Theoretical Statistics. Lecture 27.

Peter Bartlett

1. Nonparametric regression.
2. Bootstrap estimators. [vdV23]

Nonparametric regression

Suppose we observe $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$ i.i.d., and we aim to choose $\hat{f}_n : \mathcal{X} \to \mathbb{R}$ from a class F of functions, so that

$$P\ell(Y, \hat{f}_n(X)) - \inf_{f \in F} P\ell(Y, f(X))$$

is small; here, ℓ is a loss function.
e.g., $\ell(y, \hat{y}) = (y - \hat{y})^2$, $\hat{f}_n =$ least squares estimate.

Nonparametric regression: Examples

Linear regression:
$F = \{x \mapsto \beta^T x : \beta \in \Theta\}$ for $\Theta \subseteq \mathbb{R}^k$ convex.

Reproducing kernel Hilbert space:
$F = \{f \in \mathrm{span}\{k(x, \cdot)\}, \|f\|_H \leq B\}$, where k is a reproducing kernel (positive definite, symmetric).
(e.g., Splines, $F = \left\{x \mapsto f(x) : \int (f''(x))^2 \, dx \leq B\right\}$.
Radial basis functions, $k(x, y) = \exp(-(x - y)^T S(x - y)/2)$.)

Monotone regression: $F = \{f$ monotone$\}$.

Convex regression: $F = \{f$ convex$\}$.

In all of these examples, estimation with convex ℓ is finite-dimensional convex optimization.

Nonparametric regression and local complexity

Define

$$
\begin{aligned}
G &= \{(x,y) \mapsto \ell(y, f(x)) : f \in F\},\\
\hat{f}_n &= \text{ empirical risk minimizer},\\
\hat{g}_n &= \ell(y, \hat{f}_n(x)),\\
f^* &= \text{ risk minimizer},\\
g^* &= \ell(y, f^*(x))
\end{aligned}
$$

We've seen that we can use a uniform law of large numbers to bound the excess risk:

$$P\hat{g}_n - Pg^* \le 2\|P - P_n\|_G.$$

Nonparametric regression

It turns out we can sometimes get better rates by considering a local version: Define

$$G(r) = \{g \in G : Pg - Pg^* \leq r\}.$$

Then $\hat{g}_n \in G(r)$ implies $\hat{g}_n \in G(r')$ for $r' = 2\|P - P_n\|_{G(r)}$.

This leads to an improvement, for example, when functions in $G(r)$ are bounded and have variance decreasing with r (e.g., if $\sup_{g,g' \in G(r)} P(g - g')^2$ decreases with r), because in those cases, we can use Bernstein's inequality (see Lecture 3), or a functional version (*Talagrand's inequality*).

Nonparametric regression

We can iterate the argument to show that

$$P\hat{g}_n - Pg^* \leq r^*,$$

where r^* is the fixed point of $r = 2\|P - P_n\|_{G(r)}$.

(And we can replace $\|P - P_n\|_{G(r)}$ throughout by an upper bound, such as the Rademacher complexity, or the entropy integral.)

In many cases, these local Rademacher averages provide optimal bounds on the expected excess risk of empirical risk minimizers.

Nonparametric regression

One crucial condition is that as the expected excess risk, $P(g - g^*)$ gets small, the variance also gets small. For instance, we can use a condition like

$$P(g - g^*)^2 \leq cP(g - g^*).$$

Such a condition is satisfied, for example, by quadratic loss with a convex F, because this loss has large modulus of convexity:

$$\delta_\ell(\epsilon) = \inf \left\{ \frac{\ell(a) + \ell(b)}{2} - \ell\left(\frac{a+b}{2}\right) : |a - b| \geq \epsilon \right\} \geq \frac{\epsilon^2}{4},$$

which implies the risk functional $R(f) = P\ell_f$ has large modulus of convexity:

$$\delta_R(\epsilon) = \inf \left\{ \frac{R(f) + R(g)}{2} - R\left(\frac{f+g}{2}\right) : \sqrt{P(f - g)^2} \geq \epsilon \right\} \geq \frac{\epsilon^2}{4}.$$

