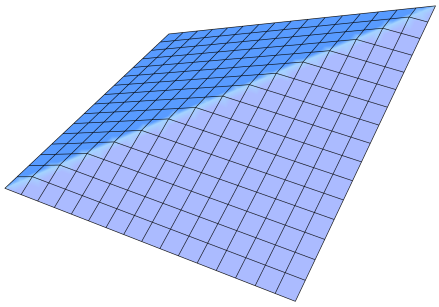


STA 3000Y: Advanced Theory of Statistics (II)

Professor Qiang Sun • Winter 2019 • University of Toronto
M 10:00AM-1:00PM in SS 581

Transcribed, LaTeXed, and edited by Rob Zimmerman

Note: These are personal lecture notes, not official course materials. They may contain errors (which by default you should attribute to me, Rob Zimmerman, rather than the course instructor). The permanent home of these — and all of my other lecture notes — is https://rob-zimmerman.github.io/coursenotes.

Contents

1 Likelihood and Classical Asymptotics **2**
Likelihood Function . . . 2
Asymptotic Statistics . . . 4
Cramér–Rao Lower Bound . . . 8
M- and Z-Estimators . . . 12
Asymptotic Normality . . . 20
Local Asymptotic Normality . . . 34
Asymptotic Cramér–Rao Lower Bound . . . 37

2 Tail Bounds and Sub-Gaussian Concentration **41**
Basic Tail Bounds . . . 41
From Markov to Chernoff . . . 42
Sub-Gaussian Random Variables . . . 44
Subexponential Random Variables and Bernstein's Bound . . . 50
Lipschitz Functions of Independent Gaussian Random Variables . . . 58

3 Measure Concentration and Empirical Processes **61**
Measure Concentration . . . 61
Separately Convex Functions and Random Variables . . . 64
Uniform Laws of Large Numbers . . . 67
Covering and Metric Entropy . . . 70
Chaining . . . 73

4 Principal Components and Sparse PCA **80**
Principal Component Analysis . . . 80
Sparse Principal Component Analysis . . . 90

5 Regression and Constrained Estimation **92**
Regression Models . . . 93
Linear Regression . . . 93
Constrained Least Squares . . . 96

Appendix: Lecture and Tutorial Index **99**

Lecture 1 — Monday, January 14, 2019

1. Likelihood and Classical Asymptotics

Likelihood Function

Let $Y_1, \ldots, Y_n$ be observations whose distribution depends on an unknown parameter θ, possibly through a parametric or semiparametric model. For example, consider the linear model

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta} + \varepsilon_i \qquad \text{and} \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

The conditional density of Y_i is then

$$\mathcal{L}_i(\boldsymbol{\theta}) \coloneqq f_{\boldsymbol{\theta}}(y_i \mid \boldsymbol{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta})^2}{2\sigma^2}\right).$$

Definition 1.1 The function $\mathcal{L}_i(\boldsymbol{\theta})$ of the unknown parameter is the **likelihood contribution** of the ith observation. If the observations are independent, their **joint likelihood** and **log likelihood** are, respectively,

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}) \qquad \text{and} \qquad \ell(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta}).$$

When $\boldsymbol{\theta} \in \mathbb{R}^d$, a **maximum likelihood estimator** is any maximizer

$$\widehat{\boldsymbol{\theta}}_n \in \operatorname*{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^d} \ell(\boldsymbol{\theta}).$$

Likelihood methods are particularly convenient when the density has a known parametric form, as in the linear model. More generally, we may model the response as $Y_i = f(X_i, \theta, \varepsilon_i)$. Important special cases include $Y_i = f(\boldsymbol{X}_i^\top \boldsymbol{\theta}, \varepsilon_i)$ and the additive model $Y_i = f(\boldsymbol{X}_i^\top \boldsymbol{\theta}) + \varepsilon_i$. Such models may be parametric or nonparametric; in the latter case, regularity assumptions such as smoothness are imposed on f.

Two broad statistical regimes will recur. In classical low dimensional asymptotics, d is fixed while n tends to infinity. Typically, $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ is asymptotically normal, which permits asymptotic inference and comparison of estimator efficiency through the limiting covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and its inverse $\mathbf{q}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$. The same ideas can sometimes accommodate a growing dimension, provided that d remains much smaller than n.

In finite sample analysis, n remains fixed and performance is expressed through probability bounds rather than limiting distributions. This viewpoint includes high dimensional problems, where d may be much larger than n and the behaviour of sample eigenvalues is often central, as well as finite population problems arising in survey sampling. The development below begins with classical asymptotics and then turns to nonasymptotic bounds.

Asymptotic Statistics

In point estimation, the objective is to construct a statistic $T = T(Y)$ that estimates a parameter of interest θ, or more generally a function $g(\theta)$. By definition, a statistic depends on the observed data but not on the unknown parameter.

Definition 1.2 A statistic $T(Y)$ is an **unbiased estimator** of $g(\theta)$ if

$$\mathbb{E}_\theta\left[T(Y)\right] = g(\theta)$$

for every θ in the parameter space.

Unbiasedness is a finite sample property rather than an asymptotic one. It is often useful, although it need not by itself identify a desirable estimator.

Example 1.3 Let $Y_1, \ldots, Y_n$ be independent and identically distributed as $\mathcal{N}(\theta, \sigma^2)$. Both Y_1 and every linear combination $\sum_{i=1}^n a_i Y_i$ whose coefficients satisfy $\sum_{i=1}^n a_i = 1$ are unbiased estimators of θ.

Solution. Linearity of expectation gives

$$\mathbb{E}_\theta\left[\sum_{i=1}^n a_i Y_i\right] = \sum_{i=1}^n a_i \mathbb{E}_\theta[Y_i] = \theta \sum_{i=1}^n a_i = \theta.$$

In particular, the sample mean

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$$

is unbiased. It improves substantially on Y_1 in variance because

$$\mathrm{Var}_\theta(\overline{Y}) = \frac{\sigma^2}{n} \qquad \text{and} \qquad \mathrm{Var}_\theta(Y_1) = \sigma^2.$$

□

This example suggests combining unbiasedness with a criterion that measures the variability of an estimator.

Definition 1.4 The **mean squared error** of an estimator $T(Y)$ of $g(\theta)$ is

$$\mathbb{E}_\theta\left[(T(Y)-g(\theta))^2\right] = \underbrace{\mathrm{Var}_\theta(T(Y))}_{\text{variance}} + \underbrace{(\mathbb{E}_\theta[T(Y)]-g(\theta))^2}_{\text{squared bias}}.$$

Thus mean squared error balances variance against squared bias.

Remark 1.5

1. Among unbiased estimators, minimizing mean squared error is equivalent to minimizing variance.
2. Without the unbiasedness restriction, an estimator that minimizes mean squared error need not be unbiased.

Definition 1.6 Let $Y_1,\dots,Y_n$ be independent and identically distributed with density f_θ. A statistic $T(Y)$ is **sufficient** for θ if the conditional distribution of $Y=(Y_1,\dots,Y_n)$ given $T(Y)$ does not depend on θ.

Theorem 1.7 (Rao–Blackwell theorem) Let $W(Y)$ be a square-integrable unbiased estimator of $g(\theta)$, let $T(Y)$ be sufficient for θ, and define

$$\phi(T) \coloneqq \mathbb{E}_\theta\left[W(Y) \mid T\right].$$

Then the following statements hold:

1. The statistic $\phi(T)$ is an unbiased estimator of $g(\theta)$.
2. For every $\theta\in\Theta$,
$$\mathrm{Var}_\theta(\phi(T)) \le \mathrm{Var}_\theta(W(Y)).$$

Proof. For (1), the tower property gives

$$\mathbb{E}_\theta[\phi(T)] = \mathbb{E}_\theta\left[\mathbb{E}_\theta[W(Y)\mid T]\right] = \mathbb{E}_\theta[W(Y)] = g(\theta).$$

For (2), the conditional variance formula yields

$$\begin{aligned}\operatorname{Var}_\theta(W(Y)) &= \operatorname{Var}_\theta\left(\mathbb{E}_\theta[W(Y) \mid T]\right) + \mathbb{E}_\theta\left[\operatorname{Var}_\theta(W(Y) \mid T)\right] \\ &= \operatorname{Var}_\theta(\phi(T)) + \underbrace{\mathbb{E}_\theta\left[\operatorname{Var}_\theta(W(Y) \mid T)\right]}_{\geq 0} \\ &\geq \operatorname{Var}_\theta(\phi(T)).\end{aligned}$$

□

Thus conditioning an unbiased estimator on a sufficient statistic never increases its variance. Completeness identifies when this improvement is the unique best unbiased estimator.

Definition 1.8 A statistic $T(Y)$ is **complete** for a family of distributions indexed by θ if, for every measurable function h,

$$\mathbb{E}_\theta[h(T)] = 0 \text{ for every } \theta \in \Theta \Longrightarrow \mathbb{P}_\theta(h(T) = 0) = 1 \text{ for every } \theta \in \Theta.$$

Theorem 1.9 (Lehmann–Scheffé theorem) If $W(Y)$ is an unbiased estimator of $g(\theta)$ and T is complete and sufficient for θ, then

$$\phi(T) \coloneqq \mathbb{E}_\theta[W(Y) \mid T]$$

is the unique uniform minimum variance unbiased estimator of $g(\theta)$, up to almost sure equality.

Remark 1.10 By the Rao–Blackwell theorem (Theorem 1.7), sufficiency permits one to improve any initial unbiased estimator. Completeness then guarantees that every such improvement produces the same function of T almost surely, so no further variance reduction within the unbiased class is possible.

Proof. Let $\widetilde{W}(Y)$ be any other unbiased estimator of $g(\theta)$ and set

$$\widetilde{\phi}(T) \coloneqq \mathbb{E}_\theta[\widetilde{W}(Y) \mid T].$$

By Theorem 1.7, both $\phi(T)$ and $\widetilde{\phi}(T)$ are unbiased, and

$$\operatorname{Var}_\theta(\widetilde{\phi}(T)) \leq \operatorname{Var}_\theta(\widetilde{W}(Y)).$$

Moreover,

$$\mathbb{E}_\theta[\phi(T) - \widetilde{\phi}(T)] = 0$$

for every θ. Completeness of T therefore implies $\phi(T) = \widetilde{\phi}(T)$ almost surely. Consequently,

$$\mathrm{Var}_\theta(\phi(T)) = \mathrm{Var}_\theta(\widetilde{\phi}(T)) \leq \mathrm{Var}_\theta(\widetilde{W}(Y)),$$

which proves both uniform minimum variance and uniqueness. □

Example 1.11 Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Poisson}(\lambda)$. The statistic

$$T = \sum_{i=1}^{n} Y_i$$

is complete and sufficient for λ. Determine the uniform minimum variance unbiased estimator of λ obtained from the initial estimator Y_1.

Solution. The estimator Y_1 is unbiased. Conditional on $T = t$, the joint Poisson mass function gives

$$\mathbb{P}_\lambda(Y_1 = k \mid T = t) = \binom{t}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{t-k}, \quad k = 0, \ldots, t.$$

Thus $Y_1 \mid T \sim \text{Bin}(T, 1/n)$ and

$$\mathbb{E}_\lambda[Y_1 \mid T] = \frac{T}{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}.$$

The Lehmann–Scheffé theorem (Theorem 1.9) shows that $\overline{Y}$ is the uniform minimum variance unbiased estimator of λ. □

Example 1.12 Let $Y \sim \text{Poisson}(\lambda)$. Find a uniform minimum variance unbiased estimator of $g(\lambda) = e^{-2\lambda}$.

Solution. The statistic $T(Y) = (-1)^Y$ is unbiased because

$$\mathbb{E}_\lambda[(-1)^Y] = e^{-\lambda} \sum_{y=0}^{\infty} \frac{(-\lambda)^y}{y!} = e^{-\lambda} e^{-\lambda} = e^{-2\lambda}.$$

The statistic Y is complete and sufficient for the Poisson family based on one observation. The Lehmann–Scheffé theorem (Theorem 1.9) therefore shows that $(-1)^Y$ is the uniform minimum variance unbiased estimator of $e^{-2\lambda}$; its value is 1 when Y is even and -1 when Y is odd. □

Cramér–Rao Lower Bound

The Lehmann–Scheffé theorem (Theorem 1.9) identifies the smallest variance attainable by an unbiased estimator when a complete sufficient statistic is available. The Cramér–Rao lower bound (Theorem 1.18) gives a general lower bound on that variance.

Definition 1.13 For a scalar parameter θ, the **score function** of a single observation Y with density f_θ is

$$\dot{\ell}_\theta(y) \coloneqq \frac{\partial}{\partial \theta} \log f_\theta(y).$$

For independent observations $Y = (Y_1, \dots, Y_n)$, the score of the full sample is

$$\dot{\ell}(Y, \theta) \coloneqq \frac{\partial}{\partial \theta} \ell(\theta) = \sum_{i=1}^{n} \dot{\ell}_\theta(Y_i).$$

Remark 1.14 When the data are clear from context, we write $\dot{\ell}(Y, \theta)$ as $\dot{\ell}(\theta)$. The additivity in Definition 1.13 follows from the additivity of the log likelihood for independent observations.

Proposition 1.15 Suppose that differentiation under the integral sign is valid and that the support of f_θ does not depend on θ. Then

$$\mathbb{E}_\theta[\dot{\ell}_\theta(Y)] = 0.$$

Proof. The assumed regularity conditions permit the derivative and integral to be interchanged. Therefore

$$\begin{aligned} \mathbb{E}_\theta[\dot{\ell}_\theta(Y)] &= \int \frac{\partial}{\partial \theta} \log f_\theta(y)\, f_\theta(y)\, \mathrm{d}y \\ &= \int \frac{\dot{f}_\theta(y)}{f_\theta(y)} f_\theta(y)\, \mathrm{d}y \\ &= \int \dot{f}_\theta(y)\, \mathrm{d}y \end{aligned}$$

$$= \frac{\partial}{\partial\theta} \int f_\theta(y)\,\mathrm{d}y$$
$$= \frac{\partial}{\partial\theta} 1 = 0.$$

□

Proposition 1.16 Under the regularity conditions of Proposition 1.15, and assuming that differentiation under the integral sign is valid a second time,

$$\mathrm{Var}_\theta(\dot{\ell}_\theta(Y)) = \mathbb{E}_\theta[\dot{\ell}_\theta(Y)^2] = -\mathbb{E}_\theta[\ddot{\ell}_\theta(Y)].$$

Proof. By Proposition 1.15, the score has mean zero, so its variance is its second moment. Moreover,

$$\ddot{\ell}_\theta(y) = \frac{\partial}{\partial\theta}\left(\frac{\dot{f}_\theta(y)}{f_\theta(y)}\right) = \frac{\ddot{f}_\theta(y)}{f_\theta(y)} - \left(\frac{\dot{f}_\theta(y)}{f_\theta(y)}\right)^2.$$

Taking expectations and interchanging differentiation and integration gives

$$\mathbb{E}_\theta[\ddot{\ell}_\theta(Y)] = \int \ddot{f}_\theta(y)\,\mathrm{d}y - \int \left(\frac{\dot{f}_\theta(y)}{f_\theta(y)}\right)^2 f_\theta(y)\,\mathrm{d}y$$
$$= \frac{\partial^2}{\partial\theta^2}\int f_\theta(y)\,\mathrm{d}y - \mathbb{E}_\theta[\dot{\ell}_\theta(Y)^2]$$
$$= -\mathbb{E}_\theta[\dot{\ell}_\theta(Y)^2].$$

□

Definition 1.17 The **Fisher information** in one observation is

$$I_1(\theta) \coloneqq \mathrm{Var}_\theta(\dot{\ell}_\theta(Y)).$$

For n independent and identically distributed observations, independence of the individual scores and Proposition 1.15 give

$$I_n(\theta) \coloneqq \mathrm{Var}_\theta(\dot{\ell}(Y,\theta)) = \sum_{i=1}^{n} \mathrm{Var}_\theta(\dot{\ell}_\theta(Y_i)) = nI_1(\theta).$$

Theorem 1.18 (Cramér–Rao lower bound) Let $T(Y)$ be a square-integrable unbiased estimator of a differentiable function $g(\theta)$. Under the regularity conditions above, suppose also that $0 < I_n(\theta) < \infty$. Then

$$\mathrm{Var}_\theta(T(Y)) \geq \frac{\dot{g}(\theta)^2}{I_n(\theta)}.$$

Proof. Differentiating the unbiasedness identity and using the same interchange of differentiation and integration as above yields

$$\dot{g}(\theta) = \frac{\partial}{\partial\theta}\mathbb{E}_\theta[T(Y)] = \mathbb{E}_\theta[T(Y)\dot{\ell}(Y,\theta)] = \mathrm{Cov}_\theta(T(Y), \dot{\ell}(Y,\theta)),$$

where the final equality follows from Proposition 1.15. The Cauchy–Schwarz inequality therefore gives

$$\dot{g}(\theta)^2 \leq \mathrm{Var}_\theta(T(Y))\,\mathrm{Var}_\theta(\dot{\ell}(Y,\theta)) = \mathrm{Var}_\theta(T(Y))I_n(\theta).$$

Rearranging proves the claim. □

In particular, when $g(\theta) = \theta$, Theorem 1.18 reduces to $\mathrm{Var}_\theta(T(Y)) \geq I_n(\theta)^{-1}$.

Example 1.19 Show that the Cramér–Rao lower bound applies as follows in three standard models:

1. $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2)$, with σ^2 known.
2. $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{Poisson}(\lambda)$.
3. The observations satisfy $Y_i = \boldsymbol{x}_i^\top\boldsymbol{\theta} + \varepsilon_i$, where the errors are independent and identically distributed as $\mathcal{N}(0, \sigma^2)$.

Solution. For (1), direct differentiation of the log likelihood gives

$$\dot{\ell}(Y,\theta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \theta) \qquad \text{and} \qquad I_n(\theta) = \frac{n}{\sigma^2}.$$

The sample mean has variance $\sigma^2/n = I_n(\theta)^{-1}$, so it attains the Cramér–Rao lower bound and is uniform minimum variance unbiased.

For (2),

$$\dot{\ell}(Y,\lambda) = \sum_{i=1}^{n}\left(\frac{Y_i}{\lambda} - 1\right) \qquad \text{and} \qquad I_n(\lambda) = \frac{n}{\lambda}.$$

Again, $\overline{Y}$ attains the bound because $\mathrm{Var}_\lambda(\overline{Y}) = \lambda/n = I_n(\lambda)^{-1}$. Its uniform minimum variance property also follows from Example 1.11.

For (3), maximizing the Gaussian likelihood is equivalent to minimizing the residual sum of squares:

$$\widehat{\boldsymbol{\theta}}_n \in \operatorname*{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{Y},\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^\top\boldsymbol{\theta})^2.$$

Under the stated zero-mean Gaussian error model and the usual full rank condition, this estimator is unbiased. For a non-Gaussian error distribution, least squares need not coincide with maximum likelihood; unbiasedness still requires appropriate zero-mean and exogeneity assumptions. This motivates the study of consistency and asymptotic normality under more general models. □

Definition 1.20 A sequence of estimators $\{\widehat{\theta}_n : n \geq 1\}$ is **consistent** for θ if

$$\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta.$$

Under suitable regularity conditions, maximum likelihood estimators satisfy the stronger conclusion

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}(0, I_1(\theta)^{-1}).$$

The next subtopic develops analogous conclusions for estimators defined by more general criterion and estimating functions, not only by negative log likelihoods.

Lecture 2 — Monday, January 21, 2019

M- and Z-Estimators

Definition 2.1 Given criterion functions m_θ, define

$$M_n(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i).$$

An estimator $\widehat{\theta}_n$ is an ***M*-estimator** if

$$\widehat{\theta}_n \in \operatorname*{argmax}_{\theta\in\Theta} M_n(\theta).$$

Definition 2.2 Given estimating functions ψ_θ, define

$$\Psi_n(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \psi_\theta(X_i).$$

An estimator $\widehat{\theta}_n$ is a ***Z*-estimator** if $\Psi_n(\widehat{\theta}_n) = 0$.

When an M-estimator is an interior maximizer of a differentiable criterion, its first-order condition frequently makes it a Z-estimator as well.

Definition 2.3 For a sequence of random variables R_n, the notation $R_n = o_{\mathbb{P}}(1)$ means that R_n is **little *o* in probability**; that is, for every $\varepsilon > 0$,

$$\mathbb{P}(|R_n| > \varepsilon) \longrightarrow 0.$$

The notation $R_n = O_{\mathbb{P}}(1)$ means that R_n is **bounded in probability**; that is, for every $\eta > 0$, there exist $M < \infty$ and n_0 such that

$$\mathbb{P}(|R_n| > M) \leq \eta$$

for every $n \geq n_0$.

Approximate M- and Z-estimators are also permitted. For example, an approximate estimator $\widetilde{\theta}_n$ may satisfy $\widetilde{\theta}_n = \widehat{\theta}_n + o_{\mathbb{P}}(1)$ for an exact optimizer or zero $\widehat{\theta}_n$.

Example 2.4 Suppose that $X_1, \ldots, X_n$ are independent and identically distributed according to $\mathbb{P}_\theta$, with density p_θ. The maximum likelihood estimator is an M-estimator with

$$m_\theta(x) = \log p_\theta(x).$$

If the log likelihood is differentiable and the maximizer is interior, it is also a Z-estimator with

$$\psi_\theta(x) = \nabla_\theta \log p_\theta(x).$$

Example 2.5 Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}[0, \theta]$. Determine the maximum likelihood estimator of θ.

Solution. The likelihood is

$$\mathcal{L}(\theta) = \theta^{-n} \mathbb{1}\left\{ \max_{1 \le i \le n} X_i \le \theta \right\}.$$

On its support, this likelihood is decreasing in θ. Consequently,

$$\widehat{\theta}_n = \max_{1 \le i \le n} X_i = X_{(n)}.$$

This is an M-estimator, but the criterion is not smooth at the sample maximum, so the estimator cannot be recovered by differentiating the criterion. □

Example 2.6 Let $X_1, \ldots, X_n$ be independent and identically distributed random variables. A general **location *Z*-estimator** solves

$$\sum_{i=1}^{n} \psi(X_i - \theta) = 0.$$

The choice $\psi(x) = x$ produces the sample mean, whereas the choice $\psi(x) = \text{sign}(x)$ produces a sample median, where

$$\text{sign}(x) = \begin{cases} -1, & x < 0, \\ 0, & x = 0, \\ 1, & x > 0. \end{cases}$$

Proposition 2.7 If $\widehat{\theta}_n$ solves

$$\sum_{i=1}^{n} \psi(X_i - \theta) = 0,$$

then $\widehat{\theta}_n + \alpha$ solves the corresponding estimating equation for the shifted sample $X_1 + \alpha, \ldots, X_n + \alpha$.

Proof. Substituting $\theta = \widehat{\theta}_n + \alpha$ into the shifted estimating equation gives

$$\sum_{i=1}^{n} \psi\big((X_i + \alpha) - (\widehat{\theta}_n + \alpha)\big) = \sum_{i=1}^{n} \psi(X_i - \widehat{\theta}_n) = 0.$$

□

Example 2.8 Suppose that independent observation pairs (X_i, Y_i) satisfy

$$Y_i = f_\theta(X_i) + \varepsilon_i \qquad \text{and} \qquad \mathbb{E}[\varepsilon_i \mid X_i] = 0.$$

The conditional mean assumption is an **exogeneity condition**; its failure is a form of **endogeneity** and may bias regression estimators. Two M-estimators are obtained by minimizing, respectively, the **least squares** criterion

$$\sum_{i=1}^{n} \big(Y_i - f_\theta(X_i)\big)^2$$

and the **least absolute deviations** criterion

$$\sum_{i=1}^{n} \left|Y_i - f_\theta(X_i)\right|.$$

Example 2.9 Consider the linear model $Y_i = \boldsymbol{X}_i^\top \boldsymbol{\theta} + \varepsilon_i$. The squared loss is sensitive to large residuals, whereas the absolute loss grows only linearly. The two loss functions are compared in Figure 1.

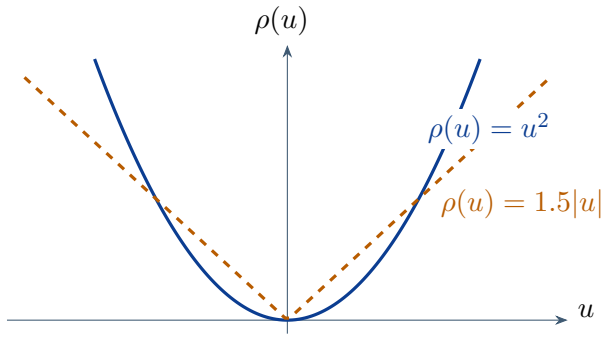


Figure 1

The **absolute loss** is more resistant to large residuals, while the **squared loss** is smooth and gives greater local sensitivity near zero. The **Huber loss** combines these behaviours:

$$\rho_\tau(u) = \begin{cases} u^2/2, & |u| \le \tau, \\ \tau|u| - \tau^2/2, & |u| > \tau. \end{cases}$$

Its quadratic center and linear tails are shown in Figure 2.

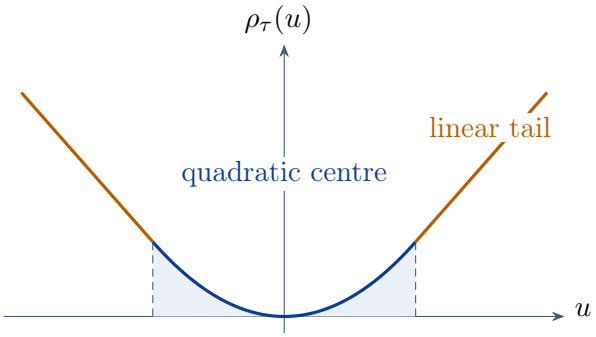


Figure 2

It is useful to distinguish **response outliers**, which have large residuals, from **leverage points**, which have unusual covariate values. An observation is **influential** when removing it substantially changes the fitted model. The Huber loss is robust to response outliers but does not by itself control leverage points. Indeed, the criterion has the form

$$M_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}),$$

whose first-order estimating equation is

$$\sum_{i=1}^{n} \psi_\tau(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}) \boldsymbol{X}_i = \mathbf{0},$$

where $\psi_\tau = \rho_\tau'$ is bounded as a function of the residual. The factor $\boldsymbol{X}_i$, however, remains unbounded. A bounded vector-valued weight function $w \colon \mathbb{R}^d \to \mathbb{R}^d$ may reduce the effect of extreme covariates by replacing the equation with

$$\sum_{i=1}^{n} \psi_\tau(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\theta}) w(\boldsymbol{X}_i) = \mathbf{0}.$$

Under suitable regularity conditions, some nonsmooth robust weighting procedures converge at a cube-root rather than a $\sqrt{n}$ rate. This behavior illustrates the general tradeoff between robustness and efficiency; it is not a universal consequence of weighting.

We now turn to consistency in the sense of Definition 1.20. For the M-estimator in Definition 2.1, define the **population criterion**

$$M(\theta) \coloneqq \mathbb{E}[m_\theta(X)] \qquad \text{and} \qquad \theta_0 \in \operatorname*{argmax}_{\theta \in \Theta} M(\theta).$$

For each fixed θ, the law of large numbers may give $M_n(\theta) \xrightarrow{\mathbb{P}} M(\theta)$. **Pointwise convergence** alone does not generally imply $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$, because the maximizer is itself random. **Uniform convergence** of the criterion is the key extra ingredient.

Theorem 2.10 Let $M_n \colon \Theta \to \mathbb{R}$ be a random criterion and let $M \colon \Theta \to \mathbb{R}$ be deterministic. Suppose that

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0$$

and that M has a well-separated maximum at θ_0: for every $\varepsilon > 0$,

$$\sup_{\theta : d(\theta, \theta_0) \geq \varepsilon} M(\theta) < M(\theta_0).$$

If

$$\widehat{\theta}_n \in \operatorname*{argmax}_{\theta \in \Theta} M_n(\theta),$$

then $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$.

Remark 2.11

1. The same conclusion holds for an approximate maximizer satisfying
$$M_n(\widehat{\theta}_n) \geq M_n(\theta_0) - o_{\mathbb{P}}(1).$$
2. The random criterion M_n need not be an empirical average of the form in Definition 2.1.

Proof. Since $\widehat{\theta}_n$ maximizes M_n,
$$\begin{aligned} 0 \leq M(\theta_0) - M(\widehat{\theta}_n) &= M(\theta_0) - M_n(\theta_0) + M_n(\theta_0) - M_n(\widehat{\theta}_n) + M_n(\widehat{\theta}_n) - M(\widehat{\theta}_n) \\ &\leq 2 \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| = o_{\mathbb{P}}(1). \end{aligned}$$

Fix $\varepsilon > 0$ and define the positive separation constant
$$\eta_\varepsilon \coloneqq M(\theta_0) - \sup_{\theta : d(\theta, \theta_0) \geq \varepsilon} M(\theta) > 0.$$

If $d(\widehat{\theta}_n, \theta_0) \geq \varepsilon$, then $M(\theta_0) - M(\widehat{\theta}_n) \geq \eta_\varepsilon$. Hence
$$\mathbb{P}\big(d(\widehat{\theta}_n, \theta_0) \geq \varepsilon\big) \leq \mathbb{P}\big(M(\theta_0) - M(\widehat{\theta}_n) \geq \eta_\varepsilon\big) \longrightarrow 0.$$

The proof for the approximate maximizer in (1) is identical after adding its $o_{\mathbb{P}}(1)$ optimization error. □

Theorem 2.12 Let $\Psi_n \colon \Theta \to \mathbb{R}^k$ be random and let $\Psi \colon \Theta \to \mathbb{R}^k$ be deterministic. Suppose that
$$\sup_{\theta \in \Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \xrightarrow{\mathbb{P}} 0$$
and that, for every $\varepsilon > 0$,
$$\inf_{\theta : d(\theta, \theta_0) \geq \varepsilon} \|\Psi(\theta)\| > 0 = \|\Psi(\theta_0)\|.$$
If $\Psi_n(\widehat{\theta}_n) = 0$, then $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$.

