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

Jaouad Mourtada   JAOUAD.MOURTADA@ENSAE.FR
CREST, ENSAE
Institut Polytechnique de Paris
Palaiseau, France

Stéphane Gaïffas   STEPHANE.GAIFFAS@LPSM.PARIS
Laboratoire de Probabilités, Statistique et Modélisation
Université Paris Diderot
Paris, France

Abstract

We introduce a procedure for conditional density estimation under logarithmic loss, which we call SMP (Sample Minmax Predictor). This estimator minimizes a new general excess risk bound for statistical learning. On standard examples, this bound scales as d/n with d the model dimension and n the sample size, and critically remains valid under model misspecification. Being an improper (out-of-model) procedure, SMP improves over within-model estimators such as the maximum likelihood estimator, whose excess risk degrades under misspecification. Compared to approaches reducing to the sequential problem, our bounds remove suboptimal $\log n$ factors and can handle unbounded classes. For the Gaussian linear model, the predictions and risk bound of SMP are governed by leverage scores of covariates, nearly matching the optimal risk in the well-specified case without conditions on the noise variance or approximation error of the linear model. For logistic regression, SMP provides a non-Bayesian approach to calibration of probabilistic predictions relying on virtual samples, and can be computed by solving two logistic regressions. It achieves a non-asymptotic excess risk of $O((d + B^2 R^2)/n)$, where R bounds the norm of features and B that of the comparison parameter; by contrast, no within-model estimator can achieve better rate than $\min(BR/\sqrt{n}, e^{BR}/n)$ in general (Hazan et al., 2014). This provides a more practical alternative to Bayesian approaches, which require approximate posterior sampling, thereby partly addressing a question raised by Foster et al. (2018).

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

1. Introduction

Consider the standard problem of density estimation: given an i.i.d. sample $Z_1, \ldots, Z_n$ from an unknown distribution P on some measurable space $\mathcal{Z}$, the goal is to produce a good approximation $\widehat{P}_n$ of P. One way to measure the quality of an estimate $\widehat{P}_n$ is through its predictive risk: given a base measure μ on $\mathcal{Z}$, the risk of a density g on $\mathcal{Z}$ with respect to μ is given by

$$R(g) = \mathbb{E}[\ell(g, Z)], \quad \text{where} \quad \ell(g, z) = -\log g(z) \tag{1}$$

for $z \in \mathcal{Z}$ and where Z is a random variable with distribution P. Letting $\mathcal{G}$ denote the set of all probability densities on $\mathcal{Z}$ with respect to μ, the loss function $\ell : \mathcal{G} \times \mathcal{Z} \to \mathbb{R}$ defined by (1), called *logarithmic* (or *negative log-likelihood*, *entropy* or *logistic*) loss, measures the error of the density $g \in \mathcal{G}$ (which can be interpreted as a probabilistic prediction of the outcome) given outcome $z \in \mathcal{Z}$. This loss function is standard in the information theory literature, due to its link with coding (Cover and Thomas, 2006). The risk of a density g can be interpreted in relation to the joint probability assigned by g to a large i.i.d. test sample $Z'_1, \ldots, Z'_m$ from P: by the law of large numbers, as m tends to infinity, almost surely

$$\prod_{j=1}^{m} g(Z'_j) = \exp\Big(-\sum_{j=1}^{m} \ell(g, Z'_j)\Big) = \exp\Big(-m[R(g) + o(1)]\Big).$$

In addition, assume that P of Z has a density $p \in \mathcal{G}$; we then have, for every $g \in \mathcal{G}$,

$$R(g) - R(p) = \mathbb{E}\Big[\log\Big(\frac{p(Z)}{g(Z)}\Big)\Big] = \int_{\mathcal{Z}} \log\Big(\frac{p}{g}\Big) p \, \mathrm{d}\mu = \mathrm{KL}(p \cdot \mu, g \cdot \mu) \geqslant 0,$$

where $\mathrm{KL}(P, Q) := \int_{\mathcal{Z}} \log\big(\frac{\mathrm{d}P}{\mathrm{d}Q}\big) \mathrm{d}P$ denotes the *Kullback-Leibler divergence* (or *relative entropy*) between distributions P and Q. In particular, the risk is minimized by the true density p (if it exists), and prediction under logarithmic loss is equivalent to density estimation under Kullback-Leibler risk.

Our aim is to find *estimators*, which associate to any sample $Z_1, \ldots, Z_n$ a density $\widehat{g}_n \in \mathcal{G}$, whose risk is controlled in some general setting. While it is typically impossible to obtain finite-sample guarantees without any assumption on the underlying distribution P (see e.g. Devroye et al., 1996, Section 7.1), oftentimes one expects this distribution to possess some structure. In such cases, it is natural to introduce inductive bias in the procedure; one standard way to do so is to select a suitable class of densities $\mathcal{F} \subset \mathcal{G}$ (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 of it.

A classical approach is then to assume that the model $\mathcal{F}$ is *well-specified*, in the sense that it contains the true density p. In this case, the problem of estimating P falls within the classical framework of parametric statistics (Ibragimov and Has'minskii, 1981; van der Vaart, 1998; Lehmann and Casella, 1998). 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 (Hájek, 1972; Le Cam, 1986; Ibragimov and Has'minskii, 1981). The same problem can also be treated for a fixed sample size, through the lens of statistical decision theory (Wald, 1949; Lehmann and Casella, 1998), which emphasizes optimal estimators in the average (Bayesian) and minimax senses. Generally speaking, these approaches offer precise descriptions of achievable rates of convergence (up to correct leading constants) and of *efficient* estimators that make the best use of available data. A major limitation of this approach, however, is that these results rely on the unrealistic 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 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. This task can naturally be cast in the framework of Statistical Learning Theory (Vapnik, 1998), where one constrains the comparison class $\mathcal{F}$ while making few modeling assumptions about the true distribution. Given a class $\mathcal{F}$ of densities, the performance of an estimator $\widehat{g}_n$ is evaluated in terms of its *excess risk* with respect to the class $\mathcal{F}$, namely

$$\mathcal{E}(\widehat{g}_n) := R(\widehat{g}_n) - \inf_{f \in \mathcal{F}} R(f) \,.$$

We say that the estimator $\widehat{g}_n$ is *proper* (or a *plug-in* estimator) when it takes value inside the class $\mathcal{F}$, otherwise $\widehat{g}_n$ will be referred to as an *improper* procedure. Below, we discuss two established approaches to this problem.

Maximum Likelihood Estimation. Arguably the simplest and most standard procedure is the *Maximum Likelihood Estimator* (MLE), or *Empirical Risk Minimizer* (ERM) with logarithmic loss, given by

$$\widehat{f}_n := \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(f, Z_i) = \arg\max_{f \in \mathcal{F}} \prod_{i=1}^n f(Z_i) \,. \tag{2}$$

Assume now that $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ is some parametric model indexed by an open subset $\Theta \subset \mathbb{R}^d$, such that the density $f_\theta(z)$ depends smoothly on θ, and denote $\widehat{f}_n = f_{\widehat{\theta}_n}$ the MLE. First, consider the well-specified case where the true distribution P belongs to the model, say $P = f_{\theta^*} \cdot \mu$, and denote $I(\theta^*) := \mathbb{E}[-\nabla^2 \log f_\theta(Z)]|_{\theta=\theta^*}$ the Fisher information matrix, assumed invertible. Then, under standard regularity and moment conditions (van der Vaart, 1998; Ibragimov and Has'minskii, 1981), we have as $n \to \infty$,

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \overset{\text{(d)}}{\to} \mathcal{N}(0, I(\theta^*)^{-1}) \quad \text{while} \quad \mathcal{E}(f_\theta) = \frac{1}{2}\|\theta - \theta^*\|_{I(\theta^*)}^2 + o(\|\theta - \theta^*\|^2) \,,$$

where we denote $\|u\|_A := \langle Au, u \rangle^{1/2}$ for any $u \in \mathbb{R}^d$ and symmetric positive matrix A. This implies that $2n\mathcal{E}(f_{\widehat{\theta}_n})$ converges in distribution to a χ_d^2 distribution; hence, under suitable domination conditions, the asymptotic excess risk of the MLE satisfies $\mathbb{E}[\mathcal{E}(\widehat{f}_n)] = d/(2n) + o(n^{-1})$. This asymptotic performance turns out to be unimprovable in the well-specified case: for instance, MLE is locally asymptotically minimax optimal (Hájek, 1972; Le Cam and Yang, 2000).

In contrast to its optimality in the well-specified case, the performance of MLE can degrade in the general misspecified case, where it depends on the true distribution P. Indeed, let $\theta^* = \arg\min_{\theta \in \Theta} R(f_{\theta^*})$ be the optimal parameter, and $G = \mathbb{E}[\nabla \ell(f_\theta, Z) \nabla \ell(f_\theta, Z)^\top]|_{\theta=\theta^*}$, $H = \mathbb{E}[\nabla^2 \ell(f_\theta, Z)]|_{\theta=\theta^*}$; when P belongs to the model, $G = H = I(\theta^*)$, but in general those matrices are distinct. In this case, under suitable conditions, it follows from general results on the asymptotic behavior of M-estimators (van der Vaart, 1998; White, 1982) that

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \overset{\text{(d)}}{\to} \mathcal{N}(0, H^{-1}GH^{-1}) \quad \text{and} \quad \mathcal{E}(f_\theta) = \frac{1}{2}\|\theta - \theta^*\|_H^2 + o(\|\theta - \theta^*\|^2) \,.$$

3

Again under suitable domination conditions, this implies that, as $n \to \infty$,

$$\mathbb{E}[\mathcal{E}(\widehat{f}_n)] = \frac{\mathrm{tr}(H^{-1/2}GH^{-1/2})}{2n} + o\Big(\frac{1}{n}\Big) = \frac{d_{\mathrm{eff}}}{2n} + o\Big(\frac{1}{n}\Big) ; \tag{3}$$

here, the constant $d_{\mathrm{eff}} := \mathrm{tr}(H^{-1/2}GH^{-1/2})$ depends on the distribution P, and can typically be arbitrarily large, as will be seen below in the case of logistic regression. In fact, degradation under model misspecification is not specific to MLE, and is typically a limitation shared by any *proper* (or *plug-in*) estimator that returns a distribution within the class $\mathcal{F}$, such as penalized MLE. Finally, let us stress that, while we adopted a simplified asymptotic viewpoint in this discussion for the sake of clarity, our focus will be on finite-sample bounds, with an explicit dependence on problem-dependent parameters.

Sequential prediction and online-to-offline conversion. In contrast, distribution-free excess risk bounds have been obtained in the literature (Barron, 1987; Catoni, 2004; Yang, 2000; Juditsky et al., 2008; Audibert, 2009) through a reduction to the comparatively much better understood setting of sequential prediction under logarithmic loss (Merhav and Feder, 1998; Cesa-Bianchi and Lugosi, 2006; Shtarkov, 1987; Grünwald, 2007). In this problem, which is connected to coding (Cover and Thomas, 2006) and the minimum description length (MDL) principle (Rissanen, 1985; Grünwald, 2007), one seeks to control *cumulative* criteria such as the cumulative excess risk, or the regret

$$\sum_{i=1}^{n} \ell(\widehat{g}_{i-1}, Z_i) - \inf_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f, Z_i)$$

over all sequences $Z_1, \ldots, Z_n \in \mathcal{Z}$, where $\widehat{g}_{i-1}$ is selected based on $Z_1, \ldots, Z_{i-1}$. The control of such cumulative quantities is significantly simplified by the observation that

$$\sum_{i=1}^{n} \ell(\widehat{g}_{i-1}, Z_i) - \inf_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f, Z_i) = -\log\left(\frac{\prod_{i=1}^{n} \widehat{g}_{i-1}(Z_i)}{\sup_{f \in \mathcal{F}} \prod_{i=1}^{n} f(Z_i)}\right),$$

where the ratio inside the logarithm can be interpreted as a ratio of joint densities over $Z_1, \ldots, Z_n$. This enables one to determine the minimax regret (Shtarkov, 1987), as well as to control the regret of specific sequential prediction strategies $\widehat{g}_0, \ldots, \widehat{g}_{n-1}$. Among those, arguably the most standard are Bayesian mixture strategies (Vovk, 1998; Littlestone and Warmuth, 1994; Merhav and Feder, 1998; Cesa-Bianchi and Lugosi, 2006) with near-optimal guarantees (Clarke and Barron, 1994; Xie and Barron, 2000; Merhav and Feder, 1998; Cesa-Bianchi and Lugosi, 2006), where given a prior distribution π on the parameter space Θ, $\widehat{g}_i$ is the *Bayesian predictive posterior*:

$$\widehat{g}_i(z) = \frac{\int_\Theta f_\theta(Z_1) \cdots f_\theta(Z_i) f_\theta(z) \pi(\mathrm{d}\theta)}{\int_\Theta f_\theta(Z_1) \cdots f_\theta(Z_i) \pi(\mathrm{d}\theta)} = \int_\Theta f_\theta(z)\, \pi(\mathrm{d}\theta | Z_1, \ldots, Z_i) . \tag{4}$$

For smooth, bounded parametric families of dimension d, the minimax cumulative excess risk and regret are known to scale as $(d \log n)/2 + C(\mathcal{F})$ for some constant $C(\mathcal{F})$ depending on the model, see Clarke and Barron (1994); Merhav and Feder (1998). Note that regret bounds hold for any sequence, and in particular do not require the sequence of observations

to be sampled from a distribution in the model. A generic procedure called *online to batch conversion* (Littlestone, 1989; Cesa-Bianchi et al., 2004) enables one to convert any guarantee on the cumulative excess risk into one on the *non-cumulative* excess risk for the average of the successive densities output by the sequential procedure, namely

$$\bar{g}_n = \frac{1}{n+1} \sum_{i=0}^{n} \widehat{g}_i \,. \tag{5}$$

When applied to Bayes mixture rules, this yields the so-called *progressive mixture* or *mirror averaging* procedure (Yang and Barron, 1999; Catoni, 1997, 2004; Juditsky et al., 2008; Audibert, 2009), with excess risk bounded by $O((d \log n)/n + C(\mathcal{F})/n)$.

   While appropriate for sequential prediction, this approach is not fully satisfactory in the statistical learning setting considered here, for the following reasons. First, the obtained $O(d \log n/n)$ rate features a suboptimal $\log n$ factor, when compared to the $O(d/n)$ rate of MLE in the well-specified case; this highlights the inefficiency of the averaged estimator $\bar{g}_n$, which mixes estimators $\widehat{g}_i$ computed with only a fraction of the sample. Obtaining bounds of $O(d/n)$ for the excess risk was posed as an open problem (first question in Grünwald and Kotłowski, 2011). Second, the minimax regret (and in particular the model-dependent constant $C(\mathcal{F})$) is typically *infinite* (Shtarkov, 1987; Clarke and Barron, 1994; Rissanen, 1996; Grünwald, 2007) for unbounded "infinite-volume" classes $\mathcal{F}$ including Gaussian models, so that no uniform guarantee can be obtained over such classes through regret minimization and online-to-offline conversion, reflecting the poor localization of such bounds. These first two limitations are shared by any approach reducing to the sequential problem, which takes into account early rounds where few observations are available. A third limitation lies in the computational requirements of such procedures: in particular, Bayesian mixture approaches involve—absent a conjugate prior allowing exact computations—approximate posterior computations, which are often significantly more expensive than maximum likelihood optimization, inhibiting practical use of such methods.

1.1 Our contributions

Let us now summarize our main contributions. Note that, while the previous discussion dealt with density estimation, most of this work in fact deals with *conditional* density estimation, where one seeks to estimate the conditional distribution of a response Y to an input variable X, under logarithmic loss $\ell(f, (X, Y)) = -\log f(Y|X)$ (see Section 2.2).

