Appendix for Message Passing Stein Variational Gradient Descent

A. Detailed Derivation and Proof for Section 3

A.1. Derivation of Eq. (4)

By using the change of variable theorem, we have

$$\begin{aligned}
&\nabla_\epsilon \mathbb{E}_{\mathbf{z}\sim q_{[\boldsymbol{T}]}}[\log p(\mathbf{z})]|_{\epsilon=0} \\
&= \nabla_\epsilon \mathbb{E}_{\mathbf{x}\sim q}[\log p(\mathbf{x}+\epsilon\boldsymbol{\phi}(\mathbf{x}))]|_{\epsilon=0} \\
&= \mathbb{E}_{\mathbf{x}\sim q}[\nabla_\epsilon \log p(\mathbf{x}+\epsilon\boldsymbol{\phi}(\mathbf{x}))|_{\epsilon=0}] \\
&= \mathbb{E}_{\mathbf{x}\sim q}[\nabla_{\mathbf{x}} \log p(\mathbf{x})^\top \boldsymbol{\phi}(\mathbf{x})] \\
&= \mathbb{E}_{\mathbf{x}\sim q}\left[\sum_{d=1}^{D} \nabla_{x_d} \log p(\mathbf{x}) \phi_d(\mathbf{x})\right] \\
&= \sum_{d=1}^{D} \mathbb{E}_{\mathbf{x}\sim q}\left[\nabla_{x_d} \log p(\mathbf{x}) \langle k(\mathbf{x},\cdot), \phi_d(\cdot)\rangle_{\mathcal{H}_0}\right] \\
&= \sum_{d=1}^{D} \langle \mathbb{E}_{\mathbf{x}\sim q}\left[\nabla_{x_d} \log p(\mathbf{x}) k(\mathbf{x},\cdot)\right], \phi_d(\cdot)\rangle_{\mathcal{H}_0}.
\end{aligned}$$

The maximum is attained when $\boldsymbol{\phi} = \boldsymbol{\phi}^*/\|\boldsymbol{\phi}^*\|_{\mathcal{H}^D}$ with $\boldsymbol{\phi}^*(\cdot) = \mathbb{E}_{\mathbf{x}\sim q}[k(\mathbf{x},\cdot)\nabla_{\mathbf{x}} \log p(\mathbf{x})]$, i.e., the kernel smoothed gradient $G(\mathbf{x}; p, q)$. This relationship holds for both q and the empirical distribution $\hat{q}_M$.

When converged, $\mathbf{G}(\mathbf{x}; p, q) \equiv \mathbf{0}$, which corresponds to

$$\begin{aligned}
&\int_{\mathcal{X}} k(\mathbf{x},\mathbf{y})\mathbf{g}(\mathbf{y})d\mathbf{y} = \mathbf{0},\ \forall \mathbf{x} \in \mathcal{X} \\
\Longrightarrow &\int_{\mathcal{X}} k(\mathbf{x},\mathbf{y})g_d(\mathbf{y})d\mathbf{y} = 0,\ \forall \mathbf{x} \in \mathcal{X},\ d \in \{1, ..., D\} \\
\Longrightarrow &\int_{\mathcal{X}} k(\mathbf{x},\mathbf{y})g_d(\mathbf{x})g_d(\mathbf{y})d\mathbf{y} = 0,\ \forall \mathbf{x} \in \mathcal{X},\ d \in \{1, ..., D\} \\
\Longrightarrow &\int_{\mathcal{X}\times\mathcal{X}} k(\mathbf{x},\mathbf{y})g_d(\mathbf{x})g_d(\mathbf{y})d\mathbf{x}d\mathbf{y} = 0,\ \forall d \in \{1, ..., D\}
\end{aligned}$$

where $\mathbf{g}(\mathbf{y}) = q(\mathbf{y})\nabla_{\mathbf{y}} \log p(\mathbf{y})$ and $g_d(\mathbf{y}) = q(\mathbf{y})\nabla_{y_d} \log p(\mathbf{y})$. Given $k(\mathbf{x},\mathbf{y})$ is strictly positive definite, i.e., $\int_{\mathcal{X}\times\mathcal{X}} k(\mathbf{x},\mathbf{y})f(\mathbf{x})f(\mathbf{y})d\mathbf{x}d\mathbf{y} = 0$ if and only if $f(\mathbf{y}) = 0$, $\forall \mathbf{y} \in \mathcal{X}$, we have $g_d(\mathbf{y}) \equiv 0$, $\forall d$, which corresponds to

$$\mathbf{g}(\mathbf{y}) = q(\mathbf{y})\nabla_{\mathbf{y}} \log p(\mathbf{y}) = 0,\ \forall \mathbf{y} \in \mathcal{X}.$$

In other words, for $\mathbf{y}$ such that $q(\mathbf{y}) \neq 0$, $\nabla_{\mathbf{y}} \log q(\mathbf{y}) = 0$, which reflects that q collapses to the modes of p.

A.2. Derivation of Eq. (5)

RBF Kernel Notice that

$$\|\mathbf{R}(\mathbf{x}; q)\|_\infty \leq \mathbb{E}_{y\sim q}\left[\exp(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2h})\frac{\|\mathbf{x}-\mathbf{y}\|_\infty}{h}\right],$$

where the inequality holds according to Jensen's inequality. For notation simplicity, let $f(h, \mathbf{x}, \mathbf{y}) = \exp(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2h})\frac{\|\mathbf{x}-\mathbf{y}\|_\infty}{h}$, we have

$$\begin{aligned}
\|\mathbf{R}(\mathbf{x}; q)\|_\infty &\leq \mathbb{E}_{y\sim q}\left[f(h, \mathbf{x}, \mathbf{y})\right] \\
&\leq \max_h \mathbb{E}_{y\sim q}\left[f(h, \mathbf{x}, \mathbf{y})\right] \\
&\leq \mathbb{E}_{y\sim q}\left[\max_h f(h, \mathbf{x}, \mathbf{y})\right].
\end{aligned}$$

By taking gradient of $f(h, \mathbf{x}, \mathbf{y})$ over h we can show that when $h = \|\mathbf{x}-\mathbf{y}\|_2^2/2$, $f(h, \mathbf{x}, \mathbf{y})$ attains its maximum, which is $f_{\max}(\mathbf{x}, \mathbf{y}) = \max_h f(h, \mathbf{x}, \mathbf{y}) = 2e^{-1}\frac{\|\mathbf{x}-\mathbf{y}\|_\infty}{\|\mathbf{x}-\mathbf{y}\|_2^2}$. And we have

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

In fact, we can bound $\|\mathbf{R}(\mathbf{x}; q)\|_r$ with any $r \geq 1$ by using the norm inequality $\|\mathbf{z}\|_r \leq D^{1/r}\|\mathbf{z}\|_\infty$, i.e.,

$$\|\mathbf{R}(\mathbf{x}, q)\|_r \leq \mathbb{E}_{y\sim q}\left[\frac{2D^{1/r}}{e}\cdot\frac{\|\mathbf{x}-\mathbf{y}\|_\infty}{\|\mathbf{x}-\mathbf{y}\|_2^2}\right].$$

IMQ Kernel For the IMQ kernel, we have

$$\|\mathbf{R}(\mathbf{x}; q)\|_\infty \leq \mathbb{E}_{y\sim q}\left[\frac{1}{2\left(1+\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2h}\right)^{3/2}}\frac{\|\mathbf{x}-\mathbf{y}\|_\infty}{h}\right].$$

Let $f(h, \mathbf{x}, \mathbf{y}) = \frac{1}{2\left(1+\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2h}\right)^{3/2}}\frac{\|\mathbf{x}-\mathbf{y}\|_\infty}{h}$ and take the maximum over h, we have $h = \frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{4}$, and corresponding

$$f_{\max}(h, \mathbf{x}, \mathbf{y}) = \frac{2}{3^{3/2}}\frac{\|\mathbf{x}-\mathbf{y}\|_\infty}{\|\mathbf{x}-\mathbf{y}\|_2^2},$$

where the only difference compared to the RBF kernel is the constant $\frac{2}{3^{3/2}}$.

