I said I would not write on SVGD anymore...

Adil Salim

Microsoft

ICERM
May 2024
Joint work with

Victor Priser

Pascal Bianchi

Ongoing work. References, suggestions, comments are welcome!
Sampling

- Bayesian inference in machine learning
- Generative models

"Pikachu eating a sandwich"
Outline

1. SVGD
2. Noisy SVGD
3. Proof
Outline

1. SVGD
2. Noisy SVGD
3. Proof
SVGD [Liu and Wang, 2016] is an algorithm to sample from $\mu_\star \propto \exp(-F)$, where $F$ $L$-smooth and nonconvex. SVGD maintains a set of $N$ particles $x^1, \ldots, x^N$.

$$x^i_{k+1} = x^i_k - \frac{\gamma}{N} \sum_{j=1}^{N} \nabla F(x^j_k) K(x^i_k, x^j_k) - \nabla_2 K(x^i_k, x^j_k),$$

where $K(x, y)$ is a kernel associated to a Reproducing Kernel Hilbert Space $H$. 
Low dimension vs high dimension

\[ \mu_*(x) \propto \exp(-F(x)) \]

Simulation from [KSA+20] (Code from Q. Liu)

However, in higher dimension SVGD particles can collapse due to the deterministic updates [Ba et al., 2021].
What do we know about the convergence of SVGD?

Let $\mu_k^N$ be the empirical measure of SVGD at iteration $k$, i.e.,

$$\mu_k^N = \frac{1}{N} \sum_{j=1}^{N} \delta_{x_k^j}$$

(1)

<table>
<thead>
<tr>
<th></th>
<th>$k$ small</th>
<th>$k$ large</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$ large</td>
<td>$\text{KSD}(\mu_k^\infty</td>
<td>\mu^*) &lt; \frac{C}{k}$</td>
</tr>
<tr>
<td></td>
<td>[KSA$^+$20]</td>
<td>[SSR22]</td>
</tr>
<tr>
<td>$N$ small</td>
<td>$\text{KSD}(\mu_k^N</td>
<td>\mu^*) &lt; \frac{C'}{k'}, k &lt; \log \log(N)$</td>
</tr>
<tr>
<td></td>
<td>[Shi and Mackey, 2024]</td>
<td></td>
</tr>
</tbody>
</table>
Remarks on the asymptotics of $\mu_k^N$ when $k \to \infty$

When $N < \infty$,

1. $(\mu_k^N)_k$ does not converge to $\mu_*$ as $k \to \infty$.
   Because $\mu_k^N$ is discrete with $N < \infty$ masses whereas $\mu_*$ has a continuous density.

2. The best hope is for $(\mu_k^N)_k$ to converge to something that converges to $\mu_*$ as $N$ grows.

3. Even if we were able to show that $(\mu_k^N)$ converges to some $\mathcal{L}_k^N$ as $k \to \infty$ (already non trivial, the particles could diverge), $\mathcal{L}_k^N$ would probably not converge to $\mu_*$ as $N \to \infty$.
   Because of particles collapse in SVGD.
Outline

1. SVGD
2. Noisy SVGD
3. Proof
We study noisy SVGD

Noisy SVGD is SVGD with Langevin regularization.

\[ x_{k+1}^i = x_k^i - \frac{\gamma_k}{N} \sum_{j=1}^{N} \nabla F(x_k^j)K(x_k^i, x_k^j) - \nabla_2 K(x_k^i, x_k^j) - \varepsilon \gamma_k \nabla F(x_k^i) + \sqrt{2\gamma_k \varepsilon \xi_k} \]

where \(\varepsilon > 0\) is noise parameter, \(\gamma_k \to 0\) and \((\xi_k^i)_{i,k}\) i.i.d standard Gaussian.

Our goal: describe the "limit" \(\mathcal{L}^N\) of noisy SVGD as \(k \to \infty\). First remark: \(\mu_* \notin \mathcal{L}^N\).
Let $\mu^N_k$ be the empirical measure of noisy SVGD and $\mu^N$ its empirical average over time $k$. We view them as random variables in the metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$.

1. **”Limit” of noisy SVGD:**
   For every $N > 0$, the sequence of r.v. $(\mu^N_k)_k$ is tight. Therefore $(\mu^N_k)_k$ converges in distribution as $k \to \infty$ to the set of its cluster points $\mathcal{L}^N$.

2. **Description of the ”limit” I:**
   The set of r.v. $\bigcup_{N > 0} \mathcal{L}^N$ is tight. Therefore $(\mathcal{L}^N)_N$ ”converges” in distribution as $N \to \infty$ to the set of its cluster points $\mathcal{L}^\infty$.