SMP: a procedure for conditional density estimation. We introduce a general procedure for predictive density estimation under entropy risk. This estimator, which we call *Sample Minmax Predictor* (SMP), is obtained by minimizing a new general excess risk bound for supervised statistical learning (Theorem 1), and in particular conditional density estimation (Theorem 2). In short, SMP is the solution of some minmax problem obtained by considering *virtual samples*. SMP satisfies an excess risk bound valid under model misspecification, and unlike previous approaches does not rely on a reduction to the sequential problem, thereby improving rates for parametric classes from $O(d \log n/n)$ to $O(d/n)$ for our considered models, addressing a question raised by Grünwald and Kotłowski (2011) in these cases ("individual risk" in this paper refers to the excess risk).

SMP for the Gaussian linear model. We apply SMP to the *Gaussian linear model* $\mathcal{F} = \{f_\theta(\cdot|x) = \mathcal{N}(\langle\theta, x\rangle, \sigma^2) : \theta \in \mathbb{R}^d\}$ for some $\sigma^2 > 0$, a classical conditional model for a scalar response $y \in \mathbb{R}$ to covariates $x \in \mathbb{R}^d$. SMP then smoothes predictions in terms of *leverage scores*, and for every distribution of covariates, 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 approximation error of the linear model or noise variance (Theorem 7). This yields an excess risk bound of $d/n + O((d/n)^2)$ over the class $\mathcal{F}$ under some regularity assumptions on covariates (Corollary 8); such a guarantee cannot be obtained for a within-model estimator, or through a regret minimization approach.

We also consider a Ridge-regularized variant of SMP, and study its performance on balls of the form $\mathcal{F}_B = \{f_\theta : \|\theta\| \leqslant B\}$ for $B > 0$. For covariates X bounded by $R > 0$, we establish two guarantees: a "finite-dimensional" bound of $O(d\log(BR/\sqrt{d})/n)$ (Proposition 10), removing an extra $\log n$ term from results of Kakade and Ng (2005) in the sequential case, and a dimension-free "nonparametric" bound (Theorem 9), where explicit dependence on d is replaced by a dependence on the covariance structure of covariates, matching well-specified minimax rates over such balls in infinite dimension (Caponnetto and De Vito, 2007).

SMP for logistic regression. We then turn to logistic regression, arguably the most standard model for a binary response $y \in \{-1, 1\}$ to covariates $x \in \mathbb{R}^d$, given by $\mathcal{F} = \{f_\theta(1|x) = \sigma(\langle\theta, x\rangle) : \theta \in \mathbb{R}^d\}$, where $\sigma(u) = e^u/(1 + e^u)$. In this case, SMP admits a simple form, and its prediction can be computed by solving two logistic regressions. Assuming that $\|X\| \leqslant R$, we show that a Ridge-penalized variant of SMP achieves excess risk $O((d + B^2 R^2)/n)$ with respect to the ball $\mathcal{F}_B = \{f_\theta : \|\theta\| \leqslant B\}$ for all $B > 0$ (Corollary 14), together with dimension-free bounds (Theorem 13). In contrast, results of Hazan et al. (2014) show that no within-model estimator can achieve better rate than $\min(BR/\sqrt{n}, e^{BR}/n)$ without further assumptions. Compared to approaches obtaining fast rates through Bayesian mixtures (Kakade and Ng, 2005; Foster et al., 2018), computation of SMP replaces posterior sampling by optimization. SMP thus provides a natural non-Bayesian approach to uncertainty quantification and calibration of probabilistic estimates, relying on virtual samples.

1.2 Related work

Well-specified density estimation. There is a rich statistical literature on predictive density estimation under entropy risk in the well-specified case (where the true distribution is assumed to belong to the model), see Harris (1989); Komaki (1996); Hartigan (1998); Aslan (2006); Liang and Barron (2004); George et al. (2006); Sweeting et al. (2006); Brown et al. (2008) and references therein. First, as mentioned above, MLE is known to be asymptotically normal and efficient (van der Vaart, 1998; Ibragimov and Has'minskii, 1981; Le Cam and Yang, 2000) in this case; its asymptotic optimality can be formalized precisely by Hájek's local asymptotic minimax theorem (Hájek, 1972; Le Cam and Yang, 2000). Beyond this optimality result, a number of refinements have been explored: improvement of Bayes predictive distributions over the MLE for finite samples (Aitchison, 1975), higher-order risk asymptotics (Hartigan, 1998; Ghosh, 1994; Aslan, 2006) and second-order minimax procedures (Aslan, 2006), exact minimax procedures for location and scale families (Liang and Barron, 2004), as well as admissibility and shrinkage for the Gaussian model (Brown et al.,

2008). While related to this line of work, our approach differs from it by relaxing the (restrictive) assumption that the distribution of interest belongs to the specified model; another difference with some of the aforementioned references is our non-asymptotic focus.

Non-asymptotic analyses of estimators under misspecification. The asymptotic behavior of MLE (including consistency and asymptotic normality) in the misspecified case is also well-understood (White, 1982; van der Vaart, 1998). Beyond the asymptotic setting, non-asymptotic analyses of MLE and related procedures have been carried (van de Geer, 1999; Birgé and Massart, 1993, 1998; Yang and Barron, 1998; Wong and Shen, 1995; Spokoiny, 2012), by using techniques from empirical process theory (van der Vaart and Wellner, 1996; Talagrand, 2014; Massart, 2007; Boucheron et al., 2013). In addition to these classical references, we mention two approaches that circumvent in different ways reliance on the machinery of empirical process theory. First, Zhang (2006) relies on information-theoretic inequalities to analyze Bayesian and penalized estimators; this approach is extended by Grünwald and Mehta (2019), who obtain bounds in terms of refined complexity measures. Second, Ostrovskii and Bach (2018) developed an analysis relying on self-concordance, which applies in particular to logistic regression. Overall, this literature differs from ours in that it studies estimators such as (penalized) MLE, which inevitably degrade for some misspecified distributions.

Sequential prediction. As mentioned previously, the sequential variant of prediction under logarithmic loss is well-studied (Shtarkov, 1987; Clarke and Barron, 1994; Merhav and Feder, 1998; Vovk, 1998; Cesa-Bianchi and Lugosi, 2006; Grünwald, 2007). These guarantees on cumulative criteria have been transported to the individual excess risk considered here (Barron, 1987; Catoni, 2004; Yang and Barron, 1999; Juditsky et al., 2008; Audibert, 2009). To the best of our knowledge, this online-to-offline conversion was the only approach to obtaining distribution-free excess risk guarantees. As mentioned above, reduction to the sequential case is suboptimal, in that it leads to extra logarithmic factors in the rate and cannot provide uniform guarantees over unbounded models. Our general guarantee for SMP provides a more "localized" risk bound adapted to such situations.

Stability. Our general bound on the excess risk is related to the approach in terms of stability of the loss of the predictor under sample changes (Bousquet and Elisseeff, 2002; Rakhlin et al., 2005; Shalev-Shwartz et al., 2010; Koren and Levy, 2015), in particular in its use of exchangeability. While close in spirit, our bounds involve a different quantity; the difference between the two is particularly apparent in the context of logistic regression, where it enables us to remove some exponential constants.

Logistic regression. An important motivation for this work was recent progress and questions on logistic regression, arguably the most common model for conditional density estimation with binary response (Berkson, 1944; McCullagh and Nelder, 1989; van der Vaart, 1998). Under boundedness assumptions, it can be seen as a special convex and Lipschitz stochastic optimization problem, for which slow rates of convergence are available (Zinkevich, 2003; Nemirovski et al., 2009; Bubeck, 2015). In addition, logistic regression is also an exp-concave problem, which enables fast rates (Hazan et al., 2007; Koren and Levy, 2015; Mehta, 2017), but with an exponential dependence on the domain radius. It is shown by Hazan et al. (2014) that such rates are unimprovable without further assumptions. To obtain improved

results, one thread of work proceeds under additional assumptions, and performs a refined analysis using (generalized) *self-concordance* of the logistic loss (Bach, 2010, 2014; Bach and Moulines, 2013; Ostrovskii and Bach, 2018; Marteau-Ferey et al., 2019a); this leads to distribution-dependent guarantees which improve for favorable distributions, but exhibit exponential dependence in the worst case. Another approach consists in using out-of-model procedures, for which the lower bound of Hazan et al. (2014) does not apply. By using Bayes mixtures strategies and reducing to the sequential problem, Kakade and Ng (2005); Foster et al. (2018) establish fast risk rates without exponential dependence on the norm, bypassing the previous lower bound; the question of finding a practical procedure enjoying such guarantees without expensive posterior sampling is left open in Foster et al. (2018). Our work is cast in the same setting under weak distributional assumptions, and provides a practical approach with fast rates guarantees in this case. We note that our analysis of SMP also relies on self-concordance, though it applies to a different estimator.

Min-max procedures for linear regression. As mentioned above, SMP provides a simple alternative to Bayes mixtures for logistic regression. Interestingly, similar procedures have been introduced for linear regression with square loss, namely the Vovk-Azoury-Warmuth (VAW) algorithm (Vovk, 2001; Azoury and Warmuth, 2001) for sequential prediction, and the Forster-Warmuth estimator (Forster and Warmuth, 2002) for statistical learning. Our estimator can be seen as a counterpart for density estimation and logistic regression to these linear regression procedures in several ways. First, they are simpler than a general aggregation approach that would "mix" over a continuous class. In addition, the predictions from these methods involve a leverage correction, like SMP for Gaussian linear density estimation (though the correction concerns the variance rather than the mean for SMP). Finally, while Vovk (2001) derives the VAW procedure from aggregation and mixability considerations, the approach of Forster and Warmuth (2002); Azoury and Warmuth (2001) also relies on a min-max loss difference.

1.3 Outline and notation

This paper is organized as follows. In Section 2, we introduce the setting and state a general excess risk bound for supervised learning (Theorem 1) and its instantiation to conditional density estimation (Theorem 2), minimized by SMP, which will be used throughout. Section 3 provides direct consequences of the previous bounds in the context of (unconditional) density estimation with multinomial and Gaussian models. In Section 4, we study SMP and its guarantees for conditional density estimation with the Gaussian linear model. We finally turn to logistic regression in Section 5. The proofs are gathered in Section 7, while Section 6 concludes.

Notation. Throughout this text, we denote $\langle x, y \rangle := x^\top y$ the canonical scalar product of $x, y \in \mathbb{R}^d$, and $\|x\| := \langle x, x \rangle^{1/2}$ the associated Euclidean norm. Likewise, for any symmetric positive semi-definite $d \times d$ matrix Σ, we let $\langle x, y \rangle := \langle \Sigma x, y \rangle$ and $\|x\|_\Sigma = \langle x, x \rangle_\Sigma^{1/2}$. We denote by $\mathrm{dist}(x, A) = \inf_{y \in A} \|x - y\|$ the distance of a point $x \in \mathbb{R}^d$ to a subset $A \subset \mathbb{R}^d$.

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, corresponding respectively to the feature, label and prediction spaces, and let $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ be a loss function. 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 also consider a 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 $f \in \mathcal{F}$. When no penalization is used ($\phi \equiv 0$) 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*

$$R(g) = \mathbb{E}[\ell(g, Z)] = \mathbb{E}[\ell(g(X), Y)] \tag{6}$$

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

$$\mathcal{E}(g) = R(g) - \inf_{f \in \mathcal{F}} R(f). \tag{7}$$

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 $Z_1^n = (Z_1, \ldots, Z_n)$ 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 $Z_1^n = (Z_1, \ldots, Z_n)$, we denote

$$\widehat{f}_{\phi,n} \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell_\phi(f, Z_i) \tag{8}$$

a (penalized) *empirical risk minimizer* (ERM); when $\phi \equiv 0$, 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 $Z_1, \ldots, Z_n$. We also introduce

$$\widehat{f}_{\phi,n}^{z} := \arg\min_{f \in \mathcal{F}} \left\{ \sum_{i=1}^{n} \ell_\phi(f, Z_i) + \ell_\phi(f, z) \right\} \tag{9}$$

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*

$$\mathbb{E}\big[\mathcal{E}_\phi(\widehat{g}_n)\big] \leqslant \mathbb{E}_{Z_1^n, X}\Big[\sup_{y \in \mathcal{Y}} \big\{ \ell(\widehat{g}_n(X), y) - \ell_\phi(\widehat{f}_{\phi,n}^{(X,y)}(X), y) \big\} \Big] \tag{10}$$

9

where $\widehat{f}_{\phi,n}^z$ is defined by (9) for $z \in \mathcal{Z}$ and $Z = (X, Y) \sim P$ is independent of Z_1^n. In addition, the right-hand side of (10) is minimized by the predictor

$$\widetilde{f}_{\phi,n}(x) = \arg\min_{\widehat{y} \in \widehat{\mathcal{Y}}} \sup_{y \in \mathcal{Y}} \left\{ \ell(\widehat{y}, y) - \ell_\phi(\widehat{f}_{\phi,n}^{(x,y)}(x), y) \right\}, \tag{11}$$

which we call SMP (Sample Minmax Predictor) whenever it exists, in which case (10) becomes

$$\mathbb{E}\big[\mathcal{E}_\phi(\widetilde{f}_{\phi,n})\big] \leqslant \mathbb{E}_{Z_1^n, X}\Big[\inf_{\widehat{y} \in \widehat{\mathcal{Y}}} \sup_{y \in \mathcal{Y}} \left\{ \ell(\widehat{y}, y) - \ell_\phi(\widehat{f}_{\phi,n}^{(X,y)}(X), y) \right\} \Big]. \tag{12}$$

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 $\widehat{f}_{\phi,n}^{(X,y)}$ 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 (12) will also be small.

Stability techniques were exploited by Bousquet and Elisseeff (2002) to establish guarantees for learning algorithms such as ERM or approximate ERM, although earlier instances of this method can be found in Vapnik and Chervonenkis (1974); Devroye and Wagner (1979); Haussler et al. (1994). Stability arguments were used by Bousquet and Elisseeff (2002); Shalev-Shwartz et al. (2010) to prove fast rates of order $O(1/n)$ for ERM in strongly convex stochastic optimization problems and more recently by Koren and Levy (2015) for exp-concave problems. While related to the notion of stability, the excess risk bound of Theorem 1 differs from the latter 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 (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010; Srebro et al., 2010; Koren and Levy, 2015). 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. The difference between the two is most apparent in the context of logistic regression (see Section 5 below), where it is critical to obtain improved guarantees that could not be derived from loss stability of regularized risk minimizers.

It is worth noting that the SMP (11) whose risk is controlled in (12) is *not* the regularized 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 fact, we will show in what follows that SMP enjoys guarantees which are not achievable by proper predictors such as regularized ERM.

