STATISTICAL THEORY SUMMARY

ANDREW TULLOCH

1. Basic Concepts

Definition (Convergence almost surely). *A sequence $X_n, n \in \mathbb{N}$ of random variables converges almost surely to a random variable X if*

$$\mathbb{P}(X_n \to X) = \mu(\omega \in \Omega | X_n(\omega) \to X(\omega)) = 1 \tag{1.1}$$

We say that $X_n \overset{as}{\to} X$.

Definition (Convergence in probability). *$X_n \overset{p}{\to} X$ (in probability) if for all $\epsilon > 0$,*

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0 \tag{1.2}$$

as $n \to \infty$. For random vectors, we define analogously with taking the norm in $\mathbb{R}^n$.

Definition (Convergence in distribution). *$X_n \overset{d}{\to} X$ or X_n converges to X in distribution if*

$$\mathbb{P}(X_n \le t) \to \mathbb{P}(X \le t) \tag{1.3}$$

whenever $t \mapsto \mathbb{P}(X \le t)$ is continuous.

Proposition. *Let $(X_n, n \in \mathbb{N})$, X taking values in $\mathcal{X} \subseteq \mathbb{R}^d$.*

(i)

$$X_n \overset{as}{\to} X \Rightarrow X_n \overset{p}{\to} X \Rightarrow X_n \overset{d}{\to} X \tag{1.4}$$

(ii) If $X_n \to X$ in any mode, and if $g : \mathcal{X} \to \mathbb{R}^d$ is continuous, then $g(X_n) \to g(X)$in the same mode.

*(iii) **Slutsky's lemma** If $X_n \overset{d}{\to} X$ and $Y_n \overset{d}{\to} c$ (a constant). Then*

- *(i) $Y_n \overset{p}{\to} c$*
- *(ii) $X_n + Y_n \overset{d}{\to} X + c$*
- *(iii) $X_n Y_n \overset{d}{\to} cX$ where $Y_n \in \mathbb{R}$.*
- *(iv) $X_n Y_n^{-1} \overset{d}{\to} c^{-1} X$ where $Y_n \in \mathbb{R}, c \neq 0$.*

(iv) If $(A_n, n \in \mathbb{N})$ are random matrices with $(A_n)_{ij} \overset{p}{\to} A_{ij}$ for all i, j and $X_n \overset{d}{\to} X$, then $A_n X_n \overset{d}{\to} AX$, and if A is invertible, $A_n^{-1} X_n \overset{d}{\to} A^{-1} X$, where $A = (A_{ij})$.

Theorem (Law of Large Numbers). let $X_1, \dots, X_n$ be IID copies of X $\mathbb{P}$ such that $\mathbb{E}(|X_i|) < \infty$, then

$$\frac{1}{n} \sum_{i=1}^n X_i \overset{as}{\to} \mathbb{E}(X) \tag{1.5}$$

Theorem (Central limit theorem). Let $X_1, \dots, X_n$ be IID copies of $X \sim \mathbb{P}$ on $\mathbb{R}$ with $\mathbb{V}(X) = \sigma^2 < \infty$. Then

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}(X) \right) \overset{d}{\to} N(0, \sigma^2) \tag{1.6}$$

In the multivariate case, where $X \sim \mathbb{P}$ on $\mathbb{R}^d$ with the covariance of X as Σ, then

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}(X) \right) \overset{d}{\to} N(0, \Sigma) \tag{1.7}$$

Theorem (Gaussian Tail Inequality). If $X \sim N(0, 1)$, then

$$\mathbb{P}(|X| > \epsilon) \le \frac{2e^{-\epsilon^2}}{\epsilon}. \tag{1.8}$$

Theorem (Chebyshev's Inequality). Let $\mu = \mathbb{E}(X)$, $\sigma^2 = \mathbb{V}(X)$. Then

$$\mathbb{P}(|X - \mu| \ge t) \le \frac{\sigma^2}{t^2} \tag{1.9}$$

Theorem (Markov's Inequality). Let X be a non-negative random variable and suppose EX exists. Then for any $t > 0$,

$$\mathbb{P}(X > t) \le \frac{\mathbb{E}(X)}{t} \tag{1.10}$$

Theorem (Hoeffding's Inequality). Suppose $EX = 0$, and $a \le X \le b$. Then

$$\mathbb{E}\left(e^{tX}\right) \le e^{\frac{t^2(b-a)^2}{8}}. \tag{1.11}$$

2. Uniform Laws of Large Numbers

Theorem. Let $\mathcal{H}$ be a class of functions from a measurable space T to $\mathbb{R}$. Assume that for every $\epsilon > 0$ there exists a finite set of brackets $[l_j, u_j]$, $j = 1, \dots, N(\epsilon)$, such that $\mathbb{E}(|l_j(X)|) < \infty$, $\mathbb{E}(|u_j(X)|) < \infty$, and $\mathbb{E}(|u_j(X) - l_j(X)|) < \epsilon$ for every j. Suppose moreover that for every $h \in H$ there exists j with $h \in [l_j, u_j] \iff h \in \{f : T \to \mathbb{R} | l(x) \le f(x) \le u(x), \forall x \in T\}$. Then we have a **uniform law of large numbers**,

$$\sup_{h \in H} \left| \frac{1}{n} \sum_{i=1}^n (h(X_i) - \mathbb{E}(h(X))) \right| \overset{as}{\to} 0 \tag{2.1}$$

Proof. Let $\epsilon > 0$, and choose brackets $[l_j, u_j]$ such that $\mathbb{E}(|u_j - l_j) (X) < \frac{\epsilon}{2}$ by hypothesis. Then for every $\omega \in T^\infty$ outside a null set A, there exists $n_0(\omega, \epsilon)$ such that

$$\max_{j=1,\dots,N(\frac{\epsilon}{2})} | \sum_{i=1}^n u_j - \mathbb{E}(u_j) | < \frac{\epsilon}{2} \tag{2.2}$$

by the strong law of large numbers and taking a union over all j.

Then this gives for $h \in \mathcal{H}$, outside a set of zero measure,

$$\frac{1}{n} \sum_{i=1}^n h(X_i) - \mathbb{E}(h(X)) \frac{1}{n} \sum_{i=1}^n u_j - \mathbb{E}(h) \tag{2.3}$$

$$= \frac{1}{n} \sum_{i=1}^n u_j - \mathbb{E}(u_j) + \mathbb{E}(u_j) - \mathbb{E}(h) \tag{2.4}$$

which is bounded above by $\frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$, as required. □

3. Consistency of M-estimators

Theorem. Let $\Theta \subseteq \mathbb{R}^p$ be compact. Let $Q : \Theta \to \mathbb{R}$ be a continuous, non-random function that has a unique minimizer $\theta_0 \in \Theta$.

