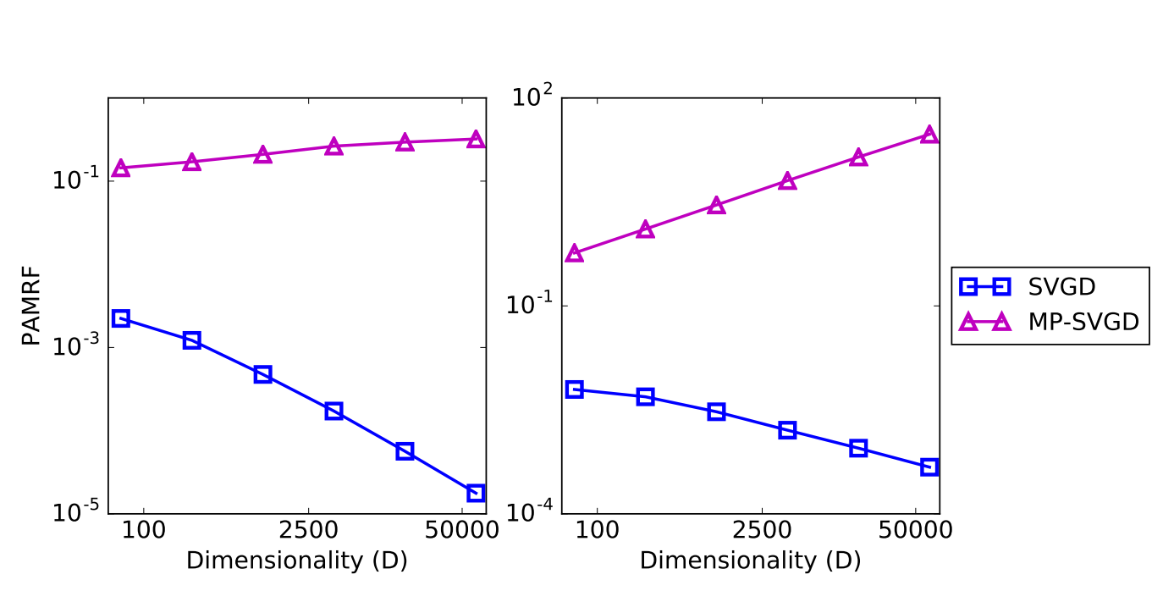


Figure 6: PAMRF $\frac{1}{M}\sum_{i=1}^M\|\mathbf{R}(\mathbf{x}^{(i)};\hat{q}_M)\|_r$ for $M=50$ converged $\{\mathbf{x}^{(i)}\}_{i=1}^M$ with $r=\infty$ (left) and $r=2$ (right) over rescaled *Lena* ranged from 8 × 8 to 256 × 256.