Some theoretical tools

-Statistical Machine Learning-

Lecturer: Darren Homrighausen, PhD

Normal means

A SIMPLER MODEL

Suppose that $Y \sim (\mu, 1)$ and let

$$L_q(\mu) = 2^{q-2}(Y - \mu)^2 + \lambda|\mu|^q$$

and

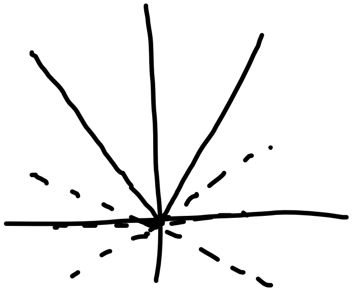
$$\hat{\mu}_q = \underset{\mu}{\operatorname{argmin}}\, L_q(\mu)$$

Then,

- $q = 0 \Rightarrow \hat{\mu}_0 = Y$
- $q = 2 \Rightarrow \hat{\mu}_2 = Y/(\lambda + 1)$
- $q = 1$?

Subdifferential

To theoretically solve this optimization problem, we use the notion of a subderivative.



We call c a subderivative of f at X_0 provided

$$f(X) - f(X_0) \geq c(X - X_0)$$

A convex function can be optimized by setting the subderivative $= 0$

The subdifferential $\partial f|_{X_0}$ is the set of subderivatives.

X_0 minimizes f if and only if $0 \in \partial f|_{X_0}$.

Subdifferential in action

For $\rho(\mu) = |\mu|$,

$$\partial\rho|_\mu = \begin{cases} \{-1\} & \text{if } \mu < 0 \\ [-1, 1] & \text{if } \mu = 0 \\ \{1\} & \text{if } \mu > 0 \end{cases}$$

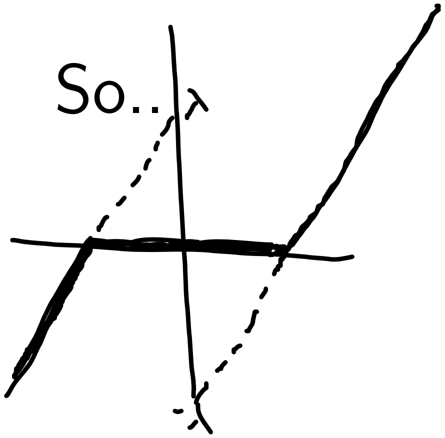
Therefore $\mu = Y - \lambda$

$$\partial L_1|_\mu = \begin{cases} \{\mu - Y - \lambda\} & \text{if } \mu < 0 \\ \{\mu - Y + \lambda z : -1 \le z \le 1\} & \text{if } \mu = 0 \\ \{\mu - Y + \lambda\} & \text{if } \mu > 0 \end{cases}$$

ℓ_1 AND SOFT-THRESHOLDING

$\hat{\mu}_1$ minimizes L_1 if and only if $0 \in \partial L_1|_{\hat{\mu}_1}$

So...



$$\hat{\mu}_1 = \begin{cases} Y + \lambda & \text{if } Y < -\lambda \\ 0 & \text{if } -\lambda \leq Y \leq \lambda \\ Y - \lambda & \text{if } Y > \lambda \end{cases}$$

This can be written

$$\hat{\mu}_1 = \mathrm{sgn}(Y)\overbrace{(|Y| - \lambda)_+}$$

This is known as soft thresholding

Orthogonal design: Example

Suppose now that ($p \leq n$)

$$Y = \mathbb{X}\beta + \epsilon,$$

where $\mathbb{X}^\top \mathbb{X}/n = I$.

Let's solve

$$\hat{\beta}_\lambda = \underset{\beta}{\text{argmin}} \frac{1}{2n}||\mathbb{X}\beta - Y||_2^2 + \lambda||\beta||_1$$

$$\frac{1}{2n}||\mathbb{X}\beta - Y||_2^2 \propto \frac{\beta^\top \mathbb{X}^\top \mathbb{X}\beta}{2n} - \frac{\beta^\top \mathbb{X}^\top Y}{n} = \frac{\beta^\top \beta}{2} - \beta^\top \hat{\beta}_{LS}$$

Now,

$$\frac{1}{2n}||\mathbb{X}\beta - Y||_2^2 + \lambda||\beta||_1 = \sum_{j=1}^{p} \left(\beta_j^2/2 - \beta_j \hat{\beta}_{LS,j} + \lambda|\beta_j|\right)$$

Orthogonal design

We can minimize this component wise:

$$L(\beta) = \beta^2/2 - \beta\hat{\beta}_{LS} + \lambda|\beta| \qquad \text{(dropping the } j\text{)}$$

This can be optimized using subdifferentials in exactly the same way

(I'll leave this as an Exercise)

This results in soft-thresholding the least squares solution.

This rationale can be extended to make the lasso gradient (coordinate) descent explicit

(And is the backbone of glmnet)

Normal means

Note that the orthogonal design linear model is an example of a normal means problem:

Let $\epsilon \sim N(0, I)$, then

$$Y = \mathbb{X}\beta + \epsilon \Leftrightarrow W \overset{D}{=} \beta + \frac{1}{\sqrt{n}}\epsilon$$

This turns out to be an even more powerful idea..

$$\mathbb{X} = UDV^T \Rightarrow U^TY = DV^T\beta + U^T\epsilon$$

$$\Leftrightarrow \quad D^{-1}U^TY = \theta + D^{-1}U^T\epsilon$$

$$\Leftrightarrow \quad W = \theta + D^{-1}\epsilon$$

Normal means

Let

- $\mathcal{H}$ be a real, separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$
- (ϕ_i) be an orthonormal basis for $\mathcal{H}$

Then we can imagine a signal h being observed with a white noise Gaussian process

$$\boxed{Y(t) = h(t) + \epsilon(t)}$$

(Technically, this doesn't exist. Rather we can observe functionals $Y(t)dt = h(t)dt + d\epsilon(t)$)

We make observations of this signal via inner products:

$Y = h + \epsilon$

$$y_i = \langle Y, \phi_i \rangle = \langle h + \epsilon, \phi_i \rangle = h_i + \epsilon_i$$

As linear operations of Gaussians are Gaussians, $\epsilon_i \sim N(0, 1)$

Concentration inequalities and empirical processes

High level overview

The core of modern machine learning theory rests with the following structure:

1. Concentration inequalities: Show that a random quantity is close to its mean with high probability
   1.1 Hoeffding's
   1.2 McDiarmid's
   1.3 Bernstein's
2. Uniform bounds: Guarantee that a set of random quantities are all simultaneously close to their means with high probability
   2.1 VC-dimension
   2.2 Rademacher complexity
   2.3 Covering/bracketing numbers

Motivation

Goal: (concentration inequalities) $+$ (complexity measure) $=$ uniform coverage of a stochastic process (i.e. $\sup_{t \in T} X_t$).

When would this be useful?

Suppose we have data $\mathcal{D}$ and a loss function ℓ_f and we wish to find a function $\hat{f}$ that can predict a new Y from an X

Form the excess risk

$$\mathcal{E}(\hat{f}) = \mathbb{P}\ell_{\hat{f}} - \inf_{f \in \mathcal{F}} \mathbb{P}\ell_f$$

and $\hat{f} = \text{argmin}_{f \in \mathcal{F}} \hat{\mathbb{P}}\ell_f$

Recall

$\hat{\mathbb{P}} = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$ is the empirical measure.