Let $Q_n : \Theta \to \mathbb{R}$ be any sequence of random functions such that

$$\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \overset{p}{\to} 0 \tag{3.1}$$

as $n \to \infty$.

If θ_n is **any** sequence of minimizers of Q_n, then $\hat{\theta}_n \overset{p}{\to} \theta_0$ as $n \to \infty$.

Proof. The key is to consider the set $S_\epsilon = \{\theta \in \Theta : \|\theta - \theta_0\| \ge \epsilon\}$. This is compact, with Q continuous on this set, and so an infimum $Q(\theta_\epsilon) > Q(\theta_0)$ is attained on this set. Choose $\delta > 0$ so $Q(\theta_\epsilon) - \delta > Q(\theta_0) + \delta$. Then consider the set $A_n(\epsilon) = \{\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| < \delta\}$. On this set, we have

$$\inf_{S_\epsilon} Q_n(\theta) \ge \inf_{S_\epsilon} Q(\theta) - \delta = Q(\theta_\epsilon) - \delta > Q(\theta_0) + \delta \ge Q_n(\theta_0). \tag{3.2}$$

So if $\hat{\theta}_n$ lay in S_ϵ, then $Q_n(\delta_0)$ would be strictly smaller than $Q_n(\hat{\theta}_n)$, contradicting that $\hat{\theta}_n$ is a minimizer. Conclude that $A_n(\epsilon) \Rightarrow \|\hat{\theta}_n - \theta_0\| < \epsilon$, but as $\mathbb{P}(A_n(\epsilon)) \to 1$, we have $\mathbb{P}\left(\|\hat{\theta}_n - \theta_0\| < \epsilon\right) \to 1$. □

Theorem. Let Θ be compact in $\mathbb{R}^p$, and let $\mathcal{X} \subseteq \mathbb{R}^d$ and consider observing $X_1, \dots, X_n$ IID from $X \sim \mathbb{P}$ on X. Let $q : \mathcal{X} \times \Theta \to \mathbb{R}$ that is continuous in θ for all x and measurable in x for all $\theta \subseteq \Theta$.

Assume

$$\mathbb{E}\left(\sup_{\theta \in \Theta} |q(X, \theta)|\right) < \infty \tag{3.3}$$

Then

$$\sup_{\theta \in \Theta} | \frac{1}{n} q(X_i, \theta) - \mathbb{E}(q(X, \theta)) | \overset{as}{\to} 0 \tag{3.4}$$

as $n \to \infty$

Proof. We seek to find suitable bracketing functions and proceed via the uniform law of large numbers. First, define the brackets $u(x, \theta, \eta) = \sup_{\theta' \in B(0, \eta)}$, so $\mathbb{E}(|u(X, \theta, \eta)|) < \infty$ by assumption. By continuity of $q(\cdot, x)$, the supremum is achieved at points $\|\theta^u(\theta) - \theta\| < \eta$, and so $\lim_{\eta \to 0} |u(x, \theta, \eta) - q(\theta, x) \to 0$ at every x and every θ, and using the dominated convergence theorem gives $\lim_{\eta \to 0} \mathbb{E}(|u(X, \theta, \eta) - q(\theta, X)) \to 0$.

Then for $\epsilon > 0$ and $\theta \in \Theta$ we can choose $\eta(\epsilon, \theta)$ small so that

$$\mathbb{E}(u(X, \theta, \eta) - l(X, \theta, \eta)) < \epsilon. \tag{3.5}$$

The open balls $\{B(\theta, \eta(\epsilon, \theta))\}$ are an open cover of the compact set Θ, so there exists a finite subcover by Heine-Borel. This finite subcover $u_j(\cdot, \theta_j, \eta(\epsilon, \theta_j))$ constitutes a bracketing set of the q, and so we apply the uniform law of large numbers. □

Theorem (Consistency of the Maximum Likelihood Estimator)**.** Consider the model $f(\theta, y)$, $\theta \in \Theta \subseteq \mathbb{R}^p$, $y \in \mathcal{Y} \subset \mathbb{R}^d$. Assume $f(\theta, y) > 0$ for all $y \in \mathcal{Y}$ and all $\theta \in \Theta$, and that $\int_{\mathcal{Y}} f(\theta, y)dy = 1$ for every $\theta \in \Theta$. Assume further that Θ is compact and that the map $\theta \mapsto f(\theta, y)$ is continuous on Θ for every $y \in \mathcal{Y}$. Let $Y_1, \ldots, Y_n$ be IID with common density $f(\theta_0)$, where $\theta_0 \in \Theta$. Suppose finally that the identification condition 13 and the domination condition

$$\int_{\mathcal{Y}} \sup_{\theta' \in \Theta} |\log f(\theta', y)| f(\theta_0, y)dy < \infty \tag{3.6}$$

hold. If $\hat{\theta}_n$ is the MLE in the model $\{f(\theta, \cdot)|\theta \in \Theta\}$ based on the sample $Y_1, \ldots, Y_n$, then $\hat{\theta}_n$ is consistent, in that $\hat{\theta}_n \to^{p_{\theta_0}} \theta_0$ as $n \to \infty$.

Proof. Setting

$$q(\theta, y) = -\log f(\theta, y), \tag{3.7}$$

$$Q(\theta) = \mathbb{E}_{\theta_0} q(\theta, Y), \tag{3.8}$$

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} q(\theta, Y_i), \tag{3.9}$$

this follows from the previous results. □

Definition (Uniform Consistency)**.** *An estimator T_n is* ***uniformly consistent*** *in $\theta \in \Theta$, if for every $\delta > 0$,*

$$\sup_{\theta_0 \in \Theta} P_{\theta_0}(\|T_n - \theta_0\| > \delta) \to 0 \tag{3.10}$$

as $n \to \infty$.

