Math 547: Mathematical Foundations of Statistical Learning Theory

Lectures 1-3 — Week 12

Instructor: S. Minsker *Scribe: S. Minsker*

1.0.1 Estimation with prior information.

Recall the definition of the Gaussian mean width of a bounded set $K \subset \mathbb{R}^p$:

$$w(K) = \mathbb{E} \sup_{z \in K - K} \langle z, g \rangle ,$$

where g has $N(0, I_p)$ distribution and $K - K = \{u - v,\ u, v \in K\}$. The Gaussian mean width satisfies many useful properties, some of which are stated below. I will ask you to prove them in your next homework assignment.

Exercise. Show that

1. $w(K) = 2\,\mathbb{E}\sup_{z\in K} \langle z, g\rangle$.
2. $w(K)$ is invariant under affine transformations, meaning that for any $y \in \mathbb{R}^p$ and any $Q \in \mathbb{R}^{p\times p}$ such that $Q^{-1} = Q^T$, $w(QK + y) = w(K)$.
3. $w(K)$ is invariant with respect to taking the convex hull: if $co(K)$ is the convex hull of K, then $w(\mathrm{co}(K)) = w(K)$.
4. Let $\mathrm{diam}(K)$ be the diameter of K. Show that

$$\sqrt{\frac{2}{\pi}}\,\mathrm{diam}(K) \le w(K) \le \sqrt{p}\,\mathrm{diam}(K).$$

Our main technical result is the following statement.

Theorem 1. Let $T \subset \mathbb{R}^p$ be bounded. Define the "ε-band"

$$T_\varepsilon = \left\{ z \in T \,\middle|\, \frac{1}{n} \left\|\mathbf{X}z\right\|_1 \le \varepsilon \right\},$$

where $\|x\|_1 = \sum_{j=1}^p |x_j|$ is the ℓ_1 norm. Then

$$\mathbb{E}\sup_{z\in T_\varepsilon} \|z\|_2 \le \sqrt{\frac{8\pi}{n}}\mathbb{E}\sup_{z\in T} |\langle g, z\rangle| + \sqrt{\frac{\pi}{2}}\varepsilon.$$

The following is an immediate corollary:

Corollary 1. Set $T = \mathcal{K} - \mathcal{K}$ and $\varepsilon = 0$ in the theorem. Then $T_\varepsilon = (K - K) \cap \ker(\mathbf{X})$, the kernel of $\mathbf{X}$, and

$$\mathbb{E} \sup_{z \in T \cap \ker(\mathbf{X})} \|z\|_2 = \mathbb{E} \sup_{z \in (\mathcal{K}-\mathcal{K}) \cap E} \|z\|_2 = \mathbb{E} \operatorname{diam}\left((\mathcal{K} - \mathcal{K}) \cap E\right) \le \sqrt{\frac{8\pi}{n}} w(\mathcal{K}).$$

Hence, we obtain an explicit bound for the estimation error in our problem.

Example 1. If $\mathcal{K}$ is a finite set, then

$$\mathbb{E} \sup_{z \in \mathcal{K}-\mathcal{K}} |\langle g, z\rangle| \le \sqrt{2}\operatorname{diam}(\mathcal{K})\sqrt{\log(2\operatorname{card}(T))}.$$

Example 2. $\mathcal{K} \subset L$, where L a d-dimensional subspace of $\mathbb{R}^p$. Then

$$w(\mathcal{K}) = \mathbb{E} \sup_{z \in \mathcal{K}-\mathcal{K}} \langle g, z\rangle \le \operatorname{diam}(\mathcal{K})\sqrt{d}.$$

Prove it using the properties of the multivariate normal distribution (namely, that a projection of a normally distributed vector is still normally distributed).