This can be interpreted in two ways:

- Expectation: Let f be a function, then we write

$$\hat{\mathbb{P}} f = \int f \, d\hat{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

- Measure: Let C be a (measurable) set, then we write

$$\hat{\mathbb{P}} C = \int \mathbf{1}_C d\hat{\mathbb{P}} = \frac{1}{n} |\{i : X_i \in C\}|$$

(These notions are used interchangeably, and motivate using $\mathbb{P}$ for both probability and expectation)

Back to motivation

Apply the $2-\epsilon$ TECHNIQUE:

$$\begin{aligned}\mathcal{E}(\hat{f}) &= \mathbb{P}\ell_{\hat{f}} - \hat{\mathbb{P}}\ell_{\hat{f}} + \hat{\mathbb{P}}\ell_{\hat{f}} - \inf_{f\in\mathcal{F}} \mathbb{P}\ell_f \\ &\leq \mathbb{P}\ell_{\hat{f}} - \hat{\mathbb{P}}\ell_{\hat{f}} + \hat{\mathbb{P}}\ell_{f_*} - \mathbb{P}\ell_{f_*} \\ &\leq 2\sup_{f\in\mathcal{F}} |\mathbb{P}\ell_f - \hat{\mathbb{P}}\ell_f|\end{aligned}$$

Where f_* is such that $\mathbb{P}\ell_{f_*} = \inf_{f\in\mathcal{F}} \mathbb{P}\ell_f$

So, fixing an $\epsilon > 0$

$$\begin{aligned}\mathbb{P}(\mathcal{E}(\hat{f}) > 2\epsilon) &\leq \mathbb{P}(|\sup_{f\in\mathcal{F}} |(\hat{\mathbb{P}} - \mathbb{P})\ell_f| > \epsilon) \\ &\leq \frac{\mathbb{E}\sup_{f\in\mathcal{F}} |(\hat{\mathbb{P}} - \mathbb{P})\ell_f|}{\epsilon}\end{aligned}$$

Motivation

Conclusion:

$$\mathbb{P}(\mathcal{E}(\hat{f}) > 2\epsilon) \leq \epsilon^{-1}\mathbb{E}\sup_{f\in\mathcal{F}}|(\hat{\mathbb{P}} - \mathbb{P})\ell_f| = \epsilon^{-1}\mathbb{E}\left|\left|\hat{\mathbb{P}} - \mathbb{P}\right|\right|_{\mathcal{F}}$$

We can bound the excess risk of an estimator $\hat{f}$ by bounding the supremum of the difference between the empirical measure and true measure

Note that:

- Using the previous notation, $X_t = (\hat{\mathbb{P}} - \mathbb{P})\ell_f$, and $T = \mathcal{F}$

  (Sometimes the index set is considered $\mathcal{L} = \{\ell_f : f \in \mathcal{F}\}$)
- The stochastic process $\mathbb{G} = \sqrt{n}(\hat{\mathbb{P}} - \mathbb{P})$ is the empirical process

Empirical process

The stochastic process $(\hat{\mathbb{P}} - \mathbb{P})\ell_f$ is zero mean and hence we know by the SLLN that for all $f \in \mathcal{F}$

$$(\hat{\mathbb{P}} - \mathbb{P})\ell_f \to 0 \text{ a.s}$$

(Assuming $\mathbb{P}\ell_f$ exists, of course)

However, this doesn't give us uniform control

(i.e: this doesn't imply that the supremum goes to zero)

We call an index set $\mathcal{F}$ a Glivenko-Cantelli class if

$$\sup_{f \in \mathcal{F}} \left|(\hat{\mathbb{P}} - \mathbb{P})\ell_f\right| = \left|\left|\hat{\mathbb{P}} - \mathbb{P}\right|\right|_{\mathcal{F}} \to 0 \text{ a.s}$$

Glivenko-Cantelli: example

A classical example is the empirical CDF

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,t]}(X_i) = \hat{\mathbb{P}} f_t$$

where $f_t(x) = \mathbf{1}_{(-\infty,t]}(x)$

Often, we are attempting to estimate a functional of the true CDF with a plug-in version using the empirical CDF
(True CDF: $F(t) = \mathbb{P}(X \leq t)$)

Sub Example: Let $\theta = \theta(\mathbb{P})$ given by the median: $\theta = \theta(\mathbb{P})$ is argmin of $\mathbb{P}(-\infty, x] = \inf_x F(x)$ subject to $F(x) \geq 1/2$

Then, we might estimate $\theta(\mathbb{P})$ with $\hat{\theta} = \theta(\hat{\mathbb{P}})$ by plugging in F_n

Glivenko-Cantelli: example

The Glivenko-Cantelli theorem says that

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \to 0 \text{ a.s.}$$

If we write $\mathcal{F} = \{f_t : f_t(x) = \mathbf{1}_{(-\infty,t]}(x), t \in \mathbb{R}\}$, then

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| = ||F_n - F||_{\mathbb{R}} = \left|\left|\hat{\mathbb{P}} - \mathbb{P}\right|\right|_{\mathcal{F}}$$

and hence $\mathcal{F}$ is a Glivenko-Cantelli (G.C.) class.

Glivenko-Cantelli: example

(Technical condition: $\mathbb{P}(-\infty, t] > 1/2$ for each $t > \theta(\mathbb{P})$. This forces a continuity property that $|$median($\mathbb{P}$) - median($\mathbb{P}$'$)| < \epsilon$ if $\mathbb{P}$ and $\mathbb{P}$' are uniformly close.)

Sub Example: As $\mathcal{F}$ is G.C., for all $\delta > 0$, for n large enough

$$\sup_t |\hat{\mathbb{P}}(-\infty, t] - \mathbb{P}(-\infty, t]| < \delta$$

Fix $\epsilon > 0$. Choose δ such that

$$\mathbb{P}(-\infty, \theta - \epsilon] < \frac{1}{2} - \delta \quad \text{(This is always possible)}$$

$$\mathbb{P}(-\infty, \theta + \epsilon] > \frac{1}{2} + \delta \quad \text{(This requires condition)}$$

Glivenko-Cantelli: example

Now,

$$\mathbb{P}(-\infty, \hat{\theta}] > \underbrace{\hat{\mathbb{P}}(-\infty, \hat{\theta}] - \delta}_{\text{uniform closeness}} \geq 1/2 - \delta$$

Hence, $\hat{\theta} > \theta - \epsilon$ as BWOC:

$$\hat{\theta} \leq \theta - \epsilon \Rightarrow \mathbb{P}(-\infty, \hat{\theta}] \leq \mathbb{P}(-\infty, \theta - \epsilon] < 1/2 - \delta$$

Also, it can be shown that $\hat{\theta} \leq \theta + \epsilon$

Hence, uniform closeness of F_n to F shows that the sample and populations medians are close

(Note, we have asked for much more than needed, sometimes this can be too much)

Glivenko-Cantelli: example

This gets refined to a rate of convergence by the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality

$$\mathbb{P}(||F_n - F||_\infty > \epsilon) \leq 2e^{-2n\epsilon^2}$$

(Both the constants 2 cannot be improved upon (Massart (1990)))

This result, along with the previous discussion gets us a rate of convergence for the median

Some lasso theory

Low assumption prediction: the lasso

Remember: Prediction risk

$$R(\beta) = \mathbb{E}_Z\left[\left(Y - X^\top\beta\right)^2\right] = \mathbb{E}_Z\left[\left(Y - X^\top\beta\right)^2|\mathcal{D}\right]$$

Define the oracle estimator

$$\beta_t^* = \operatorname*{argmin}_{\{\beta:||\beta||_1\leq t\}} R(\beta)$$

(Important: This does not assume that $\mathbb{E}Y|X$ is linear in X!)

The excess risk is

$$\mathcal{E}(\hat{\beta}_t, \beta_t^*) = R(\hat{\beta}_t) - R(\beta_t^*)$$

Persistence

A procedure is persistent for a set of measures $\mathcal{P}$ if

$$\forall \mathbb{P} \in \mathcal{P}, \qquad \mathcal{E}(\hat{\beta}_t, \beta_t^*) \xrightarrow{p} 0$$

(This is convergence in probability. What is random?)

Define the following set of distributions on $\mathbb{R} \times \mathbb{R}^p$: Let $C_{\mathcal{P}} < \infty$ and

$$\mathcal{P} = \{\mathbb{P} : \mathbb{P}Y^2 < C_{\mathcal{P}}, \text{ and } |x_j| < C_{\mathcal{P}} \text{ almost surely}, j = 1, \ldots, p\}$$

We'd like to know how fast t can grow while still maintaining persistency.

