A short note on Inference and Asymptotic Normality

John Duchi
Stanford University

Winter 2024

1 Introduction

We consider a typical supervised learning scenario, where we have data in pairs (x, y), a parameter $\theta \in \mathbb{R}^p$ of interest, and a loss

$$\ell(\theta, x, y)$$

measuring the performance of the parameter on example (x, y), where $\ell(\theta, x, y)$ is convex in θ. Typical cases include regression with the squared error, where $x \in \mathbb{R}^p$, $y \in \mathbb{R}$, and

$$\ell(\theta, x, y) = \frac{1}{2}(x^T\theta - y)^2,$$

(robust) regression with a Huber-type loss so that for some $u > 0$, we take

$$\ell(\theta, x, y) = h_u(x^T\theta - y) \;\; \text{for} \;\; h_u(t) = \begin{cases} \frac{1}{2u}t^2 & \text{if } |t| \le u \\ \frac{1}{2}|t| - \frac{u}{2} & \text{if } |t| > u, \end{cases}$$

or binary logistic regression, where $x \in \mathbb{R}^p$ and $y \in \{0, 1\}$, and for the probability model $p_\theta(y \mid x) = e^{yx^T\theta}/(1 + e^{x^T\theta})$ we have

$$\ell(\theta, x, y) = -\log p_\theta(y \mid x) = \log(1 + e^{x^T\theta}) - yx^T\theta.$$

We can abstractly let Z be any random variable or vector, and consider losses $\ell(\theta, z)$. Given such a loss and a sample $\{Z_i\}_{i=1}^n$ of size n, we can define the population and empirical losses

$$L(\theta) \coloneqq \mathbb{E}[\ell(\theta, Z)] \;\; \text{and} \;\; L_n(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^n \ell(\theta, Z_i).$$

Our main goal will be to perform inference on the parameter

$$\theta^\star \coloneqq \operatorname*{argmin}_\theta L(\theta)$$

minimizing this population loss. That is, for a given level $\alpha \in (0, 1)$, we would like to develop sample-based confidence sets $\widehat{C}_n$ such that

$$\mathbb{P}(\theta^\star \in \widehat{C}_n) \to 1 - \alpha \tag{1}$$

whenever the data Z_i are indeed drawn from the population defining the population loss L.

To develop such a result, we will require two main results:

i. Consistency of the estimated parameter

$$\widehat{\theta}_n = \operatorname*{argmin}_{\theta} L_n(\theta)$$

minimizing the empirical loss L_n, meaning that

$$\widehat{\theta}_n \to \theta^\star \text{ with probability 1.} \tag{2}$$

ii. Given consistency, the asymptotic normality result that

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \stackrel{\text{dist}}{\to} \mathsf{N}(0, \Sigma) \tag{3}$$

for a particular covariance matrix Σ we shall derive, meaning that as $n \to \infty$, if we were to draw a new sample $Z_1, \ldots, Z_n$ of n and refit $\widehat{\theta}_n$, then the distribution of the errors $\widehat{\theta}_n - \theta^\star$ would be approximately Gaussian with mean zero and covariance $\frac{1}{n}\Sigma$.[1] We write this as

$$\widehat{\theta}_n - \theta^\star \stackrel{\cdot}{\sim} \mathsf{N}\left(0, n^{-1}\Sigma\right),$$

by which we mean that for any (reasonable) set $B \subset \mathbb{R}^p$,

$$\mathbb{P}\left(\widehat{\theta}_n - \theta^\star \in B\right) \to \mathbb{P}\left(\mathsf{N}(0, n^{-1}\Sigma) \in B\right)$$

as n grows.

1.1 From convergence to inference

To move from the convergence guarantee (3) to the construction of a confidence set requires a few standard—at least to working statisticians—manipulations relating to the duality between testing and confidence sets. Recall the desidaratum (1), and let $\alpha \in (0, 1)$ be the desired level. We work in stages: first, assuming we know Σ, and then assuming we have an accurate approximation Σ_n.

Beginning with the former, suppose $W \sim \mathsf{N}(0, \Sigma)$. Then we can choose a set C such that $\mathbb{P}(W \in C) = 1 - \alpha$. The distributional convergence (3) then guarantees

$$\mathbb{P}\left(\sqrt{n}(\widehat{\theta}_n - \theta^\star) \in C\right) \to \mathbb{P}(W \in C) = 1 - \alpha,$$

and so as $\sqrt{n}(\widehat{\theta}_n - \theta^\star) \in C$ if and only if

$$\theta^\star \in \widehat{C}_n \coloneqq \widehat{\theta}_n - \frac{1}{\sqrt{n}}C = \left\{\widehat{\theta}_n - \frac{1}{\sqrt{n}}w \mid w \in C\right\},$$

we obtain

$$\mathbb{P}(\theta^\star \in \widehat{C}_n) \to 1 - \alpha.$$

[1] Distributional convergence $\stackrel{\text{dist}}{\to}$ has many equivalent definitions. In our case, the variant we use is the following: we say that random variables X_n converge in distribution to $W \sim \mathsf{N}(0, \Sigma)$ if

$$\mathbb{P}(X_n \in B) - \mathbb{P}(W \in B) \to 0$$

uniformly over all boxes $B \subset \mathbb{R}^p$. In one dimension, this means that the CDFs $F_n(t) \coloneqq \mathbb{P}(X_n \le t)$ converge to the CDF $F(t) = \mathbb{P}(W \le t)$ of $W \sim \mathsf{N}(0, \sigma^2)$.

