An improper estimator with optimal excess risk in misspecified density estimation and logistic regression

Jaouad Mourtada* Stéphane Gaïffas†

December 20, 2019

Abstract

We introduce a procedure for predictive conditional density estimation under logarithmic loss, which we call SMP (Sample Minmax Predictor). This predictor minimizes a new general excess risk bound, which critically remains valid under model misspecification. On standard examples, this bound scales as d/n where d is the dimension of the model and n the sample size, regardless of the true distribution. The SMP, which is an improper (out-of-model) procedure, improves over proper (within-model) estimators (such as the maximum likelihood estimator), whose excess risk can degrade arbitrarily in the misspecified case. For density estimation, our bounds improve over approaches based on online-to-batch conversion, by removing suboptimal $\log n$ factors, addressing an open problem from Grünwald and Kotłowski [37] for the considered models. For the Gaussian linear model, the SMP admits an explicit expression, and its expected excess risk in the general misspecified case is at most twice the minimax excess risk in the *well-specified case*, but without any condition on the noise variance or approximation error of the linear model. For logistic regression, a penalized SMP can be computed efficiently by training two logistic regressions, and achieves a non-asymptotic excess risk of $O((d+B^2R^2)/n)$, where R is a bound on the norm of the features and B the norm of the optimal linear predictor. This improves the rates of proper (within-model) estimators, since such procedures can achieve no better rate than $\min(BR/\sqrt{n}, de^{BR}/n)$ in general [43]. This also provides a computationally more efficient alternative to approaches based on online-to-batch conversion of Bayesian mixture procedures, which require approximate posterior sampling, thereby partly answering a question by Foster et al. [32].

Keywords. Density estimation, Misspecified models, Statistical Learning Theory, Logistic regression, Improper prediction.

1 Introduction

Consider the standard problem of density estimation: given an i.i.d. sample $Z_1, \dots, Z_n$ from a distribution P on some measurable space $\mathcal{Z}$, the goal is to produce a good approximation $\widehat{P}_n$ of P. While it is typically impossible to obtain finite-sample guarantees without any assumption on the underlying distribution P, oftentimes one expects this distribution to possess some structure. In such cases, it is natural to introduce some inductive bias in the procedure; one standard way to do so is to select a suitable class of distributions (often called a *statistical model*) that is susceptible to capture at least part of the structure of P, and thus provide a non-trivial approximation thereof.

A classical approach to this problem is to assume that the considered class $\mathcal{F}$ contains the true distribution P, in which case the model is said to be *well-specified*. In this case, the problem of estimating P

*MaLGa research center, LCSL, DIBRIS, University of Genova, Italy. The research leading to this work was carried while the first author was a PhD student at CMAP, École polytechnique, France.

†LPSM, UMR 8001, Université Paris Diderot, Paris, France

falls within the classical framework of parametric statistics [44, 79, 53]. This theory provides strong support for the maximum likelihood estimator (MLE), which arises as an asymptotically optimal estimator for regular models as the sample size n grows. The same problem can also be treated for a fixed sample size, through the lens of statistical decision theory [84, 53], which emphasizes optimal estimators in the average (Bayesian) and minimax senses. Generally speaking, these approaches offer remarkably precise descriptions of the achievable rates of convergence (up to the correct leading constants) and of *efficient* estimators that make the best use of available data. One major limitation of this approach, however, is that such precise results require the very strong assumption that the true distribution belongs to the selected model. Such an assumption is generally unlikely to hold, since the model usually involves a simplified representation of the phenomenon under study: it comes from a choice of the statistician, who has no control over the true underlying mechanism.

A more realistic though still beneficial situation occurs when the underlying model captures some aspects of the true distribution, such as its most salient properties, but not all of them. In other words, the statistical model provides some non-trivial approximation of the true distribution, and is thus "wrong but useful". In such a case, a meaningful objective is to approximate the true distribution (namely, to predict its realizations) almost as well as the best distribution in the model. A natural setting to formalize this task is the framework of statistical learning theory [81], where one aims to find a good predictor (in a sense measured by some loss function) given access to an i.i.d. sample from a target distribution. Specifically, one makes few assumption about the true distribution, but instead introduces inductive bias through the choice of a comparison class of predictors against which one seeks to compete. When the predictors are densities over the space $\mathcal{Z}$ of observations, and the loss function is the *logarithmic* (or *log-likelihood*) loss (namely, the negative logarithm of the probability assigned to the given observation), the problem of statistical learning amounts to density estimation under *Kullback-Leibler* (KL) risk. Below, let us review some variants of (and approaches to) this problem.

Well-specified density estimation. This problem is well-studied, and a number of remarkably precise results have been obtained in the literature: admissible, Bayes optimal and minimax procedures, as well as higher-order asymptotics are studied, see, e.g. [39, 49, 40, 2, 54, 35, 75, 21]. In the case of a parametric family, the minimax and Bayes risk asymptotically scale as $d/(2n) + o(n^{-1})$. [40] considers the asymptotics of the KL risk for smooth parametric families, and determines its second-order expansion. [2] considers asymptotically second-order minimax procedures, namely Bayes predictive distributions which asymptotically minimize the maximum of the second-order term. [54] consider location and scale families of distributions, and show that the Bayes predictive distribution under uniform prior is minimax and best invariant. However, an important limitation of these approaches is that the true distribution is assumed to belong to the specified model, and no guarantee is provided when this strong assumption does not hold.

Misspecified density estimation. Bounds on the KL excess risk under misspecification have also been studied, although known results are typically weaker and require stronger assumptions than their well-specified counterparts. The considered estimator is usually the MLE, or some penalized variant of it. [15, 16, 88, 86, 73] obtain fast $O(d/n)$ excess risk bounds for such estimators. These approaches typically rely on the theory of empirical processes [80, 78, 77, 58, 47, 17], which often leads to loose constant factors and requires some strong boundedness or exponential tail conditions on the losses. A different approach is the one from [90, 38], who rely on information-theoretic inequalities. One property of the previous bounds is that they can degrade arbitrarily in the misspecified case. For instance, in a simple Gaussian model $\{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}$, while the true distribution is $\mathcal{N}(\theta^*, \sigma^2)$, the previous excess risk bounds scale at least as σ^2/n (which is the rate of the MLE), which features the potentially unbounded distribution-depend constant σ^2.

Limitations of results for misspecified density estimation. All the aforementioned bounds on predictive density estimation degrade under misspecification, since they exhibit constants which depend on the true distribution and can therefore be arbitrarily large. Hence, these bounds can be significantly worse than the "efficient" risk bound $d/(2n)+o(1/n)$ from the well-specified parametric case. Methods such as MLE inherently suffer from such limitations in density estimation. A first problem comes from the fact that the logarithmic loss is unbounded, while bounds for empirical risk minimization (ERM, which coincides with MLE for the logarithmic loss) typically depend on the range of the loss or on its tails. A second and more fundamental limitation of the MLE is the fact that its excess risk degrades under misspecification: its risk does not scale as d/n, even asymptotically and when restricted to a bounded neighborhood of the optimal parameter. Indeed, under proper regularity conditions on the model [79], the asymptotic excess risk of the MLE scales as $d_{\text{eff}}/(2n) + o(n^{-1})$, where the *effective dimension* d_{eff} depends on the true underlying distribution. This constant, which reflects the variability of the MLE, can be arbitrarily large compared to d, such as in the Gaussian example mentioned above (where σ^2 can be arbitrarily large).

Sequential prediction under logarithmic loss. By contrast, the theory of prediction under the log-loss is well-developed for the cumulative risk (or regret), where the aim is to sequentially choose probability densities on the next observation. In this case, one considers either the cumulative excess risk, namely the sum of the individual excess risks at each time step, or the regret [63, 29] on all sequences, namely the difference in cumulative loss with the best density in the family. This problem is connected to coding [31] and to the minimum description length (MDL) principle [69, 36]. For smooth, bounded parametric families of dimension d, the minimax cumulative excess risk and regret are known to scale as $(d \log n)/2 + C + o(1)$ for some constant C that depends on the model, see [30, 63]. Note that the minimax regret bound holds for any sequence, and in particular does not require the sequence of observations to be sampled from a distribution in the model. A generic procedure called *online to batch conversion* [55, 27] enables one to derive excess risk guarantees from regret bounds. These excess risk bounds apply to the average of the successive densities output by the sequential procedure, and are valid with only the i.i.d. assumption (even under model misspecification). When applied to the standard exponential weights online algorithm [82, 56] (which coincides with the Bayes mixture for the log-loss), a standard strategy that guarantees near-minimax regret [30, 87, 63, 29], this yields the so-called *progressive mixture rule* [89, 25, 26], see also [45, 4]. However, such approaches are usually too conservative and yield inefficient predictors, as illustrated by the extra logarithmic terms in the associated rates $O(d \log n/n)$ instead of $O(d/n)$ in the case of a parametric family. In addition, the resulting bounds exhibit stronger dependence on initial conditions and on global model complexity, and may in particular be infinite for some unbounded classes. This reflects the fact that the problem of controlling the regret is harder than that of controlling the excess risk, due to its cumulative nature, which takes into account early rounds where few observations are available.

Logistic regression. Logistic regression [13, 59, 79] is arguably the most common model for conditional density estimation with binary response. Assume that $\|X\| \leqslant R$ almost surely and one wishes to compare against logistic functions with $\|\beta\| \leqslant B$. Since the logistic loss is convex and R-Lipschitz, standard results in online and stochastic optimization [91, 71, 22, 41] show that a *slow rate* of $BR/\sqrt{n}$ is achievable by properly-tuned projected stochastic gradient descent on the ball of radius B. This slow rate can also be obtained through properly-tuned squared-norm regularization, using a standard stability argument. In this paper we show that that our Ridge-regularized SMP procedure achieves an improved bound under the same assumptions.

In addition, when restricting to the ball of radius B, the logistic loss is e^{-BR}-exp-concave. This shows that the Online Newton Step algorithm [42] achieves $O(de^{BR} \log n)$ regret, and thus $O(de^{BR} \log(n)/n)$ excess risk after online-to-batch conversion. Based on the notion of self-concordance [66], which appears in the analysis of Newton-type optimization algorithms [19], Bach [5] defines pseudo-self-concordant functions, a class of functions whose third derivative is controlled by the second one. Using the fact that the logistic loss satisfies this property, he then provides excess risk bounds for the Ridge-regularized MLE in fixed-design regression, both in the well-specified (with fast rate) and misspecified (with slow rate) cases. [6] proposes a stochastic algorithm whose excess risk is $O(R^2(BR+1)^4/(\mu n))$ based on the notion of generalized self-concordance, where μ is the smallest eigenvalue of the Hessian matrix at the unique optimum. [7] shows that the rate $O(\rho^3 d(BR+1)^4/n)$ can be achieved, where ρ is a problem-dependent curvature parameter, satisfying $\Sigma \preccurlyeq \rho H$, where H is the Hessian matrix at the optimum, and Σ the covariance matrix. [68] establishes a non-asymptotic fast rate bound for the MLE of $O(d_{\text{eff}} \log(1/\delta)/n)$ with probability $1-\delta$, which matches the asymptotic rate of this estimator. These results are extended in [57] to obtain upper bounds independent of the dimension but depending on gradient and Hessian norms. However, the bounds depend on problem-dependent constants ρ and d_{eff}, which can grow exponentially with BR in the worst-case scenario. [43] shows that such a dependence is unavoidable for any *proper* (within-model) algorithm, by establishing a lower bound of $\min(BR/\sqrt{n}, de^{BR}/n)$ in the worst-case stochastic setting, answering (by the negative) an open problem from [60].

This lower bound does not hold if we allow the predictor to be chosen outside of the class of logistic predictors. This observation leads [32] to propose an improper predictor based on Bayesian mixtures whose excess risk is $O(d\log(n)/n)$ with a non-exponential dependence in B and R, namely a rate of order $O(d\log(BRn)/n)$. An analysis in the online learning setting is also performed in [46], which proposes an analysis of the regret for both logistic regression and Gaussian linear density estimation, which leads to a per-round regret of $O(d\log(BRn)/n)$ in the logistic case.

Other related work. A general framework has recently been proposed for density estimation under model misspecification, namely ρ-estimation [8, 10, 9]. The setting differs from ours in the sense that accuracy is measured using the Hellinger risk (as opposed to the Kullback-Leibler risk in our case), hence the results are complementary but not directly comparable. The theoretical guarantees therein take the form of inexact oracle inequalities, which state that the risk of the estimator is at most a constant times that of the best within the model, plus some distribution-independent $O(d/n)$ term. This implies a $O(d/n)$ excess risk whenever the approximation error of the model is itself of order $O(d/n)$, which typically occurs in a nonparametric context. However, in the misspecified case considered here where no assumption is made on the approximation error, such guarantees do not imply convergence to the risk of the best element in the model. In fact, ρ-estimators are proper (within-model) estimators, and therefore subject to the lower bounds for proper estimators in the misspecified case. Hence, in order to achieve $O(d/n)$ Kullback-Leibler excess risk, different improper procedures are required. Finally, for the logistic regression problem (which was a main motivation for the present work), accuracy is usually measured in terms of the logistic (logarithmic) loss, rather than Hellinger risk.

2 General excess risk bounds

2.1 A general excess risk bound for statistical learning

In this section, we let $\mathcal{X}, \mathcal{Y}, \widehat{\mathcal{Y}}$ be three measurable spaces, which correspond respectively to the feature, label and prediction spaces. We also let $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ be a loss function. Finally, denote by $\widehat{\mathcal{F}}$ the space of all measurable functions $\mathcal{X} \to \widehat{\mathcal{Y}}$ (also called predictors), and let $\mathcal{F} \subset \widehat{\mathcal{F}}$ be a class of predictors. We consider also an arbitrary penalization function $\phi : \mathcal{F} \to \mathbb{R}$. Denote $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and let

$$\ell_\phi(f, z) = \ell(f(x), y) + \phi(f)$$

for any $z = (x, y) \in \mathcal{Z}$ and $f \in \mathcal{F}$. Whenever no penalization is used in the loss ($\phi \equiv 0$) we simply write $\ell = \ell_0$. Let P be some probability distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The quality of a predictor $g \in \widehat{\mathcal{F}}$ is measured through its *risk*

$$R(g) = \mathbb{E}[\ell(g, Z)] = \mathbb{E}[\ell(g(X), Y)] \tag{1}$$

where $Z = (X, Y) \sim P$, whenever this expectation is well-defined and belongs to $\mathbb{R} \cup \{+\infty\}$, which we assume from now on. Also, define the *excess risk* (with respect to $\mathcal{F}$) of g as

$$\mathcal{E}(g) = R(g) - \inf_{f \in \mathcal{F}} R(f) \,. \tag{2}$$

We define similarly $R_\phi(f) = \mathbb{E}[\ell_\phi(f, Z)]$ for $f \in \mathcal{F}$ and $\mathcal{E}_\phi(g) = R(g) - \inf_{f \in \mathcal{F}} R_\phi(f)$.

In this setting, the distribution P is unknown, and we will avoid making strong assumptions on it. The aim is to produce, given an i.i.d. sample $Z_1^n = (Z_1, \dots, Z_n)$ from P, a predictor $\widehat{g}_n : \mathcal{X} \to \widehat{\mathcal{Y}}$ whose expected excess risk $\mathbb{E}[\mathcal{E}(\widehat{g}_n)]$ (where the expectation holds over the random sample) is small. In other words, $\widehat{g}_n$ should predict almost as well as the best element in $\mathcal{F}$, up to a controlled small additional term. Given a sample $Z_1^n = (Z_1, \dots, Z_n)$, we denote

$$\widehat{f}_{\phi,n} \in \operatorname*{arg\,min}_{f \in \mathcal{F}} \sum_{i=1}^n \ell_\phi(f, Z_i) \tag{3}$$

a (penalized) *empirical risk minimizer* (ERM). Whenever $\phi \equiv 0$ we simply denote the ERM as $\widehat{f}_n$. Throughout this paper, we assume to simplify that this minimum is attained. This holds in virtually all the examples considered below; in addition, the arguments naturally extend to approximate minimizers. By convention, all minimizers of the empirical risk will be chosen symmetrically in the sample points $Z_1, \dots, Z_n$. We also introduce

$$\widehat{f}^z_{\phi,n} := \operatorname*{arg\,min}_{f \in \mathcal{F}} \left\{ \sum_{i=1}^n \ell_\phi(f, Z_i) + \ell_\phi(f, z) \right\} \tag{4}$$

for any $z \in \mathcal{Z}$. Theorem 1 below introduces a new bound on the excess risk of any prediction rule, together with a predictor that minimizes it. It holds for a general loss ℓ, but in the following sections we apply it to the logarithmic loss only, for which the predictor can be made explicit.

Theorem 1 (Main excess risk bound and Sample Minmax Predictor)**.** *For any predictor $\widehat{g}_n$ depending on Z_1^n, we have*

$$\mathbb{E}\big[\mathcal{E}_\phi(\widehat{g}_n)\big] \leqslant \mathbb{E}_{Z_1^n, X}\Big[\sup_{y \in \mathcal{Y}} \Big\{ \ell(\widehat{g}_n(X), y) - \ell_\phi(\widehat{f}^{(X,y)}_{\phi,n}(X), y) \Big\}\Big] \tag{5}$$

where $\widehat{f}^z_{\phi,n}$ is defined by (4) for $z \in \mathcal{Z}$ and $Z = (X, Y) \sim P$ is independent of Z_1^n. In addition, the right-hand side of (5) is minimized by the predictor

$$\widetilde{f}_{\phi,n}(x) = \operatorname*{arg\,min}_{\widehat{y} \in \widehat{\mathcal{Y}}} \sup_{y \in \mathcal{Y}} \Big\{ \ell(\widehat{y}, y) - \ell_\phi(\widehat{f}^{(x,y)}_{\phi,n}(x), y) \Big\} \,, \tag{6}$$

which we call SMP (Sample Minmax Predictor) whenever it exists, in which case (5) becomes

$$\mathbb{E}\big[\mathcal{E}_\phi(\widetilde{f}_{\phi,n})\big] \leqslant \mathbb{E}_{Z_1^n, X}\Big[\inf_{\widehat{y} \in \widehat{\mathcal{Y}}} \sup_{y \in \mathcal{Y}} \Big\{ \ell(\widehat{y}, y) - \ell_\phi(\widehat{f}^{(X,y)}_{\phi,n}(X), y) \Big\}\Big] \,. \tag{7}$$

The proof of Theorem 1 is given in Section 7.1. The excess risk bound of Theorem 1 is related to the stability of the (regularized) empirical risk minimizer. Indeed, if the ERM $\widehat{f}^{(X,y)}_{\phi,n}$ obtained by adding a new sample (X, y) does not depend too much on the label y, *i.e.* if the set $\{\widehat{f}_{\phi,n}^{(X,y)} : y \in \mathcal{Y}\}$ is small in expectation, then the min-max quantity in the bound (7) will also be small. The use of stability to establish guarantees for learning algorithms such as ERM or approximate ERM was pioneered by [18]. Stability arguments were used by [18, 72] to prove fast rates of order $O(1/n)$ for ERM in strongly convex stochastic optimization problems and more recently by [50] for exp-concave problems.

However, while related in spirit to the notion of stability, the excess risk bound of Theorem 1 differs from standard stability bounds. Indeed, approaches based on stability control the risk in terms of variations of the loss of the output hypothesis (such as ERM) under changes of the sample [18, 72, 74, 50]. By contrast, Theorem 1 controls the risk in terms of some min-max quantity, which measures the size of the set of empirical risk minimizers obtained by adding one sample.