3. **Description of the ”limit” II:** $\mathcal{L}^\infty = \{\mu_\star\}$ a.s.
Corollaries

All together our results imply

$$\sum_{\ell=1}^{k} \gamma_{\ell} W_2(\mu_{\ell}^N, \mu^\star) \xrightarrow{\mathbb{P}} 0$$  \hspace{1cm} (2)

Under Log Sobolev Inequality,

$$W_2(\mu_k^N, \mu^\star) \xrightarrow{\mathbb{P}} 0$$  \hspace{1cm} (3)

The regime $N \ll k$ is new.
Outline

1 SVGD
2 Noisy SVGD
3 Proof
Interpolated process level

For each particle trajectory \((x^i_k)_k\) we define the interpolated trajectory \(x^i : \mathbb{R}_+ \to \mathbb{R}^d\).

Let \(\mathcal{C}\) the set of continuous functions from \(\mathbb{R}_+\) to \(\mathbb{R}^d\) endowed with the topology of uniform convergence on compact sets. Then, \(x^i \in \mathcal{C}\). Moreover, \(x^i(t + \cdot) \in \mathcal{C}\) for every \(t \geq 0\).

Next we define the empirical measure of the shifted interpolated trajectories

\[
\mu^N(t) = \frac{1}{N} \sum_{j=1}^{N} \delta_{x^j(t+\cdot)}
\]  

(4)

We view \(\mu^N(t)\) as a r.v. in the metric space \((\mathcal{P}_2(\mathcal{C}), \mathcal{W}_2)\).
”Limit” of noisy SVGD

Under growth assumption on $F$:

For every $N > 0$, the sequence of r.v. $(\mu^N(t))_t$ is tight. Therefore $(\mu^N(t))_t$ converges in distribution as $k \to \infty$ to the set of its cluster points $\mathcal{L}^N$.

This time each element of $\mathcal{L}^N$ is a random measure supported by $N$ continuous functions, i.e., trajectories (instead of $N$ points as before).
Under growth assumption on $F$:

The set of r.v. $\cup_{N>0} \mathcal{L}^N$ is tight. Therefore $(\mathcal{L}^N)_N$ ”converges” in distribution as $N \to \infty$ to the set of its cluster points $\mathcal{L}\infty$.

It remains to relate $\mathcal{L}\infty$ and $\mu\star$. 
McKean Vlasov equation

\[ dX_t = -b(X_t, \mathcal{L}(X_t))dt + \sqrt{2\varepsilon}dB_t, \]

\[ b(x, \mu) := \int \mu(dy) \nabla F(y)K(x, y) - \nabla^2 K(x, y) + \varepsilon \nabla F(x) \]

where \((B_t)\) standard Brownian motion.

The trajectory of one particle of Noisy SVGD is a discretization of MKV equation in time and space.

A McKean Vlasov measure = the law of a (weak) solution \((X_t)_t\) of the MKV equation.
McKean Vlasov measures

More precisely, $\rho \in P_2(C)$ is a MKV measure if $\rho$ solves the following martingale problem. For every $g \in C^2_c(\mathbb{R}^d)$,

$$g(X_t) - \int_0^t \langle b(X_s, \rho_s), \nabla g(X_s) \rangle + \varepsilon^2 \Delta g(X_s) \, ds$$

is a martingale, where $(X_t)_t \sim \rho$. 
Recall that the elements of $L^\infty$ are random measures over the set of continuous functions.

Under boundedness assumption of the kernel:

The elements of $L^\infty$ are a.s. McKean Vlasov measures. To understand $L^\infty$ elements (and relate them to $\mu_\star$), we need to understand MKV measures.
Asymptotics of MKV measures when $t \to \infty$

Let $\rho$ a (deterministic) MKV measure and denote $\rho_t$ its marginal distributions.

1. If $\rho$ is stationary, then $\rho_t = \mu_\star$ for every $t > 0$
2. Under LSI, $\rho_t \to \mu_\star$ uniformly.

Basically, the realizations of $\mathcal{L}^\infty$ are measures $\rho$ s.t. $\mathcal{W}_2(\rho_t, \mu_\star) \to 0$. The various conclusions we obtained follow from this observation.
Conclusion

- We provide understanding of $\mathcal{L}^N$, the limit set of noisy SVGD as $k \to \infty$.
- We show a dynamical result (convergence to MKV measures) on the way.
- Quantification of the convergence to $\mathcal{L}^N$?
- Quantification of the convergence of $\mathcal{L}^N$ to $\mathcal{L}^\infty$?
- Role of $\varepsilon$ on the convergence speed?
Understanding the variance collapse of svgd in high dimensions.
In *International Conference on Learning Representations*.

Stein variational gradient descent: A general purpose Bayesian inference algorithm.

[Shi and Mackey, 2024] Shi, J. and Mackey, L. (2024).
A finite-particle convergence rate for stein variational gradient descent.
Selected publications I