Example 1 (Confidence ellipses)**:** As know that $\Sigma^{-1/2}W \sim \mathsf{N}(0, I_p)$, and so if $\chi^2_{p,1-\alpha}$ denotes the $1-\alpha$ quantile of a χ^2 random variable with p degrees of freedom (recall this is the random variable $\|Z\|_2^2$, where Z is standard normal $Z \sim \mathsf{N}(0, I_p)$), the ellipse

$$C := \left\{w \in \mathbb{R}^p \mid \left\|\Sigma^{-1/2}w\right\|_2^2 \le \chi^2_{p,1-\alpha}\right\} = \left\{w \in \mathbb{R}^p \mid w^T\Sigma^{-1}w \le \chi^2_{p,1-\alpha}\right\}$$

satisfies $\mathbb{P}(W \in C) = 1 - \alpha$. For this particular elliptical set, we have

$$\widehat{C}_n = \left\{\theta \mid (\theta - \widehat{\theta}_n)^T\Sigma^{-1}(\theta - \widehat{\theta}_n) \le \frac{1}{n} \cdot \chi^2_{p,1-\alpha}\right\},$$

that is, the set of parameters θ close to $\widehat{\theta}_n$ in the metric defined by the covariance Σ. As $\chi^2_{p,1-\alpha} \approx p$ for large p, the confidence set $\widehat{C}_n$ corresponds to a Euclidean ball of radius roughly $\sqrt{p/n}$ around $\widehat{\theta}_n$. $\diamond$

Of course, we do not typically have access to the "true" asymptotic covariance matrix Σ and must estimate it from data. In this case, we will typically have some Σ_n such that

$$\Sigma_n \to \Sigma \;\; \text{with probability 1}$$

as n grows. Then the asymptotic normality guarantee (3) implies that

$$\sqrt{n}\Sigma_n^{-1/2}(\widehat{\theta}_n - \theta^\star) = \sqrt{n}\Sigma^{-1/2}(\widehat{\theta}_n - \theta^\star) + \underbrace{\text{error}}_{\to 0} \stackrel{\text{dist}}{\to} \mathsf{N}(0, I_p),$$

where we have elided a detail or two but used that if $W \sim \mathsf{N}(0, \Sigma)$, then $AW \sim \mathsf{N}(0, A\Sigma A^T)$ for any matrix A, choosing $A = \Sigma^{-1/2}$. We typically refer to such results as *pivotal*, meaning that the asymptotic distribution is independent of the particulars of the sampling scheme: no matter what the data process is, the final distribution is standard normal.

Writing this differently and following the "is approximately distributed as" notational convention we introduce above, we have the following key (approximate) normality result.

Theorem 1. *Assume the classical asymptotic normality result* (3) *and that* $\Sigma_n \to \Sigma$ *in probability. Then*

$$\widehat{\theta}_n - \theta^\star \stackrel{\cdot}{\sim} \mathsf{N}\left(0, n^{-1}\Sigma_n\right). \tag{4}$$

Written differently, an equivalent statement to the result (4) is that

$$\widehat{\theta}_n \stackrel{\cdot}{\sim} \mathsf{N}\left(\theta^\star, n^{-1}\Sigma_n\right).$$

Returning to Example 1, we can give an elaborated variant.

Example 2 (Confidence ellipses continued)**:** Letting the result (4) of Theorem 1 hold, we have the *pivotal* result

$$\Sigma_n^{-1/2}(\widehat{\theta}_n - \theta^\star) \stackrel{\cdot}{\sim} \mathsf{N}(0, n^{-1}I_p).$$

Then the analogue of the confidence set from Example 1 becomes

$$\widehat{C}_n := \left\{\theta \in \mathbb{R}^p \mid (\theta - \widehat{\theta}_n)^T\Sigma_n^{-1}(\theta - \widehat{\theta}_n) \le n^{-1}\chi^2_{p,1-\alpha}\right\},$$

an ellipse based on Σ_n centered at $\widehat{\theta}_n$. We have

$$\mathbb{P}(\theta^\star \in \widehat{C}_n) \to 1 - \alpha$$

as desired. $\diamond$

The key result that we demonstrate in this document is that for empirical risk minimization problems, where $\widehat{\theta}_n = \operatorname{argmin}_\theta L_n(\theta)$, the so-called "sandwich covariance," which (probably) White [1] most saliently explores, is that

$$\Sigma_n \coloneqq \left(\nabla^2 L_n(\widehat{\theta}_n)\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \nabla\ell(\widehat{\theta}_n, Z_i)\nabla\ell(\widehat{\theta}_n, Z_i)^T\right) \left(\nabla^2 L_n(\widehat{\theta}_n)\right)^{-1}$$

is a consistent estimator of the true asymptotic covariance of $\widehat{\theta}_n - \theta^\star$, which is

$$\Sigma \coloneqq \nabla^2 L(\theta^\star)^{-1}\mathrm{Cov}(\nabla\ell(\theta^\star, Z))\nabla^2 L(\theta^\star)^{-1}.$$

In the context of Theorem 1, this then gives that

$$\widehat{\theta}_n \mathrel{\dot\sim} \mathsf{N}\left(\theta^\star, n^{-1}\Sigma_n\right). \tag{5}$$

Thus, any set $\widehat{C}_n$ containing for a draw from $\mathsf{N}(\widehat{\theta}_n, n^{-1}\Sigma_n)$ with probability $1-\alpha$ is an (asymptotic) $1-\alpha$ confidence set for $\theta^\star$. We give a formal version of this result in Corollary 2.1 to follow.

1.2 Classical maximum likelihood versus general results

The result (5) holds in fairly large generality: essentially, so long as the losses have second derivatives, it holds. Classical statistical theory (which many students encounter before these results) is frequently peppered with statements such as "the maximum likelihood estimator is asymptotically normal." Here, we contrast such results a bit with the more general result (5).

In classical theory, we assume that the data Z_i come from a statistical model P_θ with density p_θ indexed by θ, and that $\theta^\star$ uniquely maximizes

$$\mathbb{E}[\log p_\theta(Z)]$$

when $Z \sim P_{\theta^\star}$. Then with a bit of handwaving, if one lets $l_\theta(z) = \log p_\theta(z)$ be the log-likelihood, one defines the Fisher Information

$$I_\theta \coloneqq \mathbb{E}[\nabla l_\theta(Z)\nabla l_\theta(Z)^T].$$

After exchanging differentiation and integration a few times we have

$$0 = \nabla 1 = \nabla \int p_\theta(z)dz = \int \nabla p_\theta(z)dz = \int \frac{\nabla p_\theta(z)}{p_\theta(z)} p_\theta(z)dz = \mathbb{E}[\nabla \log p_\theta(Z)],$$

and similarly, because $\nabla^2 \log p_\theta(z) = \frac{\nabla^2 p_\theta(z)}{p_\theta(z)} - \frac{1}{p_\theta(z)^2}\nabla p_\theta(z)\nabla p_\theta(z)^T$,

$$\begin{aligned} 0 = \nabla^2 \int p_\theta(z)dz = \int \nabla^2 p_\theta(z)dz &= \int \frac{\nabla^2 p_\theta(z)}{p_\theta(z)} p_\theta(z)dz \\ &= \int \left[\nabla^2 \log p_\theta(z) + \nabla l_\theta(z)\nabla l_\theta(z)^T\right] p_\theta(z)dz. \end{aligned}$$

