Lecture 4

July 2, 2025

Metric entropy and related ideas

- Metric entropy ideas allow us to reduce considerations over massive (uncountable) sets, to finite subsets.

Metric entropy and related ideas

- Metric entropy ideas allow us to reduce considerations over massive (uncountable) sets, to finite subsets.
- It is a measure of the size of a set; a quantitative measure of compactness.

Metric entropy and related ideas

- Metric entropy ideas allow us to reduce considerations over massive (uncountable) sets, to finite subsets.
- It is a measure of the size of a set; a quantitative measure of compactness.
- Totally bounded set in a metric space = can be covered with “finite" number of ε -balls for any $\varepsilon > 0$.

Covering and metric entropy

- How to measure the sizes of sets in metric spaces?

Covering and metric entropy

- How to measure the sizes of sets in metric spaces?
- One idea is based on measures of sets, assuming that there is a suitable measure on the space (say Lebesgue measure in $\mathbb{R}^d$).

Covering and metric entropy

- How to measure the sizes of sets in metric spaces?
- One idea is based on measures of sets, assuming that there is a suitable measure on the space (say Lebesgue measure in $\mathbb{R}^d$).
- A more topological idea: Covering one set with copies of another set.

Covering and metric entropy

- How to measure the sizes of sets in metric spaces?
- One idea is based on measures of sets, assuming that there is a suitable measure on the space (say Lebesgue measure in $\mathbb{R}^d$).
- A more topological idea: Covering one set with copies of another set.
- **Definition 1 (Covering number).** An ε-cover or ε-net of a set T w.r.t. a metric ρ is a set

$$\mathcal{N} := \{\theta^1, \dots, \theta^N\} \subset T$$

such that

$$\forall t \in T, \;\; \exists \theta^i \in \mathcal{N} \;\; \text{satisfying} \;\; \rho(\theta, \theta^i) \leq \varepsilon.$$

Covering and metric entropy

- How to measure the sizes of sets in metric spaces?
- One idea is based on measures of sets, assuming that there is a suitable measure on the space (say Lebesgue measure in $\mathbb{R}^d$).
- A more topological idea: Covering one set with copies of another set.
- **Definition 1 (Covering number).** An ε-cover or ε-net of a set T w.r.t. a metric ρ is a set

$$\mathcal{N} := \{\theta^1, \ldots, \theta^N\} \subset T$$

such that

$$\forall t \in T, \ \exists \theta^i \in \mathcal{N} \quad \text{satisfying} \quad \rho(\theta, \theta^i) \leq \varepsilon.$$

The ε-covering number $N(\varepsilon, T, \rho)$ is the cardinality of the smallest ε-cover.

Covering and metric entropy

- How to measure the sizes of sets in metric spaces?
- One idea is based on measures of sets, assuming that there is a suitable measure on the space (say Lebesgue measure in $\mathbb{R}^d$).
- A more topological idea: Covering one set with copies of another set.
- **Definition 1 (Covering number).** An ε-cover or ε-net of a set T w.r.t. a metric ρ is a set

  $$\mathcal{N} := \{\theta^1, \ldots, \theta^N\} \subset T$$

  such that

  $$\forall t \in T, \; \exists \theta^i \in \mathcal{N} \quad \text{satisfying} \quad \rho(\theta, \theta^i) \leq \varepsilon.$$

  The ε-covering number $N(\varepsilon, T, \rho)$ is the cardinality of the smallest ε-cover.
- $\log N(\varepsilon, T, \rho)$ is called the metric entropy of the (metric) space (T, ρ).

- In normed vector spaces, $\rho(x, y) = \|x - y\|$.

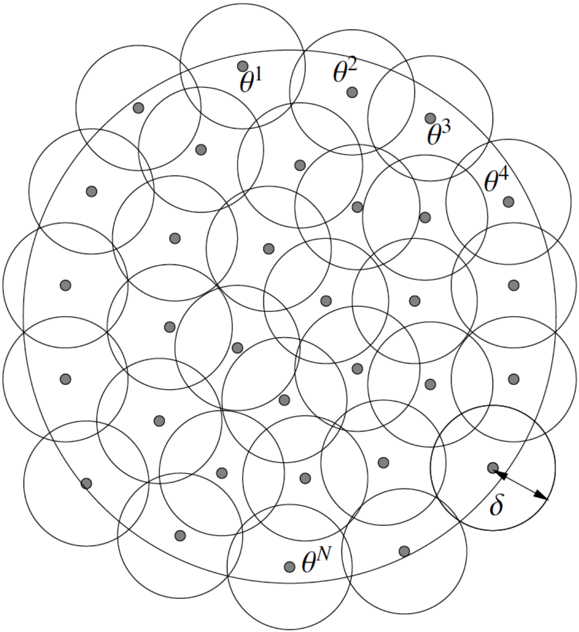
- In normed vector spaces, $\rho(x, y) = \|x - y\|$.
- Notation: $\mathbb{B} = \{\theta : \|\theta\| \leq 1\}$.

- In normed vector spaces, $\rho(x, y) = \|x - y\|$.
- Notation: $\mathbb{B} = \{\theta : \|\theta\| \leq 1\}$.
- Ball of radius ε centered at θ^j is $\theta^j + \varepsilon\mathbb{B}$.

- In normed vector spaces, $\rho(x, y) = \|x - y\|$.
- Notation: $\mathbb{B} = \{\theta : \|\theta\| \leq 1\}$.
- Ball of radius ε centered at θ^j is $\theta^j + \varepsilon\mathbb{B}$.
- $\mathcal{N} := \{\theta^1, \ldots, \theta^N\}$ is an ε-covering iff

$$T \subset \bigcup_{j=1}^{N} (\theta^j + \varepsilon\mathbb{B}),$$

i.e., covering T with shifted copies of $\varepsilon\mathbb{B}$.



θ^1
θ^2
θ^3
θ^4
θ^N
δ

Packing

- **Definition 2.** An ε-packing (ε-separated set) of a set T w.r.t. a metric ρ is a set

$$\mathcal{M} := \{\theta^1, \ldots, \theta^N\} \subset T$$

such that

$$\rho(\theta^i, \theta^j) > \varepsilon, \quad \forall i \neq j.$$

Packing

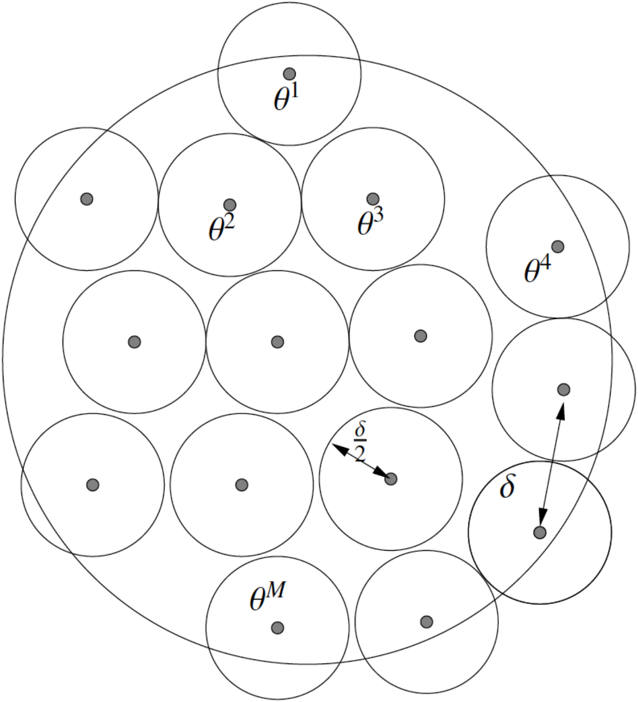
- **Definition 2.** An ε-packing (ε-separated set) of a set T w.r.t. a metric ρ is a set

$$\mathcal{M} := \{\theta^1, \ldots, \theta^N\} \subset T$$

such that

$$\rho(\theta^i, \theta^j) > \varepsilon, \quad \forall i \neq j.$$

The ε-packing number $M(\varepsilon, T, \rho)$ is the cardinality of the largest ε-packing.



θ^1
θ^2
θ^3
θ^4
$\frac{\delta}{2}$
δ
θ^M

Example

- **Example 1 (Covering ℓ_∞).** Take $T = [-1, 1]$ and $\rho(\theta, \theta') = |\theta - \theta'|$.

Example

- **Example 1 (Covering ℓ_∞).** Take $T = [-1, 1]$ and $\rho(\theta, \theta') = |\theta - \theta'|$.
- Divide $[-1, 1]$ into $L = \lfloor 1/\varepsilon \rfloor + 1$ sub-intervals centered at the points $\theta^i = -1 + (2i - 1)\varepsilon$ for $i \in [L] = \{1, \ldots, L\}$.

Example

- **Example 1 (Covering ℓ_∞).** Take $T = [-1, 1]$ and $\rho(\theta, \theta') = |\theta - \theta'|$.
- Divide $[-1, 1]$ into $L = \lfloor 1/\varepsilon \rfloor + 1$ sub-intervals centered at the points $\theta^i = -1 + (2i - 1)\varepsilon$ for $i \in [L] = \{1, \ldots, L\}$.
- Easy to verify that $\theta^1, \ldots, \theta^L$ forms an ε-net of $[-1, 1]$, hence

$$N(\varepsilon, [-1, 1], \rho) \;\leq\; L + 1 \;\leq\; \frac{1}{\varepsilon} + 1.$$

Example

- **Example 1 (Covering ℓ_∞).** Take $T = [-1, 1]$ and $\rho(\theta, \theta') = |\theta - \theta'|$.
- Divide $[-1, 1]$ into $L = \lfloor 1/\varepsilon \rfloor + 1$ sub-intervals centered at the points $\theta^i = -1 + (2i - 1)\varepsilon$ for $i \in [L] = \{1, \ldots, L\}$.
- Easy to verify that $\theta^1, \ldots, \theta^L$ forms an ε-net of $[-1, 1]$, hence
$$N(\varepsilon, [-1, 1], \rho) \le L + 1 \le \frac{1}{\varepsilon} + 1.$$
- **Exercise.** Show that this analysis can be extended to covering $[-1, 1]^d$ in the ℓ_∞ metric
$$N(\varepsilon, [-1, 1]^d, \|\cdot\|_\infty) \le \left(\frac{1}{\varepsilon} + 1\right)^d.$$

Relation between packing and covering

- **Lemma 1.** For all $\varepsilon > 0$, the packing and covering numbers are related as follows:

$$M(2\varepsilon, T, \rho) \leq N(\varepsilon, T, \rho) \leq M(\varepsilon, T, \rho).$$

Relation between packing and covering

- **Lemma 1.** For all $\varepsilon > 0$, the packing and covering numbers are related as follows:

$$M(2\varepsilon, T, \rho) \leq N(\varepsilon, T, \rho) \leq M(\varepsilon, T, \rho).$$

- **Proof:** Exercise. (Hint: any maximal packing is automatically a covering of suitable radius.).

