A Unified Framework for Consistency of Regularized Loss Minimizers

Jean Honorio JHONORIO@CSAIL.MIT.EDU
Tommi Jaakkola TOMMI@CSAIL.MIT.EDU
CSAIL, MIT, Cambridge, MA 02139, USA

Abstract

We characterize a family of *regularized loss minimization problems* that satisfy three properties: scaled uniform convergence, super-norm regularization, and norm-loss monotonicity. We show several theoretical guarantees within this framework, including loss consistency, norm consistency, sparsistency (i.e. support recovery) as well as sign consistency. A number of regularization problems can be shown to fall within our framework and we provide several examples. Our results can be seen as a concise summary of existing guarantees but we also extend them to new settings. Our formulation enables us to assume very little about the hypothesis class, data distribution, the loss, or the regularization. In particular, many of our results do not require a bounded hypothesis class, or identically distributed samples. Similarly, we do not assume boundedness, convexity or smoothness of the loss nor the regularizer. We only assume approximate optimality of the empirical minimizer. In terms of recovery, in contrast to existing results, our sparsistency and sign consistency results do not require knowledge of the sub-differential of the objective function.

1. Introduction

Several problems in machine learning can be modeled as the minimization of an *empirical loss*, which is computed from some available training data. Assuming that data samples come from some unknown arbitrary distribution, we can define the *expected loss* as the expected value of the empirical loss. The minimizers of the empirical and expected loss are called the *empirical minimizer* and the *true hypothesis*, respectively.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

One of the goals in machine learning is to infer properties of the true hypothesis by having access to a limited amount of training data. One largely used property in the context of classification and regression is *loss consistency* which measures the generalization ability of the learning algorithm. Loss consistency guarantees are usually stated as an upper bound on the difference between the expected loss of the empirical minimizer and that of the true hypothesis. Another set of properties relates to the ability to recover the true hypothesis. *Norm consistency* measures the distance between the empirical minimizer and the true hypothesis. *Sparsistency* refers to the recovery of the sparsity pattern (i.e. support recovery) of the true hypothesis, while *sign consistency* refers to the recovery of the signs of the true hypothesis. We expect these guarantees to become stronger as we have access to more data samples.

In many settings, these guarantees are made possible by the use of a *regularizer* in the learning process. Consistency guarantees are now available for several specific *regularized loss minimization* problems. We can hardly do justice to the body of prior work, and we provide a few references here. The work on linear regression includes the analysis of: the sparsity promoting ℓ_1-norm (Wainwright, 2009b), the multitask $\ell_{1,2}$ and $\ell_{1,\infty}$-norms (Negahban & Wainwright, 2011; Obozinski et al., 2011), the multitask $\ell_{1,2}$-norm for overlapping groups (Jacob et al., 2009), the *dirty* multitask regularizer (Jalali et al., 2010), the Tikhonov regularizer (Hsu et al., 2012), and the trace norm (Bach, 2008). The analysis of ℓ_1-regularization has also been performed for: the estimation of exponential family distributions (Kakade et al., 2010; Ravikumar et al., 2008; Wainwright et al., 2006), generalized linear models (Kakade et al., 2010; van de Geer, 2008; Yang et al., 2013), and SVMs and logistic regression (Rocha et al., 2009; van de Geer, 2008). These works have focused on norm consistency and sparsistency, with the exception of (Jalali et al., 2010; Obozinski et al., 2011; Ravikumar et al., 2008; Rocha et al., 2009; Wainwright, 2009b; Wainwright et al., 2006) which also analyzed sign consistency, and (Hsu et al., 2012; Kakade et al., 2010) which also analyzed loss consistency. We refer the interested reader to the article of (Negahban et al., 2012) for additional references.

There has been some notable contributions which characterize general frameworks with theoretical guarantees. Loss consistency for bounded losses and different notions of stability of the learning algorithm was analyzed in (Bousquet & Elisseeff, 2002; Mukherjee et al., 2006; Rakhlin et al., 2005; Shalev-Shwartz et al., 2010). Stability follows from the use of regularization for many different problems (Bousquet & Elisseeff, 2002). A two-level framework for loss consistency of bounded losses was provided by (van de Geer, 2005): a regularized outer-minimization is performed with respect to a set of model classes, while an unregularized inner-minimization is done with respect to functions on each class. Norm consistency for restricted strongly convex (i.e. strongly convex with respect to a subset of directions) losses and certain type of regularizers was analyzed in (Lee et al., 2013; Loh & Wainwright, 2013; Negahban et al., 2009; 2012; Yang & Ravikumar, 2013). In (Lee et al., 2013; Negahban et al., 2009; 2012; Yang & Ravikumar, 2013) the loss is differentiable and convex, and the regularizer is a mixture of decomposable norms; while in (Loh & Wainwright, 2013) the loss is differentiable and nonconvex, and the regularizer is coordinate-separable, symmetric and nondecreasing, among other technical requirements. The work of (Bousquet & Elisseeff, 2002; Mukherjee et al., 2006; Rakhlin et al., 2005; Shalev-Shwartz et al., 2010; van de Geer, 2005) focus on loss consistency and requires an *everywhere* bounded loss. On the other hand, the work of (Lee et al., 2013; Loh & Wainwright, 2013; Negahban et al., 2009; 2012; Yang & Ravikumar, 2013) focus on norm consistency and requires a differentiable loss. The framework of (van de Geer, 2005) requires a measure of complexity for each model class as well as over all classes (which are infinity for the problems that we analyze here). Finally, the availability of independent and identically distributed samples is a requirement for all these previous works.

In this paper, we characterize a family of *regularized loss minimization problems* that fulfill three properties: scaled uniform convergence, super-scale regularization and norm-loss monotonicity. We show loss consistency, norm consistency, sparsistency and sign consistency. We show that several problems in the literature fall in our framework, such as the estimation of exponential family distributions, generalized linear models, matrix factorization problems, nonparametric models and PAC-Bayes learning. Similarly, several regularizers fulfill our assumptions, such as sparsity promoting priors, multitask priors, low-rank regularizers, elastic net, total variation, dirty models, quasiconvex regularizers, among others. Note that our theoretical results imply that loss consistency, norm consistency, sparsistency and sign consistency hold for any combination of losses and regularizers that we discuss here. Many of these combinations have not been previously explored.

2. Preliminaries

We first characterize a general *regularized loss minimization* problem. To this end, we define a problem as a tuple $\Pi = (\mathcal{H}, \mathcal{D}, \widehat{\mathcal{L}}_n, \mathcal{R})$ for a hypothesis class $\mathcal{H}$, a data distribution $\mathcal{D}$, an empirical loss $\widehat{\mathcal{L}}_n$ and a regularizer $\mathcal{R}$. For clarity of presentation, we assume that $\mathcal{H}$ is a normed vector space.

