# Statistical OT Lecture 9: Dyadic Partitioning and the AKT theorem

## 1 Dyadic Partitioning

One drawback of the previous approach is that it relies strongly on the dual representation of $W_1$ – though the technique can be generalized to work for other $W_p$ distances, the connections to empirical process theory are less transparent.

**Note:** (See assignment.) I'll note in passing though that dual bounds are extremely useful in proving what are called *lower complexity adaptation* results. Roughly, suppose that we get samples $X_1, \ldots, X_n \sim \mu$, and $Y_1, \ldots, Y_n \sim \nu$ and we want to estimate the 1-Wasserstein distance. A natural estimate is $W_1(\mu_n, \nu_n)$. It turns out that the rate of convergence of $W_1(\mu_n, \nu_n)$ to $W_1(\mu, \nu)$ depends only on the intrinsic dimensionality of the simpler of the two measures (i.e. if either of these measures are supported on a lower dimensional set then we get fast rates). This phenomenon is called lower complexity adaptation, and the methods to show that this happens usually goes through the dual bounds we developed in the previous lecture.

Let us return to our goal of trying to illustrate a different proof of the convergence of $W_1(\mu_n, \mu)$ to 0. A different approach to upper bounding $W_1$ is to explicitly construct a coupling between these measures, and to evaluate its cost. These explicit constructions use an idea called dyadic partitioning.

### 1.1 Intuition and Dyadic Partitioning Upper Bound

First, we note the obvious bound that:

$$W_1(\mu, \nu) \leq \sqrt{d}.$$

Suppose we define the dyadic partition of the cube to be $\mathcal{Q}_1$ (with side length 1/2). Then we can see that, if $\mu(Q) = \nu(Q)$ for all $Q \in \mathcal{Q}_1$ then we only need to move mass within the smaller cubes so we get:

$$W_1(\mu, \nu) \leq \frac{\sqrt{d}}{2}.$$

If $\mu(Q) \neq \nu(Q)$ then we simply need to move the extra $\mu$ mass from cubes where $\mu(Q) > \nu(Q)$ to other cubes. The total amount of mass we need to move outside the small cubes is at most:

$$\Delta_1 = \sum_{Q \in \mathcal{Q}_1} (\mu(Q) - \nu(Q))_+ = \frac{1}{2} \sum_{Q \in \mathcal{Q}_1} |\mu(Q) - \nu(Q)|.$$

Putting these together we get a refined bound:

$$W_1(\mu, \nu) \leq \sqrt{d} \times \Delta_1 + \frac{\sqrt{d}}{2}.$$

Recursing this argument we obtain the dyadic partitioning upper bound. Concretely, define $\mathcal{Q}_j$ to be the dyadic partition of the cube with side length $2^{-j}$ and define:

$$\Delta_j = \frac{1}{2} \sum_{Q \in \mathcal{Q}_j} |\mu(Q) - \nu(Q)|,$$

then we have the following theorem:

**Theorem 1.** For any $\mu, \nu$ supported on the unit cube, for any $J \geq 0$,

$$W_1(\mu, \nu) \leq \sqrt{d} \sum_{j=0}^{J-1} 2^{-j} \Delta_{j+1} + \sqrt{d} 2^{-J}.$$

One can find a more formal proof of this theorem in various places (it is notationally quite heavy even though the intuition is simple). Taking this as given we can now give an alternate proof of the rates of convergence of the empirical measure in the $W_1$ distance. Note, at various points the similarities between this calculation and the chaining calculation we did earlier.

As a direct consequence of Theorem 1 we have that:

$$\begin{aligned}
\mathbb{E} W_1(\mu_n, \mu) &\leq \frac{\sqrt{d}}{2} \sum_{j=0}^{J-1} 2^{-j} \sum_{Q \in \mathcal{Q}_{j+1}} \mathbb{E}|\mu_n(Q) - \mu(Q)| + \sqrt{d} 2^{-J} \\
&\leq \frac{\sqrt{d}}{2} \sum_{j=0}^{J-1} 2^{-j} 2^{d(j+1)/2} \left( \sum_{Q \in \mathcal{Q}_{j+1}} [\mathbb{E}|\mu_n(Q) - \mu(Q)|]^2 \right)^{1/2} + \sqrt{d} 2^{-J} \\
&\leq \frac{\sqrt{d}}{2} \sum_{j=0}^{J-1} 2^{-j} 2^{d(j+1)/2} \left( \sum_{Q \in \mathcal{Q}_{j+1}} \mathbb{E}|\mu_n(Q) - \mu(Q)|^2 \right)^{1/2} + \sqrt{d} 2^{-J},
\end{aligned}$$