(Note: this set $\mathcal{P}$ is more restrictive than needed)

Useful results and observations

Let

- $Z = (x_0, x_1, \ldots, x_p)^\top$, where $x_0 = Y$
- $\gamma = (-1, \beta_1, \ldots, \beta_p)^\top$

Then, for $\ell_\beta(Z) = (Y - X^\top \beta)^2$,

$$\mathbb{P}\ell_\beta = \mathbb{P}(Y - X^\top\beta)^2 = \gamma^\top \Sigma \gamma,$$

where $\Sigma_{jk} = \mathbb{P}Z_j Z_k$ for $0 \le j, k \le p$

Likewise,

$$\hat{\mathbb{P}}\ell_\beta = \gamma^\top \hat{\Sigma}\gamma,$$

where $\hat{\Sigma}_{jk} = n^{-1} \sum_{i=1}^n Z_{ij} Z_{ik}$ for $0 \le j, k \le p$

(These can be written: $\Sigma = \mathbb{P}ZZ^\top$ and $\hat{\Sigma} = \hat{\mathbb{P}}ZZ^\top$)

Persistence theorem

Theorem

Over any $\mathbb{P}$ in $\mathcal{P}$, the procedure

$$\underset{\beta \in \{\beta : ||\beta||_1 \leq t\}}{\operatorname{argmin}} \quad \hat{\mathbb{P}}\ell_\beta$$

is persistent provided $\log p = o(n)$ *and*

$$t = t_n = o\left(\left(\frac{n}{\log p}\right)^{1/4}\right)$$

(This theorem appears in Greenshtein, Ritov (2004))

(It's worth noting that this rate is improved to a square root in Bartlett, et al. (2012). This is at the expense of higher order powers of the log terms. These logarithmic powers could be removed by bounding Talagrand's γ_2 functional directly, instead of an entropy integral)

Deterministic asymptotic notation

We write $a_n = O(b_n)$ (and say big ohh) provided

$$\frac{a_n}{b_n} = O(1),$$

where

$$c_n = O(1)$$

means

- There exists a C
- Such that for sufficiently large N
- For all $n \geq N$
- $c_n \leq C$

Deterministic asymptotic notation

We write $a_n = o(b_n)$ (and say little ohh) provided

$$\frac{a_n}{b_n} = o(1),$$

where

$$c_n = o(1)$$

means

- For all $\epsilon > 0$
- There exists an N
- Such that for all $n \geq N$
- $c_n \leq \epsilon$

Stochastic asymptotic notation

We write $a_n = O_p(b_n)$ (and say big ohh p) provided

$$\frac{a_n}{b_n} = O_p(1),$$

where

$$c_n = O_p(1)$$

means

- For all δ
- There exists a C
- Such that for sufficiently large N
- For all $n \geq N$
- $\mathbb{P}(|c_n| \geq C) \leq \delta$

(This is also called bounded in probability, and is related to convergence in distribution)

Stochastic asymptotic notation

We write $a_n = o_p(b_n)$ (and say little ohh p) provided

$$\frac{a_n}{b_n} = o_p(1),$$

where

$$c_n = o_p(1)$$

means

- For all $\epsilon > 0, \delta > 0$
- There exists an N
- Such that for all $n \geq N$
- $\mathbb{P}(|c_n| \geq \epsilon) \leq \delta$

Stochastic asymptotic notation

Note that if we have random variables (X_n) and X, then

$$X_n \to X \text{ in probability} \Leftrightarrow X_n - X = o_p(1)$$

We can also express Slutsky's theorem(s)

- $o_p(1) + O_p(1) = ?$
- $o_p(1) O_p(1) = ?$
- $o_p(1) + o_p(1) O_p(1) = ?$

Persistence proof

Note that, $\sup_{\beta \in \{b : ||b||_1 \leq t\}} ||\beta||_1 \leq t$. Also,

Lemma

Suppose $a \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p \times p}$. Then

$$a^\top A a \leq ||a||_1^2 \, ||A||_\infty \,,$$

where $||A||_\infty := \max_{i,j} |A_{ij}|$ is the entry-wise max norm.

Proof.

$$a^\top A a \underbrace{\leq}_{\text{Hölder's}} ||a||_1 \, ||Aa||_\infty \leq ||a||_1 \max_{ij} |A_{ij}| \, ||a||_1 = ||a||_1^2 \, ||A||_\infty \,,$$

□

These facts imply..

Persistence proof

$$
\begin{align}
\mathcal{E}(\hat{\beta}_t, \beta_t^*) &= \underbrace{R(\hat{\beta}_t)}_{\hat{\gamma}_t^\top \Sigma \hat{\gamma}_t} - \underbrace{R(\beta_t^*)}_{(\gamma_t^*)^\top \Sigma (\gamma_t^*)} \tag{1}\\
&= \hat{\gamma}_t^\top \Sigma \hat{\gamma}_t - \hat{\gamma}_t^\top \hat{\Sigma} \hat{\gamma}_t + \hat{\gamma}_t^\top \hat{\Sigma} \hat{\gamma}_t - (\gamma_t^*)^\top \Sigma (\gamma_t^*) \tag{2}\\
&\le \hat{\gamma}_t^\top \Sigma \hat{\gamma}_t - \hat{\gamma}_t^\top \hat{\Sigma} \hat{\gamma}_t + (\gamma_t^*)^\top \hat{\Sigma} \gamma_t^* - (\gamma_t^*)^\top \Sigma \gamma_t^* \tag{3}\\
&= \hat{\gamma}_t^\top (\Sigma - \hat{\Sigma}) \hat{\gamma}_t + (\gamma_t^*)^\top (\hat{\Sigma} - \Sigma)(\gamma_t^*) \tag{4}\\
&\le 2 \sup_{\beta \in \{b : ||b||_1 \le t\}} \gamma_t^\top (\Sigma - \hat{\Sigma}) \gamma_t \qquad \text{(}2\epsilon\text{ trick)} \tag{5}\\
&\le 2 \sup_{\beta \in \{b : ||b||_1 \le t\}} ||\gamma_t||_1^2 \left|\left| \Sigma - \hat{\Sigma} \right|\right|_\infty \qquad (\textit{Lemma}) \tag{6}\\
&\le 2(t+1)^2 \left|\left| \Sigma - \hat{\Sigma} \right|\right|_\infty \tag{7}
\end{align}
$$

Can we control the sup-norm part?

Persistence proof

Nemirovski's inequality: Let $\xi_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ be independent, zero mean, finite variance random variables with $p \geq 3$. Define $S_n = \sum_{i=1}^n \xi_i$. Then for every $q \in [2, \infty]$

$$\mathbb{E}||S_n||_q^2 \leq e(2\log(p) - 1)\min\{q, \log(p)\} \sum_{i=1}^n \mathbb{E}||\xi_i||_q^2$$

(Juditsky, Nemirovski (2000), Dümbgen, et al. (2010))

This should be compared with the naïve bound:

$$\mathbb{E}||S_n||_q^2 \leq \sum_{i=1}^n \sum_{i'=1}^n \mathbb{E}||\xi_i||_q||\xi_{i'}||_q$$

Persistence proof: Nemirovski's inequality

Motivation: Under Nemirovski's assumptions,

- $\mathbb{E}S_n^2 = \sum_{i=1}^n \mathbb{E}\xi_i^2$ $(p=1)$
- In a Hilbert space with inner product $\langle\cdot,\cdot\rangle$

$$\mathbb{E}||S_n||^2 = \sum_{i,i'}^{n} \mathbb{E}\langle\xi_i,\xi_{i'}\rangle = \sum_{i=1}^{n} \mathbb{E}||\xi_i||^2$$

- What about a Banach space (e.g. $||\cdot||_q, q \neq 2$)?

Persistence proof

Nemirovski's inequality: Let $\xi_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ be independent, zero mean, finite variance random variables with $p \geq 3$. Define $S_n = \sum_{i=1}^n \xi_i$. Then for every $q \in [2, \infty]$

$$\mathbb{E}||S_n||_q^2 \leq e(2\log(p) - 1)\min\{q, \log(p)\}\sum_{i=1}^n \mathbb{E}||\xi_i||_q^2$$

Let $\xi_i = \mathrm{vec}\left(\frac{1}{n}\left(Z_{ij}Z_{ik} - \mathbb{E}Z_jZ_k\right)\right) \in \mathbb{R}^{(p+1)^2}$ be the vectorized difference of empirical covariance and the true covariance

Then

$$\left|\left|\Sigma - \hat{\Sigma}\right|\right|_\infty = \left|\left|\sum_{i=1}^n \xi_i\right|\right|_\infty$$

Persistence proof

$$\mathbb{E}||S_n||_q^2 \leq e(2\log(p)-1)\min\{q,\log(p)\}\sum_{i=1}^n \mathbb{E}||\xi_i||_q^2$$

(Nemirovski's inequality)

$$\left(\mathbb{E}\left\|\Sigma-\hat{\Sigma}\right\|_\infty\right)^2 \leq \mathbb{E}\left\|\Sigma-\hat{\Sigma}\right\|_\infty^2 \qquad \text{(Jensen's inequality)}$$

$$= \mathbb{E}\left\|\sum_{i=1}^n \xi_i\right\|_\infty^2$$

$$\leq C\log((p+1)^2)\sum_{i=1}^n \mathbb{E}||\xi_i||_\infty^2$$

$$\leq 4CC_{\mathcal{P}}^2\log(p+1)\frac{1}{n} \qquad (\mathbb{P}\in\mathcal{P})$$

$$\lesssim \frac{\log(p)}{n}$$

Persistence proof: Conclusion

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}(\hat{\beta}_t, \beta_t^*) > \delta\right) &\leq \mathbb{E}[\mathcal{E}(\hat{\beta}_t, \beta_t^*)]\delta^{-1} \quad \text{(Markov's inequality)} \quad (8)\\ &\leq 2\delta^{-1}(t+1)^2 \mathbb{E}\left\|\Sigma - \hat{\Sigma}\right\|_\infty \quad (9)\\ &\lesssim 2\delta^{-1}(t+1)^2 \sqrt{\frac{\log p}{n}} \quad (10)\end{aligned}$$