That is, $I_\theta = -\mathbb{E}[\nabla^2 l_\theta(Z)]$. Notably, under the *extraordinary* assumption that the data honestly follow the probabilistic model $Z \sim P_{\theta^\star}$, the classical asymptotic normality of the maximum likelihood estimator that

$$\sqrt{n}\left(\widehat{\theta}_n - \theta^\star\right) \stackrel{\text{dist}}{\to} \mathsf{N}\left(0, I_{\theta^\star}^{-1}\right)$$

is simply a special case of the result (5): when $\ell(\theta, z) = -\log p_\theta(z)$, then we have

$$\nabla^2 L(\theta^\star) = I_{\theta^\star} \quad \text{and} \quad \mathrm{Cov}(\nabla \ell(\theta^\star, Z)) = I_{\theta^\star}$$

and so the sandwich covariance reduces to

$$\nabla^2 L(\theta^\star)^{-1} \mathrm{Cov}(\nabla \ell(\theta^\star, Z)) \nabla^2 L(\theta^\star)^{-1} = I_{\theta^\star}^{-1}.$$

But of course, the general covariance holds generally, without making the (frankly, bonkers) assumptions of correct model specification.

Example 3 (Linear regression)**:** In linear regression, the "true" model assumption that $y_i = x_i^T \theta^\star + \varepsilon_i$ yields a classical Fisher information (conditional on X)

$$I_\theta = \frac{1}{\sigma^2} \frac{1}{n} \sum_{i=1}^n x_i x_i^T$$

when $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$. When the variance σ^2 must be estimated, the typical choice is

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n (y_i - x_i^T \widehat{\theta}_n)^2,$$

where $\widehat{\theta}_n$ is the ordinary least-squares estimator. This gives the (approximate) Fisher information $\hat{I} = \hat{\sigma}^{-2} \cdot \frac{1}{n} \sum_{i=1}^n x_i x_i^T$ and *if the linear model is true*, then

$$\widehat{\theta}_n - \theta^\star \mathrel{\dot{\sim}} \mathsf{N}\left(0, n^{-1} \hat{I}^{-1}\right).$$

In contrast, the squared error $\ell(\theta, x, y) = \frac{1}{2}(x^T\theta - y)^2$ has $\nabla \ell(\theta, x, y) = (x^T\theta - y)x$ and $\nabla^2 \ell(\theta, x, y) = xx^T$. For the least-squares estimate $\widehat{\theta}_n$, this gives the sandwich covariance

$$\Sigma_n = \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^T\right)^{-1} \left(\sum_{i=1}^n (y_i - x_i^T \widehat{\theta}_n)^2 x_i x_i^T\right) \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^T\right)^{-1}.$$

Then if $\theta^\star = \mathrm{argmin}_\theta \, \mathbb{E}[(Y - X^T\theta)^2]$ is the population minimizer, we have

$$(\widehat{\theta}_n - \theta^\star) \mathrel{\dot{\sim}} \mathsf{N}\left(0, n^{-1}\Sigma_n\right)$$

regardless of whether the standard linear model holds. $\diamond$

1.3 Inference of individual parameters

Frequently, we wish to infer individual parameters of a larger parameter vector. For example, as in Question **??** from the homework, we can consider a causal estimation problem, where in the potential outcomes framework,

$$\tau^\star = \mathbb{E}[Y(1) - Y(0)]$$

denotes the (average) treatment effect of a treatment W on the response Y, and the vector $\theta = (\alpha, \tau, \beta)$ minimizes

$$\mathbb{E}[(Y - \alpha - X^T\beta - \tau W)^2].$$

Then we wish to infer $e_2^T\theta^\star = \theta_2^\star = \tau^\star$, the second coordinate of the vector $\theta^\star$. Recognizing that for a vector $v \in \mathbb{R}^p$ and $Z \sim \mathsf{N}(0, \Sigma)$, we have

$$v^T Z \sim \mathsf{N}(0, v^T\Sigma v),$$

we can fairly immediately develop confidence sets for linear functions of $\theta^\star$.

Corollary 1.1. *Assume the conditions of Theorem 1. Then for any vector $v \in \mathbb{R}^p$,*

$$v^T\widehat{\theta}_n \stackrel{\cdot}{\sim} \mathsf{N}\left(v^T\theta^\star, n^{-1}v^T\Sigma_n v\right).$$

Leveraging Corollary 1.1, we define $\sigma_n^2 = \frac{1}{n}v^T\Sigma_n v$ and

$$v^T(\widehat{\theta}_n - \theta^\star) \stackrel{\cdot}{\sim} \mathsf{N}(0, \sigma_n^2) \;\; \text{i.e.} \;\; \frac{1}{\sigma_n}v^T(\widehat{\theta}_n - \theta^\star) \stackrel{\cdot}{\sim} \mathsf{N}(0, 1).$$

Let z_α be the α-quantile of a standard normal (i.e., a z-score), so that $\mathbb{P}(Z \le z_\alpha) = \alpha$ for $Z \sim \mathsf{N}(0, 1)$. Then we see that a classical normal confidence set

$$\widehat{C}_n := [v^T\widehat{\theta}_n - z_{1-\alpha/2}\sigma_n, v^T\widehat{\theta}_n + z_{1-\alpha/2}\sigma_n]$$

for $v^T\theta^\star$ satisfies

$$\mathbb{P}\left(v^T\theta^\star \in \widehat{C}_n\right) \to \mathbb{P}\left(-z_{1-\alpha/2} \le \mathsf{N}(0, 1) \le z_{1-\alpha/2}\right) = 1 - \alpha.$$

Example 4 (Inferring a coordinate)**:** Because we have

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1p} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2p} \\ & & \ddots & \\ \Sigma_{p1} & \Sigma_{p2} & \cdots & \Sigma_{pp} \end{bmatrix},$$

if we are curious about a single coordinate j, we have

$$[\widehat{\theta}_n - \theta^\star]_j \stackrel{\cdot}{\sim} \mathsf{N}\left(0, n^{-1}\Sigma_{jj}\right),$$

which yields the confidence set

$$\theta_j^\star \in \widehat{C}_n := \left[(\widehat{\theta}_n)_j - \sqrt{\Sigma_{jj}/n}z_{1-\alpha/2}, (\widehat{\theta}_n)_j + \sqrt{\Sigma_{jj}/n}z_{1-\alpha/2}\right].$$