Proof. Set $M_n(\theta) = -\|\Psi_n(\theta)\|$ and $M(\theta) = -\|\Psi(\theta)\|$. The reverse triangle inequality gives

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \le \sup_{\theta\in\Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \xrightarrow{\mathbb{P}} 0.$$

Because $\widehat{\theta}_n$ is a zero of Ψ_n, it maximizes M_n. The separation assumption for Ψ is precisely the well-separated maximum condition for M. The result therefore follows from Theorem 2.10. □

Uniform convergence is substantially stronger than pointwise convergence and can be difficult to verify directly.

Definition 2.13 A class of functions $\mathcal{F}$ is a **Glivenko–Cantelli class** if

$$\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - \mathbb{P} f| \xrightarrow{\mathbb{P}} 0,$$

where

$$\mathbb{P}_n f \coloneqq \frac{1}{n}\sum_{i=1}^{n} f(X_i) \qquad \text{and} \qquad \mathbb{P} f \coloneqq \mathbb{E}[f(X)].$$

Glivenko–Cantelli theory supplies the uniform laws of large numbers required by Theorems 2.10 and 2.12.

Lemma 2.14 Let Θ be a compact metric space. Suppose that $x \mapsto f_\theta(x)$ is measurable for every $\theta \in \Theta$, that $\theta \mapsto f_\theta(x)$ is continuous for every x, and that

$$|f_\theta(x)| \le F(x)$$

for every $\theta \in \Theta$, where $\mathbb{P}F < \infty$. Then $\{f_\theta : \theta \in \Theta\}$ is a Glivenko–Cantelli class.

Proof. Fix $\eta > 0$. For $\theta \in \Theta$ and $\delta > 0$, define

$$g_{\theta,\delta}(x) \coloneqq \sup_{d(\vartheta,\theta)<\delta} |f_\vartheta(x) - f_\theta(x)|.$$

Pointwise continuity gives $g_{\theta,\delta}(x) \downarrow 0$ as $\delta \downarrow 0$, while $g_{\theta,\delta} \le 2F$. Dominated convergence therefore permits a radius $\delta_\theta > 0$ such that $\mathbb{P} g_{\theta,\delta_\theta} < \eta$. Compactness supplies finitely many centres $\theta_1, \dots, \theta_N$ whose corresponding neighborhoods cover Θ.

For any ϑ in the neighborhood of θ_j,

$$|\mathbb{P}_n f_\vartheta - \mathbb{P} f_\vartheta| \le |\mathbb{P}_n f_{\theta_j} - \mathbb{P} f_{\theta_j}| + \mathbb{P}_n g_{\theta_j,\delta_{\theta_j}} + \mathbb{P} g_{\theta_j,\delta_{\theta_j}}.$$

Taking the supremum over ϑ and then the maximum over the finitely many centres shows that the limiting superior is at most 2η in probability, by the ordinary law of large numbers. Since $\eta > 0$ was arbitrary, the class is Glivenko–Cantelli. □

Example 2.15 Let $X_1, \dots, X_n$ follow a Cauchy location model with density

$$p_\theta(x) = \frac{1}{\pi\big(1 + (x-\theta)^2\big)},$$

where θ belongs to a compact parameter set. The maximum likelihood estimator maximizes

$$M_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log\big(1 + (X_i - \theta)^2\big).$$

The compactness criterion in Lemma 2.14 gives the uniform convergence needed for consistency once the population criterion has a unique maximum.

Lemma 2.16 Let $\Theta = \mathbb{R}$. Suppose that $\Psi_n(\theta) \xrightarrow{\mathbb{P}} \Psi(\theta)$ for every θ, that each map $\theta \mapsto \Psi_n(\theta)$ is nondecreasing, and that $\Psi_n(\widehat{\theta}_n) = 0$. If, for every $\varepsilon > 0$,

$$\Psi(\theta_0 - \varepsilon) < 0 < \Psi(\theta_0 + \varepsilon),$$

then $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$.

Proof. Fix $\varepsilon > 0$. On the event

$$A_n := \{\Psi_n(\theta_0 - \varepsilon) < 0 < \Psi_n(\theta_0 + \varepsilon)\},$$

monotonicity and $\Psi_n(\widehat{\theta}_n) = 0$ imply

$$\theta_0 - \varepsilon < \widehat{\theta}_n < \theta_0 + \varepsilon.$$

Pointwise convergence at the two fixed endpoints gives $\mathbb{P}(A_n) \to 1$. Therefore

$$\mathbb{P}(|\widehat{\theta}_n - \theta_0| \ge \varepsilon) \le \mathbb{P}(A_n^c) \longrightarrow 0.$$

□

Example 2.17 Let

$$\Psi_n(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \operatorname{sign}(\theta - X_i).$$

If θ_0 is the unique population median and the distribution has a positive density in a neighborhood of θ_0, then any exact zero $\widehat{\theta}_n$ of Ψ_n is consistent for θ_0.

Solution. For every fixed θ, the law of large numbers gives

$$\Psi_n(\theta) \xrightarrow{\mathbb{P}} \Psi(\theta) \coloneqq \mathbb{P}(X < \theta) - \mathbb{P}(X > \theta).$$

The empirical functions are nondecreasing. Positivity of the density near the unique median implies

$$\Psi(\theta_0 - \varepsilon) < 0 < \Psi(\theta_0 + \varepsilon)$$

for every $\varepsilon > 0$. The conclusion follows from Lemma 2.16. □

Lecture 3 — Monday, January 28, 2019

Asymptotic Normality

Notation 3.1 For a measurable function f and a random variable X, write

$$\mathbb{P}f \coloneqq \mathbb{E}[f(X)], \qquad \mathbb{P}_n f \coloneqq \frac{1}{n}\sum_{i=1}^{n} f(X_i), \qquad \text{and} \qquad \mathbb{G}_n f \coloneqq \sqrt{n}(\mathbb{P}_n f - \mathbb{P}f).$$

Theorem 3.2 Let $\theta_0 \in \mathbb{R}$, define $\Psi_n(\theta) = \mathbb{P}_n \Psi_\theta$, and suppose that the following conditions hold:

1. There is a neighborhood B of θ_0 such that $\theta \mapsto \Psi_\theta(x)$ is twice continuously differen-

tiable on B for every x, and

$$|\ddot{\Psi}_\theta(x)| \le \ddot{\Psi}(x)$$

for every $\theta \in B$, where $\mathbb{P}\ddot{\Psi} < \infty$.

2. We have

$$\mathbb{P}\Psi_{\theta_0} = 0, \qquad \mathbb{P}\Psi_{\theta_0}^2 < \infty, \qquad \text{and} \qquad A \coloneqq \mathbb{P}\dot{\Psi}_{\theta_0} \neq 0.$$

Here $\mathbb{P}|\dot{\Psi}_{\theta_0}| < \infty$ and A is finite.

3. The estimator $\widehat{\theta}_n$ is a zero of Ψ_n and $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$.

Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}\left(0, \frac{\mathbb{P}\Psi_{\theta_0}^2}{A^2}\right).$$

The uniform second derivative envelope in (1) is the strongest of these assumptions. Its role is to control the Taylor remainder even though the intermediate point in the expansion is random.

Proof. Taylor's theorem gives a random point $\widetilde{\theta}_n$ between $\widehat{\theta}_n$ and θ_0 such that

$$0 = \Psi_n(\widehat{\theta}_n) = \Psi_n(\theta_0) + (\widehat{\theta}_n - \theta_0)\dot{\Psi}_n(\theta_0) + \frac{1}{2}(\widehat{\theta}_n - \theta_0)^2\ddot{\Psi}_n(\widetilde{\theta}_n).$$

Thus

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -\frac{\sqrt{n}\Psi_n(\theta_0)}{\dot{\Psi}_n(\theta_0) + \frac{1}{2}(\widehat{\theta}_n - \theta_0)\ddot{\Psi}_n(\widetilde{\theta}_n)}.$$

By the central limit theorem and the law of large numbers,

$$\sqrt{n}\Psi_n(\theta_0) \rightsquigarrow \mathcal{N}(0, \mathbb{P}\Psi_{\theta_0}^2) \qquad \text{and} \qquad \dot{\Psi}_n(\theta_0) \xrightarrow{\mathbb{P}} A.$$

It remains to control the second derivative at the random intermediate point. Let $E_n = \{\widetilde{\theta}_n \in B\}$. Consistency implies $\mathbb{P}(E_n) \to 1$. On E_n, the envelope assumption gives

$$|\ddot{\Psi}_n(\widetilde{\theta}_n)| \le \mathbb{P}_n\ddot{\Psi} \xrightarrow{\mathbb{P}} \mathbb{P}\ddot{\Psi} < \infty.$$

More explicitly, for $K > 0$,

$$\mathbb{P}\big(|\ddot{\Psi}_n(\widetilde{\theta}_n)| > K\big) \le \mathbb{P}(\mathbb{P}_n\ddot{\Psi} > K) + \mathbb{P}(E_n^c),$$

so $\ddot{\Psi}_n(\widetilde{\theta}_n) = O_{\mathbb{P}}(1)$. Consequently,

$$(\widehat{\theta}_n - \theta_0)\ddot{\Psi}_n(\widetilde{\theta}_n) = o_{\mathbb{P}}(1)O_{\mathbb{P}}(1) = o_{\mathbb{P}}(1).$$

Slutsky's theorem now gives the asserted limiting distribution. □

Theorem 3.3 Suppose that $\mathbb{P}\Psi_{\theta_0} = 0$, $\mathbb{P}\Psi_{\theta_0}^2 < \infty$, and the following conditions hold:

1. The map $\theta \mapsto \mathbb{P}\Psi_\theta$ is differentiable at θ_0, with nonzero derivative A.
2. For every deterministic sequence $\delta_n \downarrow 0$,

$$\sup_{|\theta-\theta_0|\le\delta_n} |\mathbb{G}_n(\Psi_\theta - \Psi_{\theta_0})| = o_{\mathbb{P}}(1).$$

If $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$ and $\mathbb{P}_n\Psi_{\widehat{\theta}_n} = 0$, then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}\left(0, \frac{\mathbb{P}\Psi_{\theta_0}^2}{A^2}\right).$$

Proof. The estimating equation can be written as

$$0 = \sqrt{n}\mathbb{P}\Psi_{\widehat{\theta}_n} + \mathbb{G}_n\Psi_{\widehat{\theta}_n}.$$

Consistency and (2) imply

$$\mathbb{G}_n\Psi_{\widehat{\theta}_n} = \mathbb{G}_n\Psi_{\theta_0} + o_{\mathbb{P}}(1) = O_{\mathbb{P}}(1).$$

Differentiability in (1), together with $A \neq 0$, gives a local constant $c > 0$ such that

$$|\mathbb{P}\Psi_\theta| \ge c|\theta - \theta_0|$$

for θ sufficiently close to θ_0. The estimating equation thus first yields $\sqrt{n}(\widehat{\theta}_n - \theta_0) = O_{\mathbb{P}}(1)$. Expanding the population map now gives

$$0 = A\sqrt{n}(\widehat{\theta}_n - \theta_0) + \mathbb{G}_n\Psi_{\theta_0} + o_{\mathbb{P}}(1),$$

and hence

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -A^{-1}\mathbb{G}_n\Psi_{\theta_0} + o_{\mathbb{P}}(1).$$

The central limit theorem and Slutsky's theorem complete the proof. □

Remark 3.4

1. The empirical process condition (2) is commonly obtained by showing that the local class is **Donsker** and asymptotically equicontinuous.
2. A useful sufficient condition is a local Lipschitz bound
$$|\Psi_{\theta_1}(y) - \Psi_{\theta_2}(y)| \leq \dot{\Psi}(y)|\theta_1 - \theta_2|$$
with an appropriately square-integrable envelope $\dot{\Psi}$.
3. In the vector-valued case, let
$$\mathbf{A} = \mathbb{P}\dot{\boldsymbol{\Psi}}_{\boldsymbol{\theta}_0} \qquad \text{and} \qquad \mathbf{B} = \mathbb{P}(\boldsymbol{\Psi}_{\boldsymbol{\theta}_0}\boldsymbol{\Psi}_{\boldsymbol{\theta}_0}^\top).$$
Then the limiting covariance matrix is $\mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^\top$.

Theorem 3.5 Suppose that the following conditions hold:

1. The population criterion has the expansion
$$\mathbb{P}m_{\boldsymbol{\theta}} - \mathbb{P}m_{\boldsymbol{\theta}_0} = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{V}_{\boldsymbol{\theta}_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + o(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2),$$
where $\mathbf{V}_{\boldsymbol{\theta}_0}$ is nonsingular and negative definite.
2. There is a square-integrable vector function $\dot{\boldsymbol{m}}_{\boldsymbol{\theta}_0}$ such that, for every $M < \infty$,
$$\sup_{\|\boldsymbol{h}\|_2 \leq M} \left| \sqrt{n}\,\mathbb{G}_n\big(m_{\boldsymbol{\theta}_0 + \boldsymbol{h}/\sqrt{n}} - m_{\boldsymbol{\theta}_0}\big) - \boldsymbol{h}^\top \mathbb{G}_n \dot{\boldsymbol{m}}_{\boldsymbol{\theta}_0} \right| = o_{\mathbb{P}}(1).$$

If $\widehat{\boldsymbol{\theta}}_n$ is a consistent approximate maximizer whose local optimization error is $o_{\mathbb{P}}(n^{-1})$, and if
$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = O_{\mathbb{P}}(1),$$
then
$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \rightsquigarrow \mathcal{N}\left(\mathbf{0}, \mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbb{P}(\dot{\boldsymbol{m}}_{\boldsymbol{\theta}_0}\dot{\boldsymbol{m}}_{\boldsymbol{\theta}_0}^\top)\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\right).$$

Proof. For bounded $\boldsymbol{h}$, (1) and (2) give the local criterion expansion

$$n\big(\mathbb{P}_n m_{\boldsymbol{\theta}_0+\boldsymbol{h}/\sqrt{n}} - \mathbb{P}_n m_{\boldsymbol{\theta}_0}\big) = \boldsymbol{h}^\top \mathbb{G}_n \dot{\boldsymbol{m}}_{\boldsymbol{\theta}_0} + \frac{1}{2}\boldsymbol{h}^\top \mathbf{V}_{\boldsymbol{\theta}_0}\boldsymbol{h} + o_{\mathbb{P}}(1),$$

uniformly on compact sets. The strictly concave quadratic on the right is maximized at

$$\boldsymbol{h}_n^* = -\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbb{G}_n \dot{\boldsymbol{m}}_{\boldsymbol{\theta}_0}.$$

The local argmax theorem therefore gives

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbb{G}_n \dot{\boldsymbol{m}}_{\boldsymbol{\theta}_0} + o_{\mathbb{P}}(1).$$

The multivariate central limit theorem gives the stated covariance matrix. □

Definition 3.6 A family of densities $\{p_\theta : \theta \in \Theta\}$ is **identifiable** if

$$p_{\theta_1} = p_{\theta_2} \text{ almost everywhere} \Longrightarrow \theta_1 = \theta_2.$$

Lemma 3.7 Let $\{p_\theta : \theta \in \Theta\}$ be identifiable, suppose that the true density is p_{θ_0}, and define

$$M(\theta) \coloneqq \mathbb{P}_{\theta_0} \log\left(\frac{p_\theta}{p_{\theta_0}}\right).$$

The expectation may take the value $-\infty$; on $\{p_{\theta_0} > 0, p_\theta = 0\}$ the logarithm is interpreted as $-\infty$, and its value on $\{p_{\theta_0} = 0\}$ is immaterial. Then M has the unique maximum $M(\theta_0) = 0$ at θ_0.

Proof. We have $M(\theta_0) = 0$. The inequality $\log x \le 2(\sqrt{x} - 1)$ for $x > 0$ yields

$$\begin{aligned} M(\theta) &= \int p_{\theta_0}(x) \log\left(\frac{p_\theta(x)}{p_{\theta_0}(x)}\right) \mathrm{d}x \\ &\le 2\int p_{\theta_0}(x)\left(\sqrt{\frac{p_\theta(x)}{p_{\theta_0}(x)}} - 1\right)\mathrm{d}x \\ &= -\int \left(\sqrt{p_\theta(x)} - \sqrt{p_{\theta_0}(x)}\right)^2 \mathrm{d}x \\ &\le 0. \end{aligned}$$

Equality holds only when $p_\theta = p_{\theta_0}$ almost everywhere, which identifiability makes equivalent to

$\theta = \theta_0$. □

Theorem 3.8 Let $X_1, \ldots, X_n$ be independent and identically distributed according to p_{θ_0}, let the model be identifiable, and define

$$M_n(\theta) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{p_\theta(X_i)}{p_{\theta_0}(X_i)} \right).$$

Suppose that M_n converges uniformly in probability to M from Lemma 3.7 and that its maximum is well separated. Then every maximum likelihood estimator

$$\widehat{\boldsymbol{\theta}}_n \in \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} M_n(\boldsymbol{\theta})$$

is consistent for $\boldsymbol{\theta}_0$.

Proof. The population criterion has its unique well-separated maximum at $\boldsymbol{\theta}_0$ by Lemma 3.7. The conclusion is therefore an immediate application of Theorem 2.10. □

Theorem 3.9 Under the vector-valued analogue of the regularity conditions in Theorem 3.3, suppose that $\mathbf{I}_1(\boldsymbol{\theta}_0)$ is nonsingular. Then a consistent interior maximum likelihood estimator satisfies

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \rightsquigarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1}).$$

Equivalently, for large n, its distribution is approximated by $\mathcal{N}(\boldsymbol{\theta}_0, (n\mathbf{I}_1(\boldsymbol{\theta}_0))^{-1})$.

Proof. The likelihood first-order condition makes the estimator a Z-estimator with

$$\boldsymbol{\Psi}_{\boldsymbol{\theta}}(x) = \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(x) = \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(x).$$

By Proposition 1.16,

$$\mathbb{P}_{\boldsymbol{\theta}_0} \dot{\boldsymbol{\Psi}}_{\boldsymbol{\theta}_0} = -\mathbf{I}_1(\boldsymbol{\theta}_0) \qquad \text{and} \qquad \mathbb{P}_{\boldsymbol{\theta}_0}(\boldsymbol{\Psi}_{\boldsymbol{\theta}_0} \boldsymbol{\Psi}_{\boldsymbol{\theta}_0}^\top) = \mathbf{I}_1(\boldsymbol{\theta}_0).$$

Substitution into the covariance formula in (3) gives $\mathbf{I}_1(\boldsymbol{\theta}_0)^{-1}$. □

The asymptotic normality theorem for maximum likelihood estimators (Theorem 3.9) suggests an asymptotic analogue of the Cramér–Rao lower bound. The following example shows why a

regularity condition on the estimator is indispensable.

Example 3.10 Let $Y_1, \dots, Y_n$ be independent and identically distributed as $\mathcal{N}(\theta, 1)$, fix $a \in (0, 1)$, and define the **Hodges estimator**

$$T_n \coloneqq \begin{cases} \overline{Y}, & |\overline{Y}| > n^{-1/4}, \\ a\overline{Y}, & |\overline{Y}| \leq n^{-1/4}. \end{cases}$$

Then

$$\sqrt{n}(T_n - \theta) \rightsquigarrow \begin{cases} \mathcal{N}(0, 1), & \theta \neq 0, \\ \mathcal{N}(0, a^2), & \theta = 0. \end{cases}$$

Thus T_n is **superefficient** at $\theta = 0$.

Solution. Set $Z_n = \sqrt{n}(\overline{Y} - \theta)$, so $Z_n \sim \mathcal{N}(0, 1)$ exactly. The threshold event can be written as

$$|\overline{Y}| \leq n^{-1/4} \iff |Z_n + \sqrt{n}\theta| \leq n^{1/4}.$$

If $\theta \neq 0$, its probability tends to zero, whereas if $\theta = 0$, its probability tends to one. On the complementary event the estimator equals $\overline{Y}$, and on the threshold event it equals $a\overline{Y}$. The two limits therefore follow from Slutsky's theorem. □

To see the failure of regularity, consider the local parameter sequence $\theta_n = c/\sqrt{n}$. Under $\mathbb{P}_{\theta_n}$, the threshold event has probability tending to one and

$$\sqrt{n}(T_n - \theta_n) = a\sqrt{n}(\overline{Y} - \theta_n) + c(a - 1) \rightsquigarrow \mathcal{N}(c(a - 1), a^2).$$

The limiting distribution depends on the local direction c and has nonzero mean whenever $c \neq 0$. The estimator is therefore not regular, despite its pointwise improvement at the origin. An asymptotic version of absolute continuity is needed to formalize such local comparisons.

Definition 3.11 Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on the same measurable space $(\Omega, \mathcal{A})$.

1. The measure $\mathbb{Q}$ is **absolutely continuous** with respect to $\mathbb{P}$, written $\mathbb{Q} \ll \mathbb{P}$, if $\mathbb{P}(A) = 0$ implies $\mathbb{Q}(A) = 0$ for every $A \in \mathcal{A}$.
2. The measures are **mutually singular**, written $\mathbb{Q} \perp \mathbb{P}$, if there is a measurable partition

$\Omega = \Omega_{\mathbb{P}} \dot\cup \Omega_{\mathbb{Q}}$ such that $\mathbb{P}(\Omega_{\mathbb{Q}}) = \mathbb{Q}(\Omega_{\mathbb{P}}) = 0$. Thus the two measures are concentrated on disjoint measurable sets.

Lemma 3.12 (Lebesgue decomposition) Suppose that $\mathbb{P}$ and $\mathbb{Q}$ have densities p and q with respect to a common dominating measure μ. Define

$$\mathbb{Q}^a(A) \coloneqq \mathbb{Q}(A \cap \{p > 0\}) \qquad \text{and} \qquad \mathbb{Q}^\perp(A) \coloneqq \mathbb{Q}(A \cap \{p = 0\}).$$

Then the following statements hold:

1. $\mathbb{Q} = \mathbb{Q}^a + \mathbb{Q}^\perp$.
2. $\mathbb{Q}^a \ll \mathbb{P}$ and $\mathbb{Q}^\perp \perp \mathbb{P}$.
3. With q/p defined to be zero on $\{p = 0\}$,
$$\mathbb{Q}^a(A) = \int_A \frac{q}{p}\, \mathrm{d}\mathbb{P}.$$
The function q/p is denoted by $\mathrm{d}\mathbb{Q}/\,\mathrm{d}\mathbb{P}$, even when $\mathbb{Q}$ is not absolutely continuous with respect to $\mathbb{P}$; it is the **Radon–Nikodym derivative** of the absolutely continuous part.
4. The following conditions are equivalent:
$$\mathbb{Q} \ll \mathbb{P}, \qquad \mathbb{Q}(\{p = 0\}) = 0, \qquad \text{and} \qquad \int_\Omega \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\, \mathrm{d}\mathbb{P} = 1.$$

Proof. The identity in (1) follows by partitioning A into $A \cap \{p > 0\}$ and $A \cap \{p = 0\}$. If $\mathbb{P}(A) = 0$, then $p = 0$ μ-almost everywhere on A, and hence $\mathbb{Q}^a(A) = 0$. Moreover, $\mathbb{P}(\{p = 0\}) = 0$, whereas $\mathbb{Q}^\perp$ is concentrated on $\{p = 0\}$. This proves (2). For every $A \in \mathcal{A}$,

$$\int_A \frac{q}{p}\, \mathrm{d}\mathbb{P} = \int_{A \cap \{p > 0\}} \frac{q}{p} p\, \mathrm{d}\mu = \mathbb{Q}^a(A),$$

which proves (3). Finally,

$$\int_\Omega \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\, \mathrm{d}\mathbb{P} = \mathbb{Q}^a(\Omega) = 1 - \mathbb{Q}(\{p = 0\}),$$

and (4) follows. □

Definition 3.13 The sequence $\{\mathbb{Q}_n\}$ is **contiguous** with respect to $\{\mathbb{P}_n\}$ if, for every sequence $A_n \in \mathcal{A}_n$,

$$\mathbb{P}_n(A_n) \longrightarrow 0 \qquad \text{implies} \qquad \mathbb{Q}_n(A_n) \longrightarrow 0.$$

This relation is denoted by $\mathbb{Q}_n \lhd \mathbb{P}_n$. The two sequences are **mutually contiguous** if each is contiguous with respect to the other; this is denoted by $\mathbb{P}_n \lhd\rhd \mathbb{Q}_n$.

Contiguity is an asymptotic analogue of absolute continuity. It transfers events whose probabilities vanish, even though the measures and measurable spaces may change with n.

Lemma 3.14 (Portmanteau lemma) For random vectors $\boldsymbol{X}_n$ and $\boldsymbol{X}$ taking values in $\mathbb{R}^k$, the following statements are equivalent:

1. $\boldsymbol{X}_n \rightsquigarrow \boldsymbol{X}$.
2. For every open set G,
$$\liminf_{n\to\infty} \mathbb{P}(\boldsymbol{X}_n \in G) \geq \mathbb{P}(\boldsymbol{X} \in G).$$
3. For every bounded continuous function f,
$$\mathbb{E}[f(\boldsymbol{X}_n)] \longrightarrow \mathbb{E}[f(\boldsymbol{X})].$$
4. The convergence in (3) holds for every bounded Lipschitz function f.
5. For every nonnegative continuous function f,
$$\liminf_{n\to\infty} \mathbb{E}[f(\boldsymbol{X}_n)] \geq \mathbb{E}[f(\boldsymbol{X})].$$

Lemma 3.15 (Helly's lemma) Any sequence of distribution functions on $\mathbb{R}^k$ has a subsequence that converges at every continuity point to a possibly **defective** distribution function F. If F is **proper**, then the corresponding probability measures converge weakly to the probability measure determined by F.

Theorem 3.16 (Prohorov's theorem) Let $\boldsymbol{X}_n$ be random vectors in $\mathbb{R}^k$. The family $\{\boldsymbol{X}_n : n \in \mathbb{N}\}$ is called **uniformly tight** if, for every $\epsilon > 0$, there is an $M > 0$ such that

$$\sup_n \mathbb{P}(\|\boldsymbol{X}_n\|_2 > M) < \epsilon.$$

Then the following statements hold:

1. If $\boldsymbol{X}_n \rightsquigarrow \boldsymbol{X}$ for some random vector $\boldsymbol{X}$, then $\{\boldsymbol{X}_n : n \in \mathbb{N}\}$ is uniformly tight.
2. If $\{\boldsymbol{X}_n : n \in \mathbb{N}\}$ is uniformly tight, then there are a subsequence $\{\boldsymbol{X}_{n_j}\}$ and a random vector $\boldsymbol{X}$ such that $\boldsymbol{X}_{n_j} \rightsquigarrow \boldsymbol{X}$ as $j \to \infty$.

Proof. For (1), fix $\epsilon > 0$. Choose $M > 0$ such that

$$\mathbb{P}(\|\boldsymbol{X}\|_2 \geq M) < \epsilon \qquad \text{and} \qquad \mathbb{P}(\|\boldsymbol{X}\|_2 = M) = 0.$$

This is possible because the tail probability tends to zero and at most countably many spheres can have positive probability. The Portmanteau lemma (Lemma 3.14) then gives

$$\mathbb{P}(\|\boldsymbol{X}_n\|_2 \geq M) \longrightarrow \mathbb{P}(\|\boldsymbol{X}\|_2 \geq M).$$

Thus there is an N such that $\mathbb{P}(\|\boldsymbol{X}_n\|_2 \geq M) < 2\epsilon$ for every $n \geq N$. Each of the finitely many random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{N-1}$ is tight, so M may be increased to make the same bound hold for every n. Since ϵ was arbitrary, the family is uniformly tight.

For (2), let

$$F_n(\boldsymbol{x}) := \mathbb{P}(\boldsymbol{X}_n \leq \boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^k,$$

where the inequality is interpreted coordinatewise. By Helly's lemma (Lemma 3.15), there is a subsequence $\{F_{n_j}\}$ that converges at every continuity point to a possibly defective distribution function F. It remains to prove that F is proper.

Fix $\epsilon > 0$. Uniform tightness gives an $M > 0$ such that

$$\inf_n \mathbb{P}(\boldsymbol{X}_n \in (-M, M)^k) > 1 - \epsilon.$$

After increasing M if necessary, we may assume that $M\mathbf{1}$ is a continuity point of F, where

$\mathbf{1} = (1, \dots, 1)^\top$. It follows that

$$F(M\mathbf{1}) = \lim_{j \to \infty} F_{n_j}(M\mathbf{1}) \geq 1 - \epsilon.$$

Hence $F(\boldsymbol{x}) \to 1$ as every coordinate of $\boldsymbol{x}$ tends to infinity. On the other hand, if $x_i \leq -M$ for some coordinate i, then

$$F_{n_j}(\boldsymbol{x}) \leq \mathbb{P}(X_{n_j,i} \leq -M) < \epsilon.$$

Passing to the limit at continuity points and using the right continuity of F shows that $F(\boldsymbol{x}) \to 0$ whenever any coordinate tends to negative infinity. Thus F is a proper distribution function. If $\boldsymbol{X}$ has distribution function F, Helly's lemma gives $\boldsymbol{X}_{n_j} \rightsquigarrow \boldsymbol{X}$. □

Remark 3.17 If $W_n \rightsquigarrow W$ and $|W_n| \leq M$ almost surely for every n, then $|W| \leq M$ almost surely and $\mathbb{E}[W_n] \to \mathbb{E}[W]$. Indeed, we may replace the identity function outside $[-M, M]$ by a bounded continuous function and then apply (3).