Therefore, we have persistence provided $\log p = o(n)$ and

$$t_n = o\left(\left(\frac{n}{\log p}\right)^{1/4}\right)$$

Alternatively

$$\mathcal{E}(\hat{\beta}_t, \beta_t^*) = O_p\left(t^2 \sqrt{\frac{\log(p)}{n}}\right)$$

Probably approximately correct (PAC)

Probability bound ⇔ high probability upper bound:

$$\mathbb{P}(\text{error} > \delta) \leq \epsilon$$

gets converted to: with probability $1 - \epsilon$

$$\text{error} \leq \delta$$

(This is known as a PAC bound)

Example:

$$\mathbb{P}(|\overline{X} - \mu| > \delta) \leq \frac{\mathbb{E}(\overline{X} - \mu)^2}{\delta^2}$$

Hence, with probability at least $1 - \frac{\sigma^2}{n\delta^2}$

$$|\overline{X} - \mu| \leq \delta$$

Lasso theory: Summary

It is important to note that we do not assume...

- a linear model
- an additive stochastic component (let alone, Gaussian errors)
- that the design is ‘almost’ uncorrelated

and we get that, with probability at least $1 - C\delta^{-1}t^2\sqrt{\frac{\log(p)}{n}}$,

$$R(\hat{\beta}_t) \leq R(\beta_t^*) + \delta$$

Orlicz norms

Orlicz norms

As alluded to previously, this type of theory is driven by the tails of the distribution This is most commonly phrased in terms of Orlicz norms

Let ψ be a non-decreasing, convex function such that $\psi(0) = 0$. Then:

$$||X||_\psi = \inf\{C > 0 : \mathbb{E}\psi\left(\frac{|X|}{C}\right) \leq 1\}$$

(Jensen's inequality shows this is a norm)

There are two main cases

- L_p norm: $\psi(x) = x^p \Rightarrow ||X||_\psi = ||X||_p = (\mathbb{E}|X|^p)^{1/p}$
- p-Orlicz: $\psi_p(x) = e^{x^p} - 1$

Orlicz norms

Two important facts:

- $||X||_{\psi_p} \leq ||X||_{\psi_q} (\log 2)^{1/q-1/p}$, for $p \leq 2$
- $||X||_p \leq p! \, ||X||_{\psi_1}$

(This allows us to interchange results about various norms, as long as we don't care about constants)

Orlicz norms

By Markov's inequality

$$
\begin{aligned}
\mathbb{P}(|X| > x) &\leq \frac{\mathbb{E}\psi(|X|)/\left|\left|X\right|\right|_\psi}{\psi(x)/\left|\left|X\right|\right|_\psi} \\
&\leq \frac{1}{\psi(x)/\left|\left|X\right|\right|_\psi} \\
&= \begin{cases} \left|\left|X\right|\right|_p x^{-p} & \text{if } \psi(x) = x^p \\ \frac{1}{e^{(x/\left|\left|X\right|\right|_\psi)^p}-1} \asymp e^{-(x/\left|\left|X\right|\right|_\psi)^p} & \text{if } \psi(x) = \psi_p(x) \end{cases}
\end{aligned}
$$

Hence, Orlicz norms allow us to encode the tail behavior of a random variable

In fact, it works as an if and only if:

If $\mathbb{P}(|X| > x) \leq Ce^{-cx^p}$ then $\left|\left|X\right|\right|_{\psi_p} \leq ((1 + C)c^{-1})^{1/p} < \infty$

Concentration inequalities

General form

For showing results about empirical processes or performance guarantees for algorithms, we want results of the form

$$\mathbb{P}\left(|f(Z_1, \ldots, Z_n) - \mu_n(f)| > \epsilon\right) < \delta_n$$

where $\delta_n \to 0$ and $\mu_n(f) = \mathbb{E} f(Z_1, \ldots, Z_n)$.

For statistical learning theory, we need uniform bounds

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |f(Z_1, \ldots, Z_n) - \mu_n(f)| > \epsilon\right) < \delta_n$$

Hoeffding's inequality

Suppose $\mu = \mathbb{E}Z < \infty$ and $\mathbb{P}(Z \geq 0) = 1$. Then for any $\epsilon > 0$

$$\mathbb{E}Z = \int_0^\infty Z d\mathbb{P} \geq \int_\epsilon^\infty Z d\mathbb{P} \geq \epsilon \int_\epsilon^\infty d\mathbb{P} = \epsilon \mathbb{P}(Z > \epsilon)$$

Yielding Markov's inequality

This can be transformed to Chebyshev's inequality by using the variance

$$\mathbb{P}(|Z - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2} \Rightarrow \mathbb{P}(|\overline{Z} - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

Observation: This is nice, but does not decay exponentially fast. However, it only makes a second moment assumption

Hoeffding's inequality

A different transformation occurs via a Chernoff bound. For any $t > 0$

$$\mathbb{P}(Z > \epsilon) = \mathbb{P}\left(e^{tZ} > e^{t\epsilon}\right) \leq e^{-t\epsilon}\mathbb{E}[e^{tZ}]$$

(This is the moment generating function. Here we see increasing moment conditions giving tighter bounds)

This can be minimized over t as it is arbitrary

$$\mathbb{P}(Z > \epsilon) \leq \inf_{t>0} e^{-t\epsilon}\mathbb{E}[e^{tZ}]$$

Hoeffding's inequality

This is the main content of the paper Hoeffding (1963)

Hoeffding's lemma: Suppose $Z \in [a, b]$, then for any t

$$\mathbb{E}[e^{tZ}] \leq e^{t\mu + t^2(b-a)^2/8}$$

Proof idea:

$$\mathbb{E}e^{tZ} \leq -\frac{a}{b-a}e^{tb} + \frac{b}{b-a}e^{ta} = E^{g(u)}$$

where $u = t(b-a)$. Write down Tayloy's theorem for g up to order 2 to get bound.

(The punchline: the bound is driven by a worst case)

Hoeffding's inequality

Hoeffding's inequality:

$$\mathbb{P}\left(|\overline{Z} - \mu| > \epsilon\right) \leq 2e^{-2n\epsilon^2/(b-a)^2}$$

Proof sketch: Let Z be zero mean

$$\begin{aligned}
\mathbb{P}\left(\overline{Z} > \epsilon\right) &= \mathbb{P}\left(e^{t\overline{Z}} > e^{t\epsilon}\right) \\
&\leq e^{-t\epsilon}\mathbb{E}e^{t\overline{Z}} \\
&= e^{-t\epsilon}\prod_{i=1}^{n}\mathbb{E}e^{tn^{-1}Z_i} \\
&\leq e^{-t\epsilon}e^{(t/n)^2(b-a)^2/8}
\end{aligned}$$

(Now, minimize over t and symmetrize)

Hoeffding's inequality: Generalizations

We can let the upper and lower limits change with i:
$Z_i \in [a_i, b_i]$

Also, we can invert this probability statement into a PAC bound: with probability at least $1 - \delta$

$$|\overline{Z} - \mu| \leq \sqrt{\frac{c}{2n} \log\left(\frac{2}{\delta}\right)}$$

where $c = n^{-1} \sum_i (b_i - a_i)^2$

Compare to Chebyshev, which has growth

$$|\overline{Z} - \mu| \leq \sqrt{\frac{\sigma^2}{n\delta}}$$

Hoeffding's inequality: Example

Let $Y_i \in \{0, 1\}$, $X_i \in \mathbb{R}^p$ and $g : \mathbb{R}^p \to \{0, 1\}$ be a classifier.

Define the training error to be

$$\hat{R}(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(Y_i \neq g(X_i))$$

and

$$R(g) = \mathbb{P}(Y_i \neq g(X_i))$$

Then, $\hat{R}(g) - R(g)$ is zero mean and in the interval $[-1, 1]$:

$$\hat{R}(g) - R(g) \leq \sqrt{\frac{1}{n} \log\left(\frac{2}{\delta}\right)}$$

with high probability (w.h.p)

Refined Hoeffding's inequality

The previous result was restricted to sample means.