It is worth noting that the SMP (6) whose risk is controlled in (7) is *not* the ERM, that is, the algorithm whose "stability" is controlled. In fact, $\widetilde{f}_{\phi,n}$ is in general an *improper* predictor, which does not belong to the class $\mathcal{F}$; it may be seen as a "center" of the set of risk minimizers obtained by adding one sample, in a sense related to the loss function.

In what follows, we show that for the logarithmic loss, the bound (7) improves over bounds based on stability of the loss, and in fact provides guarantees which are not achievable by proper predictors such as (penalized) ERM.

2.2 Conditional density estimation with the logarithmic loss

We now turn to the problem of conditional density estimation, which is the focus of this work, by considering the logarithmic loss. Let μ be a measure on $\mathcal{Y}$ and $\widehat{\mathcal{Y}}$ be the set of probability densities on $\mathcal{Y}$ with respect to μ, namely the set of measurable functions $f : \mathcal{Y} \to \mathbb{R}^+$ such that $\int_{\mathcal{Y}} f \mathrm{d}\mu = 1$. The logarithmic loss is defined as $\ell(f, y) = -\log f(y)$ for $f \in \widehat{\mathcal{Y}}$ and $y \in \mathcal{Y}$. In this setting, a predictor $f : \mathcal{X} \to \widehat{\mathcal{Y}}$ corresponds to a conditional density. We denote $f(y|x) = f(x)(y)$ and as before $\ell(f, z) = \ell(f(x), y)$ for $z = (x, y)$. Note that, in this case, the ERM (3) corresponds to the (conditional) maximum likelihood estimator (MLE). The risk of any conditional density f is

$$R(f) = -\mathbb{E}\big[\log f(Y|X)\big]$$

whenever this expectation is defined. Note that

$$R(g) - R(f) = \mathbb{E}\left[\log \frac{f(Y|X)}{g(Y|X)}\right] \tag{8}$$

for any conditional densities f and g with respect to μ, which only depends on the conditional distributions $f\mu$ and $g\mu$; and not on the measure μ which dominates them. In particular, we may choose μ such that the risk $R(f)$ is well-defined and finite for some $f \in \mathcal{F}$, and identify f and g with the corresponding conditional distributions. There exists a best predictor $f^* \in \mathcal{F}$ whenever the excess risk $\mathcal{E}(f) = \mathbb{E}[\ell(f, Z) - \ell(f^*, Z)]$ is defined and belongs to $[0, +\infty]$ for every $f \in \mathcal{F}$. Following what we did in Section 2.1, given a penalization function $\phi : \mathcal{F} \to \mathbb{R}$, we define the penalized risk R_ϕ and the penalized excess risk $\mathcal{E}_\phi$.

Theorem 2 below shows that both SMP defined in Theorem 1 and its excess risk bound (7) can be described explicitly in this case.

Theorem 2 (Excess risk bound for conditional density estimation)**.** *In the case of the logarithmic loss, the SMP $\widetilde{f}_{\phi,n}$ defined in* (6) *writes*

$$\widetilde{f}_{\phi,n}(y|x) = \frac{\widehat{f}_{\phi,n}^{(x,y)}(y|x) e^{-\phi(\widehat{f}_{\phi,n}^{(x,y)})}}{\int_{\mathcal{Y}} \widehat{f}_{\phi,n}^{(x,y')}(y'|x) e^{-\phi(\widehat{f}_{\phi,n}^{(x,y')})} \mu(\mathrm{d}y')}, \tag{9}$$

whenever the integral $\int_{\mathcal{Y}} \widehat{f}_{\phi,n}^{(X,y)}(y|X) e^{-\phi(\widehat{f}_{\phi,n}^{(X,y)})} \mu(\mathrm{d}y)$ *is finite almost surely (over* Z_1^n, X*). In addition, its excess risk bound* (7) *writes*

$$\mathbb{E}\big[\mathcal{E}_\phi(\widetilde{f}_{\phi,n})\big] \leqslant \mathbb{E}_{Z_1^n, X}\Big[\log\Big(\int_{\mathcal{Y}} \widehat{f}_{\phi,n}^{(X,y)}(y|X) e^{-\phi(\widehat{f}_{\phi,n}^{(X,y)})} \mu(\mathrm{d}y)\Big)\Big]. \tag{10}$$

Remark 1. In the non-regularized case where $\phi \equiv 0$, the SMP simply writes

$$\widetilde{f}_n(y|x) = \frac{\widehat{f}_n^{(x,y)}(y|x)}{\int_{\mathcal{Y}} \widehat{f}_n^{(x,y')}(y'|x)\mu(\mathrm{d}y')},$$

while its excess risk bound (10) takes the form:

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \mathbb{E}_{Z_1^n, X}\Big[\log\Big(\int_{\mathcal{Y}} \widehat{f}_n^{(X,y)}(y|X)\mu(\mathrm{d}y)\Big)\Big].$$

The proof of Theorem 2 is provided in Section 7.1. The SMP (9) minimizes, for every value of x, the worst-case (over $y \in \mathcal{Y}$) excess loss $\ell(\widetilde{f}_{\phi,n}(x), y) - \ell_\phi(\widehat{f}_{\phi,n}^{(x,y)}(x), y)$ with respect to the ERM on the sample $Z_1^n, (X, y)$. As explained above, the right-hand side of (10) corresponds to (the expectation of) a measure of complexity of the class $\{\widehat{f}_{\phi,n}^{(X,y)}, y \in \mathcal{Y}\}$ associated to the $\log$-loss. We will see below, in particular cases for $\mathcal{F}$, that despite being derived from a general bound for statistical learning, the excess risk bound of the SMP is remarkably tight and close to the optimal risk in the well-specified case. In fact, we will see in the case of the Gaussian linear model (Section 4.2) that the bound of the SMP is intrinsic to the hardness of the problem.

In the unconditional case, the prediction of the estimator (9) closely resembles that of a sequential prediction strategy called Sequential Normalized Maximum Likelihood (SNML), introduced by [70] and related to the Last Step Minimax algorithm (which restricts to proper predictions) from [76][1]. Interestingly, the motivation is completely different: the SNML algorithm was introduced as a computationally efficient relaxation of the minimax algorithm (in terms of cumulative regret) for sequential prediction under log-loss; its worst-case regret was shown to be almost minimax [51], and in fact minimax for some specific families [12]. By contrast, in our case the SMP estimator naturally arises as the minimizer of a novel upper bound on the *non-cumulative* excess risk.

3 Misspecified density estimation

In this section, we consider the problem of (unconditional) density estimation: the space $\mathcal{X}$ is assumed to be trivial (with a single element) and is therefore omitted, and no penalization is used ($\phi \equiv 0$). In other words, given access to an i.i.d. sample $(Y_1, \dots, Y_n)$ from a distribution P on $\mathcal{Y}$, and given a family $\mathcal{F}$ of probability densities on $\mathcal{Y}$ with respect to μ (namely, a statistical model $\mathcal{F}$), the aim is to find a predictive distribution $\widehat{g}_n$ on $\mathcal{F}$ whose excess risk with respect to $\mathcal{F}$ is as small as possible. Note that the model may be *misspecified*, in the sense that $P \notin \mathcal{F}$. Introduce the Kullback-Leibler (KL) divergence

$$\mathrm{KL}(P, Q) = \mathbb{E}_{Z\sim P}\Big[\log \frac{\mathrm{d}P}{\mathrm{d}Q}(Z)\Big]$$

between distributions P and Q (which is infinite whenever P is not absolutely continuous with respect to Q). If $\mathrm{KL}(P, f^*) < +\infty$ then $f^* = \arg\min_{f\in\mathcal{F}} \mathrm{KL}(P, f)$ and the excess risk (2) writes $\mathcal{E}(f) = \mathrm{KL}(P, f) - \mathrm{KL}(P, f^*)$ for any $f \in \mathcal{F}$. For this reason, the risk R is also called the *KL risk*.

[1] Specifically, the prediction of the SMP coincides with that of the SNML-1 algorithm from [70] at step $n + 1$, while the SNML-2 algorithm from [70] (simply called SNML in subsequent work [51, 12]) is slightly different: it minimizes the worst-case regret with respect to the next step ERM on the whole sequence, instead of just on the last sample.

In the next sections, we apply Theorem 2 to misspecified density estimation on standard families. In each case, the SMP is explicit and the excess risk bound scales as d/n irrespective of the true distribution P. These bounds are tight, since they are *within a factor of* 2 of the optimal asymptotic rate in the well-specified case. Also, we compare it with MLE and online to batch conversion using an optimal sequential prediction strategy from [27]. In all considered examples, SMP improves these strategies.

3.1 Finite alphabet: the multinomial family

In this section, we assume that $\mathcal{Y}$ is a finite set with d elements, μ is the counting measure and $\mathcal{F} = \{(p(y))_{y\in\mathcal{Y}} \in \mathbb{R}_+^{\mathcal{Y}} : \sum_{y\in\mathcal{Y}} p(y) = 1\}$, which corresponds to the multinomial model (which cannot be misspecified). For any $y \in \mathcal{Y}$, we let $N_n(y) = \sum_{i=1}^n \mathbf{1}(Y_i = y)$.

Proposition 1. *If $\mathcal{Y}$ is a finite set with d elements, then SMP corresponds to the* Laplace estimator

$$\widetilde{f}_n(y) = \frac{N_n(y)+1}{n+d}\,. \tag{11}$$

In addition, the bound (10) writes in this case

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \log\Big(\frac{n+d}{n+1}\Big) \leqslant \frac{d-1}{n}\,. \tag{12}$$

Proposition 1 is proved in Section 7.2. In this case, the SMP corresponds to the Laplace estimator, which is the Bayes predictive distribution under an uniform prior on $\mathcal{F}$. The first bound in (12) is tight: it is an equality when Y is constant almost surely.

About the MLE. The MLE is given by $\widehat{f}_n(y) = N_n(y)/n$ for the multinomial family. Its expected risk is infinite unless P is concentrated on a single point. Indeed, if P is not, then let $y_0, y_1 \in \mathcal{Y}$ be distinct elements such that $\mathbb{P}(Y = y_0), \mathbb{P}(Y = y_1) > 0$. Then, the event $E = \{Y_1 = \cdots = Y_n = y_0\}$ has positive probability. On E, we have $\widehat{f}_n(y) = \mathbf{1}(y = y_0)$, so that $\ell(\widehat{f}_n, y_1) = +\infty$ and thus $R(\widehat{f}_n) = +\infty$. Hence, $\mathbb{E}[R(\widehat{f}_n)] = +\infty$. In order to obtain non-vacuous expected risk bounds for the MLE in this case, one may restrict to $\mathcal{F}_\delta = \{p \in \mathcal{F} : \forall y \in \mathcal{Y},\ p(y) \geqslant \delta\}$ for some $\delta \in (0,1)$, so that the log ratios of the densities are bounded. In this case, whenever $p \in \mathcal{F}_\delta$, the excess risk of MLE has the asymptotically efficient rate $(d-1)/(2n) + o(n^{-1})$.

About online to batch conversion. The minimax cumulative regret with respect to the class $\mathcal{F}$ scales as $(d-1)(\log n)/2 + O(1)$ [29]. Hence, any upper bound based on online-to-batch conversion can be no better than $(d-1)(\log n)/(2n) + O(1/n)$, see [27].

3.2 The Gaussian location family

We now let $\mathcal{Y} = \mathbb{R}^d$ and consider the family of Gaussian distributions with fixed positive covariance matrix Σ, namely $\mathcal{F} = \{\mathcal{N}(\theta, \Sigma) : \theta \in \mathbb{R}^d\}$, that we call the Gaussian location family. We denote by $\|u\|$ the Euclidean norm of $u \in \mathbb{R}^d$ and $\bar{Y}_n := \frac{1}{n}\sum_{i=1}^n Y_i$.

Proposition 2. *A risk minimizer $f^* = \mathcal{N}(\theta^*, \Sigma) \in \mathcal{F}$ exists if and only if $\mathbb{E}\|Y\| < +\infty$, in which case $\theta^* = \mathbb{E}[Y]$. For $n \geqslant 1$, the SMP is given by $\widetilde{f}_n = \mathcal{N}(\bar{Y}_n, (1+1/n)^2\Sigma)$, and whenever $\mathbb{E}\|Y\| < +\infty$ the bound (10) writes*

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant d\log\Big(1+\frac{1}{n}\Big) \leqslant \frac{d}{n}\,. \tag{13}$$

In addition, when the model is well-specified, we have

$$\mathbb{E}[\mathcal{E}(\widetilde{f}_n)] = d\log\Big(1+\frac{1}{n}\Big) - \frac{d}{2n} < \frac{d}{2n}\,.$$

The proof of Proposition 2 is given in Section 7.2 below. It provides an excess risk bound valid not only under misspecification, but under the minimal hypothesis necessary to define the excess risk. In addition, this bound does not depend on the distribution of Y, and is essentially a factor of 2 above the optimal asymptotic risk $d/(2n)$ even for a worst-case distribution. In particular, this implies that finding a predictive distribution with small excess risk is feasible even when identifying the best parameter in the family is not: indeed, estimating θ^* with an accuracy independent of the true distribution of Y is not possible.

About the MLE and proper estimators. Assume that $\mathbb{E}\|Y\|^2 < +\infty$ and define $\Sigma_Y = \mathbb{E}[(Y - \mathbb{E}Y)(Y - \mathbb{E}Y)^\top]$ and $\|v\|_\Sigma^2 = v^\top \Sigma v$. The excess risk of the MLE $\widehat{f}_n = \mathcal{N}(\bar{Y}_n, \Sigma)$ is given by

$$\mathcal{E}(\widehat{f}_n) = \frac{1}{2}\mathbb{E}\big\|\bar{Y}_n - \mathbb{E}[Y]\big\|_{\Sigma^{-1}}^2 = \frac{1}{2n}\mathrm{tr}(\Sigma^{-1}\Sigma_Y).$$

In the misspecified case where $\Sigma_Y \neq \Sigma$, this quantity depends on the true distribution of Y and can be arbitrarily large depending on Σ_Y. For instance, if $\Sigma_Y = \sigma_Y^2 \Sigma$ with $\sigma_Y^2 > 0$, this quantity equals $\sigma_Y^2 d/(2n)$. In fact, this limitation is shared by any proper predictor $f_{\widehat{\theta}_n} = \mathcal{N}(\widehat{\theta}_n, \Sigma)$ for some estimator $\widehat{\theta}_n$, as explained next. Consider the family of distributions $\{P_{\theta^*} = \mathcal{N}(\theta^*, \Sigma_Y) : \theta^* \in \mathbb{R}^d\}$ for some arbitrary symmetric positive matrix Σ_Y, and the loss function $L(\theta^*, \theta) = \|\theta - \theta^*\|_{\Sigma^{-1}}^2/2$. It is a standard result in decision theory (see e.g. [53]) that the empirical mean $\bar{Y}_n$ is minimax optimal for this problem and has constant risk $\mathrm{tr}(\Sigma^{-1}\Sigma_Y)/(2n)$. Therefore, for any statistic $\widehat{\theta}_n$ (that is, any proper estimator $f_{\widehat{\theta}_n}$), we have for some $\theta^* \in \mathbb{R}^d$:

$$\mathbb{E}_{Y\sim P_{\theta^*}}\big[\mathcal{E}(f_{\widehat{\theta}_n})\big] = \frac{1}{2}\mathbb{E}\big\|\widehat{\theta}_n - \mathbb{E}[Y]\big\|_{\Sigma^{-1}}^2 \geqslant \frac{\mathrm{tr}(\Sigma^{-1}\Sigma_Y)}{2n}\,.$$

About online to batch conversion. The minimax cumulative regret with respect to the full Gaussian family $\mathcal{F}$ is infinite (see, e.g., [36]). This comes from the fact that the regret after the first step (the first prediction being made before seeing any sample) is unbounded. This difficulty does not appear in the batch setting, where one can predict conditionally on the sample, in a translation-invariant fashion. One can guarantee finite minimax regret by considering a restricted Gaussian family $\{\mathcal{N}(\theta, \Sigma) : \theta \in K\}$ for some compact set $K \subset \mathbb{R}^d$ [36], in which case the minimax regret scales as $d(\log n)/2 + C_K + o(1)$ (for some constant C_K depending on K) so that online to batch conversion yields an excess risk bound of $d(\log n)/(2n) + C_K/n + o(1/n)$, which again exhibits an extra $\log n$ factor.

Exact minimax rate in the misspecified case. In fact, for the Gaussian location family, the minimax excess risk in the general misspecified case, namely

$$\inf_{\widehat{g}_n} \sup_P \mathbb{E}_{Y\sim P}\big[\mathcal{E}(\widehat{g}_n)\big] \tag{14}$$

where the supremum spans over all probability distributions P on $\mathbb{R}^d$ such that $\mathbb{E}\|Y\|^2 < +\infty$, the infimum over density estimators $\widehat{g}_n$ and where the excess risk is under the true distribution P, can be determined exactly, together with a minimax predictor, as shown below.

Theorem 3. *For the Gaussian location model, the minimax excess risk* (14) *in the misspecified case* (*namely, over all distributions with finite second moment*) *is equal to*

$$\inf_{\widehat{g}_n} \sup_P \mathbb{E}_{Y\sim P}\big[\mathcal{E}(\widehat{g}_n)\big] = \frac{d}{2}\log\left(1 + \frac{1}{n}\right).$$

In addition, this minimax excess risk is achieved by the estimator $\widehat{f}_n = \mathcal{N}(\bar{Y}_n, (1 + 1/n)\Sigma)$*, which satisfies* $\mathbb{E}[\mathcal{E}(\widehat{f}_n)] = (d/2)\log(1 + 1/n)$ *for any distribution* P *of* Y *such that* $\mathbb{E}[\|Y\|^2] < +\infty$.

Theorem 3 is proven in Section 7.2 below. Note that $\widehat{g}_n$ corresponds to the Bayes predictive posterior under uniform prior, which is known to achieve the minimax risk in the *well-specified* case [67, 65], see also [35]. Remarkably, both the minimax excess risk and the minimax predictor remain the same in the misspecified case. This holds even though the posterior itself (a distribution on $\mathcal{F}$) does not concentrate on a neighborhood of the best parameter $\theta^* = \mathbb{E}[Y]$ in the misspecified case (contrary to the well-specified case), when the true variance is large. An explanation for this phenomenon is that the out-of-model correction of the Bayes predictive posterior (critically due to averaging over the posterior) brings it closer to distributions with high variance, thereby compensating the high variability for such distributions. As a result, the Bayes predictive posterior equalizes the excess risk across all distributions. This suggests that posterior concentration rates, which do not take into account the latter effect (and degrade under model misspecification when the true variance is large), are not enough to obtain the correct excess risk for predictive posteriors under model misspecification.

Finally, Theorem 3 shows that the worst-case excess risk bound (13) of the SMP is exactly twice the minimax excess risk for distributions with finite variance.

4 Gaussian linear conditional density estimation

In this section, we turn to conditional density estimation, starting with arguably the most standard family, namely the linear Gaussian model. After introducing the setting, notations and basic assumptions (Section 4.1), we consider the non-penalized SMP and its excess risk bounds with respect to the full unrestricted model (Section 4.2). Next, we consider in Section 4.3 the Ridge-regularized SMP and its performance, both in the finite-dimensional context and in the nonparametric one where d may be larger than n. In the latter case, the bounds only depend on the covariance structure of X and on the norm of the comparison element.