Lecture 4 — Monday, February 04, 2019

Theorem 4.1 (Le Cam's first lemma) Let $\mathbb{P}_n$ and $\mathbb{Q}_n$ be probability measures on $(\Omega_n, \mathcal{A}_n)$. The following statements are equivalent:

1. $\mathbb{Q}_n \lhd \mathbb{P}_n$.
2. Whenever, along a subsequence,

$$\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n} \rightsquigarrow U \quad \text{under } \mathbb{Q}_n,$$

   the limit satisfies $\mathbb{P}(U > 0) = 1$.
3. Whenever, along a subsequence,

$$\frac{\mathrm{d}\mathbb{Q}_n}{\mathrm{d}\mathbb{P}_n} \rightsquigarrow V \quad \text{under } \mathbb{P}_n,$$

   the limit satisfies $\mathbb{E}[V] = 1$.

4. For every sequence of statistics T_n,

$$T_n \xrightarrow{\mathbb{P}_n} 0 \qquad \text{implies} \qquad T_n \xrightarrow{\mathbb{Q}_n} 0.$$

Proof. *Equivalence of* (1) *and* (4). First suppose that (1) holds. For every $\epsilon > 0$, apply contiguity to $A_n = \{|T_n| > \epsilon\}$. This proves (4). Conversely, if $\mathbb{P}_n(A_n) \to 0$, then $T_n = \mathbb{1}_{A_n}$ converges to zero in $\mathbb{P}_n$-probability. Statement (4) therefore gives

$$\mathbb{Q}_n(A_n) = \mathbb{Q}_n\left(T_n > \tfrac{1}{2}\right) \longrightarrow 0,$$

which is (1).

(1) *implies* (2). Relabel the subsequence as $\{n\}$ and set $R_n = \mathrm{d}\mathbb{P}_n / \mathrm{d}\mathbb{Q}_n$. Choose continuity points $\epsilon_m \downarrow 0$ of the distribution function of U. By weak convergence, a diagonal argument gives integers $m(n) \to \infty$ such that

$$\mathbb{Q}_n(R_n \le \epsilon_{m(n)}) \longrightarrow \mathbb{P}(U = 0).$$

Using densities p_n and q_n with respect to a common dominating measure, define

$$A_n := \{R_n \le \epsilon_{m(n)}\} \cap \{q_n > 0\}.$$

The second restriction changes neither the $\mathbb{Q}_n$-probability nor the preceding limit, and

$$\mathbb{P}_n(A_n) = \int_{A_n} R_n \, \mathrm{d}\mathbb{Q}_n \le \epsilon_{m(n)} \longrightarrow 0.$$

Contiguity now yields $\mathbb{Q}_n(A_n) \to 0$. Consequently, $\mathbb{P}(U = 0) = 0$, which is equivalent to $\mathbb{P}(U > 0) = 1$ because U is nonnegative.

(3) *implies* (1). Suppose that $\mathbb{P}_n(A_n) \to 0$, set $B_n = \Omega_n \setminus A_n$, and let $L_n = \mathrm{d}\mathbb{Q}_n / \mathrm{d}\mathbb{P}_n$. If contiguity failed, a subsequence could be chosen on which $\mathbb{Q}_n(A_n)$ is bounded away from zero. Passing to a further subsequence if necessary, suppose that $L_n \rightsquigarrow V$ under $\mathbb{P}_n$; the absence of any such weakly convergent subsequence would force the laws of L_n to escape to infinity, which is impossible because $\mathbb{E}_{\mathbb{P}_n}[L_n] \le 1$. By Slutsky's lemma,

$$(L_n, \mathbb{1}_{B_n}) \rightsquigarrow (V, 1) \quad \text{under } \mathbb{P}_n.$$

The nonnegative function form of the Portmanteau lemma (Lemma 3.14) gives

$$1 = \mathbb{E}[V] \le \liminf_{n\to\infty} \mathbb{E}_{\mathbb{P}_n}[L_n \mathbb{1}_{B_n}] \le \liminf_{n\to\infty} \mathbb{Q}_n(B_n) \le 1.$$

Thus $\mathbb{Q}_n(B_n) \to 1$, contradicting the choice of the subsequence. Therefore (1) holds.

(2) *implies* (3). Consider a subsequence along which $L_n = \mathrm{d}\mathbb{Q}_n / \mathrm{d}\mathbb{P}_n \rightsquigarrow V$ under $\mathbb{P}_n$. Introduce the dominating probability measure

$$\mu_n := \tfrac{1}{2}(\mathbb{P}_n + \mathbb{Q}_n) \qquad \text{and} \qquad W_n := \frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mu_n}.$$

Then $0 \le W_n \le 2$ and

$$\frac{\mathrm{d}\mathbb{Q}_n}{\mathrm{d}\mu_n} = 2 - W_n.$$

By (2), a further subsequence satisfies $W_n \rightsquigarrow W$ under μ_n. Moreover, $\mathbb{E}_{\mu_n}[W_n] = 1$, so Remark 3.17 gives $\mathbb{E}[W] = 1$.

For $f \in C_b([0,\infty))$, define

$$g_f(w) = \begin{cases} f\left(\dfrac{w}{2-w}\right)(2-w), & 0 \le w < 2, \\ 0, & w = 2. \end{cases}$$

The function g_f is bounded and continuous. Hence

$$\mathbb{E}_{\mathbb{Q}_n}\left[f\left(\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n}\right)\right] = \mathbb{E}_{\mu_n}[g_f(W_n)] \longrightarrow \mathbb{E}[g_f(W)].$$

This identifies a weak limit U for the reciprocal likelihood ratios under $\mathbb{Q}_n$. Taking continuous functions f_m that decrease pointwise to $\mathbb{1}_{\{0\}}$ and applying dominated convergence gives

$$\mathbb{P}(U = 0) = 2\mathbb{P}(W = 0).$$

By (2), the left-hand side is zero, so $\mathbb{P}(W > 0) = 1$. Similarly, for every $f \in C_b([0,\infty))$,

$$\mathbb{E}[f(V)] = \mathbb{E}\left[f\left(\frac{2-W}{W}\right)W\right],$$

where the integrand is defined to be zero on $\{W = 0\}$. Apply this identity to $f_m(v) = \min\{v, m\}$ and use the monotone convergence theorem. Since $W > 0$ almost surely,

$$\mathbb{E}[V] = \mathbb{E}\left[\frac{2-W}{W}W\right] = 2 - \mathbb{E}[W] = 1.$$

This proves (3) and completes the cycle of equivalences. □

Example 4.2 Suppose that, under $\mathbb{Q}_n$,

$$\frac{\mathrm{d}\mathbb{P}_n}{\mathrm{d}\mathbb{Q}_n} \rightsquigarrow \exp(Z) \qquad \text{and} \qquad Z \sim \mathcal{N}(\mu, \sigma^2).$$

Determine the resulting contiguity relations.

Solution. Because $\exp(Z) > 0$ almost surely, Theorem 4.1(2) gives $\mathbb{Q}_n \lhd \mathbb{P}_n$. Applying Theorem 4.1(3) after interchanging the roles of the two sequences shows that $\mathbb{P}_n \lhd \mathbb{Q}_n$ precisely when

$$1 = \mathbb{E}[\exp(Z)] = \exp\left(\mu + \frac{\sigma^2}{2}\right),$$

or equivalently when $\mu = -\sigma^2/2$. Thus the sequences are mutually contiguous exactly under this condition. □

Theorem 4.3 (Le Cam's third lemma) Let $\mathbb{P}_n$ and $\mathbb{Q}_n$ be probability measures on $(\Omega_n, \mathcal{A}_n)$, and let $\boldsymbol{X}_n : \Omega_n \to \mathbb{R}^k$. Suppose that $\mathbb{Q}_n \lhd \mathbb{P}_n$ and, under $\mathbb{P}_n$,

$$\left(\boldsymbol{X}_n, \frac{\mathrm{d}\mathbb{Q}_n}{\mathrm{d}\mathbb{P}_n}\right) \rightsquigarrow (\boldsymbol{X}, V).$$

Then

$$L(B) \coloneqq \mathbb{E}[\mathbb{1}_B(\boldsymbol{X})V], \quad B \in \mathcal{B}(\mathbb{R}^k),$$

defines a probability measure, and $\boldsymbol{X}_n \rightsquigarrow L$ under $\mathbb{Q}_n$.

Proof. Set $V_n = \mathrm{d}\mathbb{Q}_n / \mathrm{d}\mathbb{P}_n$. By Theorem 4.1(3), $\mathbb{E}[V] = 1$, so L is a probability measure. Contiguity also implies that the singular mass of $\mathbb{Q}_n$ with respect to $\mathbb{P}_n$ tends to zero: apply the definition to a support of the singular part, whose $\mathbb{P}_n$-probability is zero. Consequently,

$$\mathbb{E}_{\mathbb{P}_n}[V_n] = \mathbb{Q}_n^a(\Omega_n) \longrightarrow 1 = \mathbb{E}[V].$$

For nonnegative random variables, weak convergence together with convergence of the means implies uniform integrability. Therefore, for every bounded continuous function f,

$$\mathbb{E}_{\mathbb{Q}_n}[f(\boldsymbol{X}_n)] = \mathbb{E}_{\mathbb{P}_n}[f(\boldsymbol{X}_n)V_n] + o(1) \longrightarrow \mathbb{E}[f(\boldsymbol{X})V] = \int f \,\mathrm{d}L.$$

The characterization in Lemma 3.14(3) proves the result. □

Corollary 4.4 (Gaussian form of Le Cam's third lemma) Suppose that, under $\mathbb{P}_n$,

$$\left(\boldsymbol{X}_n, \log \frac{\mathrm{d}\mathbb{Q}_n}{\mathrm{d}\mathbb{P}_n}\right) \rightsquigarrow \mathcal{N}_{k+1}\left(\begin{pmatrix}\boldsymbol{\mu} \\ -\sigma^2/2\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma} & \boldsymbol{\tau} \\ \boldsymbol{\tau}^\top & \sigma^2\end{pmatrix}\right).$$

Then, under $\mathbb{Q}_n$,

$$\boldsymbol{X}_n \rightsquigarrow \mathcal{N}_k(\boldsymbol{\mu}+\boldsymbol{\tau}, \boldsymbol{\Sigma}).$$

Proof. Write the limiting vector as $(\boldsymbol{X}, Z)$. Since $Z \sim \mathcal{N}(-\sigma^2/2, \sigma^2)$, one has $\mathbb{E}[e^Z] = 1$. Thus Theorem 4.1 gives contiguity, and Theorem 4.3 says that the limiting law is obtained by weighting the joint Gaussian law by e^Z. For $\boldsymbol{t} \in \mathbb{R}^k$,

$$\mathbb{E}\left[e^{\boldsymbol{t}^\top \boldsymbol{X}+Z}\right] = \exp\left(\boldsymbol{t}^\top\boldsymbol{\mu} - \frac{\sigma^2}{2} + \frac{1}{2}\left(\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t} + 2\boldsymbol{t}^\top\boldsymbol{\tau} + \sigma^2\right)\right) = \exp\left(\boldsymbol{t}^\top(\boldsymbol{\mu}+\boldsymbol{\tau}) + \frac{1}{2}\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}\right).$$

This is the moment generating function of $\mathcal{N}_k(\boldsymbol{\mu}+\boldsymbol{\tau}, \boldsymbol{\Sigma})$. □

Local Asymptotic Normality

Let $Y_1, \ldots, Y_n$ be independent and identically distributed according to $\mathbb{P}_{\boldsymbol{\theta}}$. Classical asymptotic theory studies the model at a fixed $\boldsymbol{\theta}$. Local asymptotic theory instead compares $\mathbb{P}_{\boldsymbol{\theta}}^{\otimes n}$ with alternatives of the form

$$\mathbb{P}_{\boldsymbol{\theta}+\boldsymbol{h}/\sqrt{n}}^{\otimes n}, \quad \boldsymbol{h} \in \mathbb{R}^k.$$

The displacement is of order $n^{-1/2}$, which is exactly the scale on which regular estimators fluctuate. A second order likelihood expansion turns these local experiments into an asymptotic Gaussian shift experiment.

Definition 4.5 Let $\Theta \subseteq \mathbb{R}^k$ be open. The model $\{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is **locally asymptotically normal** at $\boldsymbol{\theta}$ if there are random vectors $\boldsymbol{\Delta}_{n,\boldsymbol{\theta}}$ and a nonnegative definite matrix $\mathbf{I}_{\boldsymbol{\theta}}$ such that, under $\mathbb{P}_{\boldsymbol{\theta}}^{\otimes n}$,

$$\boldsymbol{\Delta}_{n,\boldsymbol{\theta}} \rightsquigarrow \mathcal{N}_k(\mathbf{0}, \mathbf{I}_{\boldsymbol{\theta}})$$

and, for every bounded sequence $\boldsymbol{h}_n$ satisfying $\boldsymbol{\theta} + \boldsymbol{h}_n/\sqrt{n} \in \Theta$,

$$\log \frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\theta}+\boldsymbol{h}_n/\sqrt{n}}^{\otimes n}}{\mathrm{d}\mathbb{P}_{\boldsymbol{\theta}}^{\otimes n}} = \boldsymbol{h}_n^\top \boldsymbol{\Delta}_{n,\boldsymbol{\theta}} - \frac{1}{2}\boldsymbol{h}_n^\top \mathbf{I}_{\boldsymbol{\theta}} \boldsymbol{h}_n + o_{\mathbb{P}_{\boldsymbol{\theta}}^{\otimes n}}(1). \tag{4.1}$$

In a smooth independent and identically distributed model,

$$\boldsymbol{\Delta}_{n,\boldsymbol{\theta}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(Y_i) = \mathbb{G}_n \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}},$$

where $\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}$ is the score vector and $\mathbf{I}_{\boldsymbol{\theta}} = \mathbb{P}_{\boldsymbol{\theta}}[\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}} \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}^{\top}]$ is the Fisher information matrix.

Definition 4.6 Suppose that $\mathbb{P}_{\boldsymbol{\theta}}$ has density $f_{\boldsymbol{\theta}}$ with respect to a dominating measure μ. The model is **differentiable in quadratic mean** at $\boldsymbol{\theta}$ if there is a measurable vector $\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}} \in L^2(\mathbb{P}_{\boldsymbol{\theta}})^k$ such that

$$\int \left(\sqrt{f_{\boldsymbol{\theta}+\boldsymbol{t}}} - \sqrt{f_{\boldsymbol{\theta}}} - \frac{1}{2} \boldsymbol{t}^{\top} \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}} \sqrt{f_{\boldsymbol{\theta}}} \right)^2 \mathrm{d}\mu = o(\|\boldsymbol{t}\|_2^2) \tag{4.2}$$

as $\boldsymbol{t} \to \mathbf{0}$ with $\boldsymbol{\theta} + \boldsymbol{t} \in \Theta$.

Theorem 4.7 If Θ is open and the model is differentiable in quadratic mean at $\boldsymbol{\theta}$, then

$$\mathbb{P}_{\boldsymbol{\theta}} \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}} = \mathbf{0} \qquad \text{and} \qquad \mathbf{I}_{\boldsymbol{\theta}} = \mathbb{P}_{\boldsymbol{\theta}}[\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}} \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}^{\top}]$$

is finite. Moreover, the model is locally asymptotically normal at $\boldsymbol{\theta}$ with central sequence

$$\boldsymbol{\Delta}_{n,\boldsymbol{\theta}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(Y_i).$$

Proof. Fix $\boldsymbol{u} \in \mathbb{R}^k$. From (4.2),

$$\frac{\sqrt{f_{\boldsymbol{\theta}+s\boldsymbol{u}}} - \sqrt{f_{\boldsymbol{\theta}}}}{s} \longrightarrow \frac{1}{2} \boldsymbol{u}^{\top} \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}} \sqrt{f_{\boldsymbol{\theta}}} \quad \text{in } L^2(\mu).$$

Multiplication by $\sqrt{f_{\boldsymbol{\theta}+s\boldsymbol{u}}} + \sqrt{f_{\boldsymbol{\theta}}}$ and the Cauchy–Schwarz inequality therefore give convergence in $L^1(\mu)$:

$$\frac{f_{\boldsymbol{\theta}+s\boldsymbol{u}} - f_{\boldsymbol{\theta}}}{s} \longrightarrow \boldsymbol{u}^{\top} \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}.$$

The integral on the left is zero. Hence $\boldsymbol{u}^{\top} \mathbb{P}_{\boldsymbol{\theta}} \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}} = 0$ for every $\boldsymbol{u}$, which proves that the score has mean zero. The finiteness of $\mathbf{I}_{\boldsymbol{\theta}}$ follows from the $L^2(\mathbb{P}_{\boldsymbol{\theta}})$ assumption in Definition 4.6.

Now let $\boldsymbol{h}_n$ be bounded and put $\boldsymbol{t}_n = \boldsymbol{h}_n/\sqrt{n}$. On the set $\{f_{\boldsymbol{\theta}} > 0\}$, define

$$s_n(y) \coloneqq \sqrt{\frac{f_{\boldsymbol{\theta}+\boldsymbol{t}_n}(y)}{f_{\boldsymbol{\theta}}(y)}} - 1.$$

Quadratic mean differentiability yields

$$n\mathbb{P}_{\boldsymbol{\theta}} \left[s_n - \frac{1}{2\sqrt{n}} \boldsymbol{h}_n^\top \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}} \right]^2 \longrightarrow 0. \tag{4.3}$$

Consequently,

$$2\sum_{i=1}^{n} \{s_n(Y_i) - \mathbb{P}_{\boldsymbol{\theta}} s_n\} = \boldsymbol{h}_n^\top \boldsymbol{\Delta}_{n,\boldsymbol{\theta}} + o_{\mathbb{P}_{\boldsymbol{\theta}}^{\otimes n}}(1), \tag{4.4}$$

$$\sum_{i=1}^{n} s_n(Y_i)^2 = \frac{1}{4} \boldsymbol{h}_n^\top \mathbf{I}_{\boldsymbol{\theta}} \boldsymbol{h}_n + o_{\mathbb{P}_{\boldsymbol{\theta}}^{\otimes n}}(1). \tag{4.5}$$

The normalization of $f_{\theta+t_n}$, together with (4.3), also gives

$$2n\mathbb{P}_{\boldsymbol{\theta}} s_n = -n\mathbb{P}_{\boldsymbol{\theta}} s_n^2 + o(1) = -\frac{1}{4} \boldsymbol{h}_n^\top \mathbf{I}_{\boldsymbol{\theta}} \boldsymbol{h}_n + o(1). \tag{4.6}$$

The same quadratic mean bound implies $\max_{1 \le i \le n} |s_n(Y_i)| \to 0$ in probability. Therefore the Taylor expansion $2\log(1+x) = 2x - x^2 + o(x^2)$, summed over the observations, gives

$$\begin{aligned}
\log \frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\theta}+\boldsymbol{t}_n}^{\otimes n}}{\mathrm{d}\mathbb{P}_{\boldsymbol{\theta}}^{\otimes n}} &= 2\sum_{i=1}^{n} \log(1 + s_n(Y_i)) \\
&= 2\sum_{i=1}^{n} \{s_n(Y_i) - \mathbb{P}_{\boldsymbol{\theta}} s_n\} + 2n\mathbb{P}_{\boldsymbol{\theta}} s_n - \sum_{i=1}^{n} s_n(Y_i)^2 + o_{\mathbb{P}}(1) \\
&= \boldsymbol{h}_n^\top \boldsymbol{\Delta}_{n,\boldsymbol{\theta}} - \frac{1}{2} \boldsymbol{h}_n^\top \mathbf{I}_{\boldsymbol{\theta}} \boldsymbol{h}_n + o_{\mathbb{P}}(1),
\end{aligned}$$

by (4.4)–(4.6). Finally, the multivariate central limit theorem gives $\boldsymbol{\Delta}_{n,\boldsymbol{\theta}} \rightsquigarrow \mathcal{N}_k(\mathbf{0}, \mathbf{I}_{\boldsymbol{\theta}})$. This is precisely Definition 4.5. □

Quadratic mean differentiability is weaker than pointwise differentiability of every density. It captures the L^2 smoothness needed for the likelihood expansion and is therefore the natural minimal regularity condition in many independent and identically distributed models.

Lecture 5 — Monday, February 11, 2019

Asymptotic Cramér–Rao Lower Bound

Definition 5.1 In the **Gaussian shift experiment**

$$\boldsymbol{Y} \sim \mathcal{N}_k(\boldsymbol{h}, \boldsymbol{\Sigma}), \quad \boldsymbol{h} \in \mathbb{R}^k,$$

where $\boldsymbol{\Sigma}$ is positive definite. A possibly randomized statistic $\boldsymbol{T}$ is **regular for estimating** $\mathbf{A}\boldsymbol{h}$ if the distribution of $\boldsymbol{T} - \mathbf{A}\boldsymbol{h}$ does not depend on $\boldsymbol{h}$.

Theorem 5.2 (Gaussian convolution theorem) Let $\boldsymbol{T}$ be regular for estimating $\mathbf{A}\boldsymbol{h}$ in the Gaussian shift experiment of Definition 5.1, and let L be the distribution of $\boldsymbol{T}$ when $\boldsymbol{h} = \mathbf{0}$. Then

$$L \overset{d}{=} \boldsymbol{Z} + \boldsymbol{W},$$

where $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$ and $\boldsymbol{Z}$ is independent of $\boldsymbol{W}$.

Proof. Let $\varphi_L(t)$ be the characteristic function of L. Regularity gives

$$\mathbb{E}_{\boldsymbol{h}}[e^{i\boldsymbol{t}^\top \boldsymbol{T}}] = e^{i\boldsymbol{t}^\top \mathbf{A}\boldsymbol{h}} \varphi_L(\boldsymbol{t}).$$

Under the distribution with $\boldsymbol{h} = \mathbf{0}$, the Gaussian likelihood ratio is

$$\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{h}}}{\mathrm{d}\mathbb{P}_{\mathbf{0}}}(\boldsymbol{Y}) = \exp\left(\boldsymbol{h}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{Y} - \frac{1}{2}\boldsymbol{h}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{h}\right).$$

Taking $\boldsymbol{h} = \boldsymbol{\Sigma}\boldsymbol{s}$ and using uniqueness of the bilateral Laplace transform of the Gaussian measure yields the joint characteristic function

$$\mathbb{E}_{\mathbf{0}}[e^{i\boldsymbol{t}^\top \boldsymbol{T} + i\boldsymbol{u}^\top \boldsymbol{Y}}] = \varphi_L(\boldsymbol{t}) \exp\left(-\frac{1}{2}\boldsymbol{u}^\top \boldsymbol{\Sigma}\boldsymbol{u} - \boldsymbol{t}^\top \mathbf{A}\boldsymbol{\Sigma}\boldsymbol{u}\right). \tag{5.1}$$

Set $\boldsymbol{W} = \boldsymbol{T} - \mathbf{A}\boldsymbol{Y}$. Replacing $\boldsymbol{u}$ by $\boldsymbol{u} - \mathbf{A}^\top \boldsymbol{t}$ in (5.1) gives

$$\mathbb{E}_{\mathbf{0}}[e^{i\boldsymbol{t}^\top \boldsymbol{W} + i\boldsymbol{u}^\top \boldsymbol{Y}}] = \varphi_L(\boldsymbol{t}) \exp\left(\frac{1}{2}\boldsymbol{t}^\top \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top \boldsymbol{t}\right) \exp\left(-\frac{1}{2}\boldsymbol{u}^\top \boldsymbol{\Sigma}\boldsymbol{u}\right).$$

The factorization shows that $\boldsymbol{W}$ and $\boldsymbol{Y}$ are independent. Since $\boldsymbol{Z} = \mathbf{A}\boldsymbol{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$ under $\mathbb{P}_{\mathbf{0}}$, one has $\boldsymbol{T} = \boldsymbol{Z} + \boldsymbol{W}$, which proves the claim. □

Definition 5.3 An estimator sequence $\boldsymbol{T}_n$ is **regular for estimating** $\boldsymbol{\psi}(\boldsymbol{\theta})$ at $\boldsymbol{\theta}$ if, for every fixed $\boldsymbol{h} \in \mathbb{R}^k$, the limit distribution under $\mathbb{P}^{\otimes n}_{\boldsymbol{\theta}+\boldsymbol{h}/\sqrt{n}}$ of

$$\sqrt{n}\left\{\boldsymbol{T}_n - \boldsymbol{\psi}\left(\boldsymbol{\theta} + \frac{\boldsymbol{h}}{\sqrt{n}}\right)\right\}$$

exists and is the same distribution $L_{\boldsymbol{\theta}}$, independent of $\boldsymbol{h}$.

Example 5.4 In the normal location model, compare the Hodges estimator from Example 3.10 with the sample mean under the local sequence $\theta_n = h/\sqrt{n}$.

Solution. By the calculation in Example 3.10, under $\mathbb{P}^{\otimes n}_{h/\sqrt{n}}$,

$$\sqrt{n}\left(T_n - \frac{h}{\sqrt{n}}\right) \rightsquigarrow \mathcal{N}(h(a-1), a^2).$$

The limit depends on h, so the Hodges estimator is not regular at zero. In contrast, if $S_n = \overline{Y}$, then

$$\sqrt{n}(S_n - \theta_n) \sim \mathcal{N}(0, 1)$$

for every n and every θ_n. Hence the sample mean is regular, with limit distribution $\mathcal{N}(0,1)$. □

Definition 5.5 A **randomized statistic** based on an observation $\boldsymbol{Y}$ is a statistic of the form $\boldsymbol{T}(\boldsymbol{Y}, U)$, where U has the uniform distribution on $[0,1]$ and is independent of $\boldsymbol{Y}$. Equivalently, it is a probability kernel from the sample space of $\boldsymbol{Y}$ to the output space.

Theorem 5.6 Suppose that $\{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is locally asymptotically normal at $\boldsymbol{\theta}$ with non-singular information $\mathbf{I}_{\boldsymbol{\theta}}$. Write $\mathbb{P}_{n,\boldsymbol{h}} = \mathbb{P}^{\otimes n}_{\boldsymbol{\theta}+\boldsymbol{h}/\sqrt{n}}$. If the statistic $\boldsymbol{S}_n$ converges in distribution under $\mathbb{P}_{n,\boldsymbol{h}}$ to a law $L_{\boldsymbol{h}}$ for every $\boldsymbol{h} \in \mathbb{R}^k$, then there is a randomized statistic $\boldsymbol{T}$ in the Gaussian shift experiment

$$\boldsymbol{Y} \sim \mathcal{N}_k(\boldsymbol{h}, \mathbf{I}_{\boldsymbol{\theta}}^{-1}), \quad \boldsymbol{h} \in \mathbb{R}^k,$$

whose distribution under $\mathcal{N}_k(\boldsymbol{h}, \mathbf{I}_{\boldsymbol{\theta}}^{-1})$ is $L_{\boldsymbol{h}}$.

Proof. Let $\boldsymbol{\Delta}_n$ be the central sequence in Definition 4.5. By (4.1), under $\mathbb{P}_{n,\mathbf{0}}$,

$$\log \frac{\mathrm{d}\mathbb{P}_{n,\boldsymbol{h}}}{\mathrm{d}\mathbb{P}_{n,\mathbf{0}}} = \boldsymbol{h}^\top \boldsymbol{\Delta}_n - \frac{1}{2}\boldsymbol{h}^\top \mathbf{I}_{\boldsymbol{\theta}} \boldsymbol{h} + o_{\mathbb{P}_{n,\mathbf{0}}}(1).$$

For fixed $\boldsymbol{h}$, the limiting log likelihood ratio is normal with mean $-\boldsymbol{h}^\top \mathbf{I}_{\boldsymbol{\theta}} \boldsymbol{h}/2$ and variance $\boldsymbol{h}^\top \mathbf{I}_{\boldsymbol{\theta}} \boldsymbol{h}$. Its exponential is positive and has expectation one. The criterion in Example 4.2 therefore shows that $\mathbb{P}_{n,\boldsymbol{h}}$ and $\mathbb{P}_{n,\mathbf{0}}$ are mutually contiguous.