Volume ratio estimates

- **Lemma 2.** Let $\|\cdot\|$ and $\|\cdot\|'$ be two norms on $\mathbb{R}^d$ with respective unit balls $\mathbb{B}$ and $\mathbb{B}'$.

Volume ratio estimates

- **Lemma 2.** Let $\|\cdot\|$ and $\|\cdot\|'$ be two norms on $\mathbb{R}^d$ with respective unit balls $\mathbb{B}$ and $\mathbb{B}'$.
- That is, $\mathbb{B} = \{\theta \in \mathbb{R}^d : \|\theta\| \leq 1\}$, and similarly for $\mathbb{B}'$.

Volume ratio estimates

- **Lemma 2.** Let $\|\cdot\|$ and $\|\cdot\|'$ be two norms on $\mathbb{R}^d$ with respective unit balls $\mathbb{B}$ and $\mathbb{B}'$.
- That is, $\mathbb{B} = \{\theta \in \mathbb{R}^d \,:\, \|\theta\| \leq 1\}$, and similarly for $\mathbb{B}'$.
- Then, ε-covering of $\mathbb{B}$ in $\|\cdot\|'$ satisfies

$$\left(\frac{1}{\varepsilon}\right)^d \frac{\mathrm{vol}(\mathbb{B})}{\mathrm{vol}(\mathbb{B}')} \leq N(\varepsilon, \mathbb{B}, \|\cdot\|') \leq M(\varepsilon, \mathbb{B}, \|\cdot\|') \leq \frac{\mathrm{vol}(\frac{2}{\varepsilon}\mathbb{B} + \mathbb{B}')}{\mathrm{vol}(\mathbb{B}')}.$$

Volume ratio estimates

- **Lemma 2.** Let $\|\cdot\|$ and $\|\cdot\|'$ be two norms on $\mathbb{R}^d$ with respective unit balls $\mathbb{B}$ and $\mathbb{B}'$.
- That is, $\mathbb{B} = \{\theta \in \mathbb{R}^d : \|\theta\| \leq 1\}$, and similarly for $\mathbb{B}'$.
- Then, ε-covering of $\mathbb{B}$ in $\|\cdot\|'$ satisfies

$$\left(\frac{1}{\varepsilon}\right)^d \frac{\mathrm{vol}(\mathbb{B})}{\mathrm{vol}(\mathbb{B}')} \leq N(\varepsilon, \mathbb{B}, \|\cdot\|') \leq M(\varepsilon, \mathbb{B}, \|\cdot\|') \leq \frac{\mathrm{vol}(\frac{2}{\varepsilon}\mathbb{B} + \mathbb{B}')}{\mathrm{vol}(\mathbb{B}')}.$$

- Important special case: Covering balls in their own metric: $\mathbb{B} = \mathbb{B}'$. Then

$$\mathrm{vol}((1 + 2/\varepsilon)\mathbb{B}) = \left(\frac{2}{\varepsilon} + 1\right)^d \mathrm{vol}(\mathbb{B}).$$

Volume ratio estimates

- **Lemma 2.** Let $\|\cdot\|$ and $\|\cdot\|'$ be two norms on $\mathbb{R}^d$ with respective unit balls $\mathbb{B}$ and $\mathbb{B}'$.
- That is, $\mathbb{B} = \{\theta \in \mathbb{R}^d \,:\, \|\theta\| \le 1\}$, and similarly for $\mathbb{B}'$.
- Then, ε-covering of $\mathbb{B}$ in $\|\cdot\|'$ satisfies

$$\left(\frac{1}{\varepsilon}\right)^d \frac{\mathrm{vol}(\mathbb{B})}{\mathrm{vol}(\mathbb{B}')} \le N(\varepsilon, \mathbb{B}, \|\cdot\|') \le M(\varepsilon, \mathbb{B}, \|\cdot\|') \le \frac{\mathrm{vol}(\frac{2}{\varepsilon}\mathbb{B} + \mathbb{B}')}{\mathrm{vol}(\mathbb{B}')}.$$

- Important special case: Covering balls in their own metric: $\mathbb{B} = \mathbb{B}'$. Then

$$\mathrm{vol}((1 + 2/\varepsilon)\mathbb{B}) = \left(\frac{2}{\varepsilon} + 1\right)^d \mathrm{vol}(\mathbb{B}).$$

and hence

$$\left(\frac{1}{\varepsilon}\right)^d \le N(\varepsilon, \mathbb{B}, \|\cdot\|) \le \frac{\mathrm{vol}(\frac{2}{\varepsilon}\mathbb{B} + \mathbb{B}')}{\mathrm{vol}(\mathbb{B}')} = \left(\frac{2}{\varepsilon} + 1\right)^d \le \left(\frac{3}{\varepsilon}\right)^d,$$

for $0 < \varepsilon \le 1$.

Proof of Lemma 2

- Let $\{\theta^1, \ldots, \theta^N\}$ be an ε-covering of $\mathbb{B}$ in $\|\cdot\|'$.

Proof of Lemma 2

- Let $\{\theta^1, \ldots, \theta^N\}$ be an ε-covering of $\mathbb{B}$ in $\|\cdot\|'$. Then

$$\mathbb{B} \subset \bigcup_{j=1}^{N} (\theta^j + \varepsilon \mathbb{B}')$$

Proof of Lemma 2

- Let $\{\theta^1, \ldots, \theta^N\}$ be an ε-covering of $\mathbb{B}$ in $\|\cdot\|'$. Then

$$\mathbb{B} \subset \bigcup_{j=1}^{N} (\theta^j + \varepsilon \mathbb{B}')$$

which gives (by union bound)

$$\mathrm{vol}(\mathbb{B}) \le N\,\mathrm{vol}(\varepsilon \mathbb{B}') = N\,\varepsilon^d \mathrm{vol}(\mathbb{B}')$$

using translation invariance of the Lebesgue measure.

Proof of Lemma 2

- Let $\{\theta^1, \ldots, \theta^N\}$ be an ε-covering of $\mathbb{B}$ in $\|\cdot\|'$. Then

$$\mathbb{B} \subset \bigcup_{j=1}^{N} (\theta^j + \varepsilon \mathbb{B}')$$

  which gives (by union bound)

$$\mathrm{vol}(\mathbb{B}) \leq N \mathrm{vol}(\varepsilon \mathbb{B}') = N \varepsilon^d \mathrm{vol}(\mathbb{B}')$$

  using translation invariance of the Lebesgue measure.

- Hence

$$\frac{\mathrm{vol}(\mathbb{B})}{\mathrm{vol}(\mathbb{B}')} \frac{1}{\varepsilon^d} \leq N.$$

- For the other direction, let $\{\theta^1, \ldots, \theta^M\}$ be a maximal ε-packing with respect to $\| \cdot \|'$.

- For the other direction, let $\{\theta^1, \ldots, \theta^M\}$ be a maximal ε-packing with respect to $\|\cdot\|'$.
- By maximality, it should also be an ε-covering.

- For the other direction, let $\{\theta^1, \ldots, \theta^M\}$ be a maximal ε-packing with respect to $\|\cdot\|'$.
- By maximality, it should also be an ε-covering.
- The sets $\theta^j + \frac{\varepsilon}{2}\mathbb{B}'$ are disjoint and contained in $\mathbb{B} + \frac{\varepsilon}{2}\mathbb{B}'$.

- For the other direction, let $\{\theta^1, \ldots, \theta^M\}$ be a maximal ε-packing with respect to $\|\cdot\|'$.
- By maximality, it should also be an ε-covering.
- The sets $\theta^j + \frac{\varepsilon}{2}\mathbb{B}'$ are disjoint and contained in $\mathbb{B} + \frac{\varepsilon}{2}\mathbb{B}'$. Hence

$$M \,\mathrm{vol}\left(\frac{\varepsilon}{2}\mathbb{B}'\right) = \sum_{i=1}^{M} \mathrm{vol}\left(\theta^i + \frac{\varepsilon}{2}\mathbb{B}'\right) \leq \mathrm{vol}\left(\frac{\varepsilon}{2}\mathbb{B}' + \mathbb{B}\right),$$

- For the other direction, let $\{\theta^1, \ldots, \theta^M\}$ be a maximal ε-packing with respect to $\|\cdot\|'$.
- By maximality, it should also be an ε-covering.
- The sets $\theta^j + \frac{\varepsilon}{2}\mathbb{B}'$ are disjoint and contained in $\mathbb{B} + \frac{\varepsilon}{2}\mathbb{B}'$. Hence

$$M \,\mathrm{vol}\left(\frac{\varepsilon}{2}\mathbb{B}'\right) = \sum_{i=1}^{M} \mathrm{vol}\left(\theta^i + \frac{\varepsilon}{2}\mathbb{B}'\right) \leq \mathrm{vol}\left(\frac{\varepsilon}{2}\mathbb{B}' + \mathbb{B}\right),$$

  and the claim follows:

$$M \leq \frac{\mathrm{vol}\left(\frac{\varepsilon}{2}\mathbb{B}' + \mathbb{B}\right)}{\mathrm{vol}\left(\frac{\varepsilon}{2}\mathbb{B}'\right)}.$$

Dudley's theorem for separable processes

- **Theorem.** Suppose (T, d) is be a separable metric space and $\{X_t, t \in T\}$ is a stochastic process that has continuous sample paths (almost surely) and such that $s, t \in T$ and $u \geq 0$,

$$\mathbb{P}\left(|X_t - X_s| \geq u\right) \leq 2\exp\left(-\frac{u^2}{2d^2(s,t)}\right).$$

  Then for every $t_0 \in T$, we have for some positive constant $C > 0$ that

$$\mathbb{E}\left(\sup_{t \in T}|X_t - X_{t_0}|\right) \leq C\int_0^{D/2} \sqrt{\log N(\epsilon, T, d)}\,d\epsilon, \qquad (1)$$

  where D is the diameter of the metric space (T, d).