2.2 Conditional density estimation with the logarithmic loss

We now turn to 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 $g : \mathcal{Y} \to \mathbb{R}^+$ such that $\int_{\mathcal{Y}} g \, d\mu = 1$. The logarithmic loss is defined as $\ell(g, y) = -\log g(y)$ for $g \in \widehat{\mathcal{Y}}$ 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 $z = (x, y)$. In this case, the ERM (8) corresponds to the

(conditional) maximum likelihood estimator (MLE). The risk of a conditional density f is

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

whenever this expectation is defined. For any conditional densities f, g with respect to μ,

$$R(g) - R(f) = \mathbb{E}\bigg[\log \frac{f(Y|X)}{g(Y|X)}\bigg], \tag{13}$$

which does not depend on the dominating measure μ (since the ratio f/g does not), but only on the conditional distributions $f\mu, g\mu$ and the distribution of X. 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 (12) 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 (11) writes*

$$\widetilde{f}_{\phi,n}(y|x) = \frac{\widehat{f}_{\phi,n}^{(x,y)}(y|x)e^{-\phi(\widehat{f}_{\phi,n}^{(x,y)})}}{\int_{\mathcal{Y}} \widehat{f}_{\phi,n}^{(x,y')}(y'|x)e^{-\phi(\widehat{f}_{\phi,n}^{(x,y')})}\mu(\mathrm{d}y')}, \tag{14}$$

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 (12) writes

$$\mathbb{E}\big[\mathcal{E}_\phi(\widetilde{f}_{\phi,n})\big] \leqslant \mathbb{E}_{Z_1^n, X}\bigg[\log \Big(\int_{\mathcal{Y}} \widehat{f}_{\phi,n}^{(X,y)}(y|X)e^{-\phi(\widehat{f}_{\phi,n}^{(X,y)})}\mu(\mathrm{d}y)\Big)\bigg]. \tag{15}$$

Remark 3 *In the non-regularized case where $\phi \equiv 0$, 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 (15) takes the form:

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

Theorem 2 is proved in Section 7.1. The SMP (14) 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 (15) 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 quite 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 (14) closely resembles that of a sequential prediction strategy called Sequential Normalized Maximum Likelihood (SNML), introduced by Roos and Rissanen (2008) and related to the Last Step Minimax algorithm (which restricts to proper predictions) from Takimoto and Warmuth (2000)[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 (Kotłowski and Grünwald, 2011), and in fact minimax for some specific families (Bartlett et al., 2013). By contrast, in our case the SMP estimator naturally arises as the minimizer of a novel upper bound on the *non-cumulative* excess risk.

Finally, the excess risk bound (15), which is specific to the statistical learning setting, provides in this case a more "local" control than bounds derived from the sequential setting, such as the averaged Bayesian mixture approach (Yang and Barron, 1999; Catoni, 2004; Juditsky et al., 2008; Audibert, 2009), as well as the normalized maximum likelihood (NML) (Shtarkov, 1987; Merhav and Feder, 1998; Cesa-Bianchi and Lugosi, 2006; Grünwald, 2007). This will be illustrated in the following sections.

We also note that a similar "last step min-max" approach was adopted by Forster and Warmuth (2002) to design a procedure for linear regression with square loss, see also Vovk (2001); Azoury and Warmuth (2001) for a related method in the online setting.

3. Some consequences for 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 thus omitted[2], and no penalization is used ($\phi \equiv 0$). In other words, given access to an i.i.d. sample $(Y_1, \ldots, Y_n)$ 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

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

between distributions P and Q (which is infinite whenever P is not absolutely continuous with respect to Q). Let $f^* \in \arg\min_{f \in \mathcal{F}} R(f)$ (assuming it exists); if $\mathrm{KL}(P, f^*) < +\infty$ then $f^* = \arg\min_{f \in \mathcal{F}} \mathrm{KL}(P, f)$ and the excess risk (7) 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 *KL risk*.

---

1. Specifically, the prediction of SMP coincides with that of the SNML-1 algorithm from Roos and Rissanen (2008) at step $n + 1$, while SNML-2 from Roos and Rissanen (2008) (simply called SNML in subsequent work Kotłowski and Grünwald (2011); Bartlett et al. (2013)) is slightly different: it minimizes worst-case regret with respect to next step ERM on the whole sequence, instead of just the last sample.
2. While conditional density estimation can be cast as a special case of density estimation, we adopt the opposite perspective since SMP exploits the conditional structure.

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 (Cesa-Bianchi et al., 2004) of sequential prediction strategies. In all considered examples, SMP improves these estimators.

3.1 Finite alphabet: the multinomial model

In this section, we assume that $\mathcal{Y}$ is a finite set with d elements, μ is the counting measure and $\mathcal{F} = \{(p(y))_{y \in \mathcal{Y}} \in \mathbb{R}_+^{\mathcal{Y}} : \sum_{y \in \mathcal{Y}} p(y) = 1\}$ is the multinomial model (which is always well-specified). For any $y \in \mathcal{Y}$, we let $N_n(y) = \sum_{i=1}^n \mathbf{1}(Y_i = y)$.

Proposition 4 *If $\mathcal{Y}$ is a finite set with d elements, then SMP corresponds to the* Laplace estimator

$$\widetilde{f}_n(y) = \frac{N_n(y) + 1}{n + d}. \tag{16}$$

In addition, the bound (15) *writes in this case*

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \log\left(\frac{n+d}{n+1}\right) \leqslant \frac{d-1}{n}. \tag{17}$$

Proposition 4 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 (17) is tight: it is an equality when Y is constant almost surely.

On MLE. The MLE is given by $\widehat{f}_n(y) = N_n(y)/n$. Its expected risk is infinite unless P is concentrated on a single point. Indeed, let $y_0, y_1 \in \mathcal{Y}$ be distinct elements such that $\mathbb{P}(Y = y_0), \mathbb{P}(Y = y_1) > 0$; with positive probability, $Y_1 = \cdots = Y_n = y_0$, so that $\widehat{f}_n(y) = \mathbf{1}(y = y_0)$, $\ell(\widehat{f}_n, y_1) = +\infty$ and thus $R(\widehat{f}_n) = +\infty$. Hence, $\mathbb{E}[R(\widehat{f}_n)] = +\infty$. In order to obtain non-vacuous expected risk for 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 log ratios of densities are bounded. In this case, whenever $p \in \mathcal{F}_\delta$, the excess risk of MLE has asymptotically efficient rate $(d-1)/(2n) + o(n^{-1})$. This reflects the fact that the model is well-specified.

On online to batch conversion. The minimax cumulative regret with respect to the class $\mathcal{F}$ scales (for fixed d and as $n \to \infty$) as $(d-1)(\log n)/2 + O(1)$ (Cesa-Bianchi and Lugosi, 2006, Remark 9.2 p. 256). Hence, any upper bound based on online-to-batch conversion (Cesa-Bianchi et al., 2004) can be no better than $(d-1)(\log n)/(2n) + O(1/n)$.

3.2 The Gaussian location model

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

Proposition 5 *A risk minimizer $f^* = \mathcal{N}(\theta^*, \Sigma) \in \mathcal{F}$ exists if and only if the random variable $Y \sim P$ satisfies $\mathbb{E}\|Y\| < +\infty$, 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 (15) writes*

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant d \log\Big(1 + \frac{1}{n}\Big) \leqslant \frac{d}{n}. \tag{18}$$

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 5 is given in Section 7.2 below. This bound is valid under misspecification: it does not depend on the distribution of Y, and is at most twice the optimal asymptotic risk $d/(2n)$.

On MLE and proper estimators. Assume that $\mathbb{E}\|Y\|^2 < +\infty$ and define $\Sigma_Y = \mathbb{E}[(Y - \mathbb{E}Y)(Y - \mathbb{E}Y)^\top]$. 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. This limitation is in fact shared by any proper estimator of the form $f_{\widehat{\theta}_n} = \mathcal{N}(\widehat{\theta}_n, \Sigma)$ for some $\widehat{\theta}_n$, 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. Lehmann and Casella, 1998) 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 proper estimator $f_{\widehat{\theta}_n}$,

$$\sup_{\theta^* \in \mathbb{R}^d} \mathbb{E}_{Y_1,\ldots,Y_n \sim P_{\theta^*}^{\otimes n}}\big[\mathcal{E}(f_{\widehat{\theta}_n})\big] = \frac{1}{2}\sup_{\theta^* \in \mathbb{R}^d} \mathbb{E}_{\theta^*}\big\|\widehat{\theta}_n - \mathbb{E}[Y]\big\|_{\Sigma^{-1}}^2 \geqslant \frac{\mathrm{tr}(\Sigma^{-1}\Sigma_Y)}{2n}.$$

On online to batch conversion. The minimax cumulative regret with respect to the full Gaussian family $\mathcal{F}$ is infinite (see, e.g., (Grünwald, 2007, Example 11.1, p. 298)): this comes from the fact that 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 model $\{\mathcal{N}(\theta, \Sigma) : \theta \in K\}$ for some compact set $K \subset \mathbb{R}^d$ (Rissanen, 1996), in which case minimax regret scales as $d(\log n)/2 + C_K + o(1)$ (for a 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 exhibits an extra $\log n$ factor.

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

$$\inf_{\widehat{g}_n} \sup_{P} \mathbb{E}_{Y_1,\ldots,Y_n \sim P^{\otimes n}}\big[\mathcal{E}(\widehat{g}_n)\big] \tag{19}$$

where the supremum is over all probability distributions P on $\mathbb{R}^d$ with $\mathbb{E}\|Y\|^2 < +\infty$, 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 estimator, as shown below.

Theorem 6 *For the Gaussian location model, the minimax excess risk* (19) *in the misspecified case* (*namely, over all distributions with finite second moment*) *is equal to*

$$\inf_{\widehat{g}_n} \sup_P \mathbb{E}_{Y_1,\ldots,Y_n \sim P^{\otimes n}} \big[\mathcal{E}(\widehat{g}_n) \big] = \frac{d}{2} \log \left(1 + \frac{1}{n} \right).$$

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

Theorem 6 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 (Ng, 1980; Murray, 1977), see also George et al. (2006). Remarkably, both the minimax excess risk and the minimax estimator 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 alone, which do not take into account the latter effect (and degrade under model misspecification when the true variance is large), fail to accurately characterize the excess risk of predictive posteriors under model misspecification.

Finally, Theorem 6 shows that the worst-case excess risk bound (18) of 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, notation 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 parameter.

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

$$\mathcal{F} = \big\{ f_\theta(\cdot|x) = \mathcal{N}(\langle \theta, x \rangle, \sigma^2) : \theta \in \mathbb{R}^d \big\} \tag{20}$$

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 log-loss with respect to the base measure $\mu = (2\pi)^{-1/2}\mathrm{d}y$ on $\mathbb{R}$, so that for $\theta \in \mathbb{R}^d$ and $(x,y) \in \mathbb{R}^d \times \mathbb{R}$:

$$\ell(f_\theta, (x,y)) = -\log f_\theta(y|x) = \frac{1}{2}(y - \langle \theta, x \rangle)^2, \tag{21}$$

and hence the risk of f_θ writes

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

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 density estimation one seeks a *probabilistic prediction* of y given x, or equivalently an estimate of the *conditional distribution* of Y given X, which includes a quantification of the uncertainty of Y given X.

- When one restricts to proper, within-model estimators (taking values in $\mathcal{F}$), the two problems are equivalent, as shown by the expression of the loss (21).

- 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) *Both X and Y are square integrable, namely*

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

We denote $\Sigma = \Sigma_X = \mathbb{E}[XX^\top]$ the second moment matrix, which we 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 $\mathbb{E}[\langle\theta, X\rangle^2] = \langle\Sigma\theta, \theta\rangle$, so that the risk $R(f_\theta)$ is finite[3] and equals:

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

with gradient $\nabla R(f_\theta) = \Sigma\theta - \mathbb{E}[YX]$. In particular, whenever Σ is invertible, the population risk minimizer $f^* \in \mathcal{F}$ is given by $f^* = f_{\theta^*}$ with $\theta^* = \Sigma^{-1}\mathbb{E}[YX]$, while the excess risk of

---

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

$f_\theta \in \mathcal{F}$ writes $\mathcal{E}(f_\theta) = \frac{1}{2} \|\theta - \theta^*\|_\Sigma^2$. Likewise, whenever the empirical covariance matrix

$$\widehat{\Sigma}_n := \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \tag{22}$$

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

$$\widehat{\theta}_n = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n (Y_i - \langle \theta, X_i \rangle)^2 = \widehat{\Sigma}_n^{-1} \widehat{S}_n \tag{23}$$

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 (23).

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 (20). This setting is relevant when $n \gg d$, especially when little is known or assumed on the optimal parameter θ^*. 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.*

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 statements and notation. In addition, a simple induction (Mourtada, 2019) shows that Assumption 2 amounts 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 saying that MLE in the model (20) is uniquely determined almost surely.

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

$$\widetilde{\Sigma}_n = \Sigma^{-1/2} \widehat{\Sigma}_n \Sigma^{-1/2} = \frac{1}{n} \sum_{i=1}^n \widetilde{X}_i \widetilde{X}_i^\top \quad \text{where} \quad \widetilde{X}_i = \Sigma^{-1/2} X_i. \tag{24}$$

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 7 below, the excess risk of SMP is connected to the fluctuations of $\widetilde{\Sigma}_n$.

Theorem 7 *Assume that Assumptions 1 and 2 are fulfilled. For the Gaussian linear family $\mathcal{F}$ given by (20), SMP is given by*

$$\widetilde{f}_n(\cdot|x) = \mathcal{N}\Big(\langle \widehat{\theta}_n, x \rangle, \big(1 + \langle (n\widehat{\Sigma}_n)^{-1} x, x \rangle\big)^2\Big). \tag{25}$$

In addition, it satisfies the following excess risk bound:

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \mathbb{E}\Big[-\log\Big(1 - \langle (n\widehat{\Sigma}_n + XX^\top)^{-1} X, X \rangle\Big)\Big] \leqslant \log\Big(1 + \frac{1}{n} \mathbb{E}\big[\mathrm{tr}(\widetilde{\Sigma}_n^{-1})\big]\Big), \tag{26}$$

where $\widetilde{\Sigma}_n$ is the rescaled empirical covariance given by (24).

17

The proof of Theorem 7 is given in Section 7.3 below. The upper bound on the excess risk depends on the distribution of the design through the term $\mathbb{E}[\text{tr}(\widetilde{\Sigma}_n^{-1})]$, namely through lower relative fluctuations of the empirical covariance matrix $\widehat{\Sigma}_n$ with respect to its population counterpart Σ. Note that this quantity is invariant under linear transformation of $X, X_1, \ldots, X_n$.

A key feature of the excess risk bound (26) 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 MLE and proper estimators. As explained above, MLE is given by $f_{\widehat{\theta}_n}$, where $\widehat{\theta}_n$ is the ordinary least-squares estimator (23). In the *well-specified case*, the minimax risk among *proper* estimators is achieved by MLE and equals $\mathbb{E}[\text{tr}(\widetilde{\Sigma}_n^{-1})]/(2n)$ (Mourtada, 2019); hence, the excess risk of SMP is only within a factor 2 of the 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 MLE scales as $\mathbb{E}_{(X,Y) \sim P}[(Y - \langle \theta^*, X \rangle)^2 \|\Sigma^{-1/2} X\|^2]/n$ up to lower-order terms, and this dependence is unavoidable for any proper estimator (Mourtada, 2019). This means that the risk of proper estimators deteriorates under misspecification, and that the minimax risk among proper estimators is infinite, since the previous quantity can be arbitrarily large.

Comparison with the well-specified case. One can in fact show that the first bound in (26) on the risk of SMP in the general misspecified case is exactly *twice* the minimax excess risk in the well-specified case, for any distribution P_X of covariates. This shows that the general excess risk bound for SMP is intrinsic to the complexity of the problem in this case. Another consequence worth pointing is 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 regret after the first observation is unbounded. Hence, one cannot 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 (26) has a natural interpretation. Indeed, the quantity $\langle (n\widehat{\Sigma}_n + XX^\top)^{-1} X, X \rangle$ is the *leverage score* of X in the sample $X_1, \ldots, X_n, X$. This means that the excess risk of SMP can be upper bounded as

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

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 (26) 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 $\mathbb{E}[\mathrm{tr}(\widetilde{\Sigma}_n^{-1})] = nd/(n - d - 1)$ (Anderson, 2003; Breiman and Freedman, 1983), giving an upper bound of $\log(1 + d/(n - d - 1))$ for SMP.