Let $\boldsymbol{\theta}$ be a hypothesis belonging to a possibly unbounded hypothesis class $\mathcal{H}$. Let $\widehat{\mathcal{L}}_n(\boldsymbol{\theta})$ be the empirical loss of n samples drawn from a distribution $\mathcal{D}$. We do not assume either independence or identical distribution of the samples. Let $\mathcal{R}(\boldsymbol{\theta})$ be a regularizer and $\lambda_n > 0$ be a penalty parameter. The *empirical minimizer* is given by:

$$\widehat{\boldsymbol{\theta}}_n^* = \arg\min_{\boldsymbol{\theta} \in \mathcal{H}} \widehat{\mathcal{L}}_n(\boldsymbol{\theta}) + \lambda_n \mathcal{R}(\boldsymbol{\theta}) \tag{1}$$

We relax this optimality assumption by defining an ξ-*approximate empirical minimizer* $\widehat{\boldsymbol{\theta}}_n$ with the following property for $\xi \geq 0$:

$$\widehat{\mathcal{L}}_n(\widehat{\boldsymbol{\theta}}_n) + \lambda_n \mathcal{R}(\widehat{\boldsymbol{\theta}}_n) \leq \xi + \min_{\boldsymbol{\theta} \in \mathcal{H}} \widehat{\mathcal{L}}_n(\boldsymbol{\theta}) + \lambda_n \mathcal{R}(\boldsymbol{\theta}) \tag{2}$$

Let $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}}[\widehat{\mathcal{L}}_n(\boldsymbol{\theta})]$ be the expected loss. The *true hypothesis* is given by:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \mathcal{H}} \mathcal{L}(\boldsymbol{\theta}) \tag{3}$$

In this paper, we do not assume boundedness, convexity or smoothness of $\widehat{\mathcal{L}}_n$, $\mathcal{R}$ or $\mathcal{L}$, although convexity is a very useful property from an optimization viewpoint.

In order to illustrate the previous setting, consider for instance a function $\ell(\mathbf{x}|\boldsymbol{\theta})$ to be the loss of a data sample $\mathbf{x}$ given $\boldsymbol{\theta}$. We can define $\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_i \ell(\mathbf{x}^{(i)}|\boldsymbol{\theta})$ to be the empirical loss of n i.i.d. samples $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ drawn from a distribution $\mathcal{D}$. Then, $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\ell(\mathbf{x}|\boldsymbol{\theta})]$ is the expected loss of $\mathbf{x}$ drawn from a distribution $\mathcal{D}$. Our framework is far more general than this specific example. We do not require identically distributed samples.

In order to gain some theoretical understanding of the general problem defined above, we make some reasonable assumptions. Therefore, we extend the problem tuple with additional terms related to these assumptions. That is, we define a problem as a tuple $\Pi' = (\mathcal{H}, \mathcal{D}, \widehat{\mathcal{L}}_n, \mathcal{R}, \varepsilon_{n,\delta}, c, r, b)$ for a hypothesis class $\mathcal{H}$, a data distribution $\mathcal{D}$, an empirical loss $\widehat{\mathcal{L}}_n$, a regularizer $\mathcal{R}$, a uniform convergence rate $\varepsilon_{n,\delta}$ and scale function c, a regularizer lower bound r and norm-loss function b. In what follows, we explain in detail the use of these additional terms.

2.1. Scaled Uniform Convergence

First, we present our definition of *scaled* uniform convergence, which differs from *regular* uniform convergence. In both uniform convergence schemes, the goal is to find a bound on the difference between the empirical and expected loss for all $\boldsymbol{\theta}$. In *regular* uniform convergence such bound is the same for all $\boldsymbol{\theta}$, while in *scaled* uniform convergence the bound depends on the "scale" of $\boldsymbol{\theta}$. For finite and infinite dimensional vector spaces as well as for function spaces, we can choose the scale to be the norm of $\boldsymbol{\theta}$. For the space of probability distributions, we can choose the scale to be the Kullback-Leibler divergence from the distribution $\boldsymbol{\theta}$ to a prior $\boldsymbol{\theta}^{(0)}$. Next, we formally state our definition.

Assumption A (Scaled uniform convergence)**.** *Let $c : \mathcal{H} \to [0; +\infty)$ be the scale function. The empirical loss $\widehat{\mathcal{L}}_n$ is close to its expected value $\mathcal{L}$, such that their absolute difference is proportional to the scale of the hypothesis $\boldsymbol{\theta}$. That is, with probability at least $1 - \delta$ over draws of n samples:*

$$(\forall \boldsymbol{\theta} \in \mathcal{H})\ |\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})| \leq \varepsilon_{n,\delta} c(\boldsymbol{\theta}) \tag{4}$$

where the rate $\varepsilon_{n,\delta}$ is nonincreasing with respect to n and δ. Furthermore, assume $\lim_{n\to+\infty} \varepsilon_{n,\delta} = 0$ for $\delta \in (0; 1)$.

In settings with a bounded complexity measure (e.g. empirical risk minimization with finite hypothesis class, VC dimension, Rademacher complexity), the *regular* uniform convergence statement is as follows $(\forall \boldsymbol{\theta} \in \mathcal{H})\ |\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})| \leq \varepsilon_{n,\delta}$. This condition is sufficient for loss consistency and regularization is unnecessary. This also occurs with a bounded hypothesis class $\mathcal{H}$, since in that case we can relax Assumption A to $(\forall \boldsymbol{\theta} \in \mathcal{H})\ |\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})| \leq \varepsilon_{n,\delta} \max_{\boldsymbol{\theta}\in\mathcal{H}} c(\boldsymbol{\theta})$.

2.2. Super-Scale Regularizers

Next, we define *super-scale regularizers*. That is, regularizers that are lower-bounded by a scale function.

Assumption B (Super-scale regularization)**.** *Let $c : \mathcal{H} \to [0; +\infty)$ be the scale function. Let $r : [0; +\infty) \to [0; +\infty)$ be a function such that:*

$$(\forall z \geq 0)\ z \leq r(z) \tag{5}$$

The regularizer $\mathcal{R}$ is bounded as follows:

$$(\forall \boldsymbol{\theta} \in \mathcal{H})\ r(c(\boldsymbol{\theta})) \leq \mathcal{R}(\boldsymbol{\theta}) < +\infty \tag{6}$$

Note that the above assumption implies that $c(\boldsymbol{\theta}) \leq \mathcal{R}(\boldsymbol{\theta})$. We opted to introduce the r function for clarity of presentation.

2.3. Norm-Loss Monotonicity

Finally, we state our last assumption of *norm-loss monotonicity*. For clarity of presentation, we use the ℓ_∞-norm in the following assumption. (The use of other norm that upper-bounds the ℓ_∞-norm, modifies the norm consistency result in Theorem 2 with respect to the new norm, and leaves the sparsistency and sign consistency result in Theorem 3 unchanged.)

Assumption C (Norm-Loss monotonicity)**.** *Let $b : [0; +\infty) \to [0; +\infty)$ be a nondecreasing function such that $b(0) = 0$. The expected loss $\mathcal{L}$ around the true hypothesis $\boldsymbol{\theta}^*$ is lower-bounded as follows:*

$$(\forall \boldsymbol{\theta} \in \mathcal{H})\ b(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_\infty) \leq \mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*) \tag{7}$$

Furthermore, we define the inverse of function b as:

$$b^\dagger(z) = \max_{b(z')=z} z' \tag{8}$$

As we discuss later, nonsmooth strongly convex functions, "minimax bounded" functions (which includes some convex functions and all strictly convex functions) and some family of nonconvex functions fulfill this assumption.