Nonparametric regression

This is equivalent to

$$\begin{aligned} \delta_R\left(\sqrt{P(f-g)^2}\right) &\le \frac{R(f)+R(g)}{2} - R\left(\frac{f+g}{2}\right) \\ \Rightarrow \qquad P(f-g)^2 &\le 4\left(\frac{R(f)+R(g)}{2} - R\left(\frac{f+g}{2}\right)\right). \end{aligned}$$

Nonparametric regression

Hence

$$\begin{aligned}
P(g-g^*)^2 &= P\left(\left(\ell(f,Y)-\ell(f^*,Y)\right)^2\right) \\
&\le L^2 P|f-f^*|^2 \qquad \text{(Lipschitz)} \\
&\le 4L^2\left(\frac{R(f)+R(f^*)}{2} - R\left(\frac{f+f^*}{2}\right)\right) \qquad \text{(convexity of } \ell\text{)} \\
&\le 4L^2\left(\frac{R(f)+R(f^*)}{2} - R(f^*)\right) \qquad \text{(convexity of } F\text{)} \\
&= 4L^2\,\frac{R(f)-R(f^*)}{2}.
\end{aligned}$$

Nonparametric regression

For example, for the following examples, we can calculate the entropy integral (which is an upper bound on the Rademacher averages) and calculate the fixed point.

For $\ell =$ quadratic loss and $F =$ the set of 1-Lipschitz functions from $[0, 1]$ to $[0, 1]$ has fixed point that scales as $n^{-2/3}$ (whereas the uniform law scales as $n^{-1/2}$).

For $\ell =$ quadratic loss and $F =$ the set of 1-Lipschitz convex functions from $[0, 1]$ to $[0, 1]$ has fixed point that scales as $n^{-4/5}$.

(It turns out that both rates are minimax optimal. Notice that convexity improves the rate...)

Bootstrap estimation

We have data $X \sim P$, and an estimate $\hat{\theta}$ of a parameter θ. We'd like to construct a confidence interval.

$$P\left(\hat{\theta} - \kappa_\alpha \hat{\sigma} \leq \theta \leq \hat{\theta} + \kappa_{1-\alpha} \hat{\sigma}\right) \geq 1 - 2\alpha,$$

where $\kappa_\alpha, \kappa_{1-\alpha}$ are the corresponding quantiles of $(\hat{\theta} - \theta)/\hat{\sigma}$.

Problem:
The distribution of $(\hat{\theta} - \theta)/\hat{\sigma}$ depends on P.
If we know the asymptotics (e.g., asymptotically normal), we can use that.
What if we want more accurate quantiles for fixed n? **Bootstrap.**

Bootstrap: definition

Estimate P using $\hat{P}$.

$$\begin{array}{rr}
\text{Under } P|\theta\text{:} & \text{Under } \hat{P}|\hat{\theta}\text{:} \\
\hat{\theta} & \hat{\theta}^* \\
\hat{\sigma} & \hat{\sigma}^* \\
\dfrac{\hat{\theta}-\theta}{\hat{\sigma}} & \underbrace{\dfrac{\hat{\theta}^*-\hat{\theta}}{\hat{\sigma}^*}}
\end{array}$$

and the quantiles of $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$ are used to define the **bootstrap confidence interval** around $\hat{\theta}$.

Bootstrap: definition

For example, if κ_α is the appropriate upper α quantile of $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$ under the bootstrap distribution, we set

$$\kappa_\alpha = \arg\min_x \left\{ \hat{P}\left(\frac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}^*} \leq x \middle| \hat{\theta} \right) \geq 1 - \alpha \right\},$$

and then assume

$$P\left(\frac{\hat{\theta} - \theta}{\hat{\sigma}} \leq \kappa_\alpha \right) \approx 1 - \alpha.$$

If $\hat{P}$ approximates P, this is a good approximation.

Bootstrap: definition

How do we estimate $\hat{P}$?