- **Proof:** Suppose that T is a finite set. Let T_l be a maximal $D2^{-l}$-packing subset of T, i.e.,

$$\min_{v,u \in T_l} d(v,u) > D2^{-l}.$$

- By construction, $|T_l| = M(D2^{-l}, T, d)$. Clearly, because of the maximality,

$$\max_{v \in T} \min_{u \in T_l} d(u,v) \leq D2^{-l}.$$

- Furthermore, $T_l = T$ for large enough l. Hence, we let

$$N = \min \{ l \geq 1 : T_l = T \}.$$

- Also, for $l \geq 1$, let π_l be the function that assigns $v \in T$ to the point in T_l closest to v. By definition,

$$d(\pi_l(v), v) \leq D2^{-l}$$

for all $v \in T$ and $l \in \mathbb{N}$. We also write $T_0 = \{v_0\}$ and so $\pi_0(v) = v_0$ for all $v \in T$.

- Next, we observe that

$$X_t - X_{t_0} = \sum_{l=1}^{N} (X_{\pi_l(t)} - X_{\pi_{l-1}(t)})$$

  for all $t \in T$.
- It follows that

$$\begin{aligned} \max_{t \in T} (X_t - X_{t_0}) &= \max_{t \in T} \sum_{l=1}^{N} (X_{\pi_l(t)} - X_{\pi_{l-1}(t)}) \\ &\le \sum_{l=1}^{N} \max_{t \in T} (X_{\pi_l(t)} - X_{\pi_{l-1}(t)}) \end{aligned}$$

- and so

$$\mathbb{E}\left(\max_{t \in T} (X_t - X_{t_0})\right) \le \sum_{l=1}^{N} \mathbb{E}\left(\max_{t \in T} (X_{\pi_l(t)} - X_{\pi_{l-1}(t)})\right)$$

- However, notice that for all $u > 0$,

$$\mathbb{P}\left(|X_{\pi_l(t)} - X_{\pi_{l-1}(t)}| \geq u\right) \leq 2\exp\left(\frac{-u^2}{2d(\pi_l(t), \pi_{l-1}(t))^2}\right)$$

- and

$$\begin{aligned} d(\pi_l(t), \pi_{l-1}(t)) &\leq d(\pi_l(t), t) + d(t, \pi_{l-1}(t)) \\ &\leq D2^{-l} + D2^{-(l-1)} \\ &= 3D2^{-l} \end{aligned}$$

- which implies, by the subGaussian maximal inequality,

$$\begin{aligned} \mathbb{E}\left(\max_{t \in T}\left(X_{\pi_l(t)} - X_{\pi_{l-1}(t)}\right)\right) &\leq \frac{3CD}{2^l}\sqrt{\log(2|T_l||T_{l-1}|)} \\ &\leq \frac{3CD}{2^l}\sqrt{\log(2|T_l|^2)} \\ &= \frac{3CD}{2^l}\sqrt{2\log(2M(D2^{-l}, T, d))} \end{aligned}$$

for some constant $C > 0$.

- Therefore, for some constant $\tilde{C} > 0$

$$
\begin{aligned}
&\mathbb{E}\left(\max_{t\in T}\,(X_t - X_{t_0})\right) \\
&\leq \sum_{l=1}^{N} \mathbb{E}\left(\max_{t\in T}\,(X_{\pi_l(t)} - X_{\pi_{l-1}(t)})\right) \\
&\leq \tilde{C}\sum_{l=1}^{N}\frac{D}{2^l}\sqrt{\log(2\,M(D2^{-l}, T, d))} \\
&\leq 2\tilde{C}\sum_{l=1}^{N}\int_{D/2^{l+1}}^{D/2^l}\sqrt{\log(2\,M(r, T, d))}dr \\
&= 2\tilde{C}\int_{D/2^{N+1}}^{D/2}\sqrt{\log(2M(r, T, d))}dr \\
&\leq 2\tilde{C}\int_{0}^{D/2}\sqrt{\log(2M(r, T, d))}dr
\end{aligned}
$$

and the claim follows.

- Suppose now that T is infinite. Let $\tilde{T}$ be a countable subset of T such that (1) holds. For each $k \geq 1$, let $\tilde{T}_k$ be the finite set obtained by taking the first k elements of $\tilde{T}$. We can ensure that $\tilde{T}_k$ contains t_0 for every $k \geq 1$.
- Applying the finite index set version of Dudley's theorem to $\tilde{T}_k$ we obtain that

$$
\begin{aligned}
\mathbb{E}\left(\sup_{t \in \tilde{T}_k} |X_t - X_{t_0}|\right) &\leq C \int_0^{\text{diamter}(\tilde{T}_k)/2} \sqrt{\log N(\epsilon, \tilde{T}_k, d)}\, d\epsilon \\
&\leq C \int_0^{D/2} \sqrt{\log N(\epsilon, T, d)}\, d\epsilon
\end{aligned}
$$

- We have shown that for every $k \geq 1$

$$\mathbb{E}\left(\sup_{t\in\tilde{T}_k}|X_t - X_{t_0}|\right) \leq C\int_0^{D/2}\sqrt{\log N(\epsilon, T, d)}d\epsilon.$$

- Note that the right hand side does not depend on k. Letting $k \to \infty$ on the left hand side, we use Fatou's lemma to obtain

$$\begin{aligned}\mathbb{E}\left(\sup_{t\in\tilde{T}}|X_t - X_{t_0}|\right) &= \lim_{k\to\infty}\mathbb{E}\left(\sup_{t\in\tilde{T}_k}|X_t - X_{t_0}|\right)\\ &\leq C\int_0^{D/2}\sqrt{\log N(\epsilon, T, d)}d\epsilon.\end{aligned}$$

- However, by the continuty of paths of X_t, we have that

$$\mathbb{E}\left(\sup_{t\in T}|X_t - X_{t_0}|\right) = \mathbb{E}\left(\sup_{t\in\tilde{T}}|X_t - X_{t_0}|\right),$$

because, almost surely,

$$\sup_{t\in T}|X_t - X_{t_0}| = \sup_{t\in\tilde{T}}|X_t - X_{t_0}|.$$

- To see why, a.s.,

$$\sup_{t\in T}|X_t - X_{t_0}| = \sup_{t\in\tilde{T}}|X_t - X_{t_0}|,$$

notice that since $\tilde{T} \subset T$,

$$a := \sup_{t\in T}|X_t - X_{t_0}| \geq \sup_{t\in\tilde{T}}|X_t - X_{t_0}|.$$

- Also, for $\delta > 0$, there exists $t \in T$ such that

$$||X_t - X_{t_0}| - a| < \delta/2$$

by definition of supremum.

- However, since X has continuous paths and $\tilde{T}$ is dense in T, there exists t' such that $|X_t - X_{t'}| < \delta/2$ and so

$$||X_{t'} - X_{t_0}| - a| \leq ||X_{t'} - X_{t_0}| - |X_t - X_{t_0}|| + ||X_t - X_{t_0}| - a| < \delta$$

which shows the claim.

Example

- Let $K \subset \mathbb{R}^n$ and $\epsilon \sim N(0, I_n)$. Then for $r > 0$ and $\theta_0 \in K$ define
$$Z(\epsilon) = \sup_{\theta \in K \,:\, \|\theta - \theta_0\|_2 \le r} \langle \epsilon, \theta - \theta_0 \rangle.$$
This is called an empirical process.

Example

- Let $K \subset \mathbb{R}^n$ and $\epsilon \sim N(0, I_n)$. Then for $r > 0$ and $\theta_0 \in K$ define
$$Z(\epsilon) \;=\; \sup_{\theta \in K \,:\, \|\theta - \theta_0\|_2 \leq r} \langle \epsilon, \theta - \theta_0 \rangle.$$
This is called an empirical process.
- It be can be shown that the function $\epsilon \to Z(\epsilon)$ is r-Lipschitz. Hence, for any $t > 0$ it holds that
$$\mathbb{P}(|Z(\epsilon) - \mathbb{E}(Z(\epsilon))| > t|) \;\leq\; 2\exp\left(-\frac{t^2}{2r^2}\right).$$

Example

- Let $K \subset \mathbb{R}^n$ and $\epsilon \sim N(0, I_n)$. Then for $r > 0$ and $\theta_0 \in K$ define
$$Z(\epsilon) \;=\; \sup_{\theta \in K \,:\, \|\theta - \theta_0\|_2 \leq r} \langle \epsilon, \theta - \theta_0 \rangle.$$
This is called an empirical process.
- It be can be shown that the function $\epsilon \to Z(\epsilon)$ is r-Lipschitz. Hence, for any $t > 0$ it holds that
$$\mathbb{P}(|Z(\epsilon) - \mathbb{E}(Z(\epsilon))| > t|) \;\leq\; 2\exp\left(-\frac{t^2}{2r^2}\right).$$
- It remains to bound $\mathbb{E}(Z(\epsilon))$, which is called the local Gaussian complexity of the set $K - \theta_0$.