The sequence $\boldsymbol{S}_n$ is tight under $\mathbb{P}_{n,\mathbf{0}}$, and $\boldsymbol{\Delta}_n \rightsquigarrow \mathcal{N}_k(\mathbf{0}, \mathbf{I}_{\boldsymbol{\theta}})$. Thus $(\boldsymbol{S}_n, \boldsymbol{\Delta}_n)$ is jointly tight. By (2), along a subsequence,

$$(\boldsymbol{S}_n, \boldsymbol{\Delta}_n) \rightsquigarrow (\boldsymbol{S}, \boldsymbol{\Delta}) \quad \text{under } \mathbb{P}_{n,\mathbf{0}},$$

where $\boldsymbol{\Delta} \sim \mathcal{N}_k(\mathbf{0}, \mathbf{I}_{\boldsymbol{\theta}})$. The Gaussian form of Le Cam's third lemma, Corollary 4.4, identifies the limit under $\mathbb{P}_{n,\boldsymbol{h}}$ as

$$L_{\boldsymbol{h}}(B) = \mathbb{E}\left[\mathbb{1}_B(\boldsymbol{S}) \exp\left(\boldsymbol{h}^\top \boldsymbol{\Delta} - \frac{1}{2}\boldsymbol{h}^\top \mathbf{I}_{\boldsymbol{\theta}} \boldsymbol{h}\right)\right]. \tag{5.2}$$

Let $K(\boldsymbol{\delta}, \cdot)$ be a regular conditional distribution of $\boldsymbol{S}$ given $\boldsymbol{\Delta} = \boldsymbol{\delta}$. Because Euclidean spaces are standard Borel spaces, the randomization lemma supplies a measurable function r such that $r(\boldsymbol{\delta}, U)$ has distribution $K(\boldsymbol{\delta}, \cdot)$ when $U \sim \mathrm{Unif}(0,1)$. Take $\boldsymbol{Y} \sim \mathcal{N}_k(\boldsymbol{h}, \mathbf{I}_{\boldsymbol{\theta}}^{-1})$ independently of U and define

$$\boldsymbol{T}(\boldsymbol{Y}, U) \coloneqq r(\mathbf{I}_{\boldsymbol{\theta}} \boldsymbol{Y}, U).$$

When $\boldsymbol{h} = \mathbf{0}$, $(\boldsymbol{T}(\boldsymbol{Y}, U), \mathbf{I}_{\boldsymbol{\theta}} \boldsymbol{Y})$ has the same joint distribution as $(\boldsymbol{S}, \boldsymbol{\Delta})$. The likelihood ratio of $\mathcal{N}_k(\boldsymbol{h}, \mathbf{I}_{\boldsymbol{\theta}}^{-1})$ with respect to $\mathcal{N}_k(\mathbf{0}, \mathbf{I}_{\boldsymbol{\theta}}^{-1})$ is

$$\exp\left(\boldsymbol{h}^\top \mathbf{I}_{\boldsymbol{\theta}} \boldsymbol{Y} - \frac{1}{2}\boldsymbol{h}^\top \mathbf{I}_{\boldsymbol{\theta}} \boldsymbol{h}\right).$$

It follows that

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{h}}(\boldsymbol{T}(\boldsymbol{Y}, U) \in B) &= \mathbb{E}_{\mathbf{0}}\left[\mathbb{1}_B(\boldsymbol{T}(\boldsymbol{Y}, U)) e^{\boldsymbol{h}^\top \mathbf{I}_{\boldsymbol{\theta}} \boldsymbol{Y} - \boldsymbol{h}^\top \mathbf{I}_{\boldsymbol{\theta}} \boldsymbol{h}/2}\right] \\
&= \mathbb{E}\left[\mathbb{1}_B(\boldsymbol{S}) e^{\boldsymbol{h}^\top \boldsymbol{\Delta} - \boldsymbol{h}^\top \mathbf{I}_{\boldsymbol{\theta}} \boldsymbol{h}/2}\right] \\
&= L_{\boldsymbol{h}}(B),
\end{aligned}$$

where the last equality is (5.2). □

Lemma 5.7 Under the assumptions of Theorem 5.6, suppose that $\boldsymbol{\psi}$ is differentiable at $\boldsymbol{\theta}$ and

$$\sqrt{n}\left\{\boldsymbol{T}_n - \boldsymbol{\psi}\left(\boldsymbol{\theta} + \frac{\boldsymbol{h}}{\sqrt{n}}\right)\right\} \rightsquigarrow L_{\boldsymbol{\theta},\boldsymbol{h}} \quad \text{under } \mathbb{P}_{n,\boldsymbol{h}}$$

for every $\boldsymbol{h} \in \mathbb{R}^k$. Then there is a randomized statistic $\boldsymbol{T}$ in $\{\mathcal{N}_k(\boldsymbol{h}, \mathbf{I}_{\boldsymbol{\theta}}^{-1}) : \boldsymbol{h} \in \mathbb{R}^k\}$ whose distribution under the parameter $\boldsymbol{h}$ is

$$L_{\boldsymbol{\theta},\boldsymbol{h}} + \dot{\psi}_{\boldsymbol{\theta}}\boldsymbol{h}.$$

Proof. Apply Theorem 5.6 to $\boldsymbol{S}_n = \sqrt{n}\{\boldsymbol{T}_n - \boldsymbol{\psi}(\boldsymbol{\theta})\}$. Differentiability gives

$$\begin{aligned}\boldsymbol{S}_n &= \sqrt{n}\left\{\boldsymbol{T}_n - \boldsymbol{\psi}\left(\boldsymbol{\theta} + \frac{\boldsymbol{h}}{\sqrt{n}}\right)\right\} + \sqrt{n}\left\{\boldsymbol{\psi}\left(\boldsymbol{\theta} + \frac{\boldsymbol{h}}{\sqrt{n}}\right) - \boldsymbol{\psi}(\boldsymbol{\theta})\right\} \\ &\rightsquigarrow L_{\boldsymbol{\theta},\boldsymbol{h}} + \dot{\psi}_{\boldsymbol{\theta}}\boldsymbol{h} \quad \text{under } \mathbb{P}_{n,\boldsymbol{h}}.\end{aligned}$$

The claimed randomized statistic is therefore the one supplied by Theorem 5.6. □

The preceding result translates the original estimator into an estimator of the linearized parameter $\dot{\psi}_{\boldsymbol{\theta}}\boldsymbol{h}$ in the limiting Gaussian shift experiment. When $\boldsymbol{\psi}$ is the identity map, the limiting statistic estimates $\boldsymbol{h}$ itself.

Theorem 5.8 (Hájek–Le Cam convolution theorem) Suppose that $\{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is locally asymptotically normal at $\boldsymbol{\theta}$ with nonsingular Fisher information $\mathbf{I}_{\boldsymbol{\theta}}$. Let $\boldsymbol{\psi}$ be differentiable at $\boldsymbol{\theta}$, and let $\boldsymbol{T}_n$ be regular for estimating $\boldsymbol{\psi}(\boldsymbol{\theta})$ at $\boldsymbol{\theta}$, with limit distribution $L_{\boldsymbol{\theta}}$. Then

$$L_{\boldsymbol{\theta}} \overset{d}{=} \boldsymbol{Z}_{\boldsymbol{\theta}} + \boldsymbol{W}_{\boldsymbol{\theta}},$$

where $\boldsymbol{Z}_{\boldsymbol{\theta}}$ and $\boldsymbol{W}_{\boldsymbol{\theta}}$ are independent and

$$\boldsymbol{Z}_{\boldsymbol{\theta}} \sim \mathcal{N}\left(\mathbf{0}, \dot{\psi}_{\boldsymbol{\theta}}\mathbf{I}_{\boldsymbol{\theta}}^{-1}\dot{\psi}_{\boldsymbol{\theta}}^{\top}\right).$$

In particular, if $\boldsymbol{\psi}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, then $\boldsymbol{Z}_{\boldsymbol{\theta}} \sim \mathcal{N}_k(\mathbf{0}, \mathbf{I}_{\boldsymbol{\theta}}^{-1})$.

Proof. By regularity, $L_{\boldsymbol{\theta},\boldsymbol{h}} = L_{\boldsymbol{\theta}}$ for every $\boldsymbol{h}$. Apply Lemma 5.7 to obtain a randomized statistic $\boldsymbol{T}$ in the Gaussian shift experiment $\mathcal{N}_k(\boldsymbol{h}, \mathbf{I}_{\boldsymbol{\theta}}^{-1})$ such that

$$\boldsymbol{T} - \dot{\psi}_{\boldsymbol{\theta}}\boldsymbol{h} \overset{d}{=} L_{\boldsymbol{\theta}}$$

for every $\boldsymbol{h}$. Thus $\boldsymbol{T}$ is regular for estimating $\dot{\psi}_{\boldsymbol{\theta}}\boldsymbol{h}$. The result follows from Theorem 5.2 with $\mathbf{A} = \dot{\psi}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma} = \mathbf{I}_{\boldsymbol{\theta}}^{-1}$. □

If $L_{\boldsymbol{\theta}}$ has a finite covariance matrix, the independent decomposition in Theorem 5.8 gives the asymptotic Cramér–Rao bound

$$\operatorname{Cov}(L_{\boldsymbol{\theta}}) \succeq \dot{\psi}_{\boldsymbol{\theta}} \mathbf{I}_{\boldsymbol{\theta}}^{-1} \dot{\psi}_{\boldsymbol{\theta}}^{\top},$$

where $\mathbf{A} \succeq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is nonnegative definite. The bound is attained if and only if the limit distribution is Gaussian with the indicated covariance matrix. In particular, if $\boldsymbol{\psi}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, then the asymptotic covariance of any regular estimator is bounded below by the inverse Fisher information matrix:

$$\operatorname{Cov}(L_{\boldsymbol{\theta}}) \succeq \mathbf{I}_{\boldsymbol{\theta}}^{-1}.$$

Lecture 6 — Monday, February 18, 2019

2. Tail Bounds and Sub-Gaussian Concentration

Basic Tail Bounds

There are three related ways to study the accuracy of statistical procedures.

1. In **classical asymptotics**, the sample size tends to infinity while the dimension and other structural quantities remain fixed. The principal objects are limiting distributions. Point estimation, consistency and asymptotic normality of M-estimators, and asymptotic Cramér–Rao bounds belong to this regime.
2. In **nonasymptotic analysis**, the sample size n, dimension d, and other problem dimensions are finite. The objective is an explicit error bound that records how accuracy depends on these quantities.
3. **Stein's method** gives quantitative bounds on the distance between a finite sample distribution and an approximating distribution. It thereby bridges the limiting perspective in (1) and the finite sample perspective in (2).

For example, let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be independent and identically distributed according to F, with mean

$\boldsymbol{\mu}_F$, and define

$$\boldsymbol{T}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

If F has covariance matrix $\boldsymbol{\Sigma}$, the central limit theorem describes the limiting distribution

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\mu}_F) \rightsquigarrow \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

A nonasymptotic result instead has the form

$$\mathbb{P}\left(\left\|\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\mu}_F)\right\|_2 > \delta\right) \leq f(n, d, \delta),$$

valid for specified finite values of n, d, and δ. More generally, we may bound a probability metric by $d(F_n, F) \leq f(n, d)$. The tail event is depicted in Figure 3; the distribution and the bound may both depend on the problem dimensions (n, d).

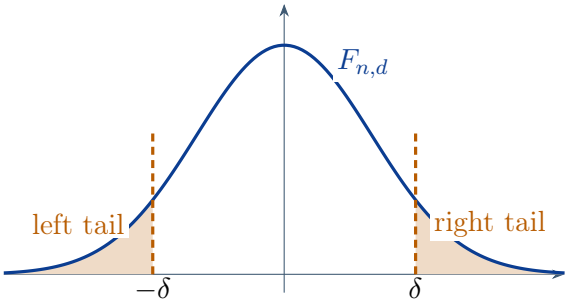


FIGURE 3

The immediate objective is to obtain upper bounds on such tail probabilities. Markov's inequality and its exponential version, the Chernoff bound, provide the basic tools.

From Markov to Chernoff

Theorem 6.1 (Markov's inequality) If Z is a nonnegative random variable, then, for every $t > 0$,

$$\mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}[Z]}{t}.$$

Consequently, if X has mean μ and a finite absolute central moment of order $k \geq 1$, then

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\mathbb{E}[|X - \mu|^k]}{t^k}. \tag{6.1}$$

Proof. The pointwise inequality $\mathbb{1}_{\{Z \geq t\}} \leq Z/t$ gives Markov's inequality after taking expectations. Apply it to $Z = |X - \mu|^k$ to obtain (6.1). □

The case $k = 1$ is a first moment bound. When $k = 2$, (6.1) becomes Chebyshev's inequality:

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\operatorname{Var}(X)}{t^2}. \tag{6.2}$$

If the moment generating function exists, Markov's inequality can instead be applied to an exponential transform.

Theorem 6.2 (Chernoff bound) Suppose that X has mean μ and that

$$\varphi(\lambda) := \mathbb{E}[e^{\lambda(X-\mu)}] < \infty$$

for $0 \leq \lambda \leq b$, where $b > 0$. Then, for $t > 0$ and $0 < \lambda \leq b$,

$$\mathbb{P}(X - \mu \geq t) \leq e^{-\lambda t}\mathbb{E}[e^{\lambda(X-\mu)}]. \tag{6.3}$$

Consequently,

$$\log \mathbb{P}(X - \mu \geq t) \leq -\sup_{0 \leq \lambda \leq b} \left\{ \lambda t - \log \mathbb{E}[e^{\lambda(X-\mu)}] \right\}. \tag{6.4}$$

Proof. Because $x \mapsto e^{\lambda x}$ is increasing for $\lambda > 0$,

$$\mathbb{P}(X - \mu \geq t) = \mathbb{P}(e^{\lambda(X-\mu)} \geq e^{\lambda t}).$$

Markov's inequality gives (6.3). Taking logarithms and optimizing over λ proves (6.4). □

The optimization in (6.4) is called the **Chernoff approach**. We may alternatively optimize the order k in (6.1). When all moments exist, the two approaches are closely related: choosing k to match the tail level often recovers the same decay rate, although the numerical constants can differ.

Sub-Gaussian Random Variables

The Chernoff bound (Theorem 6.2) depends on the growth of the moment generating function. The simplest useful regime is Gaussian growth.

Example 6.3 Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$\mathbb{E}[e^{\lambda(X-\mu)}] = e^{\lambda^2\sigma^2/2}.$$

Determine the resulting Chernoff bound.

Solution. By Theorem 6.2,

$$\log \mathbb{P}(X - \mu \geq t) \leq -\sup_{\lambda \geq 0} \left\{ \lambda t - \frac{1}{2}\lambda^2\sigma^2 \right\}.$$

The concave quadratic objective is displayed in Figure 4. Its maximum occurs at $\lambda = t/\sigma^2$ and equals $t^2/(2\sigma^2)$. Therefore

$$\mathbb{P}(X - \mu \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{6.5}$$

The exponent is optimal at the logarithmic scale, although (6.5) is an upper bound rather than the exact Gaussian tail probability.

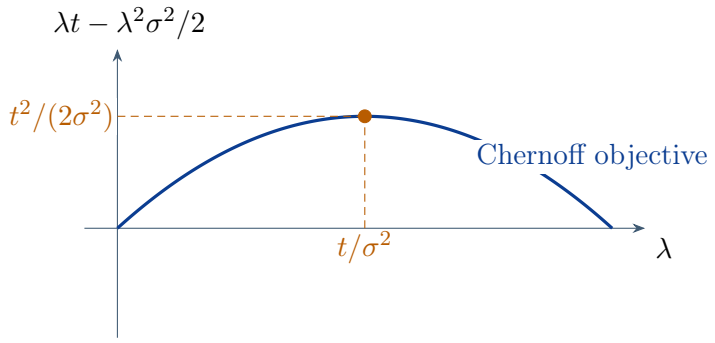


Figure 4

□

Definition 6.4 A random variable X with mean μ is **sub-Gaussian with variance proxy σ^2**, where $\sigma^2 \geq 0$, if

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\sigma^2\lambda^2/2} \tag{6.6}$$

for every $\lambda \in \mathbb{R}$.

The quantity σ^2 is called a variance proxy because differentiating (6.6) twice at zero gives $\mathrm{Var}(X) \leq \sigma^2$. A zero variance proxy forces $X = \mu$ almost surely, so tail formulas below are stated for positive proxies.

Lemma 6.5 If X has mean zero and is sub-Gaussian with variance proxy $\sigma^2 > 0$, then, for every $t \geq 0$,

$$\mathbb{P}(X \geq t) \leq e^{-t^2/(2\sigma^2)}, \tag{6.7}$$

$$\mathbb{P}(X \leq -t) \leq e^{-t^2/(2\sigma^2)}. \tag{6.8}$$

Consequently,

$$\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/(2\sigma^2)}.$$

Proof. For $\lambda > 0$, the Chernoff bound in Theorem 6.2 and (6.6) give

$$\mathbb{P}(X \geq t) \leq \exp\left(-\lambda t + \frac{1}{2}\sigma^2\lambda^2\right).$$

Choosing $\lambda = t/\sigma^2$ proves (6.7). Apply the same argument to $-X$ to obtain (6.8), and then use the union bound. □

Lemma 6.6 Suppose that $\mathbb{E}[X] = 0$ and, for every $t \geq 0$,

$$\mathbb{P}(X \geq t) \leq e^{-t^2/(2\sigma^2)} \qquad \text{and} \qquad \mathbb{P}(X \leq -t) \leq e^{-t^2/(2\sigma^2)}.$$

Then, for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}[e^{\lambda X}] \leq e^{4\sigma^2\lambda^2}.$$

Thus two-sided Gaussian tail decay implies a sub-Gaussian moment generating function bound, up to a universal constant.

Proof. For every integer $k \geq 1$, the layer-cake formula and the two-sided tail bound give

$$\mathbb{E}[|X|^k] = k \int_0^\infty t^{k-1} \mathbb{P}(|X| > t)\, \mathrm{d}t \leq 2k \int_0^\infty t^{k-1} e^{-t^2/(2\sigma^2)}\, \mathrm{d}t = k(2\sigma^2)^{k/2}\Gamma(k/2),$$

where the last identity follows from the substitution $u = t^2/(2\sigma^2)$. Since $\mathbb{E}[X] = 0$, the linear term in the exponential series vanishes, and hence

$$\mathbb{E}[e^{\lambda X}] \leq 1 + \sum_{k=2}^{\infty} \frac{|\lambda|^k k(2\sigma^2)^{k/2}\Gamma(k/2)}{k!}.$$

Separating even and odd indices and using $\Gamma(j) = (j-1)!$ and $\Gamma(j+1/2) \leq \sqrt{\pi}\, j!$ shows that the last series is at most

$$1 + \sum_{j=1}^{\infty} \frac{(4\sigma^2\lambda^2)^j}{j!} = e^{4\sigma^2\lambda^2}.$$

□

The same proof allows a fixed multiplicative constant in each tail bound; only the universal constant in the resulting moment generating function bound changes.

Thus moment generating function control yields sharp tail control, and two-sided tail control recovers a moment generating function bound with a change in the universal constant.

Theorem 6.7 Let X be a mean-zero random variable. The following statements are equivalent, with the positive constants allowed to differ by universal factors from one statement to another:

1. There is a finite $\sigma > 0$ such that
$$\mathbb{E}[e^{\lambda X}] \leq e^{\lambda^2\sigma^2/2}$$
for every $\lambda \in \mathbb{R}$.
2. There are constants $c, \tau > 0$ and $Z \sim \mathcal{N}(0, \tau^2)$ such that
$$\mathbb{P}(|X| \geq t) \leq c\mathbb{P}(|Z| \geq t)$$
for every $t \geq 0$.

3. There is a finite $\theta > 0$ such that, for every integer $k \geq 1$,

$$\mathbb{E}[X^{2k}] \leq \frac{(2k)!}{2^k k!}\theta^{2k}.$$

4. There is a finite $\theta > 0$ such that, for $0 \leq \lambda < 1$,

$$\mathbb{E}\left[e^{\lambda X^2/(2\theta^2)}\right] \leq \frac{1}{\sqrt{1-\lambda}}.$$

Proof. (1) *implies* (2). By Lemma 6.5, $\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/(2\sigma^2)}$. If $Z \sim \mathcal{N}(0, \tau^2)$ with $\tau > \sigma$, the Gaussian lower-tail estimate shows that the ratio $e^{-t^2/(2\sigma^2)}/\mathbb{P}(|Z| \geq t)$ is bounded on $[0, \infty)$. Absorbing this bound into c proves (2).

(2) *implies* (3). The layer-cake formula gives

$$\mathbb{E}[|X|^{2k}] = 2k \int_0^\infty t^{2k-1}\mathbb{P}(|X| \geq t)\,\mathrm{d}t \leq c\mathbb{E}[|Z|^{2k}] = c\frac{(2k)!}{2^k k!}\tau^{2k}.$$

Enlarging τ by a constant factor absorbs c uniformly over $k \geq 1$ and proves (3).

(3) *implies* (4). Expanding the exponential and using the moment bound yields

$$\mathbb{E}\left[e^{\lambda X^2/(2\theta^2)}\right] = \sum_{k=0}^\infty \frac{\lambda^k \mathbb{E}[X^{2k}]}{2^k\theta^{2k}k!} \leq \sum_{k=0}^\infty \frac{(2k)!}{(k!)^2}\left(\frac{\lambda}{4}\right)^k = \frac{1}{\sqrt{1-\lambda}}.$$

(4) *implies* (1). Taking $\lambda = 1/2$ in (4) and applying Markov's inequality gives

$$\mathbb{P}(|X| \geq t) \leq \sqrt{2}\, e^{-t^2/(4\theta^2)}.$$

The prefactor version of Lemma 6.6 therefore gives (1) with a modified variance proxy. □

Example 6.8 Let ε take the values -1 and 1, each with probability $1/2$. Show that ε is sub-Gaussian with variance proxy 1.

Solution. For every $\lambda \in \mathbb{R}$,

$$\mathbb{E}[e^{\lambda\varepsilon}] = \frac{e^{-\lambda} + e^{\lambda}}{2} = \cosh(\lambda) \leq e^{\lambda^2/2}.$$

The last inequality follows by comparing power series coefficients, since $(2k)! \geq 2^k k!$ for every integer $k \geq 0$. □

Example 6.9 Suppose that $a \leq X \leq b$ almost surely and $\mathbb{E}[X] = 0$. Show by symmetrization that X is sub-Gaussian with variance proxy $(b-a)^2$.

Solution. Let X^* be an independent copy of X. Jensen's inequality gives

$$\mathbb{E}[e^{\lambda X}] = \mathbb{E}_X\left[e^{\lambda\{X - \mathbb{E}_{X^*}[X^*]\}}\right] \leq \mathbb{E}_{X,X^*}[e^{\lambda(X-X^*)}].$$

The difference $X - X^*$ is symmetric, regardless of whether X itself is symmetric. If ε is an independent Rademacher random variable, then Example 6.8 gives

$$\mathbb{E}_{X,X^*}[e^{\lambda(X-X^*)}] = \mathbb{E}_{X,X^*,\varepsilon}[e^{\lambda\varepsilon(X-X^*)}] \leq \mathbb{E}_{X,X^*}\left[e^{\lambda^2(X-X^*)^2/2}\right] \leq e^{\lambda^2(b-a)^2/2}.$$

This is (6.6) with variance proxy $(b-a)^2$. □

The calculation in Example 6.9 is a basic symmetrization argument. Symmetry of X is unnecessary, and the variance proxy obtained by that argument is not sharp. The following result gives the sharp universal constant.

Lemma 6.10 (Hoeffding's lemma) If $a \leq X \leq b$ almost surely and $\mu = \mathbb{E}[X]$, then, for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\lambda^2(b-a)^2/8}. \tag{6.9}$$

Equivalently, $X - \mu$ is sub-Gaussian with variance proxy $(b-a)^2/4$.

Proof. Let $\psi(\lambda) = \log \mathbb{E}[e^{\lambda(X-\mu)}]$. Under the probability measure obtained by exponentially tilting the law of X by $e^{\lambda(X-\mu)}$, differentiation gives

$$\psi''(\lambda) = \mathrm{Var}_\lambda(X).$$

Every distribution supported on $[a, b]$ has variance at most $(b-a)^2/4$, because

$$\mathrm{Var}_\lambda(X) \leq \mathbb{E}_\lambda\left[\left\{X - \frac{a+b}{2}\right\}^2\right] \leq \frac{(b-a)^2}{4}.$$

Since $\psi(0) = \psi'(0) = 0$, Taylor's formula with integral remainder yields, for $\lambda \geq 0$,

$$\psi(\lambda) = \int_0^\lambda (\lambda - t)\psi''(t)\,\mathrm{d}t \leq \frac{\lambda^2(b-a)^2}{8}.$$

The same argument, integrating from λ to zero, applies when $\lambda < 0$. Exponentiating proves (6.9). □

Theorem 6.11 Let $X_1, \ldots, X_n$ be independent, let $\mu_i = \mathbb{E}[X_i]$, and suppose that $X_i - \mu_i$ is sub-Gaussian with variance proxy σ_i^2, with $\sum_{i=1}^n \sigma_i^2 > 0$. Then the following statements hold:

1. For every $\lambda \in \mathbb{R}$,
$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n (X_i - \mu_i)\right)\right] \leq \exp\left(\frac{\lambda^2}{2}\sum_{i=1}^n \sigma_i^2\right).$$
2. For every $t \geq 0$,
$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mu_i) \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

Proof. Independence and (6.6) give

$$\mathbb{E}\left[e^{\lambda \sum_{i=1}^n (X_i - \mu_i)}\right] = \prod_{i=1}^n \mathbb{E}[e^{\lambda(X_i-\mu_i)}] \leq \prod_{i=1}^n e^{\lambda^2\sigma_i^2/2} = e^{\lambda^2 \sum_{i=1}^n \sigma_i^2/2},$$

which proves (1). Apply the Chernoff bound and optimize at $\lambda = t/\sum_{i=1}^n \sigma_i^2$ to obtain (2). □

Corollary 6.12 (Hoeffding's inequality) If the independent random variables X_i satisfy

$a \le X_i \le b$ almost surely, where $a < b$, then

$$\mathbb{E}\left[e^{\lambda \sum_{i=1}^n (X_i - \mu_i)}\right] \le e^{\lambda^2 n (b-a)^2/8} \qquad \text{and} \qquad \mathbb{P}\left(\sum_{i=1}^n (X_i - \mu_i) \ge t\right) \le e^{-2t^2/\{n(b-a)^2\}}.$$

Proof. By Lemma 6.10, each $X_i - \mu_i$ is sub-Gaussian with variance proxy $(b-a)^2/4$. Substitute this proxy into (1) and (2). □

Lecture 7 — Monday, February 25, 2019

Subexponential Random Variables and Bernstein's Bound

Definition 7.1 A random variable X with mean μ is **subexponential with parameters (ν, α)** if $\nu, \alpha \ge 0$ and

$$\mathbb{E}[e^{\lambda(X-\mu)}] \le e^{\nu^2 \lambda^2/2} \tag{7.1}$$

whenever $|\lambda| \le 1/\alpha$. When $\alpha = 0$, the restriction is interpreted as allowing every $\lambda \in \mathbb{R}$.

Every sub-Gaussian random variable with variance proxy σ^2 is therefore subexponential with parameters $(\sigma, 0)$. The distinction is that a subexponential moment generating function need only have Gaussian growth in a neighborhood of the origin.

Example 7.2 Let $Z \sim \mathcal{N}(0,1)$. Show that Z^2 is subexponential with parameters $(2,4)$, even though it is not sub-Gaussian.

Solution. The variable Z^2 has the chi-squared distribution with one degree of freedom and mean one. For $\lambda < 1/2$,

$$\mathbb{E}[e^{\lambda(Z^2-1)}] = \frac{e^{-\lambda}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(1/2-\lambda)z^2} \, \mathrm{d}z = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}.$$

For $|\lambda| \le 1/4$, the elementary bound

$$-\lambda - \frac{1}{2}\log(1-2\lambda) \le 2\lambda^2$$

therefore gives

$$\mathbb{E}[e^{\lambda(Z^2-1)}] \le e^{2\lambda^2} = e^{4\lambda^2/2}.$$

This is (7.1) with $(\nu, \alpha) = (2, 4)$. The moment generating function of Z^2 is infinite for $\lambda \ge 1/2$, so Z^2 is not sub-Gaussian. □

Proposition 7.3 If X is subexponential with parameters $\nu > 0$ and $\alpha > 0$ and mean μ, then, for every $t \ge 0$,

$$\mathbb{P}(X - \mu \ge t) \le \begin{cases} \exp\left(-\dfrac{t^2}{2\nu^2}\right), & 0 \le t \le \dfrac{\nu^2}{\alpha}, \\ \exp\left(-\dfrac{t}{2\alpha}\right), & t > \dfrac{\nu^2}{\alpha}. \end{cases} \tag{7.2}$$

The same bounds hold for the lower tail, and hence the corresponding two-sided bound has an additional factor of 2.

Proof. For $0 \le \lambda \le 1/\alpha$, Theorem 6.2 gives

$$\mathbb{P}(X - \mu \ge t) \le \exp\left(-\lambda t + \frac{1}{2}\nu^2\lambda^2\right).$$

If $t \le \nu^2/\alpha$, the unconstrained minimizer $\lambda = t/\nu^2$ is admissible and gives the first case of (7.2). If $t > \nu^2/\alpha$, choose $\lambda = 1/\alpha$ to obtain

$$\mathbb{P}(X - \mu \ge t) \le \exp\left(-\frac{t}{\alpha} + \frac{\nu^2}{2\alpha^2}\right) \le \exp\left(-\frac{t}{2\alpha}\right).$$

Applying the argument to $-X$ proves the lower-tail assertion. □

When $\alpha = 0$, the Gaussian first regime applies for every $t \ge 0$. A zero value of ν forces $X = \mu$ almost surely, so the corresponding tail statement is immediate.

The two regimes in (7.2) explain the terminology: moderate deviations have Gaussian decay, while large deviations have exponential decay.

Definition 7.4 Let X have mean μ and variance σ^2. The **Bernstein condition with parameter b**, where $b \geq 0$, holds if, for every integer $k \geq 2$,

$$\mathbb{E}[|X-\mu|^k] \leq \frac{1}{2} k! \sigma^2 b^{k-2}. \tag{7.3}$$

Lemma 7.5 If X satisfies Definition 7.4 with $b > 0$, then, for $|\lambda| < 1/b$,

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq \exp\left(\frac{\sigma^2\lambda^2}{2(1-b|\lambda|)}\right). \tag{7.4}$$

In particular, X is subexponential with parameters $(\sqrt{2}\sigma, 2b)$.

Proof. The linear term in the exponential series vanishes. By (7.3),

$$\begin{aligned}
\mathbb{E}[e^{\lambda(X-\mu)}] &\leq 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \frac{|\lambda|^k \mathbb{E}[|X-\mu|^k]}{k!} \\
&\leq 1 + \frac{\lambda^2\sigma^2}{2} \left\{ 1 + \sum_{j=1}^{\infty} (b|\lambda|)^j \right\} \\
&= 1 + \frac{\lambda^2\sigma^2}{2(1-b|\lambda|)} \\
&\leq \exp\left(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}\right).
\end{aligned}$$

When $|\lambda| \leq 1/(2b)$, the denominator is at least $1/2$, so

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\sigma^2\lambda^2} = e^{(\sqrt{2}\sigma)^2\lambda^2/2}.$$

This proves the subexponential assertion. □

Proposition 7.6 (Bernstein's inequality) If X satisfies the Bernstein condition with $b > 0$, then, for every $t \geq 0$,

$$\mathbb{P}(X - \mu \geq t) \leq \exp\left(-\frac{t^2}{2(\sigma^2 + bt)}\right). \tag{7.5}$$

Consequently,

$$\mathbb{P}(|X-\mu| \geq t) \leq 2\exp\left(-\frac{t^2}{2(\sigma^2+bt)}\right).$$

Proof. For $0 < \lambda < 1/b$, combine Theorem 6.2 with (7.4):

$$\mathbb{P}(X-\mu \geq t) \leq \exp\left(-\lambda t + \frac{\sigma^2\lambda^2}{2(1-b\lambda)}\right).$$

The choice $\lambda = t/(\sigma^2+bt)$ is admissible and gives (7.5). Apply the same argument to $-X$ and use the union bound for the two-sided result. □

Example 7.7 Suppose that $|X-\mu| \leq b$ almost surely and $\mathrm{Var}(X) = \sigma^2$. Compare the Hoeffding and Bernstein bounds.