A.3. Derivation of Proposition 1

Here we derive the kernel smoothed gradient $\mathbf{G}(\mathbf{x}; p, q)$ the repulsive force $\mathbf{R}(\mathbf{x}; q)$ when $q(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a Gaussian distribution. This example will be useful for understanding the relationship between SVGD and dimensionality, and will be helpful for understanding the convergence condition when p is also a Gaussian.

Since $q(\mathbf{y})$ is Gaussian and $k(\mathbf{x}, \mathbf{y})$ is the RBF kernel, $q(\mathbf{y})k(\mathbf{x}, \mathbf{y})$ can be regarded as a rescaled Gaussian distribution over $\mathbf{y}$, i.e., $q(\mathbf{y})k(\mathbf{x}, \mathbf{y}) =$

$$\frac{\sqrt{\det \boldsymbol{\Sigma}^{-1}}}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) - \frac{||\mathbf{x}-\mathbf{y}||_2^2}{2h}\right)$$

$$= \frac{\sqrt{\det \boldsymbol{\Sigma}^{-1}}}{\sqrt{2\pi}} \exp\left(-\frac{||\mathbf{x}||_2^2}{2h} - \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$$

$$\cdot \exp\left(-\frac{1}{2}\mathbf{y}^\top (\boldsymbol{\Sigma}^{-1} + \frac{1}{h}\mathbf{I})\mathbf{y} + (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{\mathbf{x}}{h})^\top \mathbf{y}\right)$$

$$= \frac{\sqrt{\det \boldsymbol{\Sigma}^{-1}}}{\sqrt{\det(\boldsymbol{\Sigma}^{-1} + \frac{1}{h}\mathbf{I})}} \mathcal{N}(\mathbf{y}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$$

$$\cdot \exp\left(\frac{1}{2}\tilde{\boldsymbol{\mu}}^\top \tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}} - \frac{||\mathbf{x}||_2^2}{2h} - \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$$

$$= \frac{\sqrt{\det \boldsymbol{\Sigma}^{-1}}}{\sqrt{\det(\boldsymbol{\Sigma}^{-1} + \frac{1}{h}\mathbf{I})}} \mathcal{N}(\mathbf{y}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$$

$$\cdot \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top (\boldsymbol{\Sigma} + h\mathbf{I})^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

$$= \frac{\sqrt{\det \boldsymbol{\Sigma}^{-1}}}{\sqrt{\det(\boldsymbol{\Sigma}^{-1} + \frac{1}{h}\mathbf{I})}} \exp\left(-\frac{1}{2}d(\mathbf{x}, \boldsymbol{\mu})\right) \mathcal{N}(\mathbf{y}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$$

with $\tilde{\boldsymbol{\mu}} = (\boldsymbol{\Sigma}^{-1} + \frac{1}{h}\mathbf{I})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{h}\mathbf{x})$, $\tilde{\boldsymbol{\Sigma}} = (\boldsymbol{\Sigma}^{-1} + \frac{1}{h}\mathbf{I})^{-1}$ and $d(\mathbf{x}, \boldsymbol{\mu}) = (\mathbf{x}-\boldsymbol{\mu})^\top (\boldsymbol{\Sigma} + h\mathbf{I})^{-1}(\mathbf{x}-\boldsymbol{\mu})$.

Given $q(\mathbf{y})k(\mathbf{x}, \mathbf{y})$, the repulsive force can be computed, i.e.,

$$\mathbf{R}(\mathbf{x}; q) = \frac{\sqrt{\det \boldsymbol{\Sigma}^{-1}}}{\sqrt{\det(\boldsymbol{\Sigma}^{-1} + \frac{1}{h}\mathbf{I})}} \exp\left(-\frac{1}{2}d(\mathbf{x}, \boldsymbol{\mu})\right) \cdot \frac{\mathbf{x} - \tilde{\boldsymbol{\mu}}}{h}$$
$$= \frac{h^{D/2} \exp\left(-\frac{1}{2}d(\mathbf{x}, \boldsymbol{\mu})\right)}{\sqrt{\det(\boldsymbol{\Sigma} + h\mathbf{I})}} (\boldsymbol{\Sigma} + h\mathbf{I})^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

and thus we have

$$\|\mathbf{R}(\mathbf{x}; q)\|_2 \leq \frac{h^{D/2}}{\sqrt{\det(\boldsymbol{\Sigma} + h\mathbf{I})}} \|(\boldsymbol{\Sigma} + h\mathbf{I})^{-1}(\mathbf{x} - \boldsymbol{\mu})\|_2$$

by using the fact that $\exp\left(-\frac{1}{2}d(\mathbf{x}, \boldsymbol{\mu})\right) \leq 1$. Then, Assume the eigenvalue decomposition for $\boldsymbol{\Sigma}$ is $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ with $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_D)$ where $\lambda_1 \geq \cdots \geq \lambda_D \geq C$, we have

$$\det(\boldsymbol{\Sigma} + h\mathbf{I}) = \det(\boldsymbol{\Lambda} + h\mathbf{I}) = \prod_{d=1}^{D} (\lambda_d + h)\,.$$

and $\|(\boldsymbol{\Sigma} + h\mathbf{I})^{-1}(\mathbf{x} - \boldsymbol{\mu})\|_2 \leq \frac{1}{h+\lambda_D}\|\mathbf{x} - \boldsymbol{\mu}\|_2$.

