Sharp rates of convergence of empirical measures in Wasserstein distance

Francis Bach
INRIA - Ecole Normale Supérieure

Joint work with Jonathan Weed (MIT)
NIPS Workshop, December 2017
Wasserstein distances between distributions

- Comparing probability measures supported on a metric space

Statistical Models

Empirical Measures, i.e. data

Bags of features

Brain Activation Maps

Color Histograms

(courtesy of Marco Cuturi)
Wasserstein distances between distributions

• Comparing probability measures supported on a metric space

• Low-dimensional
  – Images, signals
  – See, e.g., Rubner et al. (2000); Solomon et al. (2015); Sandler and Lindenbaum (2011)

• High-dimensional
  – Text (see, e.g., Kusner et al., 2015; Zhang et al., 2016)
  – Statistical models (Arjovsky et al., 2017; Genevay et al., 2017)
  – Empirical measures
Wasserstein distances between distributions

• Comparing probability measures supported on a metric space

• Low-dimensional
  – Images, signals
  – See, e.g., Rubner et al. (2000); Solomon et al. (2015); Sandler and Lindenbaum (2011)

• High-dimensional
  – Text (see, e.g., Kusner et al., 2015; Zhang et al., 2016)
  – Statistical models (Arjovsky et al., 2017; Genevay et al., 2017)
  – Empirical measures

• Does it make sense to compute Wasserstein distances from samples in high dimension?
\[ W_p(\mu, \nu) := \inf_{\gamma \in \mathcal{C}(\mu, \nu)} \left( \int D(x, y)^p d\gamma(x, y) \right)^{1/p} \]

- **Wasserstein distance** of order \( p \in [1, \infty) \) between \( \mu \) and \( \nu \) on a metric space \((X, D)\)
  - \( \mathcal{C}(\mu, \nu) = couplings \gamma \) of \( \mu \) and \( \nu \) = distributions on \( X \times X \) whose first and second marginals agree with \( \mu \) and \( \nu \)
  - Metric on probability measures on \( X \) (see Santambrogio, 2015)
\[ W_p(\mu, \nu) := \inf_{\gamma \in \mathcal{C}(\mu, \nu)} \left( \int D(x, y)^p d\gamma(x, y) \right)^{1/p} \]

- **Wasserstein distance** of order \( p \in [1, \infty) \) between \( \mu \) and \( \nu \) on a metric space \((X, D)\)
  - \( \mathcal{C}(\mu, \nu) = couplings \gamma \) of \( \mu \) and \( \nu \) = distributions on \( X \times X \) whose first and second marginals agree with \( \mu \) and \( \nu \)
  - Metric on probability measures on \( X \) (see Santambrogio, 2015)

- **Estimation from samples**
  - \( \hat{\mu}_n, \hat{\nu}_n \) empirical distribution obtained from \( n \) i.i.d. samples of \( \mu, \nu \)
  
  \[
  \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta(x_i)
  \]
\[ W_p(\mu, \nu) := \inf_{\gamma \in \mathcal{C}(\mu, \nu)} \left( \int D(x, y)^p \, d\gamma(x, y) \right)^{1/p} \]

- **Wasserstein distance** of order \( p \in [1, \infty) \) between \( \mu \) and \( \nu \) on a metric space \((X, D)\)
  
  - \( \mathcal{C}(\mu, \nu) = couplings \ \gamma \) of \( \mu \) and \( \nu \) = distributions on \( X \times X \) whose first and second marginals agree with \( \mu \) and \( \nu \)
  
  - Metric on probability measures on \( X \) (see Santambrogio, 2015)

- **Estimation from samples**
  
  - \( \hat{\mu}_n, \hat{\nu}_n \) empirical distribution obtained from \( n \) i.i.d. samples of \( \mu, \nu \)
    
  \[ \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta(x_i) \]

- **Approximation:** \( |W_p(\mu, \nu) - W_p(\hat{\mu}_n, \hat{\nu}_n)| \leq W_p(\mu, \hat{\mu}_n) + W_p(\nu, \hat{\nu}_n) \)
Known properties of $W(\mu, \hat{\mu}_n)$

- **Convergence** for any $p \in [1, \infty)$: $W_p(\mu, \hat{\mu}_n) \to 0 \mu$-a.s.
  - For $X$ compact, separable and $\mu$ any Borel measure (Villani, 2008)
Known properties of $W(\mu, \hat{\mu}_n)$

- **Convergence** for any $p \in [1, \infty)$: $W_p(\mu, \hat{\mu}_n) \to 0$ $\mu$-a.s.
  