Solution. The following two facts hold:

1. Because $X-\mu \in [-b, b]$, Lemma 6.10 gives the sub-Gaussian variance proxy b^2.
2. For every integer $k \geq 2$,

$$\mathbb{E}[|X-\mu|^k] \leq b^{k-2}\mathbb{E}[(X-\mu)^2] \leq \frac{1}{2}k!\sigma^2 b^{k-2}.$$

Thus the Bernstein condition holds with parameter b.

It follows from (1) and (2) that

$$\mathbb{P}(|X-\mu| \geq t) \leq 2e^{-t^2/(2b^2)} \qquad \text{and} \qquad \mathbb{P}(|X-\mu| \geq t) \leq 2e^{-t^2/\{2(\sigma^2+bt)\}}.$$

The Hoeffding bound requires only the range. When $\sigma^2 \ll b^2$ and t is moderate, the Bernstein bound can be substantially sharper because it also uses the variance. □

Proposition 7.8 Let $X_1, \ldots, X_n$ be independent, let $\mu_i = \mathbb{E}[X_i]$, and suppose that X_i is

subexponential with parameters (ν_i, α_i). Define

$$\nu_*^2 := \sum_{i=1}^n \nu_i^2 \qquad \text{and} \qquad \alpha_* := \max_{1 \le i \le n} \alpha_i.$$

Then $\sum_{i=1}^n (X_i - \mu_i)$ is subexponential with parameters (ν_*, α_*). Moreover, for every $t \ge 0$,

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n (X_i - \mu_i) \right| \ge t \right) \le \begin{cases} 2\exp\left(-\dfrac{n^2 t^2}{2\nu_*^2} \right), & 0 \le t \le \dfrac{\nu_*^2}{n\alpha_*}, \\ 2\exp\left(-\dfrac{nt}{2\alpha_*} \right), & t > \dfrac{\nu_*^2}{n\alpha_*}. \end{cases} \tag{7.6}$$

provided $\nu_* > 0$ and $\alpha_* > 0$. If $\alpha_* = 0$, the first line applies for every $t \ge 0$; if $\nu_* = 0$, the sum is zero almost surely.

Proof. The claim is immediate if $\nu_* = 0$. If $\alpha_* = 0$, the variables are sub-Gaussian and the result follows from Theorem 6.11. Assume henceforth that $\nu_* > 0$ and $\alpha_* > 0$. For $|\lambda| \le 1/\alpha_*$, independence gives

$$\mathbb{E}\left[e^{\lambda \sum_{i=1}^n (X_i - \mu_i)} \right] = \prod_{i=1}^n \mathbb{E}[e^{\lambda(X_i - \mu_i)}] \le \prod_{i=1}^n e^{\nu_i^2 \lambda^2 / 2} = e^{\nu_*^2 \lambda^2 / 2}.$$

This proves the parameter claim. Apply Proposition 7.3 to the sum at the threshold nt and use the union bound to obtain (7.6). □

Example 7.9 Let $Y \sim \chi_n^2$. Show that, for $0 \le t \le 1$,

$$\mathbb{P}\left(\left| \frac{Y}{n} - 1 \right| \ge t \right) \le 2e^{-nt^2/8}.$$

Solution. Write $Y = \sum_{i=1}^n Z_i^2$, where $Z_1, \dots, Z_n$ are independent standard normal random variables. By Example 7.2, each $Z_i^2 - 1$ is subexponential with parameters $(2, 4)$. Hence Proposition 7.8 gives

$$\nu_*^2 = 4n \qquad \text{and} \qquad \alpha_* = 4.$$

Substitution into (7.6) proves the result. □

This chi-squared concentration bound is the main probabilistic ingredient in the random projection construction considered next.

Theorem 7.10 (Johnson–Lindenstrauss lemma) Let $\boldsymbol{u}^1, \dots, \boldsymbol{u}^N \in \mathbb{R}^d$, let $0 < \delta < 1$, and let $0 < \epsilon < 1$. Let $\mathbf{X} \in \mathbb{R}^{m \times d}$ have independent standard normal entries, and define the **random projection**

$$F(\boldsymbol{u}) \coloneqq \frac{\mathbf{X}\boldsymbol{u}}{\sqrt{m}}.$$

If

$$m \geq \frac{16}{\delta^2} \log\left(\frac{N}{\epsilon}\right), \tag{7.7}$$

then, with probability at least $1 - \epsilon$, every pair satisfies

$$(1-\delta)\left\|\boldsymbol{u}^i - \boldsymbol{u}^j\right\|_2^2 \leq \left\|F(\boldsymbol{u}^i) - F(\boldsymbol{u}^j)\right\|_2^2 \leq (1+\delta)\left\|\boldsymbol{u}^i - \boldsymbol{u}^j\right\|_2^2. \tag{7.8}$$

The map F is also called a **random sketch** or a **random linear embedding**.

The significance of Theorem 7.10 is that the target dimension depends logarithmically on the number of points and does not depend on the ambient dimension d. The geometry of this random projection is illustrated schematically in Figure 5 for the case $d = 3$ and $m = 2$. In general, a cloud of N points in $\mathbb{R}^d$ is embedded into $\mathbb{R}^m$ while all pairwise distances are preserved up to relative distortion δ.

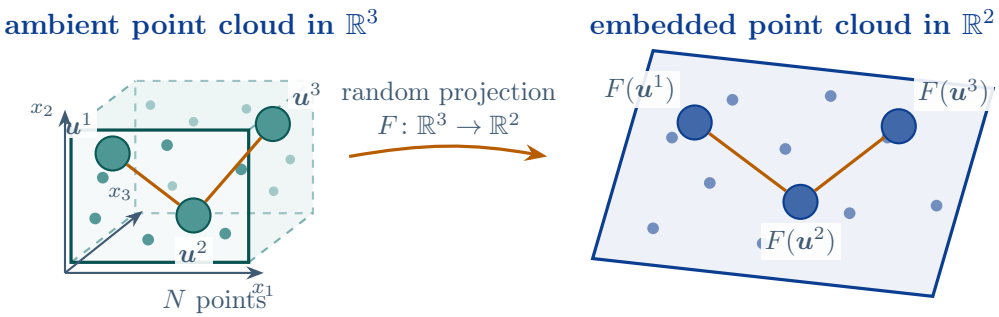


FIGURE 5

Proof. Let $\boldsymbol{X}_r^\top$ be the rth row of $\mathbf{X}$. For any fixed nonzero $\boldsymbol{u} \in \mathbb{R}^d$,

$$\left\langle \boldsymbol{X}_r, \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2} \right\rangle \sim \mathcal{N}(0,1),$$

independently over r. It follows that

$$\frac{\|\mathbf{X}\boldsymbol{u}\|_2^2}{\|\boldsymbol{u}\|_2^2} = \sum_{r=1}^{m} \left\langle \boldsymbol{X}_r, \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2} \right\rangle^2 \sim \chi^2_m. \tag{7.9}$$

By Example 7.9,

$$\mathbb{P}\left(\left| \frac{\|\mathbf{X}\boldsymbol{u}\|_2^2}{m\|\boldsymbol{u}\|_2^2} - 1 \right| \geq \delta \right) \leq 2e^{-m\delta^2/8}.$$

Since $\|F(\boldsymbol{u})\|_2^2 = \|\mathbf{X}\boldsymbol{u}\|_2^2/m$, this is equivalent to

$$\mathbb{P}\left(\frac{\|F(\boldsymbol{u})\|_2^2}{\|\boldsymbol{u}\|_2^2} \notin [1-\delta, 1+\delta] \right) \leq 2e^{-m\delta^2/8}.$$

For a distinct pair $\boldsymbol{u}^i, \boldsymbol{u}^j$, apply this bound to $\boldsymbol{u} = \boldsymbol{u}^i - \boldsymbol{u}^j$. Linearity gives $F(\boldsymbol{u}^i - \boldsymbol{u}^j) = F(\boldsymbol{u}^i) - F(\boldsymbol{u}^j)$. A union bound over the $\binom{N}{2}$ pairs shows that the probability of any failure of (7.8) is at most

$$2\binom{N}{2} e^{-m\delta^2/8} \leq N^2 e^{-m\delta^2/8}.$$

Condition (7.7) makes this quantity at most ϵ, which proves the result. If two points coincide, their distortion inequality holds trivially. □

The Johnson–Lindenstrauss lemma (Theorem 7.10) was introduced by Johnson and Lindenstrauss (1984). Matching lower bounds show that the order $\delta^{-2} \log N$ cannot generally be improved; see, for example, Larsen and Nelson's optimality result. Thus the sufficient dimension in (7.7) has the correct order up to universal constants.

Theorem 7.11 Let X be a mean-zero random variable. The following statements are equivalent, with constants allowed to change between statements:

1. There are parameters $\nu, \alpha \geq 0$ such that

$$\mathbb{E}[e^{\lambda X}] \leq e^{\nu^2\lambda^2/2}$$

whenever $|\lambda| \leq 1/\alpha$.

2. There are constants $c_1, c_2 > 0$ such that

$$\mathbb{P}(|X| \geq t) \leq c_1 e^{-c_2 t}$$

for every $t \geq 0$. For a symmetric variable, the natural two-sided constant may be taken to be twice the corresponding one-sided constant.

3. There is a constant $c > 0$ such that

$$\mathbb{E}[e^{c|X|}] < \infty.$$

Equivalently, the moment generating function is finite on a neighborhood of zero.

4. The quantity

$$\Upsilon := \sup_{k \geq 2} \left(\frac{\mathbb{E}[|X|^k]}{k!} \right)^{1/k}$$

is finite. This is the formulation underlying the Bernstein condition.

Proof. (1) *implies* (2). The two-regime bound in Proposition 7.3 is bounded above by $c_1 e^{-c_2 t}$ after adjusting c_1 to cover the bounded interval near zero. Applying the result to both X and $-X$ proves the two-sided statement.

(2) *implies* (4). For every integer $k \geq 2$, the layer-cake formula gives

$$\mathbb{E}[|X|^k] = k \int_0^\infty t^{k-1} \mathbb{P}(|X| \geq t) \, \mathrm{d}t \leq c_1 k \int_0^\infty t^{k-1} e^{-c_2 t} \, \mathrm{d}t = c_1 \frac{k!}{c_2^k}.$$

Taking kth roots proves that $\Upsilon < \infty$.

(4) *implies* (3). If $\Upsilon = 0$, then $\mathbb{E}[X^2] = 0$ and $X = 0$ almost surely. Otherwise, choose $0 < c < 1/\Upsilon$. The terms of orders at least two satisfy

$$\sum_{k=2}^\infty \frac{c^k \mathbb{E}[|X|^k]}{k!} \leq \sum_{k=2}^\infty (c\Upsilon)^k < \infty.$$

The terms of orders zero and one are finite because X is integrable. Hence $\mathbb{E}[e^{c|X|}] < \infty$.

(3) *implies* (1). Suppose that $\mathbb{E}[e^{c|X|}] < \infty$. For $|\lambda| \leq c/2$, the inequality

$$e^u \leq 1 + u + \frac{u^2}{2} e^{|u|}$$

and the identity $\mathbb{E}[X] = 0$ give

$$\mathbb{E}[e^{\lambda X}] \le 1 + \frac{\lambda^2}{2}\mathbb{E}[X^2 e^{c|X|/2}] \le \exp\left(\frac{\nu^2\lambda^2}{2}\right),$$

where $\nu^2 = \mathbb{E}[X^2 e^{c|X|/2}] < \infty$. Taking $\alpha = 2/c$ proves (1). □

Lecture 8 — Monday, March 11, 2019

Lipschitz Functions of Independent Gaussian Random Variables

Definition 8.1 A function $f : \mathbb{R}^n \to \mathbb{R}$ is ***L*-Lipschitz with respect to the Euclidean norm** if

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \le L\|\boldsymbol{x} - \boldsymbol{y}\|_2$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

Lemma 8.2 (Gaussian logarithmic Sobolev inequality) Let $\boldsymbol{G} \sim \mathcal{N}_n(\boldsymbol{0}, \mathbf{I}_n)$ and let $g : \mathbb{R}^n \to (0, \infty)$ be smooth. Then

$$\operatorname{Ent}(g(\boldsymbol{G})) \le \frac{1}{2}\mathbb{E}\left[\frac{\|\nabla g(\boldsymbol{G})\|_2^2}{g(\boldsymbol{G})}\right], \tag{8.1}$$

where

$$\operatorname{Ent}(U) := \mathbb{E}[U \log U] - \mathbb{E}[U] \log \mathbb{E}[U].$$

Proof. Let γ_n denote standard Gaussian measure and let

$$P_t g(\boldsymbol{x}) := \mathbb{E}\left[g\left(e^{-t}\boldsymbol{x} + \sqrt{1 - e^{-2t}}\boldsymbol{G}\right)\right]$$

be the Ornstein–Uhlenbeck semigroup. Gaussian measure is invariant under P_t, and $P_t g$ converges to $\mathbb{E}[g(\boldsymbol{G})]$ as $t \to \infty$. Gaussian integration by parts gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{Ent}_{\gamma_n}(P_t g) = -\mathbb{E}_{\gamma_n}\left[\frac{\|\nabla P_t g\|_2^2}{P_t g}\right].$$

The gradient commutation identity $\nabla P_t g = e^{-t} P_t(\nabla g)$ and the Cauchy–Schwarz inequality imply

$$\frac{\|\nabla P_t g\|_2^2}{P_t g} \le e^{-2t} P_t\left(\frac{\|\nabla g\|_2^2}{g}\right).$$

Integrating from zero to infinity and using invariance of γ_n yields

$$\operatorname{Ent}_{\gamma_n}(g) \le \int_0^\infty e^{-2t}\,\mathrm{d}t\,\mathbb{E}_{\gamma_n}\left[\frac{\|\nabla g\|_2^2}{g}\right] = \frac{1}{2}\mathbb{E}_{\gamma_n}\left[\frac{\|\nabla g\|_2^2}{g}\right].$$

□

Theorem 8.3 Let $\boldsymbol{G} = (G_1, \dots, G_n)$ have independent standard normal coordinates, and let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz for some $L > 0$. Then $f(\boldsymbol{G}) - \mathbb{E}[f(\boldsymbol{G})]$ is sub-Gaussian with variance proxy L^2. Equivalently, for every $\lambda \in \mathbb{R}$ and $t \ge 0$,

$$\mathbb{E}\left[e^{\lambda\{f(\boldsymbol{G}) - \mathbb{E}[f(\boldsymbol{G})]\}}\right] \le e^{\lambda^2 L^2/2}, \tag{8.2}$$

$$\mathbb{P}\left(|f(\boldsymbol{G}) - \mathbb{E}[f(\boldsymbol{G})]| \ge t\right) \le 2e^{-t^2/(2L^2)}. \tag{8.3}$$

Proof. First suppose that f is smooth. Lipschitz continuity implies $\|\nabla f\|_2 \le L$. Apply Lemma 8.2 to $g = e^{\lambda f}$. If $M(\lambda) = \mathbb{E}[e^{\lambda f(\boldsymbol{G})}]$, then

$$\lambda M'(\lambda) - M(\lambda)\log M(\lambda) = \operatorname{Ent}(e^{\lambda f(\boldsymbol{G})}) \le \frac{\lambda^2 L^2}{2} M(\lambda).$$

Set $\psi(\lambda) = \log \mathbb{E}[e^{\lambda\{f(\boldsymbol{G}) - \mathbb{E}[f(\boldsymbol{G})]\}}]$. After dividing by $M(\lambda)$, the preceding inequality becomes

$$\lambda\psi'(\lambda) - \psi(\lambda) \le \frac{\lambda^2 L^2}{2}.$$

For $\lambda > 0$,

$$\left(\frac{\psi(\lambda)}{\lambda}\right)' \le \frac{L^2}{2}.$$

Since $\psi(0) = \psi'(0) = 0$, integration from zero to λ proves (8.2) for $\lambda > 0$. Applying the same argument to $-f$ proves it for $\lambda < 0$. A general Lipschitz function is differentiable almost everywhere with gradient norm at most L; smoothing and truncation extend the result to this case. Finally, (8.3) follows from Lemma 6.5. □

Example 8.4 Let $Z_1, \dots, Z_n$ be independent standard normal random variables and let $Y = \sum_{i=1}^n Z_i^2 \sim \chi_n^2$. Use Gaussian concentration to bound the upper tail of Y/n.

Solution. Set $\boldsymbol{Z} = (Z_1, \dots, Z_n)^\top$ and $V = \|\boldsymbol{Z}\|_2/\sqrt{n} = \sqrt{Y/n}$. The map $\boldsymbol{z} \mapsto \|\boldsymbol{z}\|_2/\sqrt{n}$ is $1/\sqrt{n}$-Lipschitz. Therefore Theorem 8.3 gives

$$\mathbb{P}(V \geq \mathbb{E}[V] + \delta) \leq e^{-n\delta^2/2}.$$

Jensen's inequality yields

$$\mathbb{E}[V] \leq \sqrt{\mathbb{E}[V^2]} = \sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i^2]} = 1.$$

Hence, for every $\delta \geq 0$,

$$\mathbb{P}\left(\frac{Y}{n} \geq (1+\delta)^2\right) \leq e^{-n\delta^2/2}.$$

For $0 \leq \delta \leq 1$, one has $(1+\delta)^2 \leq 1 + 3\delta$. It follows that

$$\mathbb{P}\left(\frac{Y}{n} \geq 1 + 3\delta\right) \leq e^{-n\delta^2/2}.$$

Equivalently, for $0 \leq t \leq 3$,

$$\mathbb{P}\left(\frac{Y}{n} \geq 1 + t\right) \leq e^{-nt^2/18}.$$

□

Example 8.5 Let $G_1, \dots, G_n$ be independent standard normal random variables, ordered so that $G_{(1)} \geq \cdots \geq G_{(n)}$. Then, for every k and $\delta \geq 0$,

$$\mathbb{P}\left(|G_{(k)} - \mathbb{E}[G_{(k)}]| \geq \delta\right) \leq 2e^{-\delta^2/2}.$$

Solution. For vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, let $r = \|\boldsymbol{x} - \boldsymbol{y}\|_\infty$. Every coordinate of $\boldsymbol{x}$ lies within r of the corresponding coordinate of $\boldsymbol{y}$, so their kth order statistics satisfy

$$|x_{(k)} - y_{(k)}| \leq \|\boldsymbol{x} - \boldsymbol{y}\|_\infty \leq \|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

Thus the kth order statistic map is 1-Lipschitz. Apply Theorem 8.3 with $L = 1$. □

Example 8.6 Let $\boldsymbol{W} = (W_1, \dots, W_n)$ have independent standard normal coordinates. For $\mathcal{A} \subseteq \mathbb{R}^n$, define

$$Z(\mathcal{A}) \coloneqq \sup_{\boldsymbol{a} \in \mathcal{A}} \sum_{k=1}^{n} a_k W_k = \sup_{\boldsymbol{a} \in \mathcal{A}} \langle \boldsymbol{a}, \boldsymbol{W} \rangle.$$

Assume that $\mathcal{A}$ is nonempty, that this supremum is measurable, and that

$$\omega(\mathcal{A}) \coloneqq \sup_{\boldsymbol{a} \in \mathcal{A}} \|\boldsymbol{a}\|_2 < \infty.$$

Suppose also that $\omega(\mathcal{A}) > 0$. Then

$$\mathbb{P}\left(|Z(\mathcal{A}) - \mathbb{E}[Z(\mathcal{A})]| \geq \delta\right) \leq 2 \exp\left(-\frac{\delta^2}{2\omega(\mathcal{A})^2}\right).$$

Solution. Let $f(\boldsymbol{w}) = \sup_{\boldsymbol{a} \in \mathcal{A}} \langle \boldsymbol{a}, \boldsymbol{w} \rangle$. For $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^n$,

$$f(\boldsymbol{w}) - f(\boldsymbol{w}') \leq \sup_{\boldsymbol{a} \in \mathcal{A}} \langle \boldsymbol{a}, \boldsymbol{w} - \boldsymbol{w}' \rangle \leq \sup_{\boldsymbol{a} \in \mathcal{A}} \|\boldsymbol{a}\|_2 \left\|\boldsymbol{w} - \boldsymbol{w}'\right\|_2 = \omega(\mathcal{A}) \left\|\boldsymbol{w} - \boldsymbol{w}'\right\|_2.$$

Interchanging $\boldsymbol{w}$ and $\boldsymbol{w}'$ gives the same bound in the opposite direction, so f is $\omega(\mathcal{A})$-Lipschitz. The result follows from Theorem 8.3. □

The quantity $\omega(\mathcal{A})$ used here is the **Euclidean radius** of $\mathcal{A}$, while $\mathbb{E}[Z(\mathcal{A})]$ is its **Gaussian width**. The random supremum $Z(\mathcal{A})$ is a Gaussian analogue of Rademacher complexity.

3. Measure Concentration and Empirical Processes

Measure Concentration

Definition 8.7 Let ϕ be convex on an interval containing the range of an integrable random variable U. The **ϕ-entropy** of U is

$$H_\phi(U) \coloneqq \mathbb{E}[\phi(U)] - \phi(\mathbb{E}[U]).$$

Jensen's inequality gives $H_\phi(U) \geq 0$ whenever the displayed quantities are finite.

Example 8.8 The following choices connect ϕ-entropy with familiar quantities:

1. If $\phi(u) = u^2$, then $H_\phi(U) = \mathrm{Var}(U)$.
2. If $\phi(u) = -\log u$ for $u > 0$, then

$$H_\phi(e^{\lambda X}) = \log \mathbb{E}[e^{\lambda X}] - \lambda \mathbb{E}[X] = \log \mathbb{E}\left[e^{\lambda\{X - \mathbb{E}[X]\}}\right].$$

Thus this ϕ-entropy is the logarithm of the centered moment generating function.

Definition 8.9 For a nonnegative integrable random variable Z, define its **entropy functional** by

$$\mathrm{Ent}(Z) \coloneqq \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z],$$

with $0 \log 0$ interpreted as zero. This is the ϕ-entropy associated with

$$\phi(u) = \begin{cases} u \log u, & u > 0, \\ 0, & u = 0. \end{cases}$$

Remark 8.10 Let $M_X(\lambda) = \mathbb{E}[e^{\lambda X}]$. Direct differentiation gives

$$\mathrm{Ent}(e^{\lambda X}) = \lambda M_X'(\lambda) - M_X(\lambda) \log M_X(\lambda). \tag{8.4}$$

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$M_X(\lambda) = e^{\mu\lambda + \sigma^2\lambda^2/2} \qquad \text{and} \qquad M_X'(\lambda) = (\mu + \sigma^2\lambda) M_X(\lambda).$$

Substitution into (8.4) yields

$$\mathrm{Ent}(e^{\lambda X}) = \frac{\sigma^2\lambda^2}{2} M_X(\lambda). \tag{8.5}$$

The mean cancels, so the same identity holds for centered and noncentered normal variables.

Proposition 8.11 (Herbst argument) Suppose that, for $\lambda \geq 0$,

$$\mathrm{Ent}(e^{\lambda X}) \leq \frac{\sigma^2\lambda^2}{2} M_X(\lambda). \tag{8.6}$$

Then

$$\log \mathbb{E}\left[e^{\lambda\{X-\mathbb{E}[X]\}}\right] \leq \frac{\sigma^2\lambda^2}{2} \tag{8.7}$$

for $\lambda \geq 0$. If (8.6) holds for every $\lambda \in \mathbb{R}$, then (8.7) also holds for every $\lambda \in \mathbb{R}$, and X is sub-Gaussian with variance proxy σ^2.

Proof. Let $\psi(\lambda) = \log M_X(\lambda) - \lambda\mathbb{E}[X]$. By (8.4), dividing (8.6) by $M_X(\lambda)$ gives

$$\lambda\psi'(\lambda) - \psi(\lambda) \leq \frac{\sigma^2\lambda^2}{2}.$$

For $\lambda > 0$, this is equivalent to

$$\left(\frac{\psi(\lambda)}{\lambda}\right)' \leq \frac{\sigma^2}{2}.$$

Because $\psi(0) = \psi'(0) = 0$, integration from zero to λ yields $\psi(\lambda) \leq \sigma^2\lambda^2/2$, which is (8.7). For negative λ, repeat the argument on the interval from λ to zero, or apply the positive argument to $-X$. □

The entropy estimate is therefore converted into a moment generating function estimate, after which the Chernoff bound in Theorem 6.2 gives concentration. This conversion is the **Herbst argument**. Identity (8.5) is the model case.

Proposition 8.12 Suppose that $b, \sigma > 0$ and

$$\mathrm{Ent}(e^{\lambda X}) \leq \lambda^2\left[bM_X'(\lambda) + M_X(\lambda)\{\sigma^2 - b\mathbb{E}[X]\}\right] \tag{8.8}$$

for $0 \leq \lambda < 1/b$. Then

$$\log \mathbb{E}\left[e^{\lambda\{X-\mathbb{E}[X]\}}\right] \leq \frac{\sigma^2\lambda^2}{1-b\lambda}, \quad 0 \leq \lambda < 1/b. \tag{8.9}$$

Proof. Set

$$\psi(\lambda) \coloneqq \log M_X(\lambda) - \lambda\mathbb{E}[X].$$

After dividing (8.8) by $M_X(\lambda)$ and using (8.4), one obtains

$$\lambda\psi'(\lambda) - \psi(\lambda) \leq \lambda^2\{b\psi'(\lambda) + \sigma^2\}.$$

Consequently,

$$\left\{ \frac{(1-b\lambda)\psi(\lambda)}{\lambda} \right\}' = \frac{\lambda(1-b\lambda)\psi'(\lambda) - \psi(\lambda)}{\lambda^2} \leq \sigma^2.$$

The expression in braces tends to zero as $\lambda \downarrow 0$. Integrating therefore gives

$$\frac{(1-b\lambda)\psi(\lambda)}{\lambda} \leq \sigma^2 \lambda,$$

which is equivalent to (8.9). □

Separately Convex Functions and Random Variables

The entropy method also gives dimension-free concentration inequalities for functions of independent bounded coordinates. Convexity in each coordinate plays the role of Gaussian smoothness.

Definition 8.13 A function $f : \mathbb{R}^n \to \mathbb{R}$ is **separately convex** if, for each $k \in \{1, \ldots, n\}$ and every fixed choice of the other coordinates of $\boldsymbol{x} \in \mathbb{R}^n$, the function

$$x_k \longmapsto f(x'_1, \ldots, x'_{k-1}, x_k, x'_{k+1}, \ldots, x'_n)$$

is convex.

Lemma 8.14 Let $X_1, \ldots, X_n$ be independent and supported on $[a, b]$, and write $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$. If f is separately convex and L-Lipschitz with respect to the Euclidean norm, then, for every $\lambda \geq 0$,

$$\operatorname{Ent}(e^{\lambda f(\boldsymbol{X})}) \leq \lambda^2 L^2 (b-a)^2 \mathbb{E}[e^{\lambda f(\boldsymbol{X})}]. \tag{8.10}$$

Proof. Entropy tensorization for independent coordinates gives

$$\operatorname{Ent}(e^{\lambda f(\boldsymbol{X})}) \leq \mathbb{E}\left[\sum_{k=1}^{n} \operatorname{Ent}_{X_k}(e^{\lambda f(\boldsymbol{X})}) \right],$$

where the other coordinates are held fixed in the kth conditional entropy. For a convex function g on $[a, b]$, the one-dimensional convex entropy bound is

$$\operatorname{Ent}(e^{\lambda g(X_k)}) \leq \lambda^2 (b-a)^2 \mathbb{E}_{X_k}[e^{\lambda g(X_k)} s_k^2],$$

where the quantities s_k may be chosen as the coordinates of a subgradient of f. This bound follows

by replacing the law on $[a,b]$ by its two-endpoint convex majorant and applying the elementary two-point entropy inequality. Applying it coordinatewise and summing gives

$$\operatorname{Ent}(e^{\lambda f(\boldsymbol{X})}) \le \lambda^2 (b-a)^2 \mathbb{E}\left[e^{\lambda f(\boldsymbol{X})} \|\boldsymbol{s}(\boldsymbol{X})\|_2^2\right].$$

The Lipschitz property allows the subgradient to be chosen so that $\|\boldsymbol{s}(\boldsymbol{X})\|_2 \le L$. This proves (8.10). A standard smoothing argument removes the differentiability implicit in the subgradient notation. □

Theorem 8.15 Let $X_1, \dots, X_n$ be independent random variables supported on $[a,b]$, and write $\boldsymbol{X} = (X_1, \dots, X_n)^\top$. If $f : \mathbb{R}^n \to \mathbb{R}$ is separately convex and L-Lipschitz with respect to the Euclidean norm, then, for every $\delta > 0$,

$$\mathbb{P}\left(f(\boldsymbol{X}) \ge \mathbb{E}[f(\boldsymbol{X})] + \delta\right) \le \exp\left(-\frac{\delta^2}{4L^2(b-a)^2}\right). \tag{8.11}$$

Proof. Combine Lemma 8.14 with the Herbst argument in Proposition 8.11. This gives, for $\lambda \ge 0$,

$$\log \mathbb{E}\left[e^{\lambda\{f(\boldsymbol{X}) - \mathbb{E}[f(\boldsymbol{X})]\}}\right] \le \lambda^2 L^2 (b-a)^2.$$

The Chernoff bound in Theorem 6.2, optimized at $\lambda = \delta/\{2L^2(b-a)^2\}$, yields (8.11). □

Remark 8.16

1. The result is a bounded product analogue of Theorem 8.3.
2. Some structural assumption such as separate convexity is necessary for a dimension-free bound of this form.
3. Theorem 8.15 gives an upper tail bound. Stronger convex concentration results can give two-sided control under joint convexity and suitable assumptions on the product measure.
4. Analogous inequalities hold for broad classes of log-concave measures when the measure satisfies an appropriate Poincaré or logarithmic Sobolev inequality; log-concavity alone does not give the same universal constant.

