Complements

In this chapter, we include results that complement those in the previous two chapters. In Section 8.1, we provide the minimax excess risk for Gaussian linear density estimation in the *well-specified* setting, which relates to the result obtained in Chapter 7 for the SMP in the general *misspecified* setting; we also compare this risk to that of the best proper estimator (namely, the MLE) in high dimension. In Section 8.2, we complement the minimax lower bound for least squares in Chapter 6 by a lower bound on the Bayes risk under isotropic Gaussian prior for arbitrary signal-to-noise ratio; this amounts to a general lower bound for Stieltjes transforms of empirical spectral distributions of random vectors with identity covariance.

Contents

8.1 Gaussian linear density estimation in high dimension 301
8.2 A Marchenko-Pastur lower bound on Stieltjes transforms . . . 306

Gaussian linear density estimation in high dimension

In this complement, we determine the minimax excess risk for predictive (conditional) density estimation with respect to the linear Gaussian model in the well-specified case, which was referred to in Section 7.4.1 of Chapter 7 as well as Section 1.4.5 of the introduction.

Specifically, the setting is that of conditional density estimation, see Section 1.4 as well as Chapter 7. Here, the space of covariates is $\mathcal{X} = \mathbf{R}^d$, the response lies in $\mathcal{Y} = \mathbf{R}$. The considered (conditional) model is the *Gaussian linear model*, given by the conditional densities of the form

$$\mathcal{F} = \left\{ f_\beta(\cdot|x) := \mathcal{N}(\langle \beta, x \rangle, \sigma^2) : \beta \in \mathbf{R}^d \right\}, \tag{8.1}$$

where we set the base measure on $\mathbf{R}^d$ to be $\mu(\mathrm{d}y) = (2\pi)^{-d/2}\mathrm{d}y$ and identify densities with respect to μ with the corresponding densities. Here, σ^2 is fixed, and without loss of generality we assume that $\sigma^2 = 1$. Finally, we consider in this section the *well-specified case*, where the true conditional distribution of Y given X belongs to the class $\mathcal{F}$. The results here (and their proof) are similar in spirit to those of Chapter 6 on regression with square loss.

Setting. We assume that $(X_1, Y_1), \dots, (X_n, Y_n)$ are i.i.d. samples from a distribution P, such that the conditional distribution of Y given X belongs to the class $\mathcal{F}$, *i.e.* such that $Y =$

$\langle \beta^*, X \rangle + \varepsilon$ where $\varepsilon | X \sim \mathcal{N}(0,1)$. Hence, the corresponding set of distributions P of (X, Y) is characterized by the distribution P_X of covariates X, and is denoted $\mathcal{P} := \mathcal{P}_{\text{Gauss}}(P_X, 1)$ (with the notation of Chapter 6). Recall from Sections 1.1.1 and 1.4 of the introduction that the *risk* of a conditional density g is

$$R(g) := \mathbb{E}[\ell(g, (X,Y))] = \mathbb{E}[-\log g(Y|X)] \, ,$$

where ℓ denotes the logarithmic loss. Also, the *minimax excess risk* is by definition

$$\mathcal{E}_n^*(P_X) := \inf_{\widehat{g}_n} \sup_{P \in \mathcal{P}} \mathbb{E}[\mathcal{E}(\widehat{g}_n)] = \inf_{\widehat{g}_n} \sup_{P \in \mathcal{P}} \Big\{ \mathbb{E}[R(\widehat{g}_n)] - \inf_{\beta \in \mathcal{F}} R(f_\beta) \Big\} \, , \tag{8.2}$$

where $\widehat{g}_n$ spans all estimators of Y given X. In what follows, we assume that $\mathbb{E}[\|X\|^2] < +\infty$ and that the covariance matrix $\Sigma = \mathbb{E}[XX^\top]$ is invertible.

Main result. Theorem 8.1 below provides the minimax risk, as a function of the distribution P_X of covariates.

Theorem 8.1. *If the distribution P_X is degenerate (in the sense of Definition 6.1, Chapter 6) or if $n < d$, then the minimax risk (8.2) is infinite. If P_X is non-degenerate and $n \geqslant d$, then the minimax excess risk (8.2) in the well-specified case is given by*

$$\frac{1}{2} \mathbb{E}\Big[\log\big(1 + \langle (n\widehat{\Sigma}_n)^{-1} X, X \rangle \big) \Big] = \frac{1}{2} \mathbb{E}\big[-\log(1 - \widehat{\ell}_{n+1}) \big] \, , \tag{8.3}$$

where $\widehat{\ell}_{n+1} = \langle (n\widehat{\Sigma}_n + X_{n+1} X_{n+1}^\top)^{-1} X_{n+1}, X_{n+1} \rangle$ denotes the leverage score *of the point X_{n+1} in the sample $X_1, \dots, X_{n+1}$. This minimax risk is achieved by the Bayes predictive posterior under uniform prior on $\mathbf{R}^d$, namely*

$$\widehat{g}_n(\cdot | x) = \mathcal{N}(\langle \widehat{\beta}_n^{\text{LS}}, x \rangle, (1 + \langle (\widehat{\Sigma}_n)^{-1} x, x \rangle)) \, ,$$

where $\widehat{\beta}_n^{\text{LS}}$ is the OLS estimator.

First, it is worth noting that, as in the case of least-squares regression (Theorem 6.2, Chapter 6), the minimax excess risk in density estimation is characterized by the distribution of statistical leverage scores: the more uneven they are, the higher the minimax risk.

Second, the minimax risk in the well-specified case (Theorem 8.1) is precisely *half* the worst-case risk of the SMP estimator in the general misspecified case (Theorem 7.4 in Chapter 7). This implies in particular that the minimax risk in the misspecified case is at most twice that in the well-specified case.

High dimension and suboptimality of proper estimators. By convexity of the function $u \mapsto -\log(1-u)$, and since $\mathbb{E}[\widehat{\ell}_{n+1}] = d/(n+1)$ under the conditions of Theorem 8.1 (see Section 6.2.2), for every distribution P_X, the minimax risk (8.3) is at least

$$\mathcal{E}_n^*(P_X) \geqslant -\frac{1}{2} \log\big(1 - \mathbb{E}[\widehat{\ell}_{n+1}]\big) = -\frac{1}{2} \log\Big(1 - \frac{d}{n+1}\Big) \, .$$

On the other hand, by concavity of the log function, the minimax risk (8.3) is smaller than

$$\frac{1}{2} \log\Big(1 + \mathbb{E}\big[\langle (\widehat{\Sigma}_n)^{-1} X, X \rangle\big]\Big) = \frac{1}{2} \log\Big(1 + \mathbb{E}\big[\text{Tr}(\widetilde{\Sigma}_n^{-1})\big]\Big) \, ;$$

when the features are Gaussian, namely $X \sim \mathcal{N}(0, \Sigma)$, we have $\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})] = d/(n-d-1)$ for $n > d+1$ (Breiman and Freedman, 1983), so that

$$\mathcal{E}_n^*(P_X) \leqslant \frac{1}{2} \log\Big(1 + \frac{d}{n-d-1}\Big) = -\frac{1}{2}\log\Big(1 - \frac{d}{n-1}\Big) .$$