So, we have

$$\|\mathbf{R}(\mathbf{x}; q)\|_2 \leq \frac{(1 + \lambda_D/h)^{-1}}{h\sqrt{\prod_{d=1}^{D}(1 + \lambda_d/h)}} \|\mathbf{x} - \boldsymbol{\mu}\|_2$$
$$\leq \frac{1}{h(1 + \lambda_D/h)^{D/2+1}} \|\mathbf{x} - \boldsymbol{\mu}\|_2$$

Let $f(h) = h(1 + \lambda_D/h)^{D/2+1}$, and it is easy to show that when $h = D\lambda_D/2$, $f(h)$ attains its minimum, which is $f_{\min} = \frac{D\lambda_D}{2}(1 + 2/D)^{D/2+1} = (1 + D/2)\lambda_D(1 + 2/D)^{D/2}$. Let $\lambda_{\min}(\boldsymbol{\Sigma}) = \lambda_D$ denote the smallest eigenvalue, we have

$$\|\mathbf{R}(\mathbf{x}; q)\|_2 \leq \frac{1}{(1 + D/2)\lambda_{\min}(\boldsymbol{\Sigma})(1 + 2/D)^{D/2}} \|\mathbf{x} - \boldsymbol{\mu}\|_2.$$

By using the norm inequality that $\|\mathbf{z}\|_\infty \leq \|\mathbf{z}\|_2 \leq D^{1/2}\|\mathbf{z}\|_\infty$, we have

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

This can be further simplified by noting that $\lim_{D\to\infty}(1 + 2/D)^{D/2} = e$, so for large D, we have the following inequality

$$\|\mathbf{R}(\mathbf{x}; q)\|_2 \leq \frac{1}{(1 + D/2)\lambda_D(1 + 2/D)^{D/2}} \|\mathbf{x} - \boldsymbol{\mu}\|_2$$
$$\lesssim \frac{1}{(1 + D/2)\lambda_{\min}(\boldsymbol{\Sigma})e} \|\mathbf{x} - \boldsymbol{\mu}\|_2$$
$$\lesssim \frac{1}{D\lambda_{\min}(\boldsymbol{\Sigma})} \|\mathbf{x} - \boldsymbol{\mu}\|_2$$

and corresponding

$$\|\mathbf{R}(\mathbf{x}; q)\|_\infty \lesssim \frac{1}{\sqrt{D}\lambda_{\min}(\boldsymbol{\Sigma})} \|\mathbf{x} - \boldsymbol{\mu}\|_\infty.$$

When q is Gaussian whose smallest eigenvalue of $\boldsymbol{\Sigma}$ is greater than some constant C, corresponding $\mathbf{R}(\mathbf{x}; q)$ decreases to zero vector as $1/D$ in $\|\cdot\|_2$ or as $1/\sqrt{D}$ in $\|\cdot\|_\infty$.

A.4. Proof of Proposition 2

The inequality can be decomposed as:

$$
\begin{aligned}
& P\left(\frac{\|\mathbf{y}-\mathbf{x}\|_\infty}{\|\mathbf{y}-\mathbf{x}\|_2^2} \geq \frac{1}{D^\alpha}\right) \\
&= P\left(\frac{\max_d |y_d - x_d|}{\|\mathbf{y}-\mathbf{x}\|_2^2} \geq \frac{1}{D^\alpha}\right) \\
&= P\left(\frac{\max_d |y_d - x_d|}{\|\mathbf{y}-\mathbf{x}\|_2^2} \geq \frac{1}{D^\alpha}, \|\mathbf{y}-\mathbf{x}\|_2^2 \leq b\right) \\
&\quad + P\left(\frac{\max_d |y_d - x_d|}{\|\mathbf{y}-\mathbf{x}\|_2^2} \geq \frac{1}{D^\alpha}, \|\mathbf{y}-\mathbf{x}\|_2^2 > b\right) \\
&\leq P(\|\mathbf{y}-\mathbf{x}\|_2^2 \leq b) + P\left(\max_d |y_d - x_d| \geq \frac{b}{D^\alpha}\right) \\
&\leq P(\|\mathbf{y}-\mathbf{x}\|_2^2 \leq b) + \sum_{d=1}^D P\left(|y_d - x_d| \geq \frac{b}{D^\alpha}\right) \\
&= P(\sum_{d=1}^D (y_d - x_d)^2 \leq b) + \sum_{d=1}^D P\left(|y_d - x_d| \geq \frac{b}{D^\alpha}\right)
\end{aligned}
$$

holds for any b, even when b is a function of y, i.e. a random variable.
We bound the first term by using the Azuma-Hoeffding inequality(Azuma, 1967):

Theorem 1. *Suppose $Z_D, D \geq 1$ is a martingale such that $Z_0 = 0$ and $|Z_d - Z_{d-1}| \leq c_d, 1 \leq d \leq D$ almost surely for some constants $c_d, 1 \leq d \leq D$. Then, for every $t > 0$,*

$$
P(Z_D > t) \leq \exp\left(-\frac{t^2}{2\sum_{d=1}^D c_d^2}\right),
$$

and

$$
P(Z_D < -t) \leq \exp\left(-\frac{t^2}{2\sum_{d=1}^D c_d^2}\right).
$$

To use it, we construct that $Z_D = \sum_{d=1}^D (y_d - x_d)^2 - \sum_{d=1}^D \mathbb{E}[(y_d - x_d)^2 | y_{1:d-1}], \forall D \geq 0$ and $Z_0 = 0$ is a martingale.

First we notice that $Z_D - Z_{D-1} = (y_D - x_D)^2 - \mathbb{E}[(y_D - x_D)^2 | y_{1:D-1}]$, which satisfies $\mathbb{E}[Z_D | Z_{1:D-1}] - Z_{D-1} = \mathbb{E}[(y_D - x_D)^2 | y_{1:D-1}] - \mathbb{E}[(y_D - x_D)^2 | y_{1:D-1}] = 0$.

And then, since we assume q is with bounded support, we have

$$
\begin{aligned}
& \mathbb{E}[|Z_D|] = \mathbb{E}_{p(y_{1:D})}[|Z_D|] \\
=& \mathbb{E}_{p(y_{1:D})}\left[\left|\sum_{d=1}^D \left((y_d - x_d)^2 - \mathbb{E}_{p(y_d|y_{1:d-1})}[(y_d - x_d)^2]\right)\right|\right] \\
\leq& \sum_{d=1}^D \mathbb{E}_{p(y_{1:d})}\left[\left|(y_d - x_d)^2 - \mathbb{E}_{p(y_d|y_{1:d-1})}[(y_d - x_d)^2]\right|\right] \\
\leq& \sum_{d=1}^D \mathbb{E}_{p(y_{1:d})}\left[(y_d - x_d)^2 + \mathbb{E}_{p(y_d|y_{1:d-1})}[(y_d - x_d)^2]\right] \\
=& 2\sum_{d=1}^D \mathbb{E}_{p(y_{1:d-1})}\left[\mathbb{E}_{p(y_d|y_{1:d-1})}[(y_d - x_d)^2]\right] \\
=& 2\sum_{d=1}^D \mathbb{E}_{p(y_d)}\left[(y_d - x_d)^2\right] \\
\leq& 8DC^2 \\
\leq& \infty.
\end{aligned}
$$

So we show that $Z_D = \sum_{d=1}^D (y_d - x_d)^2 - \sum_{d=1}^D \mathbb{E}[(y_d - x_d)^2 | y_{1:d-1}]$ is a martingale.