Example 8.17 Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables, let $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$, and define

$$Z(\mathcal{A}) \coloneqq \sup_{\boldsymbol{a} \in \mathcal{A}} \langle \boldsymbol{a}, \boldsymbol{\varepsilon} \rangle \qquad \text{and} \qquad \omega(\mathcal{A}) \coloneqq \sup_{\boldsymbol{a} \in \mathcal{A}} \|\boldsymbol{a}\|_2.$$

Assume that $\mathcal{A}$ is nonempty, the displayed supremum is measurable, and $\omega(\mathcal{A}) > 0$. Then, for every $\delta > 0$,

$$\mathbb{P}\left(Z(\mathcal{A}) \geq \mathbb{E}[Z(\mathcal{A})] + \delta\right) \leq \exp\left(-\frac{\delta^2}{16\omega(\mathcal{A})^2}\right).$$

Solution. The function $f(\boldsymbol{x}) = \sup_{\boldsymbol{a} \in \mathcal{A}} \langle \boldsymbol{a}, \boldsymbol{x} \rangle$ is a supremum of linear functions and is therefore jointly convex. The calculation in Example 8.6 shows that it is $\omega(\mathcal{A})$-Lipschitz. Since each Rademacher coordinate lies in $[-1, 1]$, Theorem 8.15 applies with $b - a = 2$ and gives the stated bound. □

Remark 8.18 Changing only the kth Rademacher coordinate changes $Z(\mathcal{A})$ by at most $2 \sup_{\boldsymbol{a} \in \mathcal{A}} |a_k|$. The bounded differences inequality therefore gives the alternative martingale bound

$$\mathbb{P}\left(Z(\mathcal{A}) \geq \mathbb{E}[Z(\mathcal{A})] + \delta\right) \leq \exp\left(-\frac{\delta^2}{2\sum_{k=1}^n \sup_{\boldsymbol{a} \in \mathcal{A}} a_k^2}\right).$$

The denominators satisfy

$$\omega(\mathcal{A})^2 = \sup_{\boldsymbol{a} \in \mathcal{A}} \sum_{k=1}^n a_k^2 \leq \sum_{k=1}^n \sup_{\boldsymbol{a} \in \mathcal{A}} a_k^2.$$

The inequality can be strict by a large factor, so the entropy-based bound can be substantially sharper.

The transition from concentration inequalities to uniform laws of large numbers is summarized in Figure 6.

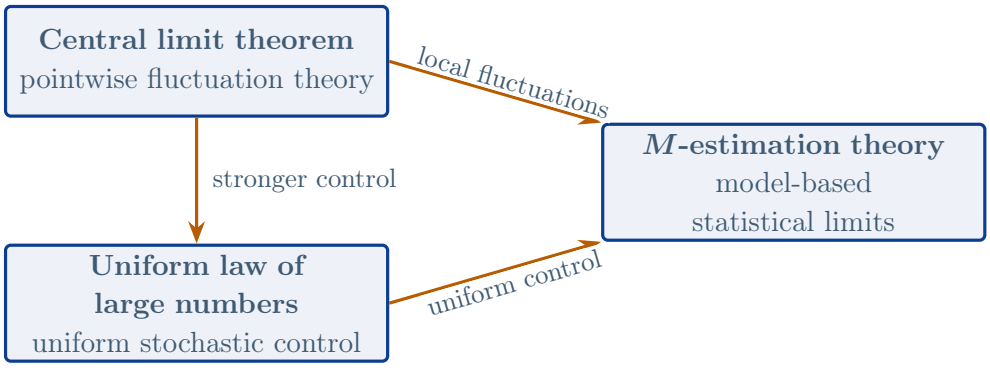


Figure 6

The central limit theorem supplies probabilistic fluctuations, while M-estimation combines these fluctuations with model structure. Uniform laws of large numbers provide the uniform control needed to pass from individual functions to entire model classes.

Uniform Laws of Large Numbers

Let $X_1, \ldots, X_n$ be independent and identically distributed copies of a random element X taking values in a measurable space $\mathcal{X}$, and let $\mathcal{F}$ be a class of measurable functions $f : \mathcal{X} \to \mathbb{R}$. A model $\{m_\theta : \theta \in \Theta\}$ from M-estimation is one important example. The maps have the form

$$(\Omega, \mathcal{A}, \mathbb{P}) \xrightarrow{X} \mathcal{X} \xrightarrow{f} \mathbb{R}.$$

Define

$$\mathbb{P}_n f := \frac{1}{n} \sum_{i=1}^{n} f(X_i), \qquad \mathbb{P} f := \mathbb{E}[f(X)], \qquad \text{and} \qquad \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - \mathbb{P} f|.$$

Thus the random variable $Z_{\mathcal{F}} = \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ measures the largest empirical error over the entire class. Two questions drive the analysis:

1. Does $Z_{\mathcal{F}}$ concentrate around its expectation?
2. Can the expectation, and hence $Z_{\mathcal{F}}$ itself, be bounded in terms of the complexity of $\mathcal{F}$ and the distribution of X?

Theorem 8.19 Suppose that $\mathcal{F}$ is nonempty, $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq b < \infty$, and the displayed

empirical supremum is measurable. Then, for every $\delta > 0$,

$$\mathbb{P}\left(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \geq \mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] + \delta\right) \leq e^{-n\delta^2/(2b^2)}. \tag{8.12}$$

The two-sided deviation from the expectation is bounded by $2e^{-n\delta^2/(2b^2)}$.

Proof. Let $g(x_1, \dots, x_n) = \sup_{f\in\mathcal{F}} |n^{-1}\sum_i f(x_i) - \mathbb{P}f|$. If $\boldsymbol{x}$ and $\boldsymbol{y}$ differ only in coordinate i, then

$$|g(\boldsymbol{x}) - g(\boldsymbol{y})| \leq \sup_{f\in\mathcal{F}} \frac{1}{n}|f(x_i) - f(y_i)| \leq \frac{2b}{n}.$$

The bounded differences inequality therefore gives

$$\mathbb{P}(g(X) - \mathbb{E}[g(X)] \geq \delta) \leq \exp\left(-\frac{2\delta^2}{n(2b/n)^2}\right) = e^{-n\delta^2/(2b^2)}.$$

Applying the same inequality to $-g$ and using the union bound gives the two-sided assertion. □

Lecture 9 — Monday, March 18, 2019

Definition 9.1 For fixed points $x^n = (x_1, \dots, x_n)$, define

$$\mathcal{F}(x^n) \coloneqq \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\} \subseteq \mathbb{R}^n.$$

If $\varepsilon_1, \dots, \varepsilon_n$ are independent Rademacher random variables, the **empirical Rademacher complexity** is

$$\widehat{\mathfrak{R}}_n(\mathcal{F}; x^n) \coloneqq \mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right|\right].$$

The **population Rademacher complexity** is

$$\mathfrak{R}_n(\mathcal{F}) \coloneqq \mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] = \mathbb{E}_X[\widehat{\mathfrak{R}}_n(\mathcal{F}; X^n)].$$

Theorem 9.2 For every function class $\mathcal{F}$ with an integrable envelope and measurable suprema,

$$\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] \le 2\mathfrak{R}_n(\mathcal{F}). \tag{9.1}$$

Proof. Let $X_1', \dots, X_n'$ be an independent copy of the sample and write $\mathbb{P}_n' f = n^{-1}\sum_i f(X_i')$. The conditional version of Jensen's inequality gives

$$\mathbb{E}_X[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] = \mathbb{E}_X\left[\sup_{f\in\mathcal{F}} \left|\mathbb{P}_n f - \mathbb{E}_{X'}[\mathbb{P}_n' f]\right|\right] \le \mathbb{E}_{X,X'}\left[\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - \mathbb{P}_n' f|\right].$$

Each pair (X_i, X_i') is exchangeable. Introducing independent Rademacher signs therefore does not change the last expectation:

$$\begin{aligned}
\mathbb{E}_{X,X'}\left[\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - \mathbb{P}_n' f|\right] &= \mathbb{E}_{X,X',\varepsilon}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\{f(X_i) - f(X_i')\}\right|\right] \\
&\le \mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] + \mathbb{E}_{X',\varepsilon}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i')\right|\right] \\
&= 2\mathfrak{R}_n(\mathcal{F}).
\end{aligned}$$

□

Combining Theorems 8.19 and 9.2 shows that, with probability at least $1-\eta$,

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \le 2\mathfrak{R}_n(\mathcal{F}) + b\sqrt{\frac{2\log(1/\eta)}{n}}.$$

For classes of indicator functions, the **Vapnik–Chervonenkis (VC) dimension** can be used to control the Rademacher complexity. Covering numbers and chaining provide tools for general real-valued classes.

Covering and Metric Entropy

Definition 9.3 A **pseudometric** on a set Θ is a map $\rho : \Theta \times \Theta \to [0, \infty)$ satisfying the following properties:

1. $\rho(x, x) = 0$ for every $x \in \Theta$.
2. $\rho(x, y) = \rho(y, x)$ for all $x, y \in \Theta$.
3. $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ for all $x, y, z \in \Theta$.

If, in addition, $\rho(x, y) = 0$ only when $x = y$, then ρ is a **metric**.

Definition 9.4 Let $T \subseteq \Theta$. A finite set $V = \{\theta^1, \ldots, \theta^N\} \subseteq \Theta$ is a ***δ*-cover** of T if, for every $\theta \in T$, there is an $i \in \{1, \ldots, N\}$ such that $\rho(\theta, \theta^i) \leq \delta$. The centers need not belong to T. The ***δ*-covering number** $N(\delta; T, \rho)$ is the smallest possible cardinality of such a cover, and

$$\log N(\delta; T, \rho)$$

is the **metric entropy**. A space is **totally bounded** precisely when its covering number is finite for every $\delta > 0$.

The map $\delta \mapsto N(\delta; T, \rho)$ is nonincreasing. For an infinite class it often diverges as $\delta \downarrow 0$, and its rate of growth measures the effective size of the class.

Example 9.5 Let $I = [-1, 1]$ with the usual distance. For $0 < \delta \leq 1$, show that

$$N(\delta; I, |\cdot|) \leq \left\lfloor \frac{1}{\delta} \right\rfloor + 1.$$

Solution. Set $L = \lfloor 1/\delta \rfloor + 1$ and use the centers $\theta^i = -1 + 2(i-1)\delta$ for $i = 1, \ldots, L$. The closed balls of radius δ about these evenly spaced centers cover $[-1, 1]$. Consequently,

$$N(\delta; I, |\cdot|) \leq L \leq \frac{1}{\delta} + 1.$$

Thus the covering number has polynomial growth as $\delta \downarrow 0$. □

Definition 9.6 A **δ-packing** of T is a set $\{\theta^1, \dots, \theta^M\} \subseteq T$ satisfying $\rho(\theta^i, \theta^j) > \delta$ whenever $i \neq j$. The **δ-packing number** $M(\delta; T, \rho)$ is the supremum of the cardinalities of all δ-packings.

The geometric distinction is shown in Figure 7. A cover asks whether radius-δ balls reach every point of T, while a packing asks whether the selected points remain more than δ apart.

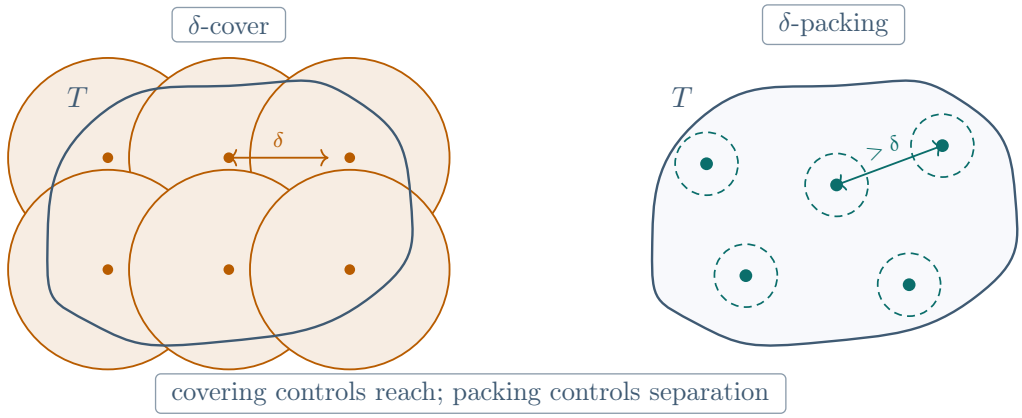


FIGURE 7

Lemma 9.7 For every $\delta > 0$,

$$M(2\delta; T, \rho) \leq N(\delta; T, \rho) \leq M(\delta; T, \rho). \tag{9.2}$$

Proof. For the second inequality, take a δ-packing that is maximal under inclusion. If it were not a δ-cover, there would be a point of T at distance greater than δ from every packing point. Adding that point would contradict maximality. Thus a maximal packing is a cover.

For the first inequality, let $\{x^1, \dots, x^M\}$ be a 2δ-packing and let $\{\theta^1, \dots, \theta^N\}$ be a δ-cover. If $M > N$, two packing points must lie in the same ball of the cover. The triangle inequality would then give

$$\rho(x^i, x^j) \leq \rho(x^i, \theta^k) + \rho(\theta^k, x^j) \leq 2\delta,$$

contradicting the packing property. □

Proposition 9.8 Let $\|\cdot\|$ be any norm on $\mathbb{R}^k$, and let $B_R = \{\boldsymbol{x} : \|\boldsymbol{x}\| \leq R\}$. For every

$\varepsilon > 0$,

$$\left(\frac{1}{\varepsilon}\right)^k \le N(\varepsilon R; B_R, \|\cdot\|) \le M(\varepsilon R; B_R, \|\cdot\|) \le \left(1 + \frac{2}{\varepsilon}\right)^k. \tag{9.3}$$

In particular, both numbers have order ε^{-k} as $\varepsilon \downarrow 0$.

Proof. Let $\{\theta^1, \dots, \theta^N\}$ be an εR-cover of B_R. Translation invariance and homogeneity of Lebesgue volume give

$$\operatorname{vol}(B_R) \le \sum_{i=1}^{N} \operatorname{vol}(B(\theta^i, \varepsilon R)) = N \operatorname{vol}(B_{\varepsilon R}).$$

Since $\operatorname{vol}(B_r) = r^k \operatorname{vol}(B_1)$, it follows that $N \ge \varepsilon^{-k}$.

The middle inequality is (9.2). For the last inequality, let $\{\boldsymbol{x}^1, \dots, \boldsymbol{x}^M\}$ be an εR-packing. The balls $B(\boldsymbol{x}^i, \varepsilon R/2)$ have disjoint interiors and their union lies in $B_{R+\varepsilon R/2}$. Therefore

$$M \left(\frac{\varepsilon R}{2}\right)^k \operatorname{vol}(B_1) \le \left(R + \frac{\varepsilon R}{2}\right)^k \operatorname{vol}(B_1),$$

which rearranges to the final inequality in (9.3). □

Example 9.9 Let $\mathbb{Q}$ be a probability measure on $\mathbb{R}^k$, let $B_R = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_2 \le R\}$, and consider

$$\mathcal{F} = \{\boldsymbol{x} \mapsto \boldsymbol{\beta}^\top \boldsymbol{x} : \boldsymbol{\beta} \in B_R\}.$$

Under the pseudometric

$$\rho(f_{\boldsymbol{\beta}}, f_{\boldsymbol{\gamma}}) = \left\{\mathbb{E}_{\mathbb{Q}}[\{(\boldsymbol{\beta} - \boldsymbol{\gamma})^\top \boldsymbol{X}\}^2]\right\}^{1/2},$$

the Cauchy–Schwarz inequality gives

$$\rho(f_{\boldsymbol{\beta}}, f_{\boldsymbol{\gamma}}) \le \left\{\mathbb{E}_{\mathbb{Q}}[\|\boldsymbol{X}\|_2^2]\right\}^{1/2} \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2.$$

Thus the metric entropy grows at most polynomially with exponent k.

Proposition 9.10 Let $\Theta \subseteq \mathbb{R}^k$ be nonempty and bounded, with Euclidean diameter D,

and let $\mathcal{F} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. Suppose that a nonnegative function Γ satisfies

$$|f_{\boldsymbol{\theta}_1}(x) - f_{\boldsymbol{\theta}_2}(x)| \le \Gamma(x)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 \tag{9.4}$$

for every $x \in \mathcal{X}$ and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$. Let $\mathbb{Q}$ be a probability measure on $\mathcal{X}$, define

$$\rho_{\mathbb{Q}}(f, g) \coloneqq \left\{\mathbb{E}_{\mathbb{Q}}[(f(X) - g(X))^2]\right\}^{1/2},$$

and assume $\|\Gamma\|_{\mathbb{Q},2} = \{\mathbb{E}_{\mathbb{Q}}[\Gamma(X)^2]\}^{1/2} < \infty$. Then

$$M(\varepsilon; \mathcal{F}, \rho_{\mathbb{Q}}) \le \left(1 + \frac{2D\|\Gamma\|_{\mathbb{Q},2}}{\varepsilon}\right)^k.$$

Proof. By (9.4),

$$\rho_{\mathbb{Q}}(f_{\boldsymbol{\theta}_1}, f_{\boldsymbol{\theta}_2}) \le \|\Gamma\|_{\mathbb{Q},2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2.$$

If $\|\Gamma\|_{\mathbb{Q},2} = 0$, all functions in $\mathcal{F}$ are identical in the pseudometric and the packing number is at most one, so the claim is immediate. Assume below that $\|\Gamma\|_{\mathbb{Q},2} > 0$. Hence every ε-separated subset of $\mathcal{F}$ corresponds to an $\varepsilon/\|\Gamma\|_{\mathbb{Q},2}$-separated subset of Θ. Fix $\boldsymbol{a} \in \Theta$. Since $\Theta \subseteq B(\boldsymbol{a}, D)$,

$$M(\varepsilon; \mathcal{F}, \rho_{\mathbb{Q}}) \le M\left(\frac{\varepsilon}{\|\Gamma\|_{\mathbb{Q},2}}; \Theta, \|\cdot\|_2\right) \le M\left(\frac{\varepsilon}{\|\Gamma\|_{\mathbb{Q},2}}; B(\boldsymbol{a}, D), \|\cdot\|_2\right) \le \left(1 + \frac{2D\|\Gamma\|_{\mathbb{Q},2}}{\varepsilon}\right)^k$$

by Proposition 9.8. □

The bound obtained by combining Theorems 8.19 and 9.2 holds for every sample size and is uniform over $f \in \mathcal{F}$. Covering numbers now provide a way to control the Rademacher complexity appearing in that bound.

Chaining

Definition 9.11 Let $T \subseteq \mathbb{R}^d$ and let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_d)$ have independent Rademacher coordinates. The **Rademacher process** indexed by T is

$$R_{\boldsymbol{\theta}} \coloneqq \langle \boldsymbol{\theta}, \boldsymbol{\varepsilon} \rangle, \quad \boldsymbol{\theta} \in T.$$

Its absolute expected supremum,

$$\mathcal{R}(T) \coloneqq \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in T} |R_{\boldsymbol{\theta}}|\right],$$

is the **Rademacher complexity** of T.

Definition 9.12 Let $\boldsymbol{g} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$. The **canonical Gaussian process** indexed by $T \subseteq \mathbb{R}^d$ is

$$G_{\boldsymbol{\theta}} \coloneqq \langle \boldsymbol{g}, \boldsymbol{\theta} \rangle, \quad \boldsymbol{\theta} \in T.$$

Its absolute expected supremum,

$$\mathcal{G}(T) \coloneqq \mathbb{E}\left[\sup_{\boldsymbol{\theta}\in T} |G_{\boldsymbol{\theta}}|\right],$$

is the **Gaussian complexity** of T. If T is centrally symmetric, the absolute values may be omitted without changing either complexity.

These processes recast the Rademacher complexity from Definition 9.1 as the expected supremum of a stochastic process.

Lemma 9.13 For every $T \subseteq \mathbb{R}^d$ with finite complexities,

$$\sqrt{\frac{2}{\pi}}\,\mathcal{R}(T) \le \mathcal{G}(T) \le \sqrt{2\log(2d)}\,\mathcal{R}(T). \tag{9.5}$$

Proof. Write $g_i = \eta_i \varepsilon_i$, where $\eta_i = |g_i|$ and the signs ε_i are independent Rademacher random variables, independent of $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d)$. Since a supremum of absolute values of linear functions is convex, the conditional version of Jensen's inequality gives

$$\mathcal{G}(T) = \mathbb{E}_{\boldsymbol{\eta},\boldsymbol{\varepsilon}}\left[\sup_{\boldsymbol{\theta}\in T}\left|\sum_{i=1}^{d} \eta_i \varepsilon_i \theta_i\right|\right] \ge \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{\boldsymbol{\theta}\in T}\left|\sum_{i=1}^{d} \mathbb{E}[\eta_i]\varepsilon_i \theta_i\right|\right] = \sqrt{\frac{2}{\pi}}\,\mathcal{R}(T).$$

For the upper bound, we establish the required comparison directly. Fix $\boldsymbol{\eta}$ and set $M = \|\boldsymbol{\eta}\|_\infty$. If $M = 0$, the comparison is immediate. Otherwise, let $S = T \cup (-T)$, so that, for any nonnegative coefficients $c_1, \ldots, c_d$,

$$\sup_{\boldsymbol{\theta}\in T}\left|\sum_{i=1}^{d} c_i \varepsilon_i \theta_i\right| = \sup_{\boldsymbol{\theta}\in S}\sum_{i=1}^{d} c_i \varepsilon_i \theta_i.$$

We compare the coefficients one coordinate at a time. Fix k and condition on all signs except ε_k. For

$$A_{\boldsymbol{\theta}} \coloneqq \sum_{i \neq k} c_i \varepsilon_i \theta_i, \quad \boldsymbol{\theta} \in S,$$

define

$$h_k(a) \coloneqq \frac{1}{2} \sup_{\boldsymbol{\theta} \in S} \bigl(A_{\boldsymbol{\theta}} + a\theta_k \bigr) + \frac{1}{2} \sup_{\boldsymbol{\theta} \in S} \bigl(A_{\boldsymbol{\theta}} - a\theta_k \bigr).$$

This is the conditional expectation over ε_k when the kth coefficient equals a. Each supremum in the definition of h_k is a convex function of a, and h_k is even. Consequently, h_k is nondecreasing on $[0, \infty)$. Thus, whenever $0 \le c_k \le M$, replacing c_k by M cannot decrease the conditional expected supremum. Beginning with $c_i = \eta_i$ and repeating this comparison for $k = 1, \dots, d$ gives

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \left[\sup_{\boldsymbol{\theta} \in T} \left| \sum_{i=1}^{d} \eta_i \varepsilon_i \theta_i \right| \right] \le M \mathbb{E}_{\boldsymbol{\varepsilon}} \left[\sup_{\boldsymbol{\theta} \in T} \left| \sum_{i=1}^{d} \varepsilon_i \theta_i \right| \right] = \|\boldsymbol{\eta}\|_\infty \mathcal{R}(T).$$

The maximum bound in Lemma 9.15 below, applied to $g_1, \dots, g_d$ and their negatives, yields $\mathbb{E}[\|\boldsymbol{\eta}\|_\infty] \le \sqrt{2 \log(2d)}$. Taking expectations proves the upper bound. □

Definition 9.14 A centered stochastic process $\{X_\theta : \theta \in T\}$ on a pseudometric space (T, ρ) is a **sub-Gaussian process with canonical pseudometric $\boldsymbol{\rho}$** if

$$\mathbb{E}\left[e^{\lambda (X_\theta - X_{\theta'})} \right] \le \exp\left(\frac{\lambda^2 \rho(\theta, \theta')^2}{2} \right) \tag{9.6}$$

for every $\lambda \in \mathbb{R}$ and $\theta, \theta' \in T$.

Both processes in Definitions 9.11 and 9.12 are sub-Gaussian under the Euclidean pseudometric because their increments are linear forms in independent sub-Gaussian coordinates.

Lemma 9.15 Let $X_1, \dots, X_m$ be mean-zero random variables, each sub-Gaussian with variance proxy σ^2. Independence is not required. Then

$$\mathbb{E}\left[\max_{1 \le i \le m} X_i \right] \le \sigma \sqrt{2 \log m}, \tag{9.7}$$

$$\mathbb{E}\left[\max_{1 \le i \le m} |X_i| \right] \le \sigma \sqrt{2 \log(2m)}. \tag{9.8}$$

Proof. If $\sigma = 0$, every X_i is zero almost surely. Assume $\sigma > 0$. When $m = 1$, the first inequality

is simply $\mathbb{E}[X_1] = 0$. For $m \geq 2$ and every $\lambda > 0$,

$$\mathbb{E}\left[\max_i X_i\right] \leq \frac{1}{\lambda} \log \mathbb{E}\left[e^{\lambda \max_i X_i}\right] \leq \frac{1}{\lambda} \log \sum_{i=1}^{m} \mathbb{E}[e^{\lambda X_i}] \leq \frac{\log m}{\lambda} + \frac{\lambda \sigma^2}{2}.$$

Optimizing at $\lambda = \sqrt{2\log(m)}/\sigma$ proves (9.7). Apply the same argument to the collection $X_1, \ldots, X_m, -X_1, \ldots, -X_m$ to obtain (9.8). □

For independent Gaussian variables, the first bound is asymptotically sharp: if $X_i \sim \mathcal{N}(0, \sigma^2)$ independently, then

$$\frac{\mathbb{E}[\max_{1 \leq i \leq m} X_i]}{\sigma\sqrt{2\log m}} \longrightarrow 1.$$

Proposition 9.16 Let $\{X_\theta : \theta \in T\}$ be a centered sub-Gaussian process with canonical pseudometric ρ, and suppose that $D = \operatorname{diam}(T, \rho) < \infty$. For $0 < \delta \leq D$,

$$\mathbb{E}\left[\sup_{\theta, \theta' \in T} (X_\theta - X_{\theta'})\right] \leq 2\mathbb{E}\left[\sup_{\substack{\tau, \tau' \in T \\ \rho(\tau, \tau') \leq \delta}} (X_\tau - X_{\tau'})\right] + 2D\sqrt{\log M(\delta; T, \rho)}. \tag{9.9}$$

Proof. Let $U = \{\phi_1, \ldots, \phi_m\}$ be a maximal δ-packing. By Lemma 9.7, it is also a δ-cover and $m \leq M(\delta; T, \rho)$. For every $\theta \in T$, choose $\pi(\theta) \in U$ with $\rho(\theta, \pi(\theta)) \leq \delta$. Then

$$\begin{aligned} X_\theta - X_{\theta'} &= \{X_\theta - X_{\pi(\theta)}\} + \{X_{\pi(\theta)} - X_{\pi(\theta')}\} + \{X_{\pi(\theta')} - X_{\theta'}\} \\ &\leq 2 \sup_{\substack{\tau, \tau' \in T \\ \rho(\tau, \tau') \leq \delta}} (X_\tau - X_{\tau'}) + \max_{\phi, \phi' \in U} (X_\phi - X_{\phi'}). \end{aligned}$$

There are at most m^2 increments in the last maximum, and each is sub-Gaussian with variance proxy at most D^2. Therefore Lemma 9.15 gives

$$\mathbb{E}\left[\max_{\phi, \phi' \in U} (X_\phi - X_{\phi'})\right] \leq D\sqrt{2\log(m^2)} = 2D\sqrt{\log m}.$$

Taking the supremum over θ, θ' and then expectations proves (9.9). □

Lecture 10 — Monday, March 25, 2019

Choosing a single resolution δ balances a local oscillation term against a finite net term. Chaining improves this strategy by using a nested sequence of increasingly fine nets.

For one index θ, Figure 8 shows the resulting telescoping path: each projection resolves a smaller amount of the remaining metric error.

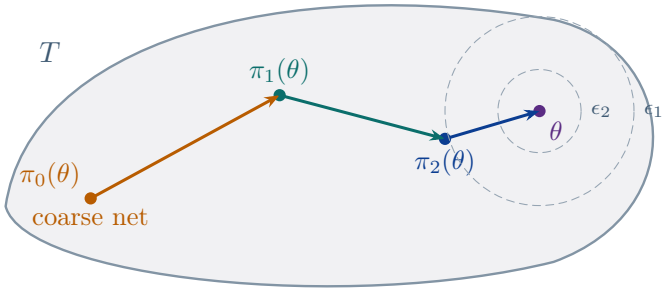


Figure 8

Definition 10.1 For $0 \le a < b \le \infty$, define the **Dudley entropy integral**

$$\mathcal{J}(a, b; T, \rho) := \int_a^b \sqrt{\log N(s; T, \rho)}\, \mathrm{d}s.$$

The Dudley entropy integral is useful precisely when the metric entropy is integrable over the indicated range; it need not be finite automatically.

