6 Rates of convergence of infinite dimensional parameters

If Θ is an infinite-dimensional set, such as a function space, then maximization of a criterion over the full space may not always be a good idea. For instance, consider fitting a function $\theta : [0, 1] \to \mathbb{R}$ to a set of observations $(z_1, Y_1), \ldots, (z_n, Y_n)$ by least squares, i.e., we minimize

$$\theta \mapsto \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \theta(z_i)\}^2.$$

If Θ consists of all functions $\theta : [0, 1] \to \mathbb{R}$, then obviously the minimum is 0, taken for any function that interpolates the data points exactly: $\theta(z_i) = Y_i$ for every $i = 1, \ldots, n$. This interpolation is typically not a good estimator, but *overfits* the data: it follows the given data exactly even though these probably contain error. The interpolation very likely gives a poor representation of the true regression function.

One way to rectify this problem is to consider minimization over a restricted class of functions. For example, the minimization can be carried out over all functions with 2 derivatives, which are bounded above by 10 throughout the interval; here the numbers 2 and (particularly) 10 are quite arbitrary. To prevent overfitting the size of the derivatives should not be too large, but can grow as we obtain more samples.

The *method of sieves* is an attempt to implement this. Sieves are subsets $\Theta_n \subset \Theta$, typically increasing in n, that can approximate any given function θ_0 that is considered likely to be "true". Given n observations the maximization is restricted to Θ_n, and as n increases this "sieve" is taken larger. In this section we extend the rate theorem in the previous section to sieved M-estimators, which include maximum likelihood estimators and least-squares estimators.

We also generalize the notation and other assumptions. In the next theorem the empirical criterion $\theta \mapsto \mathbb{P}_n m_\theta$ is replaced by a general stochastic process

$$\theta \mapsto \mathbb{M}_n(\theta).$$

It is then understood that each "estimator" $\hat{\theta}_n$ is a map defined on the same probability space as $\mathbb{M}_n$, with values in the index set Θ_n (which may be arbitrary set) of the process $\mathbb{M}_n$.

Corresponding to the criterion functions are *centering functions* $\theta \mapsto M_n(\theta)$ and "true parameters" $\theta_{n,0}$. These may be the mean functions of the processes $\mathbb{M}_n$ and their point of maximum, but this is not an assumption.

In this generality we also need not assume that Θ_n is a metric space, but measure the "discrepancy" or "distance" between θ and the true "value" $\theta_{n,0}$ by a map $\theta \mapsto d_n(\theta, \theta_{n,0})$ from Θ_n to $[0, \infty)$.

Theorem 6.1 (Rate of convergence). *For each n, let $\mathbb{M}_n$ and M_n be stochastic processes indexed by a set $\Theta_n \cup \{\theta_{n,0}\}$, and let $\theta \mapsto d_n(\theta, \theta_{n,0})$ be an arbitrary map from Θ_n to $[0, \infty)$.*

Let $\tilde{\delta}_n \geq 0$ and suppose that, for every n and $\delta > \tilde{\delta}_n$,

$$\sup_{\theta \in \Theta_n : \delta/2 < d_n(\theta, \theta_{n,0}) \leq \delta} [M_n(\theta) - M_n(\theta_{n,0})] \leq -c\delta^2, \tag{52}$$

for some $c > 0$ (for all $n \geq 1$) and

$$\mathbb{E}\left[\sup_{\theta \in \Theta_n : d_n(\theta, \theta_{n,0}) \leq \delta} \sqrt{n}\left|(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_{n,0})\right|\right] \lesssim \phi_n(\delta),$$

for increasing functions $\phi_n : [\tilde{\delta}_n, \infty) \to \mathbb{R}$ such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $0 < \alpha < 2$. Let $\theta_n \in \Theta_n$ and let δ_n satisfy

$$\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2, \qquad \delta_n^2 \geq M_n(\theta_{n,0}) - M_n(\theta_n), \qquad \delta_n \geq \tilde{\delta}_n.$$

If the sequence $\hat{\theta}_n$ takes values in Θ_n and satisfies $\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_n) - O_{\mathbb{P}}(\delta_n^2)$, then

$$d_n(\hat{\theta}_n, \theta_{n,0}) = O_{\mathbb{P}}(\delta_n).$$

Exercise (HW2): Complete the proof. Hint: The proof is similar to that of the previous rate theorem. That all entities are now allowed to depend on n asks for notational changes only, but the possible discrepancy between θ_n and $\theta_{n,0}$ requires some care.

The theorem can be applied with $\hat{\theta}_n$ and $\theta_{n,0}$ equal to the maximizers of $\theta \mapsto \mathbb{M}_n(\theta)$ over a sieve Θ_n and of $\theta \mapsto M_n(\theta)$ over a full parameter set Θ, respectively. Then (52) requires that the centering functions fall off quadratically in the "distance" $d_n(\theta, \theta_{n,0})$ as θ moves away from the maximizing value $\theta_{n,0}$. We use $\tilde{\delta}_n = 0$, and the theorem shows that the "distance" of $\hat{\theta}_n$ to $\theta_{n,0}$ satisfies

$$d_n^2(\hat{\theta}_n, \theta_{n,0}) = O_{\mathbb{P}}(\delta_n^2 + M_n(\theta_{n,0}) - M_n(\theta_n)), \tag{53}$$

for δ_n solving $\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2$ and for any $\theta_n \in \Theta_n$. Thus the rate δ_n is determined by the "modulus of continuity" $\delta \mapsto \phi_n(\delta)$ of the centered processes $\sqrt{n}(\mathbb{M}_n - M_n)$ over Θ_n and the discrepancy $M_n(\theta_{n,0}) - M_n(\theta_n)$. The latter vanishes if $\theta_n = \theta_{n,0}$ but this choice of θ_n may not be admissible (as θ_n must be an element of the sieve and $\theta_{n,0}$ need not). A natural choice of θ_n is to take θ_n as the closest element to $\theta_{n,0}$ in Θ_n, e.g., $\theta_n := \operatorname{argmin}_{\theta \in \Theta_n} d_n(\theta, \theta_{n,0})$.

Typically, small sieves Θ_n lead to a small modulus, hence fast δ_n in (53). On the other hand, the discrepancy $M_n(\theta_{n,0}) - M_n(\theta_n)$ of a small sieve will be large. Thus, the two terms in the right side of (53) may be loosely understood as a "variance" and a "squared bias" term, which must be balanced to obtain a good rate of convergence. We note that in many problems an un-sieved M-estimator actually performs well, so the trade-off should not be understood too literally: it may work well to reduce the "bias" to zero.