Theorem. An estimator is uniformly consistent if, for every $\epsilon > 0$,

$$\inf_{\theta_0 \in \Theta} \inf_{\theta \in \Theta : \|\theta - \theta_0\| \geq \epsilon} (Q(\theta) - Q(\theta_0)) > 0 \tag{3.11}$$

and that

$$\sup_{\theta_0 \in \Theta} \mathbb{P}_{\theta_0}(\sup_{\theta \in \Theta} |Q_n(\theta; Y_1, \ldots, Y_n) - Q(\theta)| > \delta) \to 0. \tag{3.12}$$

4. Asymptotic Distribution Theory

Theorem. Consider the model $f(\theta, y), \theta \in \Theta \subset \mathbb{R}^p$, $y \in \mathcal{Y} \subset \mathbb{R}^d$. Assume $f(\theta, y) > 0$for all $y \in \mathcal{Y}$ and all $\theta \in \Theta$, and that $\int_{\mathcal{Y}} f(\theta, y)dy = 1$ for every $\theta \in \Theta$. Let $Y_1, \ldots, Y_n$ be IID from density $f(\theta_0, y)$ for some $\theta_0 \in \Theta$. Assume moreover

(i) θ_0 is an interior point on Θ.

(ii) There exists an open set U satisfying $\theta_0 \in U \subset \Theta$ such that $f(\theta, y)$ is, for every $y \in \mathcal{Y}$, twice continuously differentiable with respect to θ on U,

(iii) $\mathbb{E}_{\theta_0} \frac{\partial^2 \log f(\theta_0, Y)}{\partial\theta\partial\theta^T}$ is nonsingular, and

$$\mathbb{E}_{\theta_0} \| \frac{\partial \log f(\theta_0, Y)}{\partial\theta} \|^2 < \infty \tag{4.1}$$

(iv) There exists a compact ball $K \subset U$ with nonempty interior centered at θ_0 such that

$$\mathbb{E}_{\theta_0} \sup_{\theta \in K} \| \frac{\partial^2 \log f(\theta, Y)}{\partial\theta\partial\theta^T} \| < \infty \tag{4.2}$$

$$\int_{\mathcal{Y}} \sup_{\theta \in K} \| \frac{\partial f(\theta, y)}{\partial\theta} \| dy < \infty \tag{4.3}$$

$$\int_{\mathcal{Y}} \sup_{\theta \in K} \| \frac{\partial^2 f(\theta, y)}{\partial\theta\partial\theta^T} \| dy < \infty \tag{4.4}$$

Let $\hat{\theta}_n$ be the MLE in the model $\{f(\theta, \cdot); \theta \in \Theta\}$ based on the sample $Y_1, \ldots, Y_n$, and assume $\hat{\theta}_n \to^{P_{\theta_0}} \theta_0$ as $n \to \infty$. Define the Fisher information

$$i(\theta_0) = \mathbb{E}_{\theta_0} \frac{\partial \log f(\theta_0, Y)}{\partial\theta} \frac{\partial \log f(\theta_0, Y)^T}{\partial\theta} \tag{4.5}$$

Then $i(\theta_0) = -\mathbb{E}_{\theta_0} \frac{\partial^2 \log f(\theta_0, Y)}{\partial\theta\partial\theta^T}$, and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, i^{-i}(\theta_0)) \tag{4.6}$$

as $n \to \infty$.

Proof. First, note that as $\int f(\theta, y)dy = 1$ for all $\theta \in J$, then $\frac{\partial}{\partial\theta} \int f(\theta, y)dy = \int \frac{\partial \log f(\theta, y)}{\partial\theta} f(\theta, y)dy = 0$ for every $\theta \in \operatorname{int} K$, so

$$\mathbb{E}_{\theta_0} \frac{\partial \log f(\theta_0, Y)}{\partial\theta} = 0. \tag{4.7}$$

Since $\hat{\theta}_n \xrightarrow{p} \theta_0$, we have $\hat{\theta}_n$ is an interior point of Θ on events of probability approaching one, so $\frac{\partial Q_n(\hat{\theta}_n)}{\partial\theta} = 0$. Applying the mean value theorem, we have

$$0 = \sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta} + \overline{A}_n \sqrt{n}(\hat{\theta}_n - \theta_0) \tag{4.8}$$

where $\hat{A}_n$ is the matrix of second derivatives of Q_n evaluated at a mean value $\overline{\theta}_{nj}$ on the line segment between θ_0 and $\hat{\theta}_n$.

For the first component, we have

$$\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta} = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\partial \log f(\theta_0, Y_i)}{\partial\theta} \xrightarrow{d} N(0, i(\theta_0)) \tag{4.9}$$

by the central limit theorem.

For the second component, we show $\overline{A}_n \xrightarrow{p} -\mathbb{E}\Big(\frac{\partial^2 \log f(\theta_0, Y)}{\partial\theta\partial\theta^T}\Big)$, which we do component-wise. We have $(\overline{A}_n)_{kj} = \frac{1}{n}\sum_{i=1}^{n} h_{kj}(\overline{\theta}_{nj}, Y_i)$, where h_{jk} is the second mixed partial derivative of $-\log f(\theta, Y_i)$, and we seek to show each $h_{jk} \xrightarrow{p} \mathbb{E}\big(h_{jk}(\theta_0, Y)\big)$. This follows by

$$\left| \frac{1}{n}\sum_{i=1}^{n} h_{jk}(\theta_{nj}, Y_i) - \mathbb{E}\big(h_{jk}(\theta_0, Y)\big) \right| \tag{4.10}$$

$$\leq \sup_{\theta \in K} \left| \frac{1}{n}\sum_{i=1}^{n} h_{jk}(\theta, Y_i) - \mathbb{E}\big(h_{jk}(\theta, Y)\big) \right| + \left| \mathbb{E}\Big(h_{jk}(\overline{\theta}_{nj}, Y)\Big) - \mathbb{E}\Big(h_{jk}(\theta_0, Y)\Big) \right| \tag{4.11}$$

then by the uniform law of large numbers, the first term converges to zero, and the fact that $\overline{\theta}_{nj} \to \theta_0$ in probability implies the second term converges to zero. Hence, $-\overline{A}_n \to^{p_{\theta_0}} \mathbb{E}(\theta_0) \frac{\partial^2 \log f(\theta_0, Y)}{\partial\theta\partial\theta^T} \equiv \Sigma(\theta_0)$.

As the limit is invertible we have that $\overline{A}_n$ is invertible on sets with measure approaching one, so we can rewrite the previous result as

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\overline{A}_n^{-1} \sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta} \xrightarrow{d} N(0, \Sigma^{-1}(\theta_0 i(\theta_0)\Sigma^{-1}(\theta_0))). \tag{4.12}$$

from Slutsky's lemma.