Example

- Let $K \subset \mathbb{R}^n$ and $\epsilon \sim N(0, I_n)$. Then for $r > 0$ and $\theta_0 \in K$ define
$$Z(\epsilon) \;=\; \sup_{\theta \in K \,:\, \|\theta - \theta_0\|_2 \leq r} \langle \epsilon, \theta - \theta_0 \rangle .$$
This is called an empirical process.
- It be can be shown that the function $\epsilon \rightarrow Z(\epsilon)$ is r-Lipschitz. Hence, for any $t > 0$ it holds that
$$\mathbb{P}(|Z(\epsilon) - \mathbb{E}(Z(\epsilon))| > t|) \;\leq\; 2 \exp\left(-\frac{t^2}{2r^2}\right).$$
- It remains to bound $\mathbb{E}(Z(\epsilon))$, which is called the local Gaussian complexity of the set $K - \theta_0$. Towards that end, define $X_\theta = \langle \epsilon, \theta \rangle$ and $T = K \cap \{\theta \,:\, \|\theta - \theta_0\|_2 \leq r\}$, and notice that by Dudley's inequality,
$$\mathbb{E}(Z(\epsilon)) \;\lesssim\; \int_0^r \sqrt{\log N(\kappa, T, \|\cdot\|_2)}\, d\kappa .$$

Example

- Let $K \subset \mathbb{R}^n$ and $\epsilon \sim N(0, I_n)$. Then for $r > 0$ and $\theta_0 \in K$ define
$$Z(\epsilon) = \sup_{\theta \in K \,:\, \|\theta - \theta_0\|_2 \leq r} \langle \epsilon, \theta - \theta_0 \rangle.$$
This is called an empirical process.
- It be can be shown that the function $\epsilon \rightarrow Z(\epsilon)$ is r-Lipschitz. Hence, for any $t > 0$ it holds that
$$\mathbb{P}(|Z(\epsilon) - \mathbb{E}(Z(\epsilon))| > t|) \leq 2 \exp\left(-\frac{t^2}{2r^2}\right).$$
- It remains to bound $\mathbb{E}(Z(\epsilon))$, which is called the local Gaussian complexity of the set $K - \theta_0$. Towards that end, define $X_\theta = \langle \epsilon, \theta \rangle$ and $T = K \cap \{\theta \,:\, \|\theta - \theta_0\|_2 \leq r\}$, and notice that by Dudley's inequality,
$$\mathbb{E}(Z(\epsilon)) \lesssim \int_0^r \sqrt{\log N(\kappa, T, \|\cdot\|_2)}\, d\kappa.$$
- So one only needs to bound the covering numbers of T.

Lipschitz function class

- The class of Lipschitz functions:

$$\mathcal{F}_L := \left\{ f \in [0,1] \to \mathbb{R} \mid \|f\|_{lip} \le L \right\}.$$

Recall $\|f\|_{lip} \le L$ iff $|f(x) - f(y)| \le L|x - y|$ for all $x, y \in [0, 1]$.

Lipschitz function class

- The class of Lipschitz functions:

$$\mathcal{F}_L := \left\{ f \in [0,1] \to \mathbb{R} \mid \|f\|_{lip} \leq L \right\}.$$

  Recall $\|f\|_{lip} \leq L$ iff $|f(x) - f(y)| \leq L|x - y|$ for all $x, y \in [0,1]$.
- **Theorem 1.** The entropy number of the Lipschitz class is bounded as

$$\log N(\delta, \mathcal{F}_L, \|\cdot\|_\infty) \asymp \frac{L}{\delta} \quad \text{for small } \delta.$$

Lipschitz function class

- The class of Lipschitz functions:
$$\mathcal{F}_L := \{ f \in [0,1] \to \mathbb{R} \mid \|f\|_{lip} \le L \}.$$
Recall $\|f\|_{lip} \le L$ iff $|f(x) - f(y)| \le L|x - y|$ for all $x, y \in [0,1]$.
- **Theorem 1.** The entropy number of the Lipschitz class is bounded as
$$\log N(\delta, \mathcal{F}_L, \|\cdot\|_\infty) \asymp \frac{L}{\delta} \quad \text{for small } \delta.$$
- Covering number is exponential in $1/\delta$: $N(\delta, \mathcal{F}_L, \|\cdot\|_\infty) \ge e^{cL/\delta}$.

Lipschitz function class

- The class of Lipschitz functions:

$$\mathcal{F}_L := \left\{ f \in [0,1] \to \mathbb{R} \mid \|f\|_{lip} \leq L \right\}.$$

  Recall $\|f\|_{lip} \leq L$ iff $|f(x) - f(y)| \leq L|x - y|$ for all $x, y \in [0,1]$.
- **Theorem 1.** The entropy number of the Lipschitz class is bounded as

$$\log N(\delta, \mathcal{F}_L, \|\cdot\|_\infty) \asymp \frac{L}{\delta} \quad \text{for small } \delta.$$

- Covering number is exponential in $1/\delta$: $N(\delta, \mathcal{F}_L, \|\cdot\|_\infty) \geq e^{cL/\delta}$.
- We prove the lower bound

$$N(\delta, \mathcal{F}_L, \|\cdot\|_\infty) \;\geq\; M(2\delta, \mathcal{F}_L, \|\cdot\|_\infty) \;\geq\; 2^{L/\delta}.$$

Lipschitz function class

- The class of Lipschitz functions:

$$\mathcal{F}_L := \left\{ f \in [0,1] \to \mathbb{R} \mid \|f\|_{lip} \leq L \right\}.$$

  Recall $\|f\|_{lip} \leq L$ iff $|f(x) - f(y)| \leq L|x - y|$ for all $x, y \in [0,1]$.
- **Theorem 1.** The entropy number of the Lipschitz class is bounded as

$$\log N(\delta, \mathcal{F}_L, \|\cdot\|_\infty) \asymp \frac{L}{\delta} \quad \text{for small} \quad \delta.$$

- Covering number is exponential in $1/\delta$: $N(\delta, \mathcal{F}_L, \|\cdot\|_\infty) \geq e^{cL/\delta}$.
- We prove the lower bound

$$N(\delta, \mathcal{F}_L, \|\cdot\|_\infty) \geq M(2\delta, \mathcal{F}_L, \|\cdot\|_\infty) \geq 2^{L/\delta}.$$

- Idea is to embed $\{-1, 1\}^M$ in $\mathcal{F}_L$ for M as large as we can get.

- $M = \lfloor 1/\varepsilon \rfloor$, $x_i = (i-1)\varepsilon$, $i = 1, \ldots, M$, and $x_{M+1} = M\varepsilon < 1$.

- $M = \lfloor 1/\varepsilon \rfloor$, $x_i = (i-1)\varepsilon$, $i = 1, \ldots, M$, and $x_{M+1} = M\varepsilon < 1$.
- For $\beta \in \{-1, 1\}^M$, let

$$f_\beta(y) \, = \, L\varepsilon \sum_{i=1}^{M} \beta_i \phi \left(\frac{y - x_i}{\varepsilon} \right)$$

where $\phi(u) = u$ for $u \in [0, 1]$, with continuous constant extension to $\mathbb{R}$.

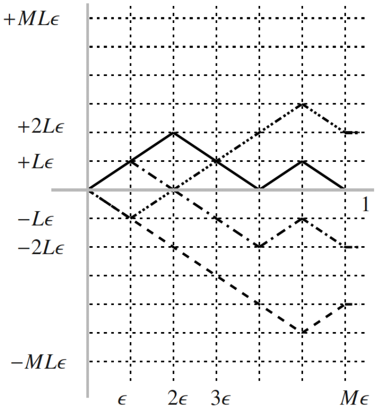


Figure 5.2 The function class $\{f_\beta, \beta \in \{-1,+1\}^M\}$ used to construct a packing of the Lipschitz class $\mathscr{F}_L$. Each function is piecewise linear over the intervals $[0,\epsilon]$, $[\epsilon,2\epsilon]$, $\ldots$, $[(M-1)\epsilon, M\epsilon]$ with slope either $+L$ or $-L$. There are 2^M functions in total, where $M = \lfloor 1/\epsilon \rfloor$.

- For $\beta \in \{-1, 1\}^M$, let

$$f_\beta(y) \;=\; L\varepsilon \sum_{i=1}^{M} \beta_i \phi\left(\frac{y - x_i}{\varepsilon}\right)$$

where $\phi(u) = u$ for $u \in [0, 1]$, with continuous constant extension to $\mathbb{R}$.

- For $\beta \in \{-1, 1\}^M$, let

$$f_\beta(y) = L\varepsilon \sum_{i=1}^{M} \beta_i \phi\left(\frac{y - x_i}{\varepsilon}\right)$$

  where $\phi(u) = u$ for $u \in [0, 1]$, with continuous constant extension to $\mathbb{R}$.
- Can verify that $\beta \in \mathcal{F}_L$.

- For $\beta \in \{-1, 1\}^M$, let

$$f_\beta(y) \;=\; L\varepsilon \sum_{i=1}^{M} \beta_i \phi\left(\frac{y - x_i}{\varepsilon}\right)$$

  where $\phi(u) = u$ for $u \in [0, 1]$, with continuous constant extension to $\mathbb{R}$.
- Can verify that $\beta \in \mathcal{F}_L$.
- $\{f_\beta \,:\, \beta \in \{-1, 1\}^M\}$ is a $2L\varepsilon$-packing of $\mathcal{F}_L$ in uniform norm, i.e. $\|f_\beta - f_{\beta'}\|_\infty \geq 2L\varepsilon$ for all $\beta \neq \beta'$.