6.1 Least squares regression on sieves

Suppose that we have data

$$Y_i = \theta_0(z_i) + \epsilon_i, \qquad \text{for } i = 1, \ldots, n, \tag{54}$$

where $Y_i \in \mathbb{R}$ is the observed response variable, $z_i \in \mathcal{Z}$ is a covariate, and ϵ_i is the unobserved error. The errors are assumed to be independent random variables with expectation $\mathbb{E}\epsilon_i = 0$ and variance $\mathrm{Var}(\epsilon_i) \leq \sigma_0^2 < \infty$, for $i = 1, \ldots, n$. The covariates $z_1, \ldots, z_n$ are fixed, i.e., we consider the case of fixed design. The function $\theta_0 : \mathcal{Z} \to \mathbb{R}$ is unknown, but we assume that $\theta_0 \in \Theta$, where Θ is a given class of regression functions.

The unknown regression function can be estimated by the *sieved-least squares estimator* (LSE) $\hat{\theta}_n$, which is defined (not necessarily uniquely) by

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta_n} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \theta(z_i))^2,$$

where Θ_n is a set of regression functions $\theta : \mathcal{Z} \to \mathbb{R}$. Inserting the expression for Y_i and calculating the square, we see that $\hat{\theta}_n$ maximizes

$$\mathbb{M}_n(\theta) = \frac{2}{n} \sum_{i=1}^{n} (\theta - \theta_0)(z_i)\epsilon_i - \mathbb{P}_n(\theta - \theta_0)^2,$$

where $\mathbb{P}_n$ is the empirical measure on the design points $z_1, \ldots, z_n$. This criterion function is not observable but is of simpler character than the sum of squares. Note that the second term is assumed non-random, the randomness solely residing in the error terms.

Under the assumption that the error variables have mean zero, the mean of $\mathbb{M}_n(\theta)$ is $M_n(\theta) = -\mathbb{P}_n(\theta - \theta_0)^2$ and can be used as a centering function. It satisfies, for every θ,

$$M_n(\theta) - M_n(\theta_0) = -\mathbb{P}_n(\theta - \theta_0)^2.$$

Thus, Theorem 6.1 applies with $d_n(\theta, \theta_0)$ equal to the $L_2(\mathbb{P}_n)$-distance on the set of regression functions. The modulus of continuity condition takes the form

$$\phi_n(\delta) \geq \mathbb{E} \sup_{\mathbb{P}_n(\theta - \theta_0)^2 \leq \delta^2, \theta \in \Theta_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\theta - \theta_0)(z_i)\epsilon_i \right|. \tag{55}$$

Theorem 6.2. *If $Y_1, \ldots, Y_n$ are independent random variables satisfying (16) for fixed design points $z_1, \ldots, z_n$ and errors $\epsilon_1, \ldots, \epsilon_n$ with mean 0, then the minimizer $\hat{\theta}_n$ over Θ_n of the least squares criterion satisfies*

$$\|\hat{\theta}_n - \theta_0\|_{\mathbb{P}_n, 2} = O_{\mathbb{P}}(\delta_n)$$

for δ_n satisfying $\delta_n \geq \|\theta_0 - \Theta_n\|_{\mathbb{P}_n, 2}$ and $\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2$ for ϕ_n in (55) such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $0 < \alpha < 2$.

Since the design points are non-random, the modulus (55) involves relatively simple multiplier processes, to which the abstract maximal inequalities may apply directly. In particular, if the error variables are sub-Gaussian, then the stochastic process $\{n^{-1/2}\sum_{i=1}^n(\theta-\theta_0)(z_i)\epsilon_i : \theta \in \Theta_n\}$ is sub-Gaussian with respect to the $L_2(\mathbb{P}_n)$- semimetric on the set of regression functions. Thus, using (41), we may choose

$$\phi_n(\delta) = \int_0^\delta \sqrt{\log N(\epsilon, \Theta_n \cap \{\theta : \mathbb{P}_n(\theta-\theta_0)^2 \le \delta^2\}, L_2(\mathbb{P}_n))}\, d\epsilon.$$

Example 6.3 (Bounded isotonic regression)**.** *Let $\Theta_n = \Theta = \{f : [0,1] \to [0,1] : f$ is nondecreasing$\}$. By Theorem 2.7.5 of [van der Vaart and Wellner, 1996] we see that*

$$\log N(\epsilon, \Theta, L_2(\mathbb{P}_n)) \le K\epsilon^{-1},$$

where $K > 0$ is a universal constant. Thus, we can take $\phi_n(\delta) = \sqrt{K}\int_0^\delta \epsilon^{-1/2}d\epsilon = 2\sqrt{K}\sqrt{\delta}$. Thus we solve $\sqrt{\delta_n} = \delta_n^2\sqrt{n}$ to obtain the rate of convergence of $\delta_n = n^{-1/3}$.

Example 6.4 (Lipschitz regression)**.** *Let $\Theta = \Theta_n := \{f : [0,1] \to [0,1] \mid f$ is 1-Lipschitz$\}$. By Lemma 2.8, we see that $\phi_n(\delta)$ can be taken[47] to be $\sqrt{\delta}$ which yields the rate of $\delta_n = n^{-1/3}$.*

Example 6.5 (Hölder smooth functions)**.** *For $\alpha > 0$, we consider the class of all functions on a bounded set $\mathcal{X} \subset \mathbb{R}^d$ that possess uniformly bounded partial derivatives up to $\lfloor\alpha\rfloor$ and whose highest partial derivates are 'Lipschitz' (actually Hölder) of order $\alpha - \lfloor\alpha\rfloor$*[48]*.*

Let $\mathcal{X} = [0,1]^d$ and let $\Theta_n = C_1^\alpha([0,1]^d)$. Then, $\log N(\epsilon, \Theta, L_2(\mathbb{P}_n)) \le \log N(\epsilon, \Theta, \|\cdot\|_\infty) \lesssim \epsilon^{-d/\alpha}$. Thus, for $\alpha > d/2$ this leads to $\phi_n(\delta) \gg \delta^{1-d/(2\alpha)}$ and hence, $\phi_n(\delta) \le \delta_n^2\sqrt{n}$

[47] Note that a ϵ-cover in the $\|\cdot\|_\infty$-norm (as in Lemma 2.8) also yields a cover in the $L_2(\mathbb{P}_n)$-seminorm.