Finally, we show $\Sigma(\theta_0) = i(\theta_0)$. This follows from interchanging integration and differentiation to show

$$\frac{\partial}{\partial\theta^T} \int \frac{\partial f(\theta, y)}{\partial\theta} dy = \int \frac{\partial^2 f(\theta, y)}{\partial\theta\partial\theta^T} dy = 0 \tag{4.13}$$

for all $\theta \in \operatorname{int} K$. Then, use the chain rule to show

$$\frac{\partial^2 \log f(\theta, y)}{\partial\theta\partial\theta^T} = \frac{1}{f(\theta, y)} \frac{\partial^2 f(\theta, y)}{\partial\theta\partial\theta^T} - \frac{\partial \log f(\theta, y)}{\partial\theta} \frac{\partial \log f(\theta, y)}{\partial\theta}^T \tag{4.14}$$

and using this identity at θ_0. □

Theorem. In the framework of the previous theorem with $p = 1$ and for $n \in \mathbb{N}$ fixed, let $\tilde{\theta} = \tilde{\theta}(Y_1, \ldots, Y_n)$, be any unbiased estimator of θ — that is, it satisfies $\mathbb{E}_\theta \tilde{\theta} = \theta$ for all $\theta \in \Theta$. Then

$$\mathbb{V}_\theta(\tilde{\theta}_n) \geq \frac{1}{ni(\theta)} \tag{4.15}$$

for all $\theta \in \operatorname{int}(\Theta)$.

Proof. Cauchy-Swartz and $\mathbb{E}_{\theta_0} \frac{\partial \log f(\theta, Y)}{\theta}$. Specifically, letting $l(\theta, Y) = \sum_{i=1}^{n} \frac{d}{d\theta} \log f(\theta, Y_i)$,

$$\mathbb{V}_\theta(\tilde{\theta}) \geq \frac{\mathrm{Cov}_\theta^2(\tilde{\theta}, l'(\theta, Y))}{\mathbb{V}_\theta(l'(\theta, Y))} = \frac{1}{ni(\theta)} \tag{4.16}$$

since

$$\mathrm{Cov}_\theta(\tilde{\theta}, l'(\theta, Y)) = \int \tilde{\theta}(y) l'(\theta, y) \prod_{i=1}^{n} f(\theta, y_i)dy \tag{4.17}$$

$$= \int \tilde{\theta}(y) \frac{d}{d\theta} f(\theta, y)dy = \frac{d}{d\theta}\mathbb{E}_\theta \tilde{\theta} = \frac{d}{d\theta}\theta = 1. \tag{4.18}$$

□

Theorem (Delta Method)**.** Let Θ be an open subset of $\mathbb{R}^p$ and let $\Phi : \Theta \to \mathbb{R}^m$ be differentiable at $\theta \in \Theta$, with derivative $D\Phi_\theta$. Let r_n be a divergent sequence of positive real numbers and let X_n be random variables taking values in Θ such that $r_n(X_n - \theta) \xrightarrow{d} X$ as $n \to \infty$. Then

$$r_n(\Phi(X_n) - \Phi(\theta)) \xrightarrow{d} D\Phi_\theta(X) \tag{4.19}$$

as $n \to \infty$. If $X \sim N(0, i^{-1}(\theta))$, then

$$D\Phi_\theta(X) \sim N(0, \Sigma(\Phi, \Theta)). \tag{4.20}$$

Definition (Likelihood Ratio test statistic)**.** *Suppose we observe* $Y_1, \dots, Y_n$ *from* $f(\theta, \cdot)$*, and consider the testing problem* $H_0 : \theta \in \Theta_0$ *against* $H_1 : \theta \in \Theta$*, where* $\Theta_0 \subset \Theta \subset \mathbb{R}^p$*. The Neyman-Pearson theory suggests to test these hypothesis by the likelihood ratio test statistic*

$$\Lambda_n(\Theta, \Theta_0) = 2 \log \frac{\sup_{\theta \in \Theta} \prod_{i=1}^n f(\theta, Y_i)}{\sup_{\theta \in \Theta_0} \prod_{i=1}^n f(\theta, Y_i)} \tag{4.21}$$

which in terms of the maximum likelihood estimators $\hat{\theta}_n, \hat{\theta}_{n,0}$ *of the models* Θ, Θ_0 *is*

$$\Lambda_n(\Theta, \Theta_0) = -2 \sum_{i=1}^n \log f(\hat{\theta}_{n,0}, Y_i) - \log f(\hat{\theta}_n, Y_i). \tag{4.22}$$

Theorem. Consider a parametric model $f(\theta, y), \theta \in \Theta \subset \mathbb{R}^p$ that satisfies the assumptions of the theorem on asymptotic normality of the MLE. Consider the simple null hypothesis $\Theta_0 = \{\theta_0\}$, $\theta_0 \in \Theta$. Then under H_0, the likelihood ratio test statistic is asymptotically chi-squared distributed, so

$$\Lambda_n(\Theta, \Theta_0) \xrightarrow{d} \chi^2_p \tag{4.23}$$

as $n \to \infty$ under P_{θ_0}.

Proof. Since $\Lambda_n(\Theta, \Theta_0) = 2nQ_n(\theta_0) - 2nQ_n(\hat{\theta}_n)$, we can expand this in a Taylor series around $\hat{\theta}_n$, obtaining

$$2n \frac{\partial Q_n(\hat{\theta}_n)}{\partial \theta}^T (\theta_0 - \hat{\theta}_n) + n(\theta_0 - \hat{\theta}_n)^T \frac{\partial^2 Q(\overline{\theta}_n)}{\partial \theta \partial \theta^T} (\theta_0 - \hat{\theta}_n) \tag{4.24}$$

for some vector $\overline{\theta}_n$ on the line segment between $\hat{\theta}_n$and θ_0.