4.1 Setting: the Gaussian linear model

Consider the spaces $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$ and the family of conditional distributions

$$\mathcal{F} = \left\{ f_\beta(\cdot|x) = \mathcal{N}(\langle \beta, x\rangle, \sigma^2) : \beta \in \mathbb{R}^d \right\} \tag{15}$$

for some $\sigma^2 > 0$; up to the change of variables $y' = y/\sigma$, we will assume without loss of generality that $\sigma^2 = 1$. Throughout this section, we consider the log-loss with respect to the base measure $\mu = (2\pi)^{-1/2} \mathrm{d}y$ on $\mathbb{R}$, so that for $\beta \in \mathbb{R}^d$ and $(x, y) \in \mathbb{R}^d \times \mathbb{R}$:

$$\ell(f_\beta, (x, y)) = -\log f_\beta(y|x) = \frac{1}{2}(y - \langle \beta, x\rangle)^2 , \tag{16}$$

and hence the risk $R(\beta) := R(f_\beta)$ of f_β writes

$$R(\beta) = \frac{1}{2}\mathbb{E}\big[(Y - \langle \beta, X\rangle)^2\big] .$$

The problem of conditional density estimation in the Gaussian linear model is intimately linked (but not equivalent) to that of linear least-squares regression, namely statistical learning with the square loss and a comparison class formed by linear predictors. Let us discuss the connection and differences between the two problems:

- In the least-squares problem, one is interested in a *point prediction* of the response y given the covariates x, or equivalently in an estimate of the *conditional expectation* $\mathbb{E}[Y|X]$ of Y given X. By contrast, in the setting of density estimation one seeks a *probabilistic prediction* of y given x, or equivalently an estimate of the *conditional distribution* of Y given X, which therefore includes a quantification of the uncertainty of Y given X.

- When one restricts to proper, or well-specified conditional distributions (i.e. that belong to $\mathcal{F}$), the two problems are equivalent, as shown by the expression of the loss (16).
- On the other hand, in the context of conditional density estimation, the possibility of using improper (out-of-model) estimators provides more flexibility. As we will see, this additional flexibility is essential to bypass lower bounds for proper estimators in the misspecified case.

Let us emphasize that in the context of conditional density estimation, well-specification refers to the fact that the conditional distribution of Y given X belongs to the model. As in the unconditional case, we are interested in bounds that do not degrade under model misspecification, and hence require only weak assumptions on this conditional distribution. Assumption 1 below will be made throughout this section, while further assumptions will be made in Sections 4.2 and 4.3 respectively.

Assumption 1 (Finite second moments)**.** We assume that both X and Y are square integrable, namely

$$\mathbb{E}\|X\|^2 < +\infty \quad \text{and} \quad \sigma_Y^2 := \mathbb{E}[Y^2] < +\infty.$$

We will denote $\Sigma = \Sigma_X = \mathbb{E}[XX^\top]$ the second-order moment matrix, which we will call (following a common abuse of terminology) the *covariance matrix* of X, even when X is not centered. Assumption 1 implies that YX is integrable (by the Cauchy-Schwarz inequality) and that $\mathbb{E}[\langle \beta, X\rangle^2] = \langle \Sigma\beta, \beta\rangle$, so that the risk $R(\beta)$ is finite[2] and equals:

$$R(\beta) = \frac{1}{2}\langle \Sigma\beta, \beta\rangle - \langle \beta, \mathbb{E}[YX]\rangle + \frac{1}{2}\mathbb{E}[Y^2]\,,$$

with gradient $\nabla R(\beta) = \Sigma\beta - \mathbb{E}[YX]$. In particular, whenever Σ is invertible, the population risk minimizer $f^* \in \mathcal{F}$ is given by $f^* = f_{\beta^*}$ with $\beta^* = \Sigma^{-1}\mathbb{E}[YX]$, while the excess risk of $f_\beta \in \mathcal{F}$ writes $\mathcal{E}(f_\beta) = \frac{1}{2}\|\beta - \beta^*\|_\Sigma^2$. Likewise, whenever the empirical covariance matrix

$$\widehat{\Sigma}_n := \frac{1}{n}\sum_{i=1}^n X_i X_i^\top \tag{17}$$

is invertible, there exists a unique empirical risk minimizer given by

$$\widehat{\beta}_n = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \sum_{i=1}^n (Y_i - \langle \beta, X_i\rangle)^2 = \widehat{\Sigma}_n^{-1}\widehat{S}_n \tag{18}$$

where $\widehat{S}_n = n^{-1}\sum_{i=1}^n Y_i X_i$. Hence, whenever $\widehat{\Sigma}_n$ is invertible (almost surely), the MLE is uniquely defined, and equals the *ordinary least squares* estimator given by (18).

4.2 The unregularized SMP

In this section, we consider uniform excess risk bounds for unpenalized SMP ($\phi \equiv 0$) with respect to the linear Gaussian class $\mathcal{F}$ given by (15). This setting is relevant when $n \gg d$, especially when little is known or assumed on the optimal parameter β^*. We will work under the following

Assumption 2 (Non-degenerate design)**.** The covariance matrix Σ is invertible and the empirical covariance matrix $\widehat{\Sigma}_n$ is invertible almost surely.

[2] The assumption $\mathbb{E}[Y^2] < +\infty$ is not strictly necessary to ensure that $R(\beta)$ is finite for some base measure μ. Indeed, taking $\mu = \mathcal{N}(0,1)$, the log-loss writes $\ell(f_\beta, (x,y)) = \langle \beta, x\rangle^2/2 - y\langle \beta, x\rangle$, and one may replace Assumption 1 by the slightly weaker assumption that YX is integrable. We choose to take a uniform dominating measure μ and to consider Assumption 1 nonetheless, in order to make the connection with the least-squares problem more explicit.

The fact that Σ is invertible amounts to assuming that X is not supported in any hyperplane of $\mathbb{R}^d$. This assumption is not restrictive, since otherwise one can simply restrict to the span of the support of X, a subspace of $\mathbb{R}^d$; we make it merely for convenience in the statements and notations. In addition, a simple induction (see [64]) shows that Assumption 2 is equivalent to assuming that $n \geqslant d$ and that $\mathbb{P}(X \in H) = 0$ for any hyperplane $H \subset \mathbb{R}^d$. Note that the latter is granted whenever X admits a density with respect to the Lebesgue measure. Moreover, as explained in Section 4.1, Assumption 2 amounts to say that the MLE is uniquely determined almost surely in the family (15).

Once again in this case, the SMP leads to an improper estimator, which can be made explicit, and satisfies a sharp excess risk bound. Let us introduce also the rescaled empirical covariance matrix

$$\widetilde{\Sigma}_n = \Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_i\widetilde{X}_i^\top \quad \text{where} \quad \widetilde{X}_i = \Sigma^{-1/2}X_i. \tag{19}$$

Note that the rescaled design $\widetilde{X}_i$ is such that $\mathbb{E}[\widetilde{X}_i\widetilde{X}_i^\top] = I_d$ for $i = 1, \dots, n$. As explained in Theorem 4 below, the excess risk of SMP is deeply connected to the spectrum of $\widetilde{\Sigma}_n$.

Theorem 4. *Assume that Assumptions 1 and 2 are fulfilled. For the Gaussian linear family $\mathcal{F}$ given by (15), the SMP is given by*

$$\widetilde{f}_n(\cdot|x) = \mathcal{N}\Big(\langle\widehat{\beta}_n, x\rangle, \big(1 + \langle (n\widehat{\Sigma}_n)^{-1}x, x\rangle\big)^2\Big). \tag{20}$$

In addition, it satisfies the following excess risk bound:

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \mathbb{E}\Big[-\log\Big(1 - \big\langle (n\widehat{\Sigma}_n + XX^\top)^{-1}X, X\big\rangle\Big)\Big] \leqslant \log\Big(1 + \frac{1}{n}\mathbb{E}\big[\mathrm{tr}(\widetilde{\Sigma}_n^{-1})\big]\Big), \tag{21}$$

where $\widetilde{\Sigma}_n$ is the rescaled empirical covariance given by (19).

The proof of Theorem 4 is given in Section 7.3 below. The upper bound on the excess risk depends on the distribution of the design through the term $\mathbb{E}[\mathrm{tr}(\widetilde{\Sigma}_n^{-1})]$, namely through the distribution of the spectrum of the rescaled empirical covariance matrix (19). Note that this quantity is invariant by linear transformation of $X, X_1, \dots, X_n$, namely it does not depend on the choice of an inner product on $\mathbb{R}^d$.

A key feature of the excess risk bound (21) on the SMP is that it only depends on the distribution of X, and *not* on the conditional distribution of Y given X. The expected risk of the SMP is therefore not affected by model misspecification, similarly to what was observed in Section 3 for unconditional densities. This is once again a strong departure from the behavior of the MLE, as explained below.

Comparison with the MLE and proper estimators. As explained above, the MLE is given in this setting by $f_{\widehat{\beta}_n}$ where $\widehat{\beta}_n$ is the ordinary least-squares estimator (18). In the *well-specified case*, the minimax risk among *proper* estimators is achieved by $f_{\widehat{\beta}_n}$ and is equal to $\mathbb{E}[\mathrm{tr}(\widetilde{\Sigma}_n^{-1})]/(2n)$ [64]. Namely, the excess risk of SMP is *only within a factor* 2 of the exact minimax risk for proper estimators in the well-specified case, despite the fact that the model can be misspecified. In the *misspecified case*, the risk of the MLE scales as $\mathbb{E}_{(X,Y)\sim P}[(Y - \langle\beta^*, X\rangle)^2\|\Sigma^{-1/2}X\|^2]/n$ up to lower-order terms, and this dependence is unavoidable for any proper estimator [64]. This means that the risk of proper estimators deteriorates under misspecification, and that the minimax risk among proper estimators is infinite, since a supremum over P can make this quantity arbitrarily large.

Comparison with the well-specified case. One can in fact show that the first bound in (21) on the risk of the SMP in the general misspecified case is exactly twice the minimax excess risk in the well-specified case. This shows that this bound is intrinsic to the difficulty of the problem, and that the minimax excess risk in the misspecified case is at most twice that of the well-specified case.

Comparison with online algorithms. The minimax regret with respect to the full linear model is infinite, since the regret after the first observation is unbounded. Hence, it is not possible to obtain any uniform excess risk bound from online-to-batch conversion of sequential procedures. We discuss non-uniform guarantees in Section 4.3.

Link with leverage scores. It is worth noting that the first part of the upper bound (21) has a nice statistical interpretation. Indeed, the quantity $\langle (n\widehat{\Sigma}_n + XX^\top)^{-1}X, X\rangle$ is the *leverage score* of X in the sample $X_1, \dots, X_n, X$. This means that the excess risk of SMP can be upper bounded as

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \mathbb{E}\big[-\log(1-\widehat{\ell}_{n+1})\big], \quad \text{where} \quad \widehat{\ell}_{n+1} = \left\langle \Big(\sum_{i=1}^{n+1} X_i X_i^\top\Big)^{-1} X_{n+1}, X_{n+1} \right\rangle$$

is the leverage score of one sample distributed as P_X among $n+1$. Intuitively, the more uneven the leverage scores are, the harder the prediction task will be, since the optimal parameter in the model will effectively be determined by smaller number of points and hence have larger variance.

Upper bounds. A first upper bound on the risk of the SMP can be obtained from (21) in the case of Gaussian covariates: when $X \sim \mathcal{N}(0,\Sigma)$, so that $\widetilde{X} \sim \mathcal{N}(0, I_d)$, we have $\mathbb{E}[\mathrm{tr}(\widetilde{\Sigma}_n^{-1})] = nd/(n-d-1)$ [1, 20], so that an upper bound for SMP is given by $\log(1+d/(n-d-1)) \leqslant d/(n-d-1)$.

We now discuss extensions to more general distributions P_X of covariates. By the law of large numbers, one has $\widetilde{\Sigma}_n \to I_d$ as $n \to \infty$ and thus $\mathrm{tr}(\widetilde{\Sigma}_n^{-1}) \to d$ almost surely. Hence, one can expect that the excess risk bound (21) of the SMP scales as $d/n + o(1/n)$. In order to turn this into an explicit, non-asymptotic bound, we need to control the lower tail of $\widetilde{\Sigma}_n$. This requires some (weak) moment assumption on $\widetilde{X} = \Sigma^{-1/2}X$ (see (23) below) together with an assumption strengthening and "quantifying" Assumption 2, namely that $\mathbb{P}(X \in H) = 0$ for any hyperplane $H \subset \mathbb{R}^d$ (see (22)).

Assumption 3. There are constants $C \geqslant 1$ and $\alpha \in (0,1]$ such that

$$\mathbb{P}(|\langle \theta, X\rangle| \leqslant t\|\theta\|_\Sigma) \leqslant (Ct)^\alpha \tag{22}$$

for any $\theta \in \mathbb{R}^d \setminus \{0\}$ and $t > 0$. Moreover, there is $\kappa \geqslant 1$ such that

$$\mathbb{E}\|\Sigma^{-1/2}X\|^4 \leqslant \kappa d^2. \tag{23}$$

Note that inequality (22) is equivalent to $\mathbb{P}(|\langle \theta, \widetilde{X}\rangle| \leqslant t) \leqslant (Ct)^\alpha$ for any $\theta \in S^{d-1}$ (Euclidean unit sphere in $\mathbb{R}^d$), which is a quantification of how close $\widetilde{X}$ is from any hyperplane $H = \{v \in \mathbb{R}^d : \langle \theta, v\rangle = 0\}$ for any $\theta \in S^{d-1}$. It is a strengthened version of the *small-ball condition* considered in [48, 62], see also [52], since it is required at all levels $t > 0$. Inequality (23) is a bound on the kurtosis of $\|\Sigma^{-1/2}X\|$. It is weaker than the following L^2–L^4 equivalence for all marginals of X: $(\mathbb{E}\langle X,\theta\rangle^4)^{1/4} \leqslant \kappa^{1/4}(\mathbb{E}\langle X,\theta\rangle^2)^{1/2}$ for all $\theta \in \mathbb{R}^d$, which implies (23) using a simple argument, see [64].

Corollary 1. *Suppose that Assumptions 1, 2 and 3 hold, and let $\widetilde{f}_n$ be the SMP for the family (15) given by (20). Then, we have*

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \frac{d}{n}\Big(1 + C'\frac{\kappa d}{n}\Big) \tag{24}$$

for $n \geqslant \min(6d/\alpha, 12\log(12/\alpha)/\alpha)$, where $C' = 28C^4 e^{1+9/\alpha}$ and where we recall that C, α come from Assumption 3.

The proof of Corollary 1 is given in Section 7. It is a direct consequence of Theorem 4, together with an upper bound from [64] on the excess risk of least-squares in the well-specified case. The bound (24) scales as $d/n + O((d/n)^2)$ as $n \to \infty$, and is therefore sharp up to constants, with an expected second order term $O((d/n)^2)$. The most technical argument is provided in [64], where a sharp control on the smallest eigenvalue of $\widetilde{\Sigma}_n$ and on $\mathrm{tr}(\widetilde{\Sigma}_n^{-1})$ is provided under Assumption 3.

4.3 Kernel Ridge SMP

We now turn to non-uniform bounds over the class $\mathcal{F}$, where some dependence on the comparison parameter $\beta \in \mathbb{R}^d$ is allowed. In particular, we will consider the SMP with Ridge regularization $\phi(\beta) = \lambda\|\beta\|^2/2$ for some $\lambda > 0$. While the bounds obtained in this setting lack the uniformity of those from the previous section, they remain meaningful in the *nonparametric* setting where d may be larger than n.

The upper bound from Theorem 5 below does not explicitly depend on the dimension d, but only on the covariance matrix Σ and on β. It extends readily to the case where $\mathbb{R}^d$ is replaced by a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$, but we will keep $\mathbb{R}^d$ in order to keep the setting and notations consistent with those of Section 4.2. We will work in this section under the following assumption.

Assumption 4 (Bounded covariates)**.** There is a constant $R > 0$ such that $\|X\| \leqslant R$ almost surely.

Assumption 4 is automatically satisfied for instance in the Reproducing Kernel Hilbert Space (RKHS) setting, where the features x are of the form $x = \Phi(x')$ where $x' \in \mathcal{X}'$ is an input variable in some measurable space $\mathcal{X}'$ and $\Phi : \mathcal{X}' \to \mathbb{R}^d$ a measurable map such that the kernel $K : \mathcal{X}' \times \mathcal{X}' \to \mathbb{R}$ given by $K(x', x'') = \langle \Phi(x'), \Phi(x'') \rangle$ is bounded: $K \leqslant R^2$.

Let us recall that we consider the family $\mathcal{F} = \{f_\beta(\cdot|x) = \mathcal{N}(\langle \beta, x \rangle, 1) : \beta \in \mathbb{R}^d\}$, together with the Ridge penalization $\phi(\beta) = \lambda\|\beta\|^2/2$ for some $\lambda > 0$. Let

$$\widehat{\beta}_{\lambda,n} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(f_\beta, (X_i, Y_i)) + \frac{\lambda}{2}\|\beta\|^2 \right\} = (\widehat{\Sigma}_n + \lambda I_d)^{-1} \widehat{S}_n$$

denote the Ridge estimator, where we recall that $\widehat{\Sigma}_n = n^{-1} \sum_{i=1}^n X_i X_i^\top$ and $\widehat{S}_n = n^{-1} \sum_{i=1}^n Y_i X_i$, and let us also define

$$\widehat{\Sigma}_\lambda^x = n\widehat{\Sigma}_n + xx^\top + \lambda(n+1) I_d, \quad \widehat{K}_\lambda^x = (\widehat{\Sigma}_\lambda^x)^{-1} \quad \text{and} \quad \lambda' = \frac{n+1}{n}\lambda$$

where $\|x\|_S^2 = \langle x, x \rangle_S = \langle Sx, x \rangle$ for a matrix $S \succ 0$. We also introduce the *degrees of freedom* of the Ridge estimator [83, 33, 85], given by

$$\mathsf{df}_\lambda(\Sigma) = \operatorname{tr}[(\Sigma + \lambda I_d)^{-1} \Sigma] \,, \tag{25}$$

and note that

$$\mathsf{df}_\lambda(\Sigma) \leqslant \operatorname{tr}[(\Sigma + \lambda I_d)^{-1}(\Sigma + \lambda I_d)] = d \,. \tag{26}$$

Theorem 5. *Let $\lambda > 0$. The penalized SMP (9) with penalty $\phi(\beta) = \frac{\lambda}{2}\|\beta\|^2$ is well-defined and writes $\widetilde{f}_{\lambda,n}(\cdot|x) = \mathcal{N}(\widetilde{\mu}_\lambda(x), \widetilde{\sigma}_\lambda^2(x))$, where*

$$\widetilde{\sigma}_\lambda(x)^2 = \left((1 - \|x\|^2_{\widehat{K}_\lambda^x})^2 + \lambda \|x\|^2_{(\widehat{K}_\lambda^x)^2} \right)^{-1} \tag{27}$$

and

$$\widetilde{\mu}_\lambda(x) = \langle \widehat{\beta}_{\lambda',n}, x \rangle - \lambda \widetilde{\sigma}_\lambda(x)^2 \langle \widehat{\beta}_{\lambda',n}, x \rangle_{\widehat{K}_\lambda^x}. \tag{28}$$

In addition, under Assumptions 1 and 4, we have

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\beta \in \mathbb{R}^d} \left\{ R(\beta) + \frac{\lambda}{2}\|\beta\|^2 \right\} \leqslant 1.25 \cdot \frac{\mathsf{df}_\lambda(\Sigma)}{n+1} \tag{29}$$

for every $\lambda \geqslant 2R^2/(n+1)$, where $\mathsf{df}_\lambda(\Sigma)$ is given by (25).

Although the space of parameters is finite dimensional (of dimension d), the bound (29) is "non-parametric" in the sense that it does not feature any explicit dependence on d; rather, it only depends on the spectral properties of Σ through $\mathsf{df}_\lambda(\Sigma)$. In particular, it remains nonvacuous even when $d \gg n$; in fact, as mentioned above, Theorem 5 remains valid (with essentially the same proof, up to slight changes in terminology and notations) in the case of an infinite-dimensional RKHS.