[48] i.e., for any vector $\mathbf{k} = (k_1, \ldots, k_d)$ of d integers the differential operator

$$D^{\mathbf{k}} = \frac{\partial^k}{\partial x_1^{k_1}\cdots\partial x_d^{k_d}},$$

where $k = \sum_{i=1}^d k_i$. Then for a function $f : \mathcal{X} \to \mathbb{R}$, let

$$\|f\|_\alpha := \max_{k \le \lfloor\alpha\rfloor} \sup_x \left|D^k f(x)\right| + \max_{k = \lfloor\alpha\rfloor} \sup_{x,y} \frac{\left|D^{\mathbf{k}} f(x) - D^{\mathbf{k}} f(y)\right|}{\|x-y\|^{\alpha - \lfloor\alpha\rfloor}},$$

where the supremum is taken over all x, y in the interior of $\mathcal{X}$ with $x \ne y$. Let $C_M^\alpha(\mathcal{X})$ be the set of all continuous functions $f : \mathcal{X} \to \mathbb{R}$ with $\|f\|_\alpha \le M$. The following lemma, proved in [van der Vaart and Wellner, 1996, Chapter 7], bounds the entropy number of the class $C_M^\alpha(\mathcal{X})$.

Lemma 6.6. *Let $\mathcal{X}$ be a bounded, convex subset of $\mathbb{R}^d$ with nonempty interior. Then there exists a constant K, depending only on α and d, and a constant K', depending only on α, diam$(\mathcal{X})$ and d, such that*

$$\begin{aligned} \log N(\epsilon, C_1^\alpha(\mathcal{X}), \|\cdot\|_\infty) &\le K\lambda(\mathcal{X}^1)\epsilon^{-d/\alpha}, \\ \log N_{[\,]}(\epsilon, C_1^\alpha(\mathcal{X}), L_r(Q)) &\le K'\epsilon^{-d/\alpha}, \end{aligned}$$

for every $\epsilon > 0$, $r \ge 1$, where $\lambda(\mathcal{X}^1)$ is the Lebesgue measure of the set $\{x : \|x - \mathcal{X}\| \le 1\}$ and Q is any probability measure on $\mathbb{R}^d$. Note that $\|\cdot\|_\infty$ denotes the supremum norm.

can be solved to obtain the rate of convergence $\delta_n \gtrsim n^{-\alpha/(2\alpha+d)}$. *The rate relative to the empirical* L_2*-norm is bounded above by*

$$n^{-\alpha/(2\alpha+d)} + \|\theta_0 - \Theta_n\|_{\mathbb{P}_{n,2}}.$$

For $\theta_0 \in C_1^\alpha([0,1]^d)$ *the second term vanishes; the first is known to be the minimax rate over this set.*

Exercise (HW2) (Convex regression): Suppose that $\theta_0 : C \to \mathbb{R}$ is known to be a convex function over its domain C, some convex and open subset of $\mathbb{R}^d$. In this case, it is natural to consider the LSE with a convexity constraint — namely

$$\hat{\theta}_n \in \operatorname*{argmin}_{f:C\to\mathbb{R} \text{ "convex"}} \frac{1}{n}\sum_{i=1}^n (Y_i - f(z_i))^2. \tag{56}$$

As stated, this optimization problem is infinite-dimensional in nature. Fortunately, by exploiting the structure of convex functions, it can be converted to an equivalent finite-dimensional problem[49]. Show that the above LSE can be computed by solving the optimization problem:

$$\min_{u_1,\ldots,u_n\in\mathbb{R};\xi_1,\ldots,\xi_n\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n (Y_i - u_i)^2 \qquad \text{s.t.} \quad u_i + \xi_i^\top (z_j - z_i) \le u_j \quad \forall i \ne j.$$

Note that this is a convex program in $N = n(d+1)$ variables, with a quadratic cost function and a total of $n(n-1)$ linear constraints. Give the form of a LSE $\hat{\theta}_n$.

Suppose now that $C = [0,1]^d$, and instead of minimizing (56) over the class of all convex functions, we minimize over the class of all L-Lipschitz convex functions. Find the rate of convergence of the LSE (over all L-Lipschitz convex functions).

6.2 Least squares regression: a finite sample inequality

In the *standard nonparametric regression model*, we assume the noise variables in (54) are drawn in an i.i.d. manner from the $N(0,\sigma^2)$ distribution, where $\sigma > 0$ is the unknown standard deviation parameter. In this case, we can write $\epsilon_i = \sigma w_i$, where $w_i \sim N(0,1)$ are i.i.d. We change our notation slightly and assume that $f^* : \mathcal{Z} \to \mathbb{R}$ is the unknown regression function (i.e., $f^* \equiv \theta_0$ in (54)).

[49] Any convex function f is *subdifferentiable* at each point in the (relative) interior of its domain C. More precisely, at any interior point $z \in C$, there exists at least one vector $\xi \in \mathbb{R}^d$ such that

$$f(z) + \xi^\top (x - z) \le f(x), \qquad \text{for all } x \in C.$$

Any such vector is known as a *subgradient*, and each point $z \in C$ can be associated with the set $\partial f(z)$ of its subgradients, which is known as the subdifferential of f at z. When f is actually differentiable at z, then the above inequality holds if and only if $\xi = \nabla f(z)$, so that we have $\partial f(z) = \{\nabla f(z)\}$. See standard references in convex analysis for more on this.

Our main result in this section yields a finite sample inequality for the $L_2(\mathbb{P}_n)$-loss of the constrained LSE

$$\hat{f}_n \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - f(z_i)\}^2;$$

i.e., we study the error $\|\hat{f}_n - f^*\|_n^2 := \frac{1}{n}\sum_{i=1}^n \{\hat{f}_n(z_i) - f^*(z_i)\}^2$. This error is expressed in terms of a *localized form of Gaussian complexity*: it measures the complexity of the function class $\mathcal{F}$, locally in a neighborhood around the true regression function f^*. More precisely, we define the set:

$$\mathcal{F}^* := \mathcal{F} - f^* = \{f - f^* : f \in \mathcal{F}\} \tag{57}$$

corresponding to an f^*-shifted version of the original function class $\mathcal{F}$. For a given radius $\delta > 0$, the *local Gaussian complexity* around f^* at scale δ is given by

$$G_n(\delta; \mathcal{F}^*) := \mathbb{E}_w \left[\sup_{g \in \mathcal{F}^* : \|g\|_n \leq \delta} \left| \frac{1}{n} \sum_{i=1}^{n} w_i g(z_i) \right| \right]$$

where the expectation is w.r.t. the variables $\{w_i\}_{i=1}^n$ which are i.i.d. $N(0,1)$.