Note that Assumption C is with respect to the *expected* loss and therefore it holds in high dimensional spaces (e.g. for domains with nonzero Lebesgue measure). On the other hand, it is trivial to provide instances where strong convexity with respect to the *empirical* loss does not hold in high dimensions, as shown by (Negahban et al., 2009; 2012; Bickel et al., 2009) in the context of linear regression.

3. Theoretical Results

3.1. Loss Consistency

First, we provide a worst-case guarantee of the difference between the expected loss of the ξ-approximate empirical minimizer $\widehat{\boldsymbol{\theta}}_n$ and that of the true hypothesis $\boldsymbol{\theta}^*$.

Theorem 1 (Loss consistency)**.** *Under Assumptions A and B,* regularized loss minimization *is loss-consistent. That is, for $\alpha \geq 1$ and $\lambda_n = \alpha\varepsilon_{n,\delta}$, with probability at least $1 - \delta$:*

$$\mathcal{L}(\widehat{\boldsymbol{\theta}}_n) - \mathcal{L}(\boldsymbol{\theta}^*) \leq \varepsilon_{n,\delta}(\alpha\mathcal{R}(\boldsymbol{\theta}^*) + c(\boldsymbol{\theta}^*)) + \xi \tag{9}$$

(See Appendix A for detailed proofs.)

Given the above result, one would be tempted to make $\mathcal{R}(\boldsymbol{\theta}) = c(\boldsymbol{\theta})$ in order to minimize the upper bound. In practice, the function $c(\cdot)$ is chosen in order to get a good rate $\varepsilon_{n,\delta}$ in Assumption A. The regularizer is chosen in order to obtain some desired structure in the empirical minimizer $\widehat{\boldsymbol{\theta}}_n$.

3.2. Norm Consistency

Here, we provide a worst-case guarantee of the distance between the ξ-approximate empirical minimizer $\widehat{\boldsymbol{\theta}}_n$ and the true hypothesis $\boldsymbol{\theta}^*$.

Theorem 2 (Norm consistency)**.** *Under Assumptions A, B and C,* regularized loss minimization *is norm-consistent. That is, for $\alpha \geq 1$ and $\lambda_n = \alpha\varepsilon_{n,\delta}$, with probability at least $1-\delta$:*

$$\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|_\infty \leq b^\dagger(\varepsilon_{n,\delta}(\alpha\mathcal{R}(\boldsymbol{\theta}^*) + c(\boldsymbol{\theta}^*)) + \xi) \quad (10)$$

As mentioned before, we use the ℓ_∞-norm for clarity of presentation. (The use of other norm that upper-bounds the ℓ_∞-norm in Assumption C, modifies Theorem 2 with respect to the new norm.)

3.3. Sparsistency and Sign Consistency

Next, we analyze the exact recovery of the sparsity pattern (i.e. support recovery or sparsistency) as well as the signs (i.e. sign consistency) of the true hypothesis $\boldsymbol{\theta}^*$, by using the ξ-approximate empirical minimizer $\widehat{\boldsymbol{\theta}}_n$ in order to infer these properties. A related problem is the estimation of the sparsity level of the empirical minimizer as analyzed in (Bickel et al., 2009; Kakade et al., 2010). Here, we are interested in the stronger guarantee of perfect recovery of the support and signs.

Our approach is to perform thresholding of the empirical minimizer. In the context of ℓ_1-regularized linear regression, thresholding has been previously used for obtaining sparsistency and sign consistency (Meinshausen & Yu, 2009; Zhou, 2009). (See Appendix B for additional discussion.)

Next, we formally define the support of a hypothesis and a thresholding operator. The *support* $\mathcal{S}$ of a hypothesis $\boldsymbol{\theta}$ is the set of its nonzero elements, i.e.:

$$\mathcal{S}(\boldsymbol{\theta}) = \{i \mid \theta_i \neq 0\} \quad (11)$$

A *hard-thresholding* operator $\mathbf{h} : \mathcal{H} \times \mathbb{R} \to \mathcal{H}$ converts to zero the elements of the hypothesis $\boldsymbol{\theta}$ that have absolute value smaller than a threshold τ. That is, for each i we have:

$$h_i(\boldsymbol{\theta}, \tau) = \theta_i 1[|\theta_i| > \tau] \quad (12)$$

In what follows, we state our sparsistency and sign consistency guarantees. The minimum absolute value of the entries in the support has been previously used in (Ravikumar et al., 2008; Tibshirani, 2011; Wainwright, 2009a;b; Zhou, 2009).

Theorem 3 (Sparsistency and sign consistency)**.** *Under Assumptions A, B and C,* regularized loss minimization *followed by hard-thresholding is sparsistent and sign-consistent. More formally, for $\alpha \geq 1$, $\lambda_n = \alpha\varepsilon_{n,\delta}$ and $\tau = b^\dagger(\varepsilon_{n,\delta}(\alpha\mathcal{R}(\boldsymbol{\theta}^*) + c(\boldsymbol{\theta}^*)) + \xi)$, the solution $\widetilde{\boldsymbol{\theta}} = \mathbf{h}(\widehat{\boldsymbol{\theta}}_n, \tau)$ has the same support and signs as the true hypothesis $\boldsymbol{\theta}^*$ provided that $\min_{i\in\mathcal{S}(\boldsymbol{\theta}^*)} |\theta_i^*| > 2\tau$. That is, with probability at least $1-\delta$:*

$$\mathcal{S}(\widetilde{\boldsymbol{\theta}}) = \mathcal{S}(\boldsymbol{\theta}^*)\,,\ (\forall i \in \mathcal{S}(\boldsymbol{\theta}^*))\ \operatorname{sgn}(\widetilde{\theta}_i) = \operatorname{sgn}(\theta_i^*) \quad (13)$$

Our result also holds for the "approximately sparse" setting by constructing a thresholded version of the true hypothesis. That is, we guarantee correct sign recovery for all i such that $|\theta_i^*| > 2\tau$.

4. Examples

4.1. Losses with Scaled Uniform Convergence

First, we show that several problems in the literature have losses that fulfill Assumption A.

Maximum Likelihood Estimation for Exponential Family Distributions. First, we focus on the problem of learning exponential family distributions (Kakade et al., 2010). This includes for instance, the problem of learning the parameters (and possibly structure) of Gaussian and discrete MRFs. While the results in (Kakade et al., 2010; Ravikumar et al., 2008) concentrate on ℓ_1-norm regularization, here we analyze arbitrary norms.

Claim i (MLE for exponential family)**.** *Let $\mathbf{t}(\mathbf{x})$ be the sufficient statistics and $\mathcal{Z}(\boldsymbol{\theta}) = \int_{\mathbf{x}} e^{\langle \mathbf{t}(\mathbf{x}), \boldsymbol{\theta}\rangle}$ be the partition function. Given n i.i.d. samples, let $\widehat{\mathbf{T}}_n = \frac{1}{n}\sum_i \mathbf{t}(\mathbf{x}^{(i)})$ and $\mathbf{T} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{t}(\mathbf{x})]$ be the empirical and expected sufficient statistics, respectively. Let $\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) = -\langle \widehat{\mathbf{T}}_n, \boldsymbol{\theta}\rangle + \log \mathcal{Z}(\boldsymbol{\theta})$ and $\mathcal{L}(\boldsymbol{\theta}) = -\langle \mathbf{T}, \boldsymbol{\theta}\rangle + \log \mathcal{Z}(\boldsymbol{\theta})$ be the empirical and expected negative log-likelihood, respectively. Assumption A holds with probability at least $1-\delta$, scale function $c(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|$ and rate $\varepsilon_{n,\delta}$, provided that the dual norm fulfills $\|\widehat{\mathbf{T}}_n - \mathbf{T}\|_* \leq \varepsilon_{n,\delta}$.*

For sub-Gaussian $\mathbf{t}(\mathbf{x})$, we can obtain a rate $\varepsilon_{n,\delta} \in \mathcal{O}(\sqrt{1/n \log 1/\delta})$ for n independent samples. While for finite variance, we can obtain a rate $\varepsilon_{n,\delta} \in \mathcal{O}(\sqrt{1/(n\delta)})$. (See Appendix D.)