- For $\beta \in \{-1, 1\}^M$, let

$$f_\beta(y) = L\varepsilon \sum_{i=1}^{M} \beta_i \phi\left(\frac{y - x_i}{\varepsilon}\right)$$

  where $\phi(u) = u$ for $u \in [0, 1]$, with continuous constant extension to $\mathbb{R}$.
- Can verify that $\beta \in \mathcal{F}_L$.
- $\{f_\beta : \beta \in \{-1, 1\}^M\}$ is a $2L\varepsilon$-packing of $\mathcal{F}_L$ in uniform norm, i.e. $\|f_\beta - f_{\beta'}\|_\infty \geq 2L\varepsilon$ for all $\beta \neq \beta'$.
  Take $\varepsilon = \delta/L$. We have 2δ-packing of size $\approx 2^{L/\delta}$.

Lipschitz Higher dimensions

The preceding example can be extended to Lipschitz functions on the unit cube in higher dimensions, meaning real-valued functions on $[0, 1]^d$ such that

$$|f(x) - f(y)| \leq L \|x - y\|_\infty \qquad \text{for all } x, y \in [0, 1]^d, \tag{5.15}$$

a class that we denote by $\mathscr{F}_L([0, 1]^d)$. An extension of our argument can then be used to show that

$$\log N_\infty(\delta\,;\, \mathscr{F}_L([0, 1]^d) \asymp (L/\delta)^d.$$

It is worth contrasting the *exponential dependence* of this metric entropy on the dimension d, as opposed to the linear dependence that we saw earlier for simpler sets (e.g., such as d-dimensional unit balls). This is a dramatic manifestation of the curse of dimensionality.

Lipschitz Higher dimensions

Example 5.11 (Higher-order smoothness classes) We now consider an example of a function class based on controlling higher-order derivatives. For a suitably differentiable function f, let us adopt the notation $f^{(k)}$ to mean the kth derivative. (Of course, $f^{(0)} = f$ in this notation.) For some integer α and parameter $\gamma \in (0, 1]$, consider the class of functions $f \colon [0, 1] \to \mathbb{R}$ such that

$$|f^{(j)}(x)| \leq C_j \qquad \text{for all } x \in [0,1],\ j = 0, 1, \ldots, \alpha, \tag{5.16a}$$

$$|f^{(\alpha)}(x) - f^{(\alpha)}(x')| \leq L\,|x - x'|^{\gamma}, \qquad \text{for all } x, x' \in [0,1]. \tag{5.16b}$$

We claim that the metric entropy of this function class $\mathscr{F}_{\alpha,\gamma}$ scales as

$$\log N(\delta\,;\, \mathscr{F}_{\alpha,\gamma}, \|\cdot\|_\infty) \asymp \left(\frac{1}{\delta}\right)^{\frac{1}{\alpha+\gamma}}. \tag{5.17}$$

(Here we have absorbed the dependence on the constants C_j and L into the order notation.) Note that this claim is consistent with our calculation in Example 5.10, which is essentially the same as the class $\mathscr{F}_{0,1}$.

Let us prove the lower bound in the claim (5.17). As in the previous example, we do so by constructing a packing $\{f_\beta, \beta \in \{-1, +1\}^M\}$ for a suitably chosen integer M. Define the function

$$\phi(y) := \begin{cases} c\, 2^{2(\alpha+\gamma)} y^{\alpha+\gamma}(1-y)^{\alpha+\gamma} & \text{for } y \in [0,1], \\ 0 & \text{otherwise.} \end{cases} \tag{5.18}$$

Nonparametric regression (least-squares)

- Fixed design points $\{x_i\}_{i=1}^n$ and response $\{y_i\}_{i=1}^n$

$$y_i = f^*(x_i) + v_i, \quad i = 1, \dots, n.$$

Nonparametric regression (least-squares)

- Fixed design points $\{x_i\}_{i=1}^n$ and response $\{y_i\}_{i=1}^n$

$$y_i = f^*(x_i) + v_i, \quad i = 1, \ldots, n.$$

- General: $f^*(x) = \mathbb{E}\,(Y|X = x)$ and $V = Y - \mathbb{E}(Y|X = x)$, so that $Y = f^*(x) + V$ conditional on $X = x$.

Nonparametric regression (least-squares)

- Fixed design points $\{x_i\}_{i=1}^n$ and response $\{y_i\}_{i=1}^n$

$$y_i = f^*(x_i) + v_i, \quad i = 1, \ldots, n.$$

- General: $f^*(x) = \mathbb{E}(Y|X=x)$ and $V = Y - \mathbb{E}(Y|X=x)$, so that $Y = f^*(x) + V$ conditional on $X = x$.
- Usual: just assume $v_i = \sigma w_i$ where $w_i \stackrel{i.i.d.}{\sim} N(0,1)$.

Nonparametric regression (least-squares)

- Fixed design points $\{x_i\}_{i=1}^n$ and response $\{y_i\}_{i=1}^n$

$$y_i = f^*(x_i) + v_i, \quad i = 1, \dots, n.$$

- General: $f^*(x) = \mathbb{E}(Y|X = x)$ and $V = Y - \mathbb{E}(Y|X = x)$, so that $Y = f^*(x) + V$ conditional on $X = x$.
- Usual: just assume $v_i = \sigma w_i$ where $w_i \overset{i.i.d.}{\sim} N(0, 1)$.
- Least-squares (LS) estimators: for some class of function $\mathcal{F}$,

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \right\}$$

- $\beta, x \in \mathbb{R}^d$, and $f_\beta(x) = \langle \beta, x \rangle$.

- $\beta, x \in \mathbb{R}^d$, and $f_\beta(x) = \langle \beta, x \rangle$. The function class:

$$\begin{aligned} \mathcal{F}_{\mathcal{C}}^{lin} &= \{f_\beta \, : \, \beta \in \mathcal{C}\} \\ &= \{x \to \langle \beta, x \rangle \, : \, \beta \in \mathcal{C}\}. \end{aligned}$$

Linear functions with normal vectors β belonging to $\mathcal{C} \subset \mathbb{R}^d$.

- $\beta, x \in \mathbb{R}^d$, and $f_\beta(x) = \langle \beta, x \rangle$. The function class:

$$\begin{aligned} \mathcal{F}_{\mathcal{C}}^{lin} &= \{f_\beta \,:\, \beta \in \mathcal{C}\} \\ &= \{x \to \langle \beta, x \rangle \,:\, \beta \in \mathcal{C}\}. \end{aligned}$$

  Linear functions with normal vectors β belonging to $\mathcal{C} \subset \mathbb{R}^d$.

- Optimizing $\ell(f)$ over $f \in \mathcal{F}$, equivalent to optimizing $\ell(f_\beta)$ over $\beta \in \mathcal{C}$.

- $\beta, x \in \mathbb{R}^d$, and $f_\beta(x) = \langle \beta, x \rangle$. The function class:

$$\begin{aligned} \mathcal{F}_{\mathcal{C}}^{lin} &= \{f_\beta : \beta \in \mathcal{C}\} \\ &= \{x \to \langle \beta, x \rangle : \beta \in \mathcal{C}\}. \end{aligned}$$

  Linear functions with normal vectors β belonging to $\mathcal{C} \subset \mathbb{R}^d$.
- Optimizing $\ell(f)$ over $f \in \mathcal{F}$, equivalent to optimizing $\ell(f_\beta)$ over $\beta \in \mathcal{C}$.
- Let $X \in \mathbb{R}^d$ be the design matrix; the rows are $x_i^\top$.

- $\beta, x \in \mathbb{R}^d$, and $f_\beta(x) = \langle \beta, x \rangle$. The function class:

$$\begin{aligned}\mathcal{F}_{\mathcal{C}}^{lin} &= \{f_\beta \,:\, \beta \in \mathcal{C}\} \\ &= \{x \to \langle \beta, x \rangle \,:\, \beta \in \mathcal{C}\}.\end{aligned}$$

  Linear functions with normal vectors β belonging to $\mathcal{C} \subset \mathbb{R}^d$.
- Optimizing $\ell(f)$ over $f \in \mathcal{F}$, equivalent to optimizing $\ell(f_\beta)$ over $\beta \in \mathcal{C}$.
- Let $X \in \mathbb{R}^d$ be the design matrix; the rows are $x_i^\top$.
- Constrained LS problem:

$$\hat{\beta} \in \arg\min_{\beta \in \mathcal{C}} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - \langle x_i, \beta \rangle)^2 \right\}$$

- $\beta, x \in \mathbb{R}^d$, and $f_\beta(x) = \langle \beta, x \rangle$. The function class:

$$\begin{aligned}\mathcal{F}_{\mathcal{C}}^{lin} &= \{f_\beta \,:\, \beta \in \mathcal{C}\} \\ &= \{x \to \langle \beta, x \rangle \,:\, \beta \in \mathcal{C}\}.\end{aligned}$$

  Linear functions with normal vectors β belonging to $\mathcal{C} \subset \mathbb{R}^d$.
- Optimizing $\ell(f)$ over $f \in \mathcal{F}$, equivalent to optimizing $\ell(f_\beta)$ over $\beta \in \mathcal{C}$.
- Let $X \in \mathbb{R}^d$ be the design matrix; the rows are $x_i^\top$.
- Constrained LS problem:

$$\begin{aligned}\hat{\beta} &\in \operatorname*{arg\,min}_{\beta \in \mathcal{C}} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - \langle x_i, \beta \rangle)^2 \right\} \\ &= \operatorname*{arg\,min}_{\beta \in \mathcal{C}} \left\{ \frac{1}{n} \|y - X\beta\|_2^2 \right\}.\end{aligned}$$

- Constrained ridge regression:

$$\mathcal{C} = \left\{ \beta \in \mathbb{R}^d \; : \; \|\beta\|_2^2 \leq R_2 \right\} \; = \; R_2^{1/2} \mathbb{B}_2^d.$$

- Constrained ridge regression:

$$\mathcal{C} = \left\{\beta \in \mathbb{R}^d \;:\; \|\beta\|_2^2 \leq R_2\right\} \;=\; R_2^{1/2}\mathbb{B}_2^d.$$

- Constrained ℓ_q ball:

$$\mathcal{C} = \left\{\beta \in \mathbb{R}^d \;:\; \|\beta\|_q^q \leq R_q\right\} \;=\; R_q^{1/q}\mathbb{B}_q^d,$$

where $\|\beta\|_q^q = \sum_{i=1}^d |\beta_i|^q$.