where both inequalities use Cauchy-Schwarz. Now, $\mathbb{E}|\mu_n(Q) - \mu(Q)|^2 \leq \mu(Q)/n$ (this is simply a Binomial variance), so we have that,

$$\mathbb{E} W_1(\mu_n, \mu) \leq \frac{\sqrt{d}}{2\sqrt{n}} \sum_{j=0}^{J-1} 2^{-j} 2^{d(j+1)/2} + \sqrt{d} 2^{-J}.$$

When $d = 1$ we can choose $J = \infty$, and obtain that $\mathbb{E} W_1(\mu_n, \mu) \lesssim 1/\sqrt{n}$. When $d = 2$, choose $J = \log n$, and we obtain that,

$$\mathbb{E} W_1(\mu_n, \mu) \lesssim \frac{\log n}{\sqrt{n}},$$

and finally for $d \geq 3$ we take $(J+1) = \lceil \log n^{1/d} \rceil$, and obtain that $\mathbb{E} W_1(\mu_n, \mu) \lesssim \sqrt{d} n^{-1/d}$.

## 2 The Case when $d = 2$

In this section, we will follow the paper of Bobkov-Ledoux (A simple Fourier analytic...) to give a sharp bound when $d = 2$. For historical context, the original paper that proved this result is quite famous in applied probability (it is known as the AKT theorem after the authors). Talagrand also wrote papers, developing and applying the "generic chaining" method to this case. The Fourier analytic proof is quite simple/elegant and will give us another way to think about Wasserstein distances (roughly, as "inverse Sobolev norms") that will also be useful to understand the case when the measures are smooth.

To apply Fourier analytic techniques we will first define a variant of the $W_1$ distance:

$$\widetilde{W}_1(\mu, \nu) = \sup_{f \in \widetilde{\mathrm{Lip}}_1} \int f d\mu - \int f d\nu,$$

where $\widetilde{\mathrm{Lip}}_1$ are the set of 1-Lipschitz, $C^\infty$ functions, which are $2\pi$ periodic on $\mathbb{R}^d$. It is obvious that, $\widetilde{W}_1(\mu, \nu) \leq W_1(\mu, \nu)$ since we have added some restrictions to the test functions. It turns out that for measures supported on $[0, 1]^d$ they in fact coincide:

**Lemma 1.** Suppose $\mu, \nu$ have support in $[0,1]^d$ then:

$$W_1(\mu,\nu) = \widetilde{W}_1(\mu,\nu).$$

*Proof.* As we discussed above $\widetilde{W}_1$ adds some restrictions to the test functions, and we need to argue that they don't change the distance. Intuitively, you should imagine that if I give you a function that is defined on $[0,1]^d$ I can extend it Lipschitzly to $[0,2\pi]^d$. Concretely, consider:

$$\widetilde{f}(y) = \sup_{x\in[0,1]^d} \{f(x) - d_{\mathbb{T}^d}(x,y)\},$$

where $d_{\mathbb{T}^d}(x,y) = \inf_{z\in\mathbb{Z}^d} \|x-y-2\pi z\|$ is the torus distance (i.e. the distance on $[0,2\pi]^d$ with ends identified). For any $x \in [0,1]^d$ the function $g(y) = f(x) - d_{\mathbb{T}^d}(x,y)$ is $2\pi$-periodic, and Lipschitz, so $\widetilde{f}$ is also $2\pi$-periodic and Lipschitz.

Now, on $[0,1]^d$, $\widetilde{f}(y) \geq f(y)$, (since we can choose $x = y$ in the supremum), and on $[0,1]^d$ we know that $d_{\mathbb{T}^d}(x,y) = \|x-y\|$, so we have that,

$$f(x) - d_{\mathbb{T}^d}(x,y) = f(x) - \|x-y\| \leq f(y),$$

so we obtain that $f(y) \geq \widetilde{f}(y)$. Thus, the integral of $\mu - \nu$ (which is supported inside $[0,1]^d$) against $f$ is identical to the integral against $\widetilde{f}$.

The restriction to $C^\infty$ functions does not change anything – we can always mollify a Lipschitz function (for instance, convolve it with a "mollifier" and let its bandwidth go to 0).