In the finite-dimensional case where $n \gg d$, one can improve the quadratic dependence on the norm $B = \|\beta\|$. This yields bounds that are appropriate when the covariate distribution is possibly degenerate, in the sense that Assumption 2 does not hold, so that excess risk bounds uniform in β are no longer achievable.

Proposition 3. *Grant Assumptions 1 and 4. Then, for any $B > 0$, the SMP $\widetilde{f}_{\lambda,n}$ of Theorem 5 with $\lambda = d/(B^2(n+1))$ satisfies*

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\beta \in \mathbb{R}^d \,:\, \|\beta\| \leqslant B} R(\beta) \leqslant \frac{5d \log\big(2 + BR/\sqrt{d}\big)}{n+1} \,. \tag{30}$$

This bound is of order $O(d \log(BR/\sqrt{d})/n)$. It is comparable to the bound from [46], when combined with online to batch conversion, but removes a superfluous $O(\log n)$ term.

Finite-dimensional case. Since $\mathsf{df}_\lambda(\Sigma) \leqslant d$ (see (26)), Theorem 5 entails, for $\lambda = 2R^2/(n+1)$ (recalling that $\|X\| \leqslant R$ a.s.), that

$$\mathbb{E}[R(\widetilde{f}_{\lambda,n})] - R(\beta) \leqslant \frac{1.25d + R^2\|\beta\|^2}{n+1} \,. \tag{31}$$

In particular, this gives an excess risk bound of $O((d + B^2R^2)/n)$ on the ball of radius B, for every $B > 0$. Proposition 3 gives an improved bound in this setting.

Slow, dimension-free rate. Since $\mathsf{df}_\lambda(\Sigma) \leqslant R^2/\lambda$ for $\lambda > 0$, Theorem 5 yields, for every $\lambda \geqslant 2R^2/(n+1)$

$$\mathbb{E}[R(\widetilde{f}_{\lambda,n})] - R(\beta) \leqslant \frac{cR^2}{\lambda(n+1)} + \frac{\lambda\|\beta\|^2}{2} \leqslant \frac{2R\|\beta\|}{\sqrt{n}} + \frac{R^2\|\beta\|^2}{n}, \tag{32}$$

where the second inequality is obtained with $\lambda = \max(2R^2/(n+1), 2R/(\|\beta\|\sqrt{n+1}))$. It looks like the standard slow rate for regression, except that it does not depend on the range of Y. This requires no assumption on the covariance Σ, except the inequality $\mathrm{tr}(\Sigma) \leqslant R^2$ induced by the assumption $\|X\| \leqslant R$.

Nonparametric case. More precise results can be obtained in terms of the spectral properties of Σ. Let b be the rate of decay of the eigenvalues of Σ such that $\mathsf{df}_\lambda(\Sigma) = O(\lambda^{-1/b})$. Then, Theorem 5 yields

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\|\beta\| \leqslant B} R(\beta) \leqslant O\Big(\frac{\lambda^{-1/b}}{n} + \lambda B^2\Big) = O(B^{2/(b+1)} n^{-b/(b+1)}) \tag{33}$$

for $\lambda \asymp (B^2 n)^{-b/(b+1)}$. This matches the standard rate for regression over balls of RKHSs in the case of unit noise, without any additional assumption on β^* [24].

5 Logistic regression

In this section we consider the problem of conditional density estimation for the logistic model, for binary classification. Section 5.1 introduces the setting. We consider the unpenalized SMP ($\phi \equiv 0$) in Section 5.2 and discuss qualitatively the predictions produced by it, compared to the MLE. this section provides also a general control of its excess risk. Section 5.3 considers the Logistic SMP procedure with Ridge penalization, for which an explicit non-asymptotic control of the excess risk can be provided.

5.1 Setting

We consider the problem of binary classification, for which $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{-1, 1\}$ and $\mu = \delta_0 + \delta_1$ is the counting measure on $\mathcal{Y}$. We consider the family of *logistic* conditional distributions given by

$$\mathcal{F} = \{f_\beta : \beta \in \mathbb{R}^d\}, \quad \text{where} \quad f_\beta(1|x) := 1 - f_\beta(-1|x) = \sigma(\langle \beta, x\rangle) \tag{34}$$

for any $x \in \mathbb{R}^d$ and where $\sigma(u) = 1/(1 + e^{-u})$ for $u \in \mathbb{R}$ is the *sigmoid* function. In what follows, the family of conditional densities (34) is called the *logistic model*. Since $\sigma(-u) = 1-\sigma(u)$, one simply has $f_\beta(y|x) = \sigma(y\langle \beta, x\rangle)$ for $x \in \mathbb{R}^d$ and $y \in \{-1, 1\}$. In this section, we will often represent the sample (x, y) by the quantity $z = -yx \in \mathbb{R}^d$. The log-loss of $f_\beta \in \mathcal{F}$ at a sample $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ writes

$$\ell(f_\beta, (x, y)) = -\log f_\beta(y|x) = \log(1 + e^{-y\langle \beta, x\rangle}) = \log(1 + e^{\langle \beta, z\rangle}) = \ell(\langle \beta, z\rangle)\,, \tag{35}$$

where we introduced the *logistic* loss $\ell(u) = \log(1 + e^u)$ for $u \in \mathbb{R}$.

Let P be some distribution on $\mathbb{R}^d \times \{-1, 1\}$ and let (X, Y) be a generic pair such that $(X, Y) \sim P$. Also, denote $Z = -Y X$ and assume that $\mathbb{E}\|X\| < +\infty$. Since $\ell'(u) = \sigma(u) \leqslant 1$ for any $u \geqslant 0$ we have $\ell(u) \leqslant \log 2 + |u|$ for any $u \in \mathbb{R}$, so that

$$\ell(\langle \beta, Z\rangle) \leqslant \log 2 + |\langle \beta, Z\rangle| \leqslant \log 2 + \|\beta\|\|X\|\,,$$

and the risk of f_β, namely

$$R(\beta) = \mathbb{E}[\ell(\langle \beta, Z\rangle)]\,, \tag{36}$$

is well-defined. Given a sample (X_i, Y_i), $1 \leqslant i \leqslant n$, an MLE $\widehat{\beta}_n$ is given by

$$\widehat{\beta}_n \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \sum_{i=1}^n \ell(\langle \beta, Z_i\rangle)\,, \tag{37}$$

with $Z_i = -Y_i X_i$. A MLE (37) does not always exist, and may not be unique. Indeed, a well-known fact (see [23] for recent results on this topic in the high-dimensional regime) is that there is no MLE (37) whenever the sets $\{X_i : Y_i = 1\}$ and $\{X_i : Y_i = -1\}$ are strictly *linearly separated* by a hyperplane, namely when one can find $\beta \in \mathbb{R}^d$ such that $-Y_i\langle \beta, X_i\rangle = \langle \beta, Z_i\rangle < 0$ for all $i = 1, \dots, n$, since in this case the empirical risk of $t\beta$ converges to 0 as $t \to +\infty$, while the empirical risk is positive. In addition, when one MLE exists in $\mathbb{R}^d$, one can see that it is unique if and only if $V = \mathrm{span}(X_1, \dots, X_n) = \mathbb{R}^d$: in this case the empirical risk is strictly convex on $\mathbb{R}^d$, since $\ell : \mathbb{R} \to \mathbb{R}$ is, while it is constant on $V^\perp$.

In light of this discussion, it is convenient to enlarge the class $\mathcal{F}$ given by (34) into a larger class $\overline{\mathcal{F}}$ in order to ensure the existence of a MLE in all cases. Specifically, consider elements of the form $\beta := (+\infty, \theta)$ for $\theta \in S^{d-1} = \{v : \mathbb{R}^d : \|v\| = 1\}$, and set $f_\beta(1|x) = 1$ if $\langle \theta, x\rangle > 0$, $f_\beta(1|x) = 0$ if $\langle \theta, x\rangle < 0$ and $f_\beta(1|x) = 1/2$ if $\langle \theta, x\rangle = 0$ (this choice of notations comes from the fact that $f_{t\theta}(1|x) \to f_{(+\infty,\theta)}(1|x)$ as $t \to +\infty$ for every $x \in \mathbb{R}^d$). By a slight abuse of notations, we denote $\sigma(\langle \beta, x\rangle) := f_\beta(1|x)$ for β of the previous form. The class $\overline{\mathcal{F}}$ obtained by adding these conditional distributions to $\mathcal{F}$ is such that a MLE always exists in $\overline{\mathcal{F}}$ but is generally not unique. Given one choice of MLE, we let

$$\begin{aligned} \widehat{\beta}_n^{(x,y)} &= \operatorname*{arg\,min}_{\beta \in \overline{\mathcal{F}}} \Big\{ \sum_{i=1}^n \ell(f_\beta, (X_i, Y_i)) + \ell(f_\beta, (x, y)) \Big\} \\ &= \operatorname*{arg\,min}_{\beta \in \overline{\mathcal{F}}} \Big\{ \sum_{i=1}^n \ell(\langle \beta, Z_i\rangle) + \ell(\langle \beta, z\rangle) \Big\} =: \widehat{\beta}_n^z \end{aligned} \tag{38}$$

for any $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ and $z = -yx$.

5.2 The SMP for logistic regression

Let us now instantiate the SMP as well as Theorem 2 to the logistic family.

Proposition 4. *For the family of logistic conditional distributions* (34)*, the SMP writes*

$$\widetilde{f}_n(y|x) = \frac{f_{\widehat{\beta}_n^{(x,y)}}(y|x)}{f_{\widehat{\beta}_n^{(x,1)}}(1|x) + f_{\widehat{\beta}_n^{(x,-1)}}(-1|x)} = \frac{\sigma(\langle \widehat{\beta}_n^{(x,y)}, yx\rangle)}{\sigma(\langle \widehat{\beta}_n^{(x,1)}, x\rangle) + \sigma(\langle \widehat{\beta}_n^{(x,-1)}, -x\rangle)} \tag{39}$$

for every $x \in \mathbb{R}^d$ and $y \in \{-1, 1\}$. Unlike the MLE (38)*, the SMP is always well-defined and unique. We always have that $\widetilde{f}_n(y|x) \in (0, 1)$ and it does not depend on the choice of a MLE in the linearly separated case. In addition, it satisfies the following excess risk bound*:

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \mathbb{E}_{Z_1^n,Z}\big[\sigma(\langle \widehat{\beta}_n^{-Z}, Z\rangle) - \sigma(\langle \widehat{\beta}_n^{Z}, Z\rangle)\big]\,, \tag{40}$$

where $Z_1, \dots, Z_n, Z$ are i.i.d. variables distributed as $-YX$.

The proof of Proposition 4 is given in Section 7.4 below. Unlike the MLE, the SMP is always well-defined and outputs predictions in $(0, 1)$. Indeed, the denominator in (39) belongs to $[1, 2]$, and whenever the points $Y_1X_1, \dots, Y_nX_n, yx$ belong to a half-space (so that the MLE does not exist in $\mathbb{R}^d$), we have $f_{\widehat{\beta}_n^{(x,y)}}(y|x) = 1$, so that the prediction of the SMP is well-defined and does not depend on the choice of the MLE in (38), see the proof of Proposition 4 for details.

Improper estimator. Once again, the SMP $\widetilde{f}_n$ is in this case an *improper* predictor, since the log odd-ratio $\log(\widetilde{f}_n(1|x)/\widetilde{f}_n(-1|x))$ is not linear in x. A strength of this improper estimator lies in its simplicity: the computation of $\widetilde{f}_n(\cdot|x)$ only requires training two logistic regressions, on a first “virtual” dataset $(X_1, Y_1), \dots, (X_n, Y_n), (x, -1)$ and on a second virtual dataset $(X_1, Y_1), \dots, (X_n, Y_n), (x, 1)$, which can be done by warm-starting an optimization algorithm at $\widehat{\beta}_n$. In particular, this algorithm is computationally much simpler than the improper estimator introduced in [32] for logistic regression, which involves a continuous mixture of logistic predictors under the Bayesian posterior, approximated by MCMC methods.

A comparison with stability approaches. Approaches based on the stability of the loss [18, 72, 74, 50] would lead to a control of the excess risk involving $\ell(\langle \widehat{\beta}_n^{-Z}, Z\rangle) - \ell(\langle \widehat{\beta}_n^{Z}, Z\rangle)$, while Proposition 4 involves $\sigma(\langle \widehat{\beta}_n^{-Z}, Z\rangle) - \sigma(\langle \widehat{\beta}_n^{Z}, Z\rangle)$, where we recall that $\ell(u) = \log(1 + e^u)$ and $\sigma(u) = 1/(1 + e^{-u})$. Whenever $u' \approx u \gg 1$, we have $\ell(u') - \ell(u) \approx \sigma(u) \cdot (u' - u) \approx u' - u$, while $\sigma(u') - \sigma(u) \approx \sigma'(u) \cdot (u' - u) \approx e^{-u} \cdot (u' - u)$. In this case, the SMP bound is exponentially smaller than the loss stability bound. This is a rough explanation of the reason why we are able to remove terms of order $e^{\|X\|}$ from our upper bound on the excess risk of SMP, provided in the next section.

5.3 Excess risk bounds for the Ridge-regularized SMP

In order to obtain explicit and precise non-asymptotic guarantees, we consider a Ridge-regularized variant of SMP for logistic regression. Specifically, for $\lambda > 0$ we consider the penalty $\phi(\beta) = \lambda\|\beta\|^2/2$. The corresponding penalized SMP can be computed as follows: for every $z \in \mathbb{R}^d$, let

$$\widehat{\beta}_{\lambda,n}^{z} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n+1}\Big(\sum_{i=1}^n \ell(\langle \beta, Z_i\rangle) + \ell(\langle \beta, z\rangle) \Big) + \frac{\lambda}{2}\|\beta\|^2 \right\}. \tag{41}$$

Note that $\widehat{\beta}_{\lambda,n}^{z} \in \mathbb{R}^d$ exists and is unique, since the regularized objective in (41) is strongly convex, hence strictly convex and diverging as $\|\beta\| \to +\infty$. As before, we let $\widehat{\beta}_{\lambda,n}^{(x,y)} = \widehat{\beta}_{\lambda,n}^{-yx}$ for $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$.

Now, following Theorem 2, the regularized SMP writes in this case

$$\widetilde{f}_{\lambda,n}(y|x) = \frac{\sigma(y\langle \widehat{\beta}^{(x,y)}_{\lambda,n}, x\rangle)\, e^{-\lambda\|\widehat{\beta}^{(x,y)}_{\lambda,n}\|^2/2}}{\sigma(\langle \widehat{\beta}^{(x,1)}_{\lambda,n}, x\rangle)\, e^{-\lambda\|\widehat{\beta}^{(x,1)}_{\lambda,n}\|^2/2} + \sigma(-\langle \widehat{\beta}^{(x,-1)}_{\lambda,n}, x\rangle)\, e^{-\lambda\|\widehat{\beta}^{(x,-1)}_{\lambda,n}\|^2/2}} \tag{42}$$

for any $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$, and comes as before at the cost of two ridge-regularized logistic regressions.

We will work under Assumption 4, namely $\|X\| \leqslant R$ almost surely, as we did in Section 4.3 for the Gaussian linear family. Our main result for the regularized SMP will be stated in a nonparametric setting, where the dependence on the dimension d is kept implicit through the degrees of freedom (25). Since $\ell''(u) = \sigma'(u) = \sigma(u)(1-\sigma(u))$ for $u \in \mathbb{R}$, the Hessian of $\beta \mapsto \ell(\langle\beta, z\rangle)$ is $\sigma(\langle\beta, z\rangle)(1-\sigma(\langle\beta, z\rangle))zz^\top$ for any $z \in \mathbb{R}^d$. Hence, the Hessian of $\beta \mapsto \ell(\langle\beta, Z\rangle)$ has Frobenius norm bounded by $\|Z\|^2 \leqslant R^2/4$, and is therefore integrable. It follows that the population risk R is twice differentiable, with Hessian

$$H(\beta) := \nabla^2 R(\beta) = \mathbb{E}\big[\sigma(\langle\beta, Z\rangle)(1 - \sigma(\langle\beta, Z\rangle))ZZ^\top\big] .$$

We now turn to our main guarantee for the Ridge-regularized SMP in the logistic setting.

Theorem 6. *Grant Assumption 4, and assume that $\lambda \geqslant 2R^2/(n + 1)$. Then, the Ridge-regularized logistic SMP given by (42) satisfies*

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] \leqslant R(\beta) + e \cdot \frac{\mathsf{df}_{4\lambda}(\Sigma)}{n} + \frac{\lambda}{2}\|\beta\|^2 \tag{43}$$

for every $\beta \in \mathbb{R}^d$, where we recall that $\mathsf{df}_\lambda(\Sigma) = \mathrm{tr}[(\Sigma + \lambda I)^{-1}\Sigma]$.

The upper bound (43) is a *fast rate* excess risk guarantee; it is worth noting that it only requires bounded covariates (Assumption 4). In particular, it requires no assumption on the conditional distribution of Y given X. Furthermore, when the feature X comes from a bounded kernel (see the discussion in Section 4.3 above), the bound (43) is valid *under no assumption* on the distribution of (X, Y).

The degrees of freedom $\mathsf{df}_{4\lambda}(\Sigma)$ in the upper bound (43) can in fact be upper bounded by $\mathsf{df}_\lambda(\widetilde{H}_{\lambda,n})$, where $\widetilde{H}_{\lambda,n} = \mathbb{E}[\widehat{H}_{n+1}(\widehat{\beta}_{\lambda,n+1})]$ with

$$\widehat{H}_{n+1}(\beta) = \nabla^2 \widehat{R}_{n+1}(\beta) = \frac{1}{n+1}\sum_{i=1}^{n+1} \sigma(\langle\beta, Z_i\rangle)(1 - \sigma(\langle\beta, Z_i\rangle))Z_i Z_i^\top$$

for $\beta \in \mathbb{R}^d$, where

$$\widehat{\beta}_{\lambda,n+1} = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^d} \left\{\frac{1}{n+1}\sum_{i=1}^{n+1} \ell(\langle\beta, Z_i\rangle) + \frac{\lambda}{2}\|\beta\|^2\right\}.$$

While (43) is obtained by using the upper bound $\widetilde{H}_{\lambda,n+1} \preccurlyeq \Sigma/4$ (valid under no assumption), for fixed λ and $n \to \infty$, using the convergence $\widehat{\beta}_{\lambda,n+1} \to \beta^*_\lambda = \arg\min_{\beta\in\mathbb{R}^d}\{R(\beta) + \lambda\|\beta\|^2/2\}$ (a consequence of general results on M-estimators) and a dominated convergence argument, one may show that $\widetilde{H}_{\lambda,n+1} \to H(\beta^*_\lambda)$. For instance, [57] obtained upper bounds in terms of $\mathsf{df}_\lambda(H(\beta^*_\lambda))$ in the *well-specified* case. It is unclear, however, if one can obtain a non-asymptotic bound $\mathsf{df}_\lambda(\widetilde{H}_{\lambda,n}) \lesssim \mathsf{df}_\lambda(H(\beta^*_\lambda))$ for $n \gtrsim \mathsf{df}_\lambda(H(\beta^*_\lambda))$ under only Assumption 4, and in particular under no assumption on the conditional distribution of Y given X.

Since $\mathsf{df}_{4\lambda}(\Sigma) \leqslant d$ for every λ, we deduce the following result in the finite-dimensional setting.