This isn't an essential part of the result, however, and is removed in a generalization known as McDiarmid's inequality

Suppose that

$$\sup_{Z_1,\ldots,Z_n,Z_i'} |f(Z_1,\ldots,Z_n) - f(Z_1,\ldots,Z_{i-1},Z_i',Z_{i+1},\ldots,Z_n)| \leq c_i$$

Then

$$\mathbb{P}\left(|f(Z) - \mathbb{E}f(Z)| > \epsilon\right) \leq 2e^{-2\epsilon^2/\sum_{i=1}^n c_i}$$

McDiarmid's inequality: Example

Let $f(Z) = \sup_A |\hat{\mathbb{P}}(A) - \mathbb{P}(A)|$.

$$\begin{aligned} |f(Z) - f(Z')| &= \left| \sup_A |\hat{\mathbb{P}}(A) - \mathbb{P}(A)| - \sup_A |\hat{\mathbb{P}}'(A) - \mathbb{P}(A)| \right| \\ &\leq \sup_A \left| |\hat{\mathbb{P}}(A) - \mathbb{P}(A)| - |\hat{\mathbb{P}}'(A) - \mathbb{P}(A)| \right| \\ &\leq \sup_A \left| \hat{\mathbb{P}}(A) - \mathbb{P}(A) - (\hat{\mathbb{P}}'(A) - \mathbb{P}(A)) \right| \\ &= \sup_A \left| \hat{\mathbb{P}}(A) - \hat{\mathbb{P}}'(A) \right| \qquad (|\,|a|-|b|\,| \leq |a-b|) \end{aligned}$$

(Either Z_i is still in A and nothing changes, or Z_i is no longer in A and hence the difference is $1/n$)

Therefore

$$\mathbb{P}\left(|f(Z) - \mathbb{E}f(Z)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

Sharper inequalities

These previous results don't used any information about where the probabilities mass lies

Hoeffding's inequality is driven by the worst case: a R.V. that puts all of its mass at the boundaries

If the variance of Z_i is small, we can get sharper inequalities

Sharper inequalities

This idea is that $\sum_{i=1}^n Z_i$ it approximately normally distributed with variance $v = \sum_{i=1}^n \mathbb{V} Z_i$

The tails of a $N(0, v)$ are of order $e^{-x^2/(2v)}$

Bernstein's inequality gives a tail bound that is a combination of a normal and a penalty for non-normality

Sharper inequalities

Lemma: Suppose that $|X| < c$ and $\mathbb{E}X = 0$. Then for any $t > 0$

$$\mathbb{E}[e^{tX}] \leq \exp\left\{t^2\sigma^2\left(\frac{e^{tc} - 1 - tc}{(tc)^2}\right)\right\}$$

where $\sigma^2 = \mathbb{V}X$

Idea: The main part of the proof relies on the inequality: for $r \geq 2$

$$\mathbb{E}X^r = \mathbb{E}X^{r-2}X^2 \leq c^{r-2}\sigma^2$$

(all higher moments than the variance are killed by the a.s. bound, while the first two moments are computed as usual)

Sharper inequalities

Bernstein's inequality: If $|Z_i| \leq c$ a.s. and $\mathbb{E}Z_i = \mu$, then for all $\epsilon > 0$

$$\mathbb{P}\left(|\overline{Z} - \mu| > \epsilon\right) \leq 2e^{-\frac{n\epsilon^2}{2\sigma^2 + 2c\epsilon/3}}$$

where $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{V}Z_i$

Compare to Hoeffding's:

$$\mathbb{P}\left(|\overline{Z} - \mu| > \epsilon\right) \leq 2e^{-\frac{n\epsilon^2}{2c^2}}$$

Sharper inequalities

- If $\sigma^2 >> 2c\epsilon/3$, then

$$\log(\text{Bernstein}) \asymp -\frac{n\epsilon^2}{2\sigma^2} \leq -\frac{n\epsilon^2}{4c^2} \asymp \log(\text{Hoeffding})$$

- If $\sigma^2 << 2c\epsilon/3$, then

$$\log(\text{Bernstein}) \asymp -\frac{3n\epsilon^2}{4c\epsilon} = -\frac{3n\epsilon}{4c} \leq -\frac{n\epsilon^2}{4c^2} \asymp \log(\text{Hoeffding})$$

Note: This implies that the Bernstein bound is like an exponential for large ϵ and normal for small ϵ

Bernstein's inequality

There is a related moment version of Bernstein's inequality:

All that is really required out of the almost sure boundedness of Z_i is bounds for the moments

Suppose that Z_i are such that $\mathbb{E}|Z_i|^m \leq m!M^{m-2}c_i/2$ for all $m \geq 2$ and constants M, c_i. Then

$$\mathbb{P}\left(|\overline{Z} - \mu| > \epsilon\right) \leq 2e^{-\frac{n\epsilon^2}{2\sigma^2 + 2c\epsilon}}$$

for $\sigma^2 \geq \frac{1}{n}\sum_{i=1}^n c_i$

Bernstein's inequality

The most useful part of Bernstein's inequality is the associated PAC bound

With probability at least $1 - \delta$

$$|\overline{Z} - \mu| \leq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}} + \frac{2c \log(1/\delta)}{3n}$$

$(|Z_i| \leq c,\ \sigma^2 = n^{-1} \sum_{i=1}^n \mathbb{V} Z_i)$

In particular, if the variance is small enough:

$$\sigma^2 \leq \frac{2c^2 \log(1/\delta)}{9n} \quad \Rightarrow \quad |\overline{Z} - \mu| \leq \frac{4c \log(1/\delta)}{3n}$$

(That is, we get a n-decay instead of $\sqrt{n}$-decay)

Bounding maximums

Finite maximums

Recall that

$$\mathbb{P}(|X| > x) \leq \begin{cases} ||X||_p \, x^{-p} & \text{if } \psi(x) = x^p \\ \frac{1}{e^{(x/||X||_\psi)^p} - 1} \asymp e^{-(x/||X||_\psi)^p} & \text{if } \psi(x) = \psi_p(x) \end{cases}$$

Hence, bounds on the sup of a norm, provide bounds on the probability

Finite maximums

For the L_p norm, a straight-forward bound exists

Suppose that we have the process X_t on $|T| < \infty$

Using the fact that $\max |X_t|^p \leq \sum |X_t|^p$, it follows that

$$\begin{aligned} \left|\left|\max_{t \in T} X_t\right|\right|_p &= \left(\mathbb{E} \max_{t \in T} |X_t|^p\right)^{1/p} \\ &\leq \left(\sum_{t \in T} \mathbb{E}|X_t|^p\right)^{1/p} \\ &\leq |T|^{1/p} \max_{t \in T} ||X_t||_p \end{aligned}$$

Finite maximums

This continues to hold, but with more complicated proof, for many Orlicz norms

Under some technical conditions on ψ that include ψ_p

$$\left|\left|\max_{t\in T} X_t\right|\right|_\psi \leq K\psi^{-1}(|T|)\max_{t\in T}||X_t||_\psi$$

Here, K only depends on ψ

Finite maximums

Some observations:

- As $\psi_p^{-1}(|T|) = (\log(1 + |T|))^{1/p}$, we get a logarithmic increase in $|T|$ while using the p-Orlicz norm, compared to polynomial for L_p
- This conversion is useless when $|T| = \infty$. This case can be handled via generic chaining whereby each R.V. is written as a sum of parts of the index space where
  - The R.Vs have low correlation between partitions
  - There aren't too many in each partition

For a random stochastic process $(X_t)_{t \in T}$, the number of links in the chain depends on the (metric) entropy of T

This is quantified via covering, packing, and bracketing numbers

Covering and packing numbers

A pseudo-metric space (T, d) is a set with a function $d : T \times T \to [0, \infty)$ such that

1. $d(t, t) = 0$
2. $d(s, t) = d(t, s)$
3. $d(s, u) \leq d(s, t) + d(t, u)$

(The pseudo part comes from not insisting that $d(s, t) = 0 \Rightarrow t = s$)

Example: Define the metric related to the empirical measure

$$d(f, g) = \frac{1}{n} \sum_{i=1}^{n} |f(Z_i) - g(Z_i)|$$

Then d is a pseudo-metric, but not a metric

Covering numbers

An ϵ-cover is a set $\tilde{T}$ comprised of ϵ-balls such that $T \subseteq \tilde{T}$. The covering number of T is:

$$N(\epsilon, T, d) = \min\{|\tilde{T}| : \tilde{T} \text{ is an } \epsilon - \text{cover}\}$$

Note that

- T is totally bounded if $N(\epsilon, T, d) < \infty$ for all $\epsilon > 0$
- The function $\epsilon \mapsto \log N(\epsilon, T, d)$ is the metric entropy of T

Packing numbers

An ϵ-packing of T is a subset $\tilde{T}$ comprised of non-overlapping ϵ-balls. The packing number is

$$M(\epsilon, T, d) = \max\{|\tilde{T}| : \tilde{T} \text{ is an } \epsilon - \text{packing of } T\}$$