As in the proof of the previous theorem, we show that $\overline{A}_n$ converges to $i(\theta_0)$ in probability. Thus, by Slutsky's lemma and consistency, $\sqrt{n}(\hat{\theta}_n - \theta_0)^T(\overline{A}_n - i(\theta_0))$ converges to zero in distribution and probability (as the limit is constant), so we can repeat the argument and obtain

$$\sqrt{n}(\hat{\theta}_n - \theta_0)^T(\overline{A}_n - i(\theta_0))\sqrt{n}(\hat{\theta}_n - \theta_0) \to^{p_{\theta_0}} 0 \tag{4.25}$$

and so $\Lambda_n(\Theta, \Theta_0)$ has the same limit distribution as the random variable

$$\sqrt{n}(\hat{\theta}_n - \theta_0)^T i(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0). \tag{4.26}$$

By continuity of $x \mapsto x^T i(\theta_0 x)$, we obtain the limiting distribution is $X^T i(\theta_0 X)$, with $X \sim N(0, i^{-1})(\theta_0)$, which is the squared Euclidean norm of the MVN $N(0, I_p)$, which has a χ^2_p distribution. □

4.1. Local Asymptotic Normality and Contiguity.

Definition (Local Asymptotic Normality)**.** *Consider a parametric model* $f(\theta) \equiv f(\theta, \cdot), \theta \in \Theta \subset \mathbb{R}^p$ *and let* $q(\theta, y) = -\log f(\theta, y)$*. Suppose* $\frac{\partial}{\partial \theta} q(\theta_0, y)$ *and the Fisher information* $i(\theta_0)$ *exist at the interior point* $\theta_0 \in \Theta$*. We say that the model* $\{f(\theta) : \theta \in \Theta\}$ *is locally asymptotically normal at* θ_0 *if for every convergent sequence* $h_n \to h$ *and for* $Y_1, \dots, Y_n$ IID $\sim f(\theta_0)$*, we have, as* $n \to \infty$*,*

$$\log \prod_{i=1}^n \frac{f(\theta_0 + \frac{h_n}{\sqrt{n}})}{f(\theta_0)}(Y_i) = -\frac{1}{\sqrt{n}} \sum_{i=1}^n h^T \frac{\partial q(\theta_0, Y_i)}{\partial \theta} - \frac{1}{2} h^T i(\theta_0) h + o_{P_{\theta_0}}(1). \tag{4.27}$$

We say that the model $\{f(\theta) : \theta \in \Theta\}$ *is locally asymptotically normal if it is locally asymptotically normal for every* $\theta \in \operatorname{int} \Theta$*.*

Theorem. Consider a parametric model $\{f(\theta), \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^p$, that satisfies the assumptions of the theorem on the asymptotic normality of the MLE. Then $\{f(\theta) : \theta \in \Theta_0\}$ is locally asymptotically normal for every open subset Θ_0 of Θ.

Proof. We prove for $n_n = h$ fixed. As before, we can expand $\log f(\theta_0 + \frac{h}{\sqrt{n}})$ about $\log f(\theta_0)$ up to second order, and obtain

$$-\frac{1}{\sqrt{n}} \sum_{i=1}^n h^T \frac{\partial q(\theta_0, Y_i)}{\partial \theta} - \frac{1}{2n} h^T \sum_{i=1}^n \frac{\partial^2 q(\overline{\theta}_n, Y_i)}{\partial \theta \partial \theta^T} h \tag{4.28}$$

for some vector $\overline{\theta}_n$ on the line segment between θ_0 and $\theta_0 + \frac{h}{\sqrt{n}}$. By the uniform law of large numbers, we have

$$\frac{1}{2n} h^T \sum_{i=1}^n \frac{\partial^2 q(\overline{\theta}_n, Y_i)}{\partial \theta \partial \theta^T} h - \frac{1}{2} h^T i(\theta_0 h) \to^{p_{\theta_0}} 0 \tag{4.29}$$

as $n \to \infty$. □

Definition (Contiguity)**.** *Let* P_n*,* Q_n *be two sequences of probability measures. We say that* Q_n *is contiguous with respect to* $P - n$ *if for every sequence of measurable sets* A_n*, the hypothesis* $P_n(A_n \to 0)$ *as* $n \to \infty$ *implies* $Q_n(A_n) \to 0$ *as* $n \to \infty$*, and write* $Q_n \triangleleft P - n$*. The sequences are mutually contiguous if both* $Q_n \triangleleft P_n$ *and* $P_n \triangleleft Q_n$*, and write* $P_n \triangleleft\triangleright Q_n$*.*

Theorem (LeCam's First Lemma)**.** Let $P - n$, Q_n be probability measures on measurable spaces $(\Omega_n, \mathcal{A}_n)$. Then the following are equivalent:

- (i) $Q_n \triangleleft P_n$.
- (ii) If $\frac{dP_n}{dQ_n} \to^d_{Q_n} U$ along a subsequence of n, then $P(U > 0) = 1$.
- (iii) If $\frac{dQ_n}{dP_n} \to^d_{P_n} V$ along a subsequence of n, then $\mathbb{E}(V) = 1$.
- (iv) For any sequence of statistics (measurable functions $T_n : \Omega_n \to \mathbb{R}^k$), we have $T_n \to^{P_n} 0$ as $n \to \infty$ implies $T_n \to^{Q_n} 0$ as $n \to \infty$.

Proof. (i) $\iff$ (iv): follows by taking $A_n = \{\|T_n\| > \epsilon\}$, so $Q_n(A_n) \to 0$ implies $T_n \to^{Q_n} 0$. Conversely, take $T_n = 1_{A_n}$. (i) $\Rightarrow$ (ii) □ Con

Theorem. Let P_n, Q_n be sequences of probability measures on measurable spaces $(\Omega_n, \mathcal{A}_n)$ such that $\frac{dP_n}{dQ_n} \to^d_{Q_n} e^X$ where $X \sim N(-\frac{1}{2}\sigma^2, \sigma^2)$, for some $\sigma^2 > 0$ as $n \to \infty$. Then $P_n \triangleleft\triangleright Q_n$.

Proof. Since $P(e^X > 0) = 1$, so $Q_n \triangleleft P_n$ from (ii) and since $\mathbb{E}\left(e^{N(\mu, \sigma^2)}\right) = 1 \iff \mu = -\frac{\sigma^2}{2}$, this follows from (iii). □

Theorem. If $\{f(\theta) : \theta \in \Theta\}$ is locally asymptotically normal and if $h_n \to h \in \mathbb{R}^p$, then the product measures $P^n_{\theta + \frac{h_n}{\sqrt{n}}}$ and P^n_θ corresponding to samples $X_1, \dots, X_n$ from densities $f(\theta + \frac{h_n}{\sqrt{n}})$ and $f(\theta)$, respectively, are mutually contiguous. In particular, if a statistic $T(Y_1, \dots, Y_n)$ converges to zero in probability under P^n_θ then it also converges to zero in $P^n_{\theta + \frac{h_n}{\sqrt{n}}}$-probability.