Generalized Linear Models. We focus on generalized linear models, which generalizes linear regression when *Gaussian* noise is assumed. This also includes for instance, logistic regression and compressed sensing with exponential-family noise (Rish & Grabarnik, 2009). For simplicity, we chose to analyze the fixed design model. That is, we analyze the case in which y is a random variable and $\mathbf{x}$ is a constant.

Claim ii (GLM with fixed design)**.** *Let $t(y)$ be the sufficient statistics and $\mathcal{Z}(\nu) = \int_y e^{t(y)\nu}$ be the partition function. Given n independent samples, let $\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_i -t(y^{(i)})\langle \mathbf{x}^{(i)}, \boldsymbol{\theta}\rangle + \log \mathcal{Z}(\langle \mathbf{x}^{(i)}, \boldsymbol{\theta}\rangle)$ be the empirical negative log-likelihood of $y^{(i)}$ given their linear predictors $\langle \mathbf{x}^{(i)}, \boldsymbol{\theta}\rangle$. Let $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(\forall i)\ y^{(i)}\sim\mathcal{D}_i}[\widehat{\mathcal{L}}_n(\boldsymbol{\theta})]$. Assumption A holds with probability at least $1-\delta$, scale function $c(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|$ and rate $\varepsilon_{n,\delta}$, provided that the dual norm fulfills $\|\frac{1}{n}\sum_i (t(y^{(i)}) - \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)])\mathbf{x}^{(i)}\|_* \le \varepsilon_{n,\delta}$.*

For sub-Gaussian $t(y)$, we can obtain a rate $\varepsilon_{n,\delta} \in \mathcal{O}(\sqrt{1/n \log 1/\delta})$ for n independent samples. While for finite variance, we can obtain a rate $\varepsilon_{n,\delta} \in \mathcal{O}(\sqrt{1/(n\delta)})$. Both cases hold for bounded $\|\mathbf{x}\|_*$. (See Appendix D.)

Matrix Factorization. We focus on two problems: exponential-family PCA and max-margin matrix factorization. We assume that the hypothesis $\boldsymbol{\theta}$ is a matrix. That is, $\boldsymbol{\theta} \in \mathcal{H} = \mathbb{R}^{n_1\times n_2}$. We assume that each entry in the random matrix $\mathbf{X} \in \mathbb{R}^{n_1\times n_2}$ is independent, and might follow a different distribution. Additionally, we assume that the matrix size grows with n. That is, we let $n = n_1 n_2$.

First, we analyze exponential-family PCA, which was introduced by (Collins et al., 2001) as a generalization of the more common *Gaussian* PCA.

Claim iii (Exponential-family PCA)**.** *Let $t(y)$ be the sufficient statistics and $\mathcal{Z}(\nu) = \int_y e^{t(y)\nu}$ be the partition function. Assume the entries of the random matrix $\mathbf{X} \in \mathbb{R}^{n_1\times n_2}$ are independent. Let $n = n_1 n_2$ and let $\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{ij} -t(x_{ij})\theta_{ij} + \log \mathcal{Z}(\theta_{ij})$ be the empirical negative log-likelihood of x_{ij} given θ_{ij}. Let $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(\forall ij)\ x_{ij}\sim\mathcal{D}_{ij}}[\widehat{\mathcal{L}}_n(\boldsymbol{\theta})]$. Assumption A holds with probability at least $1-\delta$, scale function $c(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|$ and rate $\varepsilon_{n,\delta}$, provided that the dual norm fulfills $\|\frac{1}{n}(t(x_{11}) - \mathbb{E}_{x\sim\mathcal{D}_{11}}[t(x)], \ldots, t(x_{n_1 n_2}) - \mathbb{E}_{x\sim\mathcal{D}_{n_1 n_2}}[t(x)])\|_* \le \varepsilon_{n,\delta}$.*

For sub-Gaussian $t(x_{ij})$, we can obtain a rate $\varepsilon_{n,\delta} \in \mathcal{O}(\sqrt{\log n}/n\sqrt{\log 1/\delta})$ for n independent entries. While for finite variance, we can obtain a rate $\varepsilon_{n,\delta} \in \mathcal{O}(\sqrt{1/(n\delta)})$. (See Appendix D.)

Next, we focus on max-margin matrix factorization. This problem was introduced by (Srebro et al., 2004) which used a hinge loss. We analyze the more general case of Lipschitz continuous functions, which also includes for instance, the logistic loss. Note however that the following claim also applies to nonconvex Lipschitz losses.

Claim iv (Max-margin factorization with Lipschitz loss)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a Lipschitz continuous loss function. Assume the entries of the random matrix $\mathbf{X} \in \{-1,+1\}^{n_1\times n_2}$ are independent. Let $n = n_1 n_2$ and let $\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{ij} f(x_{ij}\theta_{ij})$ be the empirical risk of predicting the binary values $x_{ij} \in \{-1,+1\}$ by using $\operatorname{sgn}(\theta_{ij})$. Let $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(\forall ij)\ x_{ij}\sim\mathcal{D}_{ij}}[\widehat{\mathcal{L}}_n(\boldsymbol{\theta})]$. Assumption A holds with probability 1 (i.e. $\delta = 0$), scale function $c(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$ and rate $\varepsilon_{n,0} \in \mathcal{O}(1/n)$.*

By using norm inequalities, Assumption A holds with probability 1 for other matrix norms besides ℓ_1.

Nonparametric Generalized Regression. Next, we analyze nonparametric regression with exponential-family noise. The goal is to learn a function, thus the hypothesis class $\mathcal{H}$ is a function space. Each function is represented in an infinite dimensional orthonormal basis. One instance of this problem is the Gaussian case, with orthonormal basis functions that depend on single coordinates, and a ℓ_1-norm prior as in (Ravikumar et al., 2005). In our nonparametric model, we allow for the number of basis functions to grow with more samples. For simplicity, we chose to analyze the fixed design model. That is, we analyze the case in which y is a random variable and $\mathbf{x}$ is a constant.