We have $\mathbb{P}(\theta_j^\star \in \widehat{C}_n) \to 1 - \alpha$. $\diamond$

2 The convergence results

The main convergence results we develop reflect the two steps we outline above: consistency and then the actual asymptotic normality result. We will be a bit fast and loose developing each of the results, though we will give the appropriate sufficient conditions and a few references for ways to make these rigorous. A key will be that have *convex* losses ℓ, meaning that for each z, the function $\ell(\theta, z)$ is convex in θ. As our losses will be twice continuously differentiable, this will be equivalent to the condition that the Hessians are positive semidefinite,

$$\nabla^2 \ell(\theta, z) \succeq 0,$$

for each z and each θ. Note, of course, that many losses are convex and even infinitely differentiable:

1. The squared error $\ell(\theta, x, y) = \frac{1}{2}(\theta^T x - y)^2$ satisfies $\nabla \ell(\theta, x, y) = (\theta^T x - y)x$, $\nabla^2 \ell(\theta, x, y) = xx^T \succeq 0$, and $\nabla^k \ell = 0$ for all $k > 2$.

2. The logistic loss for $y \in \{-1, 1\}$, $\ell(\theta, x, y) = \log(1 + \exp(-yx^T\theta))$, satisfies

$$\nabla \ell(\theta, x, y) = \frac{-yx}{1 + e^{yx^T\theta}} \quad \text{and} \quad \nabla^2 \ell(\theta, x, y) = \frac{1}{1 + e^{yx^T\theta}} \frac{1}{1 + e^{-yx^T\theta}} xx^T \succeq 0,$$

   and is infinitely differentiable.

3. The robust regression type loss, which approximates the Huber loss or the absolute error,

$$\ell(\theta, x, y) = \log(1 + \exp(\theta^T x - y)) + \log(1 + \exp(y - x^T\theta))$$

   that is, $\ell(\theta, x, y) = h(y - \theta^T x)$ for $h(t) = \log(1 + e^t) + \log(1 + e^{-t})$, satisfies

$$\nabla \ell(\theta, x, y) = h'(y - \theta^T x)x \quad \text{and} \quad \nabla^2 \ell(\theta, x, y) = h''(y - \theta^T x)xx^T \succeq 0,$$

   where $h'(t) = \frac{e^t}{1+e^t} - \frac{1}{1+e^t} = \frac{e^t - 1}{e^t + 1}$ and $h''(t) = \frac{e^t}{1+e^t} - \frac{e^{2t}-1}{(e^t+1)^2} = \frac{1}{e^t+1} > 0$.

As sums and expectations of convex functions are convex, we certainly have the convexity of L and L_n. To state the results, we will make a few standard (classical) assumptions:

Assumption A1. *The losses $\ell(\theta, z)$ satisfy the following.*

i. They are convex in θ, and for $\theta^\star = \operatorname{argmin}_\theta L(\theta)$, the Hessian $\nabla^2 L(\theta^\star) \succ 0$ is positive definite.

ii. They are twice continuously differentiable in θ, and there exists a Lipschitz constant $M(z)$ such that for all θ, θ' near $\theta^\star$,

$$\left\|\nabla \ell(\theta, z) - \nabla^2 \ell(\theta', z)\right\|_2 \le M(z) \left\|\theta - \theta'\right\|_2 \quad \text{and}$$
$$\left\|\left\|\nabla^2 \ell(\theta, z) - \nabla^2 \ell(\theta', z)\right\|\right\|_{\text{op}} \le M(z) \left\|\theta - \theta'\right\|_2$$

for which $\mathbb{E}[M(Z)] < \infty$.

iii. The expected squared norms $\mathbb{E}[\|\nabla \ell(\theta^\star, Z)\|^2]$ and $\mathbb{E}[\left\|\left\|\nabla^2 \ell(\theta^\star, Z)\right\|\right\|_{\text{op}}^2]$ are finite.

2.1 Consistency

The key to the consistency result that

$$\widehat{\theta}_n \to \theta^\star \quad \text{with probability 1}$$

is that the losses are convex and have continuous second derivatives. This means that L_n has (with high probability) some positive curvature around $\theta^\star$, and this growth dominates fluctuations in the gradients $\nabla L_n(\theta)$.

JCD Comment: TODO: Draw a picture of the proof style here, with tilts of gradients. (As in lecture)

Proposition 1. *Let Assumption A1 hold. Then*

$$\widehat{\theta}_n \to \theta^\star \quad \textit{with probability 1.}$$

Even more,

$$\limsup_n \sqrt{\frac{n}{\log\log n}} \|\widehat{\theta}_n - \theta^\star\| < \infty \quad \textit{with probability 1,}$$

and for any $\epsilon > 0$, there exists $K = K(\epsilon)$ such that $\limsup_n \mathbb{P}(\sqrt{n}\|\widehat{\theta}_n - \theta^\star\|_2 \geq K) \leq \epsilon$.

We give a heuristic argument here, as it is the intuition that is important for the result; Appendix A.1 provides a rigorous proof. The key idea is that because $\nabla^2 L(\theta^\star) \succ 0$ is strictly positive definite, the empirical loss L_n must have positive (upward) curvature in some neighborhood of $\theta^\star$ as well. Then, because

$$\nabla L_n(\theta^\star) = \frac{1}{n}\sum_{i=1}^n \nabla \ell(\theta^\star, Z_i)$$

is mean-zero and has variance

$$\mathrm{Var}(\nabla L_n(\theta^\star)) := \mathbb{E}\left[\|\nabla L_n(\theta^\star)\|_2^2\right] = \frac{1}{n}\mathbb{E}[\|\nabla \ell(\theta^\star, Z)\|_2^2] = O(1/n),$$

the upward curvature of L_n near $\theta^\star$ will dominate the first-order terms.