A function class $\mathcal{H}$ is *star-shaped* if for any $h \in \mathcal{H}$ and $\alpha \in [0,1]$, the rescaled function αh also belongs to $\mathcal{H}$. Recall the basic inequality for nonparametric least squares:

$$\frac{1}{2}\|\hat{f}_n - f^*\|_n^2 \leq \frac{\sigma}{n} \sum_{i=1}^{n} w_i \{f(z_i) - f^*(z_i)\}. \tag{58}$$

A central object in our analysis is the set of $\delta > 0$ that satisfy the *critical inequality*

$$G_n(\delta; \mathcal{F}^*) \leq \frac{\delta^2}{2\sigma}. \tag{59}$$

It can be shown that the star-shaped condition ensures existence of the critical radius[50].

[50] Let $\mathcal{H}$ be a star-shaped class of functions.

Lemma 6.7. *For any star-shaped function class $\mathcal{H}$, the function $\delta \mapsto G_n(\delta, \mathcal{H})/\delta$ is nonincreasing on the interval $(0, \infty)$. Consequently, for any constant $c > 0$, the inequality $G_n(\delta, \mathcal{H}) \leq c\delta^2$ has a smallest positive solution.*

Proof. For a pair $0 < \delta \leq t$, it suffices to show that $\frac{\delta}{t} G_n(t; \mathcal{H}) \leq G_n(\delta; \mathcal{H})$. Given any function $h \in \mathcal{H}$ with $\|h\|_n \leq t$, we may define the rescaled function $\tilde{h} = \frac{\delta}{t} h$. By construction, we have $\|\tilde{h}\|_n \leq \delta$; moreover, since $\delta \leq t$, the star-shaped assumption on $\mathcal{H}$ guarantees that $\tilde{h} \in \mathcal{H}$. Thus, write

$$\frac{1}{n} \left| \frac{\delta}{t} \sum_{i=1}^{n} w_i h(z_i) \right| = \frac{1}{n} \left| \sum_{i=1}^{n} w_i \tilde{h}(z_i) \right| \leq \sup_{g \in \mathcal{H} : \|g\|_n \leq \delta} \frac{1}{n} \left| \sum_{i=1}^{n} w_i g(z_i) \right|.$$

Taking the supremum over the set $\mathcal{H} \cap \{\|h\|_n \leq t\}$ on the left-hand side followed by expectations yields $\frac{\delta}{t} G_n(t; \mathcal{H}) \leq G_n(\delta; \mathcal{H})$, which completes the proof of the first part. As $G_n(\delta; \mathcal{H})/\delta$ is nonincreasing and $c\delta$ is nondecreasing (in δ) on $(0, \infty)$, the inequality $G_n(\delta, \mathcal{H}) \leq c\delta^2$ has a smallest positive solution. □

Theorem 6.8. *Suppose that the shifted function class $\mathcal{F}^*$ is star-shaped, and let δ_n be any positive solution to the critical inequality (59). Then for any $t \geq \delta_n$, the LSE $\hat{f}_n$ satisfies the bound*

$$\mathbb{P}\left(\|\hat{f}_n - f^*\|_n^2 \geq 16 t\delta_n\right) \leq e^{-\frac{nt\delta_n}{2\sigma^2}}.$$

Exercise (HW2): By integrating this tail bound, show that the mean-squared error in the $L_2(\mathbb{P}_n)$-semi-norm is upper bounded as

$$\mathbb{E}\left[\|\hat{f}_n - f^*\|_n^2\right] \leq c\left\{\delta_n^2 + \frac{\sigma^2}{n}\right\}$$

for some universal constant c.

Proof. Recall the basic inequality (58). In terms of the shorthand notation $\hat{\Delta} := \hat{f}_n - f^*$, it can be written as $\frac{1}{2}\|\hat{\Delta}\|_n^2 \leq \frac{\sigma}{n}\sum_{i=1}^n w_i \hat{\Delta}(z_i)$. By definition, the error function $\hat{\Delta} = \hat{f}_n - f^*$ belongs to the shifted function class $\mathcal{F}^*$. We will need the following lemma.

Lemma 6.9. *Let $\mathcal{H}$ be an arbitrary star-shaped function class, and let $\delta_n > 0$ satisfy the inequality $G_n(\delta; \mathcal{H}) \leq \delta^2/(2\sigma)$. For a given scalar $u \geq \delta_n$, define the event*

$$\mathcal{A}(u) := \left\{\exists\, g \in \{h \in \mathcal{H} : \|h\|_n \geq u\} : \left|\frac{\sigma}{n}\sum_{i=1}^n w_i g(z_i)\right| \geq 2\|g\|_n u\right\}. \tag{60}$$

Then, for all $u \geq \delta_n$, we have

$$\mathbb{P}(\mathcal{A}(u)) \leq e^{-\frac{nu^2}{2\sigma^2}}.$$

We will prove the main theorem using the lemma for the time being; we take $\mathcal{H} = \mathcal{F}^*$ and $u = \sqrt{t\delta_n}$ for some $t \geq \delta_n$, so that we can write $\mathbb{P}(\mathcal{A}^c(\sqrt{t\delta_n})) \geq 1 - e^{-\frac{nt\delta_n^2}{2\sigma^2}}$. Note that

$$\begin{aligned}
\mathbb{P}(\|\hat{\Delta}\|_n^2 \leq 16 t\delta_n) &= \mathbb{P}\left(\|\hat{\Delta}\|_n^2 \leq 16 t\delta_n, \|\hat{\Delta}\|_n^2 < t\delta_n\right) + \mathbb{P}\left(\|\hat{\Delta}\|_n^2 \leq 16 t\delta_n, \|\hat{\Delta}\|_n^2 \geq t\delta_n\right) \\
&= \mathbb{P}\left(\|\hat{\Delta}\|_n^2 < t\delta_n\right) + \mathbb{P}\left(t\delta_n \leq \|\hat{\Delta}\|_n^2 \leq 16 t\delta_n\right) \\
&\geq \mathbb{P}\left(\|\hat{\Delta}\|_n^2 < t\delta_n\right) + \mathbb{P}\left(t\delta_n \leq \|\hat{\Delta}\|_n^2 \leq 16 t\delta_n, \mathcal{A}^c(\sqrt{t\delta_n})\right) \\
&= \mathbb{P}\left(\|\hat{\Delta}\|_n^2 < t\delta_n\right) + \mathbb{P}\left(t\delta_n \leq \|\hat{\Delta}\|_n^2, \mathcal{A}^c(\sqrt{t\delta_n})\right) \qquad (61) \\
&\geq \mathbb{P}\left(\mathcal{A}^c(\sqrt{t\delta_n})\right) \geq 1 - e^{-\frac{nt\delta_n^2}{2\sigma^2}},
\end{aligned}$$

where the only nontrivial step is (61), which we explain next. Note that if $\|\hat{\Delta}\|_n^2 \geq t\delta_n$ and $\mathcal{A}^c(\sqrt{t\delta_n})$ holds, then

$$\left|\frac{1}{n}\sum_{i=1}^n w_i \hat{\Delta}(z_i)\right| \leq 2\|\hat{\Delta}\|_n \sqrt{t\delta_n}.$$

