Didn't start material until 8 mem.
* Majority of term, Mark just presented foundations

9/5/2011

at the end of the day, you'll have to decide on a target parm + estimator
↑ our $P_{st}(\Psi)$

so, let $O \sim P_0 \in \mathcal{M}$
↳ we observe $O_1, \ldots, O_n$ iid from P_0
our estimator $\Psi: \mathcal{M} \to \mathbb{R}^d$
↑ stat dist mapping
$\Psi_0 = \Psi(P_0)$ = TMLE parm value

Estimator = mapping of data to our parm space
↳ $(\hat{\Psi})$ aka an algorithm
↳ We have P_n = the empirical prob dist of $O_1, \ldots, O_n$, w/ $P_n(O = o_i) = \frac{1}{n}$, $i = 1, \ldots, n$
so for P_n, we want $\hat{\Psi}(P_n)$
* if we have randomness in our estimator, recommended to avg across runs
↳ for $\hat{\Psi}: \mathcal{M}_{NP} \to \mathbb{R}^d$
↑ non-parametric

Example: $\Psi(P) = E_P[E_P[Y|A=1, W] - E_P[Y|A=0, W]]$
↑ think he meant E_W
note: $\hat{\Psi}(P_n)$ is a random variable
* a property we'd like is $\hat{\Psi}(P_0) = \Psi_0$
↳ Consistency: $\forall \varepsilon > 0$, $P(|\hat{\Psi}(P_n) - \Psi_0| > \varepsilon) \xrightarrow{n \to \infty} 0$
↳ Doesn't tell us how fast we're learning though, i.e. how fast approaching 0
* We need Asymp. Linearity (at P_0): $\hat{\Psi}(P_n) - \Psi_0 = \frac{1}{n} \sum_{i=1}^{n} IC(P_0)(O_i) + o_p(\frac{1}{\sqrt{n}})$
behaves as empirical mean
rv, mult w/ $\sqrt{n}$ it still goes to 0 in prob
↳ cont.

unknown

cont. $IC(P_0)$ is a function of O, w/ mean 0 and finite variance
↳ It is called the influence curve of the $\hat{\Psi}$ at P_0
Note: There are parms where you won't get $\hat{\Psi}(P_n)$ that are asymp. linear

* So for n large enough, $\hat{\Psi}(P_n) - \Psi_0$ will behave like an empirical mean 0 and variance of the IC
$\Rightarrow \sqrt{n}(\hat{\Psi}(P_n) - \Psi_0) \xrightarrow{D} N(0, \sigma^2 = var(IC(P_0)(O)))$
$\Rightarrow Z_n \xrightarrow{D} Z \equiv P(Z_n \le x) \xrightarrow{n\to\infty} P(Z \le x) \ \forall x$
if dealing w/ vector, have $\sqrt{n}(\hat{\Psi}(P_n) - \Psi_0) \xrightarrow{D} N(0, \Sigma^2 = E_0(IC(O)IC(O)^T))$

This allows us to determine $P(-1.96 < \frac{\sqrt{n}(\hat{\Psi}(P_n) - \Psi_0)}{\hat{\Sigma}} < 1.96) \xrightarrow{n\to\infty} 0.95$ (i.e. 95% CI)

$P\left(\hat{\Psi}(P_n) - 1.96\frac{\hat{\Sigma}}{\sqrt{n}} < \Psi_0 < \hat{\Psi}(P_n) + 1.96\frac{\hat{\Sigma}}{\sqrt{n}}\right) \xrightarrow{n\to\infty} 0.95$

* To understand an estimator, will need to calc. Influence Curve.
↳ can estimate it w/ $\hat{IC}$
↳ gives us var-covar matrix est of $\hat{\Sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{IC}(O_i)\hat{IC}(O_i)^T$

⇒ To learn about IC's and how to calculate them, read all 3 appendix A's.

$P_n f \equiv \int f(O)\, dP_n(O) = \frac{1}{n}\sum_{i=1}^{n} f(O_i)$
↳ $\hat{\Psi}(P_n) = \hat{\Psi}(P_n f : f \in F)$

Ex. Take correlation:
We only need to estimate 5 values from the dist to est this:
$E(X), E(Y), E(X^2), E(Y^2), cov(X,Y)$
$\Rightarrow corr(X,Y) = f(E(X), E(Y), \sim, \sim, \sim)$

* The more adaptive the estimator, the more it depends on empirical data

F can be a big family of functions of O
eg. $F = \{I(O=o) : o\}$

Suppose our target parm $\Psi(P) = E_P[E_W(Y|A=1,W) - E_W(Y|A=0,W)]$
↳ says we need all empirical probabilities

For what class of func would: (for probabilistic to consider)
$\sup_{f\in F} |(P_n - P_0)f| \xrightarrow{P} 0$
↳ Cont.

cont. $G_n = (\sqrt{n}(P_n - P_0)f : f)$

↑ random func b/c for a given P_n, can calc for all values of f.

↳ $\Rightarrow$ Gaussian Process $(G) \cong (G_f : f \in \mathcal{F})$

true when certain entropy conditions hold.

Note: G_n = Glivenko class

$\mathcal{F}$ = Donsker class

* Listen to Mark prove consistency again: used $\hat{\psi}(P_n) = \vec{\psi}(P_n f : f \in \mathcal{F})$

↳ should look at section A-1

If $\hat{\psi}$ is differentiable, then we have an inheritance

$\vec{\psi}(P_n) - \vec{\psi}(P_0) \cong \hat{\psi}'(P_0)(P_n - P_0)$ ←

vectors

Recall: $f(x_n) - f(x) \cong f'(x)(x_n - x)$

that's what we do above

only instead of one var, we apply over whole vector

$= \sum_{f \in \mathcal{F}} \frac{d}{dP_0 f} \hat{\psi}(P_0 f : f \in \mathcal{F}) \cdot \underbrace{(P_n - P_0)(f)}_{= \frac{1}{n}\sum_{i=1}^{n}[f(O_i) - P_0 f]}$ ie the emp. mean

$= \frac{1}{n}\sum_{i=1}^{n} \sum_{f \in \mathcal{F}} \frac{d}{dP_0 f} \vec{\psi}(P_0)[f(O_i) - P_0 f]$

The influence curve! $= IC(P_0)(O_i)$

↳ thus, the $IC(P_0)(O_i) = \hat{\psi}'(P_0)(f(O) - P_0 f : f \in \mathcal{F})$

* can be written as sum of partial ds

Note: $\Psi'(P_0)(h) = \frac{\partial}{\partial \varepsilon}\Psi(P_0 + \varepsilon h)\big|_{\varepsilon = 0}$

↑ h being the $f(O) - P_0 f : f \in \mathcal{F}$ vector

Gateaux derivative (aka directional derivative)

* If we can apply certain chars to our estimator, then our estimator will be asymp linear

Good Properties of IC

1) CI's can be computed
2) Robust stats
3) Allows us to see what behaviours are needed for good estimators

↙

4) Suppose we have two estimators that are both asymp. linear

↳ the one w/ the lower var(IC) is the more efficient one.

↳ so IC allows us to compare estimators.

Note: We can calc. efficient IC by taking derivative of Ψ