Said differently, let us assume (for the sake of contradiction) that the minimizer $\widehat{\theta}_n$ lies outside of some ball B of radius $\epsilon > 0$ around $\theta^\star$. For θ in this ball,

$$L_n(\theta) = L_n(\theta^\star) + \nabla L_n(\theta^\star)^T(\theta - \theta^\star) + \frac{1}{2}(\theta - \theta^\star)^T \nabla^2 L_n(\widetilde{\theta})(\theta - \theta^\star),$$

where $\widetilde{\theta} \in [\theta, \theta^\star]$ lies between θ and $\theta^\star$. Now, we know that $\mathbb{E}[\|\nabla L_n(\theta^\star)\|_2^2] = O(1/n)$, and so by the Cauchy-Schwarz inequality, we have

$$\begin{aligned} L_n(\theta) &\geq L_n(\theta^\star) - \|\nabla L_n(\theta^\star)\|_2 \|\theta - \theta^\star\|_2 + \frac{1}{2}(\theta - \theta^\star)^T \nabla^2 L_n(\widetilde{\theta})(\theta - \theta^\star) \\ &\geq L_n(\theta^\star) - O(1/\sqrt{n}) \cdot \|\theta - \theta^\star\|_2 + \frac{1}{2}(\theta - \theta^\star)^T \nabla^2 L_n(\widetilde{\theta})(\theta - \theta^\star). \end{aligned}$$

Now we note that for $\widetilde{\theta}$ near enough $\theta^\star$, the smoothness of the Hessians $\nabla^2 \ell(\theta, z)$ gives

$$\nabla^2 L_n(\widetilde{\theta}) = \nabla^2 L_n(\theta^\star) + \left(\nabla^2 L_n(\widetilde{\theta}) - \nabla^2 L_n(\theta^\star)\right) \succeq \nabla^2 L_n(\theta^\star) - \left(\frac{1}{n}\sum_{i=1}^n M(Z_i)\epsilon\right) I_p,$$

because $\|\widetilde{\theta} - \theta^\star\|_2 \le \epsilon$. The first of these terms converges to $\nabla^2 L(\theta^\star)$, and the second to $\mathbb{E}[M(Z)]$ by the strong law of large numbers; as long as ϵ is small enough that $\lambda_{\min}(\nabla^2 L(\theta^\star)) > 2\epsilon$, we then obtain that for large n

$$L_n(\theta) \ge L_n(\theta^\star) - O(1/\sqrt{n}) \|\theta - \theta^\star\|_2 + \frac{1}{4}\lambda_{\min}(\nabla^2 L(\theta^\star)) \|\theta - \theta^\star\|_2^2 .$$

Immediately, we see that if $\|\theta - \theta^\star\|_2 \gg \frac{1}{\sqrt{n}}$, the sum of the right two terms is positive, and so

$$L_n(\theta) > L_n(\theta^\star).$$

That is, *no* θ with $\|\theta - \theta^\star\|_2 \gg 1/\sqrt{n}$ could minimize L_n.

2.2 Asymptotic Normality

Given consistency in the form of Proposition 1, we can now provide a proof that $\widehat{\theta}_n$ is indeed asymptotically normal and give a few different constructions of its (asymptotic) covariance matrix.

Proposition 2. *Let Assumption A1 hold. Then there exists a (random) remainder R_n such that*

$$\widehat{\theta}_n - \theta^\star = -\nabla^2 L(\theta^\star)^{-1} \nabla L_n(\theta^\star) + R_n,$$

where for any $\epsilon > 0$, $n^{1-\epsilon} R_n \to 0$ with probability 1. In particular,

$$\sqrt{n}\left(\widehat{\theta}_n - \theta^\star\right) \stackrel{\text{dist}}{\to} \mathsf{N}\left(0, \nabla^2 L(\theta^\star)^{-1} \mathrm{Cov}(\nabla \ell(\theta^\star, Z)) \nabla^2 L(\theta^\star)^{-1}\right).$$

Before giving a heuristic argument to justify Proposition 2, let us provide alternative versions of the result (we provide proofs of the proposition and these alternatives in Appendix A.2). To actually perform inference on $\theta^\star$, it is essential to estimate the actual covariance of the data. Given the first statement of the proposition and the central limit theorem, because $\nabla L_n(\theta^\star) \mathrel{\dot\sim} \mathsf{N}(0, \mathrm{Cov}(\nabla \ell(\theta^\star, Z)))$, it is natural to use a *plug-in* estimator for the covariance. Thus, we define

$$\widehat{\mathrm{Cov}}(\nabla \ell) := \frac{1}{n} \sum_{i=1}^n \nabla \ell(\widehat{\theta}_n, Z_i) \nabla \ell(\widehat{\theta}_n, Z_i)^T.$$

Because $\widehat{\theta}_n - \theta^\star \to 0$ with probability 1, this (as we show) satisfies $\widehat{\mathrm{Cov}} \to \mathrm{Cov}(\nabla \ell(\theta^\star, Z))$, and using the empirical estimate $\nabla^2 L_n(\widehat{\theta}_n)$ for $\nabla^2 L(\theta^\star)$ then gives the natural covariance estimate

$$\Sigma_n := (\nabla^2 L_n(\widehat{\theta}_n))^{-1} \widehat{\mathrm{Cov}}(\nabla \ell) (\nabla^2 L_n(\widehat{\theta}_n))^{-1}. \tag{6}$$

Frequently, we call the covariance (6) the *sandwich covariance*, because the covariance of the gradients of the losses $\mathrm{Cov}(\nabla \ell)$ is sandwiched between Hessian estimates. We record the consequences of using Σ_n as a corollary.

Corollary 2.1. *Let Assumption A1 hold. Then*

$$\widehat{\theta}_n - \theta^\star \mathrel{\dot\sim} \mathsf{N}\left(0, n^{-1}\Sigma_n\right),$$

that is,

$$\sqrt{n}\Sigma_n^{-1/2}\left(\widehat{\theta}_n - \theta^\star\right) \stackrel{\text{dist}}{\to} \mathsf{N}(0, I_p).$$

The basic idea is to proceed via a Taylor expansion, then ignore higher-order error terms. Because $\widehat{\theta}_n$ minimizes L_n, we necessarily have $0 = \nabla L_n(\widehat{\theta}_n)$. Now note that when $\widehat{\theta}_n$ is close to $\theta^\star$ (as the consistency $\widehat{\theta}_n \to \theta^\star$ guarantees), we can Taylor expand ∇L_n around $\theta^\star$ to obtain

$$0 = \nabla L_n(\widehat{\theta}_n) = \nabla L_n(\theta^\star) + (\nabla^2 L_n(\theta^\star) + E_n)(\widehat{\theta}_n - \theta^\star),$$

where E_n is some error matrix. Under Assumption A1, however, we are guaranteed that E_n is small (in fact, even more: we have $\left\|\!\left|E_n\right|\!\right\|_{\rm op} \le \frac{1}{n}\sum_{i=1}^n M(Z_i)\|\widehat{\theta}_n - \theta^\star\|_2$), and so because $\nabla^2 L_n(\theta^\star) \to \nabla^2 L(\theta^\star)$ with probability 1 by the strong law of large numbers, we have

$$\nabla^2 L_n(\theta^\star) + E_n \to \nabla^2 L(\theta^\star)$$

with probability 1, and $\nabla^2 L_n(\theta^\star) + E_n$ is positive definite for large n. We can thus rearrange the Taylor expansion to write

$$\widehat{\theta}_n - \theta^\star = -\left(\nabla^2 L_n(\theta^\star) + E_n\right)^{-1} \nabla L_n(\theta^\star). \tag{7}$$