  - For $X$ compact, separable and $\mu$ any Borel measure (Villani, 2008)

- **Rates of approximation by distributions of discrete support**
  
  - Information theory (Cover and Thomas, 2012)
  - Machine learning (Cañas and Rosasco, 2012)
Known properties of $W(\mu, \hat{\mu}_n)$

- **Convergence** for any $p \in [1, \infty)$:
  \[ W_p(\mu, \hat{\mu}_n) \to 0 \text{ } \mu\text{-a.s.} \]
  - For $X$ compact, separable and $\mu$ any Borel measure (Villani, 2008)

- **Rates of approximation by distributions of discrete support**
  - Information theory (Cover and Thomas, 2012)
  - Machine learning (Cañas and Rosasco, 2012)

- **Curse of dimensionality** (Dudley, 1968)
  - $\mu$ absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^d$:
    \[ \mathbb{E}[W_1(\mu, \hat{\mu}_n)] \gtrsim n^{-1/d} \]
    - Lower bound asymptotically tight when $d > 2$
    - Sharper results (see, e.g., Dobrić and Yukich, 1995)
Sharp asymptotic and finite-sample rates (Weed and Bach, 2017)

- Beyond measures with densities?
  - Adaptivity to low-dimensional structures
Sharp asymptotic and finite-sample rates
(Weed and Bach, 2017)

- Beyond measures with densities?
  - Adaptivity to low-dimensional structures

- Sharper finite-sample (i.e., non-asymptotic) rates?
  - Multi-scale behavior
Sharp asymptotic and finite-sample rates
(Weed and Bach, 2017)

• Beyond measures with densities?
  – Adaptivity to low-dimensional structures

• Sharper finite-sample (i.e., non-asymptotic) rates?
  – Multi-scale behavior

• Unified theoretical framework and explicit constants for all $p$

• Analysis of $\mathbb{E}[W_p(\mu, \hat{\mu}_n)] +$ new concentration inequality
**Assumptions**

- **Basic assumptions**
  - The metric space $X$ is Polish, and all measures are Borel
  - $\text{diam}(X) \leq 1$

- **Dyadic partition assumption** with parameter $\delta < 1$ (David, 1988)
  - Sequence $\{Q^k\}_{1 \leq k \leq k^*}$ with $Q^k \subseteq B(X)$ such that:
    - (a) the sets in $Q^k$ form a partition of $X$ and have diameters $\leq \delta^k$
    - (b) the $(k + 1)$th partition is a refinement of the $k$th partition
  - Main example: $X = [0, 1]^d$ with the $\ell_\infty$ metric
Assumptions

- **Basic assumptions**
  - The metric space \(X\) is Polish, and all measures are Borel
  - \(\text{diam}(X) \leq 1\)

- **Dyadic partition assumption** with parameter \(\delta < 1\) (David, 1988)
  - Sequence \(\{Q^k\}_{1 \leq k \leq k^*}\) with \(Q^k \subseteq B(X)\) such that:
    (a) the sets in \(Q^k\) form a partition of \(X\) and have diameters \(\leq \delta^k\)
    (b) the \((k+1)\)th partition is a refinement of the \(k\)th partition
  - Main example: \(X = [0, 1]^d\) with the \(\ell_\infty\) metric