Consequently, the basic inequality (58) implies that $\|\hat{\Delta}\|_n^2 \leq 4\|\hat{\Delta}\|_n\sqrt{t\delta_n}$, or equivalently, $\|\hat{\Delta}\|_n^2 \leq 16 t\delta_n$. Thus, (61) holds, thereby completing the proof.

Proof of Lemma 6.9: Our first step is to reduce the problem to controlling a supremum over a subset of functions satisfying the upper bound $\|\tilde{g}\|_n \leq u$. Suppose that there exists some $g \in \mathcal{H}$ with $\|g\|_n \geq u$ such that

$$\left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i g(z_i)\right| \geq 2\|g\|_n u. \tag{62}$$

Defining the function $\tilde{g} := \frac{u}{\|g\|_n} g$, we observe that $\|\tilde{g}\|_n = u$. Since $g \in \mathcal{H}$ and $\frac{u}{\|g\|_n} \in (0,1]$, the star-shaped assumption on $\mathcal{H}$ implies that $\tilde{g} \in \mathcal{H}$. Consequently, we have shown that if there exists a function g satisfying inequality (62), which occurs whenever the event $\mathcal{A}(u)$ is true, then there exists a function $\tilde{g} \in \mathcal{H}$ with $\|\tilde{g}\|_n = u$ such that

$$\left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i \tilde{g}(z_i)\right| = \frac{u}{\|g\|_n}\left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i g(z_i)\right| \geq 2u^2.$$

We thus conclude that

$$\mathbb{P}(\mathcal{A}(u)) \leq \mathbb{P}\left(Z_n(u) \geq 2u^2\right), \qquad \text{where} \quad Z_n(u) := \sup_{\tilde{g}\in\mathcal{H}:\|\tilde{g}\|_n \leq u} \left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i \tilde{g}(z_i)\right|.$$

Since the noise variables $w_i \sim N(0,1)$ are i.i.d., the variable $\frac{\sigma}{n}\sum_{i=1}^{n} w_i \tilde{g}(z_i)$ is zero-mean and Gaussian for each fixed $\tilde{g}$. Therefore, the variable $Z_n(u)$ corresponds to the supremum of a Gaussian process. If we view this supremum as a function of the standard Gaussian vector $(w_1, \ldots, w_n)$, then it can be verified that the associated Lipschitz constant[51] is at most $\sigma u/\sqrt{n}$. Consequently, by the *concentration of Lipschitz functions of Gaussian variables*[52],

[51] The following lemma illustrates the Lipschitz nature of Gaussian complexity.

Lemma 6.10. *Let $\{W_k\}_{k=1}^n$ be an i.i.d. sequence of $N(0,1)$ variables. Given a collection of vectors $A \subset \mathbb{R}^n$, define the random variable $Z := \sup_{a\in A} |\sum_{k=1}^n a_k W_k|$. Viewing Z as a function $(w_1, \ldots, w_n) \mapsto f(w_1, \ldots, w_n)$, we can verify that f is Lipschitz (with respect to Euclidean norm) with parameter $\sup_{a\in A\cup(-A)} \|a\|_2$.*

To see this, let $w = (w_1, \ldots, w_n), w' = (w'_1, \ldots, w'_n) \in \mathbb{R}^n$. Suppose that there exists $a^* = (a_1^*, \ldots, a_n^*)$ such that $f(w) = \sup_{a\in A} |\sum_{k=1}^n a_k w_k| = \sum_{k=1}^n a_k^* w_k$ (or $\sum_{k=1}^n (-a_k^*) w_k$, which case can also be handled similarly). Then,

$$f(w) - f(w') \leq \sum_{k=1}^n a_k^* w_k - \sum_{k=1}^n a_k^* w'_k \leq \|a^*\|_2 \|w - w'\|_2 \leq \sup_{a\in A\cup(-A)} \|a\|_2 \|w - w'\|_2.$$

The same argument holds with the roles of w and w' switched which leads to the desired result:

$$|f(w) - f(w')| \leq \sup_{a\in A\cup(-A)} \|a\|_2 \|w - w'\|_2.$$

[52] **Classical result on the concentration properties of Lipschitz functions of Gaussian variables**: Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to the Euclidean norm $\|\cdot\|_2$ if

$$|f(x) - f(y)| \leq L\|x - y\|_2, \qquad \text{for all } \; x, y \in \mathbb{R}^n.$$

The following result guarantees that any such function is sub-Gaussian with parameter at most L.

we obtain the tail bound

$$\mathbb{P}\left(Z_n(u) \geq \mathbb{E}[Z_n(u)] + s\right) \leq e^{-\frac{ns^2}{2u^2\sigma^2}},$$

valid for any $s > 0$. Setting, $s = u^2$ yields,

$$\mathbb{P}\left(Z_n(u) \geq \mathbb{E}[Z_n(u)] + u^2\right) \leq e^{-\frac{nu^2}{2\sigma^2}}. \tag{63}$$