- Constrained ridge regression:

$$\mathcal{C} = \left\{ \beta \in \mathbb{R}^d \,:\, \|\beta\|_2^2 \le R_2 \right\} \;=\; R_2^{1/2}\mathbb{B}_2^d.$$

- Constrained ℓ_q ball:

$$\mathcal{C} = \left\{ \beta \in \mathbb{R}^d \,:\, \|\beta\|_q^q \le R_q \right\} \;=\; R_q^{1/q}\mathbb{B}_q^d,$$

  where $\|\beta\|_q^q = \sum_{i=1}^d |\beta_i|^q$.
- For $q \in (0,1)$, $\mathbb{B}_q^d$ are nonconvex and approximate the ℓ_0 ball as $q \to 0$.

- Constrained ridge regression:

$$\mathcal{C} = \left\{ \beta \in \mathbb{R}^d \, : \, \|\beta\|_2^2 \leq R_2 \right\} \; = \; R_2^{1/2} \mathbb{B}_2^d.$$

- Constrained ℓ_q ball:

$$\mathcal{C} = \left\{ \beta \in \mathbb{R}^d \, : \, \|\beta\|_q^q \leq R_q \right\} \; = \; R_q^{1/q} \mathbb{B}_q^d,$$

  where $\|\beta\|_q^q = \sum_{i=1}^d |\beta_i|^q$.
- For $q \in (0, 1)$, $\mathbb{B}_q^d$ are nonconvex and approximate the ℓ_0 ball as $q \to 0$.
- They become smaller in volume as $q \to 0$.

- Constrained ridge regression:

$$\mathcal{C} = \left\{\beta \in \mathbb{R}^d \,:\, \|\beta\|_2^2 \leq R_2\right\} \;=\; R_2^{1/2}\mathbb{B}_2^d.$$

- Constrained ℓ_q ball:

$$\mathcal{C} = \left\{\beta \in \mathbb{R}^d \,:\, \|\beta\|_q^q \leq R_q\right\} \;=\; R_q^{1/q}\mathbb{B}_q^d,$$

  where $\|\beta\|_q^q = \sum_{i=1}^d |\beta_i|^q$.
- For $q \in (0, 1)$, $\mathbb{B}_q^d$ are nonconvex and approximate the ℓ_0 ball as $q \to 0$.
- They become smaller in volume as $q \to 0$.
- $q = 1$ correspond to constrained form of Lasso.

Cubic smoothing splines

- A ball in Sobolev space $H^2([0,1])$:

$$\mathcal{F}(R) := \{f : [0,1] \to \mathbb{R} : \|f''\|_{L^2}^2 \leq R\}$$

where $\|f''\|_{L^2}^2 = \int_0^1 [f''(x)]^2 dx$.

Cubic smoothing splines

- A ball in Sobolev space $H^2([0,1])$:

$$\mathcal{F}(R) := \{f : [0,1] \to \mathbb{R} : \|f''\|_{L^2}^2 \leq R\}$$

  where $\|f''\|_{L^2}^2 = \int_0^1 [f''(x)]^2 dx$.
- Penalized estimator:

$$\hat{f} \in \arg\min_f \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \int_0^1 [f''(x)]^2 dx \right\}.$$

Cubic smoothing splines

- A ball in Sobolev space $H^2([0,1])$:

$$\mathcal{F}(R) := \{f : [0,1] \to \mathbb{R} : \|f''\|_{L^2}^2 \leq R\}$$

  where $\|f''\|_{L^2}^2 = \int_0^1 [f''(x)]^2 dx$.
- Penalized estimator:

$$\hat{f} \in \arg\min_f \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda_n \int_0^1 [f''(x)]^2 dx \right\}.$$

- Minimizer is a cubic spline: piecewise cubic between design points (knots), second derivative is continuous, third derivative has jump discontinuity at the knots.

Neural networks

- For a vector $v \in \mathbb{R}^r$ we define the function $\phi_v : \mathbb{R}^r \to \mathbb{R}^r$ as

$$\phi_v \begin{pmatrix} a_1 \\ \vdots \\ a_r \end{pmatrix} = \begin{pmatrix} \phi(a_1 - v_1) \\ \vdots \\ \phi(a_r - v_r) \end{pmatrix},$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is an activation function.

Neural networks

- For a vector $v \in \mathbb{R}^r$ we define the function $\phi_v : \mathbb{R}^r \to \mathbb{R}^r$ as

$$\phi_v \begin{pmatrix} a_1 \\ \vdots \\ a_r \end{pmatrix} = \begin{pmatrix} \phi(a_1 - v_1) \\ \vdots \\ \phi(a_r - v_r) \end{pmatrix},$$

  where $\phi : \mathbb{R} \to \mathbb{R}$ is an activation function.

- A popular choice of activation function is ReLU: $\phi(x) = \max\{0, x\}$.

Neural networks

- With the previous notation, we consider neural network functions $f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}$ of the form

$$f(x) = A^{(L)}\phi_{V_L} \circ A^{(L-1)}\phi_{V_{L-1}} \circ \cdots \circ A^{(1)}\phi_{V_1} \circ A^{(0)}x, \qquad (2)$$

  where $\circ$ denotes the composition of functions, and $A^{(i)} \in \mathbb{R}^{p_{i+1} \times p_i}$, $V_i \in \mathbb{R}^{p_i}$, $p_0, \dots, p_{L+1} \in \mathbb{N}$ for $i \in \{0, 1, \dots, L+1\}$. Here the matrices $\{A^{(i)}\}$ are the weights in the network, L is the number of layers, and $(p_0, \dots, p_{L+1})^\top \in \mathbb{R}^{L+2}$ the width vector.

Neural networks

- With the previous notation, we consider neural network functions $f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}$ of the form

$$f(x) = A^{(L)}\phi_{V_L} \circ A^{(L-1)}\phi_{V_{L-1}} \circ \cdots \circ A^{(1)}\phi_{V_1} \circ A^{(0)}x, \qquad (2)$$

  where $\circ$ denotes the composition of functions, and $A^{(i)} \in \mathbb{R}^{p_{i+1} \times p_i}$, $V_i \in \mathbb{R}^{p_i}$, $p_0, \ldots, p_{L+1} \in \mathbb{N}$ for $i \in \{0, 1, \ldots, L+1\}$. Here the matrices $\{A^{(i)}\}$ are the weights in the network, L is the number of layers, and $(p_0, \ldots, p_{L+1})^\top \in \mathbb{R}^{L+2}$ the width vector.

- With $p_0 = d$, $p_{L+1} = 1$ we can let $\mathcal{F}$ be the class of functions of the form (2).

Neural networks

- With the previous notation, we consider neural network functions $f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}$ of the form

$$f(x) = A^{(L)}\phi_{V_L} \circ A^{(L-1)}\phi_{V_{L-1}} \circ \cdots \circ A^{(1)}\phi_{V_1} \circ A^{(0)}x, \qquad (2)$$

  where $\circ$ denotes the composition of functions, and $A^{(i)} \in \mathbb{R}^{p_{i+1} \times p_i}$, $V_i \in \mathbb{R}^{p_i}$, $p_0, \ldots, p_{L+1} \in \mathbb{N}$ for $i \in \{0, 1, \ldots, L+1\}$. Here the matrices $\{A^{(i)}\}$ are the weights in the network, L is the number of layers, and $(p_0, \ldots, p_{L+1})^\top \in \mathbb{R}^{L+2}$ the width vector.

- With $p_0 = d$, $p_{L+1} = 1$ we can let $\mathcal{F}$ be the class of functions of the form (2).

- We can define the estimator

$$\hat{f} \in \underset{f \in \mathcal{F}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \right\}.$$

Normal means model

- The normal means model is the prototypical example of a nonparametric model:

$$y_i = \theta_i^* + \sigma w_i, \quad w_i \overset{i.i.d}{\sim} N(0,1), \quad i = 1, \ldots, d, \tag{3}$$

  or compactly $y = \theta^* + \sigma w$ for $w \sim N(0, I_d)$. We assume that $\theta^* \in \Theta \subset \mathbb{R}^d$.

Reduction of nonparametric regression to normal means model

- We can map this model to the normal means model with $d = n$, by taking $\theta_i^* = f^*(x_i)/\sqrt{n}$ for $i = 1, \ldots, n$.

Reduction of nonparametric regression to normal means model

- We can map this model to the normal means model with $d = n$, by taking $\theta_i^* = f^*(x_i)/\sqrt{n}$ for $i = 1, \ldots, n$.
- The function class induces the following parameter space in $\mathbb{R}^d$:

$$\Theta := \{\Phi f \, : \, f \in \mathcal{F}\}, \quad \Phi f \, := \, \frac{1}{\sqrt{n}} \left(f(x_1), \ldots, f(x_n)\right)^T .$$

Reduction of nonparametric regression to normal means model

- We can map this model to the normal means model with $d = n$, by taking $\theta_i^* = f^*(x_i)/\sqrt{n}$ for $i = 1, \ldots, n$.
- The function class induces the following parameter space in $\mathbb{R}^d$:

$$\Theta := \{\Phi f \,:\, f \in \mathcal{F}\}, \quad \Phi f \,:=\, \frac{1}{\sqrt{n}}\left(f(x_1), \ldots, f(x_n)\right)^T.$$

- That is, Θ is the image of $\mathcal{F}$ under map $\Phi \,:\, \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^d$, in short $\Phi\mathcal{F} = \Theta$.