Theorem 10.2 (Dudley's entropy bound) Let $\{X_\theta : \theta \in T\}$ be a centered, separable sub-Gaussian process with canonical pseudometric ρ, and let $D = \operatorname{diam}(T, \rho)$, where $0 < D < \infty$. There is a universal constant C such that, for $0 < \delta \le D$,

$$\mathbb{E}\left[\sup_{\theta,\theta' \in T} (X_\theta - X_{\theta'})\right] \le 2\mathbb{E}\left[\sup_{\substack{\tau,\tau' \in T \\ \rho(\tau,\tau') \le \delta}} (X_\tau - X_{\tau'})\right] + C\mathcal{J}(\delta/4, D/2; T, \rho). \tag{10.1}$$

If the local oscillation tends to zero as $\delta \downarrow 0$, then

$$\mathbb{E}\left[\sup_{\theta \in T} X_\theta\right] \leq C\mathcal{J}(0, D; T, \rho). \tag{10.2}$$

Proof. Choose scales $\epsilon_j = 2^{-j}D$ and fix $\theta_0 \in T$. Set $T_0 = \{\theta_0\}$. For $j \geq 1$, start with an $(\epsilon_j/2)$-cover whose centers may lie outside T, and replace each ball that meets T by one point of its intersection with T. The resulting set $T_j \subseteq T$ is an ϵ_j-net and satisfies

$$|T_j| \leq N(\epsilon_j/2; T, \rho).$$

This step is needed because the definition of a cover permits external centers, whereas the process is indexed only by points of T. Let $\pi_j(\theta)$ be a nearest net point to θ. For a truncation level J,

$$X_\theta - X_{\theta_0} = X_\theta - X_{\pi_J(\theta)} + \sum_{j=1}^{J} \{X_{\pi_j(\theta)} - X_{\pi_{j-1}(\theta)}\}.$$

The increment in the jth sum has pseudometric length at most $\epsilon_j + \epsilon_{j-1} = 3\epsilon_j$. There are at most $|T_j||T_{j-1}|$ possible increments at that level. Since $|T_{j-1}| \leq N(\epsilon_j; T, \rho) \leq N(\epsilon_j/2; T, \rho)$, their number is at most $N(\epsilon_j/2; T, \rho)^2$. By Lemma 9.15, their expected maximum is at most

$$C_0 \epsilon_j \sqrt{\log N(\epsilon_j/2; T, \rho)}$$

for a universal constant C_0. Taking the supremum over θ and summing over j gives

$$\mathbb{E}\left[\sup_{\theta \in T}(X_{\pi_J(\theta)} - X_{\theta_0})\right] \leq C_0 \sum_{j=1}^{J} \epsilon_j \sqrt{\log N(\epsilon_j/2; T, \rho)}.$$

Because the entropy is nonincreasing in the scale, this dyadic sum is bounded by a universal constant times

$$\int_{\epsilon_J/4}^{D/2} \sqrt{\log N(s; T, \rho)}\, \mathrm{d}s.$$

Apply the same decomposition to both endpoints of $X_\theta - X_{\theta'}$. Choose J so that ϵ_J is comparable to δ, and leave the two terminal increments as the local oscillation term. A harmless rescaling of the integration limits yields (10.1). Letting $\delta \downarrow 0$ proves (10.2) whenever the local oscillation vanishes. □

Unlike the one-step estimate, Dudley's bound (Theorem 10.2) accumulates the complexity of T

across all resolutions. If the global diameter or entropy integral is infinite, localized or truncated versions may still be useful, but chaining does not make an infinite integral finite by itself.

Example 10.3 For $\theta, x \in [0,1]$, let

$$f_\theta(x) = 1 - e^{-\theta x}$$

and set $\mathcal{P} = \{f_\theta : \theta \in [0,1]\}$. For a fixed sample $x^n = (x_1, \dots, x_n)$, compare one-step discretization and chaining bounds for the Gaussian and Rademacher complexities of

$$\mathcal{P}(x^n) = \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{P}\}.$$

Solution. The derivative with respect to θ satisfies $|\partial f_\theta(x)/\partial\theta| = xe^{-\theta x} \le 1$. Hence

$$\|f_\theta - f_{\theta'}\|_\infty \le |\theta - \theta'|.$$

A radius-δ cover of $[0,1]$ therefore induces a cover of $\mathcal{P}$, and

$$N(\delta; \mathcal{P}, \|\cdot\|_\infty) \le \frac{1}{2\delta} + 2. \tag{10.3}$$

Define the empirical norms

$$\|f - g\|_{2,n} \coloneqq \left\{ \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - g(x_i))^2 \right\}^{1/2} \qquad \text{and} \qquad \|f - g\|_{\infty,n} \coloneqq \max_{1 \le i \le n} |f(x_i) - g(x_i)|.$$

Since

$$\|f - g\|_{2,n} \le \|f - g\|_{\infty,n} \le \|f - g\|_\infty,$$

one has

$$N\left(\delta; \frac{\mathcal{P}(x^n)}{\sqrt{n}}, \|\cdot\|_2\right) \le N(\delta; \mathcal{P}, \|\cdot\|_\infty) \le \frac{1}{2\delta} + 2. \tag{10.4}$$

1. For the canonical Gaussian process,

$$\mathbb{E}\left[\sup_{\|\boldsymbol{\tau} - \boldsymbol{\tau}'\|_2 \le \delta} \langle \boldsymbol{g}, \boldsymbol{\tau} - \boldsymbol{\tau}' \rangle \right] \le \delta \mathbb{E}[\|\boldsymbol{g}\|_2] \le \delta\sqrt{n}.$$

The diameter of $\mathcal{P}(x^n)/\sqrt{n}$ is at most one. Consequently, Proposition 9.16 and (10.4) give

$$\mathcal{G}\left(\frac{\mathcal{P}(x^n)}{n}\right) \le \frac{2}{\sqrt{n}}\left\{\delta\sqrt{n} + \sqrt{\log\left(\frac{1}{2\delta}+2\right)}\right\}.$$

Taking $\delta = 1/(4\sqrt{n})$ yields

$$\mathcal{G}\left(\frac{\mathcal{P}(x^n)}{n}\right) \le C\sqrt{\frac{\log n}{n}} \tag{10.5}$$

for a universal constant C.

2. Dudley's entropy bound (Theorem 10.2) and (10.4) give

$$\mathcal{G}\left(\frac{\mathcal{P}(x^n)}{n}\right) = \frac{1}{\sqrt{n}}\mathcal{G}\left(\frac{\mathcal{P}(x^n)}{\sqrt{n}}\right) \le \frac{C}{\sqrt{n}}\int_0^{1/2}\sqrt{\log\left(\frac{1}{2s}+2\right)}\,\mathrm{d}s \le \frac{C'}{\sqrt{n}},$$

because the entropy integral is finite. This improves (10.5) by a factor of order $\sqrt{\log n}$.

Taking expectations with respect to the sample preserves the same bounds. Moreover, (9.5) implies

$$\mathfrak{R}_n(\mathcal{P}) \le \sqrt{\frac{\pi}{2}}\,\mathbb{E}_X\left[\mathcal{G}\left(\frac{\mathcal{P}(X^n)}{n}\right)\right] \le \frac{C''}{\sqrt{n}}.$$

Since $0 \le f_\theta \le 1$, combining Theorems 8.19 and 9.2 shows that, for every $\delta > 0$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{P}}|\mathbb{P}_n f - \mathbb{P}f| \ge \frac{2C''}{\sqrt{n}} + \delta\right) \le e^{-n\delta^2/2}.$$

This is a nonasymptotic uniform law of large numbers with the sharp $n^{-1/2}$ complexity rate for this class. □

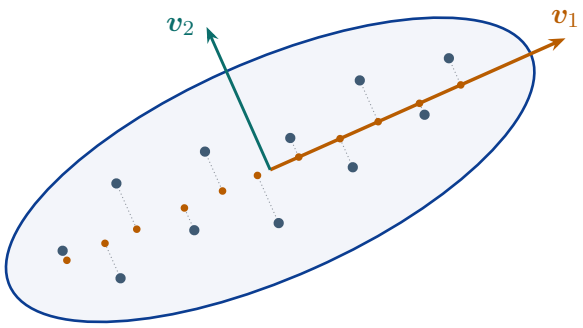
4. Principal Components and Sparse PCA

Principal Component Analysis

Let $\boldsymbol{X} \in \mathbb{R}^d$ be a random vector with mean zero and covariance matrix $\boldsymbol{\Sigma}$. **Principal component analysis** identifies directions of maximal variance.

In two dimensions, the leading principal direction is the major axis of the covariance ellipse, as

illustrated in Figure 9. Projecting onto this axis preserves more of the cloud's variation than projection onto an orthogonal direction.



the leading direction maximizes the variance of the projected coordinates

Figure 9

Definition 10.4 The **population principal directions** are defined recursively by

$$\boldsymbol{v}_k \in \operatorname*{argmax}_{\substack{\boldsymbol{v}\in S^{d-1} \\ \boldsymbol{v}\perp \boldsymbol{v}_1,\ldots,\boldsymbol{v}_{k-1}}} \boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v}, \quad k = 1, \ldots, d,$$

where the orthogonality constraint is vacuous when $k = 1$.

Thus $\boldsymbol{v}_k$ is an eigenvector associated with the kth largest eigenvalue of $\boldsymbol{\Sigma}$. When that eigenvalue is simple, $\boldsymbol{v}_k$ is identifiable up to its sign.

Definition 10.5 Suppose that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent copies of $\boldsymbol{X}$. The **empirical covariance matrix** is

$$\widehat{\boldsymbol{\Sigma}} := \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top.$$

The eigenvalues $\widehat{\lambda}_i$ and eigenvectors $\widehat{\boldsymbol{v}}_i$ of $\widehat{\boldsymbol{\Sigma}}$ estimate the corresponding population quantities. The two basic continuity questions are summarized in Figure 10:

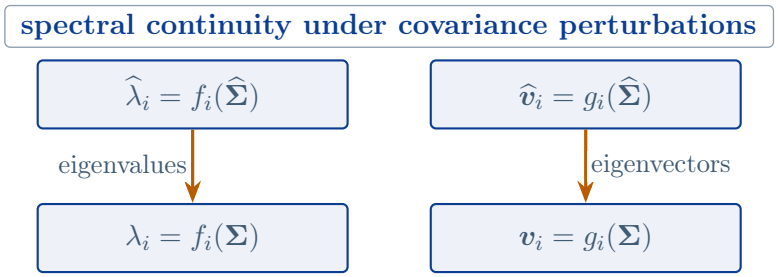


FIGURE 10

The required sample size depends both on the ambient dimension and on any additional structure, such as sparsity. Principal component analysis is used, among other applications, in ranking, mixture models, regression, manifold learning, and community detection.

Definition 10.6 The spaces of **symmetric matrices** and **positive semidefinite matrices** are, respectively,

$$\mathbb{S}^d := \{\mathbf{Q} \in \mathbb{R}^{d\times d} : \mathbf{Q} = \mathbf{Q}^\top\} \qquad \text{and} \qquad \mathbb{S}^d_+ := \{\mathbf{Q} \in \mathbb{S}^d : \mathbf{Q} \succeq 0\}.$$

For $\mathbf{Q} \in \mathbb{S}^d$, its **Rayleigh quotient** is

$$R_{\mathbf{Q}}(\boldsymbol{u}) := \frac{\boldsymbol{u}^\top \mathbf{Q} \boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{u}}, \quad \boldsymbol{u} \neq \mathbf{0}.$$

Both $\boldsymbol{\Sigma}$ and $\widehat{\boldsymbol{\Sigma}}$ belong to $\mathbb{S}^d_+$. For $\mathbf{Q} \in \mathbb{S}^d$, order its eigenvalues as

$$\gamma_{\max}(\mathbf{Q}) = \gamma_1(\mathbf{Q}) \geq \cdots \geq \gamma_d(\mathbf{Q}) = \gamma_{\min}(\mathbf{Q}).$$

Theorem 10.7 (Courant–Fischer variational characterization) Let $\mathbf{Q} \in \mathbb{S}^d$.

1. The extreme eigenvalues satisfy

$$\gamma_{\max}(\mathbf{Q}) = \max_{\boldsymbol{u}\neq\mathbf{0}} R_{\mathbf{Q}}(\boldsymbol{u}) \qquad \text{and} \qquad \gamma_{\min}(\mathbf{Q}) = \min_{\boldsymbol{u}\neq\mathbf{0}} R_{\mathbf{Q}}(\boldsymbol{u}).$$

2. For $k = 1, \dots, d$,

$$\gamma_k(\mathbf{Q}) = \min_{\substack{\mathcal{U} \subseteq \mathbb{R}^d \\ \dim(\mathcal{U}) = d-k+1}} \max_{\substack{\boldsymbol{u} \in \mathcal{U} \\ \boldsymbol{u} \neq \mathbf{0}}} R_{\mathbf{Q}}(\boldsymbol{u}) = \max_{\substack{\mathcal{U} \subseteq \mathbb{R}^d \\ \dim(\mathcal{U}) = k}} \min_{\substack{\boldsymbol{u} \in \mathcal{U} \\ \boldsymbol{u} \neq \mathbf{0}}} R_{\mathbf{Q}}(\boldsymbol{u}).$$

Proof. For (1), let $\boldsymbol{q}_1, \dots, \boldsymbol{q}_d$ be an orthonormal eigenbasis corresponding to $\gamma_1(\mathbf{Q}) \geq \cdots \geq \gamma_d(\mathbf{Q})$. If $\boldsymbol{u} = \sum_{j=1}^d a_j \boldsymbol{q}_j$ is nonzero, then

$$R_{\mathbf{Q}}(\boldsymbol{u}) = \frac{\sum_{j=1}^d \gamma_j(\mathbf{Q}) a_j^2}{\sum_{j=1}^d a_j^2}.$$

This representation proves (1).

For the first identity in (2), every subspace $\mathcal{U}$ of dimension $d - k + 1$ intersects $\operatorname{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_k)$ nontrivially. Consequently,

$$\max_{\boldsymbol{u} \in \mathcal{U} \setminus \{\mathbf{0}\}} R_{\mathbf{Q}}(\boldsymbol{u}) \geq \gamma_k(\mathbf{Q}).$$

Equality is attained by $\mathcal{U} = \operatorname{span}(\boldsymbol{q}_k, \dots, \boldsymbol{q}_d)$. Similarly, every k-dimensional subspace intersects $\operatorname{span}(\boldsymbol{q}_k, \dots, \boldsymbol{q}_d)$ nontrivially, while equality is attained by $\operatorname{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_k)$. □

The first statistical task is to control the operator norm error of the empirical covariance matrix. By (1),

$$\left\| \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \right\|_{\mathrm{op}} = \sup_{\boldsymbol{u} \in S^{d-1}} \left| \boldsymbol{u}^\top (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) \boldsymbol{u} \right| = \sup_{\boldsymbol{u} \in S^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n \langle \boldsymbol{X}_i, \boldsymbol{u} \rangle^2 - \boldsymbol{u}^\top \boldsymbol{\Sigma} \boldsymbol{u} \right|. \tag{10.6}$$

Thus covariance estimation is a uniform law of large numbers for the quadratic class $\boldsymbol{x} \mapsto \langle \boldsymbol{x}, \boldsymbol{u} \rangle^2$.

Definition 10.8 A mean-zero random vector $\boldsymbol{X} \in \mathbb{R}^d$ is **sub-Gaussian with variance proxy $\boldsymbol{\sigma^2}$**, where $\sigma^2 \geq 0$, if

$$\mathbb{E}\left[e^{\lambda \langle \boldsymbol{u}, \boldsymbol{X} \rangle} \right] \leq e^{\lambda^2 \sigma^2 / 2}$$

for every $\boldsymbol{u} \in S^{d-1}$ and every $\lambda \in \mathbb{R}$.

Example 10.9 The following are basic examples:

1. If $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ has independent standard normal coordinates, then $\boldsymbol{X}$ is sub-Gaussian with variance proxy 1.
2. If $\boldsymbol{X} \sim \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma})$, then $\boldsymbol{X}$ is sub-Gaussian with variance proxy $\|\boldsymbol{\Sigma}\|_{\mathrm{op}}$.

Indeed, $\langle \boldsymbol{u}, \boldsymbol{X} \rangle$ is normal with variance 1 in (1) and with variance $\boldsymbol{u}^\top \boldsymbol{\Sigma} \boldsymbol{u} \leq \|\boldsymbol{\Sigma}\|_{\mathrm{op}}$ in (2).

Theorem 10.10 Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent copies of a mean-zero random vector $\boldsymbol{X} \in \mathbb{R}^d$ that is sub-Gaussian with variance proxy $\sigma^2 > 0$, and let $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]$. There are universal positive constants c_0, c_1, c_2, c_3 for which the following statements hold.

1. For every λ satisfying $0 \leq \lambda < c_1 n/\sigma^2$,
$$\mathbb{E}\left[e^{\lambda \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{op}}}\right] \leq \exp\left(c_0 \frac{\lambda^2 \sigma^4}{n} + 4d\right). \tag{10.7}$$
2. For every $t \geq 0$,
$$\mathbb{P}\left(\frac{\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_{\mathrm{op}}}{\sigma^2} \geq c_1\left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right) + t\right) \leq c_2 e^{-c_3 n(t^2 \wedge t)}. \tag{10.8}$$

Proof. Set $\mathbf{Q} = \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}$. Choose a 1/8-net $\mathcal{N} = \{\boldsymbol{v}^1, \ldots, \boldsymbol{v}^N\}$ of S^{d-1} in Euclidean distance. By Proposition 9.8, it may be chosen so that $N \leq 17^d$. For $\boldsymbol{v} \in S^{d-1}$, choose $\boldsymbol{v}^j \in \mathcal{N}$ and write $\boldsymbol{v} = \boldsymbol{v}^j + \boldsymbol{\Delta}$, where $\|\boldsymbol{\Delta}\|_2 \leq 1/8$. Then
$$\begin{aligned}
\left|\boldsymbol{v}^\top \mathbf{Q} \boldsymbol{v}\right| &\leq \left|(\boldsymbol{v}^j)^\top \mathbf{Q} \boldsymbol{v}^j\right| + 2\left|\boldsymbol{\Delta}^\top \mathbf{Q} \boldsymbol{v}^j\right| + \left|\boldsymbol{\Delta}^\top \mathbf{Q} \boldsymbol{\Delta}\right| \\
&\leq \left|(\boldsymbol{v}^j)^\top \mathbf{Q} \boldsymbol{v}^j\right| + \left(2\|\boldsymbol{\Delta}\|_2 + \|\boldsymbol{\Delta}\|_2^2\right) \|\mathbf{Q}\|_{\mathrm{op}} \\
&\leq \left|(\boldsymbol{v}^j)^\top \mathbf{Q} \boldsymbol{v}^j\right| + \frac{17}{64}\|\mathbf{Q}\|_{\mathrm{op}}.
\end{aligned}$$
Taking the supremum over $\boldsymbol{v}$ and rearranging yields the convenient bound
$$\|\mathbf{Q}\|_{\mathrm{op}} \leq 2 \max_{1 \leq j \leq N} \left|(\boldsymbol{v}^j)^\top \mathbf{Q} \boldsymbol{v}^j\right|. \tag{10.9}$$

Fix $\boldsymbol{u} \in S^{d-1}$ and put $Y_i = \langle \boldsymbol{u}, \boldsymbol{X}_i \rangle^2 - \mathbb{E}[\langle \boldsymbol{u}, \boldsymbol{X}_i \rangle^2]$. The sub-Gaussian moment characterization in

Theorem 6.7 implies

$$(\mathbb{E}[|Y_i|^p])^{1/p} \leq Cp\sigma^2, \quad p \geq 1.$$

By the subexponential characterizations in Theorem 7.11, there are universal constants $C, c > 0$ such that

$$\mathbb{E}\left[\exp\left(t\,\boldsymbol{u}^\top \mathbf{Q}\boldsymbol{u}\right)\right] = \mathbb{E}\left[\exp\left(\frac{t}{n}\sum_{i=1}^n Y_i\right)\right] \leq \exp\left(C\frac{t^2\sigma^4}{n}\right) \tag{10.10}$$

whenever $|t| \leq cn/\sigma^2$.

For $\lambda \geq 0$, (10.9) and (10.10) give

$$\mathbb{E}[e^{\lambda\|\mathbf{Q}\|_{\mathrm{op}}}] \leq \sum_{j=1}^N \left\{\mathbb{E}[e^{2\lambda(\boldsymbol{v}^j)^\top \mathbf{Q}\boldsymbol{v}^j}] + \mathbb{E}[e^{-2\lambda(\boldsymbol{v}^j)^\top \mathbf{Q}\boldsymbol{v}^j}]\right\} \leq 2N\exp\left(C\frac{\lambda^2\sigma^4}{n}\right) \leq \exp\left(C\frac{\lambda^2\sigma^4}{n} + 4d\right),$$

for $\lambda \leq cn/\sigma^2$. This proves (1). Applying the Chernoff bound in Theorem 6.2 and optimizing over the permitted values of λ proves (2). □

Remark 10.11 If $\boldsymbol{\Sigma} = \mathbf{I}_d$ and $\sigma^2 = 1$, then

$$\left\|\widehat{\boldsymbol{\Sigma}} - \mathbf{I}_d\right\|_{\mathrm{op}} \lesssim \sqrt{\frac{d}{n}} + \frac{d}{n}$$

with high probability. In particular, when n is much larger than d,

$$1 - C\sqrt{\frac{d}{n}} \leq \lambda_{\min}(\widehat{\boldsymbol{\Sigma}}) \leq \lambda_{\max}(\widehat{\boldsymbol{\Sigma}}) \leq 1 + C\sqrt{\frac{d}{n}}$$

with high probability.

Lecture 11 — Monday, April 01, 2019

Definition 11.1 A covariance matrix $\boldsymbol{\Sigma} \in \mathbb{S}_+^d$ follows the **covariance model with one spike** if

$$\boldsymbol{\Sigma} = \mathbf{I}_d + \theta\boldsymbol{v}\boldsymbol{v}^\top,$$

where $\theta > 0$ and $\boldsymbol{v} \in S^{d-1}$. The vector $\boldsymbol{v}$ is the **signal eigenvector**, and θ is the **eigengap**

between the largest and second-largest eigenvalues.

The largest eigenvalue of $\boldsymbol{\Sigma}$ is $1+\theta$, with eigenvector $\boldsymbol{v}$, while all remaining eigenvalues equal 1. The sign of $\boldsymbol{v}$ is not identifiable from $\boldsymbol{\Sigma}$ because $\boldsymbol{v}\boldsymbol{v}^\top = (-\boldsymbol{v})(-\boldsymbol{v})^\top$.

Example 11.2 Let $Z \sim \mathcal{N}(0,1)$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$ be independent, and set

$$\boldsymbol{X} = \sqrt{\theta}\, Z\boldsymbol{v} + \boldsymbol{\varepsilon}.$$

Then $\mathbb{E}[\boldsymbol{X}] = \mathbf{0}$ and

$$\mathrm{Cov}(\boldsymbol{X}) = \theta\mathbb{E}[Z^2]\boldsymbol{v}\boldsymbol{v}^\top + \mathrm{Cov}(\boldsymbol{\varepsilon}) = \theta\boldsymbol{v}\boldsymbol{v}^\top + \mathbf{I}_d.$$

Thus $\boldsymbol{X}$ has the covariance matrix in Definition 11.1. A model with multiple spikes replaces $\theta\boldsymbol{v}\boldsymbol{v}^\top$ by a sum $\sum_{j=1}^{r} \theta_j \boldsymbol{v}_j \boldsymbol{v}_j^\top$ over orthonormal signal directions.

The natural estimator of $\boldsymbol{v}$ is a unit eigenvector $\widehat{\boldsymbol{v}}$ associated with the largest eigenvalue of $\widehat{\boldsymbol{\Sigma}}$. Its accuracy follows from a perturbation bound for eigenspaces.

Definition 11.3 For unit vectors $\boldsymbol{u}, \boldsymbol{v} \in S^{d-1}$, the **principal angle** between their one-dimensional spans is

$$\angle(\boldsymbol{u}, \boldsymbol{v}) \coloneqq \arccos\left|\boldsymbol{u}^\top \boldsymbol{v}\right| \in [0, \pi/2].$$

Definition 11.4 For a matrix $\mathbf{A}$ and $1 \le p < \infty$, its **Schatten p-norm** is

$$\|\mathbf{A}\|_{S_p} \coloneqq \left(\sum_{j} \sigma_j(\mathbf{A})^p\right)^{1/p},$$

where $\sigma_j(\mathbf{A})$ are the singular values. The limiting case is $\|\mathbf{A}\|_{S_\infty} = \max_j \sigma_j(\mathbf{A}) = \|\mathbf{A}\|_{\mathrm{op}}$, while $\|\mathbf{A}\|_{S_2} = \|\mathbf{A}\|_F$. The case $p = 1$ is the **nuclear norm**.

Lemma 11.5 (Von Neumann's trace inequality) Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, let $r = \min\{m, n\}$,

and arrange their singular values in nonincreasing order. Then

$$\left|\operatorname{tr}\left(\mathbf{A}^\top \mathbf{B}\right)\right| \le \sum_{i=1}^{r} \sigma_i(\mathbf{A})\sigma_i(\mathbf{B}). \tag{11.1}$$

Proof. Choose singular value decompositions

$$\mathbf{A} = \sum_{i=1}^{r} a_i \boldsymbol{u}_i \boldsymbol{v}_i^\top \qquad \text{and} \qquad \mathbf{B} = \sum_{j=1}^{r} b_j \boldsymbol{p}_j \boldsymbol{q}_j^\top,$$

where $a_i = \sigma_i(\mathbf{A})$ and $b_i = \sigma_i(\mathbf{B})$ are nonincreasing. Expanding the trace gives

$$\left|\operatorname{tr}\left(\mathbf{A}^\top \mathbf{B}\right)\right| \le \sum_{i,j=1}^{r} a_i b_j c_{ij} \qquad \text{and} \qquad c_{ij} \coloneqq \left|\boldsymbol{u}_i^\top \boldsymbol{p}_j\right| \left|\boldsymbol{v}_i^\top \boldsymbol{q}_j\right|.$$

For each i, the Cauchy–Schwarz inequality and orthonormality give

$$\sum_{j=1}^{r} c_{ij} \le \left(\sum_{j=1}^{r} (\boldsymbol{u}_i^\top \boldsymbol{p}_j)^2\right)^{1/2} \left(\sum_{j=1}^{r} (\boldsymbol{v}_i^\top \boldsymbol{q}_j)^2\right)^{1/2} \le 1.$$

The same argument with i and j interchanged shows that $\sum_{i=1}^{r} c_{ij} \le 1$ for every j. Thus $\mathbf{C} = (c_{ij})$ is **doubly substochastic**.

Set $z_i = \sum_{j=1}^{r} b_j c_{ij}$. For $1 \le k \le r$, let $s_j = \sum_{i=1}^{k} c_{ij}$. The row and column sum bounds imply that $0 \le s_j \le 1$ and $\sum_{j=1}^{r} s_j \le k$. Since the sequence $b_1, \dots, b_r$ is nonincreasing,

$$\sum_{i=1}^{k} z_i = \sum_{j=1}^{r} b_j s_j \le \sum_{j=1}^{k} b_j.$$

Taking $a_{r+1} = 0$ and applying summation by parts, we obtain

$$\sum_{i=1}^{r} a_i z_i = \sum_{k=1}^{r} (a_k - a_{k+1}) \sum_{i=1}^{k} z_i \le \sum_{k=1}^{r} (a_k - a_{k+1}) \sum_{i=1}^{k} b_i = \sum_{i=1}^{r} a_i b_i.$$

Combining this inequality with the trace expansion proves (11.1). □

Lemma 11.6 (Schatten Hölder inequality) If $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$ and $1 \le p, q \le \infty$ satisfy $1/p + 1/q = 1$, then

$$\left|\operatorname{tr}\left(\mathbf{A}^\top \mathbf{B}\right)\right| \le \|\mathbf{A}\|_{S_p} \|\mathbf{B}\|_{S_q}. \tag{11.2}$$

Proof. By Lemma 11.5,

$$\left|\operatorname{tr}\left(\mathbf{A}^\top \mathbf{B}\right)\right| \le \sum_j \sigma_j(\mathbf{A})\sigma_j(\mathbf{B}).$$

Applying the ordinary Hölder inequality to the two vectors of singular values proves (11.2). □

Theorem 11.7 (Davis–Kahan bound for leading eigenvectors) Let $\mathbf{A}, \mathbf{B} \in \mathbb{S}^d$ have ordered eigenpairs

$$(\lambda_1, \boldsymbol{u}_1), \dots, (\lambda_d, \boldsymbol{u}_d) \qquad \text{and} \qquad (\rho_1, \boldsymbol{v}_1), \dots, (\rho_d, \boldsymbol{v}_d),$$

respectively. Suppose that $\lambda_1 > \lambda_2$, and let $\boldsymbol{v}_1$ be any leading unit eigenvector of $\mathbf{B}$. Then

$$\sin \angle(\boldsymbol{u}_1, \boldsymbol{v}_1) \le \frac{2\|\mathbf{A} - \mathbf{B}\|_{\mathrm{op}}}{\lambda_1 - \lambda_2}. \tag{11.3}$$

If, in addition, $\rho_1 > \rho_2$, then the symmetric refinement is

$$\sin \angle(\boldsymbol{u}_1, \boldsymbol{v}_1) \le \frac{2\|\mathbf{A} - \mathbf{B}\|_{\mathrm{op}}}{\max\{\lambda_1 - \lambda_2, \rho_1 - \rho_2\}}. \tag{11.4}$$

Moreover, in either case,

$$\min_{\epsilon \in \{-1,1\}} \|\epsilon \boldsymbol{u}_1 - \boldsymbol{v}_1\|_2^2 \le 2 \sin^2 \angle(\boldsymbol{u}_1, \boldsymbol{v}_1). \tag{11.5}$$

Proof. Write $\alpha = \angle(\boldsymbol{u}_1, \boldsymbol{v}_1)$. The spectral decomposition of $\mathbf{A}$ gives, for every $\boldsymbol{x} \in S^{d-1}$,

$$\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x} = \lambda_1 (\boldsymbol{x}^\top \boldsymbol{u}_1)^2 + \sum_{j=2}^{d} \lambda_j (\boldsymbol{x}^\top \boldsymbol{u}_j)^2 \le \lambda_1 \cos^2 \angle(\boldsymbol{x}, \boldsymbol{u}_1) + \lambda_2 \sin^2 \angle(\boldsymbol{x}, \boldsymbol{u}_1).$$

Substituting $\boldsymbol{x} = \boldsymbol{v}_1$ and using $\boldsymbol{u}_1^\top \mathbf{A} \boldsymbol{u}_1 = \lambda_1$ yields

$$\boldsymbol{u}_1^\top \mathbf{A} \boldsymbol{u}_1 - \boldsymbol{v}_1^\top \mathbf{A} \boldsymbol{v}_1 \ge (\lambda_1 - \lambda_2) \sin^2 \alpha. \tag{11.6}$$

Because $\boldsymbol{v}_1$ maximizes $\boldsymbol{x}^\top \mathbf{B}\boldsymbol{x}$ over S^{d-1},

$$\boldsymbol{u}_1^\top \mathbf{A}\boldsymbol{u}_1 - \boldsymbol{v}_1^\top \mathbf{A}\boldsymbol{v}_1 \leq \boldsymbol{u}_1^\top(\mathbf{A}-\mathbf{B})\boldsymbol{u}_1 + \boldsymbol{v}_1^\top(\mathbf{B}-\mathbf{A})\boldsymbol{v}_1 = \left\langle \mathbf{A}-\mathbf{B}, \boldsymbol{u}_1\boldsymbol{u}_1^\top - \boldsymbol{v}_1\boldsymbol{v}_1^\top \right\rangle.$$

The matrix $\boldsymbol{u}_1\boldsymbol{u}_1^\top - \boldsymbol{v}_1\boldsymbol{v}_1^\top$ has rank at most two. Therefore, Lemma 11.6 and the inequality $\|\mathbf{M}\|_{S_1} \leq \sqrt{\operatorname{rank}(\mathbf{M})}\|\mathbf{M}\|_F$ imply

$$\begin{aligned}
\boldsymbol{u}_1^\top \mathbf{A}\boldsymbol{u}_1 - \boldsymbol{v}_1^\top \mathbf{A}\boldsymbol{v}_1 &\leq \|\mathbf{A}-\mathbf{B}\|_{\mathrm{op}} \left\| \boldsymbol{u}_1\boldsymbol{u}_1^\top - \boldsymbol{v}_1\boldsymbol{v}_1^\top \right\|_{S_1} \\
&\leq \sqrt{2}\,\|\mathbf{A}-\mathbf{B}\|_{\mathrm{op}} \left\| \boldsymbol{u}_1\boldsymbol{u}_1^\top - \boldsymbol{v}_1\boldsymbol{v}_1^\top \right\|_F \\
&= 2\|\mathbf{A}-\mathbf{B}\|_{\mathrm{op}} \sin\alpha. \qquad (11.7)
\end{aligned}$$

If $\sin\alpha = 0$, the first conclusion is immediate. Otherwise, combining (11.6) and (11.7) and cancelling $\sin\alpha$ gives

$$\sin\alpha \leq \frac{2\|\mathbf{A}-\mathbf{B}\|_{\mathrm{op}}}{\lambda_1 - \lambda_2}.$$

This proves (11.3). If $\rho_1 > \rho_2$, interchanging $\mathbf{A}$ and $\mathbf{B}$ gives the same inequality with $\rho_1 - \rho_2$ in the denominator. Taking the stronger of the two bounds proves (11.4).

