Asymptotic Guarantees for Learning Generative Models with the Sliced-Wasserstein Distance

NeurIPS 2019

Kimia Nadjahi[1], Alain Durmus[2], Umut Şimşekli[1], Roland Badeau[1]

[1] Télécom Paris, [2] ENS Paris-Saclay

Background

Minimum distance estimation (MDE)

Given i.i.d. observations $Y_{1:n} = (Y_1, \ldots, Y_n)$ and a family of distributions indexed by a parameter θ, the goal of MDE is:

$$\hat{\theta}_n = \mathrm{argmin}_{\theta \in \Theta} \ \mathbf{D}(\hat{\mu}_n, \mu_\theta) \ , \tag{1}$$

where $\mathbf{D}$: distance (or divergence) between probability measures, μ_θ: probability measure indexed by θ, Θ: parameter space, and $\hat{\mu}_n$: empirical measure of $Y_{1:n}$.

Optimal transport (OT): Wasserstein distance

For $p \geq 1$,
$\mathcal{P}_p(\mathsf{Y}) = \left\{\mu \in \mathcal{P}(\mathsf{Y}) \,:\, \int_{\mathsf{Y}} \|y - y_0\|^p \mathrm{d}\mu(y) < +\infty, \text{ for some } y_0 \in \mathsf{Y}\right\}$.

The Wasserstein distance of order p between any $\mu, \nu \in \mathcal{P}_p(\mathsf{Y})$ is,

$$\mathbf{W}_p^p(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \left\{ \int_{\mathsf{Y} \times \mathsf{Y}} \|x - y\|^p \mathrm{d}\gamma(x, y) \right\} ,$$

where $\Gamma(\mu, \nu)$: the set of probability measures γ on $(\mathsf{Y} \times \mathsf{Y}, \mathcal{Y} \otimes \mathcal{Y})$ satisfying $\gamma(\mathsf{A} \times \mathsf{Y}) = \mu(\mathsf{A})$ and $\gamma(\mathsf{Y} \times \mathsf{A}) = \nu(\mathsf{A})$ for any $\mathsf{A} \in \mathcal{B}(\mathsf{Y})$.

Problem: calculating $\mathbf{W}_p$ is computationally expensive.
. . . but $\mathbf{W}_p$ in 1D has a closed-form formula, which gives rise to an alternative OT distance: the Sliced-Wasserstein distance (SW).

Computational Optimal Transport: Sliced-Wasserstein distance

Let $\mathbb{S}^{d-1}$ be the d-dimensional unit sphere, $\langle \cdot, \cdot \rangle$ the Euclidean inner-product, and for any $u \in \mathbb{S}$, $u^\star(y) = \langle u, y \rangle$.

SW of order p between $\mu, \nu \in \mathcal{P}_p(\mathsf{Y})$ is,

$$\mathbf{SW}_p^p(\mu, \nu) = \int_{\mathbb{S}^{d-1}} \mathbf{W}_p^p(u_\sharp^\star \mu, u_\sharp^\star \nu) \mathrm{d}\boldsymbol{\sigma}(u) \tag{2}$$

where $\boldsymbol{\sigma}$: uniform distribution on $\mathbb{S}^{d-1}$, and for any measurable function $f : \mathsf{Y} \to \mathbb{R}$ and $\zeta \in \mathcal{P}(\mathsf{Y})$, $f_\sharp \zeta$: the push-forward measure of ζ by f.

$\Rightarrow$ SW has significantly lower computational requirements than the Wasserstein distance.

MDE + Computational OT

We plug $\mathbf{SW}_p$ in place of $\mathbf{D}$ in MDE.

Minimum Sliced-Wasserstein estimator (MSWE) *of order* p:

$$\hat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n, \mu_\theta)$$

Minimum expected Sliced-Wasserstein estimator (MESWE) *of order* p:

$$\hat{\theta}_{n,m} = \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) | Y_{1:n}\right]$$

Very popular and successful in recent machine learning applications, but their theoretical properties have not yet been fully established. $\Rightarrow$ We investigate them in our paper!

Contributions

Theoretical results

Topology induced by SW.

Let $p \in [1, +\infty)$. The convergence in $\mathbf{SW}_p$ implies the weak convergence of measures on $\mathbb{R}^d$.

Asymptotic properties of SW-based estimators.

- Existence and consistency of MSWE.
- Existence and consistency of MESWE.
- MESWE converges to MSWE.

Central Limit Theorems.

MSWE and the associated goodness-of-fit statistics converge to a random variable in distribution, with a rate of $\sqrt{n}$. Contrary to Wasserstein-based estimators, our result is not restricted to the 1D case, but holds for any dimension value.

Experiments

We empirically confirm our theoretical findings on:

- Multivariate Gaussians in $\mathbb{R}^{10}$: inference on the mean and scaling factor of the covariance.
- Elliptically contoured stable distributions in $\mathbb{R}^{10}$: inference on the location parameter, comparison of our estimators to Wasserstein-based estimators.
- SW-based GANs for image generation, applied on MNIST: inference on the neural network parameters.

Conclusion

Summary

- We investigated the asymptotic properties of estimators that are obtained by minimizing (expected) SW.
- We validated our theorems on both synthetic data and neural networks.

- Future work: derive analogous asymptotic guarantees for estimators based on extensions of SW.