Equality (7) leads to the desired normality results. First, $\nabla L_n(\theta^\star) = \frac{1}{n}\sum_{i=1}^n \nabla \ell(\theta^\star, Z_i)$ is an i.i.d. sum of mean-zero vectors, because $\nabla L(\theta^\star) = 0 = \mathbb{E}[\nabla \ell(\theta^\star, Z)]$. Then by the central limit theorem, we have

$$\nabla L_n(\theta^\star) \stackrel{\cdot}{\sim} \mathsf{N}\left(0, n^{-1}\mathrm{Cov}(\nabla \ell(\theta^\star, Z))\right).$$

Ignoring the higher-order error terms in the expansion (7) (which we can justify), because $\nabla^2 L_n(\theta^\star) + E_n \to \nabla^2 L(\theta^\star)$, we therefore have

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta}_n - \theta^\star\right) &= -\nabla^2 L(\theta^\star)^{-1} \nabla L_n(\theta^\star) + \text{negligible error} \\ &\stackrel{\text{dist}}{\to} \mathsf{N}\left(0, \nabla^2 L(\theta^\star)^{-1}\mathrm{Cov}(\nabla \ell(\theta^\star, Z))\nabla^2 L(\theta^\star)^{-1}\right),\end{aligned}$$

because if $W \sim \mathsf{N}(0, \Sigma)$ then $HW \sim \mathsf{N}(0, H\Sigma H^T)$, and $\nabla^2 L(\theta^\star) \succ 0$ is symmetric. This is Proposition 2.

References

[1] H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

A Formal proofs

A.1 Proof of Proposition 1

We show that any point θ that is far from $\theta^\star$ can (eventually) not be a minimizer of L_n. The key is that because ℓ is convex in its first argument, L_n is as well, and if we can demonstrate the existence of *any* radius $\varepsilon > 0$ for which $L_n(\theta) > L_n(\theta^\star)$ for all θ satisfying $\|\theta - \theta^\star\| = \varepsilon$, then any θ' farther from $\theta^\star$ cannot minimize L_n (and even more, is at least linearly larger than $L_n(\theta^\star)$). The next lemma makes this formal.

Lemma A.1. *Let $f : \mathbb{R}^p \to \mathbb{R}$ be convex and $\theta_0 \in \operatorname{dom} f$, and let $\varepsilon > 0$. Define $\Delta(\varepsilon) = \inf\{f(\theta) - f(\theta_0) \mid \|\theta - \theta_0\| = \varepsilon\}$. Then for any θ such that $\|\theta - \theta_0\| \geq \varepsilon$,*

$$f(\theta) \geq f(\theta_0) + \Delta(\varepsilon) \|\theta - \theta_0\| .$$

Proof We first fix two points $\theta_0, \theta_1 \in \operatorname{dom} f$, and consider the ray of points $\theta_t = \theta_0 + t(\theta_1 - \theta_0)$ for $t \geq 0$. If $t \geq 1$, then $\theta_t = \theta_0 + t(\theta_1 - \theta_0)$ if and only if $\theta_1 = \frac{1}{t}\theta_t + \frac{t-1}{t}\theta_0$, so that

$$f(\theta_1) \leq \frac{1}{t} f(\theta_t) + \frac{t-1}{t} f(\theta_0) \quad \text{or} \quad f(\theta_t) \geq t f(\theta_1) + (1-t) f(\theta_0) = f(\theta_0) + t(f(\theta_1) - f(\theta_0)).$$

Now, assume that for some $\varepsilon > 0$, we have $f(\theta_1) > f(\theta_0)$ for all $\theta_1 \in \theta_0 + \varepsilon \mathbb{S}^{n-1}$. Then any θ for which $\|\theta - \theta_0\| > \varepsilon$ satisfies $\theta = \theta_0 + t\frac{\varepsilon}{\|\theta - \theta_0\|}(\theta - \theta_0)$ for $t = \frac{\|\theta - \theta_0\|}{\varepsilon}$, that is, $\theta = \theta_0 + t(\theta_1 - \theta_0)$ for $\theta_1 = \theta_0 + \frac{\varepsilon}{\|\theta - \theta_0\|}(\theta - \theta_0)$, and so

$$f(\theta) \geq f(\theta_0) + t(f(\theta_1) - f(\theta_0)) \geq f(\theta_0) + \frac{\|\theta - \theta_0\|}{\varepsilon} \inf_{\|\theta_1 - \theta_0\| = \varepsilon} (f(\theta_1) - f(\theta_0)) ,$$

as desired. □

Let $\lambda = \lambda_{\min}(\nabla^2 L(\theta^\star))$ for shorthand, and let $\overline{M} = \mathbb{E}[M(Z)]$, where recall that $M(z)$ is the Lipschitz constant for the Hessian $\nabla^2 \ell(\theta, z)$ in θ. With Lemma A.1 in hand, let $0 < \varepsilon \leq \frac{\lambda}{4\overline{M}}$. By a Taylor expansion, we know that for any θ in an ε-ball around $\theta^\star$, we have

$$L_n(\theta) = L_n(\theta^\star) + \nabla L_n(\theta^\star)^T(\theta - \theta^\star) + \frac{1}{2}(\theta - \theta^\star)^T \nabla^2 L_n(\widetilde{\theta})(\theta - \theta^\star),$$

where $\widetilde{\theta} \in [\theta, \theta^\star]$. Using the $M(z)$-Lipschitz continuity of $\nabla^2 \ell(\theta, z)$ in θ, for $\overline{M}_n = \frac{1}{n}\sum_{i=1}^n M(Z_i)$ we have

$$\nabla^2 L_n(\widetilde{\theta}) \succeq \nabla^2 L_n(\theta^\star) - \overline{M}_n \big\|\widetilde{\theta} - \theta^\star\big\|_2 I_p \succeq \nabla^2 L_n(\theta^\star) - \overline{M}_n \varepsilon I_p.$$