- **Alternative definitions**
  - \(W_p(\mu, \nu) = \inf_{\gamma \in \mathcal{C}(\mu, \nu)} \left( \int D(x, y)^p d\gamma(x, y) \right)^{1/p}\)
  - \(W_1(\mu, \nu) = \sup_{f \in \text{Lip}(X)} \left| \int f d\mu - \int f d\nu \right|\) where the supremum is taken over all 1-Lipschitz functions on \(X\)
Related work

• Inherent dimension of the measure on any metric space
  – Dudley (1968): $O(n^{-1/d})$ rate with covering numbers of the support of $\mu$, using Lipschitz-function representation (for $p = 1$)
  – Boissard and Le Gouic (2014): extension to $p > 1$, not tight

• Explicit couplings on $\mathbb{R}^d$
  – Tight for measures with densities
  – Fournier and Guillin (2015); Dereich et al. (2013)

• Tail bounds
  – Direct (Boissard, 2011; Bolley, Guillin, and Villani, 2007)
  – Indirect (Boissard and Le Gouic, 2014)
Describing low-dimensional structures

- Many possible notions of dimensions (Hausdorff, Minkowski, etc.)
  
  - \( \varepsilon \)-covering number of \( S \subseteq X \): \( N_\varepsilon(S) \) = minimum \( m \) such that there exists \( m \) closed balls \( B_1, \ldots, B_m \) of diameter \( \varepsilon \) such that \( S \subseteq \bigcup_{1 \leq i \leq m} B_i \)
Describing low-dimensional structures

- Many possible notions of dimensions (Hausdorff, Minkowski, etc.)
  - $\varepsilon$-covering number of $S \subseteq X$: $\mathcal{N}_\varepsilon(S) =$ minimum $m$ such that there exists $m$ closed balls $B_1, \ldots, B_m$ of diameter $\varepsilon$ such that $S \subseteq \bigcup_{1 \leq i \leq m} B_i$
  - $\varepsilon$-dimension of $S$ equal to $d_\varepsilon(S) := \frac{\log \mathcal{N}_\varepsilon(S)}{\log(1/\varepsilon)}$
  - Minkowski’s dimension $\dim_M(S) := \limsup_{\varepsilon \to 0} d_\varepsilon(S)$

$$\mathcal{N}_\varepsilon(S) \approx C\varepsilon^{-d}$$
Describing low-dimensional structures

• Many possible notions of dimensions (Hausdorff, Minkowski, etc.)
  
  – $\varepsilon$-covering number of $S \subseteq X$: $N_{\varepsilon}(S) =$ minimum $m$ such that there exists $m$ closed balls $B_1, \ldots, B_m$ of diameter $\varepsilon$ such that $S \subseteq \bigcup_{1 \leq i \leq m} B_i$
  
  – $\varepsilon$-dimension of $S$ equal to $d_{\varepsilon}(S) := \frac{\log N_{\varepsilon}(S)}{\log(1/\varepsilon)}$
  
  – Minkowski’s dimension $\dim_M(S) := \limsup_{\varepsilon \to 0} d_{\varepsilon}(S)$

• Regular sets of dimension $d$ (Graf and Luschgy, 2007)
  
  – Nonempty, compact convex sets in dimension $d$
  
  – Relative boundaries of nonempty, compact convex sets of dimension $d + 1$
  
  – Compact $d$-dimensional differentiable manifolds
  
  – Self-similar sets with similarity dimension $d$
Lower and upper bounds (Weed and Bach, 2017)

- **Theorem:** Let $p \in [1, \infty)$. If $s > d_p^*(\mu)$, then

\[
\mathbb{E}[W_p(\mu, \hat{\mu}_n)] \lesssim n^{-1/s}
\]

If $t < d_*(\mu)$, then

\[
W_p(\mu, \hat{\mu}_n) \gtrsim n^{-1/t}
\]

- Extended notions of dimensions $d_p^*(\mu)$ and $d_*(\mu)$, equal to $\dim_M(\text{supp(\mu)})$ for regular supports
- Refinements based on covering all but a low-mass set, needed for sharpest bound valid for all $p$
- Precise results with explicit constants
- NB: lower bound holds for any discrete measure on $n$ points
Proof idea for upper bound