Claim v (Nonparametric regression)**.** *Let $\mathcal{X}$ be the domain of $\mathbf{x}$. Let $\theta : \mathcal{X} \to \mathbb{R}$ be a predictor. Let $t(y)$ be the sufficient statistics and $\mathcal{Z}(\nu) = \int_y e^{t(y)\nu}$ be the partition function. Given n independent samples, let $\widehat{\mathcal{L}}_n(\theta) = \frac{1}{n}\sum_i -t(y^{(i)})\theta(\mathbf{x}^{(i)}) + \log \mathcal{Z}(\theta(\mathbf{x}^{(i)}))$ be the empirical negative log-likelihood of $y^{(i)}$ given their predictors $\theta(\mathbf{x}^{(i)})$. Let $\mathcal{L}(\theta) = \mathbb{E}_{(\forall i)\ y^{(i)}\sim\mathcal{D}_i}[\widehat{\mathcal{L}}_n(\theta)]$. Let $\psi_1, \ldots, \psi_\infty : \mathcal{X} \to \mathbb{R}$ be an infinitely dimensional orthonormal basis, and let $\boldsymbol{\psi}(\mathbf{x}) = (\psi_1(\mathbf{x}), \ldots, \psi_\infty(\mathbf{x}))$. Assumption A holds with probability at least $1-\delta$, scale function $c(\theta) = \|\theta\|$ and rate $\varepsilon_{n,\delta}$, provided that the dual norm fulfills $\|\frac{1}{n}\sum_i (t(y^{(i)}) - \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)])\boldsymbol{\psi}(\mathbf{x}^{(i)})\|_* \le \varepsilon_{n,\delta}$.*

Let $\gamma \in (0; 1/2)$. For sub-Gaussian $t(y)$, we can obtain a rate $\varepsilon_{n,\delta} \in \mathcal{O}((1/n^{1/2-\gamma})\sqrt{\log 1/\delta})$ for n independent samples and $O(e^{n^{2\gamma}})$ basis functions. While for finite variance, we can obtain a rate $\varepsilon_{n,\delta} \in \mathcal{O}((1/n^{1/2-\gamma})\sqrt{1/\delta})$ for $O(n^{2\gamma})$ basis functions. Both cases hold for bounded $\|\boldsymbol{\psi}(\mathbf{x})\|_*$. (See Appendix D.)

Nonparametric Clustering with Exponential Families. We consider a version of the clustering problem, where the number of clusters is not fixed (and possibly infinite), and where the goal is to estimate the exponential-family parameters of each cluster. Thus, the hypothesis class $\mathcal{H}$ is an infinite dimensional vector space. An analysis for the Gaussian case, with fixed covariances and number of clusters (i.e. k-means) was given by (Sun et al., 2012). In our nonparametric model, we allow for the number of clusters to grow with more samples. For simplicity, we chose to analyze the case of "balanced" clusters. That is, each cluster contains the same number of training samples.

Claim vi (Nonparametric clustering)**.** *Let $\boldsymbol{\theta}^{(j)}$ be the pa-*

rameters of cluster j. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(\infty)})$ be the concatenation of an infinite set of clusters. Let $\mathbf{t}(\mathbf{x})$ be the sufficient statistics and $\mathcal{Z}(\boldsymbol{\nu}) = \int_{\mathbf{x}} e^{\langle \mathbf{t}(\mathbf{x}), \boldsymbol{\nu} \rangle}$ be the partition function. Given n i.i.d. samples, let $\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_i \min_j -\langle \mathbf{t}(\mathbf{x}^{(i)}), \boldsymbol{\theta}^{(j)} \rangle + \log \mathcal{Z}(\boldsymbol{\theta}^{(j)})$ be the empirical negative log-likelihood of $\mathbf{x}^{(i)}$ on its assigned cluster. Let $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[\widehat{\mathcal{L}}_n(\boldsymbol{\theta})]$. Let $\mathcal{X}$ be the domain of $\mathbf{x}$. Assumption A holds with probability at least $1 - \delta$, scale function $c(\boldsymbol{\theta}) = \sum_{j=1}^{\infty} \|\boldsymbol{\theta}^{(j)}\|$ and rate $\varepsilon_{n,\delta}$, provided that for all partitions $\mathcal{X}^{(1)}, \dots, \mathcal{X}^{(\infty)}$ of $\mathcal{X}$, the dual norm fulfills $(\forall j)$ $\| \frac{1}{n} \sum_i 1[\mathbf{x}^{(i)} \in \mathcal{X}^{(j)}] \mathbf{t}(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[1[\mathbf{x} \in \mathcal{X}^{(j)}] \mathbf{t}(\mathbf{x})] \|_* \leq \varepsilon_{n,\delta}$.

For sub-Gaussian $\mathbf{t}(\mathbf{x})$, we can obtain a rate $\varepsilon_{n,\delta} \in \mathcal{O}(\sqrt{\log n / n \log 1/\delta})$ for n independent samples and $O(\sqrt{n})$ clusters. For finite variance, we were not able to obtain a decreasing rate $\varepsilon_{n,\delta}$ with respect to n. (See Appendix D.)

PAC-Bayes Learning. In the PAC-Bayes framework, θ is a probability distribution of predictors f in a hypothesis class $\mathcal{F}$. Thus, the hypothesis class $\mathcal{H}$ is the space of probability distributions of support $\mathcal{F}$. After observing a training set, the task is to choose a posterior distribution $\widehat{\theta}_n$. PAC-Bayes guarantees are then given with respect to a prior distribution $\theta^{(0)}$. Next, we show a connection between PAC-Bayes learning and Kullback-Leibler regularization (Bousquet & Elisseeff, 2002; Germain et al., 2009). The following theorem applies to tasks such as classification as well as structured prediction.

Claim vii (PAC-Bayes learning)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be the domain of $\mathbf{x}$ and y respectively. Let $f : \mathcal{X} \to \mathcal{Y}$ be a predictor and $d : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ be a distortion function. Let θ be a probability distribution of predictors. Given n i.i.d. samples, let $\widehat{\mathcal{L}}_n(\theta) = \frac{1}{n} \sum_i \mathbb{E}_{f \sim \theta}[d(y^{(i)}, f(\mathbf{x}^{(i)}))]$ be the empirical risk of predicting $y^{(i)}$ by using the Gibbs predictor $f(\mathbf{x}^{(i)})$. Let $\mathcal{L}(\theta) = \mathbb{E}_{(y,\mathbf{x}) \sim \mathcal{D}}[\widehat{\mathcal{L}}_n(\theta)]$. Let $\theta^{(0)}$ be a prior distribution. Assumption A holds with probability at least $1 - \delta$, scale function $c(\theta) = KL(\theta || \theta^{(0)}) + 1$ and rate $\varepsilon_{n,\delta} \in \mathcal{O}(\sqrt{\log n / n \log 1/\delta})$.*

4.2. Super-Scale Regularizers

In what follows, we show that several regularizers commonly used in the literature fulfill Assumption B. We also provide yet unexplored priors with guarantees.

Norms. Norms regularizers (i.e. $\mathcal{R}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|$) fulfill Assumption B for $c(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|$ and $r(z) = z$. These regularizers include: the sparsity promoting regularizers, such as the ℓ_1-norm (e.g. Ravikumar et al. 2008) and the k-support norm (Argyriou et al., 2012), the multitask $\ell_{1,2}$ and $\ell_{1,\infty}$-norms for overlapping groups (Jacob et al., 2009; Mairal et al., 2010) as well as for non-overlapping groups (Negahban & Wainwright, 2011; Obozinski et al., 2011), and the trace norm for low-rank regularization (Bach, 2008; Srebro et al., 2004).

