# Width Comparison Homework

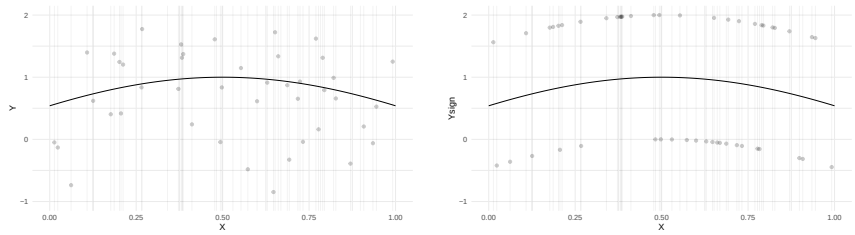
QTM 490R: Machine Learning Theory

## 1 Random-Sign Width and Contraction Inequalities

Our goal here is to move away from the gaussian-noise model, toward more realistic assumptions. But our starting point will be another stylized model: the *random-sign noise* model. In this model, our observations are independent and what we observe is, depending on the outcome of a coin flip, either $Y_i = \mu(X_i)+1$ or $Y_i = \mu(X_i) - 1$.

$$Y_i = \mu(X_i) + s_i \quad \text{where} \quad s_i = \begin{cases} +1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases}.$$

To get a sense of what this looks like, I've plotted observations with noise gaussian noise and random sign noise below.




What we'll show is that it's roughly equally hard to estimate $\mu$ when we have random sign noise and when we have gaussian noise. In particular, what we'll show that for a vector $s$ of independent random signs,

$$\underset{\approx 1.25}{1/\,\mathrm{E}|\varepsilon_i|}\, \mathrm{w}_s(\mathcal{V}) \leq \mathrm{w}_\varepsilon(\mathcal{V}) \leq \mathrm{E}\, \underset{\propto \sqrt{\log(n)}}{\max_{i \leq n}} |\varepsilon_i|\, \mathrm{w}_s(\mathcal{V}) \tag{1}$$

This random-sign width, which is sometimes called the *Rademacher Complexity* of the set $\mathcal{V}$, is something people talk a lot about. Like gaussian width, there are a lot of techniques out there for calculating and bounding it for various sets $\mathcal{V}$. But absent a technique specific to random-sign width, we can always resort

to using the bound above in combination with tecniques for calculating and bounding gaussian width.

In fact, what we'll prove is more general. We'll prove inequalities that (1), ignoring the underset gaussian specific contants, holds for any vector $\varepsilon$ of independent *symmetric* Here, by symmetric, we mean that $\varepsilon_i$ and $-\varepsilon_i$ have the same distribution.

This is relevant because it allows us to bound the difficulty of estimating $\mu$ in one noise model in terms of another.

**Exercise 1** *Suppose we have noise vectors $\varepsilon$ and $\nu$ for which we know that $\alpha\,\mathrm{w}_\nu(\mathcal{V}) \le \mathrm{w}_\varepsilon(\mathcal{V}) \le \beta\,\mathrm{w}_\nu(\mathcal{V})$ for all sets $\mathcal{V}$. If $s_\varepsilon^2 \ge 2c_\delta\,\mathrm{w}_\varepsilon(\mathcal{M}_s)$ and $s_\nu^2 \ge 2c_\delta\,\mathrm{w}_\nu(\mathcal{V})$ respectively, then we have the bounds $\|\hat{\mu}-\mu\|_{L_2(\mathrm{P_n})} < s_\varepsilon$ and $\|\hat{\mu}-\mu\|_{L_2(\mathrm{P_n})} < s_\nu$ these two noise models. Find two-sided bound of the form $f(\alpha,\beta)s_\nu \le s_\varepsilon \le g(\alpha,\beta)s_\nu$ for some functions $f$ and $g$.*

To prove our comparison inequality (1) we'll use another bound that's useful in its own right.

$$\mathrm{E}\max_{v\in\mathcal{V}}\sum_i a_i s_i v_i \le \mathrm{E}\max_{v\in\mathcal{V}}\sum_i s_i v_i \quad \text{for} \quad a_i \ \text{ with } \ |a_i|\le 1. \tag{2}$$

We call this the *contraction inequality* for random-sign width. It says that if we transform each vector in a set $\mathcal{V}$ by scaling down each element $v_i$ by some fixed amount $|a_i| \le 1$ and/or flipping its sign, we cannot get a set of larger random-sign width.

## 1.1 Proving the Contraction Inequality

Let's think about the function $f(u) = \mathrm{E}\max_{v\in\mathcal{V}} u_i s_i v_i$. In terms of this function, the left side of (2) is $f(a)$. Here's a sketch of our argument.

1. We'll observe that $f(a) \le \max_{u\in\mathcal{U}} f(u)$ for any set $\mathcal{U}$ that contains $a$. In particular, it holds for the cube $\mathcal{U} = \{u : |u_i| \le B\}$ for $B = \max_i |a_i|$.
2. We'll show that this function $f$ is convex.
3. We'll show that the maximum of a convex function over this cube must occur at an *extreme point*, i.e., at a vector $u$ with $u_i = \pm B$ for all $i$.
4. We'll observe that because $+Bs_i$ and $-Bs_i$ have the same distribution, and therefore the random vector $us$ with elements $u_i s_i$ has the same distribution for all extreme points $u$. Consequently, for every vector $v$, the distribution of the sum $\sum_i u_i s_i v_i$ is the same at all extreme points $u$.