Now we show that

$$
\begin{aligned}
|Z_d - Z_{d-1}| =& |(y_d - x_d)^2 - \mathbb{E}_{p(y_d|y_{1:d-1})}[(y_d - x_d)^2]| \\
\leq& (y_d - x_d)^2 + \mathbb{E}_{p(y_d|y_{1:d-1})}[(y_d - x_d)^2] \\
\leq& 8C^2.
\end{aligned}
$$

So, by choosing $b = \sum_{d=1}^D \mathbb{E}[(y_d - x_d)^2 | y_1, ..., y_{d-1}] - t$ (notice that b here is indeed a random variable) and using the inequality, we have

$$
\begin{aligned}
& P(\sum_{d=1}^D (y_d - x_d)^2 < b) \\
&= P(\sum_{d=1}^D (y_d - x_d)^2 - \sum_{d=1}^D \mathbb{E}[(y_d - x_d)^2 | y_1, ..., y_{d-1}] < -t) \\
&\leq \exp\left(-\frac{t^2}{128DC^4}\right)
\end{aligned}
$$

When $t = DC_0/2$, we have

$$
P(\sum_{d=1}^D (y_d - x_d)^2 < b) \leq \exp\left(-\beta D\right),
$$

where $\beta = C_0^2/(256C^4)$.

Now we bound the second term. Notice that $b = \sum_{d=1}^D \mathbb{E}[(y_d - x_d)^2 | y_1, ..., y_{d-1}] - \frac{1}{2}DC_0 \geq \frac{1}{2}\sum_{d=1}^D C_0 = DC_0/2 = b'$ almost surely as the assumption (b is a random variable while b' is a constant), we have

$$
\begin{aligned}
& P\left(|y_d - x_d| \geq \frac{b}{D^\alpha}\right) \leq P\left(|y_d - x_d| \geq \frac{b'}{D^\alpha}\right) \\
&= P\left(y_d \geq x_d + \frac{b'}{D^\alpha}\right) + P\left(y_d \leq x_d - \frac{b'}{D^\alpha}\right) \\
&= P\left(\exp(t_1 y_d) \geq \exp(t_1(\frac{b'}{D^\alpha} + x_d))\right) \\
&\quad + P\left(\exp(-t_2 y_d) \geq \exp(t_2(\frac{b'}{D^\alpha} - x_d))\right) \\
&\leq \frac{\mathbb{E}[\exp(t_1 y_d)]}{\exp\left(t_1(\frac{b'}{D^\alpha} + x_d)\right)} + \frac{\mathbb{E}[\exp(-t_2 y_d)]}{\exp\left(t_2(\frac{b'}{D^\alpha} - x_d)\right)} \\
&\leq \frac{\exp(t_1\mu_d + \frac{1}{2}t_1^2 C^2)}{\exp\left(t_1(\frac{b'}{D^\alpha} + x_d)\right)} + \frac{\exp(-t_2\mu_d + \frac{1}{2}t_2^2 C^2)}{\exp\left(t_2(\frac{b'}{D^\alpha} - x_d)\right)} \\
&= \frac{\exp(t_1\mu_d + \frac{1}{2}t_1^2 C^2)}{\exp\left(t_1(\frac{1}{2}D^{1-\alpha}C_0^2 + x_d)\right)} + \frac{\exp(-t_2\mu_d + \frac{1}{2}t_2^2 C^2)}{\exp\left(t_2(\frac{1}{2}D^{1-\alpha}C_0^2 - x_d)\right)} \\
&= \frac{\exp(\frac{1}{2}t_1^2 C^2 + t_1(\mu_d - x_d))}{\exp\left(\frac{1}{2}t_1 D^{1-\alpha}C_0^2\right)} + \frac{\exp(\frac{1}{2}t_2^2 C^2 - t_2(\mu_d - x_d))}{\exp\left(\frac{1}{2}t_2 D^{1-\alpha}C_0^2\right)}
\end{aligned}
$$

holds for any $t_1, t_2 > 0$, where the first inequality holds because of the Markov inequality, and the second inequality holds according to the definition of the sub-Gaussian distribution. According to Hoeffding's Lemma, any bounded random variables $|Z| \leq C$ corresponds to C-sub-Gaussian distribution, which satisfies $\mathbb{E}[e^{t(Z-\mu)}] \leq \exp(t^2C^2/2)$ for any $t \in \mathbb{R}$.

Now let $t = t_1 = t_2$ and $\mu_d' = \mu_d - x_d$, we have

$$
\begin{aligned}
& P\left(|y_d - x_d| \geq \frac{b}{D^\alpha}\right) \\
&\leq \frac{\exp(t^2C^2/2 + t\mu_d')}{\exp\left(tD^{1-\alpha}C_0^2/2\right)} + \frac{\exp(t^2C^2/2 - t\mu_d')}{\exp\left(5tD^{1-\alpha}\sigma^2\right)} \\
&= \exp\left(-tD^{1-\alpha}C_0^2/2 + t^2C^2/2\right)\left(e^{t\mu_d'} + e^{-t\mu_d'}\right) \\
&\leq 2\exp\left(-tD^{1-\alpha}C_0^2/2 + t^2C^2/2\right)\cosh(t\mu_d')
\end{aligned}
$$

By choosing $t = 2/C_0^2$, we have

$$
P\left(|y_d - x_d| \geq \frac{b}{D^\alpha}\right) \leq L\exp(-D^{1-\alpha})
$$

where $2\exp\left(2C^2/C_0^2\right)\cosh(2\mu_{z_d}/C_0^2) \leq 2\exp\left(2C^2/C_0^2\right)\cosh(2\|\boldsymbol{\mu} - \mathbf{x}\|_\infty/C_0^2) \leq 2\exp\left(2C^2/C_0^2\right)\cosh(4C/C_0^2) = L$.

Combining these two terms, we have

$$
P\left(\frac{\|\mathbf{y}-\mathbf{x}\|_\infty}{\|\mathbf{y}-\mathbf{x}\|_2^2} \geq \frac{1}{D^\alpha}\right) \leq e^{-\beta D} + LDe^{-D^{1-\alpha}}
$$

for some $\beta, L \geq 0$. Now, we'd like to give a clean (but loose) bound by noticing that

$$
e^{-\beta D} + LDe^{-D^{1-\alpha}} \leq (L+1)D\max\{e^{-D^{1-\alpha}}, e^{-\beta D}\}.
$$

By using some derivations, we can get another bound

$$
De^{-D^{1-\alpha}} \leq e^{-(1-1/e)D^{1-\alpha}},
$$

and

$$
De^{-\beta D} \leq \frac{1}{\beta}e^{-\frac{1}{2}\beta D}.
$$

Let $\delta' \geq (L+1)\max\{e^{-(1-1/e)D^{1-\alpha}}, \frac{1}{\beta}e^{-\frac{1}{2}\beta D}\}$, we have

$$
D \geq \max\{\exp(\frac{1}{1-\alpha})\frac{1}{1-1/e}\log\frac{L+1}{\delta'}, \frac{2}{\beta}\log\frac{L+1}{\beta\delta'}\}.
$$

As a result, for any $\delta' \in (0,1)$, there exists $D_0 = \max\{\exp(\frac{1}{1-\alpha})\frac{1}{1-1/e}\log\frac{L+1}{\delta'}, \frac{2}{\beta}\log\frac{L+1}{\beta\delta'}\}$, such that for any $D > D_0$, we have $\frac{\|\mathbf{y}-\mathbf{x}\|_\infty}{\|\mathbf{y}-\mathbf{x}\|_2^2} \leq \frac{1}{D^\alpha}$ with at least probability $1-\delta'$.

