Stats 300b: Theory of Statistics **Winter 2017**

Lecture 20 – March 15

Lecturer: John Duchi *Scribe: Teng Zhang*

Warning: *these notes may contain factual errors*

Reading: VDV Chapters 7 and 8; Notes on class website.

Outline

- limiting Gaussian experiments
- local asymptotic minimax theorem

1 Recap

Definition 1.1. A collection $\{P_{\theta,n}\}_{\theta\in\Theta,\, n\in\mathbb{N}}$ is locally asymptotically normal (LAN) at $\theta_0 \in int(\Theta)$ with precision/information $K \in \mathbb{R}^{d\times d}$ if there exists $\Delta_n \in \mathbb{R}^d$ such that:

$$\log\left(\frac{dP_{\theta_0+\frac{h}{\sqrt{n}},n}}{dP_{\theta_0,n}}(X^n)\right) = h^T\Delta_n - \frac{1}{2}h^T Kh + o_{P_{\theta,n}}(||h||)$$

where $\Delta_n \xrightarrow[P_{\theta_0,n}]{d} \mathcal{N}(0,K)$.

Le Cam's third lemma implies that,

$$\Delta_n \xrightarrow[P_{\theta_0+\frac{h}{\sqrt{n}},n}]{d} \mathcal{N}(Kh,K)$$

or, with $Z_n = K^{-1}\Delta_n$, $Z_n \xrightarrow[P_{\theta_0+\frac{h}{\sqrt{n}},n}]{d} \mathcal{N}(h,K^{-1})$

The **goal** is to show how to use this to get asymptotic optimality/lower bounds in estimation problems. We will look at estimating h in local model $P_{\theta_0+\frac{h}{\sqrt{n}}}$ as h varies.

2 Limiting Gaussianity

Throughout, we assume that $\theta_0 = 0$ (wlog). We want to show that "local" experiments $P_{\frac{h}{\sqrt{n}},n}$ asymptotically look like Gaussian location family experiments/observations. To do that, we first provide some heuristics.

In LAN, if we want to estimate h, asymptotically Δ_n should be sufficient. In other words, if we only want to estimate h, then $h^T\Delta_n - \frac{1}{2}h^TKh$ contains all the relevant information.

Say, we want to estimate h in a Bayesian model: draw $h \sim N(0, \Gamma)$ denoted by $\pi(\cdot)$, and then sample $X^n | h \sim P_{\frac{h}{\sqrt{n}}, n}$. We can do an approximation for the posterior distribution of $h|X^n$ by

$$
\begin{aligned}
\pi(h|X^n) &\propto \frac{dP_{\frac{h}{\sqrt{n}},n}}{dP_{0,n}}(X^n)\pi(h) \\
&\approx \exp(h^T\Delta_n - \frac{1}{2}h^T K h)\exp(\frac{1}{2}h^T\Gamma^{-1}h) \\
&= \exp(-\frac{1}{2}(h - (K + \Gamma^{-1})^{-1}\Delta_n)^T(K + \Gamma^{-1})(h - (K + \Gamma^{-1})^{-1}\Delta_n) + \mathit{function}(\Delta, K))
\end{aligned}
$$

where we use $o_{P_{0,n}}(\|h\|) = 0$ as the approximation. i.e we have $h|X^n \sim N((K + \Gamma^{-1})^{-1}\Delta_n, (K + \Gamma^{-1})^{-1})$. Then take $\Gamma \to \infty$ (diffuse prior on h), then $h|X^n \sim N(K^{-1}\Delta_n, K^{-1})$ in some asymptotic sense.