□

## 2.1 Fourier Analysis

For a measure $\mu$ we can define its characteristic function:

$$\phi_\mu(m) = \mathbb{E}\exp(im^T X), \quad m \in \mathbb{Z}^d.$$

Then we can show that:

**Lemma 2.**

$$\widetilde{W}_1(\mu,\nu)^2 \leq \sum_{m\in\mathbb{Z}^d\setminus\{0\}} \frac{1}{\|m\|^2}|\phi_\mu(m) - \phi_\nu(m)|^2.$$

This representation is the key. It is similar in some ways to the dyadic partitioning bound – we are computing how much the distributions differ at higher and higher frequencies (just like in the dyadic partitioning), and then downweighting the higher frequency discrepancies, and putting it together. It will turn out however that this bound is sharper when $d = 2$. For some heuristic background that is maybe more familiar to statisticians, it is useful to think about the 1D case (everything carries over to the general case). Given a (square integrable) function, we can imagine writing it in a nice orthonormal basis (like the Fourier basis):

$$f = \sum_{j=1}^{\infty} \theta_j \phi_j,$$

where $\theta_j$ are some coefficients, and $\phi_j$ are orthonormal basis functions. Now, we could imagine smoothness (say Hölder or Sobolev smoothness), as restrictions on the coefficients. Roughly, a $s$-Hölder function will have coefficients which satisfy:

$$\sum_{j=1}^{\infty} j^{2s}\theta_j^2 \leq M < \infty,$$

i.e. the coefficients "decay" so there is very little contribution from higher-order terms. Now, suppose I took the measures $\mu, \nu$ and suppose they had densities (abusing notation call them $\mu, \nu$), then we could expand them in the basis to see that,

$$\int f(d\mu - d\nu) = \sum_{j=1}^{\infty} (\alpha_j - \beta_j)\theta_j,$$

where $\alpha_j$ and $\beta_j$ are the coefficients of $\mu, \nu$ in our basis. You could now solve the problem:

$$\sup_{\theta: \sum_{j=1}^{\infty} j^{2s}\theta_j^2 \leq M} \sum_{j=1}^{\infty} (\alpha_j - \beta_j)\theta_j.$$

When $s = 1$, this would correspond to $W_1$, and you could explicitly solve the above maximization to obtain that,

$$W_1(\mu, \nu) \leq \sqrt{\sum_{j=1}^{\infty} (\alpha_j - \beta_j)^2 j^{-2s}}.$$

This is sometimes called an "inverse" Sobolev norm – in our case it is just the Lipschitz IPM – but such expressions are also true of other $W_p$ distances under some assumptions. It is saying that $W_1$ roughly corresponds to measuring the difference in a basis (like the Fourier basis), but rather than do so with the usual $L_2$-norm we now down-weigh the higher-frequency deviations.

*Proof.* The proof roughly makes some of the intuition above precise. Concretely, given a function $f$ which is $C^\infty$ and $2\pi$ periodic we can write it in terms of its Fourier series:

$$f(x) = \sum_{m \in \mathbb{Z}^d} \widehat{f}(m) \exp(im^T x).$$

We can differentiate through this expression (the coefficients converge to 0 super-algebraically), and apply Parseval's identity to see that:

$$\frac{1}{(2\pi)^d} \int_{[0,2\pi]^d} (\partial_i f(x))^2 dx = \sum_{m \in \mathbb{Z}^d} m_i^2 |\widehat{f}(m)|^2,$$

and summing we obtain that,

$$\frac{1}{(2\pi)^d} \int_{[0,2\pi]^d} \|\nabla f\|^2 dx = \sum_{m \in \mathbb{Z}^d} \|m\|^2 |\widehat{f}(m)|^2.$$

For a 1-Lipschitz, $C^\infty$ function, $\|\nabla f(x)\| \leq 1$, so we obtain that,

$$\sum_{m \in \mathbb{Z}^d} \|m\|^2 |\widehat{f}(m)|^2 \leq 1.$$

Continuing we obtain that,

$$\int f(d\mu - d\nu) = \sum_{m \in \mathbb{Z}^d} \widehat{f}(m)(\phi_\mu(m) - \phi_\nu(m)),$$

and we know that $\phi_\mu(0) = \phi_\nu(0)$, so applying Cauchy-Schwarz we obtain that,

$$\int f(d\mu - d\nu) \leq \sqrt{\sum_{m \in \mathbb{Z}^d \setminus \{0\}} \frac{1}{\|m\|^2} |\phi_\mu(m) - \phi_\nu(m)|^2}.$$

□

## 2.2 Convolution Smoothing

It will turn out that for the empirical measure the upper bound from Lemma 2 diverges (you will see hints of why in the sequel), so we need to instead upper bound the distance between a smoothed empirical measure and the truth using the Lemma, and then bridge the gap.