Functions of Norms. The Tikhonov regularizer (i.e. $\mathcal{R}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2 + 1/4$) fulfills Assumption B for $c(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2$ and $r(z) = z^2 + 1/4$. We can define some instances that have not been explored yet, but that have theoretical guarantees. Consider, for instance a polynomial bound $r(z) = z^\gamma - \gamma^{-\gamma/(\gamma-1)} + \gamma^{-1/(\gamma-1)}$ for $\gamma > 1$, a γ-insensitive bound $r(z) = \max(0, z - \gamma) + \gamma$, a logistic bound $r(z) = \log(1 + e^z)$, an exponential bound $r(z) = e^z - 1$, as well as an entropy bound $r(z) = z \log z + 1$.

Mixture of Norms. The *sparse and low-rank* prior (Richard et al., 2012) of the form $\mathcal{R}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1 + \|\boldsymbol{\theta}\|_{\text{tr}}$, fulfills Assumption B by making either $c(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$ or $c(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_{\text{tr}}$, and $r(z) = z$. The *elastic net* (Zou & Hastie, 2005) of the form $\mathcal{R}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1 + \|\boldsymbol{\theta}\|_2^2 + 1/4$, fulfills Assumption B by making $c(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$ and $r(z) = z$; or $c(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2$, $r(z) = z^2 + 1/4$.

Dirty Models. The *dirty* multitask prior (Jalali et al., 2010) of the form $\mathcal{R}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}^{(1)}\|_1 + \|\boldsymbol{\theta}^{(2)}\|_{1,\infty}$ where $\boldsymbol{\theta} = \boldsymbol{\theta}^{(1)} + \boldsymbol{\theta}^{(2)}$, fulfills Assumption B with $c(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_{1,\infty}$ and $r(z) = z$. (See Appendix C.)

Other Priors. The Kullback-Leibler regularizer (Bousquet & Elisseeff, 2002; Germain et al., 2009) fulfills Assumption B for $c(\theta) = KL(\theta || \theta^{(0)})$ and $r(z) = z$, where $\theta^{(0)}$ is a prior distribution. Any regularizer of the form $\mathcal{R}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\| + f(\boldsymbol{\theta})$ where $f(\boldsymbol{\theta}) \geq 0$, fulfills Assumption B with $c(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|$ and $r(z) = z$. This includes the mixture of norms, and the *total variation* prior (Kolar et al., 2010; 2009; Zhang & Wang, 2010). Since f is not required to be convex, quasiconvex regularizers of the form $\mathcal{R}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1 + \|\boldsymbol{\theta}\|_p$ for $p < 1$, fulfill Assumption B.

4.3. Norm-Loss Monotonicity

Here, we show some specific conditions on the expected loss in order fulfill Assumption C. We also provide yet unexplored cases with theoretical guarantees.

Strong Convexity. First, we show that strongly convex expected losses are a special case in our framework. We consider strongly convex functions that are not necessarily smooth.

Several authors have shown different flavors of strong convexity for specific problems. *Almost* strong convexity with respect to a small neighborhood around the true minimizer $\boldsymbol{\theta}^*$ was shown in (Kakade et al., 2010) for maximum likelihood estimation of exponential family distributions. Strong convexity of SVMs and logistic regression for Gaussian

Table 1. Rates for different losses and regularizers (See Appendix D for further details.)

Rates $\varepsilon_{n,\delta}$ for n samples, with probability at least $1-\delta$. We show dependence with respect to dimension, i.e. $\boldsymbol{\theta} \in \mathcal{H} = \mathbb{R}^p$. For exponential-family PCA and max-margin matrix factorization: $\boldsymbol{\theta} \in \mathcal{H} = \mathbb{R}^{n_1 \times n_2}$ where $n = n_1 n_2$. Rates were not optimized. All rates follow from the ℓ_1-norm regularizer and norm inequalities.
NA: not applicable, NG: no guarantees, $\gamma \in (0; 1/2)$.

	Sparsity (ℓ_1) Elastic net Total variation Sparsity and low-rank Quasiconvex ($\ell_1+\ell_p$, $p<1$)	Sparsity (k-support norm)	Tikhonov Multitask ($\ell_{1,\infty}$) Dirty multitask	Multitask ($\ell_{1,2}$)	Overlap multitask ($\ell_{1,2}$) g is maximum group size	Overlap multitask ($\ell_{1,\infty}$) g is maximum group size	Low-rank
MLE for exponential family: sub-Gaussian ($\sqrt{\log 1/\delta}$)	$\sqrt{\frac{\log p}{n}}$	$\sqrt{\frac{k \log p}{n}}$	$\sqrt{\frac{p \log p}{n}}$	$\frac{p^{1/4}\sqrt{\log p}}{\sqrt{n}}$	$\sqrt{\frac{g \log p}{n}}$	$\frac{g\sqrt{\log p}}{\sqrt{n}}$	$\sqrt{\frac{p \log p}{n}}$
Finite variance ($\sqrt{1/\delta}$)	$\sqrt{\frac{p}{n}}$	$\sqrt{\frac{kp}{n}}$	$\frac{p}{\sqrt{n}}$	$\frac{p^{3/4}}{\sqrt{n}}$	$\sqrt{\frac{gp}{n}}$	$\frac{g\sqrt{p}}{\sqrt{n}}$	$\frac{p}{\sqrt{n}}$
GLM with fixed design: sub-Gaussian ($\sqrt{\log 1/\delta}$)	$\sqrt{\frac{\log p}{n}}$	$\sqrt{\frac{k \log p}{n}}$	$\sqrt{\frac{p \log p}{n}}$	$\frac{p^{1/4}\sqrt{\log p}}{\sqrt{n}}$	$\sqrt{\frac{g \log p}{n}}$	$\frac{g\sqrt{\log p}}{\sqrt{n}}$	NA
Finite variance ($\sqrt{1/\delta}$)	$\sqrt{\frac{p}{n}}$	$\sqrt{\frac{kp}{n}}$	$\frac{p}{\sqrt{n}}$	$\frac{p^{3/4}}{\sqrt{n}}$	$\sqrt{\frac{gp}{n}}$	$\frac{g\sqrt{p}}{\sqrt{n}}$	NA
Exponential-family PCA: sub-Gaussian ($\sqrt{\log 1/\delta}$)	$\frac{\sqrt{\log n}}{n}$	NA	$\sqrt{\frac{\log n}{n}}$	$\frac{\sqrt{\log n}}{n^{3/4}}$	NA	NA	$\sqrt{\frac{\log n}{n}}$
Finite variance ($\sqrt{1/\delta}$)	$\frac{1}{\sqrt{n}}$	NA	NG	$\frac{1}{n^{1/4}}$	NA	NA	NG
Max-margin factorization with Lipschitz loss ($\delta=0$)	$\frac{1}{n}$	NA	$\frac{1}{\sqrt{n}}$	$\frac{1}{n^{3/4}}$	NA	NA	$\frac{1}{\sqrt{n}}$
Nonparametric regression: sub-Gaussian ($\sqrt{\log 1/\delta}$)	$\frac{\sqrt{\log p}}{n^{1/2-\gamma}}$	$\frac{\sqrt{k \log p}}{n^{1/2-\gamma}}$	$\frac{p\sqrt{\log p}}{n^{1/2-\gamma}}$	$\frac{\sqrt{p \log p}}{n^{1/2-\gamma}}$	$\frac{\sqrt{g \log p}}{n^{1/2-\gamma}}$	$\frac{g\sqrt{\log p}}{n^{1/2-\gamma}}$	NA
Finite variance ($\sqrt{1/\delta}$)	$\frac{\sqrt{p}}{n^{1/2-\gamma}}$	$\frac{\sqrt{kp}}{n^{1/2-\gamma}}$	$\frac{p^{3/2}}{n^{1/2-\gamma}}$	$\frac{p}{n^{1/2-\gamma}}$	$\frac{\sqrt{gp}}{n^{1/2-\gamma}}$	$\frac{g\sqrt{p}}{n^{1/2-\gamma}}$	NA
Nonparametric clustering: sub-Gaussian ($\sqrt{\log 1/\delta}$)	$\sqrt{\frac{\log np}{n}}$	$\sqrt{\frac{k \log np}{n}}$	$\frac{p\sqrt{\log np}}{\sqrt{n}}$	$\sqrt{\frac{p \log np}{n}}$	$\sqrt{\frac{g \log np}{n}}$	$\frac{g\sqrt{\log np}}{\sqrt{n}}$	$\sqrt{\frac{p \log np}{n}}$
PAC-Bayes learning ($\sqrt{\log 1/\delta}$)	Kullback-Leibler regularization $\sqrt{\frac{\log n}{n}}$						