Covering and packing numbers

For almost all purposes, the difference between these two concepts are unimportant:

For instance: $\forall \epsilon > 0$

$$M(\epsilon) \leq N(\epsilon)$$

Related lower bounds for M in terms of N are possible with scalar transformations of ϵ

Covering and packing numbers

Computing either covering or packing numbers can be very difficult

I advise looking up known results if you go this route, such as

- Covering numbers of the unit ball of $\mathbb{R}^d$ (such as with ridge or lasso)
- Compact sets of functions

Examples

Example: Uniform risk

Suppose we want to minimize $\hat{\mathbb{P}}\ell_f = n^{-1}\sum_{i=1}^n |f(X_i) - Y_i|^2$ over a set of functions $\mathcal{F}_n$

The main tool in showing consistency/rates of convergence for such estimators is to show that the empirical risk looks like the true risk, uniformly over $\mathcal{F}_n$:

$$\sup_{f\in\mathcal{F}_n}\left|\hat{\mathbb{P}}\ell_f - \mathbb{P}\ell_f\right| \to 0 \text{ a.s.}$$

It's easier to generate the set of functions[1] $\mathcal{L}_n = \{\ell_f : f \in \mathcal{F}_n\}$ and look at

$$\sup_{\ell\in\mathcal{L}_n}\left|\hat{\mathbb{P}}\ell - \mathbb{P}\ell\right|$$

[1] Often, $\mathcal{L}_n$ is closed

Example: Uniform risk

Using Hoeffding's inequality shows that if $\ell(Z) \in [0, b]$ a.s., then

$$\mathbb{P}\left(\left|\hat{\mathbb{P}}\ell - \mathbb{P}\ell\right| > \epsilon\right) \leq 2\exp\left\{-\frac{2n\epsilon^2}{b^2}\right\}$$

By a union bound (with $\mathcal{L}$ finite)

$$\mathbb{P}\left(\sup_{\ell \in \mathcal{L}_n}\left|\hat{\mathbb{P}}\ell - \mathbb{P}\ell\right| > \epsilon\right) \leq 2|\mathcal{L}_n|\exp\left\{-\frac{2n\epsilon^2}{b^2}\right\}$$

$(\mathbb{P} \cup_{j=1}^{J} A_j \leq \sum_{j=1}^{J} \mathbb{P}(A_j))$

This pairing is quite common in basic theory

Example: Uniform risk

This gives us an in probability convergence

This can be strengthened to an almost sure convergence via the Borel-Cantelli lemma

If $|\mathcal{L}_n|$ grows slowly enough for

$$\sum_{n=1}^{\infty} |\mathcal{L}_n| \exp\left\{-\frac{2n\epsilon^2}{b^2}\right\} < \infty$$

Then

$$\sup_{\ell \in \mathcal{L}_n} \left|\hat{\mathbb{P}}\ell - \mathbb{P}\ell\right| \to 0 \text{ a.s.}$$

Example: Uniform risk

Though this will work for some situations, $|\mathcal{L}_n|$ is really infinite

In this case, the goal becomes to find a finite set $\mathcal{L}_{n,\epsilon}$ such that

$$\left\{ \sup_{\ell \in \mathcal{L}_n} \left| \hat{\mathbb{P}}\ell - \mathbb{P}\ell \right| > \epsilon \right\} \subseteq \left\{ \sup_{\ell \in \mathcal{L}_{n,\epsilon}} \left| \hat{\mathbb{P}}\ell - \mathbb{P}\ell \right| > \epsilon' \right\}$$

Letting ϵ' be a function of ϵ but not n

We construct $\mathcal{L}_{n,\epsilon}$ using covering numbers with respect to different metrics $||\cdot||$

(This corresponds to $(T, d) \leftrightarrow (\mathcal{L}_n, ||\cdot||)$)

Example: Uniform risk

The most basic choice is

$$d_\infty(\ell, \ell') = ||\ell - \ell'||_\infty = \sup_z |\ell(z) - \ell'(z)|$$

An ϵ-cover of $\mathcal{L}_n$ with respect to d_∞ is such that for every $\ell \in \mathcal{L}_n$, $\exists \ell_\epsilon \in \mathcal{L}_{n,\epsilon}$ such that

$$d_\infty(\ell, \ell_\epsilon) = ||\ell - \ell_\epsilon||_\infty = \sup_z |\ell(z) - \ell_\epsilon(z)| < \epsilon$$

We'll define the L_∞ covering number

$$N_\infty(\epsilon, \mathcal{L}_n) = N_\infty(\epsilon, \mathcal{L}_n, ||\cdot||_\infty)$$

to be the minimal ϵ-cover

(Alternatively known as the uniform covering number (not to be confused with uniform convergence))

Example: Uniform risk

Putting this together, we get

$$\mathbb{P}\left(\sup_{\ell\in\mathcal{L}_n}\left|\hat{\mathbb{P}}\ell-\mathbb{P}\ell\right|>\epsilon\right)\leq 2N_\infty(\epsilon/3,\mathcal{L}_n)\exp\left\{-\frac{2n\epsilon^2}{9b^2}\right\}$$

Basic idea: Let $\mathcal{L}_{n,\epsilon/3}$ be a minimal $\epsilon/3$-cover of $\mathcal{L}_n$. Fix a $\ell\in\mathcal{L}_n$, then $\exists\ell'\in\mathcal{L}_{n,\epsilon/3}$ such that $||\ell-\ell'||_\infty<\epsilon/3$

$$\begin{aligned}\left|\hat{\mathbb{P}}\ell-\mathbb{P}\ell\right| &\leq \left|\hat{\mathbb{P}}\ell-\hat{\mathbb{P}}\ell'\right|+\left|\hat{\mathbb{P}}\ell'-\mathbb{P}\ell'\right|+|\mathbb{P}\ell'-\mathbb{P}\ell| \\ &\leq ||\ell-\ell'||_\infty+\left|\hat{\mathbb{P}}\ell'-\mathbb{P}\ell'\right|+||\ell-\ell'||_\infty \\ &\leq 2\epsilon/3+\left|\hat{\mathbb{P}}\ell'-\mathbb{P}\ell'\right|\end{aligned}$$

Now, we do as before, but with $N_\infty(\epsilon/3,\mathcal{L}_n)=|\mathcal{L}_{n,\epsilon/3}|$

Example: Uniform risk

Unfortunately, $N_\infty(\epsilon/3, \mathcal{L}_n)$ is often too large for

$$\sum_{n=1}^{\infty} N_\infty(\epsilon/3, \mathcal{L}_n) \exp\left\{-\frac{2n\epsilon^2}{9b^2}\right\} < \infty$$

(The uniform norm is quite strong)

At issue is that

$$\left|\hat{\mathbb{P}}\ell - \hat{\mathbb{P}}\ell'\right| + |\mathbb{P}\ell' - \mathbb{P}\ell| \leq ||\ell - \ell'||_\infty + ||\ell - \ell'||_\infty$$

is quite coarse

A solution is to appeal to random L_1 norm covers

Example: Uniform risk

The idea is based around fictively creating a ghost sample $Z_1, \ldots, Z_{2n}$ such that $Z_i \overset{i.i.d}{\sim} \mathbb{P}$ for $i = 1, \ldots, 2n$

Now, we will think of $\mathbb{P}\ell \approx \hat{\mathbb{P}}_{n+1}^{2n}\ell$ and form

$$\left\{ \sup_{\ell \in \mathcal{L}_n} \left| \hat{\mathbb{P}}\ell - \hat{\mathbb{P}}_{n+1}^{2n}\ell \right| > \epsilon \right\} \subseteq \left\{ \sup_{\ell \in \mathcal{L}_{n,\epsilon}} \left| \hat{\mathbb{P}}\ell - \hat{\mathbb{P}}_{n+1}^{2n}\ell \right| > \epsilon' \right\}$$

where now $\mathcal{L}_{n,\epsilon}$ is going to be a data-dependent set

(i.e. a random variable)

Now, we cover $\mathcal{L}_n$ with ϵ-covers with respect to

$$||\ell - \ell'||_n = \int |\ell - \ell'| d\hat{\mathbb{P}}_1^n$$

with covering number $N_1(\epsilon, \mathcal{L}_n) = N(\epsilon, \mathcal{L}_n, ||\cdot||_n)$

Example: Uniform risk

The randomness of the covering number can be dealt with by following four steps

1. Ghost sample
2. Introduction of additional randomness via Rademacher random variables
3. Conditioning to introduce covering number
4. Hoeffding