Proof of the theorem. Assume we can show that

$$\mathbb{E} \sup_{z \in T} \left| \frac{1}{n} \sum_{j=1}^n \left|\langle X_j, z\rangle\right| - \sqrt{\frac{2}{\pi}} \|z\|_2 \right| \le \frac{4}{\sqrt{n}} \mathbb{E} \sup_{z \in T} |\langle g, z\rangle|, \tag{1.1}$$

where $g \sim \mathcal{N}(0, \mathbf{I}_p)$ and $T' \subset \mathbb{R}^p$. Note that, since $T_\varepsilon \subset T$,

$$\mathbb{E} \sup_{z \in T_\varepsilon} \left| \frac{1}{n} \sum_{j=1}^n \left|\langle X_j, z\rangle\right| - \sqrt{\frac{2}{\pi}} \|z\|_2 \right| \le \mathbb{E} \sup_{z \in T} \left| \frac{1}{n} \sum_{j=1}^n \left|\langle X_j, z\rangle\right| - \sqrt{\frac{2}{\pi}} \|z\|_2 \right|.$$

Moreover, for $z \in T_\varepsilon$,

$$\frac{1}{n} \sum_{j=1}^n |\langle X_j, z\rangle| = \frac{1}{n} \left\|\mathbf{X} z\right\|_1 \le \varepsilon,$$

which implies that

$$\mathbb{E} \sup_{z \in T_\varepsilon} \|z\|_2 \le \sqrt{\frac{\pi}{2}} \varepsilon + \sqrt{\frac{\pi}{2}} \frac{4}{\sqrt{n}} \mathbb{E} \sup_{z \in T} |\langle g, z\rangle|.$$

It remains to establish the inequality (1.1). First, note that $\mathbb{E}|\langle X_1, z\rangle| = \sqrt{2/\pi}\|z\|_2$ since X_1 has standard normal distribution. Next, by the symmetrization and contraction inequalities (applied to $\phi(x) = |x|$),

$$\begin{aligned}
\mathbb{E}\sup_{z\in T}\left|\frac{1}{n}\sum_{j=1}^{n}\left|\langle X_j, z\rangle\right| - \sqrt{\frac{2}{\pi}}\|z\|_2\right| &= \mathbb{E}\sup_{z\in T}\left|\frac{1}{n}\sum_{j=1}^{n}\left|\langle X_j, z\rangle\right| - \mathbb{E}\left|\langle X_j, z\rangle\right|\right| \\
&\le 2\mathbb{E}\sup_{z\in T}\left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j|\langle X_j, z\rangle|\right| \\
&= 4\mathbb{E}\sup_{z\in T}\left|\frac{1}{n}\sum_{j=1}^{n}\langle \varepsilon_j \cdot X_j, z\rangle|\right| \\
&= \frac{4}{\sqrt{n}}\mathbb{E}\sup_{z\in T}\left|\left\langle \underbrace{\sum_{j=1}^{n}\frac{1}{\sqrt{n}}\varepsilon_j X_j}_{:=g\sim\mathcal{N}(0,\mathbf{I}_p)}, z\right\rangle\right| = \frac{4}{\sqrt{n}}\mathbb{E}\sup_{z\in T}\left|\left\langle g, z\right\rangle\right|,
\end{aligned}$$

where we used the fact that $\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\varepsilon_j X_j$ has $\mathcal{N}(0,\mathbf{I}_p)$ distribution (check!). □

Remark 1. Note that for any $z_0 \in T$ and T not necessarily symmetric,

$$\begin{aligned}
\mathbb{E}\sup_{z\in T}|\langle g, z\rangle| &\le \mathbb{E}\sup_{z\in T}|\langle g, z - z_0\rangle| + |\langle g, z_0\rangle| \\
&\le \mathbb{E}\sup_{z,z_0\in T}|\langle g, z - z_0\rangle| + \sqrt{\mathbb{E}|\langle g, z_0\rangle|^2} \\
&= \mathbb{E}\sup_{z,z_0\in T}\langle g, z - z_0\rangle + \sqrt{\mathbb{E}|\langle g, z_0\rangle|^2} \\
&= w(T) + \|z_0\|_2.
\end{aligned}$$