We now discuss extensions to 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 (26) 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 conditions on the distribution of X, in order to ensure even finiteness of $\mathbb{E}[\mathrm{tr}(\widetilde{\Sigma}_n^{-1})]$:

Assumption 3 (Small ball) *There exist constants $C \geqslant 1$ and $\alpha \in (0, 1)$ such that, for any hyperplane $H \subset \mathbb{R}^d$ and $t > 0$,*

$$\mathbb{P}(\mathrm{dist}(\Sigma^{-1/2}X, H) \leqslant t) \leqslant (Ct)^\alpha. \tag{27}$$

Assumption 3 quantifies Assumption 2, which states that $\mathbb{P}(X \in H) = 0$ for any hyperplane $H \subset \mathbb{R}^d$. It is equivalent to $\mathbb{P}(|\langle \theta, X \rangle| \leqslant t\|\theta\|_\Sigma) \leqslant (Ct)^\alpha$ for every $\theta \in \mathbb{R}^d$ and $t \in (0, 1)$. This condition is a strengthened version of the *small-ball condition* considered in Koltchinskii and Mendelson (2015); Mendelson (2015); Lecué and Mendelson (2016), which amounts to requiring this for a single $t < C^{-1}$. A matching lower bound to (27) holds with $\alpha = 1$ and $C = 0.025$ for any distribution of X when $d \geqslant 2$ (Mourtada, 2019).

Assumption 4 (Kurtosis) $\mathbb{E}\|\Sigma^{-1/2}X\|^4 \leqslant \kappa d^2$ *for some $\kappa \geqslant 1$.*

Assumption 4 is a bound on the kurtosis of $\|\Sigma^{-1/2}X\|$, since $\mathbb{E}\|\Sigma^{-1/2}X\|^2 = d$. It is weaker than the following L^2–L^4 equivalence for one-dimensional 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 $\theta \in \mathbb{R}^d$ (Oliveira, 2016), and a significantly weaker requirement on X than a sub-Gaussian assumption (Vershynin, 2012, Section 3.4).

Corollary 8 *Suppose that Assumptions 1, 2, 3 and 4 hold, and let $\widetilde{f}_n$ be the SMP given by (25). Then, denoting $C' = 28C^4 e^{1 + 9/\alpha}$, for $n \geqslant \min(6d/\alpha, 12\log(12/\alpha)/\alpha)$ we have*

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \frac{d}{n}\Big(1 + C'\frac{\kappa d}{n}\Big). \tag{28}$$

The proof of Corollary 8 is given in Section 7. It is a direct consequence of Theorem 7, together with an upper bound from Mourtada (2019) on the excess risk of the ordinary least-squares estimator in the well-specified case. The bound (28) deduced from Theorem 7 scales as $d/n + O((d/n)^2)$ as $d = o(n)$, with exact first-order constant and order-optimal second-order term $O((d/n)^2)$. The most technical argument is provided in Mourtada (2019), where a tight control on the smallest eigenvalue of $\widetilde{\Sigma}_n$ and on $\mathbb{E}[\mathrm{tr}(\widetilde{\Sigma}_n^{-1})]$ is obtained under Assumptions 3 and 4.

4.3 Ridge-regularized SMP

In the previous section, we considered uniform excess risk bounds with respect to the full Gaussian linear model $\mathcal{F}$. We now turn to non-uniform bounds over $\mathcal{F}$, where some dependence on the comparison parameter $\theta \in \mathbb{R}^d$ is allowed. Such guarantees are relevant

when uniform bounds over $\mathcal{F}$ are not possible, which occurs either when $d > n$, or when the distribution of covariates X does not satisfy the regularity condition (Assumption 2 or 3) ensuring finite minimax risk.

Specifically, we investigate excess risk bounds with respect to balls of the form $\mathcal{F}_B = \{f_\theta : \|\theta\| \leqslant B\}$ for some $B > 0$. For this purpose, we will consider SMP with Ridge regularization $\phi(\theta) = \lambda \|\theta\|^2/2$ for some $\lambda > 0$. One advantage of the bounds obtained in this setting is that they remain meaningful in the *nonparametric* setting where d may be larger than n.

The upper bound from Theorem 9 below does not explicitly depend on the dimension d, but only on the covariance matrix Σ and on $\|\theta\|$. It extends readily to the case where $\mathbb{R}^d$ is replaced by a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$, but we keep $\mathbb{R}^d$ in order to keep the setting and notation consistent with those of Section 4.2. We work in this section under the following assumption.

Assumption 5 (Bounded covariates) $\|X\| \leqslant R$ *almost surely for some constant $R > 0$.*

Assumption 5 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: $K \leqslant R^2$.

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

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

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

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

We also introduce the *degrees of freedom* of the Ridge estimator (Wahba, 1990; Friedman et al., 2001; Wasserman, 2006), given by

$$\mathsf{df}_\lambda(\Sigma) = \mathrm{tr}[(\Sigma + \lambda I_d)^{-1} \Sigma], \tag{29}$$

and note that

$$\mathsf{df}_\lambda(\Sigma) \leqslant \mathrm{tr}[(\Sigma + \lambda I_d)^{-1}(\Sigma + \lambda I_d)] = d. \tag{30}$$

Theorem 9 *Let $\lambda > 0$. The penalized SMP (14) with penalty $\phi(\theta) = \frac{\lambda}{2}\|\theta\|^2$ is well-defined and writes $\widetilde{f}_{\lambda,n}(\cdot|x) = \mathcal{N}(\widetilde{\mu}_\lambda(x), \widetilde{\sigma}_\lambda^2(x))$, where*

$$\widetilde{\sigma}_\lambda(x)^2 = \left((1 - \|x\|_{\widehat{K}_\lambda^x}^2)^2 + \lambda \|x\|_{(\widehat{K}_\lambda^x)^2}^2 \right)^{-1} \tag{31}$$

and

$$\widetilde{\mu}_\lambda(x) = \langle \widehat{\theta}_{\lambda',n}, x \rangle - \lambda \widetilde{\sigma}_\lambda(x)^2 \langle \widehat{\theta}_{\lambda',n}, x \rangle_{\widehat{K}_\lambda^x}. \tag{32}$$

In addition, under Assumptions 1 and 5, we have

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\theta \in \mathbb{R}^d} \left\{ R(f_\theta) + \frac{\lambda}{2}\|\theta\|^2 \right\} \leqslant 1.25 \cdot \frac{\mathsf{df}_\lambda(\Sigma)}{n+1} \tag{33}$$

for every $\lambda \geqslant 2R^2/(n+1)$, where $\mathsf{df}_\lambda(\Sigma)$ is given by (29).

Although the space of parameters is finite-dimensional (of dimension d), the bound (33) 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 9 remains valid (with the same proof, up to minor changes in terminology and notation) in the case of an infinite-dimensional RKHS. Its proof, which is provided in Section 7.3, relies on a combination of the general SMP excess risk bound, exchangeability, and convexity of matrix functions.

Let us now discuss some consequences of Theorem 9.

- *Finite-dimensional case.* Since $\mathsf{df}_\lambda(\Sigma) \leqslant d$ (see (30)), Theorem 9 entails, for $\lambda = 2R^2/(n+1)$, that

$$\mathbb{E}[R(\widetilde{f}_{\lambda,n})] - \inf_{\|\theta\| \leqslant B} R(f_\theta) \leqslant \frac{1.25d + B^2 R^2}{n+1} \tag{34}$$

  for every $B > 0$. This gives an excess risk bound of $O((d + B^2 R^2)/n)$. Proposition 10 below further refines this finite-dimensional bound.

- *Slow, dimension-free rate.* Since $\mathsf{df}_\lambda(\Sigma) \leqslant \mathrm{tr}(\Sigma)/\lambda \leqslant R^2/\lambda$ for $\lambda > 0$, Theorem 9 yields, for every $\lambda \geqslant 2R^2/(n+1)$ and $B > 0$,

$$\mathbb{E}[R(\widetilde{f}_{\lambda,n})] - \inf_{\|\theta\| \leqslant B} R(f_\theta) \leqslant \frac{1.25R^2}{\lambda(n+1)} + \frac{\lambda B^2}{2} \leqslant \frac{2BR}{\sqrt{n}} + \frac{B^2 R^2}{n}, \tag{35}$$

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

- *Nonparametric case.* More precise results can be obtained in terms of spectral properties of Σ. Let $b > 1$ be a rate of decay of the eigenvalues of Σ, such that $\mathsf{df}_\lambda(\Sigma) \leqslant C\lambda^{-1/b}$. Then, Theorem 9 yields

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\|\theta\| \leqslant B} R(f_\theta) \lesssim \frac{C\lambda^{-1/b}}{n} + \lambda B^2 \asymp C^{b/(b+1)} B^{2/(b+1)} n^{-b/(b+1)} \tag{36}$$

  for $\lambda \asymp (B^2 n/C)^{-b/(b+1)}$. This matches the minimax rate for regression with unit noise over balls of RKHSs in the well-specified case, without additional assumptions on θ (Caponnetto and De Vito, 2007).

21

In the finite-dimensional case where $n \gg d$, one can improve the quadratic dependence on the norm $B = \|\theta\|$. 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 θ are no longer achievable.

Proposition 10 *Grant Assumptions 1 and 5. Then, for any $B > 0$, the Ridge-SMP $\widetilde{f}_{\lambda,n}$ of Theorem 9 with $\lambda = d/(B^2(n+1))$ satisfies*

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\theta \in \mathbb{R}^d \,:\, \|\theta\| \leqslant B} R(f_\theta) \leqslant \frac{5d \log \big(2 + BR/\sqrt{d}\big)}{n+1} \,. \tag{37}$$

This bound is of order $O(d \log(BR/\sqrt{d})/n)$. It improves a guarantee obtained by Kakade and Ng (2005) (with suitable parameters and online-to-batch conversion) of $O(d \log(B^2 R^2 n/d)/n)$ from the sequential setting through Bayesian mixture strategies, by removing an extra $O(\log n)$ term (the bound from Kakade and Ng, 2005 holds in the online setting, where the $\log n$ term is unavoidable). In addition, the bound from Theorem 9 is dimension-free, whereas the guarantee from Kakade and Ng (2005) has explicit dependence on d.

Remark 11 (Parameter scaling) *The previous results are valid for arbitrary parameters BR, d, n. In order to make these bounds more concrete, we now discuss some natural scaling for the norm BR. Consider the finite-dimensional case where $n \gg d$, and assume that Σ is well-conditioned, in the sense that $c := \|\Sigma\|_{\mathrm{op}} \cdot \|\Sigma^{-1}\|_{\mathrm{op}} = O(1)$. This means that X is approximately isotropic, or equivalently that the chosen norm on $\mathbb{R}^d$ does not favor specific directions, but rather controls signal strength $\|\theta\|_\Sigma \asymp \|\theta\|$; this can be ensured in practice by rescaling covariates. Also, assume that $\|\Sigma^{-1/2} X\| \leqslant \rho \sqrt{d}$ for some $\rho \geqslant 1$, a bounded leverage condition (Hsu et al., 2014), and let $\psi := \|\theta\|_\Sigma = \mathbb{E}[\langle \theta, X \rangle^2]^{1/2}$ denote signal strength. Then,*

$$\|\theta\| \cdot \|X\| \leqslant \|\Sigma^{-1/2}\|_{\mathrm{op}} \cdot \|\Sigma^{1/2}\theta\| \cdot \|\Sigma^{1/2}\|_{\mathrm{op}} \cdot \|\Sigma^{-1/2}X\| \leqslant c^{1/2} \rho \psi \sqrt{d} \,,$$

so that $BR \leqslant c^{1/2} \rho \psi \sqrt{d} = O(\sqrt{d})$.

On the other hand, one can have $BR \ll \sqrt{d}$: this occurs in the "nonparametric" case where Σ has eigenvalue decay, and θ lies close to the space spanned by the leading eigenvectors of Σ; in this case, $\mathsf{df}_\lambda(\Sigma) \ll d$, and it is beneficial to replace d by $\mathsf{df}_\lambda(\Sigma)$ as in Theorem 9.

We close this section by pointing out that, in the well-conditioned finite-dimensional regime where $BR = O(\sqrt{d})$, the bounds (34) and (37) both yield a $O(d/n)$ guarantee, while the latter has an improved dependence on signal strength.

5. Logistic regression

In this section, we consider conditional density estimation with a binary response, using the logistic model. Section 5.1 introduces the setting. We consider the unpenalized SMP ($\phi \equiv 0$) in Section 5.2 and contrast its predictions with those of MLE. In Section 5.3 we introduce the Logistic SMP procedure with Ridge penalization, and establish a non-asymptotic bound on its excess risk.

5.1 Setting

We consider binary labels in $\mathcal{Y} = \{-1, 1\}$, with counting measure $\mu = \delta_0 + \delta_1$, while $\mathcal{X} = \mathbb{R}^d$. The *logistic model* is the family of conditional distributions given by

$$\mathcal{F} = \{f_\theta : \theta \in \mathbb{R}^d\}, \quad \text{where} \quad f_\theta(1|x) := 1 - f_\theta(-1|x) = \sigma(\langle \theta, x \rangle) \tag{38}$$

for any $x \in \mathbb{R}^d$, with $\sigma(u) = e^u/(1 + e^u)$ for $u \in \mathbb{R}$ the *sigmoid* function. Since $\sigma(-u) = 1 - \sigma(u)$, one simply has $f_\theta(y|x) = \sigma(y\langle \theta, x \rangle)$ for $x \in \mathbb{R}^d$ and $y \in \{-1, 1\}$. The log-loss of $f_\theta \in \mathcal{F}$ at a sample $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ writes

$$\ell(f_\theta, (x, y)) = -\log f_\theta(y|x) = \log(1 + e^{-y\langle \theta, x \rangle}) = \ell(-y\langle \theta, x \rangle), \tag{39}$$

where we introduced the *logistic* loss $\ell(u) = \log(1 + e^u)$ for $u \in \mathbb{R}$. Let (X, Y) have distribution P on $\mathbb{R}^d \times \{-1, 1\}$, such that $\mathbb{E}\|X\| < +\infty$. Since $\ell'(u) = \sigma(u) \in [0, 1]$ for any $u \in \mathbb{R}$, we have $0 \leqslant \ell(u) \leqslant \log 2 + |u|$ so that $\ell(-Y\langle \theta, X \rangle) \leqslant \log 2 + \|\theta\|\|X\|$, and the risk of f_θ, namely

$$R(f_\theta) = \mathbb{E}[\ell(-Y\langle \theta, X \rangle)], \tag{40}$$

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

$$\widehat{\theta}_n \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(-Y_i\langle \theta, X_i \rangle), \tag{41}$$

A MLE (41) does not always exist, and may not be unique. Indeed, it is a well-known fact (see Candès and Sur, 2020 for recent results on this topic in the high-dimensional regime) that there is no MLE (41) 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 $\theta \in \mathbb{R}^d$ such that $Y_i\langle \theta, X_i \rangle > 0$ for all $i = 1, \ldots, n$ (indeed, in this case the empirical risk of $t\theta$ converges to 0 as $t \to +\infty$, while the empirical risk is positive on $\mathbb{R}^d$). In addition, when a MLE exists in $\mathbb{R}^d$, one can see that it is unique if and only if $V = \text{span}(X_1, \ldots, X_n) = \mathbb{R}^d$: in this case, the empirical risk is strictly convex on $\mathbb{R}^d$ since $\ell : \mathbb{R} \to \mathbb{R}$ is.