Finally, by definition of $Z_n(u)$ and $G_n(u;\mathcal{H})$, we have $\mathbb{E}[Z_n(u)] = \sigma G_n(u;\mathcal{H})$. By Lemma 6.7, the function $v \mapsto G_n(v;\mathcal{H})/v$ is nonincreasing, and since $u \geq \delta_n$ by assumption, we have

$$\sigma\frac{G_n(u;\mathcal{H})}{u} \leq \sigma\frac{G_n(\delta_n;\mathcal{H})}{\delta_n} \leq \frac{\delta_n}{2} \leq \delta_n,$$

where the 2nd inequality used the critical condition (59). Putting together the pieces, we have shown that $\mathbb{E}[Z_n(u)] \leq u\delta_n$. Combined with the tail bound (63), we obtain

$$\mathbb{P}(Z_n(u) \geq 2u^2) \leq \mathbb{P}(Z_n(u) \geq u\delta_n + u^2) \leq \mathbb{P}\left(Z_n(u) \geq \mathbb{E}[Z_n(u)] + u^2\right) \leq e^{-\frac{nu^2}{2\sigma^2}},$$

where we have used the fact that $u^2 \geq u\delta_n$. □

Exercise (HW2): Suppose that $\mathcal{F}^*$ is star-shaped. Show that for any $\delta \in (0, \sigma]$ such that

$$\frac{16}{\sqrt{n}}\int_{\delta^2/(4\sigma)} \sqrt{\log N(t, \mathcal{F}^* \cap \{h : \|h\|_n \leq \delta\}, \|\cdot\|_n)}dt \leq \frac{\delta^2}{4\sigma} \tag{64}$$

satisfies the critical inequality (59) and hence the conclusion of of Theorem 6.8 holds.

Exercise (HW2) [Linear regression]: Consider the standard linear regression model $Y_i = \langle \theta^*, z_i\rangle + w_i$, where $\theta^* \in \mathbb{R}^d$, and fixed x_i are d-dimensional covariates. Although this example can be studied using direct linear algebraic arguments, we will use our general theory in analysis this model. The usual LSE corresponds to optimizing over the class of all linear functions

$$\mathcal{F}_{\text{lin}} := \{f_\theta = \langle \theta, \cdot\rangle : \theta \in \mathbb{R}^d\}. \tag{65}$$

Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ denote the design matrix with $z_i \in \mathbb{R}^d$ as its i-th row. Let $\hat{\theta}$ be the LSE. Show that

$$\|f_{\hat{\theta}} - f_{\theta^*}\|_n^2 = \frac{\|\mathbf{X}(\hat{\theta} - \theta^*)\|_2^2}{n} \lesssim \sigma^2\frac{\text{rank}(\mathbf{X})}{n}$$

Theorem 6.11. *Let $X = (X_1, \ldots, X_n)$ be a vector of i.i.d. standard Gaussian variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz with respect to the Euclidean norm. Then the variable $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most L, and hence*

$$\mathbb{P}\big(|f(X) - \mathbb{E}[f(X)]| \geq t\big) \leq 2e^{-\frac{t^2}{2L^2}} \qquad \text{for all } t \geq 0.$$

Note that this result is truly remarkable: it guarantees that any L-Lipschitz function of a standard Gaussian random vector, regardless of the dimension, exhibits concentration like a scalar Gaussian variable with variance L^2. See Section 13.4 for more details about this result and a proof.

with high probability.

Hint: First note that the shifted class $\mathcal{F}^*_{\text{lin}} = \mathcal{F}_{\text{lin}}$ for any choice of $f_{\theta^*} \in \mathcal{F}_{\text{lin}}$. Moreover, $\mathcal{F}^*_{\text{lin}}$ is convex and hence star-shaped around any point. To use (64) to find δ_n so that Theorem 6.8 applies in this setting, we have to find $N(t, \mathcal{F}^*_{\text{lin}} \cap \{h : \|h\|_n \leq \delta\}, \|\cdot\|_n)$. Show that the required covering number can be bounded by $(1 + \frac{2\delta}{t})^r$ where $r := \text{rank}(\mathbf{X})$.

6.3 Oracle inequalities

In our analysis thus far, we have assumed that the true regression function f^* belongs to the function class $\mathcal{F}$ over which the constrained LSE is defined. In practice, this assumption might be violated. In such settings, we expect the performance of the LSE to involve both the *estimation error* that arises in Theorem 6.8, and some additional form of *approximation error*, arising from the fact that $f^* \notin \mathcal{F}$. A natural way in which to measure approximation error is in terms of the best approximation to f^* using functions from $\mathcal{F}$ — the error in this best approximation is given by $\inf_{f \in \mathcal{F}} \|f - f^*\|_n^2$. Note that this error can only be achieved by an "oracle" that has direct access to the samples $\{f^*(x_i)\}_{i=1}^n$. For this reason, results that involve this form of approximation error are referred to as *oracle inequalities*. With this setup, we have the following generalization of Theorem 6.8. We define

$$\partial\mathcal{F} := \{f - g : f, g \in \mathcal{F}\}.$$

Theorem 6.12. *Assume that $\partial\mathcal{F}$ is star-shaped. Let $\delta_n > 0$ be any solution to*

$$G_n(\delta; \partial\mathcal{F}) \leq \frac{\delta^2}{2\sigma}. \tag{66}$$

Then for any $t \geq \delta_n$, the LSE $\hat{f}$ satisfies the bound

$$\|\hat{f} - f^*\|_n^2 \leq 2 \inf_{f \in \mathcal{F}} \|f - f^*\|_n^2 + 36 t\delta_n \tag{67}$$

with probability greater than $1 - e^{-\frac{nt\delta_n}{2\sigma^2}}$.

Proof. Recall the definition of $\mathcal{A}(u)$ in (60). We apply Lemma 6.9 with $u = \sqrt{t\delta_n}$ and $\mathcal{H} = \partial\mathcal{F}$ to conclude that $\mathbb{P}\left(\mathcal{A}^c(\sqrt{t\delta_n})\right) \geq 1 - e^{-\frac{nt\delta_n}{2\sigma^2}}$. We will assume below that the event $\mathcal{A}^c(\sqrt{t\delta_n})$ holds.

Given an arbitrary $\tilde{f} \in \mathcal{F}$, since $\tilde{f}$ is feasible and $\hat{f}$ is optimal, we have

$$\frac{1}{2n}\sum_{i=1}^n \{Y_i - \hat{f}(z_i)\}^2 \leq \frac{1}{2n}\sum_{i=1}^n \{Y_i - \tilde{f}(z_i)\}^2.$$

Using the relation $Y_i = f^*(z_i) + \sigma w_i$, some algebra yields

$$\frac{1}{2}\|\hat{\Delta}\|_n^2 \leq \frac{1}{2}\|\tilde{f} - f^*\|_n^2 + \left|\frac{\sigma}{n}\sum_{i=1}^n w_i \tilde{\Delta}(z_i)\right|, \tag{68}$$

where $\hat{\Delta} := \hat{f} - f^*$ and $\tilde{\Delta} := \hat{f} - \tilde{f}$. It remains to analyze the term on the right-hand side involving $\tilde{\Delta}$. We break our analysis into two cases.