By the strong law of large numbers, we have $\overline{M}_n \overset{a.s.}{\to} \mathbb{E}[M(Z)]$ and so eventually (with probability 1, for all large enough n) $\overline{M}_n \leq 2\overline{M}$ and $\varepsilon \overline{M}_n \leq \frac{\lambda}{2}$. In particular, there exists a (random but finite) sample size N such that for $n \geq N$, all θ satisfying $\|\theta - \theta^\star\|_2 \leq \varepsilon$ have

$$\nabla^2 L_n(\theta) \succeq \frac{\lambda}{2} I_p.$$

Returning to the Taylor expansion, we obtain that there exists a random $N < \infty$ such that $n \geq N$ implies

$$L_n(\theta) = L_n(\theta^\star) + \nabla L_n(\theta^\star)^T(\theta - \theta^\star) + \frac{\lambda}{4}\|\theta - \theta^\star\|_2^2$$

for all $\|\theta - \theta^\star\|_2 \leq \varepsilon$. Now we apply the strong law of large numbers again. We have $\nabla L_n(\theta^\star) \overset{a.s.}{\to} 0$, and so for any $\delta > 0$, there exists (a random) $N' < \infty$ such that $\|L_n(\theta^\star)\|_2 \leq \delta$ for $n \geq N'$. Applying Cauchy-Schwarz, we have

$$\begin{aligned} L_n(\theta) &\geq L_n(\theta^\star) - \|\nabla L_n(\theta^\star)\|_2 \|\theta - \theta^\star\|_2 + \frac{\lambda}{4}\|\theta - \theta^\star\|_2^2 \\ &\geq L_n(\theta^\star) - \delta \|\theta - \theta^\star\|_2 + \frac{\lambda}{4}\|\theta - \theta^\star\|_2^2 \end{aligned}$$

whenever $\|\theta - \theta^\star\|_2 \le \varepsilon$. But of course $\frac{\lambda}{4}t^2 - \delta t > 0$ whenever $t > \frac{4\delta}{\lambda}$; if we take δ such that $\frac{4\delta}{\lambda} < \varepsilon$ then for the shell $S = \{\theta \mid \frac{4\delta}{\lambda} < \|\theta - \theta^\star\|_2 \le \varepsilon\}$, we have $L_n(\theta) > L_n(\theta^\star)$ for all $\theta \in S$. Apply Lemma A.1 to see that we must therefore have

$$\|\widehat{\theta}_n - \theta^\star\|_2 \le \frac{4\delta}{\lambda}.$$

This is the first part of Proposition 1, as $\delta > 0$ was arbitrary.

To obtain the sharper convergence result, we must take $\delta = \delta_n$ converging to 0. In this case, the central limit theorem implies that $\sqrt{n}L_n(\theta^\star)$ converges in distribution to a Gaussian. Even more the law of the iterated logarithm gives that

$$\limsup_n \sqrt{\frac{n}{\log\log n}} \|\nabla L_n(\theta^\star)\| < \infty$$

for any norm $\|\cdot\|$ with probability 1. Taking $\delta_n = \|\nabla L_n(\theta^\star)\|_2$, we obtain that there exists a constant $C < \infty$ such that eventually $\delta_n \le C\sqrt{\log\log n/n}$. Repeating the same argument above, *mutatis mutandis*, we have

$$L_n(\theta) \ge L_n(\theta^\star) - \delta_n \|\theta - \theta^\star\|_2 + \frac{\lambda}{4}\|\theta - \theta^\star\|_2^2$$

for all $\|\theta - \theta^\star\|_2 \le \varepsilon$. Solving gives we must have $\|\widehat{\theta}_n - \theta^\star\|_2 \le \frac{4\delta_n}{\lambda}$ eventually.

A.2 Proof of Proposition 2

By Proposition 1, we have $n^\gamma(\widehat{\theta}_n - \theta^\star) \stackrel{a.s.}{\to} 0$ for all $\gamma < \frac{1}{2}$. Thus we eventually have

$$0 = \nabla L_n(\widehat{\theta}_n) = \nabla L_n(\theta^\star) + (\nabla^2 L_n(\theta^\star) + E_n)(\widehat{\theta}_n - \theta^\star),$$

where $\|\!|E_n|\!\|_{\rm op} \le \frac{1}{n}\sum_{i=1}^n M(Z_i)\|\widehat{\theta}_n - \theta^\star\|_2$. Then because of the consistency guarantee and that $\frac{1}{n}\sum_{i=1}^n M(Z_i) \stackrel{a.s.}{\to} \mathbb{E}[M(Z)]$, we have $E_n \stackrel{a.s.}{\to} 0$, $\nabla^2 L_n(\theta^\star) \stackrel{a.s.}{\to} \nabla^2 L(\theta^\star)$, and so $\nabla^2 L_n(\theta^\star) + E_n$ is invertible eventually and

$$\widehat{\theta}_n - \theta^\star = (\nabla^2 L_n(\theta^\star) + E_n)^{-1}\nabla L_n(\theta^\star),$$

that is, expansion (7) holds.

In a standard proof of asymptotic normality, one would now invoke the central limit theorem and Slutsky's convergence theorems. We can avoid these by giving a more direct proof as well. To that end, we apply an expansion of $A \mapsto A^{-1}$. If $A \succ 0$ and $\|\!|E|\!\|_{\rm op} < \lambda_{\min}(A)$, then

$$(A + E)^{-1} = A^{-1} + A^{-1}\sum_{i=1}^\infty (-1)^i (EA^{-1})^i,$$

or $(A + E)^{-1} = A^{-1} - A^{-1}EA^{-1} + O(\|E\|^2)$, and because $\sum_{i=2}^\infty \delta^i = \delta^2/(1-\delta)$ for $|\delta| < 1$, we have

$$\left\|\!\left|(A + E)^{-1} - A^{-1} + A^{-1}EA^{-1}\right|\!\right\|_{\rm op} \le \|\!|A^{-1}|\!\|_{\rm op} \sum_{i=2}^\infty \|\!|EA^{-1}|\!\|_{\rm op}^i \le \frac{\|\!|A^{-1}|\!\|_{\rm op}^3 \|\!|E|\!\|_{\rm op}^2}{1 - \|\!|A^{-1}|\!\|_{\rm op}\|\!|E|\!\|_{\rm op}}.$$