Reduction of nonparametric regression to normal means model

- We can map this model to the normal means model with $d = n$, by taking $\theta_i^* = f^*(x_i)/\sqrt{n}$ for $i = 1, \ldots, n$.
- The function class induces the following parameter space in $\mathbb{R}^d$:

$$\Theta := \{\Phi f \,:\, f \in \mathcal{F}\}, \quad \Phi f := \frac{1}{\sqrt{n}} \left(f(x_1), \ldots, f(x_n)\right)^T.$$

- That is, Θ is the image of $\mathcal{F}$ under map $\Phi : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^d$, in short $\Phi\mathcal{F} = \Theta$.
- We can rewrite the model as

$$\tilde{y}_i = \theta_i^* + \tilde{\sigma} w_i, \quad \text{where} \quad \tilde{y}_i = \frac{y_i}{\sqrt{n}} \quad \text{and} \quad \tilde{\sigma} = \frac{\sigma}{\sqrt{n}}.$$

Reduction of nonparametric regression to normal means model

- We can map this model to the normal means model with $d = n$, by taking $\theta_i^* = f^*(x_i)/\sqrt{n}$ for $i = 1, \ldots, n$.
- The function class induces the following parameter space in $\mathbb{R}^d$:

$$\Theta := \{\Phi f \,:\, f \in \mathcal{F}\}, \quad \Phi f \;:=\; \frac{1}{\sqrt{n}}\,(f(x_1), \ldots, f(x_n))^T\,.$$

- That is, Θ is the image of $\mathcal{F}$ under map $\Phi \,:\, \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^d$, in short $\Phi\mathcal{F} = \Theta$.
- We can rewrite the model as

$$\tilde{y}_i \;=\; \theta_i^* + \tilde{\sigma} w_i, \quad \text{where} \quad \tilde{y}_i = \frac{y_i}{\sqrt{n}} \quad \text{and} \quad \tilde{\sigma} = \frac{\sigma}{\sqrt{n}}.$$

- Let $\hat{\theta} = \Phi\hat{f}$ so that $\hat{\theta}_i = \hat{f}(x_i)/\sqrt{n}$.

Reduction of nonparametric regression to normal means model

- We can map this model to the normal means model with $d = n$, by taking $\theta_i^* = f^*(x_i)/\sqrt{n}$ for $i = 1, \ldots, n$.
- The function class induces the following parameter space in $\mathbb{R}^d$:

$$\Theta := \{\Phi f \,:\, f \in \mathcal{F}\}, \quad \Phi f := \frac{1}{\sqrt{n}} \left(f(x_1), \ldots, f(x_n)\right)^T.$$

- That is, Θ is the image of $\mathcal{F}$ under map $\Phi : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^d$, in short $\Phi\mathcal{F} = \Theta$.
- We can rewrite the model as

$$\tilde{y}_i = \theta_i^* + \tilde{\sigma} w_i, \quad \text{where} \quad \tilde{y}_i = \frac{y_i}{\sqrt{n}} \quad \text{and} \quad \tilde{\sigma} = \frac{\sigma}{\sqrt{n}}.$$

- Let $\hat{\theta} = \Phi \hat{f}$ so that $\hat{\theta}_i = \hat{f}(x_i)/\sqrt{n}$. We note that

$$\hat{\theta} \in \operatorname*{arg\,min}_{\theta \in \Theta} \sum_{i=1}^{n} (\tilde{y}_i - \theta_i)^2.$$

Notation

- For any $\Theta \subset \mathbb{R}^d$, let

$$\Theta(u) := \Theta \cap \mathbb{B}_2(u) = \{\theta \in \Theta : \|\theta\|_2 \leq u\}.$$

Notation

- For any $\Theta \subset \mathbb{R}^d$, let
$$\Theta(u) := \Theta \cap \mathbb{B}_2(u) = \{\theta \in \Theta : \|\theta\|_2 \le u\}.$$
- We also write
$$\Theta_{\theta_0} := \Theta - \theta_0 = \{\theta - \theta_0 : \theta \in \Theta\}, \quad \Theta_{\theta_0}(u) := \Theta_{\theta_0} \cap \mathbb{B}_2(u).$$
Note that Θ_{θ_0} is Θ translated to be “centered" at θ_0.

Notation

- For any $\Theta \subset \mathbb{R}^d$, let

$$\Theta(u) \;:=\; \Theta \cap \mathbb{B}_2(u) \;=\; \{\theta \in \Theta \;:\; \|\theta\|_2 \;\leq\; u\}.$$

- We also write

$$\Theta_{\theta_0} \;:=\; \Theta - \theta_0 \;=\; \{\theta - \theta_0 \;:\; \theta \in \Theta\}, \quad \Theta_{\theta_0}(u) := \Theta_{\theta_0} \cap \mathbb{B}_2(u).$$

  Note that Θ_{θ_0} is Θ translated to be "centered" at θ_0.
- Recall the Gaussian complexity of set $T \subset \mathbb{R}^d$, defined as

$$\gamma(T) \;=\; \mathbb{E}\left(\sup_{\theta \in T} |\langle w, \theta \rangle|\right), \quad w \sim N(0, I_d).$$

General theorem

- For $y \in \mathbb{R}^d$, we consider the following projection estimator

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} \|y - \theta\|_2^2. \tag{4}$$

General theorem

- For $y \in \mathbb{R}^d$, we consider the following projection estimator

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} \|y - \theta\|_2^2. \tag{4}$$

- **Theorem 1.** Assume that y is generated from the normal means model in (3), with $\theta^* \in \Theta \subset \mathbb{R}^d$ such that Θ_{θ^*} is star-shaped with respect to the origin. Consider the projection estimator $\hat{\theta}$ in (4). Then, for any $u > 0$ and $t \geq 0$, with probability at least $1 - e^{-t^2/2}$

$$\|\theta^* - \hat{\theta}\| \;\leq\; \max\left\{\frac{2\sigma G(u,t)}{u}, \sqrt{2\sigma G(u,t)}\right\}$$

where $G(u,t) := \gamma(\Theta_{\theta^*}(u)) + tu$.

Proof

- **Lemma 1.** Notice that the function

$$w \rightarrow \sup_{\Delta \in \Theta_{\theta*}(u)} |w^\top \Delta|$$

is u- Lipschitz.

Proof

- **Lemma 1.** Notice that the function

$$w \to \sup_{\Delta \in \Theta_{\theta*}(u)} |w^\top \Delta|$$

is u- Lipschitz. Hence, by the Gaussian concentration inequality for Lipschitz functions, since $w = v/\sigma$,

$$\sup_{\Delta \in \Theta_{\theta*}(u)} |(v/\sigma)^\top \Delta| \leq \mathbb{E}\left(\sup_{\Delta \in \Theta_{\theta*}(u)} |w^\top \Delta| \right) + ut$$

with probability at least $1 - e^{-t^2/2}$.

Proof

- **Lemma 1.** Notice that the function

$$w \to \sup_{\Delta \in \Theta_{\theta*}(u)} |w^\top \Delta|$$

is u- Lipschitz. Hence, by the Gaussian concentration inequality for Lipschitz functions, since $w = v/\sigma$,

$$\begin{aligned} \sup_{\Delta \in \Theta_{\theta*}(u)} |(v/\sigma)^\top \Delta| &\leq \mathbb{E}\left(\sup_{\Delta \in \Theta_{\theta*}(u)} |w^\top \Delta| \right) + ut \\ &= \gamma(\Theta_{\theta*}(u)) + ut, \end{aligned}$$

with probability at least $1 - e^{-t^2/2}$.

Proof

- **Lemma 1.** Notice that the function

$$w \to \sup_{\Delta \in \Theta_{\theta*}(u)} |w^\top \Delta|$$

is u- Lipschitz. Hence, by the Gaussian concentration inequality for Lipschitz functions, since $w = v/\sigma$,

$$\begin{aligned} \sup_{\Delta \in \Theta_{\theta*}(u)} |(v/\sigma)^\top \Delta| &\leq \mathbb{E}\left(\sup_{\Delta \in \Theta_{\theta*}(u)} |w^\top \Delta|\right) + ut \\ &= \gamma(\Theta_{\theta*}(u)) + ut, \end{aligned}$$

with probability at least $1 - e^{-t^2/2}$.

- **Lemma 2.** Let

$$f(u) = \sup_{\theta \in \Theta_{\theta*}(u)} |\langle w, \theta \rangle|.$$

Then

$$|\langle w, \theta \rangle| \leq \max\{\frac{\|\theta\|}{u}, 1\} \cdot f(u), \quad \forall \theta \in \Theta_{\theta*}.$$

- **Proof of Lemma 2.** If $\theta \in \Theta_{\theta^*}$ and $\|\theta\| \leq u$, then $\theta \in \Theta_{\theta^*}(u)$, hence by definition $|\langle w, \theta \rangle| \leq f(u)$.