Now, we begin to prove our proposition. By using the conclusion in section A.2, we can bound $\|\mathbf{R}(\mathbf{x}; \hat{q}_M)\|_\infty$ as

$$
\|\mathbf{R}(\mathbf{x}; \hat{q}_M)\|_\infty \leq \frac{2}{Me}\sum_{i=1}^{M}\frac{\|\mathbf{x}-\mathbf{y}\|_\infty}{\|\mathbf{x}-\mathbf{y}\|_2^2}.
$$

According to the union bound, we have

$$
\begin{aligned}
& P\left(\max_i \frac{\|\mathbf{x}^{(i)}-\mathbf{x}\|_\infty}{\|\mathbf{x}^{(i)}-\mathbf{x}\|_2^2} \geq D^{-\alpha}\right) \\
&\leq \sum_{i=1}^{M} P\left(\frac{\|\mathbf{x}^{(i)}-\mathbf{x}\|_\infty}{\|\mathbf{x}^{(i)}-\mathbf{x}\|_2^2} \geq D^{-\alpha}\right) \\
&\leq MP\left(\frac{\|\mathbf{y}-\mathbf{x}\|_\infty}{\|\mathbf{y}-\mathbf{x}\|_2^2} \geq D^{-\alpha}\right), \mathbf{y} \sim q
\end{aligned}
$$

where the last inequality holds since $\{\mathbf{x}^{(i)}\}_{i=1}^M$ are samples from q. Then, we can directly apply the conclusion with $\delta = M\delta'$. Then, we have, for any $\delta \in (0,1)$, there exists $D_0 = \max\{\exp(\frac{1}{1-\alpha})\frac{1}{1-1/e}\log\frac{(L+1)M}{\delta}, \frac{2}{\beta}\log\frac{(L+1)M}{\beta\delta}\}$, such that for any $D > D_0$, we have

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

with at least probability $1-\delta$.

B. Detailed Derivation and Proof for Section 4

B.1. Derivation of Sub-KL Divergence

Given the condition that

$$
\mathrm{KL}(q(x_d|\mathbf{x}_{-d})q(\mathbf{x}_{-d})\|p(x_d|\mathbf{x}_{\Gamma_d})q(\mathbf{x}_{-d}))] = 0,\ \forall d,
$$

we have $q(x_d|\mathbf{x}_{-d}) = p(x_d|\mathbf{x}_{\Gamma_d}) = p(x_d|\mathbf{x}_{-d}),\ \forall d$. When both p and q are differentiable, we have $\nabla_{x_d}\log q(x_d|\mathbf{x}_{-d}) = \nabla_{x_d}\log p(x_d|\mathbf{x}_{-d}), \forall d$. In other words, we have $\nabla_\mathbf{x}\log q(\mathbf{x}) = \nabla_\mathbf{x}\log p(\mathbf{x})$, and thus $q(\mathbf{x}) = e^C p(\mathbf{x})$. By using the normalization property of distribution, we have $C = 0$ and thus $q(\mathbf{x}) = p(\mathbf{x})$.

B.2. Derivation of $q_{[\boldsymbol{T}]}(\mathbf{z}_{\neg d}) = q(\mathbf{z}_{\neg d})$

Recall the change of variable theorem, we have

$$q_{[\boldsymbol{T}]}(\mathbf{z}) = q(\boldsymbol{T}^{-1}(\mathbf{z}))\left|\det(\nabla_{\mathbf{z}}\boldsymbol{T}^{-1})\right|.$$

Since $\boldsymbol{T}_{\neg d}$ is an identity mapping from $\mathbf{x}_{\neg d}$ to $\mathbf{x}_{\neg d}$, $\nabla_{\mathbf{z}}\boldsymbol{T}^{-1}$ is a block-wise triangular matrix and the determinant $\det(\nabla_{\mathbf{z}}\boldsymbol{T}^{-1})$ satisfies

$$\begin{aligned}\det(\nabla_{\mathbf{z}}\boldsymbol{T}^{-1}) &= \det(\nabla_{\mathbf{z}_{\neg d}}\boldsymbol{T}_{\neg d}^{-1}) \cdot \det(\nabla_{z_d}T_d^{-1})\\ &= \det(\nabla_{z_d}T_d^{-1}).\end{aligned}$$

As a result, we have

$$q_{[\boldsymbol{T}]}(\mathbf{z}) = q(\boldsymbol{T}^{-1}(\mathbf{z}))\left|\det(\nabla_{z_d}T_d^{-1})\right|.$$

So $q_{[T]}(\mathbf{z}_{\neg d}) =$

$$\begin{aligned}&\textstyle\int q_{[T]}(\mathbf{z})dz_d\\ &= q(\mathbf{z}_{\neg d})\textstyle\int q(T_d^{-1}(z_d)|\mathbf{z}_{\neg d})\left|\det(\nabla_{z_d}T_d^{-1})\right|dz_d\\ &= q(\mathbf{z}_{\neg d})\textstyle\int q_{[T_d]}(z_d|\mathbf{z}_{\neg d})dz_d = q(\mathbf{z}_{\neg d}).\end{aligned}$$

B.3. Proof of Proposition 3

First we prove that

$$\begin{aligned}&\nabla_\epsilon \mathrm{KL}(q_{[\boldsymbol{T}]}\|p) =\\ &\quad\nabla_\epsilon \mathrm{KL}\big(q_{[T_d]}(z_d|\mathbf{z}_{\Gamma_d})q(\mathbf{z}_{\Gamma_d})\big\|p(z_d|\mathbf{z}_{\Gamma_d})q(\mathbf{z}_{\Gamma_d})\big).\end{aligned}$$

Given $\mathbf{z} = \boldsymbol{T}(\mathbf{x}) = [x_1, ..., T_d(x_d), ..., x_D]^\top$, as proved in Section B.2, we have $q_{[\boldsymbol{T}]}(\mathbf{z}) = q(\boldsymbol{T}^{-1}(\mathbf{z}))|\det(\nabla_{z_d}T_d^{-1})|$ and thus

$$\begin{aligned}\mathrm{KL}(q_{[\boldsymbol{T}]}\|p) = &\ \mathrm{KL}\big(q_{[T_d]}(z_d|\mathbf{z}_{\neg d})q(\mathbf{z}_{\neg d})\big\|p(z_d|\mathbf{z}_{\Gamma_d})q(\mathbf{z}_{\neg d})\big)\\ &+ \mathrm{KL}\big(q(\mathbf{z}_{\neg d})\|p(\mathbf{z}_{\neg d})\big).\end{aligned} \tag{1}$$

When $T_d : x_d \to x_d + \epsilon\phi_d(x_{S_d})$ where $S_d = \{d\} \cup \Gamma_d$, we can further decompose the right handside of Eq. (1), i.e., $\mathrm{KL}\big(q_{[T_d]}(z_d|\mathbf{z}_{\neg d})q(\mathbf{z}_{\neg d})\big\|p(z_d|\mathbf{z}_{\Gamma_d})q(\mathbf{z}_{\neg d}\big) =$

$$\begin{aligned}&\mathrm{KL}\big(q_{[T_d]}(z_d|\mathbf{z}_{\neg d})q(\mathbf{z}_{\neg d})\big\|q_{[T_d]}(z_d|\mathbf{z}_{\Gamma_d})q(\mathbf{z}_{\neg d})\big)\\ &+ \mathrm{KL}\big(q_{[T_d]}(z_d|\mathbf{z}_{\Gamma_d})q(\mathbf{z}_{\neg d})\big\|p(z_d|\mathbf{z}_{\Gamma_d})q(\mathbf{z}_{\neg d})\big).\end{aligned}$$