It is convenient to enrich the class $\mathcal{F}$ given by (38) to ensure existence (though not uniqueness) of MLE in the separated case. Specifically, define the model $\overline{\mathcal{F}}$ obtained by adding to $\mathcal{F}$ the conditional densities $f_{\infty,\theta}$ for $\theta \in \mathbb{R}^d$, $\|\theta\| = 1$, defined by $f_{\infty,\theta}(1|x) = 1$ if $\langle \theta, x \rangle > 0$, 0 if $\langle \theta, x \rangle < 0$ and $1/2$ if $\langle \theta, x \rangle = 0$. Denote by $\overline{\Theta}$ the parameter space obtained by adding to $\mathbb{R}^d$ elements of the form (∞, θ). We note that MLE exists in $\overline{\mathcal{F}}$ in the separated case, although it is not unique since it depends on the choice of a separating hyperplane defined by θ. Given a choice of MLE, we let

$$\widehat{\theta}_n^{(x,y)} = \arg\min_{\theta \in \overline{\Theta}} \left\{ \sum_{i=1}^{n} \ell(f_\theta, (X_i, Y_i)) + \ell(f_\theta, (x, y)) \right\} \tag{42}$$

for any $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$. It is also convenient to let $Z_i = -Y_i X_i$; then, one has $\widehat{\theta}_n^{(x,y)} = \widehat{\theta}_n^{-yx}$, where for $z \in \mathbb{R}^d$ we define (with a slight abuse of notation for $\theta \in \overline{\Theta} \setminus \mathbb{R}^d$)

$$\widehat{\theta}_n^z = \arg\min_{\theta \in \overline{\Theta}} \left\{ \sum_{i=1}^{n} \ell(\langle \theta, Z_i \rangle) + \ell(\langle \theta, z \rangle) \right\}. \tag{43}$$

23

5.2 SMP for logistic regression

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

Proposition 12 *For the family of logistic conditional distributions* (38), *SMP writes*

$$\widetilde{f}_n(y|x) = \frac{f_{\widehat{\theta}_n^{(x,y)}}(y|x)}{f_{\widehat{\theta}_n^{(x,1)}}(1|x) + f_{\widehat{\theta}_n^{(x,-1)}}(-1|x)} = \frac{\sigma(\langle \widehat{\theta}_n^{(x,y)}, yx \rangle)}{\sigma(\langle \widehat{\theta}_n^{(x,1)}, x \rangle) + \sigma(\langle \widehat{\theta}_n^{(x,-1)}, -x \rangle)} \tag{44}$$

for every $x \in \mathbb{R}^d$ and $y \in \{-1, 1\}$. Unlike the MLE (42), *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:*

$$\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \mathbb{E}_{Z_1^n, Z}\big[\sigma(\langle \widehat{\theta}_n^{-Z}, Z \rangle) - \sigma(\langle \widehat{\theta}_n^{Z}, Z \rangle)\big], \tag{45}$$

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

The proof of Proposition 12 is given in Section 7.4 below. Unlike MLE, SMP is always well-defined and outputs predictions in $(0, 1)$. Indeed, the numerator in (44) belongs to $(0, 1]$, and whenever the points $Y_1 X_1, \ldots, Y_n X_n, yx$ belong to a half-space passing through the origin (so that MLE does not exist in $\mathbb{R}^d$), we have $f_{\widehat{\theta}_n^{(x,y)}}(y|x) = 1$, so that the prediction of SMP is well-defined and does not depend on the choice of MLE in (42), see the proof of Proposition 12 for details.

Comparison with MLE. SMP corrects a well-known deficiency of MLE, which tends to produce overly confident and ill-calibrated predictions (Sur and Candès, 2019). To emphasize this effect, consider the case of a point x for which the virtual datasets $(X_1, Y_1), \ldots, (X_n, Y_n), (x, y)$ are separated for both $y = -1$ and $y = 1$. Then, the prediction $\widehat{f}_n(1|x)$ of an MLE $\widehat{f}_n \in \overline{\mathcal{F}}$ can be either 1 or 0, both being possible depending on the specific choice of separating hyperplane. Hence, in this case the prediction of MLE is both highly confident and dependent on an arbitrary choice. By contrast, in this situation SMP gives equal probability $1/2$ to both classes, reflecting the uncertainty for such points x.

A non-Bayesian approach to calibration. As for the Gaussian linear model (Section 4), SMP returns more uncertain conditional distributions for input points x with high "leverage", namely strong influence on the prediction of MLE at this point. This provides a simple and natural approach to calibration of probabilistic predictions for logistic regression, which does not rely on Bayesian methods. Such an approach is appealing on computational grounds, since the prediction $\widetilde{f}(\cdot|x)$ of SMP is obtained by solving two logistic regressions (42) and does not require approximate posterior sampling.

Comparison with stability approaches. Approaches based on stability of the loss (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010; Srebro et al., 2010; Koren and Levy, 2015) would lead to a control of the excess risk involving $\ell(\langle \widehat{\theta}_n^{-Z}, Z \rangle) - \ell(\langle \widehat{\theta}_n^{Z}, Z \rangle)$, while Proposition 12 involves $\sigma(\langle \widehat{\theta}_n^{-Z}, Z \rangle) - \sigma(\langle \widehat{\theta}_n^{Z}, Z \rangle)$, 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 $\ell(u') - \ell(u) \approx \ell'(u) \cdot (u' - u) \approx u' - u$, 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 roughly explains why we are able to remove terms of order e^{BR} from our upper bound on the excess risk of SMP, provided in the next section.

5.3 Excess risk bounds for 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 $\phi(\theta) = \lambda\|\theta\|^2/2$. The corresponding penalized SMP can be computed as follows: for every $z \in \mathbb{R}^d$, let

$$\widehat{\theta}^z_{\lambda,n} := \underset{\theta \in \mathbb{R}^d}{\arg\min}\left\{\frac{1}{n+1}\Big(\sum_{i=1}^n \ell(\langle\theta, Z_i\rangle) + \ell(\langle\theta, z\rangle)\Big) + \frac{\lambda}{2}\|\theta\|^2\right\}. \tag{46}$$

Note that $\widehat{\theta}^z_{\lambda,n} \in \mathbb{R}^d$ exists and is unique, since the regularized objective in (46) is strongly convex, hence strictly convex and diverging as $\|\theta\| \to +\infty$. As before, we let $\widehat{\theta}^{(x,y)}_{\lambda,n} = \widehat{\theta}^{-yx}_{\lambda,n}$ for $(x,y) \in \mathbb{R}^d \times \{-1,1\}$. Now, following Theorem 2, the regularized SMP writes in this case

$$\widetilde{f}_{\lambda,n}(y|x) = \frac{\sigma(y\langle\widehat{\theta}^{(x,y)}_{\lambda,n}, x\rangle)\,e^{-\lambda\|\widehat{\theta}^{(x,y)}_{\lambda,n}\|^2/2}}{\sigma(\langle\widehat{\theta}^{(x,1)}_{\lambda,n}, x\rangle)\,e^{-\lambda\|\widehat{\theta}^{(x,1)}_{\lambda,n}\|^2/2} + \sigma(-\langle\widehat{\theta}^{(x,-1)}_{\lambda,n}, x\rangle)\,e^{-\lambda\|\widehat{\theta}^{(x,-1)}_{\lambda,n}\|^2/2}} \tag{47}$$

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 5, namely $\|X\| \leqslant R$ almost surely, as in Section 4.3 for the Gaussian linear model. Our main guarantee for Ridge-regularized SMP is stated in a nonparametric setting, where dependence on the dimension d is kept implicit through the degrees of freedom (29).

Theorem 13 *Grant Assumption 5, and assume that $\lambda \geqslant 2R^2/(n+1)$. Then, the Ridge-regularized logistic SMP given by* (47) *satisfies*

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\theta \in \mathbb{R}^d}\left\{R(f_\theta) + \frac{\lambda}{2}\|\theta\|^2\right\} \leqslant e \cdot \frac{\mathsf{df}_{4\lambda}(\Sigma)}{n}, \tag{48}$$

where we recall that $\mathsf{df}_\lambda(\Sigma) = \mathrm{tr}[(\Sigma + \lambda I)^{-1}\Sigma]$.

The upper bound (48) is a *fast rate* excess risk guarantee; it is worth noting that it only requires bounded covariates (Assumption 5). 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 (48) is valid under no assumption on the distribution of (X,Y).

We note that Marteau-Ferey et al. (2019b) established nonparametric fast rate guarantees akin to (48) for the Ridge-regularized estimator in the *well-specified* case. Compared to (48), their bias term, while also equal to λB^2 under the sole assumption $\|\theta\| \leqslant B$, can be further improved under stronger assumptions on θ (namely, faster coefficient decay, or *source condition* Caponnetto and De Vito, 2007). On the other hand, this result relies on the assumption of a well-specified model, and under our general assumptions such rates would exhibit exponential dependence in BR (Hazan et al., 2014).

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

Corollary 14 *Under Assumption 5, the Ridge-regularized logistic SMP $\widetilde{f}_{\lambda,n}$ (47) with $\lambda = 2R^2/(n+1)$ satisfies, for every $B > 0$,*

$$\mathbb{E}[R(\widetilde{f}_{\lambda,n})] - \inf_{\|\theta\| \leqslant B} R(f_\theta) \leqslant \frac{e \cdot d + B^2 R^2}{n}. \qquad (49)$$

Note that under the well-conditioned scaling of dimension d with constant signal strength, namely $BR = O(\sqrt{d})$ (see Remark 11 from Section 4.3), Corollary 14 yields an excess risk of $O(d/n)$.

Bypassing a lower bound. Under Assumption 5, Corollary 14 leads to an upper bound for Ridge SMP of $O((d + B^2R^2)/n)$ with respect to the ball $\|\theta\| \leqslant B$. By contrast, Hazan et al. (2014) showed a lower bound for any *proper* estimator (including the norm-constrained or Ridge-penalized MLE, or any stochastic optimization procedure) of order $\min(BR/\sqrt{n}, de^{BR}/n)$ in the worst case. We note that SMP is an improper estimator, as the log-odds ratio $\log(\widetilde{f}_{\lambda,n}(1|x)/\widetilde{f}_{\lambda,n}(-1|x))$ is nonlinear in x, and that it bypasses the lower bound for proper estimators.

A practical improper estimator. Fast rates of order $O(d\log(BRn)/n)$ are obtained by Kakade and Ng (2005); Foster et al. (2018) under Assumption 5, by applying online-to-offline conversion (averaging) to a Bayes mixture sequential procedure, with prior on θ uniform over the ball of radius B (Foster et al., 2018) or Gaussian (Kakade and Ng, 2005). This bound has an even better dependence on B (logarithmic instead of quadratic) than Corollary 14, although it also has a slightly worse dependence in n (additional $\log n$ factor); Theorem 13 additionally replaces d by $\mathsf{df}_{4\lambda}(\Sigma)$. The main advantage of 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 Foster et al. (2018), about finding an efficient alternative with fast rate, at least in the batch statistical learning case. We note however that SMP is still more computationally demanding at prediction time than MLE, because of the required updates of the logistic risk minimization problem.

Overview of learning rates for logistic regression. Logistic regression with bounded features $\|X\| \leqslant R$ over the ball $\mathcal{F}_B = \{f_\theta : \|\theta\| \leqslant B\}$ is (when restricting to proper estimators) a convex and R-Lipschitz stochastic optimization problem over a bounded domain. This implies that a *slow rate* of $O(BR/\sqrt{n})$ can be achieved by properly-tuned averaged projected online gradient descent (Robbins and Monro, 1951; Zinkevich, 2003; Shalev-Shwartz, 2012; Bubeck, 2015; Hazan, 2016), Ridge-regularized ERM over $\mathbb{R}^d$ (Bousquet and Elisseeff, 2002; Sridharan et al., 2009), or (as a linear prediction problem) constrained ERM over $\mathcal{F}_B$ (Kakade et al., 2009; Bartlett and Mendelson, 2002; Meir and Zhang, 2003). Under the same assumptions, the logistic loss is e^{-BR}-exp-concave over $\mathcal{F}_B$, implying that a rate of $O(de^{BR}/n)$ can be achieved (up to potential $\log n$ factors) through the (averaged) Exponential Weights (Cover, 1991; Hazan et al., 2007; Vovk, 1998) or Online Newton Step algorithms (Hazan et al., 2007; Mahdavi et al., 2015), as well as ERM over $\mathcal{F}_B$ (Koren and Levy, 2015; Gonen and Shalev-Shwartz, 2018; Mehta, 2017). The improved dependence on n in this bound is typically outweighed by the prohibitive exponential dependence on

parameter norm. As mentioned before, a lower bound of Hazan et al. (2014) shows that, without further assumptions, no proper (within model $\mathcal{F}$) estimator can improve over the $O(\min(BR/\sqrt{n}, de^{BR}/n))$ guarantee. In order to bypass this lower bound, one has to resort to improper procedures (Foster et al., 2018). This is the approach taken by Foster et al. (2018); Kakade and Ng (2005) and ourselves, enabling improved guarantees without further assumptions, as discussed above.

Another line of work (Bach, 2010, 2014; Bach and Moulines, 2013; Ostrovskii and Bach, 2018; Marteau-Ferey et al., 2019b) studies the behavior of specific (within-model) estimators, such as Ridge-regularized MLE or stochastic approximation procedures, in a distribution-dependent fashion. A key technique in these refined analyses is the use of (generalized) *self-concordance* of logistic loss, introduced by Bach (2010), namely a control of the third derivative in terms of the second. Following progress in Bach (2010, 2014), Bach and Moulines (2013) introduces a stochastic approximation algorithm with excess risk $O(\rho^3 d(BR)^4/n)$, where ρ is a distribution-dependent curvature parameter. This bound eliminates dependence on the smallest eigenvalue of the Hessian at the optimum (Bach, 2014), but does not lead to the correct scaling in the finite-dimensional case with $BR = O(\sqrt{d})$, or in the nonparametric setting due to dependence on d instead of $\mathsf{df}_\lambda(\Sigma)$ (see Remark 11). In finite dimension, a tight non-asymptotic guarantee for MLE is obtained by Ostrovskii and Bach (2018), with an excess risk of $O(d_{\text{eff}}/n)$ for $n \gtrsim \max(\rho d_{\text{eff}}, d \log d)$, where d_{eff} denotes the *effective dimension* characterizing the asymptotic risk of MLE (3). These results are extended in Marteau-Ferey et al. (2019b) in the well-specified nonparametric setting, with sharp risk bounds for the Ridge-regularized MLE. In the worst case, the distribution-dependent constants ρ and d_{eff} scale with e^{BR} (Bach and Moulines, 2013), although they can be much smaller for more favorable distributions. Despite the difference in assumptions, from a technical point of view, our analysis of the bound on the SMP excess risk also uses self-concordance.

In addition to these non-asymptotic analyses, a recent line of work (Sur and Candès, 2019; Barbier et al., 2019; Candès and Sur, 2020; Salehi et al., 2019) studies logistic regression under high-dimensional asymptotics where $d \asymp n$. This asymptotic approach differs from the non-asymptotic one in that it provides an exact characterization of the error, but under highly specific distributional assumptions (well-specified model and Gaussian or jointly independent features).

6. Conclusion

In this paper, we derive a general excess risk bound for predictive density estimation under logarithmic loss. Minimizing this bound 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 with estimators taking values within the model, whose performance typically degrade under misspecification. This estimator provides an alternative to approaches based on online-to-offline conversion (Barron, 1987; Catoni, 2004; Cesa-Bianchi et al., 2004; Audibert, 2009) of sequential procedures, whose rates feature an additional logarithmic dependence on the sample size, and may be infinite for unbounded models.

We apply SMP to the Gaussian linear model. In this case, it 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 and finite-dimensional fast rates.

We then consider logistic regression. Here, (Ridge-penalized) SMP is a simple and 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 (Hazan et al., 2014). In the statistical learning setting, this provides a practical alternative to the improper estimator from Foster et al. (2018), which relies on Bayesian mixtures. This work leaves a number of open questions 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 (Catoni, 2004; Audibert, 2008, 2009; Foster et al., 2018). In particular, the high-probability bound stated by Foster et al. (2018) for a procedure based on a "confidence boosting" technique (Mehta, 2017) 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. Designing efficient procedures that achieve high (exponential) probability excess risk bounds for misspecified logistic regression is an interesting direction for future work.