Corollary 2. *Under Assumption 4, the Ridge-regularized logistic SMP $\widetilde{f}_{\lambda,n}$ (42) with $\lambda = 2R^2/(n+1)$ satisfies, for every $\beta \in \mathbb{R}^d$,*

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] \leqslant R(\beta) + e \cdot \frac{d}{n} + \frac{R^2\|\beta\|^2}{n} \tag{44}$$

Note that under the "well-conditioned" scaling of dimension d, namely $R = O(\sqrt{d})$ and $\|\beta^*\| = O(1)$ (constant signal strength), Corollary 2 yields an excess risk of $O(d/n)$.

Bypassing a lower bound. Under Assumption 4, Corollary 2 leads to an upper bound for Ridge SMP of $O((d + R^2\|\beta\|^2)/n)$ whenever $\lambda \asymp R^2/n$. By contrast, [43] showed a lower bound for any *proper* estimator (including the norm-constrained or Ridge-penalized MLE, or any stochastic optimization procedure) of order $\min(BR/\sqrt{n}, de^{BR}/n)$ (where $B = \|\beta\|$) in the worst case. This means that SMP, which is an *improper* predictor, bypasses the lower bound of proper estimators.

Computationally simpler than previous estimators. [32] shows an upper bound of order $d(\log n + \log(BR))/n$ by applying online-to-offline conversion (iterate averaging) to a Bayes mixture sequential procedure, with uniform prior over the ball of radius B; [46] obtains similar bounds using Gaussian priors. This bound has an even better dependence in B than the penalized SMP (logarithmic instead of quadratic), although it also has a slightly worse dependence in n (additional $\log n$ factor). The main advantage of the SMP over Bayes is that it is computationally less demanding: it replaces a problem of posterior sampling by one of optimization, since it requires training two updated logistic regressions, starting for instance at the Ridge-penalized MLE. Therefore, we partly answer an open problem from [32], about finding an efficient alternative with fast rate, at least in the batch statistical learning case. Note however that the SMP is still more computationally demanding at prediction time than the MLE, because of the required updates of the logistic risk minimization problem.

6 Conclusion

In this paper, we derive excess risk bounds for predictive density estimation under the logarithmic loss, which hold under misspecification. Minimizing these excess risk bounds naturally leads to a new improper (out-of-model) procedure, which we call *Sample Minmax Predictor* (SMP). On several problems, we show that the resulting bound, which is based on a refinement of the stability argument tailored for the logarithmic loss, scales as d/n, irrespective of the true distribution. This contrasts to approaches based on estimating the best parameter in the considered class of distributions, whose performance typically degrade under misspecification, where it exhibits unbounded constants. This estimator provides an alternative to approaches based on online-to-offline conversion [11, 26, 27, 4] of sequential procedures, whose rates feature an additional logarithmic dependence on sample size, and may be infinite for unbounded classes.

We apply the SMP to the class given by the Gaussian linear conditional model. In this case, the SMP can be described explicitly, and achieves in the general misspecified case at most twice the minimax risk in the well-specified case, for every distribution of covariates. We then consider a Ridge-regularized variant, which achieves nonparametric fast rates, as well as a bound with a logarithmic dependence on the diameter of the comparison class in the finite-dimensional regime with degenerate design.

We then consider the problem of logistic regression. Then, (penalized) SMP is a simple explicit procedure, whose predictions can be computed at the cost of two logistic regressions. From a statistical perspective, it achieves fast excess risk rates even for worst-case distributions; such guarantees are known to be out of reach for any *proper* procedure [43]. In the batch i.i.d. case, this provides a more efficient alternative to the improper estimator from [32], which relies on exponential weights, thereby partly addressing an open question from this article. This work leaves a number of open problems and future directions:

- First, the excess risk bounds in this paper only hold in expectation, and not with exponential probability. This limitation is shared by procedures relying on online-to-batch conversion [26, 3, 4, 32]. In particular, the high-probability bound stated by [32] for a procedure based on a "confidence boosting" technique [61] appears to be incorrect: specifically, Equation (17) herein is obtained by applying Markov's inequality to the excess risk; however, this quantity can take negative values since the predictor is outside the class. A direction for future work is to design procedures that achieve high (exponential) probability excess risk bounds that do not degrade under model misspecification.

- Second, it would be interesting to see if the proposed method could be adapted to the problem of online logistic regression with individual sequences, with valid regret bound. We believe this to be feasible by adapting the algorithm and the proof technique, and leave this task to future work.

- Another possibility is to apply SMP to other large non-parametric classes with polynomial entropy numbers, beyond the balls of RKHSs considered here. One interest of doing that is that exponential weights typically yield suboptimal rates in this case [4], owing to their reliance on volumetric arguments and the fact that they do not combine well with the chaining technique. In [28, 34, 32], improved rates are obtained through more involved techniques. On the other hand, the excess risk bound for the SMP depends on some supremum over the class, which may lead to the correct rate of convergence for large classes.

- Finally, Theorem 3 shows that in the Gaussian model, the Bayes predictive posterior (under uniform prior) *equalizes* the excess risk over all distributions (even in the misspecified case). This reveals the critical role of averaging in the misspecified case (which compensates the increased variance, namely the slower posterior concentration rate). It would be interesting to extend this finding to other models. We conjecture that, by local asymptotic normality, this behavior should extend asymptotically (at the first order) to regular models with smooth priors. Next, an interesting direction would be to study under which conditions on the model and the prior (starting with exponential families) a uniform and non-asymptotic bound (such as Theorem 3 or our guarantees for SMP) can be obtained for Bayesian or pseudo-Bayesian methods.

On a more general note, the problem of statistical learning with logarithmic loss (that is, predictive density estimation under model misspecification) possesses some specific properties, which can be exploited to obtain more precise results than generic approaches applicable to general loss functions (which often suffer from the unboundedness of the logarithmic loss). This has been exploited successfully in the sequential case where cumulative criteria are considered [63, 29]; while the present work may be understood as an attempt to obtain similar guarantees for the statistical learning setting, we expect that further advances are possible on this subject.

7 Proofs

7.1 Proofs of general excess risk bounds (Section 2)

Proof of Theorem 1. Let Z_1^n, Z denote $n+1$ i.i.d. variables distributed as P. We have

$$\begin{aligned}\mathbb{E}\big[\mathcal{E}_\phi(\widehat{g}_n)\big] &= \mathbb{E}_{Z_1^n,Z}[\ell(\widehat{g}_n, Z)] - \inf_{f\in\mathcal{F}} \mathbb{E}_{Z_1^n,Z}\bigg[\frac{1}{n+1}\bigg\{\sum_{i=1}^n \ell_\phi(f,Z_i) + \ell_\phi(f,Z)\bigg\}\bigg] \\ &= \mathbb{E}_{Z_1^n,Z}[\ell(\widehat{g}_n, Z)] - \mathbb{E}_{Z_1^n,Z}\bigg[\inf_{f\in\mathcal{F}} \frac{1}{n+1}\bigg\{\sum_{i=1}^n \ell_\phi(f,Z_i) + \ell_\phi(f,Z)\bigg\}\bigg] - \Delta_n\end{aligned}$$

where we denoted

$$\Delta_n = \inf_{f\in\mathcal{F}} \mathbb{E}\bigg[\frac{1}{n+1}\bigg\{\sum_{i=1}^n \ell_\phi(f,Z_i) + \ell_\phi(f,Z)\bigg\}\bigg] - \mathbb{E}\bigg[\inf_{f\in\mathcal{F}} \frac{1}{n+1}\bigg\{\sum_{i=1}^n \ell_\phi(f,Z_i) + \ell_\phi(f,Z)\bigg\}\bigg] \geqslant 0. \tag{45}$$

In particular, by definition of $\widehat{f}^Z_{\phi,n}$,

$$\mathbb{E}\big[\mathcal{E}_\phi(\widehat{g}_n)\big] + \Delta_n = \mathbb{E}_{Z_1^n,Z}[\ell(\widehat{g}_n, Z)] - \frac{1}{n+1}\mathbb{E}\bigg[\sum_{i=1}^n \ell_\phi(\widehat{f}^Z_{\phi,n}, Z_i) + \ell_\phi(\widehat{f}^Z_{\phi,n}, Z)\bigg] . \tag{46}$$

Since the distribution of the i.i.d. sample $(Z_1, \dots, Z_n, Z)$ is preserved by exchanging Z and Z_i, we have $\mathbb{E}[\ell_\phi(\widehat{f}^Z_{\phi,n}, Z_i)] = \mathbb{E}[\ell_\phi(\widehat{f}^Z_{\phi,n}, Z)]$ for $i = 1, \dots, n$ (recall that $\widehat{f}^Z_{\phi,n}$ is chosen symmetrically in $Z_1, \dots, Z_n, Z$). Hence, (46) becomes

$$\begin{aligned}\mathbb{E}\big[\mathcal{E}_\phi(\widehat{g}_n)\big] + \Delta_n &= \mathbb{E}_{Z_1^n,Z}\big[\ell(\widehat{g}_n, Z) - \ell_\phi(\widehat{f}^Z_{\phi,n}, Z)\big] \\ &= \mathbb{E}_{Z_1^n,X}\mathbb{E}_{Y|X}\big[\ell(\widehat{g}_n(X), Y) - \ell_\phi(\widehat{f}^{(X,Y)}_{\phi,n}(X), Y)\big] \\ &\leqslant \mathbb{E}_{Z_1^n,X}\Big[\sup_{y\in\mathcal{Y}}\big\{\ell(\widehat{g}_n(X), y) - \ell_\phi(\widehat{f}^{(X,y)}_{\phi,n}(X), y)\big\}\Big] ,\end{aligned} \tag{47}$$

which implies the bound (5) since $\Delta_n \geqslant 0$. The remaining claims follow directly. □

Proof of Theorem 2. In the case of the logarithmic loss $\ell(p, (x, y)) = -\log p(y|x)$, we have for every density p on $\mathcal{Y}$ and $x \in \mathcal{X}$:

$$\sup_{y\in\mathcal{Y}}\big\{\ell(p, y) - \ell_\phi(\widehat{f}^{(x,y)}_{\phi,n}(x), y)\big\} = \sup_{y\in\mathcal{Y}} \log \frac{\widehat{f}^{(x,y)}_{\phi,n}(y|x)e^{-\phi(\widehat{f}^{(x,y)}_{\phi,n})}}{p(y)} . \tag{48}$$

Now, Theorem 2 follows from Theorem 1 together with Lemma 1 below, where we consider $g(y) = \widehat{f}^{(x,y)}_{\phi,n}(y|x)e^{-\phi(\widehat{f}^{(x,y)}_{\phi,n})}$. □

Lemma 1. *Let $g : \mathcal{Y} \to [0, +\infty]$ be a measurable function such that $\int_\mathcal{Y} g\mathrm{d}\mu \in \mathbb{R}^*_+$. Then,*

$$\inf_p \sup_{y\in\mathcal{Y}} \log \frac{g(y)}{p(y)} = \log\bigg(\int_\mathcal{Y} g(y)\mu(\mathrm{d}y)\bigg) , \tag{49}$$

where the infimum in (49) spans over all probability densities $p : \mathcal{Y} \to \mathbb{R}^+$ with respect to μ, and the infimum is reached at

$$p^* = \frac{g}{\int_\mathcal{Y} g\mathrm{d}\mu} . \tag{50}$$

Proof. For every density p, denote $C(p) = \sup_{y\in\mathcal{Y}} \log g(y)/p(y)$. By definition, $p(y) \geqslant e^{-C(p)}g(y)$, so that since p is a density

$$1 = \int_\mathcal{Y} p(y)\mu(\mathrm{d}y) \geqslant e^{-C(p)}\int_\mathcal{Y} g(y)\mu(\mathrm{d}y) ,$$

so that $C(p) \geqslant \log\big(\int_\mathcal{Y} g\mathrm{d}\mu\big)$. Since $C(p^*) = \log\big(\int_\mathcal{Y} g\mathrm{d}\mu\big)$, this concludes the proof. □

We will sometimes also use the following observation:

Lemma 2. *The expected excess risk of the SMP is equal to*:

$$\mathbb{E}\big[\mathcal{E}_\phi(\widetilde{f}_{\phi,n})\big] = \mathbb{E}_{Z_1^n,X}\bigg[\log\bigg(\int_\mathcal{Y} \widehat{f}^{(X,y)}_{\phi,n}(y|X)e^{-\phi(\widehat{f}^{(X,y)}_{\phi,n})}\mu(\mathrm{d}y)\bigg)\bigg] - \Delta_n , \tag{51}$$

where, letting $Z_1, \dots, Z_{n+1}$ be i.i.d. sample from P and f^* a risk minimizer (*when it exists*),

$$\begin{aligned}\Delta_n &= \frac{1}{n+1}\inf_{f\in\mathcal{F}}\mathbb{E}\bigg[\sum_{i=1}^{n+1}\ell_\phi(f, Z_i) - \sum_{i=1}^{n+1}\ell_\phi(\widehat{f}_{\phi,n+1}, Z_i)\bigg] \\ &= \frac{1}{n+1}\mathbb{E}\bigg[\sum_{i=1}^{n+1}\ell_\phi(f^*, Z_i) - \sum_{i=1}^{n+1}\ell_\phi(\widehat{f}_{\phi,n+1}, Z_i)\bigg].\end{aligned} \tag{52}$$

Proof. This follows from the fact that inequality (47) is an equality when $\widehat{g}_n = \widetilde{f}_{\phi,n}$ (see Lemma 1). $\square$

7.2 Proofs for density estimation (Section 3)

Proof of Proposition 1. Since the MLE $\widehat{f}_n$ writes $\widehat{f}_n(y) = N_n(y)/n$, we have for every $y \in \mathcal{Y}$:

$$\widehat{f}_n^y(y) = \frac{N_n(y)+1}{n+1} \propto N_n(y) + 1\,, \tag{53}$$

so that, since $\sum_{y\in\mathcal{Y}} N_n(y) = n$,

$$\sum_{y\in\mathcal{Y}} \widehat{f}_n^y(y) = \frac{n+d}{n+1}\,. \tag{54}$$

It proves that the SMP $\widetilde{f}_n$ (9) is the Laplace estimator (11) and that the excess risk bound (10) becomes $\mathbb{E}[\mathcal{E}(\widetilde{f}_n)] \leqslant \log\frac{n+d}{n+1} \leqslant \frac{d-1}{n}$ (since $\log(1+u) \leqslant u$ for $u \geqslant 0$). $\square$

Proof of Proposition 2. First, let us prove that a risk minimizer $f_{\theta^*,\Sigma} \in \mathcal{F}$ exists if and only if $\mathbb{E}[\|Y\|] < +\infty$ and that $\theta^* = \mathbb{E}[Y]$ in this case. Let μ be the distribution $\mathcal{N}(0,\Sigma)$, and define the log loss with respect to μ. Then, for every $\theta, y \in \mathbb{R}^d$, $\ell(f_{\theta,\Sigma}, y) = -\langle \Sigma^{-1}\theta, y\rangle + \frac{1}{2}\theta^\top\Sigma^{-1}\theta$. Assume that there exists $\theta^* \in \mathbb{R}^d$ such that $\mathbb{E}[\ell(f_{\theta^*+\theta,\Sigma}, Y) - \ell(f_{\theta^*,\Sigma}, Y)]$ is well-defined and in $[0,+\infty]$ for every $\theta \in \mathbb{R}^d$. This implies that $\mathbb{E}[(\ell(f_{\theta^*+\theta,\Sigma}, Y) - \ell(f_{\theta^*,\Sigma}, Y))_-] < +\infty$, and hence that $\mathbb{E}[(\langle\Sigma^{-1}\theta, Y\rangle)_-] < +\infty$. Taking $\theta = \pm e_j$, $1 \leqslant j \leqslant d$ (where $(e_j)_{1\leqslant j\leqslant d}$ is the canonical basis of $\mathbb{R}^d$), this implies that $\mathbb{E}[|Y_j|] < +\infty$ for $1 \leqslant j \leqslant d$, and hence that $\mathbb{E}\|Y\| \leqslant \mathbb{E}\|Y\|_1 = \sum_{j=1}^d \mathbb{E}[|Y_j|] < +\infty$. Conversely, if $\mathbb{E}[\|Y\|] < +\infty$, so that $\mathbb{E}[Y] \in \mathbb{R}^d$ exists, then for every $\theta \in \mathbb{R}^d$, $R(f_{\theta,\Sigma}) = \mathbb{E}[\ell(f_{\theta,\Sigma}, Y)] = -\langle\Sigma^{-1}\theta, \mathbb{E}[Y]\rangle + \frac{1}{2}\theta^\top\Sigma^{-1}\theta$, which is minimized by $\theta^* = \mathbb{E}[Y]$.

We now proceed to determine the SMP and establish the excess risk bound (13). The MLE is $f_{\bar{Y}_n,\Sigma} = \mathcal{N}(\bar{Y}_n, \Sigma)$, so that for $y \in \mathbb{R}^d$, $\widehat{f}_n^y = f_{\widehat{\theta}_n^y,\Sigma}$ with $\widehat{\theta}_n^y = \frac{n\bar{Y}_n + y}{n+1}$. Since $y - \widehat{\theta}_n^y = \frac{n}{n+1}(y - \bar{Y}_n)$, we have, considering densities with respect to the measure $(2\pi)^{-d/2}\mathrm{d}y$:

$$\begin{aligned}
f_{\widehat{\theta}_n^y}(y) &= (\det\Sigma)^{-1/2}\exp\Big(-\frac{1}{2}(y - \widehat{\theta}_n^y)^\top\Sigma^{-1}(y - \widehat{\theta}_n^y)\Big) \\
&= (\det\Sigma)^{-1/2}\exp\Big(-\frac{1}{2}\Big(\frac{n}{n+1}\Big)^2(y - \bar{Y}_n)^\top\Sigma^{-1}(y - \bar{Y}_n)\Big) \\
&= (\det\Sigma)^{-1/2}\det((1+1/n)^2\Sigma)^{1/2} f_{\bar{Y}_n,(1+1/n)^2\Sigma}(y) \\
&= \Big(1 + \frac{1}{n}\Big)^d f_{\bar{Y}_n,(1+1/n)^2\Sigma}(y)\,,
\end{aligned} \tag{55}$$

so that (after normalization) $\widetilde{f}_n = \mathcal{N}(\bar{Y}_n, (1+1/n)^2\Sigma)$ and

$$\int_{\mathbb{R}^d} f_{\widehat{\theta}_n^y}(y)(2\pi)^{-d/2}\mathrm{d}y = \int_{\mathbb{R}^d}\Big(1+\frac{1}{n}\Big)^d f_{\bar{Y}_n,(1+1/n)^2\Sigma}(y)(2\pi)^{-d/2}\mathrm{d}y = \Big(1+\frac{1}{n}\Big)^d\,, \tag{56}$$

which yields the excess risk bound (13) using Theorem 2.