In particular, in the high-dimensional asymptotic regime where $n, d \to \infty$ while the "hardness" of the problem is fixed, namely $d/n \to \gamma \in (0,1)$ (if $\gamma > 1$, the minimax risk is infinite by Theorem 8.1), both the above distribution-independent lower bound and the upper bound for Gaussian covariates converge to the same limit, namely

$$\frac{1}{2}\log\Big(1 + \frac{\gamma}{1-\gamma}\Big) = -\frac{1}{2}\log(1-\gamma) . \tag{8.4}$$

As in the least-squares problem (Chapter 6), Gaussian covariates are almost the "easiest" covariates in terms of minimax risk in high dimension, owing to the fact that the distribution of leverage scores converges to a Dirac mass at γ.

In addition, when restricting to proper (within $\mathcal{F}$) conditional distributions, the problem is equivalent to least-squares regression, with square loss $\ell(\beta, (x, y)) = \frac{1}{2}(y - \langle \beta, x\rangle)^2$ (see e.g. Section 7.3.2). In particular, by the results in Section 6.2, the minimax *proper* estimator is the MLE $\widehat{f}_n(\cdot|x) = \mathcal{N}(\langle \widehat{\beta}_n^{\mathrm{LS}}, x\rangle, 1)$, with risk

$$\frac{\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})]}{2n} \geqslant \frac{d}{2(n-d+1)} .$$

In particular, if $d/n \to \gamma$, this quantity is asymptotically at least $\gamma/(2(1-\gamma))$, with equality in the case of Gaussian covariates. This limiting risk is strictly larger than the one for general (improper) estimators (8.4). Hence, even when the true distribution belongs to the model, using improper estimators can be advantageous in high dimension. This contrasts with the asymptotic optimality of the MLE in the classical regime where d is fixed while $n \to \infty$, and complements the results of Chapter 7 which highlight the degradation of the MLE under model misspecification.

Proof of Theorem 8.1. The proof of Theorem 8.1 follows from similar arguments as that of Theorem 6.1, hence we only highlight the part that differs. The main difference is the computation of the risk of Bayes estimators under Gaussian prior.

Lemma 8.1 (Risk of Bayes predictive posteriors)**.** *Let $\lambda > 0$. The Bayes predictive posterior under Gaussian prior $\Pi_\lambda = \mathcal{N}(0, (\lambda n)^{-1} I_d)$ on β^* is*

$$\widehat{g}_{\lambda,n}(\cdot|x) = \mathcal{N}\big(\langle \widehat{\beta}_{\lambda,n}, x\rangle, 1 + n^{-1}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} x, x\rangle\big) , \tag{8.5}$$

where $\widehat{\beta}_{\lambda,n}$ denotes the Ridge estimator (6.33); when P_X is non-degenerate and $n \geqslant d$, the above is well-defined for $\lambda = 0$ and equals $\widehat{g}_n$. Then, if $\lambda > 0$ or if the previous conditions apply, we have, assuming that $Y = \langle \beta^*, X\rangle + \varepsilon$ with $\varepsilon | X \sim \mathcal{N}(0,1)$,

$$\begin{aligned}\mathbb{E}[\mathcal{E}(\widehat{g}_{\lambda,n})] = \frac{1}{2}\mathbb{E}\big[\log\big(1 + n^{-1}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} X, X\rangle\big)\big] + \frac{\lambda^2}{2} \cdot \mathbb{E}\Bigg[\frac{\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} X, \beta^*\rangle^2}{1 + n^{-1}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} X, X\rangle}\Bigg] \\ - \frac{\lambda}{2} \cdot \mathbb{E}\Bigg[\frac{n^{-1}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-2} X, X\rangle}{1 + n^{-1}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} X, X\rangle}\Bigg] .\end{aligned} \tag{8.6}$$

Proof of Lemma 8.1. A standard computation shows that the posterior distribution $\widehat{\Pi}_\lambda$ on $\beta^* \in \mathbf{R}^d$ is $\mathcal{N}(\widehat{\beta}_{\lambda,n}, n^{-1}(\widehat{\Sigma}_n + \lambda I_d)^{-1})$. The predictive posterior given $x \in \mathbf{R}^d$ is then the distribution of $Y_x \sim \mathcal{N}(\langle \beta, x\rangle, 1)$, where $\beta \sim \widehat{\Pi}_\lambda$. Now, $\beta = \widehat{\beta}_{\lambda,n} + n^{-1/2}(\widehat{\Sigma}_n + \lambda I_d)^{-1/2} Z$, where $Z \sim \mathcal{N}(0, I_d)$, while $Y_x = \langle \beta, x\rangle + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0,1)$ independent of Z. Hence, $Y_x = \langle \widehat{\beta}_{\lambda,n}, x\rangle + n^{-1/2}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1/2} Z, x\rangle + \varepsilon$ is Gaussian (conditionally on $\widehat{\beta}_{\lambda,n}, \widehat{\Sigma}_n$) with mean $\langle \widehat{\beta}_{\lambda,n}, x\rangle$ and variance $\mathrm{Var}(\varepsilon) + \mathrm{Var}(n^{-1/2}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1/2} x, Z\rangle) = 1 + n^{-1}\|(\widehat{\Sigma}_n + \lambda I_d)^{-1/2} x\|^2$, *i.e.* $Y_x \sim \widehat{g}_{\lambda,n}$.