By using the change of variable, we can find out that

$$\begin{aligned}&\mathrm{KL}\big(q_{[T_d]}(z_d|\mathbf{z}_{\neg d})q(\mathbf{z}_{\neg d})\big\|q(z_d|\mathbf{z}_{\Gamma_d})q(\mathbf{z}_{\neg d})\big)\\ &= \int q_{[\boldsymbol{T}]}(\mathbf{z})\log\frac{q_{T_d}(z_d|\mathbf{z}_{\neg d})}{q_{T_d]}(z_d|\mathbf{z}_{\Gamma_d})}d\mathbf{z}\\ &= \int q(\mathbf{x})\log\frac{q(x_d|\mathbf{x}_{\neg d})/|\det(\nabla_{x_d}T_d)|}{q(x_d|\mathbf{x}_{\Gamma_d})/|\det(\nabla_{x_d}T_d)|}d\mathbf{x}\\ &= \int q(\mathbf{x})\log\frac{q(x_d|\mathbf{x}_{\neg d})}{q(x_d|\mathbf{x}_{\Gamma_d})}d\mathbf{x}\\ &= \mathrm{KL}\big(q(x_d|\mathbf{x}_{\neg d})q(\mathbf{x}_{\neg d})\big\|q(x_d|\mathbf{x}_{\Gamma_d})q(\mathbf{x}_{\neg d})\big),\end{aligned}$$

which is unrelated with T_d (and thus unrelated with ϵ). As a result, we have

$$\begin{aligned}&\nabla_\epsilon \mathrm{KL}(q_{[\boldsymbol{T}]}\|p) =\\ &\quad\nabla_\epsilon \mathrm{KL}\big(q_{[T_d]}(z_d|\mathbf{z}_{\Gamma_d})q(\mathbf{z}_{\Gamma_d})\big\|p(z_d|\mathbf{z}_{\Gamma_d})q(\mathbf{z}_{\Gamma_d})\big).\end{aligned}$$

Now we derive the optimal ϕ_d^* for $\min_{\|\phi_d\|_{\mathcal{H}_d}\le 1}\nabla_\epsilon \mathrm{KL}(q_{[\boldsymbol{T}]}\|p)|_{\epsilon=0}$. Notice that

$$\begin{aligned}&\mathrm{KL}\big(q_{[T_d]}(z_d|\mathbf{z}_{\Gamma_d})q(\mathbf{z}_{\neg d})\big\|p(z_d|\mathbf{z}_{\Gamma_d})q(\mathbf{z}_{\neg d})\big)\\ &= \int q_{[T_d]}(z_d|\mathbf{z}_{\Gamma_d})q(\mathbf{z}_{\Gamma_d})\log\frac{q_{[T_d]}(z_d|\mathbf{z}_{\Gamma_d})}{p(z_d|\mathbf{z}_{\Gamma_d})}d\mathbf{z}\\ &= \mathbb{E}_{q(\mathbf{z}_{\Gamma_d})}\Big[\mathrm{KL}\big(q_{[T_d]}(z_d|\mathbf{z}_{\Gamma_d})\|p(z_d|\mathbf{z}_{\Gamma_d})\big)\Big].\end{aligned}$$

Following the proof of Theorem 3.1 in (Liu & Wang, 2016), we have

$$\begin{aligned}&\nabla_\epsilon \mathrm{KL}\big(q_{[T_d]}(z_d|\mathbf{z}_{\Gamma_d})\|p(z_d|\mathbf{z}_{\Gamma_d})\big)|_{\epsilon=0}\\ &= -\mathbb{E}_{q(y_d|\mathbf{y}_{\Gamma_d})}\big[\phi_d(\mathbf{y}_{S_d})\nabla_{y_d}\log p(y_d|\mathbf{y}_{\Gamma_d}) + \nabla_{y_d}\phi_d(\mathbf{y}_{S_d})\big].\end{aligned}$$

Combing the above three equations together, we have $\nabla_\epsilon \mathrm{KL}(q_{[\boldsymbol{T}]}\|p)|_{\epsilon=0} =$

$$-\mathbb{E}_{q(y_d|\mathbf{y}_{\Gamma_d})q(\mathbf{y}_{\Gamma_d})}\big[\phi_d(\mathbf{y}_{S_d})\nabla_{y_d}\log p(y_d|\mathbf{y}_{\Gamma_d}) + \nabla_{y_d}\phi_d(\mathbf{y}_{S_d})\big]$$

and $\min_{\|\phi_d\|_{\mathcal{H}_d}\le 1}\nabla_\epsilon \mathrm{KL}(q_{[\boldsymbol{T}]}\|p)|_{\epsilon=0}$ corresponds to

$$\max_{\|\phi_d\|_{\mathcal{H}_d}\le 1}\mathbb{E}_{q(\mathbf{y}_{S_d})}\big[\phi_d(\mathbf{y}_{S_d})\nabla_{y_d}\log p(y_d|\mathbf{y}_{\Gamma_d}) + \nabla_{y_d}\phi_d(\mathbf{y}_{S_d})\big].$$

By using the reproducing property of the RKHS $\mathcal{H}_d$, we have $\phi_d(\mathbf{y}_{S_d}) = \langle\phi_d(\cdot), k_d(\cdot, \mathbf{y}_{S_d})\rangle_{\mathcal{H}_d}$, and thus

$$\begin{aligned}&\mathbb{E}_{q(\mathbf{y}_{S_d})}\big[\phi_d(\mathbf{y}_{S_d})\nabla_{y_d}\log p(y_d|\mathbf{y}_{\Gamma_d}) + \nabla_{y_d}\phi_d(\mathbf{y}_{S_d})\big]\\ &= \Big\langle\phi_d(\cdot), \mathbb{E}_{q(\mathbf{y}_{S_d})}\big[k_d(\cdot, \mathbf{y}_{S_d})\nabla_{y_d}\log p(y_d|\mathbf{y}_{\Gamma_d}) + \nabla_{y_d}k_d(\cdot, \mathbf{y}_{S_d})\big]\Big\rangle_{\mathcal{H}_d}.\end{aligned}$$

Following the derivation in (Liu et al., 2016) and (Chwialkowski et al., 2016), we can show the optimal solution is $\phi_d^*/\|\phi_d^*\|_{\mathcal{H}_d}$ where

$$\begin{aligned}\phi_d^*(\mathbf{x}_{S_d}) =&\mathbb{E}_{q(\mathbf{y}_{S_d})}\big[k_d(\mathbf{x}_{S_d}, \mathbf{y}_{S_d})\nabla_{y_d}\log p(y_d|\mathbf{y}_{\Gamma_d})\\ &+ \nabla_{y_d}k_d(\mathbf{x}_{S_d}, \mathbf{y}_{S_d})\big].\end{aligned}$$

C. More Experimental Results

C.1. Toy Example for SVGD with the IMQ Kernel

Fig. 1 shows the toy example for SVGD with the IMQ kernel. We can find out that the behavior of the IMQ kernel resembles that of the RBF kernel.