Now, assume that the model is well-specified, namely $Y \sim \mathcal{N}(\theta^*, \Sigma)$ for some $\theta^* \in \mathbb{R}^d$. Using Lemma 2, we have

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] = \mathbb{E}\Big[\log\Big(\int_{\mathbb{R}^d} f_{\widehat{\theta}_n^y}(y)(2\pi)^{-d/2}\mathrm{d}y\Big)\Big] - \Delta_n = d\log\Big(1+\frac{1}{n}\Big) - \Delta_n\,,$$

where Δ_n is defined as in (45), *i.e.*

$$\begin{aligned}
\Delta_n &= \frac{1}{n+1}\mathbb{E}\bigg[\sum_{i=1}^{n+1}\ell(f_{\theta^*,\Sigma},Y_i) - \inf_{\theta\in\mathbb{R}^d}\sum_{i=1}^{n+1}\ell(f_{\theta,\Sigma},Y_i)\bigg] \\
&= \frac{1}{2}\mathbb{E}\bigg[\frac{1}{n+1}\sum_{i=1}^{n+1}(Y_i-\theta^*)^\top\Sigma^{-1}(Y_i-\theta^*) - \frac{1}{n+1}\sum_{i=1}^{n+1}(\bar{Y}_{n+1}-Y_i)^\top\Sigma^{-1}(\bar{Y}_{n+1}-Y_i)\bigg] \\
&= \frac{1}{2}\mathbb{E}\big[(\bar{Y}_{n+1}-\theta^*)^\top\Sigma^{-1}(\bar{Y}_{n+1}-\theta^*)\big] \\
&= \frac{1}{2}\mathrm{tr}\Big(\Sigma^{-1}\mathbb{E}\big[(\bar{Y}_{n+1}-\theta^*)(\bar{Y}_{n+1}-\theta^*)^\top\big]\Big) \\
&= \frac{1}{2}\mathrm{tr}\Big(\Sigma^{-1}\times\frac{1}{n+1}\Sigma\Big) = \frac{d}{2(n+1)}
\end{aligned}$$

where we used the fact that $\mathbb{E}[(Y-\theta^*)(Y-\theta^*)^\top] = \Sigma$. It follows that $\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] = d\log(1+1/n) - d/(2n) \leqslant d/(2n)$, which completes the proof of Proposition 2. □

Proof of Theorem 3. Define the densities and the log-loss with respect to the measure $(2\pi)^{-d/2}\mathrm{d}y$ on $\mathbb{R}^d$. For every $\sigma^2>0$, $\theta\in\mathbb{R}$ and $y\in\mathbb{R}^d$, we have

$$\ell(f_{\theta,\sigma^2\Sigma},y) = -\log f_{\theta,\sigma^2\Sigma}(y) = \frac{d}{2}\log\sigma^2 + \frac{1}{2}\log\det(\Sigma) + \frac{1}{2\sigma^2}\left(y-\theta\right)^\top\Sigma^{-1}\left(y-\theta\right)$$

so that, denoting $\theta^* = \mathbb{E}[Y]$ and $\Sigma_Y := \mathbb{E}[(Y-\theta^*)(Y-\theta^*)^\top]$, we obtain

$$\begin{aligned}
&R(f_{\theta,\sigma^2\Sigma}) - \frac{1}{2}\log\det(\Sigma) = \frac{d}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\mathbb{E}\big[(Y-\theta)^\top\Sigma^{-1}(Y-\theta)\big] \\
&= \frac{d}{2}\log\sigma^2 + \frac{1}{2\sigma^2}(\theta-\theta^*)^\top\Sigma^{-1}(\theta-\theta^*) + \frac{1}{2\sigma^2}\mathbb{E}\,\mathrm{tr}\big(\Sigma^{-1}(Y-\theta^*)(Y-\theta^*)^\top\big) \\
&= \frac{d}{2}\log\sigma^2 + \frac{1}{2\sigma^2}(\theta-\theta^*)^\top\Sigma^{-1}(\theta-\theta^*) + \frac{1}{2\sigma^2}\mathrm{tr}\big(\Sigma^{-1}\Sigma_Y\big)
\end{aligned}$$

so that

$$\begin{aligned}
\mathcal{E}(f_{\theta,\sigma^2\Sigma}) &= R(f_{\theta,\sigma^2\Sigma}) - R(f_{\theta^*,\Sigma}) \\
&= \frac{d}{2}\log\sigma^2 + \frac{1}{2\sigma^2}(\theta-\theta^*)^\top\Sigma^{-1}(\theta-\theta^*) + \frac{1}{2}\Big(\frac{1}{\sigma^2}-1\Big)\mathrm{tr}\big(\Sigma^{-1}\Sigma_Y\big)\,. \qquad (57)
\end{aligned}$$

Now, since

$$\mathbb{E}\big[(\bar{Y}_n-\theta^*)^\top\Sigma^{-1}(\bar{Y}_n-\theta^*)\big] = \mathrm{tr}\Big(\Sigma^{-1}\mathbb{E}\big[(\bar{Y}_n-\theta^*)(\bar{Y}_n-\theta^*)^\top\big]\Big) = \frac{\mathrm{tr}(\Sigma^{-1}\Sigma_Y)}{n}\,,$$

equation (57) implies that, for $\sigma^2 = 1+1/n$,

$$\mathbb{E}\big[\mathcal{E}(f_{\bar{Y}_n,\sigma^2\Sigma})\big] = \frac{d}{2}\log\sigma^2 + \frac{1}{2}\Big[\Big(1+\frac{1}{n}\Big)\frac{1}{\sigma^2}-1\Big]\mathrm{tr}(\Sigma^{-1}\Sigma_Y) = \frac{d}{2}\log\Big(1+\frac{1}{n}\Big)\,. \qquad (58)$$

In order to conclude that $\widehat{f}_n = \mathcal{N}(\bar{Y}_n,(1+1/n)\Sigma)$, which has constant risk, achieves minimax excess risk over the class of distributions of Y with finite variance, it suffices to note that $\widehat{f}_n$ achieves minimax excess risk for Y a Gaussian from $\{\mathcal{N}(\theta^*,\Sigma) : \theta^*\in\mathbb{R}^d\}$ (*i.e.*, in the well-specified case). Indeed, if $Y\sim\mathcal{N}(\theta^*,\Sigma)$, then $\mathcal{E}(f) = \mathrm{KL}(\mathcal{N}(\theta^*,\Sigma),f)$ for every density f, and $\widehat{g}_n$ achieves minimax KL-risk on the Gaussian location family [67, 65]. □

7.3 Proofs for the Gaussian linear model (Section 4)

Proof of Theorem 4. Let us first recall that $\mathcal{F} = \{f_\beta(y|x) = \mathcal{N}(\langle \beta, x\rangle, 1) : \beta \in \mathbb{R}^d\}$ and that $\widehat{\Sigma}_n = n^{-1}\sum_{i=1}^n X_i X_i^\top$ and $\widehat{S}_n = n^{-1}\sum_{i=1}^n Y_i X_i$. The MLE is given by $\widehat{\beta}_n = \widehat{\Sigma}_n^{-1}\widehat{S}_n$ and, for every $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$,

$$\widehat{\beta}_n^{(x,y)} = (n\widehat{\Sigma}_n + xx^\top)^{-1}(n\widehat{S}_n + yx).$$

Hence, we have

$$\begin{aligned} y - \langle \widehat{\beta}_n^{(x,y)}, x\rangle &= y - \big\langle (n\widehat{\Sigma}_n + xx^\top)^{-1}(n\widehat{S}_n + yx), x\big\rangle \\ &= \big(1 - \big\langle (n\widehat{\Sigma}_n + xx^\top)^{-1}x, x\big\rangle\big) y - \big\langle (n\widehat{\Sigma}_n + xx^\top)^{-1} n\widehat{S}_n, x\big\rangle \\ &= \sigma_n(x)^{-1}(y - \mu_n(x)), \end{aligned}$$

where we defined

$$\sigma_n(x) = \big(1 - \big\langle (n\widehat{\Sigma}_n + xx^\top)^{-1}x, x\big\rangle\big)^{-1} \quad \text{and} \quad \mu_n(x) = \frac{\big\langle (n\widehat{\Sigma}_n + xx^\top)^{-1} n\widehat{S}_n, x\big\rangle}{1 - \big\langle (n\widehat{\Sigma}_n + xx^\top)^{-1}x, x\big\rangle}.$$

Note that both quantities are well-defined under since $\widehat{\Sigma}_n$ is invertible almost surely by Assumption 2. Moreover, these quantities can be simplified thanks to the following lemma.

Lemma 3. *Assume that S is a symmetric positive d-dimensional matrix and that $v \in \mathbb{R}^d$. Then, one has*

$$\big(1 - \langle (S + vv^\top)^{-1}v, v\rangle\big)^{-1} = 1 + \langle S^{-1}v, v\rangle, \tag{59}$$

and, for any $u \in \mathbb{R}^d$,

$$\frac{\big\langle (S + vv^\top)^{-1}Su, v\big\rangle}{1 - \big\langle (S + vv^\top)^{-1}v, v\big\rangle} = \langle u, v\rangle\,. \tag{60}$$

The proof of Lemma 3 is given below. It also follows from the Sherman-Morrison formula. Using (59) with $S = n\widehat{\Sigma}_n$ and $v = x$ leads to

$$\sigma_n(x) = 1 + \big\langle (n\widehat{\Sigma}_n)^{-1}x, x\big\rangle$$

while the fact that $\widehat{S}_n = \widehat{\Sigma}_n\widehat{\beta}_n$ together with (60) for $S = n\widehat{\Sigma}_n$, $v = x$ and $u = \widehat{\beta}_n$ leads to

$$\mu_n(x) = \frac{\big\langle (n\widehat{\Sigma}_n + xx^\top)^{-1} n\widehat{S}_n, x\big\rangle}{1 - \big\langle (n\widehat{\Sigma}_n + xx^\top)^{-1}x, x\big\rangle} = \frac{\big\langle (n\widehat{\Sigma}_n + xx^\top)^{-1} n\widehat{\Sigma}_n\widehat{\beta}_n, x\big\rangle}{1 - \big\langle (n\widehat{\Sigma}_n + xx^\top)^{-1}x, x\big\rangle} = \langle \widehat{\beta}_n, x\rangle.$$

Consider the dominating measure $\mu(\mathrm{d}y) = (2\pi)^{-1/2}\mathrm{d}y$ on $\mathbb{R}$. The computations above entail that for every $y \in \mathbb{R}$, we have

$$f_{\widehat{\beta}_n^{(x,y)}}(y|x) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\big(y - \langle \widehat{\beta}_n^{(x,y)}, x\rangle\big)^2\Big) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2\sigma_n^2(x)}\big(y - \mu_n(x)\big)^2\Big).$$

Note that

$$\int_{\mathbb{R}} f_{\widehat{\beta}_n^{(x,y)}}(y|x)\mu(\mathrm{d}y) = \sigma_n(x),$$

which shows after normalization (9) that the SMP is given by

$$\widetilde{f}_n(y|x) = \mathcal{N}(\mu_n(x), \sigma_n^2(x)) \tag{61}$$

and that its excess risk writes

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \mathbb{E}\big[\log\sigma_n(X)\big] = \mathbb{E}\Big[-\log\Big(1 - \big\langle (n\widehat{\Sigma}_n + XX^\top)^{-1}X, X\big\rangle\Big)\Big]\,. \tag{62}$$

This proves the first inequality in (21). Let us prove now the second inequality in (21). Let us recall that the covariance Σ and rescaled design $\widetilde{X}$, $\widetilde{X}_i$ and rescaled covariance $\widetilde{\Sigma}_n$ are given by (17) and (19). We have

$$\begin{aligned}\big\langle (n\widehat{\Sigma}_n + XX^\top)^{-1}X, X\big\rangle &= \big\langle \Sigma^{1/2}(n\widehat{\Sigma}_n + XX^\top)^{-1}\Sigma^{1/2}\Sigma^{-1/2}X, \Sigma^{-1/2}X\big\rangle \\ &= \big\langle (n\widetilde{\Sigma}_n + \widetilde{X}\widetilde{X}^\top)^{-1}\widetilde{X}, \widetilde{X}\big\rangle,\end{aligned} \tag{63}$$

hence, combining (62), (63) and (59), we have

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \mathbb{E}\Big[-\log\Big(1 - \big\langle (n\widetilde{\Sigma}_n + \widetilde{X}\widetilde{X}^\top)^{-1}\widetilde{X}, \widetilde{X}\big\rangle\Big)\Big] = \mathbb{E}\Big[\log\Big(1 + \big\langle (n\widetilde{\Sigma}_n)^{-1}\widetilde{X}, \widetilde{X}\big\rangle\Big)\Big],$$

which leads, using Jensen's inequality, together with $\mathbb{E}[\widetilde{X}\widetilde{X}^\top] = I_d$ and the fact that $\widetilde{\Sigma}_n$ and $\widetilde{X}$ are independent, to

$$\begin{aligned}\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] &\leqslant \log\Big(1 + \frac{1}{n}\mathbb{E}\big[\mathrm{tr}\big(\widetilde{\Sigma}_n^{-1}\widetilde{X}\widetilde{X}^\top\big)\big]\Big) = \log\Big(1 + \frac{1}{n}\mathrm{tr}\big\{\mathbb{E}[\widetilde{\Sigma}_n^{-1}]\mathbb{E}[\widetilde{X}\widetilde{X}^\top]\big\}\Big) \\ &= \log\Big(1 + \frac{1}{n}\mathbb{E}[\mathrm{tr}(\widetilde{\Sigma}_n^{-1})]\Big).\end{aligned}$$

This concludes the proof of Theorem 4. □

Proof of Lemma 3. First, (60) clearly holds if $v = 0$. Now, for $u, v \in \mathbb{R}^d$, $v \neq 0$:

$$\begin{aligned}\big\langle (S + vv^\top)^{-1}Su, v\big\rangle &= \big\langle (S + vv^\top)^{-1}(S + vv^\top - vv^\top)u, v\big\rangle \\ &= \big\langle (I_d - (S + vv^\top)^{-1}vv^\top)u, v\big\rangle \\ &= \langle u, v\rangle\big(1 - \langle (S + vv^\top)^{-1}v, v\rangle\big)\,.\end{aligned} \tag{64}$$

Letting $u = S^{-1}v$ in (64), the left-hand side is $\big\langle (S + vv^\top)^{-1}v, v\big\rangle > 0$ (since $S + vv^\top \succcurlyeq S$ which is positive, and $v \neq 0$) so that the right-hand side is non-zero and therefore $1 - \langle (S + vv^\top)^{-1}v, v\rangle \neq 0$. Dividing both sides of (64) by this quantity establishes (60), which implies (59) by taking $u = S^{-1}v$. □

Proof of Theorem 5 and Proposition 3. Let us recall that we consider the family $\mathcal{F} = \{f_\beta(\cdot|x) = \mathcal{N}(\langle \beta, x\rangle, \sigma^2) : \beta \in \mathbb{R}^d\}$, together with the Ridge penalization $\phi(\beta) = \lambda\|\beta\|^2/2$ for some $\lambda > 0$. Let

$$\widehat{\beta}_{\lambda,n} := \operatorname*{arg\,min}_{\beta\in\mathbb{R}^d}\Big\{\frac{1}{n}\sum_{i=1}^n \ell(f_\beta, (X_i, Y_i)) + \frac{\lambda}{2}\|\beta\|^2\Big\} = (\widehat{\Sigma}_n + \lambda I_d)^{-1}\widehat{S}_n,$$

denote the Ridge estimator, where $\widehat{\Sigma}_n$ and $\widehat{S}_n$ are the same as in the proof of Theorem 4. Defining

$$\widehat{\Sigma}_\lambda^x = n\widehat{\Sigma}_n + xx^\top + \lambda(n+1)I_d \quad \text{and} \quad \widehat{K}_\lambda^x = (\widehat{\Sigma}_\lambda^x)^{-1},$$

we have

$$\widehat{\beta}_{\lambda,n}^{(x,y)} = \big(n\widehat{\Sigma}_n + xx^\top + \lambda(n+1)I_d\big)^{-1}(n\widehat{S}_n + yx) = \widehat{K}_\lambda^x(n\widehat{S}_n + yx)$$

for any $y \in \mathbb{R}$ and $x \in \mathbb{R}^d$. Note that we have

$$y - \langle \widehat{\beta}_{\lambda,n}^{(x,y)}, x\rangle = y - \big\langle \widehat{K}_\lambda^x(n\widehat{S}_n + yx), x\big\rangle = \big(1 - \|x\|^2_{\widehat{K}_\lambda^x}\big)y - \langle n\widehat{S}_n, x\rangle_{\widehat{K}_\lambda^x} \tag{65}$$

and that

$$\begin{aligned}\lambda\|\widehat{\beta}_{\lambda,n}^{(x,y)}\|^2 &= \lambda\|\widehat{K}_\lambda^x(n\widehat{S}_n + yx)\|^2 = \lambda\|n\widehat{S}_n + yx\|^2_{(\widehat{K}_\lambda^x)^2} \\ &= y^2\lambda\|x\|^2_{(\widehat{K}_\lambda^x)^2} + 2y\lambda\langle n\widehat{S}_n, x\rangle_{(\widehat{K}_\lambda^x)^2} + \lambda\|n\widehat{S}_n\|^2_{(\widehat{K}_\lambda^x)^2}\,.\end{aligned}$$

The SMP is given in this setting by

$$\widetilde{f}_{\lambda,n}(y|x) = \frac{f_{\widehat{\beta}^{(x,y)}_{\lambda,n}}(y|x) e^{-\lambda \|\widehat{\beta}^{(x,y)}_{\lambda,n}\|^2/2}}{\int_{\mathbb{R}} f_{\widehat{\beta}^{(x,y')}_{\lambda,n}}(y'|x) e^{-\lambda \|\widehat{\beta}^{(x,y')}_{\lambda,n}\|^2/2} \mu(\mathrm{d}y')},$$

where $\mu(\mathrm{d}y) = (2\pi)^{-1/2}\mathrm{d}y$, see (9), and where

$$f_{\widehat{\beta}^{(x,y)}_{\lambda,n}}(y|x) e^{-\lambda \|\widehat{\beta}^{(x,y)}_{\lambda,n}\|^2/2} = \exp\Big(-\frac{1}{2}\Big\{ \big(y - \langle \widehat{\beta}^{(x,y)}_{\lambda,n}, x\rangle\big)^2 + \lambda \|\widehat{\beta}^{(x,y)}_{\lambda,n}\|^2 \Big\}\Big).$$

Now, the equality (65) gives, after a straightforward computation,

$$\big(y - \langle \widehat{\beta}^{(x,y)}_{\lambda,n}, x\rangle\big)^2 + \lambda \|\widehat{\beta}^{(x,y)}_{\lambda,n}\|^2 = \frac{1}{\sigma_\lambda(x)^2}\big(y - \mu_\lambda(x)\big)^2 + C,$$

where C is a quantity that does not depend on y and where we introduced, respectively,

$$\begin{aligned} \sigma_\lambda(x)^2 &= \Big((1 - \|x\|^2_{\widehat{K}^x_\lambda})^2 + \lambda \|x\|^2_{(\widehat{K}^x_\lambda)^2} \Big)^{-1} \\ \mu_\lambda(x) &= \frac{(1 - \|x\|^2_{\widehat{K}^x_\lambda}) \langle n\widehat{S}_n, x\rangle_{\widehat{K}^x_\lambda} - \lambda \langle n\widehat{S}_n, x\rangle_{(\widehat{K}^x_\lambda)^2}}{(1 - \|x\|^2_{\widehat{K}^x_\lambda})^2 + \lambda \|x\|^2_{(\widehat{K}^x_\lambda)^2}}. \end{aligned}$$

This entails that the SMP is given by

$$\widetilde{f}_{\lambda,n}(\cdot|x) = \mathcal{N}\big(\mu_\lambda(x), \sigma_\lambda(x)^2\big). \tag{66}$$