Case 1: First suppose that $\|\tilde{\Delta}\|_n \leq \sqrt{t\delta_n}$. Then,

$$\begin{aligned} \|\hat{\Delta}\|_n^2 = \|\hat{f} - f^*\|_n^2 &= \|(\tilde{f} - f^*) + \tilde{\Delta}\|_n^2 \\ &\leq \big\{\|\tilde{f} - f^*\|_n + \sqrt{t\delta_n}\big\}^2 \\ &\leq 2\|\tilde{f} - f^*\|_n^2 + 2t\delta_n \quad (\text{taking } \beta = 1) \end{aligned}$$

where in the first inequality above we have used the triangle inequality, and the second inequality follows from the fact that $(a+b)^2 \leq 2(a^2+b^2)$ (for $a, b \in \mathbb{R}$).

Case 2: Suppose now that $\|\tilde{\Delta}\|_n > \sqrt{t\delta_n}$. Note that $\tilde{\Delta} \in \partial\mathcal{F}$ and as the event $\mathcal{A}^c(\sqrt{t\delta_n})$ holds, we get

$$\Big|\frac{\sigma}{n}\sum_{i=1}^n w_i \tilde{\Delta}(z_i)\Big| \leq 2\sqrt{t\delta_n}\|\tilde{\Delta}\|_n.$$

Combining with the basic inequality (68), we find that, with probability at least $1 - e^{-\frac{nt\delta_n}{2\sigma^2}}$, the squared error is bounded as

$$\begin{aligned} \|\hat{\Delta}\|_n^2 &= \|\tilde{f} - f^*\|_n^2 + 4\sqrt{t\delta_n}\|\tilde{\Delta}\|_n \\ &\leq \|\tilde{f} - f^*\|_n^2 + 4\sqrt{t\delta_n}\big\{\|\hat{\Delta}\|_n + \|\tilde{f} - f^*\|_n\big\} \\ &\leq \|\tilde{f} - f^*\|_n^2 + 2\Big[\frac{t\delta_n}{\beta} + \beta\|\hat{\Delta}\|_n^2\Big] + 2\Big[\frac{t\delta_n}{\beta} + \beta\|\tilde{f} - f^*\|_n^2\Big] \\ \Rightarrow \quad (1-2\beta)\|\hat{\Delta}\|_n^2 &\leq (1+2\beta)\|\tilde{f} - f^*\|_n^2 + 4\frac{t\delta_n}{\beta} \end{aligned}$$

where the second step follows from the triangle inequality and the next step follows from multiple usage of the fact that $2ab \leq \beta a^2 + b^2/\beta$ (for $a, b \in \mathbb{R}$ and $\beta > 0$). Taking $\beta = 1/6$, we have $\frac{(1+2\beta)}{(1-2\beta)} = 2$, and thus we get

$$\|\hat{\Delta}\|_n^2 \leq 2\|\tilde{f} - f^*\|_n^2 + 36t\delta_n.$$

Combining the pieces we get that, under the event $\mathcal{A}^c(\sqrt{t\delta_n})$, the above inequality holds for any $\tilde{f} \in \mathcal{F}$. Thus, (67) holds. □

Remark 6.1. *We can, in fact, have a slightly more general form of* (67) *where the 'oracle' approximation term* $2\|\tilde{f} - f^*\|_n^2$ *can be replaced by* $\frac{1+\gamma}{1-\gamma}\|\tilde{f} - f^*\|_n^2$ *for any* $\gamma \in (0,1)$ *(with appropriate adjustments to the 'estimation' error term* $36t\delta_n$*).*

Note that the guarantee (67) is actually a family of bounds, one for each $f \in \mathcal{F}$. When $f^* \in \mathcal{F}$, then we can set $f = f^*$, so that the bound (67) reduces to asserting that $\|\hat{f} - f^*\|_n^2 \lesssim t\delta_n$ with high probability, where δ_n satisfies the critical inequality (66). Thus, up to constant factors, we recover Theorem 6.8 as a special case of Theorem 6.12. By integrating the tail bound, we are guaranteed that

$$\mathbb{E}\Big[\|\hat{f} - f^*\|_n^2\Big] \lesssim \inf_{f \in \mathcal{F}} \|f - f^*\|_n^2 + \delta_n^2 + \frac{\sigma^2}{n}. \tag{69}$$

The bound (69) guarantees that the LSE $\hat{f}$ has prediction error that is at most a constant multiple of the oracle error, plus a term proportional to δ_n^2. The term $\inf_{f \in \mathcal{F}} \|f - f^*\|_n^2$ can be viewed a form of approximation error that decreases as the function class $\mathcal{F}$ grows, whereas the term δ_n^2 is the estimation error that increases as $\mathcal{F}$ becomes more complex.

6.3.1 Best sparse linear regression

Consider the standard linear model $Y_i = f_{\theta^*}(z_i) + \sigma w_i$, where $f_\theta(z) := \langle \theta, z \rangle$ is an unknown linear regression function, and $w_i \overset{iid}{\sim} N(0,1)$ is an i.i.d. noise sequence. Here $\theta^* \in \mathbb{R}^d$ is the unknown parameter. For some sparsity index $s \in \{1, 2, \ldots, d\}$, consider the class of all linear regression functions based on s-sparse vectors — namely, the class

$$\mathcal{F}_{\text{spar}}(s) := \{f_\theta : \theta \in \mathbb{R}^d, \|\theta\|_0 \leq s\},$$

where $\|\theta\|_0 := \sum_{j=1}^d I(\theta_j \neq 0)$ counts the number of non-zero coefficients in the vector $\theta \in \mathbb{R}^d$. Disregarding computational considerations, a natural estimator of θ^* is given by

$$\hat{\theta} \equiv f_{\hat{\theta}} \in \arg \min_{f_\theta \in \mathcal{F}_{\text{spar}(s)}} \sum_{i=1}^n \{Y_i - f_\theta(z_i)\}^2, \tag{70}$$

corresponding to performing least squares over the set of all regression vectors with at most s non-zero coefficients. As a corollary of Theorem 6.12, we claim that the $L_2(\mathbb{P}_n)$-error of this estimator is upper bounded as

$$\|f_{\hat{\theta}} - f_{\theta^*}\|_n^2 \lesssim \inf_{\theta \in \mathcal{F}_{\text{spar}(s)}} \|f_{\hat{\theta}} - f_{\theta^*}\|_n^2 + \sigma^2 \frac{s \log(\frac{ed}{s})}{n}, \tag{71}$$

with high probability; here $\delta_n^2 = \sigma^2 \frac{s \log(\frac{ed}{s})}{n}$. Consequently, up to constant factors, its error is as good as the best s-sparse predictor plus the 'estimation' error term δ_n^2. Note that this 'estimation' error term grows linearly with the sparsity s, but only logarithmically in the dimension d, so that it can be very small even when the dimension is exponentially larger than the sample size n. In essence, this result guarantees that we pay a relatively small price for not knowing in advance the best s-sized subset of coefficients to use.