C.2. The Impact of Bandwidth

Bandwidth plays an important role in kernel methods. In this section, we provide additional experimental results for the impact of bandwidth over the performance of SVGD.

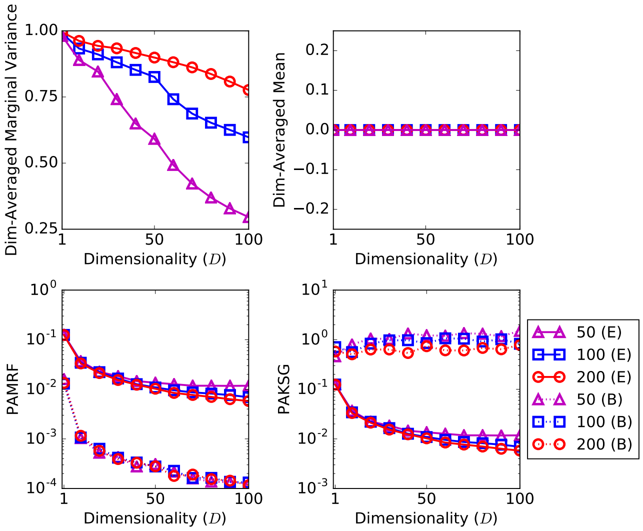

Figure 1. Results for inferring $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I})$ using SVGD with the IMQ kernel, where particles are initialized by $\mathcal{N}(\mathbf{x}|\mathbf{0}, 25\mathbf{I})$. Top two figures show the dimension-averaged marginal variance $\frac{1}{D}\sum_{d=1}^{D} \mathrm{Var}_{\hat{q}_M}(x_d)$ and mean $\frac{1}{D}\sum_{d=1}^{D} \mathbb{E}_{\hat{q}_M}[x_d]$ respectively, and bottom two figures show the particle-averaged magnitude of the repulsive force (PAMRF) $\frac{1}{M}\sum_{i=1}^{M} \|\mathbf{R}(\mathbf{x}^{(i)}; \hat{q}_M)\|_\infty$ and kernel smoothed gradient (PAKSG) $\frac{1}{M}\sum_{i=1}^{M} \|\mathbf{G}(\mathbf{x}^{(i)}; p, \hat{q}_M)\|_\infty$ respectively, at both the beginning (dotted;B) and the end of iterations (solid;E) with different number of particles $M = 50, 100$ and 200.

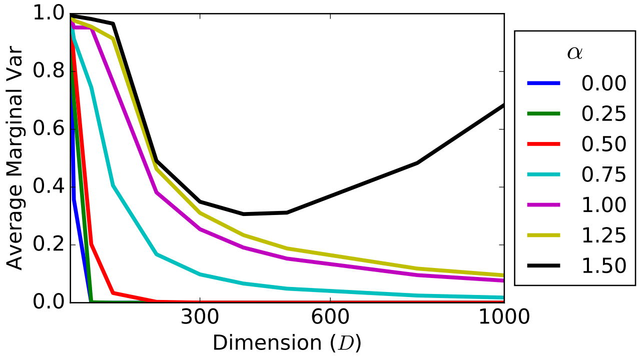

Figure 2. The average marginal variance of particles generated by SVGD with different bandwidth versus dimension.

In this experiment, we set the target to be $p(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$ as a D dimensional isotropic Gaussian distribution, and use $M = 100$ particles initialized as i.i.d examples from $q_0(x) = \mathcal{N}(x|0, 25\mathbf{I})$. We use the RBF kernel $k(x, y) = \exp(\frac{-\|x-y\|_2^2}{2h})$, in which the bandwidth $h = D^{\alpha-1} \cdot \mathrm{med}^2$ with $\alpha = 1$ the median heuristic, $\alpha > 1$ the overestimated bandwidth and $\alpha < 1$ the underestimated bandwidth. We evaluate the quality of particles in marginal approximation by using the average marginal variance $\frac{1}{D}\sum_{d=1}^{D} \mathrm{Var}_{\hat{q}_M}(x_d)$, which measures the extent to which the particles are diverse to each other in marginals. The average marginal variance of $p(\mathbf{x})$ is 1. For all experiments, we use Adagrad (Duchi et al., 2011) for step size and execute 10000 iterations to get final particles.

Fig. 2 demonstrates the relationship among marginal particle diversity, bandwidth choices and dimensions. An interesting observation is that there exists an inflection point around $D = 400$ in the curve of overestimated bandwidth ($\alpha = 1.5$). The reason is that larger bandwidth leads to smaller $\hat{\phi}^*(\mathbf{x})$ and thus slower convergence, and this phenomenon deteriorates as dimension increases. Thus, for the bandwidth $\alpha = 1, 5$ and $D > 400$, SVGD cannot converge with 10000 iterations. Excluding this unconverged case, we can find out that as dimension increases, the approximation deteriorates no matter which bandwidth is chosen.

Fig. 3 demonstrates the dynamic of SVGD with different bandwidth. To highlight the difference, we use the log scale in Y axis. As is shown, bandwidth plays an important role in the convergence of SVGD: smaller bandwidth leads to faster convergence. And the gap between different bandwidth becomes larger as dimension increases. Another observation is, when converged, larger bandwidth corresponds to higher marginal variance, which implies more diverse particles and better marginal approximation. Among these bandwidth choices, the median heuristic ($\alpha = 1$) is somehow the best one for two reasons: (1) It converges almost as fast as the underestimated bandwidth ($\alpha < 1$); (2) It achieves almost the best marginal variance. Though overestimated bandwidth ($\alpha > 1$) achieves slightly better performance than the median heuristic when converged, the gap is not as large as that between the median heuristic and underestimated bandwidth. For example, in the rightmost figure, the gap of the average marginal variance between $\alpha = 1.25$ and $\alpha = 1$ is much smaller than that between $\alpha = 0.75$ and $\alpha = 1$.

C.3. Synthetic Markov Random Fields

Fig. 4 compares EP with other methods mentioned in the main body. Due to the strong Gaussian assumption, EP achieves the highest RMSE compared to other methods.

C.4. Image Denoising

Fig. 5 shows more denoising examples apart from *Lena* in the main body.

References

Azuma, K. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

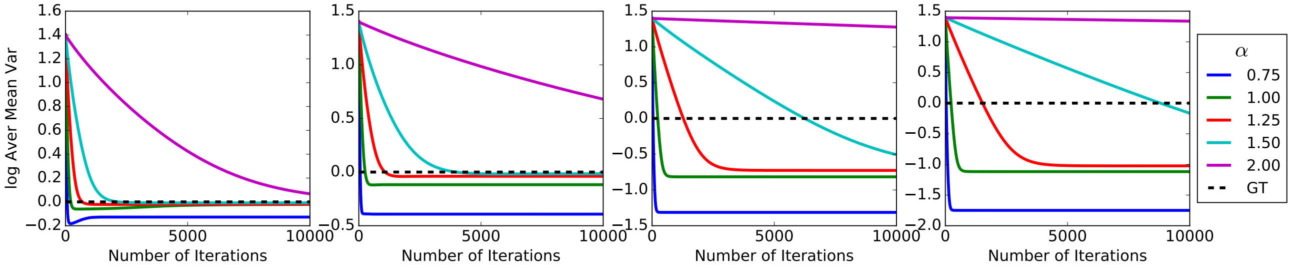

Figure 3. The convergence performance of SVGD with different bandwidth is evaluated for different dimension $D = 50, 100, 500, 1000$ arrange from left to right. "GT" denotes the ground truth, which equals one.