First, consider a conditional density of the form $g(\cdot|x) = \mathcal{N}(\mu(x), \sigma^2(x))$, so that $g(y|x) = \sigma(x)^{-1}\exp(-[y-\mu(x)]^2/[2\sigma^2(x)])$. We have, for every $(x,y) \in \mathbf{R}^d \times \mathbf{R}$,

$$\ell(g,(x,y)) = \frac{1}{2}\log \sigma^2(x) + \frac{1}{2\sigma^2(x)}(y - \mu(x))^2 \,,$$

so that

$$R(g) = \frac{1}{2}\mathbb{E}\big[\log \sigma^2(X)\big] + \frac{1}{2}\mathbb{E}\Big[\frac{(Y-\mu(X))^2}{\sigma^2(X)}\Big]$$

In particular, this risk is minimized by $g(\cdot|x) = \mathcal{N}(\langle \beta^*, x\rangle, 1)$, for which it equals $1/2$. Hence, in the case of $\widehat{g}_{\lambda,n}$, we get

$$2\,\mathbb{E}[\mathcal{E}(\widehat{g}_{\lambda,n})] = \mathbb{E}[\log(1 + n^{-1}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} X, X\rangle)] + \mathbb{E}\Big[\frac{(Y - \langle \widehat{\beta}_{\lambda,n}, X\rangle)^2}{1 + n^{-1}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} X, X\rangle}\Big] - 1. \quad (8.7)$$

Now, we have

$$\begin{aligned}\mathbb{E}\Big[\frac{(Y - \langle \widehat{\beta}_{\lambda,n}, X\rangle)^2}{1 + n^{-1}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} X, X\rangle}\Big] &= \mathbb{E}\Big[\frac{(\langle \widehat{\beta}_{\lambda,n} - \beta^*, X\rangle - \varepsilon)^2}{1 + n^{-1}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} X, X\rangle}\Big] \\ &= \mathbb{E}\Big[\frac{\langle \widehat{\beta}_{\lambda,n} - \beta^*, X\rangle^2 + 1}{1 + n^{-1}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} X, X\rangle}\Big] \qquad (8.8) \\ &= \mathbb{E}\Big[\frac{\mathbb{E}[\langle \widehat{\beta}_{\lambda,n} - \beta^*, X\rangle^2 | X_{1:n}, X] + 1}{1 + n^{-1}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} X, X\rangle}\Big] \qquad (8.9)\end{aligned}$$

where $X_{1:n} = (X_1, \dots, X_n)$ and (8.8) comes from the fact that, conditionally on $(X_{1:n}, Y_{1:n}, X)$, ε is centered with unit variance. Now since

$$\widehat{\beta}_{\lambda,n} = (\widehat{\Sigma}_n + \lambda I_d)^{-1} \cdot \frac{1}{n}\sum_{i=1}^n Y_i X_i = (\widehat{\Sigma}_n + \lambda I_d)^{-1}\widehat{\Sigma}_n \beta^* + (\widehat{\Sigma}_n + \lambda I_d)^{-1} \cdot \frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i \,,$$

we have

$$\langle \widehat{\beta}_n - \beta^*, X\rangle = \frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} X_i, X\rangle - \lambda \langle (\widehat{\Sigma}_n + \lambda I_d)^{-1}\beta^*, X\rangle \,,$$

so that, using that $\mathbb{E}[\varepsilon_i | X_{1:n}, X] = 0$ and $\mathbb{E}[\varepsilon_i^2 | X_{1:n}, X] = 1$,

$$\begin{aligned}\mathbb{E}[\langle \widehat{\beta}_{\lambda,n} - \beta^*, X\rangle^2 | X_{1:n}, X] &= \frac{1}{n^2}\sum_{i=1}^n \langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} X_i, X\rangle^2 + \lambda^2 \langle (\widehat{\Sigma}_n + \lambda I_d)^{-1}\beta^*, X\rangle^2 \\ &= \frac{1}{n}\langle (\widehat{\Sigma}_n + \lambda I_d)^{-1}\widehat{\Sigma}_n (\widehat{\Sigma}_n + \lambda I_d)^{-1} X, X\rangle + \lambda^2 \langle (\widehat{\Sigma}_n + \lambda I_d)^{-1} X, \beta^*\rangle^2.\end{aligned}$$

Plugging this into (8.9), we get

$$\begin{aligned}
&\mathbb{E}\Big[\frac{(Y-\langle\widehat{\beta}_{\lambda,n},X\rangle)^2}{1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle}\Big]-1\\
&=\mathbb{E}\Big[\frac{n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}\widehat{\Sigma}_n(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle+\lambda^2\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,\beta^*\rangle^2+1}{1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle}\Big]-1\\
&=\mathbb{E}\Big[\frac{n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}\widehat{\Sigma}_n(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle+\lambda^2\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,\beta^*\rangle^2-n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle}{1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle}\Big]\\
&=\lambda^2\cdot\mathbb{E}\Big[\frac{\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,\beta^*\rangle^2}{1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle}\Big]-\lambda\cdot\mathbb{E}\Big[\frac{n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-2}X,X\rangle}{1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle}\Big],
\end{aligned}$$

which together with (8.7) establishes Lemma 8.1. □

In particular, when P_X is non-degenerate and $n \geqslant d$, we obtain by setting $\lambda = 0$ in Lemma 8.1:

$$\mathbb{E}[\mathcal{E}(\widehat{g}_n)] = \frac{1}{2}\mathbb{E}\Big[\log\big(1+\langle(n\widehat{\Sigma}_n)^{-1}X,X\rangle\big)\Big] = \frac{1}{2}\mathbb{E}\big[-\log(1-\widehat{\ell}_{n+1})\big]\,,$$

where the second inequality comes from the Sherman-Morrison identity (Horn and Johnson, 1990), see Lemma 6.1. This establishes an upper bound on the minimax risk.

A matching lower bound on the minimax risk (including in the case where P_X is degenerate or $n < d$) is then obtained similarly to Theorem 6.1, from the following:

Corollary 8.1 (Bayes risk under Gaussian prior). *Let $\lambda > 0$. Then, the Bayes optimal risk under Gaussian prior $\Pi_\lambda = \mathcal{N}(0,(\lambda n)^{-1}I_d)$ equals*

$$\frac{1}{2}\mathbb{E}\big[\log\big(1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle\big)\big]\,.$$

Proof. Let $\mathcal{L}(\beta^*,\widehat{f}_n) = \mathbb{E}_{\beta^*}[\mathcal{E}(\widehat{f}_n)]$ denote the Kullback-Leibler expected excess risk of the estimator $\widehat{f}_n$ under the distribution P_{β^*}, namely when $Y|X \sim \mathcal{N}(\langle\beta^*,X\rangle,1)$. The Bayes optimal estimator under prior Π_λ and under Kullback-Leibler loss is simply the predictive posterior (Berger, 1985; Lehmann and Casella, 1998), which is $\widehat{g}_{\lambda,n}$. Hence, we have

$$\begin{aligned}
&\inf_{\widehat{f}_n}\mathbb{E}_{\beta^*\sim\Pi_\lambda}[\mathbb{E}_{\beta^*}[\mathcal{E}(\widehat{f}_n)]] = \mathbb{E}_{\beta^*\sim\Pi_\lambda}[\mathcal{L}(\beta^*,\widehat{g}_{\lambda,n})]\\
&=\frac{1}{2}\,\mathbb{E}\big[\log\big(1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle\big)\big]+\\
&+\frac{\lambda^2}{2}\cdot\mathbb{E}_{\beta^*\sim\Pi_\lambda}\Big[\mathbb{E}_{X_{1:n},X}\Big[\frac{\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,\beta^*\rangle^2}{1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle}\Big]\Big]-\frac{\lambda}{2}\cdot\mathbb{E}\Big[\frac{n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-2}X,X\rangle}{1+n^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,X\rangle}\Big]
\end{aligned}$$