Making the posterior limit rigorous (Le Cam, Le Cam & Yang):
Define, for $K \succeq 0$, $\Gamma \succeq 0$, the Gaussian distribution

$$G_{K,\Gamma}(\cdot|z) = \mathcal{N}\left((K + \Gamma^{-1})^{-1}Kz,\, (K + \Gamma^{-1})^{-1}\right)$$

Remark This is the posterior distribution of $h|z$ in the model $h \sim N(0, \Gamma), z|h \sim N(h, K^{-1})$.
Idea In LAN family, let $Z_n = K^{-1}\Delta_n$. Then shift h in $dP_{\frac{h}{\sqrt{n}},n}$ should have asymptotic posterior $G_{K,\Gamma}(\cdot|Z_n)$.
Let $\pi^{\Gamma,c}$ be Gaussian distribution $N(0, \Gamma)$ truncated to set $\{h \in \mathbb{R}^d : \|h\| \leq c\}$ and renormalized.
Theorem 2.1 Assume that data X^n satisfy $X^n|h \sim P_{\frac{h}{\sqrt{n}},n}$. Denote $Z_n := K^{-1}\Delta_n(X^n)$ (LAN family), $\bar{P}_n(\cdot) := \int P_{\frac{h}{\sqrt{n}},n}(\,.\,)d\pi^{\Gamma,c}(h)$ the marginal distribution of X^n, $\pi^{\Gamma,c}(\cdot|X^n) :=$ the posterior on h condition on X^n. Then, for all $\epsilon > 0$, there exist $C, N < +\infty$ such that for all $n \geq N$, $c \geq C$,

$$\int ||G_{K,\Gamma}(\cdot|z_n(x^n)) - \pi^{\Gamma,c}(\cdot|x^n)||_{TV} d\bar{P}_n(x^n) \leq \epsilon$$

Proof See notes.
Remark The true posterior of a LAN family, under truncated Gaussian prior, is, on average, really close to a Gaussian distribution, conditioned on $Z_n = K^{-1}\Delta_n(x^n)$. (not the only notion of limiting Gaussianity for LAN families)

3 Local asymptotic minimax theorem

We now reduce everything to estimation in Gaussian shift experiments $N(h, K^{-1})$ as h varies in $\mathbb{R}^d$.
Definition 3.1. A function $L : \mathbb{R}^d \mapsto \mathbb{R}$ is quasi-convex if for all $\alpha \in \mathbb{R}$, the α-sublevel set $\{x : L(x) \leq \alpha\}$ is convex.

Example 3.1. $L(x) = \frac{1}{2}||x||_2^2 \wedge B$ is quasi-convex for any $B \in \mathbb{R}$.

Lemma 3.1. **(Anderson)** Let L be symmetric (i.e. $L(z) = L(-z)$) and quasi-convex. Let $A \in \mathbb{R}^{d\times k}$ and $X \sim \mathcal{N}(\mu, \Sigma)$. Then:

$$\inf_{v\in\mathbb{R}^k} \mathbb{E}\left[L(AX - v)\right] = \mathbb{E}\left[L(A(X - \mu))\right] = \mathbb{E}\left[L(A\Sigma^{\frac{1}{2}}W)\right]$$

where $W \sim \mathcal{N}(0, I)$.

Theorem 3.1. (Local asymptotic minimax)
Let $L : \mathbb{R}^d \mapsto \mathbb{R}$ be quasi-convex, symmetric and bounded. Let $\{P_{\theta,n}\}$ be LAN at θ_0 with information $K \succeq 0$. Then, with $W \sim \mathcal{N}(0, I)$ and $Z \sim \mathcal{N}(0, K^{-1})$,

$$\liminf_{c\to\infty} \liminf_{n\to\infty} \inf_{\hat{\theta}_n} \sup_{\|\theta-\theta_0\|\leq \frac{c}{\sqrt{n}}} \mathbb{E}_{P_{\theta,n}} \left[L(\sqrt{n}(\hat{\theta}_n - \theta))\right] \geq \mathbb{E}\left[L(K^{-\frac{1}{2}}W)\right] = \mathbb{E}\left[L(Z)\right]$$