Returning to the expansion (7), for any $\gamma < \frac{1}{2}$, that

$$\|\!|E_n|\!\|_{\rm op} \le \overline{M}_n \left\|\widehat{\theta}_n - \theta^\star\right\|_2$$

implies $n^\gamma \|E_n\|_{\text{op}} \overset{a.s.}{\to} 0$ by Proposition 1, and similarly, that $\mathbb{E}[\|\nabla \ell(\theta^\star, Z)\|_2^2] < \infty$ implies that $n^\gamma \|\nabla L_n(\theta^\star)\| \overset{a.s.}{\to} 0$ by the law of the iterated logarithm. The law of the iterated logarithm also gives $\nabla^2 L_n(\theta^\star) - \nabla^2 L(\theta^\star) = O(1/n^\gamma)$ for any $\gamma < \frac{1}{2}$ (with probability 1). So defining $H = \nabla^2 L(\theta^\star)$ and $H_n = \nabla^2 L_n(\theta^\star)$ as shorthand for the Hessians, we have

$$\begin{aligned}(H_n + E_n)^{-1} &= (H + (H_n + E_n - H))^{-1} \\ &= H^{-1} - H^{-1}(H_n + E_n - H)H^{-1} + O(\|H_n + E_n - H\|^2) = H^{-1} + O(n^{-\gamma})\end{aligned}$$

with probability 1, for any $\gamma < \frac{1}{2}$. That is,

$$\widehat{\theta}_n - \theta^\star = \nabla^2 L(\theta^\star)^{-1} \nabla L_n(\theta^\star) + R_n$$

where the remainder term $R_n \in \mathbb{R}^p$ satisfies $n^{2\gamma} R_n \overset{a.s.}{\to} 0$ for any $\gamma < \frac{1}{2}$. This completes the proof of Proposition 2.

Proof of Corollary 2.1

To formalize the result in Corollary 2.1, it suffices to show that

$$\Sigma_n - \Sigma \to 0 \;\text{ with probability 1,}$$

as in this case we have (with probability 1) that

$$\sqrt{n}\Sigma_n^{-1/2}\left(\widehat{\theta}_n - \theta^\star\right) = \sqrt{n}\Sigma^{-1/2}\left(\widehat{\theta}_n - \theta^\star\right) + \underbrace{\sqrt{n}(\Sigma_n^{-1/2} - \Sigma^{-1/2})\left(\widehat{\theta}_n - \theta^\star\right)}_{=o(1)}.$$

To that end, let $\gamma < \frac{1}{2}$ be otherwise arbitrary. Note that

$$\left\|\nabla^2 L_n(\widehat{\theta}_n) - \nabla^2 L_n(\theta^\star)\right\|_{\text{op}} \leq \overline{M}_n \left\|\widehat{\theta}_n - \theta^\star\right\|_2$$

and $n^\gamma \|\widehat{\theta}_n - \theta^\star\|_2 \overset{a.s.}{\to} 0$ by Proposition 2. Similarly, $n^\gamma \|\nabla^2 L_n(\theta^\star) - \nabla^2 L(\theta^\star)\|_{\text{op}} \overset{a.s.}{\to} 0$ by the law of the iterated logarithm, and so because $\overline{M}_n \overset{a.s.}{\to} \mathbb{E}[M(Z)]$, we have

$$n^\gamma \left\|\nabla^2 L_n(\widehat{\theta}_n) - \nabla^2 L(\theta^\star)\right\|_{\text{op}} \overset{a.s.}{\to} 0.$$

To control the empirical covariance term, note that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \nabla \ell(\widehat{\theta}_n, Z_i)\nabla \ell(\widehat{\theta}_n, Z_i)^T = \frac{1}{n}\sum_{i=1}^n (\nabla \ell(\theta^\star, Z_i) + e_i)(\nabla \ell(\theta^\star, Z_i) + e_i)^T \\ &= \frac{1}{n}\sum_{i=1}^n \nabla \ell(\theta^\star, Z_i)\nabla \ell(\theta^\star, Z_i)^T + \frac{1}{n}\sum_{i=1}^n (\nabla \ell(\theta^\star, Z_i)e_i^T + e_i \nabla \ell(\theta^\star, Z_i)^T) + \frac{1}{n}\sum_{i=1}^n e_i e_i^T\end{aligned}$$

where $e_i = \nabla \ell(\widehat{\theta}_n, Z_i) - \nabla \ell(\theta^\star, Z_i)$. Using Assumption A1, we have $\|\nabla \ell(\widehat{\theta}_n, z) - \nabla \ell(\theta^\star, z)\|_2 \leq M(z)\|\widehat{\theta}_n - \theta^\star\|_2$. Applying Cauchy-Schwarz, we obtain

$$\left\|\frac{1}{n}\sum_{i=1}^n \nabla \ell(\theta^\star, Z_i)e_i^T\right\|_{\text{op}} \leq \sqrt{\frac{1}{n}\sum_{i=1}^n \|\nabla \ell(\theta^\star, Z_i)\|_2^2}\sqrt{\frac{1}{n}\sum_{i=1}^n M(Z_i)^2}\left\|\widehat{\theta}_n - \theta^\star\right\|_2 \overset{a.s.}{\to} 0,$$

and a similar calculation shows $n^{-1}\sum_{i=1}^n e_i e_i^T \stackrel{a.s.}{\to} 0$. Tracking the error rates a bit more carefully and noting that $\frac{1}{n}\sum_{i=1}^n \nabla\ell(\theta^\star, Z_i)\nabla\ell(\theta^\star, Z_i)^T \stackrel{a.s.}{\to} \mathrm{Cov}(\nabla\ell(\theta^\star, Z))$, we have shown that

$$n^\gamma \left(\widehat{\mathrm{Cov}}(\nabla\ell) - \mathrm{Cov}(\nabla\ell(\theta^\star, Z))\right) \to 0 \text{ with probability 1.}$$

Because $(H, C) \mapsto H^{-1}CH^{-1}$ is continuous, this shows that

$$\Sigma_n \to \Sigma = \nabla^2 L(\theta^\star)^{-1}\mathrm{Cov}(\nabla\ell(\theta^\star, Z))\nabla^2 L(\theta^\star)^{-1}$$

with probability 1, as desired.