Proof. This follows from the fact that the asymptotic expansion converges to $N(-\frac{h^T i(\theta) h}{2}, h^T i(\theta) h)$ under P_θ. Then we can apply (iv). □

4.2. Bayesian Inference.

Definition. $\|P - Q\|_{TV} = \sup_{B \in \mathcal{B}(\mathbb{R}^p)} |P(B) - Q(B)|$ *is the total variation distance on the set of probability measures on the Borel* σ*-algebra* $\mathcal{B}(\mathbb{R}^p)$ *of* $\mathbb{R}^p$*.*

Theorem (Bernstein-von Mises Theorem)**.** Consider a parametric model $\{f(\theta), \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^p$, that satisfies the assumptions of the theorem on the asymptotic normality of the MLE. Suppose the model admits a uniformly consistent estimator T_n. Let $X_1, \dots, X_n$ be IID from density $f(\theta_0)$, let $\hat{\theta}_n$ be the MLE based on that sample, assume the prior measure Π is defined on the Borel sets of $\mathbb{R}^p$ and that Π possesses a Lebesgue-density π that is continuous and positive in a neighborhood of θ_0. Then, if $\Pi(\cdot | X_1, \dots, X_n)$ is the posterior distribution given the sample, we have

$$\|\Pi(\cdot | X_1, \dots, X_n) - N(\hat{\theta}_n, \frac{1}{n} i^{-1}(\theta_0))\|_{TV} \to^{P_{\theta_0}} 0 \tag{4.30}$$

as $n \to \infty$.

5. High Dimensional Linear Models

Here, we consider the model $Y = X\theta = \epsilon$, $\epsilon \sim N(0, \sigma^2 I_n)$, $\theta \in \Theta = \mathbb{R}^p, \sigma^2 > 0$, where X is an $n \times p$ design matrix, and ϵ is a standard Gaussian noise vector in $\mathbb{R}^n$. Throughout, we denote the resulting $p \times p$ **Gram matrix** as $\hat{\Sigma} = \frac{1}{n} X^T X$ which is symmetric and positive semidefinite.

Write $a \lesssim b$ for $a \leq Cb$ for some fixed (ideally harmless) constant $C > 0$.

Theorem. In the case $p \leq n$, the classical least squares estimator introduced by Gauss solves the problem

$$\min_{\theta \in \mathbb{R}^p} \frac{\|Y - X\theta\|^2}{n} \tag{5.1}$$

with the solution $\hat{\theta} = (X^TX)^{-1}X^TY \sim N(0, \sigma^2(X^TX)^{-1})$ where we rely on X having full column rank so X^TX is invertible. Assuming $\frac{X^TX}{n} = I_p$, we have

$$\frac{1}{n}\mathbb{E}_\theta\|X(\hat{\theta}-\theta)\|_{2e^2} = \mathbb{E}_\theta\|\hat{\theta}-\theta\|_2^2 \frac{\sigma^2}{n}\operatorname{tr}(I_p) = \frac{\sigma^2 p}{n}. \tag{5.2}$$

Definition. *$\theta^0 \in B_0(k) \equiv \{\theta \in \mathbb{R}^p, \text{at most } k \text{ nonzero entries}\}$.*
For $\theta^0 \in B_0(k)$, call $S_0 = \{j : \theta^0_j \neq 0\}$ the ***active set*** *of θ^0.*

5.1. The LASSO.

Definition (The LASSO)**.** *The $\tilde{\theta} = \tilde{\theta}_{LASSO} = \arg\min_{\theta\in\mathbb{R}^p} \frac{\|Y-X\theta\|_2^2}{n} + \lambda\|\theta\|_1$.*

Theorem (The LASSO performs almost as well as the LS estimator)**.** Let $\theta^0 \in B_0(k)$ be a k-sparse vector in $\mathbb{R}^p$ with active set S_0. Suppose $Y = X\theta_0 + \epsilon$ where $\epsilon \sim N(0, I_n)$, and let $\tilde{\theta}$ be the LASSO estimator with penalization parameter

$$\lambda = 4\overline{\sigma}\sqrt{\frac{t^2+2\log p}{n}}, \overline{\sigma}^2 = \max_{j=1,\dots,p}\hat{\Sigma}_{jj}, \tag{5.3}$$

and assume the $n\times p$ matrix X is such that, for some $r_0 > 0$,

$$\frac{1}{n}\|\tilde{\theta}_{S_0}-\theta^0\|_1^2 \leq kr_0(\tilde{\theta}-\theta^0)^T\hat{\Sigma}(\tilde{\theta}-\theta^0) \tag{5.4}$$

on an event of probability at least $1-\beta$. Then with probability at least $1-\beta-\exp^{-\frac{t^2}{2}}$ we have

$$\frac{1}{n}\|X(\tilde{\theta}-\theta^0)\|_2^2 + \lambda\|\tilde{\theta}-\theta^0\|_1 \leq 4\lambda^2 kr_0 \lesssim \frac{k}{n}\times\log p. \tag{5.5}$$

Proof. Note that by definition we have

$$\frac{1}{n}\|Y-X\tilde{\theta}\|_2{}^2 + \lambda\|\tilde{\theta}\|_1 \leq \frac{1}{n}\|Y-X\theta^0\|_2^2 + \lambda\|\theta^0\|_1 \tag{5.6}$$

$$\frac{1}{n}\|X(\theta^0-\tilde{\theta})\|_2^2 + \lambda\|\tilde{\theta}\|_1 \leq \frac{2}{n}\epsilon^TX(\tilde{\theta}-\theta^0) + \lambda\|\theta^0\|_1. \tag{5.7}$$

by inserting the model equation.

