Statistical OT Lecture 7: Dual Bounds on W_1

1 Estimation of and under the Wasserstein distance

We will begin our journey by considering the following question: suppose we are given $X_1, \ldots, X_n \sim \mu$, where μ is supported on $[0,1]^d$. How well can we estimate μ in the Wasserstein sense. This is a question of density estimation where our loss is measured using the Wasserstein distance.

Recall that in the last class we saw that it might make sense to simply use the empirical measure:

$$\widehat{\mu} := \mu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i},$$

as an estimate of the underlying distribution μ.

We will in the next few lectures give a couple of different proofs of the following result (building up the necessary background), and then discuss its tightness:

Theorem 1. Suppose that μ is supported on $[0,1]^d$ then:

$$\mathbb{E}W_1(\mu_n, \mu) \lesssim \sqrt{d} \begin{cases} n^{-1/2} & \text{if d} = 1 \\ \sqrt{\frac{\log n}{n}} & \text{if d} = 2 \\ n^{-1/d} & \text{if } d \geq 3. \end{cases}$$

2 Chaining

One natural way to approach our goal is to recall the dual representation of W_1:

$$W_1(\mu_n, \mu) = \sup_{f \in \text{Lip}_1} \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f.$$

So to bound the Wasserstein distance, we need to study the (uniform) deviation of a collection of averages from their expectation. This is the central quantity of study in *empirical process theory*, and we will develop one of the most important tools in empirical process theory (chaining).

To begin with, let us define the covering number of a class of functions $\mathcal{F}$. An ε-cover is a collection of functions $\{f_1, \ldots, f_N\}$ such that, for any $f \in \mathcal{F}$, there is a function $f_j \in \{f_1, \ldots, f_N\}$ such that $\rho(f, f_j) \leq \varepsilon$ for some metric ρ. We will primarily focus on the case when the metric is the sup-norm, but most of the ideas generalize straightforwardly. The cardinality of the minimum ε-cover is the covering number:

$$N(\varepsilon, \mathcal{F}, \rho) = \min\{N : \exists \{f_1, \ldots, f_N\} \text{ an } \varepsilon\text{-cover of } \mathcal{F}\}.$$

Covering numbers give us a sensible way of discussing the "size" of (potentially) infinite collections of functions.

With this in place, we can state a useful result:

Theorem 2. Suppose that $\mathcal{F}$ is a collection of functions, with $\|f\|_\infty \leq R$, then:

$$\mathbb{E}\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f \lesssim \inf_{\tau > 0}\left[\tau + \frac{1}{\sqrt{n}}\int_{\tau}^{2R} \sqrt{\log N(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)}\, d\varepsilon\right].$$

The second term on the right is known as the entropy integral. We remark again, we use ℓ_∞-covering numbers in the result above, but the proof is far more general. Before we turn our attention to proving this result we state another useful result giving the covering

numbers for the class of 1-Lipschitz functions. We note that we are interested in Lipschitz functions on $[0,1]^d$, and the quantity of interest is invariant to shifting each of the functions by a constant so without loss of generality we can focus on 1-Lipschitz functions on $[0,1]^d$ for which $f(0,\ldots,0)=0$. Denote this class of functions as $\widetilde{\mathcal{F}}$.

Lemma 1.

$$\log N(\varepsilon, \widetilde{\mathcal{F}}, \|\cdot\|_\infty) \lesssim (4\sqrt{d}/\varepsilon)^d.$$

With these two results in place, it is straightforward to prove Theorem 1 up to an extra log factor when $d=2$. We will return to this log factor at some point.

