Learning Generative Models with Sinkhorn Divergences

Supplementary Material

Anonymous

October 20, 2017

Abstract

These supplementary materials present numerical evidence of the low-sample complexity of Sinkhorn divergence, and of its positivity.

1 Numerical Exploration of the Sinkhorn Divergence

1.1 Sample Complexity

To better grasp the statistical tradeoff offered by the entropic regularization, we study numerically the so-called sample complexity of these divergence. We consider

\[ \hat{\mu}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i} \quad \text{and} \quad \hat{\nu}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{y_i}, \]

which are random measures, where the \((x_i)\) and \((y_i)\) are points independently drawn from the same distribution \(\xi\). In the numerical experiments, \(\xi\) is the uniform distribution on \([0, 1]^d\) where \(d \in \mathbb{N}^*\) is the ambient dimension.

We recall that

\[ \bar{W}_{c,\epsilon} \left( \mu, \nu \right) \overset{\text{def}}{=} 2W_{c,\epsilon}(\mu, \nu) - W_{c,\epsilon}(\mu, \mu) - W_{c,\epsilon}(\nu, \nu) \]

where

\[ W_{c,\epsilon}(\mu, \nu) \overset{\text{def}}{=} \int c(x, y) d\gamma_{\epsilon} \]

where \(\gamma_{\epsilon}\) is the unique solution of the entropy-regularization optimal transport problem between \(\mu\) and \(\nu\). In the following, we consider \(c(x, y) = \|x - y\|^p\) for \(p = \frac{3}{2}\) for \((x, y) \in (\mathbb{R}^d)^2\).

As shown in the paper, one has

\[ W_{c,\epsilon}(\mu, \nu) \xrightarrow{\epsilon \rightarrow 0} 2W_p(\mu, \nu)^p \quad \text{and} \quad W_{c,\epsilon}(\mu, \nu) \xrightarrow{\epsilon \rightarrow +\infty} \|\mu - \nu\|^2_{\text{ED}(p)} \]
Figure 1: Influence of the regularization $\varepsilon$ on the sample complexity rate. The plot displays $\log_{10}(R_{\varepsilon,d}(N))$ as a function of $\log(N)$.

where $W_p$ is the Wasserstein-$p$ distance while $\|\xi\|_{ED}^2 = \int -\|x-y\|^p d\xi(x)d\xi(y)$ is the Energy Distance, which is a special case of MMD norm for $0 < p < 2$.

The goal is to study numerically the decay rate toward zero of

$$R_{\varepsilon,d}(N) \overset{\text{def.}}{=} \mathbb{E}(\tilde{W}_{c,\varepsilon}(\hat{\mu}_N, \hat{\nu}_N))$$

and also analyze the standard deviation

$$S^2_{\varepsilon,d}(N) \overset{\text{def.}}{=} \mathbb{E}(|\tilde{W}_{c,\varepsilon}(\hat{\mu}_N, \hat{\nu}_N) - R_{\varepsilon,d}(N)|^2).$$

In these formula, the expectation $\mathbb{E}$ with respect to random draws of $(x_i)_i$ and $(y_i)_i$ is estimated numerically by averaging over $10^3$ drawings. For optimal transport, i.e. $\varepsilon = 0$, it is well-known (we refer to the references given in the paper) that $R_{0,d}(N) = O(\frac{1}{N^{\frac{p}{d}}})$, while for MMD norm, i.e. $\varepsilon = +\infty$, one has $R_{+\infty,d}(N) = O(\frac{1}{N})$.

Figure 2 (resp. 1) display in log-log plot the decay of $R_{\varepsilon,d}(N)$ with $N$, and allows to compare on a single plot the influence of $d$ (resp. $\varepsilon$) for a fixed $\varepsilon$ (resp. $d$) on each plot.

From these experiments, one can conclude on this distribution $\xi$ that:

- $\mathbb{W}_{c,\varepsilon}(\mu, \nu) \geq 0$ (more on this in the following section).
- $R_{\varepsilon,d}(N)$ as a polynomial decay of the form $1/N^{\kappa_{\varepsilon,d}}$.
- One recovers the known rates $\kappa_{0,d} = p/d$ (here for $p = 3/2$) and $\kappa_{\infty,d} = 1$.
- Small values of $\varepsilon < 1$ have rates $\kappa_{\varepsilon,d}$ close to the rate of OT $\kappa_{0,d}$. 2
Large values of \( \varepsilon > 1 \) have rates \( \kappa_{\varepsilon,d} \) matching almost exactly the rate of MMD \( \kappa_{+\infty,d} = 1 \).

- The variance \( S_{\varepsilon,d}^2(N) \) is significantly smaller for small values of \( \varepsilon \) (i.e. close to OT).

Note that similar conclusion are obtained when testing on other distributions \( \xi \) (e.g. a Gaussian).

### 1.2 Positivity

For \( \varepsilon \in \{0, +\infty\} \), both OT and MMD are distances, so that \( \bar{W}_{\varepsilon,c}(\mu, \nu) = 0 \) if and only if \( \mu = \nu \). It not known whether this property is true for \( 0 < \varepsilon < +\infty \), and this seems a very difficult problem to tackle. We investigate numerically this question by looking at small modification of a discrete input measure \( \mu = \sum_i a_i \sum_{i=1}^N a_i \delta_{x_i} \) where the \( x_i \) are i.i.d. points drawn in \([0, 1]^2\) and \( (a_i) \) are i.i.d. number drawn uniformly in \([1/2, 1]\), and perform a small modification

\[
\mu_t \triangleq \frac{1}{\sum_i a_i, t} \sum_{i=1}^N a_i \delta_{x_i,t} \quad \text{where} \quad \left\{ \begin{array}{l}
a_{i,t} = a_{i,t} + t b_i, \\
x_{i,t} = x_i + t z_i,
\end{array} \right.
\]

where \( (b_i) \subset \mathbb{R} \) are i.d.d. Gaussian distributed \( \mathcal{N}(0,1) \) and where \( (z_i) \subset \mathbb{R}^2 \) are i.d.d. Gaussian distributed \( \mathcal{N}(0, \text{Id}_2) \).
Figure (3) shows, on a single realization of \((a_i, x_i, b_i, z_i)\), that \(\bar{W}_{c,e}(\mu, \mu_t) > 0\) for \(t \neq 0\). Testing for \(10^4\) other realizations gives the same results, showing that experimentally \(\bar{W}_{c,e}\) is locally strictly positive for discrete measures.