predictors was proved in (Rocha et al., 2009). *Restricted* strong convexity (i.e. strong convexity with respect to a subset of directions) was shown in (Negahban et al., 2009; 2012) for generalized linear models under sparsity, group-sparsity and low-rank promoting regularizers. For simplicity, we focus on the regular form of strong convexity.

Claim viii (Strong convexity)**.** *Assumption C holds for $b(z) = \frac{\nu}{2} z^2$ provided that the expected loss $\mathcal{L}$ is strongly convex with parameter ν. Moreover, if $\mathcal{L}$ is twice continuously differentiable, Assumption C holds if the Hessian of $\mathcal{L}$ is positive definite, i.e. if there is $\nu > 0$ such that $(\forall \boldsymbol{\theta} \in \mathcal{H})\ \frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta}^2}(\boldsymbol{\theta}) \succeq \nu \mathbf{I}$.*

Note that the function $b(z) = \frac{\nu}{2} z^2$ is strictly increasing, and its inverse function is $b^\dagger(z) = \sqrt{2z/\nu}$. Furthermore, since $b^\dagger(0) = 0$, Theorem 2 guarantees exact recovery of the true hypothesis in the asymptotic case with exact minimization. That is, since we require $\lim_{n\to+\infty} \varepsilon_{n,\delta} = 0$ in Assumption A and for $\xi = 0$, we have $\lim_{n\to+\infty} \|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|_\infty = 0$.

The constant ν has a problem-specific meaning. In linear regression, ν is the minimum eigenvalue of the expected covariance matrix of the predictors (Wainwright, 2009b). In the estimation of Gaussian MRFs, ν is the squared minimum eigenvalue of the true covariance matrix (Ravikumar et al., 2008).

Minimax Boundedness. Next, we provide an approach for creating a "minimax" lower bound for an arbitrary expected loss. Note that we consider functions that are not necessarily smooth or convex. On the other hand, any strictly convex function is a "minimax bounded" function.

Our constructed lower bound resembles a cone without apex. First, we create a flat disk of a prescribed radius centered at the true hypothesis $\boldsymbol{\theta}^*$. Then, we create a linear lower bound (i.e. linear in $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2$) with minimum slope across the maximum over all possible directions. This linear function is the lower bound of the expected loss outside the flat disk region.

Claim ix (Minimax boundedness)**.** *Let $\nu_1 > 0$ be a fixed radius around the true hypothesis $\boldsymbol{\theta}^*$. Let $\mathcal{M}(\boldsymbol{\theta}) = \sup_{\gamma \geq 0}\{\gamma \mid \mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*) \geq \gamma(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 - \nu_1)\}$ be the maximum slope for a linear lower bound of the expected loss $\mathcal{L}$ in the direction of $\boldsymbol{\theta} - \boldsymbol{\theta}^*$. Let $\nu_2 = \inf_{\boldsymbol{\theta} \in \mathcal{H}, \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 > \nu_1} \mathcal{M}(\boldsymbol{\theta})$ be the "minimax" slope across all possible directions. Assumption C holds for $b(z) = \nu_2 \max(0, z - \nu_1)$ provided that $\nu_2 > 0$.*

Note that the function $b(z) = \nu_2 \max(0, z - \nu_1)$ is not strictly increasing for $z \in (0; \nu_1)$, and its inverse function is $b^\dagger(z) = \frac{z}{\nu_2} + \nu_1$. Furthermore, since $b^\dagger(0) = \nu_1$, Theorem 2 only guarantees recovery of the true hypothesis up to a small region in the asymptotic case, even with exact optimization. That is, for $\lim_{n\to+\infty} \varepsilon_{n,\delta} = 0$ and $\xi = 0$, we have $\lim_{n\to+\infty} \|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|_\infty = \nu_1$.

Other Types of Nonconvexity. By using our previous results, one can devise some yet unexplored settings for which our framework provides theoretical guarantees. Functions such as the square root (i.e. $b(z) = \sqrt{z}$) and the logarithm (i.e. $b(z) = \log(1 + z)$) can be used for nonconvex problems.

We construct a lower bound of the expected loss as follows. First, we define a "transformed" expected loss $\widetilde{\mathcal{L}}(\boldsymbol{\theta}) = b^{\dagger}(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*))$. Then, we invoke strong convexity (Claim viii) or minimax boundedness (Claim ix) for the "transformed" expected loss $\widetilde{\mathcal{L}}$. Thus, by using the square root function (i.e. $b(z) = \sqrt{z}$ and $b^{\dagger}(z) = z^2$), we define the family of "squared strongly convex" and "squared minimax bounded" functions. By using the logarithmic function (i.e. $b(z) = \log(1 + z)$ and $b^{\dagger}(z) = e^z - 1$), we define the "exponential strongly convex" and "exponential minimax bounded" functions. As expected, in order to obtain theoretical guarantees, scaled uniform convergence has to be shown with respect to the "transformed" expected loss $\widetilde{\mathcal{L}}$.

4.4. Rates and Novel Results

Table 1 shows the rates for the different losses and regularizers. We have not optimized these rates. All the rates follow mainly from the ℓ_1-norm regularizer and norm inequalities. Thus, while we match some rates in the literature, some are not better. Note that our framework is more general and uses less assumptions than previous analyses for specific problems.

New results include the four types of consistency of MLE of exponential family distributions, and GLMs for other priors besides ℓ_1. We prove the four types of consistency of regularizers that are a norm plus a nonnegative function (elastic net, total variation, dirty models, sparsity and low-rank), of relatively new regularizers (k-support norm, multitask priors with overlapping groups), and of a proposed quasiconvex regularizer. The analysis of matrix factorization problems is novel and without the i.i.d assumption. Our analysis of max-margin matrix factorization does not assume convexity. We provide a new analysis of nonparametric models, such as generalized regression and clustering. All the problems above have unbounded hypothesis class and unbounded loss. Finally, we show a connection between PAC-Bayes learning and Kullback-Leibler regularization.