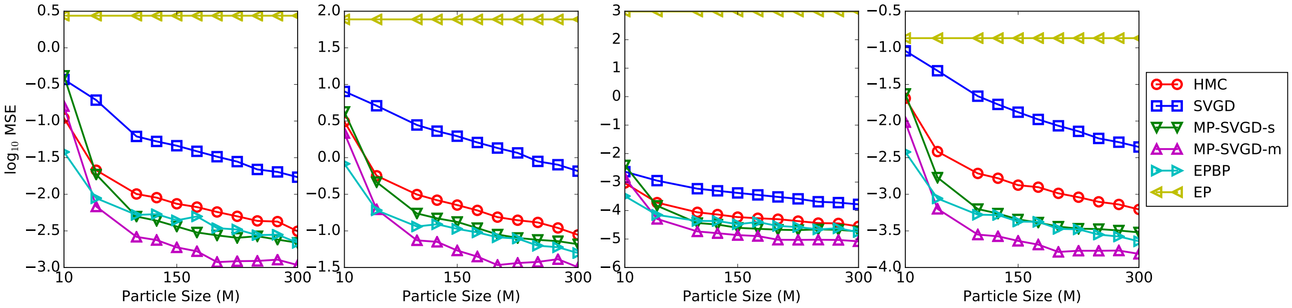

Figure 4. A quantitative comparison of inference methods with varying number of particles. Performance is measured by the MSE of the estimation of expectation $\mathbb{E}_{\mathbf{x}\sim\hat{q}_M}[\mathbf{f}(\mathbf{x})]$ for test functions $\mathbf{f}(\mathbf{x}) = \mathbf{x}$, $\mathbf{x}^2$, $1/(1+\exp(\boldsymbol{\omega}\circ\mathbf{x}+\mathbf{b}))$ and $\cos(\boldsymbol{\omega}\circ\mathbf{x}+\mathbf{b})$, arranged from left to right, where $\circ$ denotes the element-wise product. Results are averaged over 10 random draws of $\boldsymbol{\omega}$ and $\mathbf{b}$, where $\boldsymbol{\omega}, \mathbf{b} \in \mathbb{R}^{100}$ with $\omega_d \sim \mathcal{N}(0, 1)$ and $b_d \in \mathrm{Uniform}[0, 2\pi]$, $\forall d \in \{1, ..., 100\}$.

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. In *Proceedings of the International Conference on Machine Learning*, 2016.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, 2016.

Liu, Q., Lee, J., and Jordan, M. A kernelized stein discrepancy for goodness-of-fit tests. In *Proceedings of the International Conference on Machine Learning*, 2016.

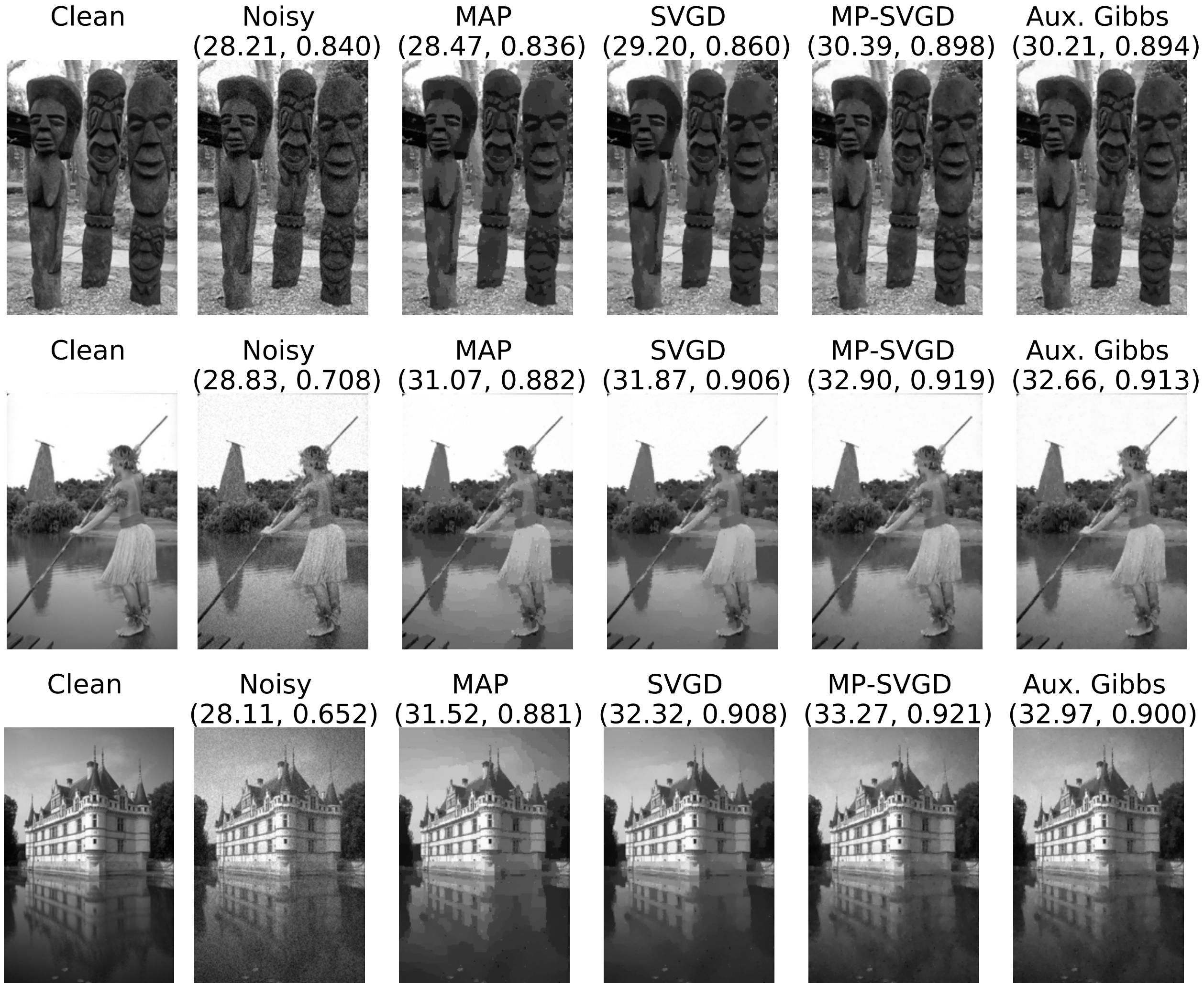

Figure 5. Extra denoising results on BSD dataset using 50 particles, 240×160 pixels, $\sigma_n = 10$. The number in bracket is PSNR and SSIM.