Proof. Using the chaining result with the covering number bound we obtain that when $d=1$:

$$\mathbb{E}W_1(\mu_n,\mu) \lesssim \inf_{\tau>0}\left[\tau + \frac{1}{\sqrt{n}}\int_\tau^1 (4/\varepsilon)^{1/2} d\varepsilon\right],$$

and choosing $\tau=0$, we see that $\mathbb{E}W_1(\mu_n,\mu) \lesssim n^{-1/2}$. When $d\geq 2$ the entropy integral diverges so we need to truncate it (i.e. select $\tau>0$). Choosing $\tau = 4\sqrt{d}n^{-1/d}$, we obtain that,

$$\mathbb{E}W_1(\mu_n,\mu) \lesssim 4\sqrt{d}n^{-1/d} + \frac{1}{\sqrt{n}}\int_{4\sqrt{d}n^{-1/d}}^{\sqrt{d}} (4\sqrt{d}/\varepsilon)^{d/2} d\varepsilon.$$

When $d=2$ the integral is $O(\log n)$ and we see that $\mathbb{E}W_1(\mu_n,\mu) \lesssim (\log n)/\sqrt{n}$. For $d>2$ we have the bound

$$\mathbb{E}W_1(\mu_n,\mu) \lesssim \sqrt{d}n^{-1/d},$$

as desired. □

We will devote the rest of our lecture to understanding Theorem 2 and Lemma 1.

3 Proof of Theorem 2

3.1 Preliminaries

It will be useful to first define a sub-Gaussian process:

Definition 1. A collection of random variables $\{X_t\}_{t\in T}$ is a sub-Gaussian process if $\mathbb{E}[X_t]=0$, and

$$\mathbb{E}\exp(\lambda(X_s - X_t)) \leq \exp(\lambda^2\rho(s,t)^2/2), \quad \forall\ \lambda\geq 0, s,t\in T.$$

Throughout we will let D denote the diameter, i.e. $D = \sup_{s,t\in T}\rho(s,t)$. Our second definition is that of a Lipschitz process:

Definition 2. A collection of random variables $\{X_t\}_{t\in T}$ is a Lipschitz process with respect to a metric ρ if for some $v>0$,

$$|X_s - X_t| \leq v\rho(s,t) \quad \forall\ s,t\in T.$$

We were interested in a quantity of the form:

$$\mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f,$$

and for each $f \in \mathcal{F}$ defining the random variable $X_f = \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f$. We see that this collection of random variables has mean 0, and furthermore:

$$\begin{aligned}\mathbb{E}\exp(\lambda(X_f - X_g)) &= \prod_{i=1}^n \mathbb{E}\exp(\lambda/n(f(X_i) - g(X_i) - \mathbb{E}f(X_i) + \mathbb{E}g(X_i))) \\ &\leq \prod_{i=1}^n \exp\left(\frac{\lambda^2\|f-g\|_\infty^2}{2n^2}\right) = \exp\left(\frac{\lambda^2\|f-g\|_\infty^2}{2n}\right),\end{aligned}$$

where the inequality follows from Hoeffding's lemma. This shows that the empirical process we are interested in studying is sub-Gaussian with respect to the metric $\rho(s,t) = \|s-t\|_\infty/\sqrt{n}$. Furthermore, it is easy to see from the above proof that:

$$|X_f - X_g| \leq 2\|f-g\|_\infty,$$

which shows that it is also a Lipschitz process with respect to the metric $\rho(s,t) = \|s-t\|_\infty/\sqrt{n}$, with $v = 2\sqrt{n}$ (it will be convenient to use the same metric, and absorb the $\sqrt{n}$ scaling into v). Finally, the diameter of our collection of random variables as measured by $\rho(s,t) = \|s-t\|_\infty/\sqrt{n}$ is $\sqrt{d/n}$.

For some intuition, one should view definition 1 as capturing the "in-probability" Lipschitzness of the process, i.e. directly from the definition (and an application of Markov's inequality) we see that for a sub-Gaussian process:

$$\mathbb{P}(|X_s - X_t| \geq t\rho(s,t)) \leq \exp(-t^2/2).$$

On the other hand, definition 2 captures the "almost-sure" Lipschitzness of the process. What our example above shows (and this is typical) is that many processes of interest exhibit much stronger "in-probability" Lipschitzness versus "almost-sure" Lipschitzness. In the former case the Lipschitz constant is $O(1)$, and in the latter case it is $O(\sqrt{n})$ when the metric is chosen to be $\|s-t\|_\infty/\sqrt{n}$). We might obtain better bounds if we can correctly exploit this fact.