1.1 Estimation from noisy observations.

In this section, we will extend the previous results on noiseless measurements to the case of noisy observations. Assume that

$$Y = \mathbf{X}\lambda_* + \nu \quad \text{s.t.} \quad \frac{1}{n}\|\nu\|_1 \le \varepsilon.$$

Here, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the design matrix such that $X_{i,j} \sim \mathcal{N}(0, \mathbf{I}_p)$, and $\nu \in \mathbb{R}^n$ is the noise vector. Note that

$$\frac{1}{n}\|\nu\|_2 = \frac{1}{n}\sqrt{\sum_{j=1}^{n}\nu_j^2} \le \frac{1}{n}\sqrt{\left(\sum_{j=1}^{n}|\nu_j|\right)^2} \le \varepsilon.$$

Let $\widehat{\lambda} \in \mathbb{R}^p$ satisfy (a) $\widehat{\lambda} \in \mathcal{K}$, and (b) $\frac{1}{n}\left\|Y - \mathbf{X}\widehat{\lambda}\right\|_1 \le \varepsilon$.

Theorem 2. For any $\widehat{\lambda}$ that satisfies (a) and (b),

$$\mathbb{E}\sup_{\lambda\in\mathcal{K}}\|\widehat{\lambda} - \lambda_*\|_2 \le \sqrt{8\pi}\left(\frac{w(K)}{\sqrt{n}} + \frac{\varepsilon}{2}\right).$$

Proof. Set $T = \mathcal{K} - \mathcal{K}$ and $\varepsilon' = 2\varepsilon$, and apply Theorem 1to get that

$$\mathbb{E} \sup_{u \in T_{\varepsilon'}} \|u\|_2 \leq \sqrt{\frac{8\pi}{n}} \left(\mathbb{E} \sup_{u \in K - K} \langle g, u \rangle \right) + \sqrt{\frac{\pi}{2}} \varepsilon'.$$

Observe that $\widehat{\lambda} - \lambda \in T_{2\varepsilon}$ for any $\lambda \in \mathcal{K}$. Indeed,

$$\begin{aligned} \frac{1}{n} \left\| \mathbf{X}(\widehat{\lambda} - \lambda) \right\|_1 &= \frac{1}{n} \left\| \mathbf{X}\widehat{\lambda} - Y + \nu \right\|_1 \\ &\leq \frac{1}{n} \left\| \mathbf{X}\widehat{\lambda} - Y \right\|_1 + \frac{1}{n} \|\nu\|_1 \leq 2\varepsilon. \end{aligned}$$

Hence, $\|\widehat{\lambda} - \lambda_*\|_2 \leq \sup_{u \in T_{\varepsilon'}} \|u\|_2$, and the result follows. □

1.1.1 Estimation via convex optimization.

The next question we address is related to computational side of the problem, namely, how to evaluate $\widehat{\lambda}$ numerically? To this end, we will make an additional assumption stating that the set $\mathcal{K}$ is *star-shaped*, meaning that $tK \subseteq K$ for $t \in [0, 1]$.

Definition 1. The gauge (or the Minkowski functional) of associated to the set $\mathcal{K}$ is

$$\|x\|_{\mathcal{K}} := \inf \left\{ t > 0 : \frac{x}{t} \in \mathcal{K} \right\}.$$

Remark 2. $x \in K \iff \|x\|_{\mathcal{K}} \leq 1$.