Now, by Fubini's theorem and since

$$\begin{aligned}
\mathbb{E}_{\beta^*\sim\Pi_\lambda}\big[\langle(\widehat{\Sigma}_n+\lambda I_d)^{-1}X,\beta^*\rangle^2\big] &= \mathbb{E}_{\beta^*\sim\Pi_\lambda}\big[X^\top(\widehat{\Sigma}_n+\lambda I_d)^{-1}\beta^*(\beta^*)^\top(\widehat{\Sigma}_n+\lambda I_d)^{-1}X\big]\\
&= X^\top(\widehat{\Sigma}_n+\lambda I_d)^{-1}\mathbb{E}_{\beta^*\sim\Pi_\lambda}[\beta^*(\beta^*)^\top](\widehat{\Sigma}_n+\lambda I_d)^{-1}X\\
&= (\lambda n)^{-1}X^\top(\widehat{\Sigma}_n+\lambda I_d)^{-1}(\widehat{\Sigma}_n+\lambda I_d)^{-1}X\\
&= (\lambda n)^{-1}\langle(\widehat{\Sigma}_n+\lambda I_d)^{-2}X,X\rangle\,,
\end{aligned}$$

the second and third terms of the above sum compensate. This proves Corollary 8.1. □

8.2 A Marchenko-Pastur lower bound on Stieltjes transforms of ESDs of covariance matrices

In this section, we let X be a random vector in $\mathbf{R}^d$, with unit covariance: $\mathbb{E}[XX^\top] = I_d$. Given n i.i.d. variables $X_1, \dots, X_n$ distributed as X, define the *sample covariance matrix* as

$$\widehat{\Sigma}_n := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top \, . \tag{8.10}$$

$\widehat{\Sigma}_n$ is a symmetric, positive semi-definite $d \times d$ matrix. Let $\lambda_1(\widehat{\Sigma}_n) \geqslant \dots \geqslant \lambda_d(\widehat{\Sigma}_n)$ denote the (ordered) eigenvalues of $\widehat{\Sigma}_n$, and denote $\widehat{\lambda}_{j,n} = \lambda_j(\widehat{\Sigma}_n)$ for $1 \leqslant j \leqslant d$. The *empirical spectral distribution* (ESD) of $\widehat{\Sigma}_n$ is by definition the distribution $\widehat{\mu}_n = (1/d) \sum_{j=1}^d \delta_{\widehat{\lambda}_{j,n}}$, with cumulative distribution function

$$\widehat{F}_n(x) = \frac{1}{d} \sum_{j=1}^{d} \mathbf{1}(\widehat{\lambda}_{j,n} \leqslant x)$$

for $x \in \mathbf{R}$. The celebrated Marchenko-Pastur theorem (Marchenko and Pastur, 1967) states that, if $X \sim \mathcal{N}(0, I_d)$, as $d, n \to \infty$ while $d/n \to \gamma \in (0,1)$, the ESD $\widehat{\mu}_n$ converges almost surely in distribution to the *Marcenko-Pastur distribution* μ_γ^{MP}, with density

$$x \longmapsto \frac{\sqrt{(b_\gamma - x)(x - a_\gamma)}}{2\pi\gamma x} \cdot \mathbf{1}(a_\gamma \leqslant x \leqslant b_\gamma)$$

with respect to the Lebesgue measure, where $a_\gamma = (1-\sqrt{\gamma})^2$ and $b_\gamma = (1+\sqrt{\gamma})^2$. This behavior has a form of *universality*, in the sense that it remains true whenever the coordinates of X are independent, centered and with unit variance (Wachter, 1978; Yin, 1986). On the other hand, the independence assumption that underlies this “universal” behavior is quite strong, especially in high dimension where it implies a very specific “incoherent” geometry for the X_i's (including near-constant norm and pairwise orthogonality, see Section 1.3.2).

In this section, we show a form of *extremality* of the Marchenko-Pastur distribution among ESDs of empirical covariance matrices of general (unit covariance) random vectors in $\mathbf{R}^d$. Define the *Stieltjes transform* $S_\mu : \mathbf{R}_+^* \to \mathbf{R}$ of a probability distribution μ supported on $\mathbf{R}^+$ by

$$S_\mu(\lambda) := \int_{\mathbf{R}} (x+\lambda)^{-1} \mu(\mathrm{d}x) \, .$$

The Stieltjes transform (extended to $\lambda \in \mathbf{C} \setminus \mathbf{R}^-$) plays an important role in the spectral analysis of random matrices, and in particular in the proof of the Marchenko-Pastur law (Bai and Silverstein, 2010). Also, define the *expected ESD* $\bar{\mu}_n = \mathbb{E}[\widehat{\mu}_n]$ (such that $\bar{\mu}_n(A) = (1/d) \sum_{j=1}^d \mathbb{P}(\widehat{\lambda}_{j,n} \in A)$ for every measurable subset A of $\mathbf{R}$) and its cumulative distribution function $\bar{F}_n(x) := \mathbb{E}[\widehat{F}_n(x)] = (1/d) \sum_{j=1}^d \mathbb{P}(\widehat{\lambda}_{j,n} \leqslant x)$. Our main result is the following:

Theorem 8.2 (Marchenko-Pastur lower bound)**.** *Let X be a random vector in $\mathbf{R}^d$ such that $\mathbb{E}[XX^\top] = I_d$. Then, the expected Stieltjes transform of the ESD $\widehat{\mu}_n$ is lower bounded in terms of that of the Marchenko-Pastur distribution $\mu_{\gamma'}^{\mathsf{MP}}$ with $\gamma' = d/(n+1)$. Specifically, for every*

$\lambda > 0$, denoting $\lambda' = [n/(n+1)]\lambda$,

$$S_{\bar{\mu}_n}(\lambda) = \frac{1}{d}\,\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\big\}\big] \geqslant \frac{n}{n+1}\,\frac{-(1-\gamma'+\lambda') + \sqrt{(1-\gamma'+\lambda')^2 + 4\gamma'\lambda'}}{2\lambda'\gamma'}$$
$$= \frac{n}{n+1} S_{\mu^{\mathsf{MP}}_{\gamma'}}(\lambda') \,. \tag{8.11}$$

In particular, if $n, d \to \infty$ with $d/n \to \gamma \in (0,1)$, $\liminf_{n\to\infty} \inf_{P_X} S_{\bar{\mu}_n}(\lambda) \geqslant S_{\mu^{\mathsf{MP}}_{\gamma}}(\lambda)$ for every $\lambda > 0$.