Finally,

$$\min_{\epsilon\in\{-1,1\}} \|\epsilon\boldsymbol{u}_1 - \boldsymbol{v}_1\|_2^2 = 2 - 2\left|\boldsymbol{u}_1^\top \boldsymbol{v}_1\right| \leq 2 - 2(\boldsymbol{u}_1^\top \boldsymbol{v}_1)^2 = 2\sin^2\alpha,$$

which proves (11.5). □

Corollary 11.8 Assume the setting of Theorem 10.10, suppose that $\boldsymbol{\Sigma} = \mathbf{I}_d + \theta\boldsymbol{v}\boldsymbol{v}^\top$, and let $\widehat{\boldsymbol{v}}$ be a leading unit eigenvector of $\widehat{\boldsymbol{\Sigma}}$. If the sub-Gaussian variance proxy is $\sigma^2 = 1+\theta$, then, for every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\min_{\epsilon\in\{-1,1\}} \|\epsilon\widehat{\boldsymbol{v}} - \boldsymbol{v}\|_2 \leq C\frac{1+\theta}{\theta}\left\{\sqrt{\frac{d+\log(2/\delta)}{n}} + \frac{d+\log(2/\delta)}{n}\right\}, \qquad (11.8)$$

where C is a universal constant.

Proof. The population eigengap is θ. Applying Theorem 11.7 with $\mathbf{A} = \boldsymbol{\Sigma}$ and $\mathbf{B} = \widehat{\boldsymbol{\Sigma}}$, and then using (11.5), gives

$$\min_{\epsilon\in\{-1,1\}} \|\epsilon\widehat{\boldsymbol{v}} - \boldsymbol{v}\|_2 \leq \frac{2\sqrt{2}}{\theta}\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_{\mathrm{op}}.$$

Choose the deviation level in (10.8) so that its right-hand side is at most δ, and absorb universal constants to obtain (11.8). $\square$

Remark 11.9 If $(1+\theta)/\theta$ remains bounded and $d/n \to 0$, then (11.8) shows that $\angle(\widehat{\boldsymbol{v}}, \boldsymbol{v}) \xrightarrow{\mathbb{P}} 0$. When d is of the same order as or much larger than n, ordinary principal component analysis need not consistently estimate the signal direction. In a simple factor model, Johnstone and Lu (2009) showed that the ordinary leading principal component is consistent precisely when $d/n \to 0$. Sparsity can restore consistency in substantially higher dimensions.

Sparse Principal Component Analysis

When the ambient dimension d is much larger than the sample size n, ordinary principal component analysis cannot generally estimate the leading eigenvector consistently. A structural assumption can reduce the effective dimension. In the model with one spike, suppose that the signal eigenvector is s-sparse:

$$\|\boldsymbol{v}\|_0 := \#\{j \in \{1, \dots, d\} : v_j \neq 0\} \leq s.$$

This motivates the sparsity-constrained estimator

$$\widehat{\boldsymbol{v}}_s \in \operatorname*{argmax}_{\substack{\boldsymbol{u} \in S^{d-1} \\ \|\boldsymbol{u}\|_0 \leq s}} \boldsymbol{u}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{u}. \tag{11.9}$$

The optimization problem is nonconvex, but its statistical behavior displays the gain from sparsity particularly clearly.

The constraint in (11.9) is represented schematically in Figure 11. Ordinary principal component analysis may distribute weight across all d coordinates; sparse principal component analysis searches only among directions supported on at most s coordinates.

Theorem 11.10 Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be independent copies of a mean-zero sub-Gaussian random vector with variance proxy $1+\theta$ and covariance matrix

$$\boldsymbol{\Sigma} = \mathbf{I}_d + \theta \boldsymbol{v}\boldsymbol{v}^\top, \quad \theta > 0.$$

Suppose that $\|\boldsymbol{v}\|_0 \leq s \leq d/2$, and let $\widehat{\boldsymbol{v}}_s$ be any solution of (11.9). For every $\delta \in (0,1)$,

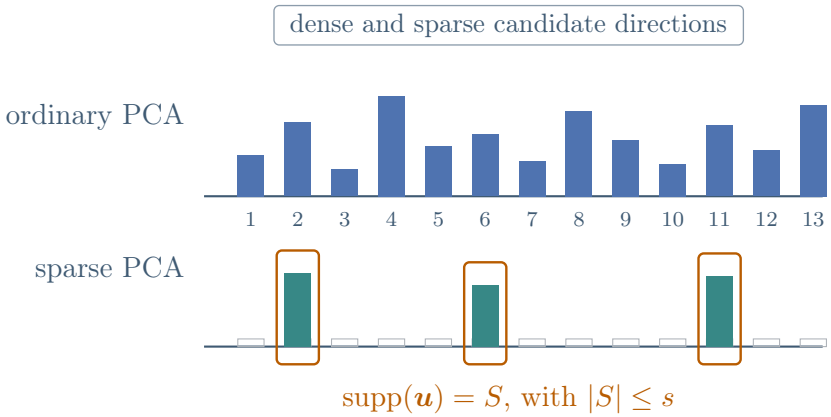


Figure 11

with probability at least $1-\delta$,

$$\min_{\epsilon\in\{-1,1\}} \|\epsilon\widehat{\boldsymbol{v}}_s - \boldsymbol{v}\|_2 \le C\frac{1+\theta}{\theta}\left\{\sqrt{\frac{s\log(ed/s)+\log(2/\delta)}{n}} + \frac{s\log(ed/s)+\log(2/\delta)}{n}\right\}, \tag{11.10}$$

where C is a universal constant.

Proof. For a set $A \subseteq \{1,\dots,d\}$, let $\boldsymbol{w}_A$ denote the restriction of a vector $\boldsymbol{w}$ to A, and let $\mathbf{M}_{AA}$ denote the corresponding principal submatrix of a matrix $\mathbf{M}$. Since $\widehat{\boldsymbol{v}}_s$ maximizes the empirical quadratic form and $\boldsymbol{v}$ is feasible,

$$\widehat{\boldsymbol{v}}_s^\top \widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{v}}_s \ge \boldsymbol{v}^\top \widehat{\boldsymbol{\Sigma}}\boldsymbol{v}.$$

Write $\alpha = \angle(\widehat{\boldsymbol{v}}_s, \boldsymbol{v})$. The curvature of the population objective around its maximizer is

$$\boldsymbol{v}^\top\boldsymbol{\Sigma}\boldsymbol{v} - \widehat{\boldsymbol{v}}_s^\top\boldsymbol{\Sigma}\widehat{\boldsymbol{v}}_s = \theta\left\{1-(\boldsymbol{v}^\top\widehat{\boldsymbol{v}}_s)^2\right\} = \theta\sin^2\alpha.$$

Combining these two observations gives

$$\theta\sin^2\alpha \le \left\langle \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}, \widehat{\boldsymbol{v}}_s\widehat{\boldsymbol{v}}_s^\top - \boldsymbol{v}\boldsymbol{v}^\top\right\rangle. \tag{11.11}$$

Let $S = \mathrm{supp}(\widehat{\boldsymbol{v}}_s)\cup\mathrm{supp}(\boldsymbol{v})$, so $|S| \le 2s$. The inner product in (11.11) depends only on the coordinates in S. The two projection matrices differ by a rank-two matrix whose nuclear norm equals $2\sin\alpha$. Hence

$$\theta\sin^2\alpha \le \left\|(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{SS}\right\|_{\mathrm{op}}\left\|(\widehat{\boldsymbol{v}}_s)_S(\widehat{\boldsymbol{v}}_s)_S^\top - \boldsymbol{v}_S\boldsymbol{v}_S^\top\right\|_{S_1} = 2\left\|(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{SS}\right\|_{\mathrm{op}}\sin\alpha.$$

It follows that

$$\sin\alpha \le \frac{2}{\theta}\max_{\substack{A\subseteq\{1,\dots,d\}\\ |A|=2s}} \left\|(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{AA}\right\|_{\mathrm{op}}. \tag{11.12}$$

Here a set with fewer than $2s$ elements may be enlarged to one with exactly $2s$ elements.

For each fixed A with $|A|=2s$, the restricted vectors $(\boldsymbol{X}_i)_A$ are sub-Gaussian with the same variance proxy. Applying Theorem 10.10 in dimension $2s$ and then taking a union bound over the $\binom{d}{2s}$ possible supports gives

$$\begin{aligned}
\mathbb{P}\left(\max_{|A|=2s}\left\|(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})_{AA}\right\|_{\mathrm{op}} > (1+\theta)\left[C\left\{\sqrt{\frac{s}{n}}+\frac{s}{n}\right\}+t\right]\right) &\\
\le c_2\binom{d}{2s}e^{-c_3 n(t^2\wedge t)} &\\
\le c_2\exp\left(2s\log\left(\frac{ed}{2s}\right)-c_3 n(t^2\wedge t)\right).&
\end{aligned}$$

Choosing

$$t = C'\left\{\sqrt{\frac{s\log(ed/s)+\log(2/\delta)}{n}}+\frac{s\log(ed/s)+\log(2/\delta)}{n}\right\}$$

makes the last probability at most δ. Finally, (11.5) and (11.12) yield (11.10). □

The rate in (11.10) depends on the effective dimension $s\log(ed/s)$ rather than on d. This agrees with the minimax theory for sparse leading eigenvectors developed by Vu and Lei.

Remark 11.11 A computationally tractable alternative estimates the rank-one projector $\boldsymbol{v}\boldsymbol{v}^\top$ by solving

$$\widehat{\boldsymbol{\Pi}}_\lambda \in \operatorname*{argmax}_{\substack{\boldsymbol{\Pi}\succeq 0\\ \operatorname{tr}(\boldsymbol{\Pi})=1}}\left\{\langle\widehat{\boldsymbol{\Sigma}},\boldsymbol{\Pi}\rangle-\lambda\|\boldsymbol{\Pi}\|_{1,1}\right\}, \tag{11.13}$$

where $\|\boldsymbol{\Pi}\|_{1,1}=\sum_{j,k}|\Pi_{jk}|$. The trace constraint is a convex relaxation of the rank-one normalization, and the entrywise ℓ_1 penalty encourages a sparse projector. This type of semidefinite relaxation was introduced by d'Aspremont et al. (2007).

5. Regression and Constrained Estimation

Regression Models

Let (X, Y) take values in $\mathcal{X} \times \mathcal{Y}$, where X is a covariate and $\mathcal{Y} \subseteq \mathbb{R}$. The covariate space $\mathcal{X}$ may be an abstract measurable space, although $\mathcal{X} \subseteq \mathbb{R}^d$ is the most common setting. Under squared error loss, the target of regression is defined as follows.

Definition 11.12 The **regression function** under squared error loss is

$$f^*(x) \coloneqq \mathbb{E}[Y \mid X = x], \quad x \in \mathcal{X},$$

where a suitable version of the conditional expectation is understood.

Linear Regression

Definition 11.13 The **fixed design linear model** is

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}. \tag{11.14}$$

Here $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$, $\boldsymbol{\theta}^* \in \mathbb{R}^d$, and the ith row of the deterministic **design matrix** $\mathbf{X} \in \mathbb{R}^{n \times d}$ is $\boldsymbol{x}_i^\top$. Equivalently,

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta}^* + \varepsilon_i, \quad i = 1, \ldots, n.$$

The mean-zero noise vector $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is assumed to be sub-Gaussian with variance proxy σ^2 in the sense of Definition 10.8.

Definition 11.14 A **least squares estimator** is any solution

$$\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\theta}\|_2^2. \tag{11.15}$$

Although the coefficient vector need not be unique when $\mathbf{X}$ is rank deficient, its fitted value $\mathbf{X}\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}$ is unique.

Definition 11.15 The **fixed design prediction loss** is

$$\mathcal{E}_{\mathbf{X}}(\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}, \boldsymbol{\theta}^*) := \frac{1}{n}\left\|\mathbf{X}(\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \boldsymbol{\theta}^*)\right\|_2^2 = (\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \boldsymbol{\theta}^*)^\top \left(\frac{\mathbf{X}^\top \mathbf{X}}{n}\right)(\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \boldsymbol{\theta}^*). \tag{11.16}$$

Theorem 11.16 Let (11.14) hold, and suppose that $\boldsymbol{\varepsilon}$ is sub-Gaussian with variance proxy σ^2. If $r = \mathrm{rank}(\mathbf{X})$, then

$$\mathbb{E}\left[\mathcal{E}_{\mathbf{X}}(\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}, \boldsymbol{\theta}^*)\right] \le \frac{\sigma^2 r}{n}. \tag{11.17}$$

If, in addition, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, equality holds in (11.17). Moreover, for every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\mathcal{E}_{\mathbf{X}}(\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}, \boldsymbol{\theta}^*) \le C\sigma^2 \frac{r + \log(2/\delta)}{n}, \tag{11.18}$$

where C is a universal constant.

Proof. Set $\boldsymbol{\Delta} = \widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \boldsymbol{\theta}^*$. The defining optimality inequality in (11.15) gives

$$\left\|\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}\right\|_2^2 \le \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\theta}^*\|_2^2.$$

Substituting (11.14) and expanding both sides yields

$$\|\mathbf{X}\boldsymbol{\Delta}\|_2^2 \le 2\boldsymbol{\varepsilon}^\top \mathbf{X}\boldsymbol{\Delta}. \tag{11.19}$$

If $r = 0$, the column space of $\mathbf{X}$ is $\{\mathbf{0}\}$, so the prediction loss is identically zero and every conclusion is immediate. Assume henceforth that $r \ge 1$.

Let $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_r) \in \mathbb{R}^{n \times r}$ have orthonormal columns spanning the column space of $\mathbf{X}$. There is a vector $\boldsymbol{a} \in \mathbb{R}^r$ such that $\mathbf{X}\boldsymbol{\Delta} = \boldsymbol{\Phi}\boldsymbol{a}$. If $\mathbf{X}\boldsymbol{\Delta} \neq \mathbf{0}$, then

$$\frac{\boldsymbol{\varepsilon}^\top \mathbf{X}\boldsymbol{\Delta}}{\|\mathbf{X}\boldsymbol{\Delta}\|_2} = \frac{\boldsymbol{\varepsilon}^\top \boldsymbol{\Phi}\boldsymbol{a}}{\|\boldsymbol{\Phi}\boldsymbol{a}\|_2} = \frac{(\boldsymbol{\Phi}^\top \boldsymbol{\varepsilon})^\top \boldsymbol{a}}{\|\boldsymbol{a}\|_2} \le \left\|\boldsymbol{\Phi}^\top \boldsymbol{\varepsilon}\right\|_2.$$

Consequently, (11.19) implies

$$\|\mathbf{X}\boldsymbol{\Delta}\|_2 \le 2\left\|\boldsymbol{\Phi}^\top \boldsymbol{\varepsilon}\right\|_2.$$

This is the **basic inequality argument** and remains useful for constrained estimators.

For ordinary least squares we can sharpen the identity. The fitted vector is the orthogonal projection of $\boldsymbol{Y}$ onto the column space of $\mathbf{X}$. Writing $\mathbf{P}_{\mathbf{X}} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\top}$ for this projection and observing that $\mathbf{X}\boldsymbol{\theta}^{*}$ already lies in that column space, we obtain

$$\mathbf{X}\boldsymbol{\Delta} = \mathbf{P}_{\mathbf{X}}\boldsymbol{\varepsilon} \qquad \text{and} \qquad \mathcal{E}_{\mathbf{X}}(\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}, \boldsymbol{\theta}^{*}) = \frac{1}{n}\left\|\boldsymbol{\Phi}^{\top}\boldsymbol{\varepsilon}\right\|_2^2. \tag{11.20}$$

Put $\boldsymbol{Z} = \boldsymbol{\Phi}^{\top}\boldsymbol{\varepsilon}$. Each coordinate $Z_j = \boldsymbol{\phi}_j^{\top}\boldsymbol{\varepsilon}$ is sub-Gaussian with variance proxy σ^2, so $\mathbb{E}[Z_j^2] \le \sigma^2$. Therefore,

$$\mathbb{E}[\|\boldsymbol{Z}\|_2^2] = \sum_{j=1}^{r} \mathbb{E}[Z_j^2] \le r\sigma^2,$$

which proves (11.17). For Gaussian noise, $\boldsymbol{Z} \sim \mathcal{N}_r(\mathbf{0}, \sigma^2\mathbf{I}_r)$, and the inequality is an equality.

If $\sigma = 0$, every coordinate of $\boldsymbol{Z}$ is zero almost surely, and the high-probability claim is immediate. Assume henceforth that $\sigma > 0$.

For completeness, choose a $1/2$-net $\mathcal{N}$ of S^{r-1}. By Proposition 9.8, it may be chosen with $|\mathcal{N}| \le 5^r$, and one-step discretization gives

$$\|\boldsymbol{Z}\|_2 \le 2 \max_{\boldsymbol{u} \in \mathcal{N}} \left|\boldsymbol{u}^{\top}\boldsymbol{Z}\right|.$$

For every $\boldsymbol{u} \in S^{r-1}$, the vector $\boldsymbol{\Phi}\boldsymbol{u}$ has Euclidean norm one, so $\boldsymbol{u}^{\top}\boldsymbol{Z} = (\boldsymbol{\Phi}\boldsymbol{u})^{\top}\boldsymbol{\varepsilon}$ is sub-Gaussian with variance proxy σ^2. A union bound and Lemma 6.5 therefore give

$$\mathbb{P}\left(\|\boldsymbol{Z}\|_2 \ge 2t\right) \le 2 \cdot 5^r e^{-t^2/(2\sigma^2)}.$$

Taking $t = \sigma\sqrt{2\{r\log 5 + \log(2/\delta)\}}$ and using (11.20) proves (11.18). $\square$

Remark 11.17 If a scalar random variable W is sub-Gaussian with variance proxy σ^2, then

$$\left(\mathbb{E}[|W|^k]\right)^{1/k} \le C\sigma\sqrt{k}, \quad k \ge 1.$$

This follows either by integrating the sub-Gaussian tail bound or from Theorem 6.7. The case $k = 2$ is the moment estimate used in the proof of Theorem 11.16.

Corollary 11.18 Suppose that $\mathbf{H} = \mathbf{X}^\top \mathbf{X}/n$ is positive definite. Under the assumptions of Theorem 11.16, with probability at least $1 - \delta$,

$$\left\| \widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \boldsymbol{\theta}^* \right\|_2^2 \le \frac{\mathcal{E}_{\mathbf{X}}(\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}, \boldsymbol{\theta}^*)}{\lambda_{\min}(\mathbf{H})} \le \frac{C\sigma^2}{\lambda_{\min}(\mathbf{H})} \frac{d + \log(2/\delta)}{n}. \tag{11.21}$$

Proof. For every $\boldsymbol{\Delta} \in \mathbb{R}^d$, $\boldsymbol{\Delta}^\top \mathbf{H} \boldsymbol{\Delta} \ge \lambda_{\min}(\mathbf{H}) \|\boldsymbol{\Delta}\|_2^2$. Apply this inequality with $\boldsymbol{\Delta} = \widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \boldsymbol{\theta}^*$ and then use (11.18). □

In particular, if $d/n \to 0$ and $\lambda_{\min}(\mathbf{H})$ is bounded away from zero, then $\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}^*$.

Constrained Least Squares

Now suppose that $\boldsymbol{\theta}^*$ is s-sparse, so $\|\boldsymbol{\theta}^*\|_0 \le s$. The direct **sparsity-constrained least squares estimator** is

$$\widehat{\boldsymbol{\theta}}_s \in \operatorname*{argmin}_{\substack{\boldsymbol{\theta} \in \mathbb{R}^d \\ \|\boldsymbol{\theta}\|_0 \le s}} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\theta}\|_2^2. \tag{11.22}$$

This optimization problem is nonconvex. A standard convex surrogate is the **least absolute shrinkage and selection operator**, usually called the **lasso**. Its constrained and penalized forms are

$$\widehat{\boldsymbol{\theta}}_{1,R} \in \operatorname*{argmin}_{\substack{\boldsymbol{\theta} \in \mathbb{R}^d \\ \|\boldsymbol{\theta}\|_1 \le R}} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\theta}\|_2^2, \tag{11.23}$$

$$\widehat{\boldsymbol{\theta}}_{1,\lambda} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \left\{ \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 \right\}. \tag{11.24}$$

For every penalized solution, taking $R = \left\| \widehat{\boldsymbol{\theta}}_{1,\lambda} \right\|_1$ makes it a constrained solution. Conversely, the Karush–Kuhn–Tucker conditions associate a multiplier $\lambda \ge 0$ with each constrained solution under the usual constraint qualification. Thus the formulations describe the same regularization path, although arbitrary preassigned values of R and λ need not correspond.

In two dimensions, the constrained geometry in Figure 12 explains why the ℓ^1 constraint promotes exact zeros: a loss contour can first touch the feasible diamond at a corner.

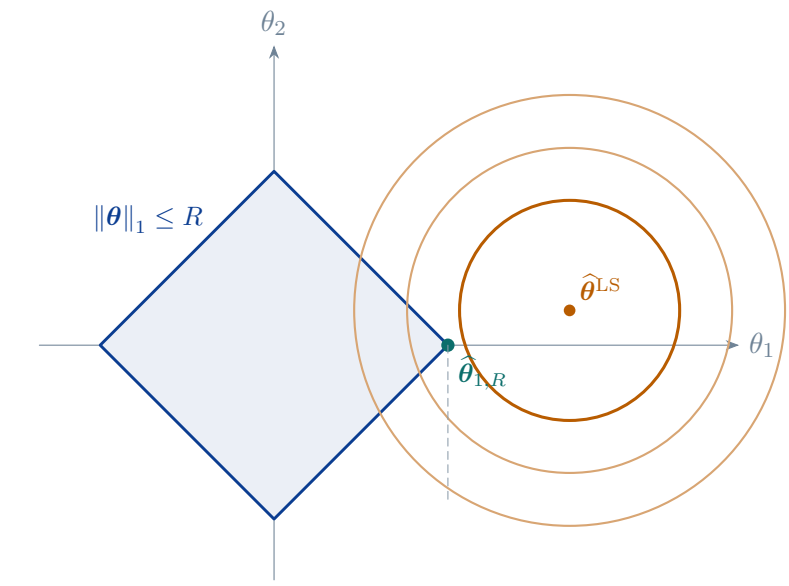


Figure 12

Lemma 11.19 For integers $d \ge 1$ and $1 \le k \le d$,

$$\binom{d}{k} \le \left(\frac{ed}{k}\right)^k. \tag{11.25}$$

Proof. Since the logarithm is increasing,

$$\log(k!) = \sum_{j=1}^{k} \log j \ge \int_1^k \log x \, \mathrm{d}x = k \log k - k + 1 \ge k \log k - k.$$

It follows that $k! \ge (k/e)^k$. Therefore,

$$\binom{d}{k} = \frac{d(d-1)\cdots(d-k+1)}{k!} \le \frac{d^k}{k!} \le \left(\frac{ed}{k}\right)^k,$$

which proves (11.25). □

Theorem 11.20 Suppose that (11.14) holds, $\|\boldsymbol{\theta}^*\|_0 \leq s$ for an integer $1 \leq s \leq d/2$, and $\boldsymbol{\varepsilon}$ is sub-Gaussian with variance proxy σ^2. For every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\mathcal{E}_{\mathbf{X}}(\widehat{\boldsymbol{\theta}}_s, \boldsymbol{\theta}^*) \leq C\frac{\sigma^2}{n}\left\{s + \log\binom{d}{2s} + \log(2/\delta)\right\} \leq C'\frac{\sigma^2}{n}\left\{s\log\left(\frac{ed}{s}\right) + \log(2/\delta)\right\}, \quad (11.26)$$

where C and C' are universal constants.

Proof. Set $\boldsymbol{\Delta} = \widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}^*$. Both vectors in this difference are s-sparse, so $\boldsymbol{\Delta}$ is supported on a set S with $|S| \leq 2s$. The optimality of $\widehat{\boldsymbol{\theta}}_s$ gives the same basic inequality as in (11.19):

$$\|\mathbf{X}\boldsymbol{\Delta}\|_2^2 \leq 2\boldsymbol{\varepsilon}^\top \mathbf{X}\boldsymbol{\Delta}.$$

For a set $A \subseteq \{1,\dots,d\}$, let $\mathbf{X}_A$ be the submatrix containing the columns indexed by A, and let $\mathbf{P}_A$ be the orthogonal projection onto the column space of $\mathbf{X}_A$. Since $\mathbf{X}\boldsymbol{\Delta}$ belongs to the column space of $\mathbf{X}_S$,

$$\boldsymbol{\varepsilon}^\top \mathbf{X}\boldsymbol{\Delta} = (\mathbf{P}_S\boldsymbol{\varepsilon})^\top \mathbf{X}\boldsymbol{\Delta} \leq \|\mathbf{P}_S\boldsymbol{\varepsilon}\|_2 \|\mathbf{X}\boldsymbol{\Delta}\|_2.$$

Therefore,

$$\|\mathbf{X}\boldsymbol{\Delta}\|_2^2 \leq 4 \max_{\substack{A \subseteq \{1,\dots,d\} \\ |A| = 2s}} \|\mathbf{P}_A\boldsymbol{\varepsilon}\|_2^2. \quad (11.27)$$

For a fixed A, the rank r_A of $\mathbf{P}_A$ is at most $2s$. Repeating the 1/2-net argument from the proof of Theorem 11.16 in this r_A-dimensional subspace shows that

$$\mathbb{P}\left(\|\mathbf{P}_A\boldsymbol{\varepsilon}\|_2^2 > C\sigma^2\{2s + t\}\right) \leq 2e^{-t}$$

for every $t \geq 0$, after adjusting the universal constant C. A union bound over the $\binom{d}{2s}$ possible supports, with

$$t = \log\left(\frac{2}{\delta}\binom{d}{2s}\right),$$

and (11.27) give the first bound in (11.26). Applying Lemma 11.19 with $k = 2s$ gives

$$\log\binom{d}{2s} \leq 2s\log\left(\frac{ed}{2s}\right) \leq 2s\log\left(\frac{ed}{s}\right).$$

Since $\log(ed/s) \geq 1$, the term s is absorbed into the same order, which proves the second bound. □

Appendix: Lecture and Tutorial Index

1. Lecture 1 — Monday, January 14, 2019
2. Lecture 2 — Monday, January 21, 2019
3. Lecture 3 — Monday, January 28, 2019
4. Lecture 4 — Monday, February 04, 2019
5. Lecture 5 — Monday, February 11, 2019
6. Lecture 6 — Monday, February 18, 2019
7. Lecture 7 — Monday, February 25, 2019
8. Lecture 8 — Monday, March 11, 2019
9. Lecture 9 — Monday, March 18, 2019
10. Lecture 10 — Monday, March 25, 2019
11. Lecture 11 — Monday, April 01, 2019