- Second, it could be interesting to adapt the proposed method to online logistic regression, with a regret bound for individual sequences. Since the initial release of this work, Jézéquel et al. (2020) proposed a related (though distinct) practical sequential algorithm also relying on virtual samples, with a per-round regret of $O(dBR\log(BRn)/n)$. In finite dimension with $BR = O(\sqrt{d})$, this implies a $\widetilde{O}(d\sqrt{d}/n)$ bound up to logarithmic terms, leaving room for further improvement in the online setting.

- Another possibility is to apply SMP to other (conditional or otherwise) models beyond the Gaussian linear and logistic ones considered here, including other generalized linear models (McCullagh and Nelder, 1989).

- Finally, it would be interesting to investigate conditions on the model and prior under which Bayes predictive posteriors (without iterate averaging) achieve uniform non-asymptotic bounds (such as Theorem 6 or our guarantees for SMP).

On a more general note, density estimation under Kullback-Leibler risk possesses 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 logarithmic loss). This has been successfully exploited in the sequential case, where cumulative criteria are considered. Beyond the present work, we expect that further advances are possible in the statistical setting. In particular, the idea of using a notion of leverage to quantify uncertainty for conditional density estimation, as done by SMP, as an alternative to Bayesian posteriors, may have broader applicability beyond the models we consider.

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

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

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

where we denoted

$$
\Delta_n = \inf_{f \in \mathcal{F}} \mathbb{E}\bigg[\frac{1}{n+1}\bigg\{\sum_{i=1}^n \ell_\phi(f, Z_i) + \ell_\phi(f, Z)\bigg\}\bigg] - \mathbb{E}\bigg[\inf_{f \in \mathcal{F}} \frac{1}{n+1}\bigg\{\sum_{i=1}^n \ell_\phi(f, Z_i) + \ell_\phi(f, Z)\bigg\}\bigg] \geqslant 0.
\tag{50}
$$

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

$$
\mathbb{E}\big[\mathcal{E}_\phi(\widehat{g}_n)\big] + \Delta_n = \mathbb{E}_{Z_1^n, Z}[\ell(\widehat{g}_n, Z)] - \frac{1}{n+1}\mathbb{E}\bigg[\sum_{i=1}^n \ell_\phi(\widehat{f}_{\phi,n}^Z, Z_i) + \ell_\phi(\widehat{f}_{\phi,n}^Z, Z)\bigg].
\tag{51}
$$

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 $\mathbb{E}[\ell_\phi(\widehat{f}_{\phi,n}^Z, Z_i)] = \mathbb{E}[\ell_\phi(\widehat{f}_{\phi,n}^Z, Z)]$ for $i = 1, \dots, n$ (recall that $\widehat{f}_{\phi,n}^Z$ is chosen symmetrically in $Z_1, \dots, Z_n, Z$). Hence, (51) becomes

$$
\mathbb{E}\big[\mathcal{E}_\phi(\widehat{g}_n)\big] + \Delta_n = \mathbb{E}_{Z_1^n, Z}\big[\ell(\widehat{g}_n, Z) - \ell_\phi(\widehat{f}_{\phi,n}^Z, Z)\big]
$$

$$
= \mathbb{E}_{Z_1^n, X}\mathbb{E}_{Y|X}\big[\ell(\widehat{g}_n(X), Y) - \ell_\phi(\widehat{f}_{\phi,n}^{(X,Y)}(X), Y)\big]
$$

$$
\leqslant \mathbb{E}_{Z_1^n, X}\bigg[\sup_{y \in \mathcal{Y}}\big\{\ell(\widehat{g}_n(X), y) - \ell_\phi(\widehat{f}_{\phi,n}^{(X,y)}(X), y)\big\}\bigg],
\tag{52}
$$

which implies the bound (10) 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 $x \in \mathcal{X}$:

$$
\sup_{y \in \mathcal{Y}}\big\{\ell(p, y) - \ell_\phi(\widehat{f}_{\phi,n}^{(x,y)}(x), y)\big\} = \sup_{y \in \mathcal{Y}} \log \frac{\widehat{f}_{\phi,n}^{(x,y)}(y|x) e^{-\phi(\widehat{f}_{\phi,n}^{(x,y)})}}{p(y)}.
\tag{53}
$$

Now, Theorem 2 follows from Theorem 1 together with Lemma 15 below, where we consider $g(y) = \widehat{f}_{\phi,n}^{(x,y)}(y|x) e^{-\phi(\widehat{f}_{\phi,n}^{(x,y)})}$. ∎

Lemma 15 Let $g : \mathcal{Y} \to [0, +\infty]$ be a measurable function such that $\int_{\mathcal{Y}} g\,\mathrm{d}\mu \in \mathbb{R}_+^*$. Then,

$$
\inf_p \sup_{y \in \mathcal{Y}} \log \frac{g(y)}{p(y)} = \log\bigg(\int_{\mathcal{Y}} g(y)\mu(\mathrm{d}y)\bigg),
\tag{54}
$$

where the infimum in (54) spans over all probability densities $p : \mathcal{Y} \to \mathbb{R}^+$ with respect to μ, and the infimum is reached at

$$p^* = \frac{g}{\int_{\mathcal{Y}} g \mathrm{d}\mu} \,. \tag{55}$$

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 $C(p) \geqslant \log \left(\int_{\mathcal{Y}} g \mathrm{d}\mu \right)$. Since $C(p^*) = \log \left(\int_{\mathcal{Y}} g \mathrm{d}\mu \right)$, this concludes the proof. ∎

We will sometimes also use the following observation:

Lemma 16 *The expected excess risk of the SMP is equal to*:

$$\mathbb{E}\big[\mathcal{E}_\phi(\widetilde{f}_{\phi,n})\big] = \mathbb{E}_{Z_1^n, X}\Big[\log \Big(\int_{\mathcal{Y}} \widehat{f}_{\phi,n}^{(X,y)}(y|X) e^{-\phi(\widehat{f}_{\phi,n}^{(X,y)})} \mu(\mathrm{d}y) \Big) \Big] - \Delta_n \,, \tag{56}$$

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

$$
\begin{aligned}
\Delta_n &= \frac{1}{n+1} \inf_{f \in \mathcal{F}} \mathbb{E}\Bigg[\sum_{i=1}^{n+1} \ell_\phi(f, Z_i) - \sum_{i=1}^{n+1} \ell_\phi(\widehat{f}_{\phi,n+1}, Z_i) \Bigg] \\
&= \frac{1}{n+1} \mathbb{E}\Bigg[\sum_{i=1}^{n+1} \ell_\phi(f^*, Z_i) - \sum_{i=1}^{n+1} \ell_\phi(\widehat{f}_{\phi,n+1}, Z_i) \Bigg] .
\end{aligned}
\tag{57}
$$

Proof This follows from the fact that inequality (52) is an equality when $\widehat{g}_n = \widetilde{f}_{\phi,n}$ (see Lemma 15). ∎

7.2 Proofs for density estimation (Section 3)

Proof of Proposition 4 Since the MLE $\widehat{f}_n$ writes $\widehat{f}_n(y) = N_n(y)/n$, we have for every $y \in \mathcal{Y}$:

$$\widehat{f}_n^y(y) = \frac{N_n(y) + 1}{n+1} \propto N_n(y) + 1 \,, \tag{58}$$

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

$$\sum_{y \in \mathcal{Y}} \widehat{f}_n^y(y) = \frac{n+d}{n+1} \,. \tag{59}$$

It proves that the SMP $\widetilde{f}_n$ (14) is the Laplace estimator (16) and that the excess risk bound (15) 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 $u \geqslant 0$). ∎

Proof of Proposition 5 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 $\theta, y \in \mathbb{R}^d$, $\ell(f_{\theta,\Sigma}, y) = -\langle \Sigma^{-1}\theta, y \rangle + \frac{1}{2}\|\theta\|_{\Sigma^{-1}}^2$. 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 $\theta = \pm\Sigma e_j$ for $1 \leqslant j \leqslant d$ (where $(e_j)_{1\leqslant j\leqslant d}$ is the canonical basis of $\mathbb{R}^d$), this implies that $\mathbb{E}|Y_j| < +\infty$, 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 (18). 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 $(2\pi)^{-d/2}\mathrm{d}y$:

$$
\begin{aligned}
f_{\widehat{\theta}_n^y}(y) &= (\det \Sigma)^{-1/2} \exp\Big(-\frac{1}{2}\|y - \widehat{\theta}_n^y\|_{\Sigma^{-1}}^2 \Big) \\
&= (\det \Sigma)^{-1/2} \exp\Big(-\frac{1}{2}\Big(\frac{n}{n+1}\Big)^2 \|y - \bar{Y}_n\|_{\Sigma^{-1}}^2 \Big) \\
&= (\det \Sigma)^{-1/2} \det((1 + 1/n)^2\Sigma)^{1/2} f_{\bar{Y}_n,(1+1/n)^2\Sigma}(y) \\
&= \Big(1 + \frac{1}{n}\Big)^d f_{\bar{Y}_n,(1+1/n)^2\Sigma}(y) \,,
\end{aligned}
\tag{60}
$$

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

$$
\int_{\mathbb{R}^d} f_{\widehat{\theta}_n^y}(y)(2\pi)^{-d/2}\mathrm{d}y = \int_{\mathbb{R}^d} \Big(1 + \frac{1}{n}\Big)^d f_{\bar{Y}_n,(1+1/n)^2\Sigma}(y)(2\pi)^{-d/2}\mathrm{d}y = \Big(1 + \frac{1}{n}\Big)^d, \tag{61}
$$

which yields the excess risk bound (18) 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 16, we have

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

where Δ_n is defined as in (50), *i.e.*

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

where we used the fact that $\mathbb{E}[(Y - \theta^*)(Y - \theta^*)^\top] = \Sigma$. It follows that $\mathbb{E}[\mathcal{E}(\widetilde{f}_n)] = d\log(1 + 1/n) - d/(2n) \leqslant d/(2n)$, which completes the proof of Proposition 5. ∎

Proof of Theorem 6 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

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

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

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

so that

$$
\begin{aligned}
\mathcal{E}(f_{\theta,\sigma^2\Sigma}) &= R(f_{\theta,\sigma^2\Sigma}) - R(f_{\theta^*,\Sigma}) \\
&= \frac{d}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\|\theta - \theta^*\|_{\Sigma^{-1}}^2 + \frac{1}{2}\Big(\frac{1}{\sigma^2} - 1\Big)\mathrm{tr}\big(\Sigma^{-1}\Sigma_Y\big). \tag{62}
\end{aligned}
$$

Now, since

$$\mathbb{E}[\|\bar{Y}_n - \theta^*\|_{\Sigma^{-1}}^2] = \mathrm{tr}\Big(\Sigma^{-1}\mathbb{E}[(\bar{Y}_n - \theta^*)(\bar{Y}_n - \theta^*)^\top]\Big) = \frac{\mathrm{tr}(\Sigma^{-1}\Sigma_Y)}{n},$$

equation (62) implies that, for $\sigma^2 = 1 + 1/n$,

$$\mathbb{E}[\mathcal{E}(f_{\bar{Y}_n,\sigma^2\Sigma})] = \frac{d}{2}\log\sigma^2 + \frac{1}{2}\Big[\Big(1 + \frac{1}{n}\Big)\frac{1}{\sigma^2} - 1\Big]\mathrm{tr}(\Sigma^{-1}\Sigma_Y) = \frac{d}{2}\log\Big(1 + \frac{1}{n}\Big). \tag{63}$$

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 (Ng, 1980; Murray, 1977). ∎

7.3 Proofs for the Gaussian linear model (Section 4)

Proof of Theorem 7 Let us first recall that $\mathcal{F} = \{f_\theta(y|x) = \mathcal{N}(\langle\theta, x\rangle, 1) : \theta \in \mathbb{R}^d\}$ 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 $\widehat{\theta}_n = \widehat{\Sigma}_n^{-1}\widehat{S}_n$ and, for every $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$,

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

Hence, we have

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

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

$$
\big(1 - \big\langle (S + vv^\top)^{-1}v, v \big\rangle\big)^{-1} = 1 + \langle S^{-1}v, v \rangle, \tag{64}
$$

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

$$
\frac{\big\langle (S + vv^\top)^{-1}Su, v \big\rangle}{1 - \big\langle (S + vv^\top)^{-1}v, v \big\rangle} = \langle u, v \rangle. \tag{65}
$$

The proof of Lemma 17 is given below. It also follows from the Sherman-Morrison formula. Using (64) 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 $\widehat{S}_n = \widehat{\Sigma}_n \widehat{\theta}_n$ together with (65) for $S = n\widehat{\Sigma}_n$, $v = x$ and $u = \widehat{\theta}_n$ leads to

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

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

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

Note that

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

which shows after normalization (14) that the SMP is given by

$$
\widetilde{f}_n(y|x) = \mathcal{N}(\mu_n(x), \sigma_n^2(x)) \tag{66}
$$

and that its excess risk writes

$$
\mathbb{E}\big[\mathcal{E}(\widetilde{f}_n)\big] \leqslant \mathbb{E}\big[\log \sigma_n(X)\big] = \mathbb{E}\Big[-\log\big(1 - \big\langle (n\widehat{\Sigma}_n + XX^\top)^{-1}X, X \big\rangle\big)\Big]. \tag{67}
$$

This proves the first inequality in (26). Let us prove now the second inequality in (26). Let us recall that the covariance Σ and rescaled design $\widetilde{X}, \widetilde{X}_i$ and rescaled covariance $\widetilde{\Sigma}_n$ are given by (22) and (24). We have

$$
\begin{aligned}
\langle (n\widehat{\Sigma}_n + XX^\top)^{-1}X, X \rangle &= \langle \Sigma^{1/2}(n\widehat{\Sigma}_n + XX^\top)^{-1}\Sigma^{1/2}\Sigma^{-1/2}X, \Sigma^{-1/2}X \rangle \\
&= \langle (n\widetilde{\Sigma}_n + \widetilde{X}\widetilde{X}^\top)^{-1}\widetilde{X}, \widetilde{X} \rangle,
\end{aligned}
\tag{68}
$$

hence, combining (67), (68) and (64), we have

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

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 7. ∎

Proof of Lemma 17 First, (65) clearly holds if $v = 0$. Now, for $u, v \in \mathbb{R}^d$, $v \neq 0$:

$$
\begin{aligned}
\langle (S + vv^\top)^{-1}Su, v \rangle &= \langle (S + vv^\top)^{-1}(S + vv^\top - vv^\top)u, v \rangle \\
&= \langle (I_d - (S + vv^\top)^{-1}vv^\top)u, v \rangle \\
&= \langle u, v \rangle \big(1 - \langle (S + vv^\top)^{-1}v, v \rangle\big).
\end{aligned}
\tag{69}
$$

Letting $u = S^{-1}v$ in (69), the left-hand side is $\langle (S + vv^\top)^{-1}v, v \rangle > 0$ (since $S + vv^\top \succcurlyeq S$ is positive, and $v \neq 0$) so that the right-hand side is positive and thus $1 - \langle (S + vv^\top)^{-1}v, v \rangle > 0$. Dividing both sides of (69) by this quantity establishes (65), which implies (64) by taking $u = S^{-1}v$. ∎

Proof of Theorem 9 and Proposition 10 Let us recall that we consider the family $\mathcal{F} = \{f_\theta(\cdot|x) = \mathcal{N}(\langle \theta, x \rangle, \sigma^2) : \theta \in \mathbb{R}^d\}$, together with the Ridge penalization $\phi(\theta) = \lambda\|\theta\|^2/2$ for some $\lambda > 0$. Let