Theorem 8.2 states that the Marchenko-Pastur law, which is a limiting distribution of ESDs of vectors with *independent coordinates*, also provides a non-asymptotic lower bound (in terms of associated Stieltjes transforms) for ESDs of *general* random vectors in $\mathbf{R}^d$.

Before giving the proof of Theorem 8.2 (which is elementary and relies on a combination of the Sherman-Morrison formula with a fixed-point argument), let us indicate some consequences for least-squares regression and Gaussian linear density estimation.

Let us fix a distribution P_X of covariates X such that $\Sigma := \mathbb{E}[XX^\top]$ is invertible. For $\sigma^2 > 0$, consider the statistical model $\mathcal{P} = \mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2) = \{P_{(X,Y)} : Y|X \sim \mathcal{N}(\langle\beta^*, X\rangle, \sigma^2), \beta^* \in \mathbf{R}^d\}$. For $\lambda > 0$, define the prior distribution $\Pi_\lambda = \mathcal{N}(0, \sigma^2/(\lambda n)\Sigma^{-1})$ on β^*. Π_λ has constant density on the sets $\{\beta^* \in \mathbf{R}^d : \|\beta^*\|_\Sigma = t\}$ of constant *signal strength* $\|\beta^*\|_\Sigma = \mathbb{E}[\langle\beta^*, X\rangle^2]^{1/2}$. Let us also define the *signal-to-noise ratio* (SNR) $\eta^2 = \eta^2(\lambda) := \mathbb{E}_{\beta^*\sim\Pi_\lambda}[\|\beta^*\|_\Sigma^2]/\sigma^2 = d/(\lambda n)$.

Corollary 8.2 (Lower bound on Bayes risk in regression in terms of SNR)**.** *Let $\lambda > 0$, and $\eta := \eta(\lambda)$ be the corresponding SNR. Then for every distribution P_X such that $\mathbb{E}[XX^\top] = \Sigma$, the Bayes optimal risk $B_{d,n}(P_X, \eta, \sigma^2)$ under prior Π_λ for prediction under square loss $\ell(\beta, (x,y)) = (y - \langle\beta, x\rangle)^2$ is lower bounded as*

$$B_{d,n}(P_X, \eta, \sigma^2) \geqslant \sigma^2 \cdot \frac{-\big(n+1-d+d/\eta^2\big) + \sqrt{(n+1-d+d/\eta^2)^2 + 4d^2/\eta^2}}{2d/\eta^2} \,. \tag{8.12}$$

In particular, under the limit scaling $n, d \to \infty$ with $d/n \to \gamma \in (0,1)$, the Bayes risk is asymptotically lower bounded by

$$\liminf_{n\to\infty,\ d/n\to\gamma}\ \inf_{P_X} B_{d,n}(P_X, \eta, \sigma^2) \geqslant \sigma^2 \cdot \frac{-(1-\gamma+\gamma/\eta^2) + \sqrt{(1-\gamma+\gamma/\eta^2)^2 + 4\gamma^2/\eta^2}}{2\gamma/\eta^2} \,.$$

This lower bound is tight: indeed, when $X \sim \mathcal{N}(0, \Sigma)$, the Bayes risk converges to this limit; for fixed $\lambda > 0$ this follows from the Marchenko-Pastur law and dominated convergence, see Bai et al. (2003); Dicker (2016) for rates of convergence. This extends the observation that the minimax risk is approximately minimized in the case of Gaussian covariates (see Section 6.2.2 of Chapter 6) to the Bayes risk with arbitrary signal strength.

Proof of Theorem 8.2. First, write

$$S_{\bar{\mu}_n}(\lambda) = \mathbb{E}\bigg[\int_{\mathbf{R}} (x+\lambda)^{-1} \widehat{\mu}_n(\mathrm{d}x)\bigg] = \mathbb{E}\bigg[\frac{1}{d}\sum_{j=1}^d (\widehat{\lambda}_{j,n} + \lambda)^{-1}\bigg] = \frac{1}{d}\,\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\big\}\big] \,.$$

Then, denoting $\rho := (d/n)S_{\bar{\mu}_n}(\lambda)$, we have

$$\begin{aligned}
\rho &= \frac{1}{n}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\big\}\big] \\
&= \mathbb{E}\big[\langle (n\widehat{\Sigma}_n + \lambda n I_d)^{-1} X, X\rangle\big] \\
&= \mathbb{E}\left[\frac{\langle (n\widehat{\Sigma}_n + XX^\top + \lambda n I_d)^{-1} X, X\rangle}{1 - \langle (n\widehat{\Sigma}_n + XX^\top + \lambda n I_d)^{-1} X, X\rangle}\right] \qquad (8.13) \\
&\geqslant \frac{\mathbb{E}\big[\langle (n\widehat{\Sigma}_n + XX^\top + \lambda n I_d)^{-1} X, X\rangle\big]}{1 - \mathbb{E}\big[\langle (n\widehat{\Sigma}_n + XX^\top + \lambda n I_d)^{-1} X, X\rangle\big]} \qquad (8.14)
\end{aligned}$$

where (8.13) uses the Sherman-Morrison identity, while (8.14) comes from the convexity of $x \mapsto x/(1-x)$. Now, by exchangeability of $(X_1, \dots, X_n, X)$, letting $\lambda' = [n/(n+1)]\lambda$,

$$\begin{aligned}
\mathbb{E}\big[\langle (n\widehat{\Sigma}_n + XX^\top + \lambda n I_d)^{-1} X, X\rangle\big] &= \frac{1}{n+1}\mathbb{E}\left[\sum_{i=1}^{n+1}\big\langle \big((n+1)\widehat{\Sigma}_{n+1} + \lambda n I_d\big)^{-1} X_i, X_i\big\rangle\right] \\
&= \mathbb{E}\big[\mathrm{Tr}\big\{\big((n+1)\widehat{\Sigma}_{n+1} + \lambda n I_d\big)^{-1}\widehat{\Sigma}_{n+1}\big\}\big] \\
&= \frac{1}{n+1}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_{n+1} + \lambda' I_d)^{-1}\widehat{\Sigma}_{n+1}\big\}\big] \\
&= \frac{d}{n+1} - \lambda' \cdot \frac{1}{n+1}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_{n+1} + \lambda' I_d)^{-1}\big\}\big] \\
&\geqslant \frac{d}{n+1} - \lambda' \rho\,, \qquad (8.15)
\end{aligned}$$

where (8.15) comes from the fact that, since $(n+1)\widehat{\Sigma}_{n+1} \geqslant n\widehat{\Sigma}_n$,

$$\begin{aligned}
\frac{1}{n+1}\mathbb{E}\big[\mathrm{Tr}\big\{\big(\widehat{\Sigma}_{n+1} + \lambda' I_d\big)^{-1}\big\}\big] &= \mathbb{E}\big[\mathrm{Tr}\big\{\big((n+1)\widehat{\Sigma}_{n+1} + \lambda n I_d\big)^{-1}\big\}\big] \\
&\leqslant \mathbb{E}\big[\mathrm{Tr}\big\{\big(n\widehat{\Sigma}_n + \lambda n I_d\big)^{-1}\big\}\big] = \rho\,. \qquad (8.16)
\end{aligned}$$