Let's go over each of these briefly

Example: Uniform risk

Ghost sample: It can be shown that for $n \geq 2b^2/\epsilon^2$

$$\mathbb{P}\left(\sup_{\ell\in\mathcal{L}_n}\left|\hat{\mathbb{P}}\ell - \mathbb{P}\ell\right| > \epsilon\right) \leq 2\mathbb{P}\left(\sup_{\ell\in\mathcal{L}_n}\left|\hat{\mathbb{P}}\ell - \hat{\mathbb{P}}_{n+1}^{2n}\ell\right| > \epsilon/2\right)$$

(See pages 136-138 of Györfi et al. (2002) for details)

Example: Uniform risk

Introduction of additional randomness: As all the R.V.s are iid, their difference is invariant to a random sign change with Rademacher R.V.s

$$\mathbb{P}\left(\sup_{\ell\in\mathcal{L}_n}\left|\hat{\mathbb{P}}\ell-\hat{\mathbb{P}}_{n+1}^{2n}\ell\right|>\frac{\epsilon}{2}\right)$$

$$=\mathbb{P}\left(\sup_{\ell\in\mathcal{L}_n}\left|\sum_{i=1}^{n}\epsilon_i\left(\ell(Z_i)-\ell(Z_{i+n})\right)\right|>\frac{n\epsilon}{2}\right)$$

$$\leq\mathbb{P}\left(\sup_{\ell\in\mathcal{L}_n}\left|\sum_{i=1}^{n}\epsilon_i\ell(Z_i)\right|>\frac{n\epsilon}{4}\right)+\mathbb{P}\left(\sup_{\ell\in\mathcal{L}_n}\left|\sum_{i=1}^{n}\epsilon_i\ell(Z_{i+n})\right|>\frac{n\epsilon}{4}\right)$$

$$=2\mathbb{P}\left(\sup_{\ell\in\mathcal{L}_n}\left|\sum_{i=1}^{n}\epsilon_i\ell(Z_i)\right|>\frac{n\epsilon}{4}\right)$$

(This is known as symmetrization. We'll return to this again)

Example: Uniform risk

Conditioning: To introduce the covering number, observe that

$$\mathbb{P}\left(\sup_{\ell\in\mathcal{L}_n}\left|\sum_{i=1}^n \epsilon_i \ell(Z_i)\right| > \frac{n\epsilon}{4}\right)$$
$$= \mathbb{E}_Z \mathbb{P}_\epsilon\left(\exists \ell \in \mathcal{L}_n : \left|\sum_{i=1}^n \epsilon_i \ell(z_i)\right| > \frac{n\epsilon}{4} \middle| (Z_i)_{i=1}^n = (z_i)_{i=1}^n\right)$$

Now, we can apply the complexity part: Let $\mathcal{L}_{n,\epsilon/8}$ be an $\epsilon/8$-cover of $\mathcal{L}_n$ with respect to $||\cdot||_n$:

$$\frac{1}{n}\sum_{i=1}^n |\ell(z_i) - \ell'(z_i)| < \epsilon/8$$

Example: Uniform risk

Conditioning: Fix an ℓ and find its $\ell' \in \mathcal{L}_{n,\epsilon/8}$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \ell(z_i) \right| \leq \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \ell'(z_i) \right| + \epsilon/8$$

Showing (conditional on $(Z_i)_{i=1}^n = (z_i)_{i=1}^n$)

$$\begin{aligned} & \mathbb{P}\left(\exists \ell \in \mathcal{L}_{n,\epsilon/8} : \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \ell(z_i) \right| + \epsilon/8 > \frac{\epsilon}{4} \right) \\ & \leq N_1(\epsilon, \mathcal{L}_n) \max_{\ell \in \mathcal{L}_{n,\epsilon/8}} \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \ell(z_i) \right| > \frac{\epsilon}{8} \right) \end{aligned}$$

Example: Uniform risk

Hoeffding: Lastly, we just need to bound

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \ell(z_i)\right| > \frac{\epsilon}{8}\right)$$

This can be done with Hoeffding

(Recall, though, that Bernstein's tends to give better bounds)

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \ell(z_i)\right| > \frac{\epsilon}{8}\right) \leq 2\exp\left\{-\frac{2n(\epsilon/8)^2}{(2b)^2}\right\} \leq 2\exp\left\{-\frac{2n\epsilon^2}{128b^2}\right\}$$

Example: Uniform risk

In conclusion:

$$\mathbb{P}\left(\sup_{\ell \in \mathcal{L}_n} \left|\hat{\mathbb{P}}\ell - \mathbb{P}\ell\right| > \epsilon\right) \leq 8\mathbb{E}N_1(\epsilon, \mathcal{L}_n)\exp\left\{-\frac{2n\epsilon^2}{128b^2}\right\}$$

Now, we need to know the expected L_1 covering number

This can be difficult to calculate. Hence, we can turn to a few notions

- VC dimension
- Bracketing numbers

Example: Uniform risk

Recall the VC dimension of a class of sets $\mathcal{A}$ is the largest number of points $\mathcal{G}_n$ which we can shatter

(That is, the largest n such that $|\{G \subset \mathcal{G}_n : G = \mathcal{G}_n \cap A, A \in \mathcal{A}\}| = 2^n$)

Let

$$\mathcal{G} = \{(z, t) \in \mathbb{R}^{p+1} \times \mathbb{R} : t \leq \ell(z), \ell \in \mathcal{L}_n\}$$

(This is the set of all subgraphs of functions in $\mathcal{L}_n$)

We can bound the L_q packing numbers by $VC_{\mathcal{G}} =$ VC dimension of $\mathcal{G}$

Example: Uniform risk

Let ν be a probability measure on $\mathbb{R}^{p+1}$ and let $\sup_{\ell \in \mathcal{L}_n} ||\ell||_\infty \leq b$. Then for any $\epsilon \in (0, b/4)$

$$M(\epsilon, \mathcal{L}_n, ||\cdot||_{L_q(\nu)}) \leq 3 \left(\frac{2eb^q}{\epsilon^q} \log \left(\frac{3eb^q}{\epsilon^q} \right) \right)^{VC_{\mathcal{G}}}$$

(Remember: packing numbers are essentially covering numbers for our purposes. Note that the bound on the RHS doesn't depend on ν)

Hence, we can leverage the "agreement" between the empirical L_q norm and packing numbers to bound the random quantity with the nonrandom VC dimension

Unfortunately,

- the gap between these bounds can be huge
- VC dimension can frequently only be upper-bounded itself

Example: Uniform risk

Alternatively, we can use bracketing numbers

For a given function class $\mathcal{F}$, we can make a bracket:

$$[L, U] = \{f \in \mathcal{F} : L(x) \leq f(x) \leq U(x), \forall x\}$$

and a bracketing of $\mathcal{F}$ is a collection of brackets such that

$$\mathcal{F} \subseteq \bigcup [L_j, U_j]$$

We are most interested in $\epsilon - L_q(\mathbb{P})$-bracketings: for all j

$$\left(\int |U_j - L_j|^q d\mathbb{P} \right)^{1/q} \leq \epsilon$$

This smallest $\epsilon - L_q(\mathbb{P})$-bracketing is the bracketing number

$$N_{[]}(\epsilon, \mathcal{F}, ||\cdot||_{L_q(\mathbb{P})})$$

Example: Uniform risk

Bracketing numbers are a bit larger that covering numbers

$$N(\epsilon, \mathcal{F}, ||\cdot||_{L_q(\mathbb{P})}) \leq N_{[\,]}(2\epsilon, \mathcal{F}, ||\cdot||_{L_q(\mathbb{P})})$$

But they provide stronger control over complexity. To wit:

Fundamental GC theorem: If $N_{[\,]}(\epsilon, \mathcal{F}, ||\cdot||_{L_1(\mathbb{P})}) < \infty$ for all $\epsilon > 0$. Then $\mathcal{F}$ is GC

Proof idea: For any $f \in \mathcal{F}$, $\exists j$

$$(\hat{\mathbb{P}} - \mathbb{P})f \leq (\hat{\mathbb{P}} - \mathbb{P})U_j + (\mathbb{P}U_j - f) \leq (\hat{\mathbb{P}} - \mathbb{P})U_j + \epsilon$$

Hence

$$\sup_{f \in \mathcal{F}}(\hat{\mathbb{P}} - \mathbb{P})f \leq \max_j(\hat{\mathbb{P}} - \mathbb{P})U_j + \epsilon \leq 2\epsilon \quad \text{(a.s. for large enough } n)$$

Example: Uniform risk

We can get an exponential bound using bracket, just like for covering numbers/VC dimension:

$$\mathbb{P}\left(\sup_{\ell \in \mathcal{L}_n} |(\hat{\mathbb{P}} - \mathbb{P})\ell| > \epsilon\right) \leq 4N_{[\,]}(\epsilon, \mathcal{F}, ||\cdot||_{L_1(\mathbb{P})}) \exp\left\{-\frac{96n\epsilon^2}{76Fb}\right\}$$

where

- $F = \sup_\ell ||\ell||_{L_1(\mathbb{P})}$
- $b = \sup_\ell ||\ell||_{L_\infty(\mathbb{P})}$

The main utility of this bound is that, in my experience, bracketing numbers are easier to bound/compute

Example: Uniform risk

Important case: Lipschitz in a parameter

Suppose the $\mathcal{L} = \{\ell_\beta : \beta \in \mathcal{B}\}$.