Using the tail bound on the next theorem, we have on an event A,

$$|\frac{2\epsilon^TX(\tilde{\theta}-\theta^0)}{n}| \leq \frac{\lambda}{2}\|\tilde{\theta}-\theta^0\|_1. \tag{5.8}$$

and thus combining with the above result obtain

$$\frac{2}{n}\|X(\theta^0-\tilde{\theta})\|_2^2 + 2\lambda\|\tilde{\theta}\|_1 \leq \backslash\|\tilde{\theta}-\theta^0\|_1 + 2\lambda\|\theta^0\|_1. \tag{5.9}$$

Using $\|\tilde{\theta}\|_1 = \|\tilde{\theta}_{S_0}\|_1 + \|\tilde{\theta}_{S_0^c}\|_1 \geq \|\theta^0_{S_0}\|_1 - \|\tilde{\theta}_{S_0}-\theta^0_{S_0}\|_1 + \|\tilde{\theta}^c_{S_0}\|_1$ we obtain on this event, noting $\theta^0_{S_0^c} = 0$ by definition of S_0

$$\frac{2}{n}\|X(\theta^0-\tilde{\theta})\|_2^2 + 2\lambda\|\tilde{\theta}_{S_0^c}\|_1 \leq 3\lambda\|\tilde{\theta}_{S_0}-\theta^0_{S_0}\|_1 + \lambda\|\tilde{\theta}_{S_0^c}\|_1 \tag{5.10}$$

and so

$$\frac{2}{n}\|X(\theta^0-\tilde{\theta})\|_2^2 + \lambda\|\tilde{\theta}_{S_0^c}\|_1 \leq 3\lambda\|\tilde{\theta}_{S_0}-\theta^0_{S_0}\|_1 \tag{5.11}$$

holds on the event.

Then we have

$$\frac{2}{n}\|X(\tilde{\theta}-\theta^0)\|_2^2 + \lambda\|\tilde{\theta}-\theta^0\|_1 = \frac{2}{n}\|X(\tilde{\theta}-\theta^0)\|_2^2 + \lambda\ |\tilde{\theta}_{S_0}-\theta^0_{S_0}\|_1 + \lambda\|\tilde{\theta}_{S_0^c}\|_1 \tag{5.12}$$

$$\leq 4\lambda\|\tilde{\theta}_{S_0}-\theta^0_{S_0}\|_1 \tag{5.13}$$

$$\leq 4\lambda\sqrt{\frac{kr_0}{n}}\|X(\tilde{\theta}-\theta^0)\|_2 \tag{5.14}$$

$$\leq \frac{1}{n}\|X(\tilde{\theta}-\theta^0)\|_2^2 + 4\lambda^2 kr_0 \tag{5.15}$$

using the previous inequalities and $4ab \leq a^2+4b^2$. □

Theorem. Let $\lambda_0 = \frac{\lambda}{2}$. The for all $t > 0$,

$$\mathbb{P}\left(\max_{j=1,\dots,p}\frac{2}{n}|(\epsilon^TX)_j| \leq \lambda_0\right) \geq 1-\exp(-\frac{t^2}{2}). \tag{5.16}$$

Proof. Note that $\frac{\epsilon^TX}{\sqrt{n}}$ are $N(0,\hat{\Sigma})$ distributed. We then have the probability in questions exceeds one minus

$$\mathbb{P}\left(\max_{j=1,\dots,p}\frac{1}{\sqrt{n}}|(\epsilon^TX)| > \overline{\sigma}\sqrt{t^2+2\log p}\right) \tag{5.17}$$

$$\leq \sum_{j=1}^p \mathbb{P}\left(|Z| > \sqrt{t^2+2\log p}\right) \tag{5.18}$$

$$\leq pe^{-\frac{t^2}{2}}\exp^{-\log p} = e^{-\frac{t^2}{2}}. \tag{5.19}$$

□

5.2. Coherence Conditions for Design Matrices.

The critical condition is

$$\|\tilde{\theta}_{S_0}-\theta^0\|_1^2 \leq kr_0(\tilde{\theta}-\theta^0)^T\hat{\Sigma}(\tilde{\theta}-\theta^0) \tag{5.20}$$

holding true with high probability.

Theorem. The theorem on LASSO holds true with the crucial condition (5.20) replaced with the following condition: For S_0, the active set of $\theta^0 \in B_0(k)$, $k \leq p$, assume the $n\times p$ matrix X satisfies, for all θ in

$$\{\theta \in \mathbb{R}^p : \|\theta_{S_0^c}\|_1 \leq 3\|\theta_{S_0}-\theta^0_{S_0}\|_1\} \tag{5.21}$$

and some universal constant r_0,

$$\|\theta_{S_0}-\theta^0\|_1^2 \leq kr_0(\theta-\theta^0)^T\hat{\Sigma}(\theta-\theta^0). \tag{5.22}$$

Theorem. Let the $n\times p$ matrix X have entries $(X_{ij}) \sim^{\text{IID}} N(0,1)$ and let $\hat{\Sigma} = \frac{X^TX}{n}$. Suppose $\frac{n}{\log p} \to \infty$ as $\min(p,n) \to \infty$. Then for every $k \in \mathbb{N}$ fixed and every $0 < C < \infty$, there exists n large enough such that $\mathbb{P}\left(\theta^T\hat{\Sigma}\theta \geq \frac{1}{2}\|\theta\|_2^2 \forall\theta \in B_0(k)\right) \geq 1-2\exp(-Ck\log p)$.

Proof. For $\theta = 0$, the result is trivial. Thus, it suffices to bound

$$\mathbb{P}\left(\theta^T\hat{\Sigma}\theta \geq \frac{1}{2}\|\theta\|_2^2\forall\theta \in B_0(k)\backslash\{0\}\right) \tag{5.23}$$

$$= \mathbb{P}\left(\frac{\theta^T\hat{\Sigma}\theta}{\|\theta\|_2^2} - 1 \geq -\frac{1}{2}\forall\theta \in B_0(k)\backslash\{0\}\right) \tag{5.24}$$

$$\geq \mathbb{P}\left(\sup_{\theta\in B_0(k),\|\theta\|_2^2\neq 0}\left|\frac{\theta^T\hat{\Sigma}\theta}{\theta^T\theta} - 1\right| \leq \frac{1}{2}\right) \tag{5.25}$$

from below by $1-2\exp(-Ck\log p)$. We can then do this over each k-dimensional subspace $\mathbb{R}^p_S$ for each $S \subset \{1,\dots,p\}$ with $|S| = k$, then use

$$\mathbb{P}\left(\sup_{\theta\in B_0(k),\|\theta\|_2^2\neq 0}\left|\frac{\theta^T\hat{\Sigma}\theta}{\theta^T\theta} - 1\right| \leq \frac{1}{2}\right) \tag{5.26}$$

$$\leq \sum_{S\subseteq\{1,\dots,p\}}\mathbb{P}\left(\sup_{\theta\in\mathbb{R}^p_S,\|\theta\|_2^2\neq 0}\left|\frac{\theta^T\hat{\Sigma}\theta}{\theta^T\theta} - 1\right| \geq \frac{1}{2}\right). \tag{5.27}$$

Then we just need a bound of $2e^{-(C+1)k\log p)} = 2e^{-Ck\log p}p^{-k}$ and sum over the $\binom{p}{k} \leq p^k$ subsets.