Let $\gamma' := d/(n+1)$. It follows from (8.14) and (8.15) that ρ satisfies:

$$\rho \geqslant \frac{\gamma' - \lambda'\rho}{1 - \gamma' + \lambda'\rho}\,,$$

which after rearranging (using $1 - \gamma' + \lambda'\rho > 0$) amounts to

$$\lambda'\rho^2 + (1 - \gamma' + \lambda')\rho - \gamma' \geqslant 0\,. \qquad (8.17)$$

Since the polynomial of order 2 in ρ in equation (8.17) equals $-\gamma' < 0$ at 0, it has a positive and a negative root. Since $\rho > 0$, (8.17) is equivalent to saying that ρ is larger than the positive root of the polynomial in (8.17), which writes:

$$\rho \geqslant \frac{-(1 - \gamma' + \lambda') + \sqrt{(1 - \gamma' + \lambda')^2 + 4\gamma'\lambda'}}{2\lambda'}\,. \qquad (8.18)$$

Theorem 8.2 ensues since $S_{\bar{\mu}_n}(\lambda) = (n/d)\rho = [n/(n+1)]\rho/\gamma'$, while the Stieltjes transform of the Marchenko-Pastur distribution may be found in Bai and Silverstein (2010, Lemma 3.11). □

Proof of Corollary 8.2. Up to the changes of variables $\widetilde{X} = \Sigma^{-1/2} X$ and $\widetilde{Y} = Y/\sigma$, we reduce to the case where $\Sigma = I_d$ and $\sigma^2 = 1$. As shown in the proof of Theorem 6.1 (Chapter 6), the Bayes risk under prior $\Pi_\lambda = \mathcal{N}(0, (\lambda n)^{-1} I_d)$ is equal to $n^{-1} \mathbb{E}[\mathrm{Tr}\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\}]$. Corollary 8.2 then follows from Theorem 8.2 after substituting for λ', γ'. □

Conclusion and future work

In this thesis, we studied problems and methods for learning and prediction.

The first part was devoted to Random forest methods, specifically to a variant called *Mondrian forests* (MF) introduced by Lakshminarayanan et al. (2014). Our main contribution was a statistical analysis of this nonparametric procedure, relying on exact computations of relevant local and global properties of the underlying recursive partitions. We deduced minimax rates of convergence for both trees and forests; these rates extend results from Arlot and Genuer (2014) for purely uniformly random forests in dimension one, and highlight an advantage of forests over single trees by a bias reduction due to smoothing of discontinuities (Chapter 2). We also amended the original Mondrian forest procedure, which was introduced for computational reasons (in order to obtain an efficient online algorithm), in order to obtain an exact procedure with risk guarantees, and experimented with this method in a conditional density estimation context (Chapter 3).

The second part dealt with sequential prediction with expert advice. Our main contribution was an analysis of the behavior of the standard exponential weights algorithm, tuned for the worst-case adversarial setting, in the stochastic setting. We showed that the variant with decreasing learning rate achieves optimal adaptation to the sub-optimality gap of the stochastic instance in the same way as more sophisticated algorithms, but unlike the latter fails to adapt to more general Bernstein conditions on losses, and can therefore perform worse in the presence of near-optimal experts (Chapter 4). We also studied a variant of the problem with growing expert classes, and designed efficient algorithms with optimal regret in this setting (Chapter 5).

In the third part, we investigated problems of regression and density estimation, with an emphasis on linear methods. Our first main contribution was a study of random-design least-squares prediction, where we considered the minimax excess risk with respect to linear classes, as a function of the distribution of covariates. We showed that the ordinary least-squares (OLS) estimator is exactly minimax optimal in the absence of an approximation error, and asymptotically as $d = o(n)$ in the general case. In addition, we expressed the minimax excess risk in terms of the distribution of leverage scores, and deduced tight lower bounds for this problem, highlighting the fact that Gaussian design is nearly most favorable for prediction in high dimension. We also obtained upper bounds in expectation for the OLS estimator for non-Gaussian design, under weak distributional assumptions. These latter results relied on a study of the lower tail and negative moments of sample covariance matrices, for which we obtained matching upper and lower bounds under a minimal "small-ball" regularity condition (Chapter 6). Our second main contribution is the introduction of a procedure for statistical learning under logarithmic loss, which satisfies a general excess risk bound valid under model misspecification. This procedure, called *Sample Minmax Predictor* (SMP), is improper and improves over guarantees achievable by proper (within-model) predictors (which degrade under model misspecification) as well as ones obtained through online-to-offline comparison (whose rates contain additional $\log n$ terms, and which cannot achieve uniform bounds over some unbounded classes), partially answering an open problem from Grünwald and Kotłowski (2011). We investigated this procedure in detail for conditional density estimation, with comparison classes formed by the Gaussian linear and logistic models. For logistic regression, the SMP is a simple procedure whose predictions can be computed at the cost of two logistic regressions, and it achieves a fast risk rate under weak assumptions, partly addressing an open problem from Foster et al. (2018) on efficient algorithms with such guarantees (Chapter 7). We complemented these results by a minimax analysis of Gaussian linear density estimation in the well-specified case, showing an advantage of improper estimators even in this case, provided that the dimension is moderately large. In addition, we established a non-asymptotic lower bound for the Stieltjes transform of empirical spectral distributions of sample covariance matrices of general isotropic random vectors in terms of that of the Marchenko-Pastur distribution, which extends the minimax lower bound of Chapter 6 for least-squares regression to Bayes risks depending on the signal-to-noise ratio (Chapter 8).

This work leaves a number of open questions for future research:

1. The results of Chapter 2 for Mondrian Forests, as well as those of Arlot and Genuer (2014) for Purely Random Forests, show that for the proposed (stylized) variants of Random Forests and in the considered regime, the advantage of forests over single trees lies in a *bias* reduction. As highlighted in Section 1.5.2 of the introduction, this result runs counter to the initial motivation for introducing forests, which was to reduce the *variance* of the procedure. Indeed, the results on bias reduction suggest to use *shallower* trees inside a forest than for single trees, which contrasts with the use of deep, completely developed individual trees in Random forests (Breiman, 2001a). To the best of our knowledge, no currently available result justifies the use of fully developed randomized trees with the parameters for bagging used in practice. One way to investigate this could be to consider partly randomized but more adaptive partitions, whose splits are partly data-dependent, and show some variance reduction in this case. This would formalize the intuition put forward by Breiman (1996) for bagging decision trees, that the data-dependent choice of splits makes this procedure unstable, so that bagging may reduce variance.