- Construct explicit coupling between $\mu$ and $\hat{\mu}$
- Moving mass between elements of the partition to correct for unequal mass
- Moving mass within elements of partition
- Done recursively
Finite-sample bounds and multiscale behavior

- Single dimension not enough to characterize behavior

- Real datasets typically exhibit structures at multiple scales
  - See Little, Maggioni, and Rosasco (2016)
Finite-sample bounds and multiscale behavior

- Single dimension not enough to characterize behavior
Finite-sample bounds and multiscale behavior

• Single dimension not enough to characterize behavior

• Previous result
  
  \[ \eta_n = W_p(\mu, \hat{\mu}_n) \approx n^{-1/d} = \exp\left(-\frac{\log n}{d}\right) \]
  
  \[ \eta_n = W_p(\mu, \hat{\mu}_n) \approx n^{-1/d} = \lim_{\varepsilon \to 0} \exp\left(-\log\left(\frac{1}{\varepsilon}\right)\frac{\log n}{\log N_\varepsilon(X)}\right) \]

  Choosing \( \varepsilon \) such that \( n \approx N_\varepsilon(X) \) leads to \( \eta_n = \varepsilon \)
Finite-sample bounds and multiscale behavior

• Single dimension not enough to characterize behavior

• Previous result
  - $\eta_n = W_p(\mu, \hat{\mu}_n) \approx n^{-1/d} = \exp\left(-\frac{\log n}{d}\right)$
  - $\eta_n = W_p(\mu, \hat{\mu}_n) \approx n^{-1/d} = \lim_{\varepsilon \to 0} \exp\left(-\log(1/\varepsilon) \frac{\log n}{\log N_\varepsilon(X)}\right)$
  - Choosing $\varepsilon$ such that $n \approx N_\varepsilon(X)$ leads to $\eta_n = \varepsilon$

• “Proposition”: for $p \in [1, \infty)$, let $d_n = \frac{\log N_{\varepsilon_n}(X)}{\log(1/\varepsilon_n)}$, with $\varepsilon_n$ so that $N_{\varepsilon_n}(X) \approx n$. If $d_n > 2p$, then

$$\mathbb{E}[W_p(\mu, \hat{\mu}_n)] \lesssim n^{-1/d_n},$$
Finite-sample bounds and multiscale behavior

• Single dimension not enough to characterize behavior

• Previous result
  - $\eta_n = W_p(\mu, \hat{\mu}_n) \approx n^{-1/d} = \exp\left(-\frac{\log n}{d}\right)$
  - $\eta_n = W_p(\mu, \hat{\mu}_n) \approx n^{-1/d} = \lim_{\epsilon \to 0} \exp\left(-\log(1/\epsilon) \frac{\log n}{\log N_\epsilon(X)}\right)$
  - Choosing $\epsilon$ such that $n \approx N_\epsilon(X)$ leads to $\eta_n = \epsilon$

• “Proposition”: for $p \in [1, \infty)$, let $d_n = \frac{\log N_{\epsilon_n}(X)}{\log(1/\epsilon_n)}$, with $\epsilon_n$ so that $N_{\epsilon_n}(X) \approx n$. If $d_n > 2p$, then
  $$\mathbb{E}[W_p(\mu, \hat{\mu}_n)] \lesssim n^{-1/d_n},$$

• “Proposition”: All reasonable sequences $d_n$ can be achieved by a certain density
Clusterable distributions

- **Definition**: A distribution \( \mu \) is \((m, \Delta)\)-clusterable if \( \text{supp}(\mu) \) lies in the union of \( m \) balls of radius at most \( \Delta \).
Clusterable distributions

- **Definition**: A distribution $\mu$ is $(m, \Delta)$-clusterable if $\text{supp}(\mu)$ lies in the union of $m$ balls of radius at most $\Delta$.