Let $\varphi_\varepsilon$ denote a multivariate Gaussian with variance $\epsilon I_d$. Then we have the following result:

**Lemma 3.** For any $\varepsilon > 0$

$$\widetilde{W}_1(\mu, \nu) \leq \widetilde{W}_1(\mu, \nu \star \varphi_\varepsilon) + \sqrt{d\varepsilon}.$$

*Proof.* $\widetilde{W}_1$ satisfies the triangle inequality so we obtain that,

$$\widetilde{W}_1(\mu, \nu) \leq \widetilde{W}_1(\mu, \nu \star \varphi_\varepsilon) + \widetilde{W}_1(\nu, \nu \star \varphi_\varepsilon),$$

and since $\widetilde{W}_1 \leq W_1$ we have,

$$\widetilde{W}_1(\mu, \nu) \leq \widetilde{W}_1(\mu, \nu \star \varphi_\varepsilon) + W_1(\nu, \nu \star \varphi_\varepsilon).$$

Now, we can construct a coupling of these distributions by sampling $X \sim \nu$ and setting $Y = X + \sqrt{\varepsilon}Z$, where $Z$ is an independent standard Gaussian, to obtain that:

$$\widetilde{W}_1(\mu, \nu) \leq \widetilde{W}_1(\mu, \nu \star \varphi_\varepsilon) + \sqrt{\varepsilon}\mathbb{E}\|Z\|,$$

and the result follows. □

## 2.3 Final Calculation

Applying these results twice (smoothing both measures), and noting that:

$$\phi_{\nu \star \varphi_\varepsilon}(m) = \mathbb{E}\exp(im^T(X + \sqrt{\varepsilon}Z)) = \exp(-\varepsilon\|m\|^2/2)\phi_\nu(m),$$

we see that, for any $\varepsilon > 0$:

$$\widetilde{W}_1(\mu, \nu) \leq 2\sqrt{d\varepsilon} + \sqrt{\sum_{m \in \mathbb{Z}^d \setminus \{0\}} \frac{1}{\|m\|^2} \exp(-\varepsilon\|m\|^2)|\phi_\mu(m) - \phi_\nu(m)|^2}.$$

Applying this to the empirical measure we get:

$$\mathbb{E}W_1(\mu_n, \mu) \lesssim \sqrt{d\varepsilon} + \sqrt{\sum_{m \in \mathbb{Z}^d \setminus \{0\}} \frac{1}{\|m\|^2} \exp(-\varepsilon\|m\|^2)\mathbb{E}|\phi_{\mu_n}(m) - \phi_\mu(m)|^2},$$

and noting that $\phi_{\mu_n}(m) - \phi_\mu(m) = \frac{1}{n}\sum_{i=1}^n \exp(im^T X_i) - \mathbb{E}\exp(im^T X)$, which are bounded in modulus, so:

$$\mathbb{E}|\phi_{\mu_n}(m) - \phi_\mu(m)|^2 \leq \frac{1}{n}.$$

Putting these together we get that,

$$\mathbb{E}W_1(\mu_n, \mu) \lesssim \sqrt{d\varepsilon} + \frac{1}{\sqrt{n}}\sqrt{\sum_{m \in \mathbb{Z}^d \setminus \{0\}} \frac{1}{\|m\|^2} \exp(-\varepsilon\|m\|^2)}.$$

Noting that the sum can be upper bounded by the integral $\int_{\|x\| \geq 1} \|x\|^{-2}\exp(-\varepsilon\|m\|^2)dx \lesssim \log(1/\varepsilon)$, we have the bound,

$$\mathbb{E}W_1(\mu_n, \mu) \lesssim \sqrt{d\varepsilon} + \frac{1}{\sqrt{n}}\sqrt{\log(1/\varepsilon)},$$

and the result for $d = 2$ with sharp log-factor follows by setting $\varepsilon = 1/n$.