2. Another way to study this problem, which may be more amenable to precise analysis, would be to investigate the effect of bagging and features subsampling on simpler methods. A natural choice is linear predictors with enough variables that they can fit the dataset (in the same way as fully developed trees), which may be simple enough to be analytically tractable, yet rich enough that they can convey insights about the behavior of complex predictors in the interpolating regime, as shown by recent work (Advani and Saxe, 2017; Liang and Rakhlin, 2018; Bartlett et al., 2019; Hastie et al., 2019; Belkin et al., 2019; Muthukumar et al., 2019).

3. The bound on the lower tail of covariance matrices of Chapter 6 holds in the regime where $n \geqslant 6d$ (that is, for "tall" rectangular design matrices); while we did not attempt to optimize the factor 6, our argument does not extend to square of nearly square matrices with $d \sim n$. Extending the bound to this regime (in order to control moments of the condition number for such matrices) may require refining the proof by using the techniques from Rudelson and Vershynin (2008, 2009); Tao and Vu (2009b,a).

4. The non-asymptotic Marchenko-Pastur lower bound of Section 8.2 on Stieltjes transforms of expected ESDs of general empirical covariance matrices can be seen as a form of extremality of the Marchenko-Pastur distribution among such ESDs. Indeed, it states that
$$\int_{\mathbf{R}} f(x)\bar{\mu}_n(\mathrm{d}x) \geqslant \frac{n}{n+1}\int_{\mathbf{R}} f(x)\mu^{\mathsf{MP}}_{d/(n+1)}(\mathrm{d}x) \tag{8.19}$$
for $f = f_\lambda : x \mapsto (x+\lambda)^{-1}$, for every $\lambda > 0$ (and therefore for any positive linear combination of such functions, which is necessarily convex and in fact totally positive, in the sense that $(-1)^p f^{(p)} \geqslant 0$ for every $n \geqslant 0$). In a restricted and somewhat imprecise sense, this suggests that the Marchenko-Pastur distribution is essentially the least spread-out expected ESD given the aspect ratio d/n in high dimension. It would be interesting to see if this statement could be made more precise, by extending inequality (8.19) (or a similar one) to a wider subclass of convex functions f.

5. The distribution-free excess risk guarantees for the SMP in Chapter 7 hold in expectation, similarly to those obtained through online-to-batch conversion. It would be interesting to complement this by a procedure that achieves distribution-free high (exponential) probability excess risk bounds, for instance in the case of logistic regression.

6. Theorem 7.3 in Chapter 7 shows that the Bayes predictive posterior on the Gaussian location model under uniform prior equalizes the expected excess risk across all (misspecified) distributions. By local asymptotic normality, we expect this behavior to extend asymptotically to smooth parametric models with smooth priors. It would be interesting to see if non-asymptotic distribution-free expected excess risk bounds (similar to those of the SMP) can be obtained for Bayes predictive posteriors (possibly with a learning rate smaller than 1) without averaging, for more general exponential families with suitable (presumably log-concave) priors.

7. Finally, another research direction is to extend or refine the results on the logistic SMP. One possible direction is to obtain analogous regret guarantees for the online problem with individual sequences; the sequential analogue of the SMP may be the Sequentially Normalized Maximum Likelihood (SNML) algorithm (Roos and Rissanen, 2008; Kotłowski and Grünwald, 2011). Another direction would be to obtain excess risk bounds (in the batch setting) with logarithmic rather than quadratic dependence on the norm $\|\beta\|$, similarly to the bound for (computationally expensive) Bayes mixtures (Kakade and Ng, 2005; Foster et al., 2018) or the Ridge SMP in for the Gaussian linear model (Proposition 7.3 Chapter 7). In our setting, a technical difficulty arises in the case of logistic regression due to the use of self-concordance to control the error of the local quadratic approximation, which prevents using regularization parameters as small as in the Gaussian case. A last important question is whether fast rates under weak distributional assumptions such as those of SMP or Bayes mixtures with averaging can be obtained in a computationally more efficient way, and more generally to better understand the possible tradeoffs between computational efficiency and statistical robustness to worst-case distributions.

Bibliography

Adamczak, R., Litvak, A., Pajor, A., and Tomczak-Jaegermann, N. (2010). Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23(2):535–561.

Adamczak, R., Litvak, A. E., Pajor, A., and Tomczak-Jaegermann, N. (2011). Sharp bounds on the rate of convergence of the empirical covariance matrix. *Comptes Rendus Mathematique*, 349(3-4):195–200.

Advani, M. S. and Saxe, A. M. (2017). High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*.

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, 62(3):547–554.

Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation*, 9(7):1545–1588.

Anderson, G. W., Guionnet, A., and Zeitouni, O. (2010). *An introduction to random matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley New York.

Arlot, S. (2008). V-fold cross-validation improved: V-fold penalization. *arXiv preprint arXiv:0802.0566*.

Arlot, S. and Genuer, R. (2014). Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*.

Aslan, M. (2006). Asymptotically minimax Bayes predictive densities. *The Annals of Statistics*, 34(6):2921–2938.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Audibert, J.-Y. (2004). *PAC-Bayesian statistical learning theory*. PhD thesis, Université Paris VI.

Audibert, J.-Y. (2008). Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems 20*, pages 41–48.

Audibert, J.-Y. (2009). Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646.

Audibert, J.-Y. and Catoni, O. (2010). Linear regression through PAC-Bayesian truncation. *arXiv preprint arXiv:1010.0072*.

Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794.

Auer, P., Cesa-Bianchi, N., and Gentile, C. (2002). Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75.

Auer, P. and Chiang, C.-K. (2016). An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *29th Annual Conference on Learning Theory (COLT)*, volume 49, pages 116–120.

Azoury, K. S. and Warmuth, M. K. (2001). Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246.

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367.

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414.

Bach, F. (2014). Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627.

Bach, F. and Moulines, É. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems 26*, pages 773–781.

Bai, Z., Miao, B., and Yao, J.-F. (2003). Convergence rates of spectral distributions of large sample covariance matrices. *SIAM Journal on Matrix Analysis and Applications*, 25(1):105–127.

Bai, Z. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer-Verlag New York, 2 edition.

Bai, Z. and Yin, Y. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Annals of Probability*, 21(3):1275–1294.

Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. (2019). Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460.

Barron, A. R. (1987). Are bayes rules consistent in information? In *Open Problems in Communication and Computation*, pages 85–91. Springer.

Barron, A. R., Rissanen, J. J., and Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

Bartlett, P. L., Grünwald, P. D., Harremoës, P., Hedayati, F., and Kotłowski, W. (2013). Horizon-independent optimal prediction with log-loss in exponential families. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, pages 639–661.