By definition of $\widehat{\beta}_{\lambda,n}$ we have

$$n\widehat{S}_n = \big(n\widehat{\Sigma}_n + \lambda(n+1)I_d\big)\widehat{\beta}_{\lambda',n}$$

where $\lambda' = (n+1)\lambda/n$, so that for $\alpha \in \{1,2\}$ we have

$$\begin{aligned} \langle n\widehat{S}_n, x\rangle_{(\widehat{K}^x_\lambda)^\alpha} &= \big\langle \big(n\widehat{\Sigma}_n + xx^\top + \lambda(n+1)I_d\big)^\alpha n\widehat{S}_n, x\big\rangle \\ &= \big\langle \big(n\widehat{\Sigma}_n + xx^\top + \lambda(n+1)I_d\big)^\alpha (n\widehat{\Sigma}_n + \lambda(n+1)I_d + xx^\top - xx^\top)\widehat{\beta}_{\lambda',n}, x\big\rangle \\ &= \langle \widehat{\beta}_{\lambda',n}, x\rangle_{(\widehat{K}^x_\lambda)^{\alpha-1}} - \langle \widehat{\beta}_{\lambda',n}, x\rangle\, \|x\|^2_{(\widehat{K}^x_\lambda)^\alpha}, \end{aligned}$$

namely

$$\langle n\widehat{S}_n, x\rangle_{\widehat{K}^x_\lambda} = \big(1 - \|x\|^2_{\widehat{K}^x_\lambda}\big)\langle \widehat{\beta}_{\lambda',n}, x\rangle \quad \text{and} \quad \langle n\widehat{S}_n, x\rangle_{(\widehat{K}^x_\lambda)^2} = \langle \widehat{\beta}_{\lambda',n}, x\rangle_{\widehat{K}^x_\lambda} - \langle \widehat{\beta}_{\lambda',n}, x\rangle\, \|x\|^2_{(\widehat{K}^x_\lambda)^2}.$$

This allows, after straightforward computations, to express $\mu_\lambda(x)$ as a function of $\widehat{\beta}_{\lambda',n}$ as follows:

$$\mu_\lambda(x) = \langle \widehat{\beta}_{\lambda',n}, x\rangle - \lambda\sigma_\lambda(x)^2 \langle \widehat{\beta}_{\lambda',n}, x\rangle_{\widehat{K}^x_\lambda}.$$

We know from Theorem 2 that the penalized excess risk of SMP satisfies

$$\begin{aligned} \mathbb{E}\big[\mathcal{E}_\lambda(\widetilde{f}_{\lambda,n})\big] &\leqslant \mathbb{E}_{Z_1^n, X}\Big[\log\Big(\int_{\mathbb{R}} f_{\widehat{\beta}^{(X,y)}_{\lambda,n}}(y|X) e^{-\lambda\|\widehat{\beta}^{(X,y)}_{\lambda,n}\|^2/2}\mu(\mathrm{d}y)\Big)\Big] \\ &\leqslant \mathbb{E}_{Z_1^n, X}\Big[\log\Big(\int_{\mathbb{R}} f_{\widehat{\beta}^{(X,y)}_{\lambda,n}}(y|X)\mu(\mathrm{d}y)\Big)\Big]. \end{aligned}$$

We know from the computations above that

$$\big(y - \langle \widehat{\beta}_{\lambda,n}^{(x,y)}, x\rangle\big)^2 = \big(1 - \|x\|^2_{\widehat{K}_\lambda^x}\big)^2 \big(y - \langle \widehat{\beta}_{\lambda',n}, x\rangle\big)^2,$$

so that, after integrating with respect to y,

$$\mathbb{E}\big[\mathcal{E}_\lambda(\widetilde{f}_{\lambda,n})\big] \leqslant \mathbb{E}_{X_1^n,X}\left[\log\left(\frac{1}{1-\|X\|^2_{\widehat{K}_\lambda^X}}\right)\right] = \mathbb{E}_{X_1^n,X}\Big[-\log\big(1 - \langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle\big)\Big]. \tag{67}$$

Note that, by the identity (59) from Lemma 3, and since $\|X\| \leqslant R$ almost surely (Assumption 4) we have

$$\langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle = \frac{\langle(n\widehat{\Sigma}_n + \lambda(n+1)I_d)^{-1}X, X\rangle}{1 + \langle(n\widehat{\Sigma}_n + \lambda(n+1)I_d)^{-1}X, X\rangle} \leqslant \frac{R^2/\big(\lambda(n+1)\big)}{1 + R^2/\big(\lambda(n+1)\big)}\,. \tag{68}$$

In addition, the function $g(u) = -\log(1-u)/u$ defined on $(0,1)$ is nondecreasing, since its derivative writes:

$$g'(u) = \frac{1}{u^2}\left[\frac{u}{1-u} - \log\left(1 + \frac{u}{1-u}\right)\right] \geqslant 0\,,$$

where we used the inequality $\log(1+v) \leqslant v$ for $v \geqslant 0$. Combining this fact with (68) shows that

$$-\log\big(1 - \langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle\big) \leqslant g\left(\frac{R^2/\big(\lambda(n+1)\big)}{1 + R^2/\big(\lambda(n+1)\big)}\right) \cdot \langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle\,. \tag{69}$$

Next, by exchangeability of $(X_1, \dots, X_n, X)$, we have

$$\begin{aligned}
\mathbb{E}\big[\langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle\big] &= \frac{1}{n+1}\,\mathbb{E}\left[\sum_{i=1}^n \langle(\widehat{\Sigma}_\lambda^X)^{-1}X_i, X_i\rangle + \langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle\right] \\
&= \frac{1}{n+1}\,\mathbb{E}\left[\mathrm{tr}\left\{\left(\sum_{i=1}^n X_iX_i^\top + XX^\top + \lambda(n+1)I_d\right)^{-1}\left(\sum_{i=1}^n X_iX_i^\top + XX^\top\right)\right\}\right].
\end{aligned} \tag{70}$$

In addition, the function $A \mapsto \mathrm{tr}((A + I_d)^{-1}A)$ is concave on positive matrices. Indeed, it writes $d - \mathrm{tr}[(A + I_d)^{-1}]$, and $A \mapsto \mathrm{tr}(A^{-1})$ is convex on positive matrices since $x \mapsto x^{-1}$ is convex on $\mathbb{R}_+^*$, by a general result on the convexity of trace functionals, see e.g. [14, 19]. Hence, applying Jensen's inequality to (70) and using the fact that

$$\mathbb{E}\left[\sum_{i=1}^n X_iX_i^\top + XX^\top\right] = (n+1)\Sigma\,,$$

we obtain:

$$\mathbb{E}\big[\langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle\big] \leqslant \frac{\mathsf{df}_\lambda(\Sigma)}{n+1}\,. \tag{71}$$

Finally, combining the bounds (67), (69) and (71) yields:

$$\mathbb{E}\big[\mathcal{E}_\lambda(\widetilde{f}_{\lambda,n})\big] \leqslant g\left(\frac{R^2/\big(\lambda(n+1)\big)}{1 + R^2/\big(\lambda(n+1)\big)}\right) \cdot \frac{\mathsf{df}_\lambda(\Sigma)}{n+1}\,. \tag{72}$$

Nonparametric rates (Theorem 5). Assume that $\lambda(n+1) \geqslant 2R^2$. The quantity inside $g(\cdot)$ in (72) is then bounded by $(1/2)/(1+1/2) = 1/3$, and since $g(1/3) = 3\log(3/2) \leqslant 1.25$, (72) becomes, by definition of $\mathcal{E}_\lambda$:

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\beta\in\mathbb{R}^d}\left\{R(\beta) + \frac{\lambda}{2}\|\beta\|^2\right\} \leqslant 1.25 \cdot \frac{\mathsf{df}_\lambda(\Sigma)}{n+1}\,. \tag{73}$$

which is precisely the announced bound (29).

Finite-dimensional case: improved dependence on the norm (Proposition 3). Now, let $\lambda = d/\big(B^2(n+1)\big)$ for some $B > 0$ (which will be a bound on the norm of the comparison parameter β). Then, $R^2/\big(\lambda(n+1)\big) = B^2R^2/d$. Now, note that for every $v > 0$

$$g\left(\frac{v}{1+v}\right) = \frac{-\log\big(1 - v/(1+v)\big)}{v/(1+v)} = \frac{(1+v)\log(1+v)}{v}\,.$$

In addition, if $v \leqslant 1$, then $(1+v)\log(1+v)/v \leqslant 1+v \leqslant 2$. On the other hand, if $v \geqslant 1$, then $(1+v)/v \leqslant 2$; it follows that for every $v > 0$:

$$g\left(\frac{v}{1+v}\right) \leqslant 2\log(e+v) \leqslant 2\log(4+4\sqrt{v}+v) = 4\log(2+\sqrt{v})\,. \tag{74}$$

Now, the excess risk bound (72) implies that, for every $\beta \in \mathbb{R}^d$ such that $\|\beta\| \leqslant B$,

$$\begin{aligned}
\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - R(\beta) &\leqslant g\left(\frac{B^2R^2/d}{1+B^2R^2/d}\right)\cdot\frac{\mathsf{df}_\lambda(\Sigma)}{n+1} + \frac{\lambda}{2}\|\beta\|^2 \\
&\leqslant 4\log\left(2+\frac{BR}{\sqrt{d}}\right)\times\frac{d}{n+1} + \frac{d}{B^2(n+1)}\times\frac{B^2}{2} && (75)\\
&= \frac{d}{n+1}\left\{4\log\left(2+\frac{BR}{\sqrt{d}}\right)+\frac{1}{2}\right\} \\
&\leqslant \frac{5d\log\big(2+BR/\sqrt{d}\big)}{n+1} && (76)
\end{aligned}$$

where inequality (75) uses the bound (74) with $v = B^2R^2/d$, the bound $\mathsf{df}_\lambda(\Sigma) \leqslant d$ (26) and the fact that $\|\beta\| \leqslant B$, while inequality (76) uses the fact that $1/2 \leqslant \log 2$. □

7.4 Proofs for logistic regression (Section 5)

Proof of Proposition 4. Let us first discuss the properties of predictions produced by the SMP, and compare it to the MLE. First, if the points $Z_1, \dots, Z_n$ do not lie within a half-space, the MLE is uniquely determined and belongs to $\mathbb{R}^d$; in addition, for any $x \in \mathbb{R}^d$ and $y \in \{-1, 1\}$, $Z_1, \dots, Z_n, -yx$ are not separated either, so $\widehat{\beta}_n^{(x,y)} \in \mathbb{R}^d$ is also well-defined and unique, and so is the prediction $\widetilde{f}_n(1|x) \in (0,1)$.

Let $\Lambda_n = \{\sum_{1\leqslant i\leqslant n}\lambda_i Z_i : \lambda_i \in \mathbb{R}^+, 1 \leqslant i \leqslant n\}$ denote the convex cone generated by $Z_1, \dots, Z_n$. Assume that $\Lambda_n \cap (-\Lambda_n) = \{0\}$ and that all Z_i are distinct from 0. Then, convex separation implies that there exists $\beta \in \mathbb{R}^d$ such that $\langle \beta, z\rangle < 0$ for all $z \in \Lambda_n \setminus \{0\}$, so that the Z_i lie within a strict half-space: $\langle \beta, Z_i\rangle < 0$ for all i. Hence, any MLE $f_{\widehat{\beta}_n}$ in $\overline{\mathcal{F}}$ belongs to $\overline{\mathcal{F}} \setminus \mathcal{F}$, and corresponds to a separating hyperplane $(+\infty, \widehat{\theta}_n)$ for some $\widehat{\theta}_n \in S^{d-1}$ (such that $\langle \widehat{\theta}_n, z\rangle < 0$ for all $z \in \Lambda_n \setminus \{0\}$). Its predictions $f_{\widehat{\beta}_n}(1|x)$ are as follows:

- If $x = 0$, then $f_{\widehat{\beta}_n}(1|x) = 1/2$.
- If $x \in \Lambda_n \setminus \{0\}$, then $\langle \widehat{\theta}_n, x\rangle < 0$ and thus $f_{\widehat{\beta}_n}(1|x) = 0$. Likewise, if $x \in (-\Lambda_n) \setminus \{0\}$, then $f_{\widehat{\beta}_n}(1|x) = 1$;
- If $x \in \mathbb{R}^d \setminus [\Lambda_n \cup (-\Lambda_n)]$, then both x and $-x$ are linearly separated from Λ_n. Hence, one can choose $\widehat{\theta}_n$ with $\langle \widehat{\theta}_n, z\rangle < 0$ for $z \in \Lambda_n \setminus \{0\}$ such that either $\langle \widehat{\theta}_n, x\rangle > 0$ or $\langle \widehat{\theta}_n, x\rangle < 0$ (or even $\langle \widehat{\theta}_n, x\rangle = 0$). In other words, one can choose an MLE $\widehat{\beta}_n$ such that $f_{\widehat{\beta}_n}(1|x)$ is either 1, 0 or $1/2$: the prediction of the MLE is ill-determined in this region, since it depends on the specific choice of the MLE.

By contrast, let us consider the prediction of the SMP $\widetilde{f}_n$. Let $z = -yx \in \mathbb{R}^d \setminus \{0\}$. As before, if $z \in \mathbb{R}^d \setminus (-\Lambda_n)$, then there exists β with $\langle \beta, z \rangle < 0$ and $\langle \beta, Z_i \rangle = -\langle \beta, -Z_i \rangle < 0$. Hence, $f_{\widehat{\beta}_n^{(x,y)}}(y|x) = 1$. On the other hand, if $z \in (-\Lambda_n) \setminus \{0\}$, then the dataset $Z_1, \dots, Z_n, z$ is not separated, so that $f_{\widehat{\beta}_n^{(x,y)}}(y|x) \in (0,1)$. Hence, for $x \in \mathbb{R}^d$:

- If $x = 0$, then $\widetilde{f}_n(1|x) = 1/2$.
- If $x \in \Lambda_n$, then $-x \in (-\Lambda_n)$ so that $f_{\widehat{\beta}_n^{(x,1)}}(1|x) \in (0,1)$, while $x \in \mathbb{R}^d \setminus (-\Lambda_n)$ so that $f_{\widehat{\beta}_n^{(x,-1)}}(-1|x) = 1$; hence, $\widetilde{f}_n(1|x) \in (0, 1/2)$. Likewise, if $x \in (-\Lambda_n)$, then $\widetilde{f}_n(1|x) \in (1/2, 1)$.
- If $x \in \mathbb{R}^d \setminus [\Lambda_n \cup (-\Lambda_n)]$, then $f_{\widehat{\beta}_n^{(x,1)}}(1|x) = f_{\widehat{\beta}_n^{(x,-1)}}(-1|x) = 1$, so that $\widetilde{f}_n(1|x) = 1/2$.

Finally, the excess risk bound (40) is established in the proof of Theorem 5 below, letting $\lambda = 0$. $\square$

Proof of Theorem 6. Let (X, Y) be a test sample, and $Z = -YX$. Since $\{Z, -Z\} = \{X, -X\}$, the excess risk bound (10) of the SMP $\widetilde{f}_{\lambda,n}$ (42) writes:

$$
\begin{aligned}
&\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\beta \in \mathbb{R}^d} \Big\{ R(\beta) + \frac{\lambda}{2}\|\beta\|^2 \Big\} \\
&\leqslant \mathbb{E}\left[\log\left(\sigma(\langle \widehat{\beta}_{\lambda,n}^{(X,1)}, X \rangle)\, e^{-\lambda\|\widehat{\beta}_{\lambda,n}^{(X,1)}\|^2/2} + \sigma(-\langle \widehat{\beta}_{\lambda,n}^{(X,-1)}, X \rangle)\, e^{-\lambda\|\widehat{\beta}_{\lambda,n}^{(X,-1)}\|^2/2}\right)\right] \\
&= \mathbb{E}\left[\log\left(\sigma(\langle \widehat{\beta}_{\lambda,n}^{-Z}, Z \rangle)\, e^{-\lambda\|\widehat{\beta}_{\lambda,n}^{-Z}\|^2/2} + \sigma(-\langle \widehat{\beta}_{\lambda,n}^{Z}, Z \rangle)\, e^{-\lambda\|\widehat{\beta}_{\lambda,n}^{Z}\|^2/2}\right)\right] \\
&\leqslant \mathbb{E}\left[\log\left(1 + \sigma(\langle \widehat{\beta}_{\lambda,n}^{-Z}, Z \rangle) - \sigma(\langle \widehat{\beta}_{\lambda,n}^{Z}, Z \rangle)\right)\right] && (77) \\
&\leqslant \mathbb{E}\left[\sigma(\langle \widehat{\beta}_{\lambda,n}^{-Z}, Z \rangle) - \sigma(\langle \widehat{\beta}_{\lambda,n}^{Z}, Z \rangle)\right] && (78)
\end{aligned}
$$

where inequality (77) is obtained by lower-bounding $e^{-\lambda\|\cdot\|^2/2} \leqslant 1$ and using the identity $\sigma(-u) = 1 - \sigma(u)$. Now, defining for $\beta \in \mathbb{R}^d$

$$
\widehat{R}_{\lambda,n}^{Z}(\beta) := \frac{1}{n+1}\left\{ \sum_{i=1}^n \ell(\langle \beta, Z_i \rangle) + \ell(\langle \beta, Z \rangle) \right\} + \frac{\lambda}{2}\|\beta\|^2 ,
$$

we have, respectively,

$$
\widehat{\beta}_{\lambda,n}^{Z} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \widehat{R}_{\lambda,n}^{Z}(\beta) \tag{79}
$$

$$
\widehat{\beta}_{\lambda,n}^{-Z} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \left\{ \widehat{R}_{\lambda,n}^{Z}(\beta) - \frac{1}{n+1}\langle \beta, Z \rangle \right\} , \tag{80}
$$

where (80) comes from the fact that $\ell(-u) = \ell(u) - u$ for $u \in \mathbb{R}$.