For example, for regression:

$$\ell_\beta(Z) = (Y - X^\top \beta)^2$$

So, the squared error loss class is indexed by the regression coefficients

Sometimes, the complexity of $\mathcal{L}$ can be translated to the complexity of $\mathcal{B}$

(For constrained least squares, this is a huge win as the complexity of norm balls in Euclidean space is well known)

Example: Uniform risk

Important case: Lipschitz in a parameter

Intuitively, a bracket on $\mathcal{B}$ can make a bracket on $\mathcal{L}$, provided $\ell_\beta - \ell_{\beta'}$ can't change too much relative to $\beta - \beta'$.

Translated into mathematics, this is a Lipschitz-type constraint:

Let $(\mathcal{B}, ||\cdot||)$ be a normed subset of $\mathbb{R}^{p+1}$. If there is a function m where

$$|\ell_\beta(z) - \ell_{\beta'}(z)| \leq m(z)\,||\beta - \beta'||$$

then

$$N_{[]}(\epsilon, \mathcal{F}, ||\cdot||_{L_q(\mathbb{P})}) \leq \left(\frac{4\sqrt{p+1}\,\mathrm{diam}(\mathcal{B})\,||m||^q_{L_q(\mathbb{P})}}{\epsilon} \right)^{p+1}$$

(Often this approach won't work in "high dimensions" as the dimension exponential dominates the sample size exponential)

Example: Uniform risk

To get a more careful bound, we need the contraction theorem

(See Ledoux and Talagrand (1991). Often we need to make more assumptions to bound the symmetrized process, as in the last example in this lecture)

If the loss ℓ is Lipschitz, then for any function $f_* \in \mathcal{F}$ and nonrandom z_i

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i\left(\ell_f(z_i)-\ell_{f_*(z_i)}\right)\right|\right)$$
$$\leq 2\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i\left(f(z_i)-f_*(z_i)\right)\right|\right)$$

(See van der Geer (2008) for an interesting application of this technique)

Sub-Gaussian bounds

Sub-Gaussian

A sub-Gaussian random variable is one that has tail decay at least as fast a Gaussian

Note that $X \sim N(0,1)$ has $\mathbb{P}(|X| > x) \leq 2\exp\{-x^2/2\}$ and in fact $||X||_{\psi_2} = \sqrt{8/3}$

Any R.V. that obeys this tail quantity is a sub-Gaussian R.V.

A sub-Gaussian process is a $\{X_t\}_{t \in T}$ where

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

Note that a process is sub-Gaussian with respect to the pseudo-metric on the index set

(This is crucial to remember)

Sub-Gaussian

Reminder: $\mathbb{P}(|X| > x) \leq Ce^{-cx^p}$ iff $||X||_{\psi_p} < \infty$. Hence:

$$||X_s - X_t||_{\psi_2} \leq \sqrt{6}d(s,t)$$

(Here, we can see our approach. If our process isn't necessarily Lipschitz, but it is sub-Gaussian, then it is Lipschitz with respect to the ψ_2-Orlicz norm and d)

Example: Any Gaussian process is sub-Gaussian with respect to the (standard deviation) pseudo-metric $d(s,t) = \sigma(X_s - X_t)$

Example: Let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d uniform($\{-1,1\}$) and $a = (a_1, \ldots, a_n)^\top \in \mathbb{R}^n$. Then

$$X_a = \sum_{i=1}^{n} a_i \epsilon_i$$

is a sub-Gaussian process w.r.t. $d(a,b) = ||a-b||_2$ known as the Rademacher process

Sub-Gaussian

This last example follows from an important special case of Hoeffding's inequality

$$\mathbb{P}(|X_a| > x) \leq 2\exp\left\{-\frac{x^2}{2\,||a||_2^2}\right\}$$

It turns out the innocuous looking Rademacher process is crucial. But first, let's state the main result

Sub-Gaussian

Suppose we sub-Gaussian process $(X_t)_{t \in T}$ such that $||X_s - X_t||_{\psi_2} \leq Cd(s,t)$

Also, let the diameter of T be

$$D = \sup_{s,t \in T} d(s,t)$$

Then[2] if X_t is zero mean

$$\mathbb{E} \sup_{t \in T} X_t \leq K \int_0^D \sqrt{\log(N(\epsilon, T, d))} d\epsilon$$

(See Chapter 1.2 of Talagrand's Generic Chaining. For the symmetrizing statement for lower bounds, see Lemma 1.2.8)

[2] There are some other technical conditions, notably separability, to this result (the sup of a measurable process isn't necessarily measurable)

Sub-Gaussian uniform bound

$$\mathbb{E}\sup_{t\in T} X_t \leq K \int_0^D \sqrt{\log(N(\epsilon, T, d))} d\epsilon$$

Note: The $\sqrt{\log}$ part comes from the inverse of the ψ_2 function. The $\int$ comes from 'adding' together subsets of T

Some notes:

- This is known as Dudley's inequality
- The upper bound can be improved with Talagrand's Γ function, as demonstrated in his book Generic Chaining
- Often, Dudley's bound gives the appropriate rate, but with unnecessary log factors

Sub-Gaussian uniform bound

The most important process is the empirical process:

$$\left|\left|\hat{\mathbb{P}} - \mathbb{P}\right|\right|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\hat{\mathbb{P}}f - \mathbb{P}f|$$

Unfortunately, the empirical process is not sub-Gaussian and we can't use the previous bound

In fact, by Bernstein's inequality, the typical tail behavior is a mixture of sub-Gaussian and sub-exponential tails

The main tool in this case is called symmetrization

(Theorem 1.2.7 in Generic Chaining provides the relevant upper bound for processes that are mixtures of ψ_2 and ψ_1 Orlicz norm bounds. I've never seen anyone go this route in statistics, though)

Sub-Gaussian uniform bound

For Rademacher R.V.s ϵ_i

$$\mathbb{E}\left|\left|\hat{\mathbb{P}}-\mathbb{P}\right|\right|_{\mathcal{F}} \leq \frac{2}{n}\mathbb{E}_Z\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i f(Z_i)\right|$$

This generates a random process that is conditionally sub-Gaussian (Hoeffding's inequality) on the observed data, indexed by the (random) coordinate evaluation $f(Z_i)$

Sub-Gaussian must always be stated with respect to a metric on the indexing set.

This turns out to be rather complicated in this case

Sub-Gaussian uniform bound

Let's return to the generalization error bound for the lasso

Previously, we used Nemirovski's inequality to bound the process

It was conjectured in Greenshtein, Ritov (2004) that the induced $n^{1/4}$ rate could be increased to $n^{1/2}$

(Additionally, they showed that under Gaussian design, they could in fact get the $1/2$ rate)

Recently, this was answered in the affirmative by Bartlett et al. (2012), using the techniques we have discussed so far

(In fact, they were able to make weaker assumptions as well)

Sub-Gaussian uniform bound

The approach is to

1. Symmetrize with Rademacher r.v.s
2. Bound with Dudley's inequality $\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left|\sum_{i=1}^n \epsilon_i f(Z_i)\right|$
3. Compute the $\sqrt{\log(N, \epsilon, T, d)}$, where d is now the complicated norm from the symmetrization
4. Integrate this bound up to the diameter of the indexing set

If we do this, we can increase t like

$$t_n = o\left(\frac{\sqrt{n}}{\log^{3/2}(n)\log^{3/2}(np)}\right)$$

which, compared with Greenstein's result provides a faster rate

$$t_n = o\left(\left(\frac{n}{\log(p)}\right)^{1/4}\right)$$

Bookkeeping

- The best known concentration inequality of the Bernstein-type can be found in Bousquet (2001)
- The best possible bounds for the $\mathbb{E}\sup_{t\in T} X_t$ is given by Talagrand's Γ-function, though I've never seen this used
- symmetrization $+$ Dudley's bound is the more popular route