Bartlett, P. L., Koolen, W. M., Malek, A., Takimoto, E., and Warmuth, M. K. (2015). Minimax fixed-design linear regression. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, pages 226–239.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2019). Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Bartlett, P. L. and Mendelson, S. (2006). Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334.

Belkin, M., Hsu, D., and Xu, J. (2019). Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*.

Bennett, G., Dor, L. E., Goodman, V., Johnson, W. B., and Newman, C. M. (1977). On uncomplemented subspaces of l^p, $1 < p < 2$. *Israel Journal of Mathematics*, 26(2):178–187.

Benveniste, A., Métivier, M., and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximations*, volume 22 of *Stochastic Modelling and Applied Probability*. Springer-Verlag Berlin Heidelberg.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer Series in Statistics. Springer-Verlag New York.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39(227):357–365.

Bhatia, R. (2009). *Positive Definite Matrices*, volume 16 of *Princeton Series in Applied Mathematics*. Princeton University Press.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13(1):1063–1095.

Biau, G., Cérou, F., and Guyader, A. (2010). On the rate of convergence of the bagged nearest neighbor estimate. *Journal of Machine Learning Research*, 11(Feb):687–712.

Biau, G. and Devroye, L. (2010). On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101(10):2499–2518.

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9:2015–2033.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2):197–227.

Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97(1-2):113–150.

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375.

Blackwell, D. (1956). An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8.

Blanchard, G. (1999). The "progressive mixture" estimator for regression trees. *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, 35(6):793–820.

Boucheron, S., Bousquet, O., and Lugosi, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford.

Boulesteix, A.-L., Janitza, S., Kruppa, J., and König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):493–507.

Bousquet, O., Boucheron, S., and Lugosi, G. (2004). Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 169–207. Springer.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526.

Bousquet, O. and Warmuth, M. K. (2002). Tracking a small set of experts by mixing past posteriors. *Journal of Machine Learning Research*, 3:363–396.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Breiman, L. (2000). Some infinity theory for predictor ensembles. Technical Report 577, Statistics departement, University of California Berkeley.

Breiman, L. (2001a). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231.

Breiman, L. (2004). Consistency for a simple model of random forests. Technical Report 670, Statistics departement, University of California Berkeley.

Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation? *Journal of the American Statistical Association*, 78(381):131–136.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. The Wadsworth statistics/probability series. CRC, Monterey, CA.

Brown, L. D., George, E. I., and Xu, X. (2008). Admissible predictive density estimation. *The Annals of Statistics*, pages 1156–1170.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

Bubeck, S. and Slivkins, A. (2012). The best of both worlds: stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, volume 23, pages 42.1–42.23.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, 30(4):927–961.

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2007). Aggregation for gaussian regression. *The Annals of Statistics*, 35(4):1674–1697.

Candès, E. J. and Sur, P. (2018). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv preprint arXiv:1804.09753*.

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368.

Catoni, O. (1997). The mixture approach to universal model selection. Technical report, École Normale Supérieure.

Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI - 2001*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag Berlin Heidelberg.

Catoni, O. (2007). *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Institute of Mathematical Statistics.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185.

Cesa-Bianchi, N., Conconi, A., and Gentile, C. (2004). On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057.

Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D. P., Schapire, R. E., and Warmuth, M. K. (1997). How to use expert advice. *Journal of the ACM*, 44(3):427–485.

Cesa-Bianchi, N., Gaillard, P., Lugosi, G., and Stoltz, G. (2012). Mirror descent meets fixed share (and feels no regret). In *Advances in Neural Information Processing Systems 25*, pages 980–988.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, New York, USA.

Cesa-Bianchi, N., Mansour, Y., and Stoltz, G. (2007). Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66:321–352.

Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity analysis in linear regression*, volume 327 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, New York.

Chaudhuri, K., Freund, Y., and Hsu, D. J. (2009). A parameter-free hedging algorithm. In *Advances in Neural Information Processing Systems 22*, pages 297–305.

Chernov, A. and Vovk, V. (2009). Prediction with expert evaluators' advice. In *Proceedings of the 20th conference on Algorithmic Learning Theory (ALT)*, pages 8–22.

Chernov, A. and Vovk, V. (2010). Prediction with advice of unknown number of experts. In *Proceedings of the 26th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 117–125.

Chernov, A. and Zhdanov, F. (2010). Prediction with expert advice under discounted loss. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 255–269.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Clarke, B. S. and Barron, A. R. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1):37–60.

Clémençon, S., Depecker, M., and Vayatis, N. (2013). Ranking forests. *Journal of Machine Learning Research*, 14(Jan):39–73.

Cover, T. M. (1968). Rates of convergence for nearest neighbor procedures. In *Proceedings of the Hawaii International Conference on Systems Sciences*, pages 413–415.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, New York, USA, 2nd edition.

Criminisi, A., Shotton, J., and Konukoglu, E. (2012). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision*, 7(2–3):81–227.

Cucker, F. and Smale, S. (2002a). Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computational Mathematics*, 2(4):413–428.

Cucker, F. and Smale, S. (2002b). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49.

Cui, Y., Zhu, R., Zhou, M., and Kosorok, M. (2017). Some asymptotic results of survival tree and forest models. *arXiv preprint arXiv:1707.09631*.

Cutler, A. and Zhao, G. (2001). PERT–Perfect random tree ensembles. *Computing Science and Statistics*, 33:490–497.

Dasgupta, S. and Freund, Y. (2008). Random projection trees and low dimensional manifolds. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 537–546. ACM.

de Rooij, S., van Erven, T., Grünwald, P., and Koolen, W. M. (2014). Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15:1281–1316.

De Vito, E., Caponnetto, A., and Rosasco, L. (2005). Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, 5(1):59–85.

Denil, M., Matheson, D., and de Freitas, N. (2013). Consistency of online random forests. In *Proceedings of the 30th Annual International Conference on Machine Learning (ICML)*, pages 1256–1264.

Denil, M., Matheson, D., and de Freitas, N. (2014). Narrowing the gap: Random forests in theory and in practice. In *Proceedings of the 31st Annual International Conference on Machine Learning (ICML)*, pages 665–673.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). A bayesian CART algorithm. *Biometrika*, 85(2):363–377.

Devroye, L. (1982). Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 61(4):467–481.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics*. Springer-Verlag.

Devroye, L. and Lugosi, G. (1995). Lower bounds in pattern recognition and learning. *Pattern recognition*, 28(7):1011–1018.

Devroye, L. and Wagner, T. (1979a). Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207.

Devroye, L. and Wagner, T. (1979b). Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604.

Dicker, L. H. (2016). Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1–37.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157.

Dieuleveut, A. and Bach, F. (2016). Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399.

Dobriban, E. and Wager, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279.

Domingos, P. and Hulten, G. (2000). Mining high-speed data streams. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 71–80.