- **Proof of Lemma 2.** If $\theta \in \Theta_{\theta^*}$ and $\|\theta\| \le u$, then $\theta \in \Theta_{\theta^*}(u)$, hence by definition $|\langle w, \theta \rangle| \le f(u)$.
- If $\theta \in \Theta_{\theta^*}$ and $\|\theta\| > u$, then $\frac{\theta}{\|\theta\|} \cdot u \in \Theta_{\theta^*}$ because Θ_{θ^*} is star shaped. Also, $\|\frac{\theta}{\|\theta\|} \cdot u\| = u$. Hence,

$$|\langle w, \frac{\theta}{\|\theta\|} \cdot u \rangle| \le f(u).$$

- **Proof of Lemma 2.** If $\theta \in \Theta_{\theta^*}$ and $\|\theta\| \leq u$, then $\theta \in \Theta_{\theta^*}(u)$, hence by definition $|\langle w, \theta \rangle| \leq f(u)$.
- If $\theta \in \Theta_{\theta^*}$ and $\|\theta\| > u$, then $\frac{\theta}{\|\theta\|} \cdot u \in \Theta_{\theta^*}$ because Θ_{θ^*} is star shaped. Also, $\|\frac{\theta}{\|\theta\|} \cdot u\| = u$. Hence,

$$|\langle w, \frac{\theta}{\|\theta\|} \cdot u \rangle| \leq f(u).$$

which implies

$$|\langle w, \theta \rangle| \leq \frac{\|\theta\|}{u} \cdot f(u).$$

- **Proof of Lemma 2.** If $\theta \in \Theta_{\theta^*}$ and $\|\theta\| \leq u$, then $\theta \in \Theta_{\theta^*}(u)$, hence by definition $|\langle w, \theta \rangle| \leq f(u)$.
- If $\theta \in \Theta_{\theta^*}$ and $\|\theta\| > u$, then $\frac{\theta}{\|\theta\|} \cdot u \in \Theta_{\theta^*}$ because Θ_{θ^*} is star shaped. Also, $\|\frac{\theta}{\|\theta\|} \cdot u\| = u$. Hence,

$$|\langle w, \frac{\theta}{\|\theta\|} \cdot u \rangle| \leq f(u).$$

  which implies

$$|\langle w, \theta \rangle| \leq \frac{\|\theta\|}{u} \cdot f(u).$$

- Lemma 2 follows combining the two cases:

$$|\langle w, \theta \rangle| \leq \max\left\{ \frac{\|\theta\|}{u}, 1 \right\} f(u),$$

  for $\theta \in \Theta_{\theta^*}$.

- **Back to the proof of the theorem.** By the basic inequality,
$$\|y - \hat{\theta}\|^2 \le \|y - \theta^*\|^2.$$

- **Back to the proof of the theorem.** By the basic inequality,

$$\|y - \hat{\theta}\|^2 \leq \|y - \theta^*\|^2.$$

- This is equivalent to

$$\|(\theta^* + \sigma w) - \hat{\theta}\|^2 \leq \|(\theta^* + \sigma w) - \theta^*\|^2,$$

- **Back to the proof of the theorem.** By the basic inequality,

$$\|y - \hat{\theta}\|^2 \le \|y - \theta^*\|^2.$$

- This is equivalent to

$$\|(\theta^* + \sigma w) - \hat{\theta}\|^2 \le \|(\theta^* + \sigma w) - \theta^*\|^2,$$

which gives

$$\frac{1}{2}\|\hat{\theta} - \theta^*\|^2 \le \sigma\langle w, \hat{\theta} - \theta^*\rangle.$$

- **Back to the proof of the theorem.** By the basic inequality,

$$\|y - \hat{\theta}\|^2 \;\leq\; \|y - \theta^*\|^2.$$

- This is equivalent to

$$\|(\theta^* + \sigma w) - \hat{\theta}\|^2 \;\leq\; \|(\theta^* + \sigma w) - \theta^*\|^2,$$

  which gives

$$\frac{1}{2}\|\hat{\theta} - \theta^*\|^2 \;\leq\; \sigma\langle w, \hat{\theta} - \theta^*\rangle.$$

- Notice that $\hat{\Delta} := \hat{\theta} - \theta^* \in \Theta_{\theta^*}$.

- **Back to the proof of the theorem.** By the basic inequality,

$$\|y - \hat{\theta}\|^2 \leq \|y - \theta^*\|^2.$$

- This is equivalent to

$$\|(\theta^* + \sigma w) - \hat{\theta}\|^2 \leq \|(\theta^* + \sigma w) - \theta^*\|^2,$$

which gives

$$\frac{1}{2}\|\hat{\theta} - \theta^*\|^2 \leq \sigma\langle w, \hat{\theta} - \theta^*\rangle.$$

- Notice that $\hat{\Delta} := \hat{\theta} - \theta^* \in \Theta_{\theta^*}$.
- By Lemma 1, with probability at least $1 - e^{-t^2/2}$, we have that

$$\sup_{\Delta \in \Theta_{\theta^*}(u)} |\langle w, \Delta\rangle| \leq G(u, t) := \gamma(\Theta_{\theta^*}(u)) + tu.$$

- **Back to the proof of the theorem.** By the basic inequality,

$$\|y - \hat{\theta}\|^2 \le \|y - \theta^*\|^2.$$

- This is equivalent to

$$\|(\theta^* + \sigma w) - \hat{\theta}\|^2 \le \|(\theta^* + \sigma w) - \theta^*\|^2,$$

which gives

$$\frac{1}{2}\|\hat{\theta} - \theta^*\|^2 \le \sigma\langle w, \hat{\theta} - \theta^*\rangle.$$

- Notice that $\hat{\Delta} := \hat{\theta} - \theta^* \in \Theta_{\theta^*}$.
- By Lemma 1, with probability at least $1 - e^{-t^2/2}$, we have that

$$\sup_{\Delta \in \Theta_{\theta^*}(u)} |\langle w, \Delta\rangle| \le G(u,t) := \gamma(\Theta_{\theta^*}(u)) + tu.$$

- Hence, by Lemma 2, with probability at least $1 - e^{-t^2/2}$,

$$|\langle w, \Delta\rangle| \le \max\left\{\frac{\|\Delta\|}{u}, 1\right\} G(u,t), \ \ \forall \Delta \in \Theta_{\theta^*}.$$

- Therefore,

$$\frac{1}{2}\|\hat{\theta} - \theta^*\|^2 \quad \leq \quad \sigma\langle w, \hat{\Delta}\rangle$$

- Therefore,

$$\begin{aligned} \frac{1}{2}\|\hat{\theta} - \theta^*\|^2 &\leq \sigma\langle w, \hat{\Delta}\rangle \\ &\leq \sigma|\langle w, \hat{\Delta}\rangle| \end{aligned}$$

- Therefore,

$$\begin{aligned}\frac{1}{2}\|\hat{\theta} - \theta^*\|^2 &\leq \sigma\langle w, \hat{\Delta}\rangle \\ &\leq \sigma|\langle w, \hat{\Delta}\rangle| \\ &\leq \sigma \max\left\{\frac{\|\hat{\Delta}\|}{u}, 1\right\} \cdot G(u, t).\end{aligned}$$

with probability at least $1 - e^{-t^2/2}$.

- Therefore,

$$\begin{aligned}\frac{1}{2}\|\hat{\theta} - \theta^*\|^2 &\le \sigma\langle w, \hat{\Delta}\rangle \\ &\le \sigma|\langle w, \hat{\Delta}\rangle| \\ &\le \sigma \max\left\{\frac{\|\hat{\Delta}\|}{u}, 1\right\} \cdot G(u,t).\end{aligned}$$

  with probability at least $1 - e^{-t^2/2}$.
- Now, if $\|\hat{\Delta}\|/u \le 1$, then

$$\frac{1}{2}\|\hat{\theta} - \theta^*\|^2 \le \sigma G(u,t).$$

- Therefore,

$$\begin{aligned}\tfrac{1}{2}\|\hat{\theta}-\theta^*\|^2 &\leq \sigma\langle w, \hat{\Delta}\rangle \\ &\leq \sigma|\langle w, \hat{\Delta}\rangle| \\ &\leq \sigma \max\left\{\frac{\|\hat{\Delta}\|}{u}, 1\right\} \cdot G(u,t).\end{aligned}$$

  with probability at least $1 - e^{-t^2/2}$.
- Now, if $\|\hat{\Delta}\|/u \leq 1$, then

$$\frac{1}{2}\|\hat{\theta}-\theta^*\|^2 \leq \sigma G(u,t).$$

  or

$$\|\hat{\theta}-\theta^*\| \leq \sqrt{2\sigma G(u,t)}.$$

- Therefore,
$$\begin{aligned}\tfrac{1}{2}\|\hat{\theta}-\theta^*\|^2 &\le \sigma\langle w, \hat{\Delta}\rangle \\ &\le \sigma|\langle w, \hat{\Delta}\rangle| \\ &\le \sigma \max\left\{\tfrac{\|\hat{\Delta}\|}{u}, 1\right\} \cdot G(u,t).\end{aligned}$$
with probability at least $1 - e^{-t^2/2}$.
- Now, if $\|\hat{\Delta}\|/u \le 1$, then
$$\frac{1}{2}\|\hat{\theta}-\theta^*\|^2 \le \sigma G(u,t).$$
or
$$\|\hat{\theta}-\theta^*\| \le \sqrt{2\sigma G(u,t)}.$$
- If $\|\hat{\Delta}\|/u > 1$, then
$$\frac{1}{2}\|\hat{\theta}-\theta^*\|^2 \le \left(\frac{\|\hat{\theta}-\theta^*\|}{u}\right)\sigma G(u,t).$$

- Therefore,

$$\begin{aligned} \tfrac{1}{2}\|\hat{\theta} - \theta^*\|^2 &\le \sigma\langle w, \hat{\Delta}\rangle \\ &\le \sigma|\langle w, \hat{\Delta}\rangle| \\ &\le \sigma \max\left\{\tfrac{\|\hat{\Delta}\|}{u}, 1\right\} \cdot G(u,t). \end{aligned}$$

  with probability at least $1 - e^{-t^2/2}$.
- Now, if $\|\hat{\Delta}\|/u \le 1$, then

$$\frac{1}{2}\|\hat{\theta} - \theta^*\|^2 \le \sigma G(u,t).$$

  or

$$\|\hat{\theta} - \theta^*\| \le \sqrt{2\sigma G(u,t)}.$$

- If $\|\hat{\Delta}\|/u > 1$, then

$$\frac{1}{2}\|\hat{\theta} - \theta^*\|^2 \le \left(\frac{\|\hat{\theta} - \theta^*\|}{u}\right)\sigma G(u,t).$$

  or

$$\|\hat{\theta} - \theta^*\| \le \frac{2\sigma G(u,t)}{u}.$$

- Combining the two cases we obtain:

$$\|\theta^* - \hat{\theta}\| \;\leq\; \max\left\{\frac{2\sigma G(u,t)}{u}, \sqrt{2\sigma G(u,t)}\right\}$$

with probability at least $1 - e^{-t^2/2}$. The claim then follows.