3.2 Maxima of Sub-Gaussian RVs

We will need one more detour to build our intuition. Our general goal is to understand the maxima of sub-Gaussian processes, where the index set T is typically infinite. It will be useful to understand the case where T is finite better first.

Notice, that:

$$\mathbb{E}\sup_t X_t \leq \mathbb{E}\sum_t |X_t| \leq |T|\sup_t \mathbb{E}|X_t|.$$

This seems like an extremely naive bound. If X_i had bounded p-th moment, then we could strengthen this:

$$\mathbb{E}\sup_t X_t \leq [\mathbb{E}\sup_t |X_t|^p]^{1/p} \leq |T|^{1/p}\sup_t [\mathbb{E}|X_t|^p]^{1/p}.$$

It should be intuitive that if each of the random variables were σ-sub Gaussian (all moments are finite) then we could do better:

Lemma 2. If $\{X_t\}$ is a collection of σ-sub Gaussian random variables then,

$$\mathbb{E}\sup_t X_t \leq \sqrt{2\sigma^2 \log|T|}.$$

Proof. For any $\lambda > 0$,

$$\begin{aligned}
\mathbb{E} \sup_t X_t &\leq \frac{1}{\lambda} \log \mathbb{E} \exp(\lambda \sup_t X_t) \\
&\leq \frac{1}{\lambda} \log \sum_t \mathbb{E} \exp(\lambda X_t) \\
&\leq \frac{1}{\lambda} \left[\log |T| + \frac{\sigma^2 \lambda^2}{2} \right] \\
&\leq \sqrt{2\sigma^2 \log |T|},
\end{aligned}$$

where the first inequality is Jensen's inequality. □

3.3 One-step Discretization

Before we get to the full chaining bound let us prove the following simpler lemma:

Lemma 3. Suppose that $\{X_t\}_{t \in T}$ is a sub-Gaussian process and Lipschitz process with respect to the metric ρ then:

$$\mathbb{E}[\sup_t X_t] \lesssim \inf_{\tau > 0} \left[v\tau + D\sqrt{\log N(T, \rho, \tau)} \right].$$

Proof. Fix any $X_0 \in \{X_t\}$, and let us denote by $X_{\pi(t)}$ the nearest element to X_t in a minimal τ cover of X_t. Then:

$$\begin{aligned}
\mathbb{E} \sup_t X_t &= \mathbb{E} \sup_t (X_t - X_0) \\
&\leq \mathbb{E} \sup_t |X_t - X_{\pi(t)}| + \mathbb{E} \sup_t |X_{\pi(t)} - X_0|.
\end{aligned}$$

Now, the first term is simple to control using the Lipschitzness of the process. The second term is the supremum of $N(T, \rho, \tau)$ D-sub-Gaussian random variables, and we can bound it using Lemma 2. Together we obtain the desired result. □

This result already illustrates the two main ideas we use to bound the supremum of a sub-Gaussian process (approximation, and a union bound). It turns out however that this bound is not tight enough on its own. A useful exercise to try is to instantiate this bound for the collection of Lipschitz functions in 1-D – you will obtain an $n^{-1/3}$ rate, instead of the correct $n^{-1/2}$ rate.

The main problem is that we are not using the fact that this is a sub-Gaussian process at all (instead, we are simply using the fact that the collection of random variables is sub-Gaussian, but we have a lot more structure left to exploit). An alternate perspective is simply that we could attempt to improve this bound, by using a similar technique (of one-step discretization) to improve our bound on the second term (and then recurse this).