In order to derive this result as a consequence of Theorem 6.12, we need to compute the local Gaussian complexity $G_n(\delta; \partial\mathcal{F}_{\text{spar}}(s))$. Making note of the inclusion $\partial\mathcal{F}_{\text{spar}}(s) \subset \mathcal{F}_{\text{spar}}(2s)$, we have $G_n(\delta; \partial\mathcal{F}_{\text{spar}}(s)) \subset G_n(\delta; \mathcal{F}_{\text{spar}}(2s))$. Now let $S \subset \{1, \ldots, d\}$ be an arbitrary $2s$-sized subset of indices, and let $\mathbf{X}_S \in \mathbb{R}^{n \times 2s}$ denote the submatrix with columns indexed by S. We can then write

$$G_n(\delta; \mathcal{F}_{\text{spar}}(2s)) = \mathbb{E}_w \left[\sup_{g \in \mathcal{F}_{\text{spar}}(2s); \|g\|_n \leq \delta} \left| \frac{1}{n} \sum_{i=1}^n w_i g(z_i) \right| \right] = \mathbb{E}_w \left[\max_{|S| \leq 2s} Z_n(S) \right],$$

where

$$Z_n(S) := \sup_{\theta_S \in \mathbb{R}^{2s}; \frac{\|\mathbf{X}_S \theta_S\|_2}{\sqrt{n}} \leq \delta} \frac{1}{n} \left| w^\top \mathbf{X}_S \theta_S \right|$$

as, for $g \in \mathcal{F}_{\text{spar}}(2s)$, $g(z) \equiv g_\theta(z) = \langle \theta, z \rangle = \langle \theta_S, z_S \rangle$, if θ has nonzero entries in the subset $S \subset \{1, \ldots, d\}$, and $\|g\|_n^2 = \frac{1}{n}\sum_{i=1}^n \langle \theta, z_i \rangle^2 = \frac{1}{n}\|\mathbf{X}_S \theta_S\|_2^2$ (here $\|\cdot\|_2$ denotes the usual Euclidean norm).

Viewed as a function of the standard Gaussian vector $w \in \mathbb{R}^n$, the variable $Z_n(S)$ is Lipschitz with parameter at most $\delta/\sqrt{n}$ (by Lemma 6.10), from which Theorem 6.11 implies the tail bound

$$\mathbb{P}(Z_n(S) \geq \mathbb{E}[Z_n(S)] + t\delta) \leq e^{\frac{-t^2\delta^2}{(2\delta^2/n)}} = e^{\frac{-nt^2}{2}}, \qquad \text{for all } t > 0. \tag{72}$$

We now upper bound the expectation. Consider the singular value decomposition $\mathbf{X}_S = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{n \times 2s}$ and $\mathbf{V} \in \mathbb{R}^{d \times 2s}$ are matrices of left and right singular vectors, respectively, and $\mathbf{D} \in \mathbb{R}^{2s \times 2s}$ is a diagonal matrix of the singular values. Noting that $\|\mathbf{X}_S \theta_S\|_2 = \|\mathbf{D}\mathbf{V}^\top \theta_S\|_2$, we arrive at the upper bound

$$\mathbb{E}[Z_n(S)] \leq \mathbb{E}\Big[\sup_{\beta \in \mathbb{R}^{2s}: \|\beta\|_2 \leq \delta} \frac{1}{\sqrt{n}} |\langle \mathbf{U}^\top w, \beta \rangle|\Big] \leq \frac{\delta}{\sqrt{n}} \mathbb{E}\left[\|\mathbf{U}^\top w\|_2\right]$$

where we have taken $\beta = \frac{\mathbf{D}\mathbf{V}^\top \theta_S}{\sqrt{n}}$. Since $w \sim N(0, I_n)$ and the matrix $\mathbf{U}$ has orthogonal columns, we have $\mathbf{U}^\top w \sim N(0, I_{2s})$, and therefore $\mathbb{E}\left[\|\mathbf{U}^\top w\|_2\right] \leq \sqrt{2s}$. Combining this upper bound with the earlier tail bound (72), an application of the union bound yields, for all $t > 0$,

$$\mathbb{P}\left[\max_{|S|=2s} Z_n(S) \geq \frac{\delta\sqrt{2s}}{\sqrt{n}} + t\delta\right] \leq \binom{d}{2s} e^{\frac{-nt^2}{2}}.$$

By integrating this tail bound, we find that

$$\frac{G_n(\delta; \mathcal{F}_{\text{spar}}(2s))}{\delta} = \frac{\mathbb{E}_w\left[\max_{|S|=2s} Z_n(S)\right]}{\delta} \lesssim \sqrt{\frac{s}{n}} + \sqrt{\frac{\log\binom{d}{2s}}{n}} \lesssim \sqrt{\frac{\log \frac{ed}{s}}{n}},$$

so that the critical inequality (66) is satisfied for $\delta_n^2 \asymp \sigma^2 \frac{s \log(\frac{ed}{s})}{n}$, as claimed.

6.4 Density estimation via maximum likelihood

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a density p_0 that belongs to a set $\mathcal{P}$ of densities with respect to a measure μ on some measurable space. In this subsection the parameter is the density p_0 itself (and we denoted a generic density by p instead of θ).

The *sieved maximum likelihood estimator* (MLE) $\hat{p}_n$ based on $X_1, \ldots, X_n$ maximizes the log-likelihood $p \mapsto \mathbb{P}_n \log p$ over a sieve $\mathcal{P}_n$, i.e.,

$$\hat{p}_n = \operatorname*{argmax}_{p \in \mathcal{P}_n} \mathbb{P}_n[\log p].$$

Although it is natural to take the objective (criterion) function we optimize (i.e., $\mathbb{M}_n(\cdot)$ in our previous notation) as $\mathbb{P}_n \log p$, for some technical reasons (explained below) we consider a slightly modified function.