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

Kimia Nadjahi[1] Alain Durmus[2] Umut Şimşekli[1,3] Roland Badeau[1]

[1] Télécom Paris [2] ENS Paris-Saclay [3] University of Oxford

Minimum Distance Estimation

$$\hat{}_n = \operatorname{argmin}_{2} \; D(\hat{}_n ; \quad)$$

D : distance between distributions

$\hat{}_n$: empirical distribution of data points $Y_1; : : : ; Y_n$ i.i.d from $_?$

: distribution parametrized by 2

Minimum Distance Estimation

$$\hat{}_n = \mathrm{argmin}_{\;2} \; D(\hat{}_n; \quad)$$

D : distance between distributions

Example: Generative Modeling

$\hat{}_n$: empirical distribution of data points $Y_1; : : : ; Y_n$ i.i.d from $_?$

: distribution parametrized by 2

Minimum Expected Distance Estimation

Directly optimizing is often not possible (e.g. GANs)

$$\hat{\ }_{n;m} = \operatorname{argmin}_{\ 2} \ E[D(\hat{\ }_{n} ; \hat{\ }_{;m}) \ j \ Y_{1:n}]$$

$\hat{\ }_{;m}$: empirical distribution of a sample $Z_1 ; : : : ; Z_m$ i.i.d. from

Minimum Wasserstein Estimation

Choose $D = W_p$ (Wasserstein distance of order p 1)

- X Robust and increasingly popular estimators: Wasserstein GAN [1], Wasserstein auto-encoders [2]
- X Asymptotic guarantees [3]

[1] Arjovsky et al., 2017 [2] Tolstikhin et al., 2018 [3] Bernton et al., 2019

Minimum Wasserstein Estimation

Choose $D = W_p$ (Wasserstein distance of order p 1)

- X Robust and increasingly popular estimators: Wasserstein GAN [1], Wasserstein auto-encoders [2]
- X Asymptotic guarantees [3]

[1] Arjovsky et al., 2017 [2] Tolstikhin et al., 2018 [3] Bernton et al., 2019

W_p: expensive + curse of dimensionality

Central limit theorem in [3] valid in 1D

Sliced-Wasserstein distance

In 1D, W_p has an analytical form $\Rightarrow$ Motivates a practical alternative:

$$\boxed{SW_p^p(\;\; ; \;\;) = \int_{S^{d-1}} W_p^p({\color{red}u^\star_\sharp} \;\; ; {\color{red}u^\star_\sharp} \;\;)d \;\; ({\color{red}u})}$$

Minimum Sliced-Wasserstein Estimation

$$\hat{}_{n} = \mathrm{argmin}_{\;2} \quad \mathrm{SW}_{p}(\hat{}_{n};\quad)$$

$$\hat{}_{n;m} = \mathrm{argmin}_{\;2} \quad \mathrm{E}\,[\mathrm{SW}_{p}(\hat{}_{n};\hat{}_{;m})\ j\ Y_{1:n}]$$

Successful in generative modeling applications(e.g., SW-GAN, Deshpande et al., 2018)

Minimum Sliced-Wasserstein Estimation

$$\hat{\ }_n = \operatorname{argmin}_{\ 2} \quad \mathrm{SW}_p(\hat{\ }_n; \quad)$$

$$\hat{\ }_{n;m} = \operatorname{argmin}_{\ 2} \quad \mathrm{E}\,[\mathrm{SW}_p(\hat{\ }_n; \hat{\ }_{;m}) \ j \ Y_{1:n}]$$

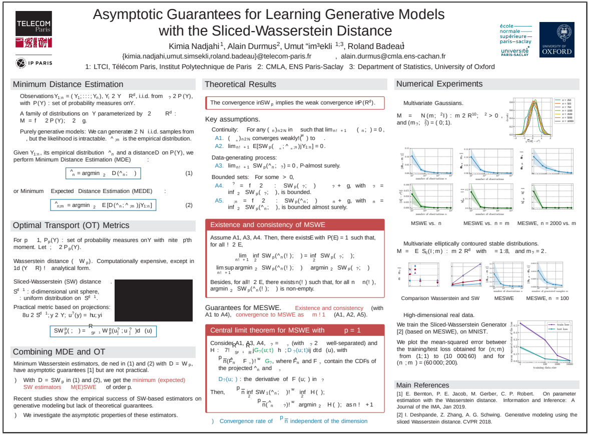
Successful in generative modeling applications (e.g., SW-GAN, Deshpande et al., 2018)

Our contributions:

- Convergence in SW_p) weak convergence of probability measures
- Existence and consistency of $\hat{\ }_n$, $\hat{\ }_{n;m}$
- Central limit theorem for $\hat{\ }_n$: p $\overline{\text{n}}$ convergence rate for any dimension

Thank you!

Our Poster: East Exhibition Hall B + C #226



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

Kimia Nadjahi[1], Alain Durmus[2], Umut “im³ekli [1;3], Roland Badeau[1]
{kimia.nadjahi,umut.simsekli,roland.badeau}@telecom-paris.fr , alain.durmus@cmla.ens-cachan.fr
1: LTCI, Télécom Paris, Institut Polytechnique de Paris 2: CMLA, ENS Paris-Saclay 3: Department of Statistics, University of Oxford

Minimum Distance Estimation

Observations $Y_{1:n}$ = (Y_1; : : : ; Y_n), Y_i 2 Y R^d, i.i.d. from $_?$ 2 P (Y), with P(Y) : set of probability measures onY.