- **Proposition**: If $\mu$ is $(m, \Delta)$ clusterable, then for all $n \leq m(2\Delta)^{-2p}$,

  $$\mathbb{E}[W_p^p(\mu, \hat{\mu}_n)] \lesssim \sqrt{\frac{m}{n}}$$

  - Usual bound still holds $\mathbb{E}[W_p^p(\mu, \hat{\mu}_n)] \lesssim n^{-p/d}$ for all $n$
  - Extension to Gaussian mixtures

- **Extension to approximately low-dimensional sets**

  - Initial convergence at the rate of the low-dimensional set
Concentration

• **Previous work:** Bolley et al. (2007); Boissard (2011) obtain tail bounds of the form

\[ \mathbb{P}[W_p^p(\mu, \hat{\mu}_n) \geq t] \leq \psi_n(t) \]

where \( \psi_n(t) \) has sub-Gaussian subgaussian decay, with unclear dependence on ambient dimension

– Two-step approach by Boissard and Le Gouic (2014) with different tools
Concentration

- **Previous work:** Bolley et al. (2007); Boissard (2011) obtain tail bounds of the form
  \[ \mathbb{P}[W_p^p(\mu, \hat{\mu}_n) \geq t] \leq \psi_n(t) \]
  where \( \psi_n(t) \) has sub-Gaussian subgaussian decay, with unclear dependence on ambient dimension
  - Two-step approach by Boissard and Le Gouic (2014) with different tools

- **Simple new result:** For all \( n \geq 0 \) and \( 0 \leq p < \infty \),
  \[ \mathbb{P}[W_p^p(\mu, \hat{\mu}_n) \geq \mathbb{E}W_p^p(\mu, \hat{\mu}_n) + t] \leq \exp(-2nt^2) \]
  - Consequence of Mac Diarmid inequality
  - Concentration phenomenon
“Applications”

- Quadrature
  
  From the representation $W_1(\mu, \nu) = \sup_{f \in \text{Lip}(X)} \left| \int f \, d\mu - \int f \, d\nu \right|$: 

  $\mathbb{E} \sup_{f \in \text{Lip}(X)} \left| \int f(x) \, d\mu(x) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right| \lesssim n^{-1/d}$
“Applications”

- **Quadrature**
  - From the representation \( W_1(\mu, \nu) = \sup_{f \in \text{Lip}(X)} \left| \int f \, d\mu - \int f \, d\nu \right| : \)
    \[
    \mathbb{E} \sup_{f \in \text{Lip}(X)} \left| \int f(x) \, d\mu(x) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right| \lesssim n^{-1/d}
    \]

- **\( k \)-means clustering** (Cañas and Rosasco, 2012)
  - Approximation of distributions by finitely supported distributions
  - Equivalence to approximation with \( W_2 \)
  - Consequence: approximation by empirical measure asymptotically optimal with explicit bounds for regular supports
Conclusion and Future Work

• Summary
  – Sharper / explicit rates for the convergence of $W_p(\hat{\mu}_n, \mu)$
  – Both in asymptotic and finite-sample settings
  – Adaptivity to low-dimensional structures, otherwise exponentially slow convergence
Conclusion and Future Work

• Summary
  – Sharper / explicit rates for the convergence of $W_p(\mu_n, \mu)$
  – Both in asymptotic and finite-sample settings
  – Adaptivity to low-dimensional structures, otherwise exponentially slow convergence

• Extensions
  – Wasserstein distance with entropic penalty (Cuturi, 2013; Solomon et al., 2015; Carlier et al., 2017; Rolet et al., 2016)
    \[
    \inf_{\gamma \in \mathcal{C}(\mu, \nu)} \int D(x, y)^p d\gamma(x, y) + \lambda \int \log \frac{d\gamma(x,y)}{d\gamma_0(x,y)} d\gamma(x,y)
    \]
  – Link with stochastic optimization (Genevay, Cuturi, Peyré, and Bach, 2016) for directly computing $W(\mu, \nu)$ in a single pass
  – Importance sampling
References