Now, the function $\widehat{R}_n^Z$ is λ-strongly convex, as the sum of a convex function (recall that ℓ is convex since $\ell'' = \sigma(1-\sigma) \geqslant 0$) and a $\lambda\|\beta\|^2/2$ term. It follows from Lemma 4 that

$$
R \cdot \|\widehat{\beta}_{\lambda,n}^{-Z} - \widehat{\beta}_{\lambda,n}^{Z}\| \leqslant R \cdot \frac{\|Z/(n+1)\|}{\lambda} \leqslant \frac{R^2}{\lambda(n+1)} \leqslant \frac{1}{2} , \tag{81}
$$

where we used the assumption that $\lambda \geqslant 2R^2/(n+1)$. In addition, still by Lemma 4,

$$
0 \leqslant \langle \widehat{\beta}_{\lambda,n}^{-Z} - \widehat{\beta}_{\lambda,n}^{Z}, Z \rangle \leqslant 1/2 . \tag{82}
$$

Now, since $(\log \sigma')' = \sigma''/\sigma' = 1 - 2\sigma \leqslant 1$, we have for every $u \in \mathbb{R}$ and $v \in [0, 1/2]$, $\log \sigma'(u+v) - \log \sigma'(u) \leqslant v$, namely $\sigma'(u+v) \leqslant e^v \sigma'(u) \leqslant e \cdot \sigma'(u)$. Hence, $\sigma(u+v) \leqslant e \cdot \sigma'(u) \cdot v$ for every $u \in \mathbb{R}$ and $v \in [0, 1/2]$. By (82), applying this inequality to $u = \langle \widehat{\beta}^Z_{\lambda,n}, Z \rangle$ and $v = \langle \widehat{\beta}^{-Z}_{\lambda,n} - \widehat{\beta}^Z_{\lambda,n}, Z \rangle$ yields:

$$\sigma\big(\langle \widehat{\beta}^{-Z}_{\lambda,n}, Z \rangle\big) - \sigma\big(\langle \widehat{\beta}^{Z}_{\lambda,n}, Z \rangle\big) \leqslant e^{1/2} \cdot \sigma'\big(\langle \widehat{\beta}^{Z}_{\lambda,n}, Z \rangle\big) \cdot \langle \widehat{\beta}^{-Z}_{\lambda,n} - \widehat{\beta}^Z_{\lambda,n}, Z \rangle \,. \tag{83}$$

Let us now consider the function $\widehat{R}^Z_{\lambda,n}$; its third derivative can be controlled in terms of its Hessian, as shown by [5]. Fix $\beta, \theta \in \mathbb{R}^d$, and define the function $g(t) = \widehat{R}^Z_{\lambda,n}(\beta + t\theta)$ for $t \in \mathbb{R}$. We have respectively, denoting $\beta_t = \beta + t\theta$,

$$g''(t) = \langle \nabla^2 \widehat{R}^Z_{\lambda,n}(\beta_t)\theta, \theta \rangle = \frac{1}{n+1} \left\{ \sum_{i=1}^n \sigma'(\langle \beta_t, Z_i \rangle)\langle \theta, Z_i \rangle^2 + \sigma'(\langle \beta_t, Z \rangle)\langle \theta, Z \rangle^2 \right\} + \lambda \|\theta\|^2$$

$$g'''(t) = \nabla^3 \widehat{R}^Z_{\lambda,n}(\beta_t)[\theta, \theta, \theta] = \frac{1}{n+1} \left\{ \sum_{i=1}^n \sigma''(\langle \beta_t, Z_i \rangle)\langle \theta, Z_i \rangle^3 + \sigma''(\langle \beta_t, Z \rangle)\langle \theta, Z \rangle^3 \right\}$$

Now, since $|\sigma''| = |\sigma(1-\sigma)(1-2\sigma)| \leqslant \sigma(1-\sigma) = \sigma'$ (as $0 \leqslant \sigma \leqslant 1$), and since by the Cauchy-Schwarz inequality $|\langle \theta, Z_i \rangle| \leqslant R\|\theta\|$ ($1 \leqslant i \leqslant n$) and $|\langle \theta, Z \rangle| \leqslant R\|\theta\|$, we have

$$\begin{aligned} |g'''(t)| &= \frac{1}{n+1} \left\{ \sum_{i=1}^n \left|\sigma''(\langle \beta_t, Z_i \rangle)\langle \theta, Z_i \rangle^3\right| + \left|\sigma''(\langle \beta_t, Z \rangle)\langle \theta, Z \rangle^3\right| \right\} \\ &\leqslant R\|\theta\| \cdot \frac{1}{n+1} \left\{ \sum_{i=1}^n \sigma'(\langle \beta_t, Z_i \rangle)\langle \theta, Z_i \rangle^2 + \sigma'(\langle \beta_t, Z \rangle)\langle \theta, Z \rangle^2 \right\} \leqslant R\|\theta\| \cdot g''(t) \,. \end{aligned} \tag{84}$$

The property (84) is the pseudo-self-concordance condition introduced by [5]; in particular, by Proposition 1 therein, we have for every $\beta, \theta \in \mathbb{R}^d$:

$$\nabla^2 \widehat{R}^Z_{\lambda,n}(\beta + \theta) \succcurlyeq e^{-R\|\theta\|} \cdot \nabla^2 \widehat{R}^Z_{\lambda,n}(\beta) \,. \tag{85}$$

It follows from (85) (letting $\beta = \widehat{\beta}^Z_{\lambda,n}$ and $\theta = \beta' - \widehat{\beta}^Z_{\lambda,n}$) that $\widehat{R}^Z_{\lambda,n}$ is $e^{-(1/2+\varepsilon)}\nabla^2 \widehat{R}^Z_{\lambda,n}(\widehat{\beta}^Z_{\lambda,n})$-strongly convex on the open convex ball $\Omega_\varepsilon = \{\beta' \in \mathbb{R}^d : R\|\beta' - \widehat{\beta}^Z_{\lambda,n}\| < 1/2 + \varepsilon\}$ for every $\varepsilon > 0$. In addition, the inequality (81) shows that the function $\widehat{R}^Z_{\lambda,n}(\beta) - \langle \beta, Z \rangle/(n+1)$ reaches its minimum $\widehat{\beta}^{-Z}_{\lambda,n}$ on Ω_ε, so that by Lemma 4,

$$\langle \widehat{\beta}^{-Z}_{\lambda,n} - \widehat{\beta}^Z_{\lambda,n}, Z/(n+1) \rangle \leqslant e^{1/2+\varepsilon} \left\| \frac{Z}{n+1} \right\|^2_{\nabla^2 \widehat{R}^Z_{\lambda,n}(\widehat{\beta}^Z_{\lambda,n})^{-1}} .$$

Taking $\varepsilon \to 0$ in the above bound and multiplying by $n+1$, we obtain:

$$\langle \widehat{\beta}^{-Z}_{\lambda,n} - \widehat{\beta}^Z_{\lambda,n}, Z \rangle \leqslant \frac{e^{1/2}}{n+1} \cdot \langle \nabla^2 \widehat{R}^Z_{\lambda,n}(\widehat{\beta}^Z_{\lambda,n})^{-1} Z, Z \rangle \,, \tag{86}$$

so that by combining inequalities (83) and (86),

$$\sigma\big(\langle \widehat{\beta}^{-Z}_{\lambda,n}, Z \rangle\big) - \sigma\big(\langle \widehat{\beta}^{Z}_{\lambda,n}, Z \rangle\big) \leqslant \frac{e}{n+1} \cdot \sigma'\big(\langle \widehat{\beta}^{Z}_{\lambda,n}, Z \rangle\big) \cdot \langle \nabla^2 \widehat{R}^Z_{\lambda,n}(\widehat{\beta}^Z_{\lambda,n})^{-1} Z, Z \rangle \,. \tag{87}$$

It thus remains to control the expectation of the right-hand side of (87). By exchangeability of

$(Z_1, \dots, Z_n, Z)$ (and since $\widehat{R}^Z_{\lambda,n}, \widehat{\beta}^Z_{\lambda,n}$ are unchanged after permutation of Z_i and Z), we have:

$$\begin{aligned}
&\mathbb{E}\big[\sigma'\big(\langle \widehat{\beta}^Z_{\lambda,n}, Z\rangle\big) \cdot \langle \nabla^2 \widehat{R}^Z_{\lambda,n}(\widehat{\beta}^Z_{\lambda,n})^{-1} Z, Z\rangle\big] \\
&= \frac{1}{n+1}\mathbb{E}\Bigg[\sum_{i=1}^n \sigma'\big(\langle \widehat{\beta}^Z_{\lambda,n}, Z_i\rangle\big) \cdot \langle \nabla^2 \widehat{R}^Z_{\lambda,n}(\widehat{\beta}^Z_{\lambda,n})^{-1} Z_i, Z_i\rangle + \sigma'\big(\langle \widehat{\beta}^Z_{\lambda,n}, Z\rangle\big) \cdot \langle \nabla^2 \widehat{R}^Z_{\lambda,n}(\widehat{\beta}^Z_{\lambda,n})^{-1} Z, Z\rangle\Bigg] \\
&= \mathbb{E}\Bigg[\mathrm{tr}\Bigg\{\nabla^2 \widehat{R}^Z_{\lambda,n}(\widehat{\beta}^Z_{\lambda,n})^{-1} \cdot \frac{1}{n+1}\Bigg(\sum_{i=1}^n \sigma'\big(\langle \widehat{\beta}^Z_{\lambda,n}, Z\rangle\big) Z_i Z_i^\top + \sigma'\big(\langle \widehat{\beta}^Z_{\lambda,n}, Z\rangle\big) Z Z^\top\Bigg)\Bigg\}\Bigg] \\
&= \mathbb{E}\left[\mathrm{tr}\Big\{\big[\nabla^2 \widehat{R}^Z_n(\widehat{\beta}^Z_{\lambda,n}) + \lambda I_d\big]^{-1} \nabla^2 \widehat{R}^Z_n(\widehat{\beta}^Z_{\lambda,n})\Big\}\right] ; \qquad (88)
\end{aligned}$$

in (88), we defined

$$\widehat{R}^Z_n(\beta) = \widehat{R}^Z_n(\beta) - \frac{\lambda}{2}\|\beta\|^2 = \frac{1}{n+1}\Bigg\{\sum_{i=1}^n \ell(\langle \beta, Z_i\rangle) + \ell(\langle \beta, Z\rangle)\Bigg\},$$

whose Hessian writes

$$\nabla^2 \widehat{R}^Z_n(\beta) = \frac{1}{n+1}\Bigg\{\sum_{i=1}^n \sigma'(\langle \beta, Z_i\rangle) Z_i Z_i^\top + \sigma'(\langle \beta, Z\rangle) Z Z^\top\Bigg\}.$$

Finally, by concavity of the map $A \mapsto \mathrm{tr}[(A + \lambda I_d)^{-1} A]$ on positive matrices (shown in the proof of Theorem 5), denoting $\widetilde{H}_{\lambda,n} := \mathbb{E}[\nabla^2 \widehat{R}^Z_n(\widehat{\beta}^Z_{\lambda,n})] = \mathbb{E}[\nabla^2 \widehat{R}_{n+1}(\widehat{\beta}_{\lambda,n+1})]$ we have

$$\mathbb{E}\left[\mathrm{tr}\Big\{\big[\nabla^2 \widehat{R}^Z_n(\widehat{\beta}^Z_{\lambda,n}) + \lambda I_d\big]^{-1} \nabla^2 \widehat{R}^Z_n(\widehat{\beta}^Z_{\lambda,n})\Big\}\right] \leqslant \mathrm{tr}\big\{[\widetilde{H}_{\lambda,n} + \lambda I_d]^{-1} \widetilde{H}_{\lambda,n}\big\} = \mathsf{df}_\lambda(\widetilde{H}_{\lambda,n})\,. \qquad (89)$$

Combining inequalities (78), (87), (88) and (89), we conclude that

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\beta \in \mathbb{R}^d}\Big\{R(\beta) + \frac{\lambda}{2}\|\beta\|^2\Big\} \leqslant e \cdot \frac{\mathsf{df}_\lambda(\widetilde{H}_{\lambda,n})}{n+1}\,. \qquad (90)$$

Finally, the bound (43) is obtained by noting that, by exchangeability and since $\sigma' = \sigma(1-\sigma) \leqslant 1/4$ and $Z_1 Z_1^\top = X_1 X_1^\top$,

$$\widetilde{H}_{\lambda,n+1} = \mathbb{E}\big[\sigma'(\langle \widehat{\beta}_{\lambda,n+1}, Z_1\rangle) Z_1 Z_1^\top\big] \leqslant \mathbb{E}\big[X_1 X_1^\top\big]/4 = \Sigma/4\,,$$

so that $\mathsf{df}_\lambda(\widetilde{H}_{\lambda,n}) \leqslant \mathsf{df}_\lambda(\Sigma/4) = \mathsf{df}_{4\lambda}(\Sigma)$. □

Lemma 4 (Stability). *Let Ω be a nonempty open convex subset of $\mathbb{R}^d$, and $F : \Omega \to \mathbb{R}$ a differentiable function. Assume that F is* Σ-strongly convex *on Ω (where Σ is a $d \times d$ symmetric positive matrix), in the sense that, for every $x, x' \in \Omega$,*

$$F(x') \geqslant F(x) + \langle \nabla F(x), x' - x\rangle + \frac{1}{2}\|x' - x\|^2_\Sigma\,. \qquad (91)$$

Assume that F reaches its minimum at $x^* \in \Omega$. Let $g \in \mathbb{R}^d$, and assume that the function $x \mapsto F(x) - \langle g, x\rangle$ reaches its minimum at some $\widetilde{x} \in \Omega$. Then,

$$\|\widetilde{x} - x^*\|_\Sigma \leqslant \|g\|_{\Sigma^{-1}}\,, \qquad \langle g, \widetilde{x} - x^*\rangle \leqslant \|g\|^2_{\Sigma^{-1}}\,. \qquad (92)$$

Proof. First, since $\widetilde{x} \in \Omega$ minimizes the function $x \mapsto F(x) - \langle g, x \rangle$, we have $0 = \nabla F(\widetilde{x}) - g$. This implies

$$\langle \nabla F(\widetilde{x}), \widetilde{x} - x^* \rangle = \langle g, x \rangle \,. \tag{93}$$

Now, by substituting x' and x in inequality (91) and adding the resulting inequality to (91), we obtain for every $x, x' \in \Omega$,

$$\langle \nabla F(x') - \nabla F(x), x' - x \rangle \geqslant \|x' - x\|_\Sigma^2 \,.$$

Setting $x' = \widetilde{x}$ and $x = x^*$, and using that $\nabla F(x^*) = 0$ (since $x^* \in \Omega$ minimizes F), we obtain $\langle \nabla F(\widetilde{x}), \widetilde{x} - x^* \rangle \geqslant \|\widetilde{x} - x^*\|_\Sigma^2$. On the other hand, the Cauchy-Schwarz inequality implies that

$$\langle g, \widetilde{x} - x^* \rangle \leqslant \|g\|_{\Sigma^{-1}} \cdot \|\widetilde{x} - x^*\|_\Sigma \,. \tag{94}$$

Plugging the previous inequalities in (93) yields $\|x' - x\|_\Sigma^2 \leqslant \|g\|_{\Sigma^{-1}} \cdot \|\widetilde{x} - x^*\|_\Sigma$, hence $\|x' - x\|_\Sigma \leqslant \|g\|_{\Sigma^{-1}}$; the inequality $\langle g, \widetilde{x} - x^* \rangle \leqslant \|g\|_{\Sigma^{-1}}^2$ then follows by (94). $\square$

References

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley New York, 2003.

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

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

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

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

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

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

[8] Y. Baraud and L. Birgé. Rho-estimators for shape restricted density estimation. *Stochastic Processes and their Applications*, 126(12):3888–3912, 2016.

[9] Y. Baraud and L. Birgé. Rho-estimators revisited: General theory and applications. *The Annals of Statistics*, 46(6B):3767–3804, 2018.

[10] Y. Baraud, L. Birgé, and M. Sart. A new method for estimation and model selection: ρ-estimation. *Inventiones mathematicae*, 207(2):425–517, 2017.

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

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

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

[14] R. Bhatia. *Positive Definite Matrices*, volume 16 of *Princeton Series in Applied Mathematics*. Princeton University Press, 2009.

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

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

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

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

[19] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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

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

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

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

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

[25] O. Catoni. The mixture approach to universal model selection. Technical report, École Normale Supérieure, 1997.

[26] O. Catoni. *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, 2004.

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

[28] N. Cesa-Bianchi and G. Lugosi. Worst-case bounds for the logarithmic loss of predictors. *Machine Learning*, 43(3):247–264, 2001.

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

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

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

[32] D. J. Foster, S. Kale, H. Luo, M. Mohri, and K. Sridharan. Logistic regression: the importance of being improper. In *Proceedings of the 31st Conference On Learning Theory (COLT)*, pages 167–208, 2018.

[33] J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*. Springer series in statistics, New York, 2001.

[34] P. Gaillard and S. Gerchinovitz. A chaining algorithm for online nonparametric regression. In *Proceedings of the 28th Annual Conference on Learning Theory (COLT)*, pages 764–796, 2015.

[35] E. I. George, F. Liang, and X. Xu. Improved minimax predictive densities under Kullback-Leibler loss. *The Annals of Statistics*, 34(1):78–91, 2006.

[36] P. D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

[37] P. D. Grünwald and W. Kotłowski. Open problem: Bounds on individual risk for log-loss predictors. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, volume 19, pages 813–816. PMLR, 2011.

[38] P. D. Grünwald and N. A. Mehta. Fast rates for general unbounded loss functions: from ERM to generalized Bayes. *arXiv preprint arXiv:1605.00252*, 2016.

[39] I. R. Harris. Predictive fit for natural exponential families. *Biometrika*, 76(4):675–684, 1989.

[40] J. A. Hartigan. The maximum likelihood prior. *The Annals of Statistics*, 26(6):2083–2103, 1998.

[41] E. Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

[42] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[43] E. Hazan, T. Koren, and K. Y. Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Proceedings of the 27th Conference on Learning Theory (COLT)*, pages 197–209, 2014.

[44] I. A. Ibragimov and R. Z. Has'minskii. *Statistical estimation: asymptotic theory*. Springer Science & Business Media, 1981.

[45] A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.

[46] S. M. Kakade and A. Y. Ng. Online bounds for Bayesian algorithms. In *Advances in Neural Information Processing Systems 17*, pages 641–648, 2005.

[47] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033 of *École d'Été de Probabilités de Saint-Flour*. Springer-Verlag Berlin Heidelberg, 2011.

[48] V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, 2015.

[49] F. Komaki. On asymptotic properties of predictive distributions. *Biometrika*, 83(2):299–313, 1996.

[50] T. Koren and K. Levy. Fast rates for exp-concave empirical risk minimization. In *Advances in Neural Information Processing Systems 28*, pages 1477–1485, 2015.

[51] W. Kotłowski and P. D. Grünwald. Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, pages 457–476, 2011.

[52] G. Lecué and S. Mendelson. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534, 2016.

[53] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer, 1998.

[54] F. Liang and A. R. Barron. Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Transactions on Information Theory*, 50(11):2708–2726, 2004.

[55] N. Littlestone. From on-line to batch learning. In *Proceedings of the 2nd annual workshop on Computational Learning Theory (COLT)*, pages 269–284. Morgan Kaufmann Publishers Inc., 1989.

[56] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

[57] U. Marteau-Ferey, D. Ostrovskii, F. Bach, and A. Rudi. Beyond least-squares: fast rates for regularized empirical risk minimization through self-concordance. *arXiv preprint arXiv:1902.03046*, 2019.

[58] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, 2007.

[59] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall/CRC, 2 edition, 1989.

[60] H. B. McMahan and M. Streeter. Open problem: Better bounds for online logistic regression. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, volume 23, pages 44.1–44.3, 2012.

[61] N. Mehta. Fast rates with high probability in exp-concave statistical learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1085–1093, 2017.

[62] S. Mendelson. Learning without concentration. *Journal of the ACM*, 62(3):21, 2015.

[63] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44:2124–2147, 1998.

[64] J. Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *Preprint*, 2019.

[65] G. D. Murray. A note on the estimation of probability density functions. *Biometrika*, 64(1):150–152, 1977.

[66] Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. Society of Industrial and Applied Mathematics, 1994.

[67] V. M. Ng. On the estimation of parametric density functions. *Biometrika*, 67(2):505–506, 1980.

[68] D. Ostrovskii and F. Bach. Finite-sample analysis of M-estimators using self-concordance. *arXiv preprint arXiv:1810.06838*, 2018.

[69] J. J. Rissanen. *Minimum description length principle*. Wiley Online Library, 1985.

[70] T. Roos and J. J. Rissanen. On sequentially normalized maximum likelihood models. In *Workshop on Information Theoretic Methods in Science and Engineering*, 2008.

[71] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

[72] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

[73] V. Spokoiny. Parametric estimation. finite sample theory. *The Annals of Statistics*, 40(6):2877–2909, 2012.

[74] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems 23*, pages 2199–2207, 2010.

[75] T. J. Sweeting, G. S. Datta, and M. Ghosh. Nonsubjective priors via predictive relative entropy regret. *The Annals of Statistics*, pages 441–468, 2006.

[76] E. Takimoto and M. K. Warmuth. The last-step minimax algorithm. In *International conference on Algorithmic Learning Theory (ALT)*, pages 279–290, 2000.

[77] M. Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

[78] S. van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, Cambridge, 1999.

[79] A. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998.

[80] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996.

[81] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[82] V. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173, 1998.

[83] G. Wahba. *Spline Models for Observational Data*, volume 59. SIAM, 1990.

[84] A. Wald. Statistical decision functions. *The Annals of Mathematical Statistics*, 20(2):165–205, 1949.

[85] L. Wasserman. *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[86] W. H. Wong and X. Shen. Probability inequalities for likelihood ratios and convergence rates of sieve mles. *The Annals of Statistics*, 23(2):339–362, 1995.

[87] Q. Xie and A. R. Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Transactions on Information Theory*, 46(2):431–445, 2000.

[88] Y. Yang and A. R. Barron. An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, 44(1):95–116, 1998.

[89] Y. Yang and A. R. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

[90] T. Zhang. From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.

[91] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 928–936, 2003.