Empirical bootstrap: $\hat{P} = P_n^n,$

$$P_n = \frac{1}{n} \sum_i \delta_{X_i}.$$

Parametric bootstrap: $\hat{P} = P_{\hat{\theta}}^n.$

We use $\hat{P}$ to generate many (say m) bootstrap samples, each of the form $(X_1^*, \ldots, X_n^*)$, each gives a single $(\hat{\theta}^*, \hat{\sigma}^*)$ pair, and the empirical quantiles of $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$ are calculated. The number m affects the accuracy, but it can be made as large as desired, so we consider the 'population case', that is, the distribution of $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$ under $\hat{P}$.

Bootstrap: consistency

We want the bootstrap-estimated confidence intervals to be accurate as $n \to \infty$:

$$\sup_x \left| P\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \le x \right) - \hat{P}_n\left(\frac{\hat{\theta}_n^* - \hat{\theta}}{\hat{\sigma}_n^*} \le x \right) \right| \xrightarrow{P} 0.$$

Under the assumption that

$$P\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \le x \right) \to F(x),$$

where F is continuous, it suffices if, for each x,

$$\hat{P}_n\left(\frac{\hat{\theta}_n^* - \hat{\theta}}{\hat{\sigma}_n^*} \le x \right) \xrightarrow{P} F(x).$$

Bootstrap: consistency

For example, we might have F normal, and then it suffices to show that the bootstrap statistic converges weakly to a normal.

The $1/\hat{\sigma}_n^*$ factor is easy to dispense with, so we can consider just the asymptotics of $\sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}_n\right)$.

Theorem: If $X_1, X_2, \ldots$ are i.i.d. with mean μ and covariance Σ, almost surely on the sequence, the conditional distribution of the bootstrap estimate of the mean satisfies

$$\sqrt{n}\left(\bar{X}_n^* - \bar{X}_n\right) \rightsquigarrow N(0, \Sigma).$$

Bootstrap

Theorem: If

1. $\hat{\theta}_n \overset{as}{\rightarrow} \theta$,
2. $\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow T$,
3. Conditionally almost surely, $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \rightsquigarrow T$,
4. ϕ is continuously differentiable near θ,

then conditionally almost surely,

$$\sqrt{n}\left(\phi\left(\hat{\theta}_n^*\right) - \phi\left(\hat{\theta}_n\right)\right) \rightsquigarrow \phi_\theta'(T).$$

(And, of course, $\sqrt{n}\left(\phi\left(\hat{\theta}_n\right) - \phi(\theta)\right) \rightsquigarrow \phi_\theta'(T)$.)

Bootstrap

So we can go from consistency of the bootstrap in estimating the (normal) distribution of $\hat{\theta}_n - \theta$ to consistency in estimating the distribution of a smooth function of θ. For instance, we can prove consistency of the bootstrap for estimating the distribution of a variance estimate. And we can show that a bootstrap empirical process converges to the same weak limit as the usual empirical process (a Brownian bridge), and exploit the functional delta method to show that suitable functions of the bootstrap empirical distribution function (for example, sample quantiles) have the same weak limit as that of the corresponding functions of the usual empirical distribution function.

Bootstrap

Notice that these results have the flavor of a sanity check: they show only that, if we have normal asymptotics, then the bootstrap will asymptotically also have normal asymptotics. But we could have used a confidence interval based on the normal asymptotics! We might expect the bootstrap estimates to be an improvement on the asymptotics. But to demonstrate an improvement, we need a more refined result.

Bootstrap

One approach is based on an Edgeworth expansion, which is a refined version of the central limit theorem:

$$\begin{aligned} P^n\left(\sqrt{n}\frac{\bar{X}_n-\mu}{\hat{\sigma}_n}\leq x\right) &= \Phi(x) \\ &\quad + \phi(x)\left(\frac{p_1(x;\mu_3)}{\sqrt{n}}+\frac{p_2(x;\mu_3,\mu_4)}{n}+O\left(\frac{1}{n^{3/2}}\right)\right), \end{aligned}$$

where the μ_i are the moments of P. By showing that the bootstrap matches the correct distribution to higher order than $1/\sqrt{n}$, we can demonstrate an improvement over the normal approximation.

Final Exam

- Thursday, May 16, 8am-11am, Evans 334.
- Open book: Bring any material you like.
- Grade is total of best $n - 1$ of n questions.