As before, assume that $Y = \mathbf{X}\lambda + \nu$. Let $\widehat{\lambda}$ be a solution to the problem

$$\|\lambda'\|_{\mathcal{K}} \to \min \quad \text{subject to } \frac{1}{n} \|Y - \mathbf{X}\lambda'\|_1 \leq \varepsilon. \tag{1.2}$$

Theorem 3. Solution $\widehat{\lambda}$ of the problem (1.2) satisfies the inequality

$$\mathbb{E} \sup_{\lambda \in \mathcal{K}} \|\widehat{\lambda} - \lambda\|_2 \leq \sqrt{8\pi} \left(\frac{w(\mathcal{K})}{\sqrt{n}} + \frac{\varepsilon}{2} \right).$$

Proof. It follows from Theorem 1 that it is enough to show that $\widehat{\lambda} \in \mathcal{K}$. The latter follows since

$$\|\widehat{\lambda}\|_{\mathcal{K}} \leq \|\lambda\|_{\mathcal{K}} \leq 1$$

by the definition of $\widehat{\lambda}$. □

If $\mathcal{K}$ is convex, then $\|\cdot\|_{\mathcal{K}}$ is also convex, and (1.2) is a convex problem that can be solved efficiently (say, by the gradient descent). What if $\mathcal{K}$ is not convex? A natural idea is to replace $\mathcal{K}$ by the smallest convex set that contains $\mathcal{K}$, namely its convex hull $\text{co}(\mathcal{K})$.

$$\|\lambda'\|_{\text{co}(\mathcal{K})} \to \min \quad \text{subject to } \frac{1}{n} \|Y - \mathbf{X}\lambda'\|_1 \leq \varepsilon. \tag{1.3}$$

It follows from Theorem 3 that

$$\mathbb{E} \sup_{\lambda \in \mathcal{K}} \|\widehat{\lambda} - \lambda\|_2 \leq \mathbb{E} \sup_{\lambda \in \text{co}(\mathcal{K})} \|\widehat{\lambda} - \lambda\|_2 \leq \sqrt{8\pi} \left(\frac{w(\mathcal{K})}{\sqrt{n}} + \frac{\varepsilon}{2} \right)$$

since $w(\text{co}(\mathcal{K})) = w(\mathcal{K})$ by the property of the Gaussian mean width.

Example 3. Assume that λ is sparse, so that

$$J(\lambda) = \{j \in \{1, \dots, p\} : \lambda_j \neq 0\} \quad \text{satisfies} \quad |J(\lambda)| = s \ll p.$$

We know that

$$\lambda \in K_\lambda := \{\lambda' \in \mathbb{R}^p : \ |J(\lambda')| \leq s, \ \|\lambda'\|_1 \leq \|\lambda\|_1\}.$$

The extreme points of the set K_λ consist precisely of the rescaled basis vectors

$$\{\pm\|\lambda\|_1 \, e_j, \ j = 1, \dots, p\},$$

hence its convex hull (check!) is

$$\mathrm{co}(K_\lambda) := \mathcal{K} = \|\lambda\|_1 B_{\|\cdot\|_1}(0, 1),$$

where $B_{\|\cdot\|_1}(0, 1)$ is the unit ball for ℓ_1 norm. Consider the convex minimization problem

$$\|\lambda'\|_{\mathcal{K}} \to \min \quad \text{subject to } \mathbf{X}\lambda' = Y. \tag{1.4}$$

But, since minimizing $\|\cdot\|_{\mathcal{K}}$ it equivalent to minimizing $\|\cdot\|_{c\mathcal{K}}$ for any $c > 0$, problem (1.4) is in turn equivalent to

$$\|\lambda'\|_1 \to \min \quad \text{subject to } \mathbf{X}\lambda' = Y. \tag{1.5}$$

Let $\widehat{\lambda}$ be a solution. It immediately follows from Theorem 3 that

$$\mathbb{E}\|\widehat{\lambda} - \lambda\|_2 \leq \sqrt{8\pi} \frac{w(\mathcal{K})}{\sqrt{n}}.$$

One can check (we did this in class) that $w\left(\|\lambda\|_1 B_{\|\cdot\|_1}(0, 1)\right) \leq \sqrt{2}\|\lambda\|_1 \sqrt{\log(2p)}$.