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

denote the Ridge estimator, where $\widehat{\Sigma}_n$ and $\widehat{S}_n$ are the same as in the proof of Theorem 7. 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

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

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

$$y - \langle \widehat{\theta}_{\lambda,n}^{(x,y)}, x \rangle = y - \langle \widehat{K}_\lambda^x (n\widehat{S}_n + yx), x \rangle = \left(1 - \|x\|_{\widehat{K}_\lambda^x}^2\right) y - \langle n\widehat{S}_n, x \rangle_{\widehat{K}_\lambda^x} \tag{70}$$

and that

$$\lambda \|\widehat{\theta}_{\lambda,n}^{(x,y)}\|^2 = \lambda \|\widehat{K}_\lambda^x (n\widehat{S}_n + yx)\|^2 = \lambda \|n\widehat{S}_n + yx\|_{(\widehat{K}_\lambda^x)^2}^2$$

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

The SMP is given in this setting by

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

where $\mu(dy) = (2\pi)^{-1/2} dy$, see (14), and where

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

Now, the equality (70) gives, after a straightforward computation,

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

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

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

$$\mu_\lambda(x) = \frac{\left(1 - \|x\|_{\widehat{K}_\lambda^x}^2\right)\langle n\widehat{S}_n, x \rangle_{\widehat{K}_\lambda^x} - \lambda \langle n\widehat{S}_n, x \rangle_{(\widehat{K}_\lambda^x)^2}}{\left(1 - \|x\|_{\widehat{K}_\lambda^x}^2\right)^2 + \lambda \|x\|_{(\widehat{K}_\lambda^x)^2}^2}.$$

This entails that the SMP is given by

$$\widetilde{f}_{\lambda,n}(\cdot|x) = \mathcal{N}\left(\mu_\lambda(x), \sigma_\lambda(x)^2\right). \tag{71}$$

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

$$n\widehat{S}_n = \left(n\widehat{\Sigma}_n + \lambda(n+1)I_d\right)\widehat{\theta}_{\lambda',n}$$

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

$$\langle n\widehat{S}_n, x \rangle_{(\widehat{K}_\lambda^x)^\alpha} = \left\langle \left(n\widehat{\Sigma}_n + xx^\top + \lambda(n+1)I_d\right)^\alpha n\widehat{S}_n, x \right\rangle$$

$$= \left\langle \left(n\widehat{\Sigma}_n + xx^\top + \lambda(n+1)I_d\right)^\alpha (n\widehat{\Sigma}_n + \lambda(n+1)I_d + xx^\top - xx^\top)\widehat{\theta}_{\lambda',n}, x \right\rangle$$

$$= \langle \widehat{\theta}_{\lambda',n}, x \rangle_{(\widehat{K}_\lambda^x)^{\alpha-1}} - \langle \widehat{\theta}_{\lambda',n}, x \rangle \|x\|_{(\widehat{K}_\lambda^x)^\alpha}^2,$$

namely

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

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

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

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

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

$$\leqslant \mathbb{E}_{Z_1^n,X}\Big[\log\Big(\int_{\mathbb{R}} f_{\widehat{\theta}_{\lambda,n}^{(X,y)}}(y|X)\mu(\mathrm{d}y)\Big)\Big].$$

We know from the computations above that

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

so that, after integrating with respect to y,

$$\mathbb{E}\big[\mathcal{E}_\lambda(\widetilde{f}_{\lambda,n})\big] \leqslant \mathbb{E}_{X_1^n,X}\Big[\log\Big(\frac{1}{1 - \|X\|^2_{\widehat{K}_\lambda^X}}\Big)\Big] = \mathbb{E}_{X_1^n,X}\Big[-\log\big(1 - \langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle\big)\Big]. \quad (72)$$

Note that, by the identity (64) from Lemma 17, and since $\|X\| \leqslant R$ almost surely (Assumption 5) we have

$$\langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle = \frac{\langle(n\widehat{\Sigma}_n + \lambda(n+1)I_d)^{-1}X, X\rangle}{1 + \langle(n\widehat{\Sigma}_n + \lambda(n+1)I_d)^{-1}X, X\rangle} \leqslant \frac{R^2/\big(\lambda(n+1)\big)}{1 + R^2/\big(\lambda(n+1)\big)}. \quad (73)$$

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

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

where we used the inequality $\log(1+v) \leqslant v$ for $v \geqslant 0$. Combining this fact with (73) shows that

$$-\log\big(1 - \langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle\big) \leqslant g\Big(\frac{R^2/\big(\lambda(n+1)\big)}{1 + R^2/\big(\lambda(n+1)\big)}\Big) \cdot \langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle. \quad (74)$$

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

$$\mathbb{E}\big[\langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle\big] = \frac{1}{n+1}\mathbb{E}\Big[\sum_{i=1}^n\langle(\widehat{\Sigma}_\lambda^X)^{-1}X_i, X_i\rangle + \langle(\widehat{\Sigma}_\lambda^X)^{-1}X, X\rangle\Big]$$

$$= \frac{1}{n+1}\mathbb{E}\Big[\mathrm{tr}\Big\{\Big(\sum_{i=1}^n X_iX_i^\top + XX^\top + \lambda(n+1)I_d\Big)^{-1}\Big(\sum_{i=1}^n X_iX_i^\top + XX^\top\Big)\Big\}\Big]. \quad (75)$$

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. Bhatia (2009); Boyd and Vandenberghe (2004). Hence, applying Jensen's inequality to (75) and using the fact that

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

we obtain:

$$\mathbb{E}\big[\langle (\widehat{\Sigma}_\lambda^X)^{-1} X, X \rangle\big] \leqslant \frac{\mathsf{df}_\lambda(\Sigma)}{n+1}. \tag{76}$$

Finally, combining the bounds (72), (74) and (76) yields:

$$\mathbb{E}\big[\mathcal{E}_\lambda(\widetilde{f}_{\lambda,n})\big] \leqslant g\bigg(\frac{R^2/\big(\lambda(n+1)\big)}{1 + R^2/\big(\lambda(n+1)\big)} \bigg) \cdot \frac{\mathsf{df}_\lambda(\Sigma)}{n+1}. \tag{77}$$

Nonparametric rates (Theorem 9). Assume that $\lambda(n+1) \geqslant 2R^2$. The quantity inside $g(\cdot)$ in (77) is then bounded by $(1/2)/(1 + 1/2) = 1/3$, and since $g(1/3) = 3\log(3/2) \leqslant 1.25$, (77) becomes, by definition of $\mathcal{E}_\lambda$:

$$\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\theta \in \mathbb{R}^d} \bigg\{ R(f_\theta) + \frac{\lambda}{2}\|\theta\|^2 \bigg\} \leqslant 1.25 \cdot \frac{\mathsf{df}_\lambda(\Sigma)}{n+1}. \tag{78}$$

which is precisely the announced bound (33).

Finite-dimensional case: improved dependence on the norm (Proposition 10). 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 θ). Then, $R^2/\big(\lambda(n+1)\big) = B^2 R^2/d$. Now, note that for every $v > 0$

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

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 $v > 0$:

$$g\bigg(\frac{v}{1+v} \bigg) \leqslant 2\log(e + v) \leqslant 2\log(4 + 4\sqrt{v} + v) = 4\log(2 + \sqrt{v}). \tag{79}$$

Now, the excess risk bound (77) implies that, for every $\theta \in \mathbb{R}^d$ such that $\|\theta\| \leqslant B$,

$$\begin{aligned}
\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - R(f_\theta) &\leqslant g\bigg(\frac{B^2 R^2/d}{1 + B^2 R^2/d} \bigg) \cdot \frac{\mathsf{df}_\lambda(\Sigma)}{n+1} + \frac{\lambda}{2}\|\theta\|^2 \\
&\leqslant 4\log\bigg(2 + \frac{BR}{\sqrt{d}} \bigg) \times \frac{d}{n+1} + \frac{d}{B^2(n+1)} \times \frac{B^2}{2} \tag{80} \\
&= \frac{d}{n+1} \bigg\{ 4\log\bigg(2 + \frac{BR}{\sqrt{d}} \bigg) + \frac{1}{2} \bigg\} \\
&\leqslant \frac{5d\log\big(2 + BR/\sqrt{d}\big)}{n+1} \tag{81}
\end{aligned}$$

37

where inequality (80) uses the bound (79) with $v = B^2 R^2/d$, the bound $\mathsf{df}_\lambda(\Sigma) \leqslant d$ (30) and the fact that $\|\theta\| \leqslant B$, while inequality (81) uses the fact that $1/2 \leqslant \log 2$. ∎

7.4 Proofs for logistic regression (Section 5)

Proof of Proposition 12 Let us first discuss the properties of predictions produced by the SMP, and compare it to the MLE. First, if the points $Z_1, \ldots, Z_n$ 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 $y \in \{-1, 1\}$, $Z_1, \ldots, Z_n, -yx$ are not separated either, so $\widehat{\theta}_n^{(x,y)} \in \mathbb{R}^d$ 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 $Z_1, \ldots, Z_n$. Assume that $\Lambda_n \cap (-\Lambda_n) = \{0\}$ and that all Z_i are distinct from 0. Then, convex separation implies that there exists $\theta \in \mathbb{R}^d$ such that $\langle \theta, z \rangle < 0$ for all $z \in \Lambda_n \setminus \{0\}$, so that the Z_i lie within a strict half-space: $\langle \theta, Z_i \rangle < 0$ for all i. Hence, any MLE $f_{\widehat{\theta}_n}$ 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 $f_{\widehat{\theta}_n}(1|x)$ are as follows:

- If $x = 0$, then $f_{\widehat{\theta}_n}(1|x) = 1/2$.

- If $x \in \Lambda_n \setminus \{0\}$, then $\langle \widehat{\theta}_n, x \rangle < 0$ and thus $f_{\widehat{\theta}_n}(1|x) = 0$. Likewise, if $x \in (-\Lambda_n) \setminus \{0\}$, then $f_{\widehat{\theta}_n}(1|x) = 1$;

- 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 $\widehat{\theta}_n$ such that $f_{\widehat{\theta}_n}(1|x)$ 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 θ with $\langle \theta, z \rangle < 0$ and $\langle \theta, Z_i \rangle = -\langle \theta, -Z_i \rangle < 0$. Hence, $f_{\widehat{\theta}_n^{(x,y)}}(y|x) = 1$. On the other hand, if $z \in (-\Lambda_n) \setminus \{0\}$, then the dataset $Z_1, \ldots, Z_n, z$ is not separated, so that $f_{\widehat{\theta}_n^{(x,y)}}(y|x) \in (0, 1)$. 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 $f_{\widehat{\theta}_n^{(x,1)}}(1|x) \in (0, 1)$, while $x \in \mathbb{R}^d \setminus (-\Lambda_n)$ so that $f_{\widehat{\theta}_n^{(x,-1)}}(-1|x) = 1$; 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 $f_{\widehat{\theta}_n^{(x,1)}}(1|x) = f_{\widehat{\theta}_n^{(x,-1)}}(-1|x) = 1$, so that $\widetilde{f}_n(1|x) = 1/2$.

Finally, the excess risk bound (45) is established in the proof of Theorem 9 below, letting $\lambda = 0$. ∎

Proof of Theorem 13 Let (X, Y) be a test sample, and $Z = -YX$. Since $\{Z, -Z\} = \{X, -X\}$, the excess risk bound (15) of the SMP $\widetilde{f}_{\lambda,n}$ (47) writes:

$$
\mathbb{E}\big[R(\widetilde{f}_{\lambda,n})\big] - \inf_{\theta \in \mathbb{R}^d} \left\{ R(f_\theta) + \frac{\lambda}{2} \|\theta\|^2 \right\}
$$

$$
\leqslant \mathbb{E}\left[\log\left(\sigma(\langle \widehat{\theta}_{\lambda,n}^{(X,1)}, X \rangle)\, e^{-\lambda \|\widehat{\theta}_{\lambda,n}^{(X,1)}\|^2/2} + \sigma(-\langle \widehat{\theta}_{\lambda,n}^{(X,-1)}, X \rangle)\, e^{-\lambda \|\widehat{\theta}_{\lambda,n}^{(X,-1)}\|^2/2} \right) \right]
$$

$$
= \mathbb{E}\left[\log\left(\sigma(\langle \widehat{\theta}_{\lambda,n}^{-Z}, Z \rangle)\, e^{-\lambda \|\widehat{\theta}_{\lambda,n}^{-Z}\|^2/2} + \sigma(-\langle \widehat{\theta}_{\lambda,n}^{Z}, Z \rangle)\, e^{-\lambda \|\widehat{\theta}_{\lambda,n}^{Z}\|^2/2} \right) \right]
$$

$$
\leqslant \mathbb{E}\left[\log\left(1 + \sigma(\langle \widehat{\theta}_{\lambda,n}^{-Z}, Z \rangle) - \sigma(\langle \widehat{\theta}_{\lambda,n}^{Z}, Z \rangle) \right) \right] \tag{82}
$$

$$
\leqslant \mathbb{E}\left[\sigma(\langle \widehat{\theta}_{\lambda,n}^{-Z}, Z \rangle) - \sigma(\langle \widehat{\theta}_{\lambda,n}^{Z}, Z \rangle) \right] \tag{83}
$$

where inequality (82) is obtained by lower-bounding $e^{-\lambda \|\cdot\|^2/2} \leqslant 1$ and using the identity $\sigma(-u) = 1 - \sigma(u)$. Now, defining for $\theta \in \mathbb{R}^d$

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

we have, respectively,

$$
\widehat{\theta}_{\lambda,n}^{Z} = \arg\min_{\theta \in \mathbb{R}^d} \widehat{R}_{\lambda,n}^{Z}(\theta) \tag{84}
$$

$$
\widehat{\theta}_{\lambda,n}^{-Z} = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \widehat{R}_{\lambda,n}^{Z}(\theta) - \frac{1}{n+1} \langle \theta, Z \rangle \right\}, \tag{85}
$$

where (85) 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 $\lambda \|\theta\|^2/2$ term. It follows from Lemma 18 that

$$
R \cdot \big\| \widehat{\theta}_{\lambda,n}^{-Z} - \widehat{\theta}_{\lambda,n}^{Z} \big\| \leqslant R \cdot \frac{\|Z/(n+1)\|}{\lambda} \leqslant \frac{R^2}{\lambda(n+1)} \leqslant \frac{1}{2} \,, \tag{86}
$$

where we used the assumption that $\lambda \geqslant 2R^2/(n+1)$. In addition, still by Lemma 18,

$$
0 \leqslant \langle \widehat{\theta}_{\lambda,n}^{-Z} - \widehat{\theta}_{\lambda,n}^{Z}, Z \rangle \leqslant 1/2 \,. \tag{87}
$$

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 $\sigma'(u+v) \leqslant e^v \sigma'(u) \leqslant e^{1/2} \cdot \sigma'(u)$. Hence, $\sigma(u+v) \leqslant e^{1/2} \cdot \sigma'(u) \cdot v$ for every $u \in \mathbb{R}$ and $v \in [0, 1/2]$. By (87), applying this inequality to $u = \langle \widehat{\theta}_{\lambda,n}^{Z}, Z \rangle$ and $v = \langle \widehat{\theta}_{\lambda,n}^{-Z} - \widehat{\theta}_{\lambda,n}^{Z}, Z \rangle$ yields:

$$
\sigma\big(\langle \widehat{\theta}_{\lambda,n}^{-Z}, Z \rangle\big) - \sigma\big(\langle \widehat{\theta}_{\lambda,n}^{Z}, Z \rangle\big) \leqslant e^{1/2} \cdot \sigma'\big(\langle \widehat{\theta}_{\lambda,n}^{Z}, Z \rangle\big) \cdot \langle \widehat{\theta}_{\lambda,n}^{-Z} - \widehat{\theta}_{\lambda,n}^{Z}, Z \rangle \,. \tag{88}
$$