A family of distributions on Y parameterized by 2 R^d : M = f 2 P (Y); 2 g.

Purely generative models: We can generatem 2 N i.i.d. samples from , but the likelihood is intractable. ^ $_{;m}$ is the empirical distribution.

Given $Y_{1:n}$, its empirical distribution ^$_n$ and a distanceD on P(Y), we perform Minimum Distance Estimation (MDE) :

^$_n$ = argmin $_2$ D(^$_n$;) (1)

or Minimum Expected Distance Estimation (MEDE) :

^$_{n;m}$ = argmin $_2$ E[D(^$_n$; ^ $_{;m}$)jY$_{1:n}$] (2)

Optimal Transport (OT) Metrics

For p 1, P$_p$(Y) : set of probability measures onY with nite p'th moment. Let ; 2 P$_p$(Y).

Wasserstein distance (W$_p$). Computationally expensive, except in 1d (Y R) ! analytical form.

Sliced-Wasserstein (SW) distance .

[redacted]

S^d ^{1} : d-dimensional unit sphere, : uniform distribution on S^d ^{1}.

Practical metric based on projections: 8u 2 S^d ^{1}; y 2 Y; u$^?$(y) = hu; yi

SW^p_p(;) = R S^d $_1$ W^p_p(u$^?_]$; u$^?_]$)d (u)

Combining MDE and OT

Minimum Wasserstein estimators, de ned in (1) and (2) with D = W$_p$, have asymptotic guarantees [1] but are not practical.

) With D = SW$_p$ in (1) and (2), we get the minimum (expected) SW estimators M(E)SWE of order p.

Recent studies show the empirical success of SW-based estimators on generative modeling but lack of theoretical guarantees.

) We investigate the asymptotic properties of these estimators.

Theoretical Results

The convergence inSW$_p$ implies the weak convergence inP(R^d).

Key assumptions.

Continuity: For any ($_n$)$_{n2N}$ in such that lim$_{n!+1}$ ($_n$;) = 0 ,
A1. ($_n$)$_{n2N}$ converges weakly!^{w}( ) to .
A2. lim$_{n!+1}$ E[SW$_p$($_n$; ^ $_{n;n}$)jY$_{1:n}$] = 0 .

Data-generating process:
A3. lim$_{n!+1}$ SW$_p$(^$_n$; $_?$) = 0 , P-almost surely.

Bounded sets: For some > 0,
A4. $^?$ = f 2 : SW$_p$($_?$;) $_?$ + g, with $_?$ = inf $_2$ SW$_p$($_?$;), is bounded.
A5. $_{;n}$ = f 2 : SW$_p$(^$_n$;) $_n$ + g, with $_n$ = inf $_2$ SW$_p$(^$_n$;), is bounded almost surely.

Existence and consistency of MSWE

Assume A1, A3, A4. Then, there existsE with P(E) = 1 such that, for all ! 2 E,

lim$_{n!+1}$ inf$_2$ SW$_p$(^$_n$(!);) = inf$_2$ SW$_p$($_?$;);

lim sup$_{n!+1}$ argmin $_2$ SW$_p$(^$_n$(!);) argmin $_2$ SW$_p$($_?$;)

Besides, for all! 2 E, there existsn(!) such that, for all n n(!), argmin $_2$ SW$_p$(^$_n$(!);) is non-empty.

Guarantees for MESWE. Existence and consistency (with A1 to A4), convergence to MSWE as m ! 1 (A1, A2, A5).

Central limit theorem for MSWE with p = 1

Consider A1, A3, A4, $_?$ = , (with $_?$ 2 well-separated) and H : 7! R S^d $_1$ R R jG$_?$(u; t) h ; D$_?$(u; t)ij dtd (u), with

p $\overline{n}$($\hat{F}_n$ F $_,$)!^{w} G$_?$, where $\hat{F}_n$ and F $_,$ contain the CDFs of the projected ^$_n$ and $_?$

D$_?$(u;) : the derivative of F (u;) in $_?$

Then, p $\overline{n}$ inf$_2$ SW$_1$(^$_n$;)!^{w} inf$_2$ H( );
p $\overline{n}$(^$_n$ $_?$)!^{w} argmin $_2$ H( ); as n ! +1

) Convergence rate of p $\overline{n}$ independent of the dimension

Numerical Experiments

Multivariate Gaussians.

M = N(m; ^{2}I) : m 2 R^{10}; ^{2} > 0 , and (m$_?$; $^2_?$) = (0; 1).

MSWE vs. n MESWE vs. n = m MESWE, n = 2000 vs. m

Multivariate elliptically contoured stable distributions.
M = E S$_c$(I ; m) : m 2 R^d with = 1:8, and m$_?$ = 2.

Comparison Wasserstein and SW MESWE MESWE, n = 100

High-dimensional real data.

We train the Sliced-Wasserstein Generator [2] (based on MESWE), on MNIST.

We plot the mean-squared error between the training/test loss obtained for (n; m) from (1; 1) to (10 000; 60) and for (n ; m) = (60 000; 200).

Main References

[1] E. Bernton, P. E. Jacob, M. Gerber, C. P. Robert. On parameter estimation with the Wasserstein distance. Information and Inference: A Journal of the IMA, Jan 2019.

[2] I. Deshpande, Z. Zhang, A. G. Schwing. Generative modeling using the sliced Wasserstein distance. CVPR 2018.