That more or less does it. Because this set of extreme points includes the vector $u$ with $u_i = B$ for all $i$, it follows that for any extreme point $u$, $\sum_i u_i s_i v_i$ has the same distribution as $B\sum_i s_i v_i$. It follows that the maximum of these two sums over $v \in \mathcal{V}$ will have the same distribution and therefore the same expected value, and we can conclude that $f(a) \le \max_{u\in\mathcal{U}} f(u) = B\,\mathrm{E}\sum_i s_i v_i$.

To conclude, we've got two things to show. Let's get to it.

**Exercise 2** *Show that the function $f$ is convex.*

**Exercise 3** ***Optional.*** *Show that the maximum of a convex function $f$ over the cube $\mathcal{U} = \{u : |u_i| \le B\}$ must occur at an extreme point of that set, i.e. a point with $|u_i| = B$ for all $i$.*

*You may assume two facts without proof.*

- *Every point $u$ in the set $\mathcal{U}$ can be written as a weighted average of the extreme points $u_1, u_2, \ldots \in \mathcal{U}$, i.e., $u = \sum_i \lambda_i u_i$ where $\sum_i \lambda_i = 1$.*[1]
- *For a convex function $f$, $f(\sum_i \lambda_i u_i) \le \sum_i \lambda_i f(u_i)$ for all $\lambda$ with $\sum_i \lambda_i = 1$.*[2]

## 1.2 Widths involving random sign and symmetric noise

Using the contraction inequality (2), proving our bounds (1) relating random-sign and symmetric-noise width is relatively easy.

The key observation is that we can multiply a symmetric random variable $\varepsilon_i$ by an independent random sign $s_i$ without changing its distribution. It follows that if $\varepsilon$ is a vector of independent symmetric random variables and $s$ a vector of random signs that are independent of each other and $\varepsilon$, then the random vector $s\varepsilon$ with elements $s_i\varepsilon_i$ has the same distribution as $\varepsilon$. The implication relevant to us is that for any vector $v$, $\sum_i s_i\varepsilon_i v_i$ and $\sum_i s_i\varepsilon_i v_i$ have the same distribution and therefore, for set of vectors $\mathcal{V}$, $\mathrm{E}\max_{v\in\mathcal{V}} \sum_i s_i\varepsilon_i v_i = \mathrm{E}\max_{v\in\mathcal{V}} \sum_i \varepsilon_i v_i$.

**Exercise 4** *Prove these two claims. You can take the observation above as a given.*

**Exercise 5** ***Optional.*** *Prove the observation above, i.e., that $s\varepsilon$ and $\varepsilon$ have the same distribution.*

**Tip.** Two random variables have the same distribution if they have the same cumulative distribution function, so for the first claim what we want to show is that $P(s_i\varepsilon_i \le t) = P(\varepsilon_i \le t)$ for all $t \in \mathbb{R}$. To do this, use the *law of total probability*: $P(s_i\varepsilon_i \le t) = \sum_{s=\pm 1} P(s_i\varepsilon_i \le t \mid s_i = s)P(s_i = s)$.

[1] This is obvious enough for a one-dimensional cube (i.e. an interval). We can extend that to a two dimensional cube by observing that each point is a weighted average of points of the form $(u_1, 0)$ and $(0, u_2)$ which are themselves weighted averages of the points $[-B, 0]$ and $[B, 0]$ and the points $[0, -B]$ and $[0, B]$ respectively. Then extend it to a three dimensional cube by observing that each point is a weighted average of points $(u_1, u_2, -B)$ and $(u_1, u_2, B)$ that we can write as weighted averages of extreme points. And so on. We prove it by induction on dimension.

[2] Our definition of convexity tells us this is the case when only $\lambda_1$ and $\lambda_2$ are nonzero. We can show it's true when $\lambda_3$ is also nonzero by observing that $f\{(\lambda_1 u_1 + \lambda_2 u_2) + \lambda_3 u_3\} \le f(\lambda_1 u_1 + \lambda_2 u_2) + f(\lambda_3 u_3)$ and treat the case when $\lambda_4$ is also nonzero analogously. We prove it by induction on dimension.

**Tip.** Two random vectors have the same distribution if they have the same multivariate cumulative distribution function, so for the second claim what we want to show is that $P(s_1\varepsilon_1 \leq t_1, \ldots, s_n\varepsilon_n \leq t_n)$ for $t \in \mathbb{R}^n$. To show this, use our first claim and the independence of the random variables $s_1\varepsilon_1 \ldots s_n\varepsilon_n$.

Now, if you like, take a minute to think about what you can say about other types of symmetrically-distributed noise, e.g.

- Noise uniformly distributed on the interval $[-M, +M]$.
- Symmetrically-distributed $\sigma$-subgaussian noise, i.e., symmetric noise satisfying the bound $P(|\nu| \geq t) \leq e^{-t^2/2\sigma^2}$.
- *Laplace* noise, i.e., noise $\nu$ with the density $P(|\nu| \geq t) = e^{-t}$.

**Exercise 6** ***Optional.*** *For each of these noise types, compare the associated width to random-sign width and gaussian width. Use these comparisons to establish bounds,* in terms of gaussian width*, on the error* $\|\hat{\mu} - \mu_\star\|_{L_2(\mathrm{P_n})}$ *for the least squares estimator with that type of noise, much like you did in Exercise 1.*