Let us now consider the function $\widehat{R}_{\lambda,n}^{Z}$; its third derivative can be controlled in terms of its Hessian, as shown by Bach (2010). Fix $\theta, \beta \in \mathbb{R}^d$, and define the function $g(t) = \widehat{R}_{\lambda,n}^{Z}(\theta + t\beta)$

for $t \in \mathbb{R}$. We have respectively, denoting $\theta_t = \theta + t\beta$,

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

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

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 $|\langle \beta, Z_i \rangle| \leqslant R\|\beta\|$ $(1 \leqslant i \leqslant n)$ and $|\langle \beta, Z \rangle| \leqslant R\|\beta\|$, we have

$$|g'''(t)| = \frac{1}{n+1}\left\{\sum_{i=1}^n \left|\sigma''(\langle \theta_t, Z_i \rangle)\langle \beta, Z_i \rangle^3\right| + \left|\sigma''(\langle \theta_t, Z \rangle)\langle \beta, Z \rangle^3\right|\right\}$$

$$\leqslant R\|\theta\| \cdot \frac{1}{n+1}\left\{\sum_{i=1}^n \sigma'(\langle \theta_t, Z_i \rangle)\langle \beta, Z_i \rangle^2 + \sigma'(\langle \theta_t, Z \rangle)\langle \beta, Z \rangle^2\right\} \leqslant R\|\beta\| \cdot g''(t). \quad (89)$$

The property (89) is the pseudo-self-concordance condition introduced by Bach (2010); in particular, by Proposition 1 therein, we have for every $\theta, \beta \in \mathbb{R}^d$:

$$\nabla^2 \widehat{R}_{\lambda,n}^Z(\theta + \beta) \succcurlyeq e^{-R\|\beta\|} \cdot \nabla^2 \widehat{R}_{\lambda,n}^Z(\theta). \quad (90)$$

It follows from (90) (letting $\theta = \widehat{\theta}_{\lambda,n}^Z$ and $\beta = \theta' - \widehat{\theta}_{\lambda,n}^Z$) that $\widehat{R}_{\lambda,n}^Z$ is $e^{-(1/2+\varepsilon)}\nabla^2 \widehat{R}_{\lambda,n}^Z(\widehat{\theta}_{\lambda,n}^Z)$-strongly convex on the open convex ball $\Omega_\varepsilon = \{\theta' \in \mathbb{R}^d : R\|\theta' - \widehat{\theta}_{\lambda,n}^Z\| < 1/2 + \varepsilon\}$ for every $\varepsilon > 0$. In addition, the inequality (86) shows that the function $\widehat{R}_{\lambda,n}^Z(\theta) - \langle \theta, Z \rangle/(n+1)$ reaches its minimum $\widehat{\theta}_{\lambda,n}^{-Z}$ on Ω_ε, so that by Lemma 18,

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

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

$$\langle \widehat{\theta}_{\lambda,n}^{-Z} - \widehat{\theta}_{\lambda,n}^Z, Z \rangle \leqslant \frac{e^{1/2}}{n+1} \cdot \langle \nabla^2 \widehat{R}_{\lambda,n}^Z(\widehat{\theta}_{\lambda,n}^Z)^{-1}Z, Z \rangle, \quad (91)$$

so that by combining inequalities (88) and (91),

$$\sigma(\langle \widehat{\theta}_{\lambda,n}^{-Z}, Z \rangle) - \sigma(\langle \widehat{\theta}_{\lambda,n}^Z, Z \rangle) \leqslant \frac{e}{n+1} \cdot \sigma'(\langle \widehat{\theta}_{\lambda,n}^Z, Z \rangle) \cdot \langle \nabla^2 \widehat{R}_{\lambda,n}^Z(\widehat{\theta}_{\lambda,n}^Z)^{-1}Z, Z \rangle. \quad (92)$$

It thus remains to control the expectation of the right-hand side of (92). By exchangeability of $(Z_1, \ldots, Z_n, Z)$ (and since $\widehat{R}_{\lambda,n}^Z, \widehat{\theta}_{\lambda,n}^Z$ are unchanged after permutation of Z_i and Z), we have:

$$\mathbb{E}\big[\sigma'(\langle \widehat{\theta}_{\lambda,n}^Z, Z \rangle) \cdot \langle \nabla^2 \widehat{R}_{\lambda,n}^Z(\widehat{\theta}_{\lambda,n}^Z)^{-1}Z, Z \rangle\big]$$

$$= \frac{1}{n+1}\mathbb{E}\bigg[\sum_{i=1}^n \sigma'(\langle \widehat{\theta}_{\lambda,n}^Z, Z_i \rangle) \cdot \langle \nabla^2 \widehat{R}_{\lambda,n}^Z(\widehat{\theta}_{\lambda,n}^Z)^{-1}Z_i, Z_i \rangle + \sigma'(\langle \widehat{\theta}_{\lambda,n}^Z, Z \rangle) \cdot \langle \nabla^2 \widehat{R}_{\lambda,n}^Z(\widehat{\theta}_{\lambda,n}^Z)^{-1}Z, Z \rangle\bigg]$$

$$= \mathbb{E}\bigg[\mathrm{tr}\Big\{\nabla^2 \widehat{R}_{\lambda,n}^Z(\widehat{\theta}_{\lambda,n}^Z)^{-1} \cdot \frac{1}{n+1}\Big(\sum_{i=1}^n \sigma'(\langle \widehat{\theta}_{\lambda,n}^Z, Z \rangle)Z_i Z_i^\top + \sigma'(\langle \widehat{\theta}_{\lambda,n}^Z, Z \rangle)ZZ^\top\Big)\Big\}\bigg]$$

$$= \mathbb{E}\Big[\mathrm{tr}\Big\{\big[\nabla^2 \widehat{R}_n^Z(\widehat{\theta}_{\lambda,n}^Z) + \lambda I_d\big]^{-1}\nabla^2 \widehat{R}_n^Z(\widehat{\theta}_{\lambda,n}^Z)\Big\}\Big]; \quad (93)$$

in (93), we defined

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

whose Hessian writes

$$\nabla^2 \widehat{R}_n^Z(\theta) = \frac{1}{n+1}\left\{\sum_{i=1}^n \sigma'(\langle\theta, Z_i\rangle)Z_i Z_i^\top + \sigma'(\langle\theta, Z\rangle)ZZ^\top\right\}.$$

Finally, by concavity of the map $A \mapsto \operatorname{tr}[(A + \lambda I_d)^{-1}A]$ on positive matrices (as in the proof of Theorem 9), denoting $\widetilde{H}_{\lambda,n} := \mathbb{E}[\nabla^2\widehat{R}_n^Z(\widehat{\theta}_{\lambda,n}^Z)] = \mathbb{E}[\nabla^2\widehat{R}_{n+1}(\widehat{\theta}_{\lambda,n+1})]$ we have

$$\mathbb{E}\left[\operatorname{tr}\left\{\left[\nabla^2\widehat{R}_n^Z(\widehat{\theta}_{\lambda,n}^Z) + \lambda I_d\right]^{-1}\nabla^2\widehat{R}_n^Z(\widehat{\theta}_{\lambda,n}^Z)\right\}\right] \leqslant \operatorname{tr}\left\{[\widetilde{H}_{\lambda,n} + \lambda I_d]^{-1}\widetilde{H}_{\lambda,n}\right\} = \mathsf{df}_\lambda(\widetilde{H}_{\lambda,n}). \quad (94)$$

Combining inequalities (83), (92), (93) and (94), we conclude that

$$\mathbb{E}\left[R(\widetilde{f}_{\lambda,n})\right] - \inf_{\theta\in\mathbb{R}^d}\left\{R(f_\theta) + \frac{\lambda}{2}\|\theta\|^2\right\} \leqslant e \cdot \frac{\mathsf{df}_\lambda(\widetilde{H}_{\lambda,n})}{n+1}. \quad (95)$$

Finally, the bound (48) is obtained by noting that, by exchangeability and since $\sigma' = \sigma(1-\sigma) \leqslant 1/4$ and $Z_1 Z_1^\top = X_1 X_1^\top$,

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

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

Lemma 18 (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 $x, x' \in \Omega$,*

$$F(x') \geqslant F(x) + \langle\nabla F(x), x' - x\rangle + \frac{1}{2}\|x' - x\|_\Sigma^2. \quad (96)$$

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,

$$\|\widetilde{x} - x^*\|_\Sigma \leqslant \|g\|_{\Sigma^{-1}}, \qquad \langle g, \widetilde{x} - x^*\rangle \leqslant \|g\|_{\Sigma^{-1}}^2. \quad (97)$$

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

$$\langle\nabla F(\widetilde{x}), \widetilde{x} - x^*\rangle = \langle g, x\rangle. \quad (98)$$

Now, by substituting x' and x in inequality (96) and adding the resulting inequality to (96), we obtain for every $x, x' \in \Omega$,

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

41

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

$$\langle g, \widetilde{x} - x^* \rangle \leqslant \|g\|_{\Sigma^{-1}} \cdot \|\widetilde{x} - x^*\|_\Sigma. \tag{99}$$

Plugging the previous inequalities in (98) 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 (99). ∎

References

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

Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley New York, 2003.

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

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

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

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

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

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

Francis Bach and Éric 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.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Andrew R. Barron. Are Bayes rules consistent in information? In *Open Problems in Communication and Computation*, pages 85–91. Springer, 1987.

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

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

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

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

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

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

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

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

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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

Lawrence D. Brown, Edward I. George, and Xinyi Xu. Admissible predictive density estimation. *The Annals of Statistics*, pages 1156–1170, 2008.

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

Emmanuel J. Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48 (1):27–42, 2020.

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

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

Olivier 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.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, New York, USA, 2006.

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

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

Thomas M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–19, 1991.

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

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

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics*. Springer-Verlag, 1996.

Jürgen Forster and Manfred K. Warmuth. Relative expected instantaneous loss bounds. *Journal of Computer and System Sciences*, 64(1):76–102, 2002.

Dylan J. Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: the importance of being improper. In *Proceedings of the 31st Conference On Learning Theory*, pages 167–208, 2018.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*. Springer series in statistics, New York, 2001.

Edward I. George, Feng Liang, and Xinyi Xu. Improved minimax predictive densities under Kullback-Leibler loss. *The Annals of Statistics*, 34(1):78–91, 2006.

Jayanta K Ghosh. *Higher order asymptotics*. Institute of Mathematical Statistics, 1994.

Alon Gonen and Shai Shalev-Shwartz. Average stability is invariant to data preconditioning. implications to exp-concave empirical risk minimization. *Journal of Machine Learning Research*, 18(222):1–13, 2018.

Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

Peter D. Grünwald and Wojciech Kotłowski. Open problem: Bounds on individual risk for log-loss predictors. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 813–816, 2011.

Peter D. Grünwald and Nishant A. Mehta. A tight excess risk bound via a unified PAC-Bayesian-Rademacher-Shtarkov-MDL complexity. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, pages 433–465, 2019.

Jaroslav Hájek. Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 175–194, 1972.

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

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

David Haussler, Nick Littlestone, and Manfred K. Warmuth. Predicting {0, 1}-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

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

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Elad Hazan, Tomer Koren, and Kfir Y. Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Proceedings of the 27th Conference on Learning Theory*, pages 197–209, 2014.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

Ildar Abdulovič Ibragimov and Rafail Zalmanovich Has'minskii. *Statistical estimation: asymptotic theory*. Springer Science & Business Media, 1981.

Rémi Jézéquel, Pierre Gaillard, and Alessandro Rudi. Efficient improper learning for online logistic regression. *arXiv preprint arXiv:2003.08109*, 2020.

Anatoli Juditsky, Philippe Rigollet, and Alexandre B. Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.

Sham M. Kakade and Andrew Y. Ng. Online bounds for Bayesian algorithms. In *Advances in Neural Information Processing Systems 17*, pages 641–648, 2005.

Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems 21*, pages 793–800, 2009.

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

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

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

Wojciech Kotłowski and Peter D. Grünwald. Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 457–476, 2011.

Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer-Verlag New York, 1986.

Lucien Le Cam and Grace Lo Yang. *Asymptotics in statistics: some basic concepts*. Springer Series in Statistics. Springer-Verlag New York, 2000.

Guillaume Lecué and Shahar Mendelson. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534, 2016.

Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer, 1998.

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

Nick Littlestone. From on-line to batch learning. In *Proceedings of the 2nd annual workshop on Computational Learning Theory*, pages 269–284, 1989.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

Mehrdad Mahdavi, Lijun Zhang, and Rong Jin. Lower and upper bounds on the generalization of stochastic exponentially concave optimization. In *Proceedings of the 28th Conference on Learning Theory*, pages 1305–1320, 2015.

Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Globally convergent Newton methods for ill-conditioned generalized self-concordant losses. In *Advances in Neural Information Processing Systems 32*, pages 7634–7644, 2019a.

Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least-squares: fast rates for regularized empirical risk minimization through self-concordance. *Proceedings of the 32nd Conference on Learning Theory*, pages 2294–2340, 2019b.

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

Peter McCullagh and John A. Nelder. *Generalized Linear Models*. Chapman and Hall/CRC, 2 edition, 1989.

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

Ron Meir and Tong Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4(Oct):839–860, 2003.

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

Neri Merhav and Meir Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44:2124–2147, 1998.

Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *arXiv preprint arXiv:1912.10754*, 2019.

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

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Vee Ming Ng. On the estimation of parametric density functions. *Biometrika*, 67(2):505–506, 1980.

Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3):1175–1194, 2016.

Dmitrii Ostrovskii and Francis Bach. Finite-sample analysis of M-estimators using self-concordance. *arXiv preprint arXiv:1810.06838*, 2018.

Alexander Rakhlin, Sayan Mukherjee, and Tomaso Poggio. Stability results in learning theory. *Analysis and Applications*, 3(4):397–417, 2005.

Jorma J. Rissanen. *Minimum description length principle*. Wiley Online Library, 1985.

Jorma J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, 1996.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Teemu Roos and Jorma J. Rissanen. On sequentially normalized maximum likelihood models. In *Workshop on Information Theoretic Methods in Science and Engineering*, 2008.

Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*, pages 11982–11992, 2019.

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

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

Yuri M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3–17, 1987.

Vladimir Spokoiny. Parametric estimation. Finite sample theory. *The Annals of Statistics*, 40(6):2877–2909, 2012.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems 23*, pages 2199–2207, 2010.

Karthik Sridharan, Shai Shalev-Shwartz, and Nathan Srebro. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems 21*, pages 1545–1552, 2009.

Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29): 14516–14525, 2019.

Trevor J. Sweeting, Gauri S. Datta, and Malay Ghosh. Nonsubjective priors via predictive relative entropy regret. *The Annals of Statistics*, pages 441–468, 2006.

Eiji Takimoto and Manfred K. Warmuth. The last-step minimax algorithm. In *International conference on Algorithmic Learning Theory*, pages 279–290, 2000.

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

Sara van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, Cambridge, 1999.

Aad van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Vladimir N. Vapnik and Alexey Ya. Chervonenkis. *Theory of Pattern Recognition: Statistical Learning Problems (in Russian)*. Nauka, Moscow, 1974.

Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, pages 210–268. Cambridge University Press, Cambridge, 2012.

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

Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Grace Wahba. *Spline Models for Observational Data*, volume 59. SIAM, 1990.

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

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

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50 (1):1–25, 1982.

Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics*, 23(2):339–362, 1995.

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

Yuhong Yang. Mixing strategies for density estimation. *The Annals of Statistics*, 28(1): 75–87, 2000.

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

Yuhong Yang and Andrew R. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

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

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936, 2003.