Using the below result and taking $t = (C+1)k\log p$ is then sufficient. □

Theorem. Under the conditions of the previous theorem, we have for some universal constant $c_0 > 0$, every $S \subset \{1,\dots,p\}$ such that $|S| = k$ and every $t > 0$,

$$\mathbb{P}\left(\sup_{\theta\in\mathbb{R}^p_S,\|\theta\|_2^2\neq 0}|\frac{\theta^T\hat{\Sigma}\theta}{\theta^T\theta} - 1| \geq 18(\sqrt{\frac{t+c_0k}{n}} + \frac{t+c_0k}{n})\right) \leq 2e^{-t}. \tag{5.28}$$

Proof (Nontrivial!). Note

$$\sup_{\theta\in\mathbb{R}^p_S,\|\theta\|_2^2\neq 0}\left|\frac{\theta^T\hat{\Sigma}\theta}{\theta^T\theta} - 1\right| = \sup_{\theta\in\mathbb{R}^p_S,\|\theta\|_2^2\leq 1}|\theta^T(\hat{\Sigma}-1)\theta| \tag{5.29}$$

By compactness, we can cover the unit ball $B(S) = \{\theta \in \mathbb{R}^p_S : \|\theta\|_2 \leq 1\}$ by a net of points θ^l such that for every $\theta \in B(S)$ there exists l with $\|\theta-\theta^l\|_2 \leq \delta$. Then with $\Phi = \hat{\Sigma} - I$, we have

$$\theta^T\Phi\theta = (\theta-\theta^l)\Phi(\theta-\theta^l) + (\lambda^l)^T\Phi\theta^l + 2(\theta-\theta^l)^T\Phi\theta^l. \tag{5.30}$$

The second term is bounded by $\delta^2\sup_{v\in B(S)}|v^T\Phi v|$. The third term is bounded by $2\delta\sup_{v\in B(S)}|v^T\Phi v|$. Thus,

$$\sup_{\theta\in B(S)}|\theta^T\Phi\theta| \leq \max_{l\in 1,\dots,n(\theta)}|\theta^l\Phi\theta^l| + (\delta^2+2\delta)\sup_{v\in B(S)}|v^T\Phi v| \tag{5.31}$$

which gives the bound

$$\sup_{\theta \in B(S)} |\theta^T \Phi \theta| \leq \frac{9}{2} \max_{l=1,\dots,N(\delta)} |\theta^l \Phi \theta^l| \tag{5.32}$$

at $\delta = \frac{1}{3}$.

At $\theta^l \in B(S)$ fixed, we have

$$(\theta^l)^T \Phi \theta^l = \frac{1}{n} \sum_{i=1}^{n} ((X\theta^l)_i^2 - \mathbb{E}\Big((X\theta^l)_i^2\Big)) \tag{5.33}$$

and the random variables $(X\theta^l)_i$ are IID $N(0, \|\theta^l\|_2^2)$ distributed with variance $\|\theta^l\|_2^2 \leq 1$. Thus for $g_i \text{IID} N(0,1)$, we have

$$\mathbb{P}\left(\frac{9}{2} \max_{l=1,\dots,N(\frac{1}{3})} |(\theta^l)^T \Phi \theta^l| > 18 \left(\sqrt{\frac{t + c_0 k}{n}} + \frac{t + c_0 k}{n}\right)\right) \tag{5.34}$$

$$\leq \sum_{l=1}^{N(\frac{1}{3})} \mathbb{P}\left(|(\theta^l)^T \Phi \theta^l| > 4\|\theta^l\|_2^2 \left(\sqrt{\frac{t + c_0 k}{n}} + \frac{t + c_0 k}{n}\right)\right) \tag{5.35}$$

$$= \sum_{l=1}^{N(\frac{1}{3})} \mathbb{P}\left(\left|\sum_{i=1}^{n} (g_i^2 - 1)\right| > 4(\sqrt{n(t + c_0 k)} + t + c_0 k)\right) \tag{5.36}$$

$$\leq 2N(\frac{1}{3}) e^{-t} e^{-c_0 k} \tag{5.37}$$

$$\leq 2e^{-t} \tag{5.38}$$

where we apply the next inequality with $z = t + c_0 k$, and rely on the covering numbers of the unit ball in k-dimensional Euclidean space satisfying $N(\delta) \leq \left(\frac{A}{\delta}\right)^k$ for some universal $A > 0$. $\square$

Theorem. Let g_i,i = 1, . . . , n be IID $N(0,1)$, and set $X = \sum_{i=1}^{n} (g_i^2 - 1)$. Then for all $t \geq 0$ and $n \in \mathbb{N}$,

$$\mathbb{P}(|X| \geq t) \leq 2 \exp(-\frac{t^2}{4(n+t)}). \tag{5.39}$$

Proof. For $|\lambda| < \frac{1}{2}$, we can compute the MGF of $\mathbb{E}\Big(e^{\lambda(g^2-1)}\Big) = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} = \exp(\frac{1}{2} - \log(1-2\lambda) - 2\lambda)$.

Then taking Taylor expansions, we have

$$\frac{1}{2}[-\log(1-2\lambda) - 2\lambda] = \lambda^2(1 + \frac{2}{3} 2\lambda + \cdots + \frac{2}{k+2}(2\lambda)^k + \dots) \leq \frac{\lambda^2}{1-2\lambda} \tag{5.40}$$

and by IID, $\log \mathbb{E}\Big(e^{\lambda X}\Big) \leq \frac{n\lambda^2}{1-2\lambda}$.

Then by Markov's inequality, $\mathbb{P}(X > t) \leq \mathbb{E}\Big(e^{\lambda X - \lambda t}\Big) \leq \exp(\frac{n\lambda^2}{1-2\lambda} - \lambda t) = \exp(-\frac{t^2}{4(n+t)})$, when taking $\lambda = \frac{t}{2n+2t}$.

Then taking $t = 4(\sqrt{nz} + z)$, we obtain the required result. $\square$

References