Remark We can replace supreme over $\|\theta - \theta_0\| \leq \frac{c}{\sqrt{n}}$ with average of θ over $\pi_{c,n}$ = truncated normal $N(\theta_0, \frac{c^2}{n}I)$ truncated to $\|\theta - \theta_0\| \leq \frac{c}{\sqrt{n}}$.
Corollary 3.1. Consider a quadratic mean differentiable family $\{P_\theta\}_{\theta\in\Theta}$ with Fisher information I_{θ_0} at parameter θ_0. Then the theorem implies that:

$$\liminf_{c\to\infty} \liminf_{n\to\infty} \inf_{\hat{\theta}_n} \int \mathbb{E}_{P_\theta^n} \left[L(\sqrt{n}(\hat{\theta}_n - \theta))\right] d\pi_{c,n}(\theta) \geq \mathbb{E}[L(Z)]$$

with $Z \sim \mathcal{N}(0, I_{\theta_0}^{-1})$.
So efficient estimators (i.e. estimator $\hat{\theta}_n$ satisfying $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I_{\theta_0}^{-1}))$ exist and are locally asymptotically optimal. Anytime you have $\hat{\theta}_n = \theta + \frac{1}{n} I_\theta^{-1} \sum_{i=1}^n \dot{\ell}(X_i) + o_{P_\theta}(\frac{1}{\sqrt{n}})$ under P_θ, then $\hat{\theta}_n$ is efficient and achieves LAMT bound for all θ by contiguity.
Proof of Theorem 3.1.
Without loss of generality, assume that L takes values in $[0, 1]$ and $\theta_0 = 0$.
Observe that

$$\sup_{||h||\leq c} \mathbb{E}_{P_{\frac{h}{\sqrt{n}},n}} \left[L(\sqrt{n}(\hat{\theta}_n - \theta))\right] \geq \int \mathbb{E}_{P_{\frac{h}{\sqrt{n}},n}} \left[L(\sqrt{n}\hat{\theta}_n - h)\right] d\pi(h)$$

where $\theta = \frac{h}{\sqrt{n}}$, for any π with support in $\{h : ||h|| \leq c\}$.

Consider $\pi := \pi^{\Gamma,c}$, prior of h, to be the normal distribution $\mathcal{N}(0, \Gamma)$, truncated to $\{h : ||h|| \leq c\}$ and denote the marginal distribution of X^n:

$$\bar{P}_n(\cdot) = \int P_{\frac{h}{\sqrt{n}},n}(\cdot) d\pi^{\Gamma,c}(h)$$

where $X^n|h \sim P_{\frac{h}{\sqrt{n}},n}$ with posterior $\pi(h \in \cdot|X^n)$ on h. Then, the left hand-side $(*)$ of the last inequality satisfies:

$$(*) \geq \int \mathbb{E}\left[L(\sqrt{n}\hat{\theta}_n - h) \,|\, X^n = x^n\right] d\bar{P}_n(x^n)$$

Using the previous notation $G_{K,\Gamma}$, we get:

$$(*) \geq \int \inf_{\hat{h}} \mathbb{E}_{G_{K,\Gamma}} \left[L(\hat{h} - h) \,|\, x^n\right] d\bar{P}_n(x^n) - \int \sup_{h,\hat{h}} L(\hat{h} - h) \left(G_{K,\Gamma}(h \,|\, x^n) - \pi(h \,|\, x^n)\right) d\bar{P}_n(x^n)$$

Observe that for the second term:

$$\int \sup_{h,\hat{h}} L(\hat{h} - h) \left(dG_{K,\Gamma}(h \,|\, x^n) - \pi(h \,|\, x^n)\right) d\bar{P}(x^n) \leq \int ||G_{K,\Gamma}(\,.\,|\, x^n) - \pi(\,.\,|x^n)||_{TV} d\bar{P}_n(x^n)$$

and that, by Theorem 2.1, the right-hand side of the last inequality is less than ϵ, for any $\epsilon > 0$, c appropriately chosen and n sufficiently large.

Moreover, by Anderson's lemma, for the first term we have:

$$\int \inf_{\hat{h}} \mathbb{E}_{G_{K,\Gamma}} \left[L(\hat{h} - h) \,|\, x^n \right] d\bar{P}_n(x^n) \geq \int \mathbb{E} \left[L(\mathcal{N}(0, (K + \Gamma^{-1})^{-1})) \right] d\bar{P}_n$$

Taking $\Gamma \to \infty$, we get:

$$(*) \geq \mathbb{E} \left[L(Z) \right] - \epsilon$$