5. Concluding Remarks

There are several ways of extending this research. While we focused on the exact recovery of the entire sparsity pattern, approximate sparsistency should also be studied. The use of a surrogate loss and theoretical guarantees with respect to the original loss is a challenging open problem. Most consistency results, including ours, do not provide a data-dependent mechanism for setting λ_n. Extending the results on cross-validation for ℓ_1-penalized linear regression (Homrighausen & McDonald, 2013) is one of our future goals. We provided examples for the i.i.d. and the independent sampling settings. We plan to analyze examples for the non-i.i.d. setting as in (London et al., 2013; Mohri & Rostamizadeh, 2010).

References

Argyriou, A., Foygel, R., and Srebro, N. Sparse prediction with the k-support norm. *NIPS*, 2012.

Bach, F. Consistency of trace norm minimization. *JMLR*, 2008.

Bickel, P., Ritov, Y., and Tsybakov, A. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 2009.

Bousquet, O. and Elisseeff, A. Stability and generalization. *JMLR*, 2002.

Collins, M., Dasgupta, S., and Schapire, R. A generalization of principal component analysis to the exponential family. *NIPS*, 2001.

Germain, P., Lacasse, A., Laviolette, F., Marchand, M., and Shanian, S. From PAC-Bayes bounds to KL regularization. *NIPS*, 2009.

Homrighausen, D. and McDonald, D. The lasso, persistence, and cross-validation. *ICML*, 2013.

Hsu, D., Kakade, S., and Zhang, T. Random design analysis of ridge regression. *COLT*, 2012.

Jacob, L., Obozinski, G., and Vert, J. Group lasso with overlap and graph lasso. *NIPS*, 2009.

Jalali, A., Ravikumar, P., Sanghavi, S., and Ruan, C. A dirty model for multi-task learning. *NIPS*, 2010.

Kakade, S., Shamir, O., Sridharan, K., and Tewari, A. Learning exponential families in high-dimensions: Strong convexity and sparsity. *AISTATS*, 2010.

Kolar, M., Song, L., and Xing, E. Sparsistent learning of varying-coefficient models with structural changes. *NIPS*, 2009.

Kolar, M., Song, L., Ahmed, A., and Xing, E. Estimating time-varying networks. *Annals of Applied Statistics*, 2010.

Lee, J., Sun, Y., and Taylor, J. On model selection consistency of M-estimators with geometrically decomposable penalties. *NIPS*, 2013.

Loh, P. and Wainwright, M. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *NIPS*, 2013.

London, B., Huang, B., Taskar, B., and Getoor, L. Collective stability in structured prediction: Generalization from one example. *ICML*, 2013.

Mairal, J., Jenatton, R., Obozinski, G., and Bach, F. Network flow algorithms for structured sparsity. *NIPS*, 2010.

Maurer, A. A note on the PAC-Bayesian theorem. *ArXiv*, 2004.

Meinshausen, N. and Yu, B. Lasso-type recovery of sparse representations for high dimensional data. *Annals of Statistics*, 2009.

Mohri, M. and Rostamizadeh, A. Stability bounds for stationary φ-mixing and β-mixing processes. *JMLR*, 2010.

Mukherjee, S., Niyogi, P., Poggio, T., and Rifkin, R. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 2006.

Negahban, S. and Wainwright, M. Simultaneous support recovery in high dimensions: Benefits and perils of block ℓ_1/ℓ_∞-regularization. *IEEE Transactions on Information Theory*, 2011.

Negahban, S., Ravikumar, P., Wainwright, M., and Yu, B. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *NIPS*, 2009.

Negahban, S., Ravikumar, P., Wainwright, M., and Yu, B. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 2012.

Obozinski, G., Wainwright, M., and Jordan, M. Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, 2011.

Rakhlin, S., Mukherjee, S., and Poggio, T. Stability results in learning theory. *Analysis and Applications*, 2005.

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. Spam: Sparse additive models. *NIPS*, 2005.

Ravikumar, P., Raskutti, G., Wainwright, M., and Yu, B. Model selection in Gaussian graphical models: High-dimensional consistency of ℓ_1-regularized MLE. *NIPS*, 2008.

Richard, E., Savalle, P., and Vayatis, N. Estimation of simultaneously sparse and low rank matrices. *ICML*, 2012.

Rish, I. and Grabarnik, G. Sparse signal recovery with exponential-family noise. *Allerton*, 2009.

Rocha, G., Xing, W., and Yu, B. Asymptotic distribution and sparsistency for ℓ_1-penalized parametric M-estimators with applications to linear SVM and logistic regression. *TR0903, Indiana University*, 2009.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Learnability, stability and uniform convergence. *JMLR*, 2010.

Srebro, N., Rennie, J., and Jaakkola, T. Maximum-margin matrix factorization. *NIPS*, 2004.

Sun, W., Wang, J., and Fang, Y. Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electronic Journal of Statistics*, 2012.

Tibshirani, R. Regression shrinkage and selection via the lasso: a retrospective (comments from Bühlmann). *J. Royal Statistical Society*, 2011.

van de Geer, S. A survey on empirical risk minimization. *Oberwolfach Reports*, 2005.

van de Geer, S. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 2008.

Wainwright, M. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 2009a.

Wainwright, M. Sharp thresholds for high-dimensional and noisy sparsity recovery using constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 2009b.

Wainwright, M., Ravikumar, P., and Lafferty, J. High dimensional graphical model selection using ℓ_1-regularized logistic regression. *NIPS*, 2006.

Yang, E. and Ravikumar, P. Dirty statistical models. *NIPS*, 2013.

Yang, E., Tewari, A., and Ravikumar, P. On robust estimation of high dimensional generalized linear models. *IJCAI*, 2013.

Zhang, B. and Wang, Y. Learning structural changes of Gaussian graphical models in controlled experiments. *UAI*, 2010.

Zhao, P. and Yu, B. On model selection consistency of lasso. *JMLR*, 2006.

Zhou, S. Thresholding procedures for high dimensional variable selection and statistical estimation. *NIPS*, 2009.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *J. Royal Statistical Society*, 2005.

A. Detailed Proofs

In this section, we state the proofs of all the theorems and claims in our manuscript.

A.1. Proof of Theorem 1

Proof. By optimality of the empirical minimizer, we have:

$$\begin{aligned}\widehat{\mathcal{L}}_n(\widehat{\boldsymbol{\theta}}_n) + \lambda_n \mathcal{R}(\widehat{\boldsymbol{\theta}}_n) &\leq \widehat{\mathcal{L}}_n(\widehat{\boldsymbol{\theta}}_n^*) + \lambda_n \mathcal{R}(\widehat{\boldsymbol{\theta}}_n^*) + \xi \\ &\leq \widehat{\mathcal{L}}_n(\boldsymbol{\theta}^*) + \lambda_n \mathcal{R}(\boldsymbol{\theta}^*) + \xi\end{aligned}$$

or equivalently $\widehat{\mathcal{L}}_n(\widehat{\boldsymbol{\theta}}_n) - \widehat{\mathcal{L}}_n(\boldsymbol{\theta}^*) \leq -\lambda_n \mathcal{R}(\widehat{\boldsymbol{\theta}}_n) + \lambda_n \mathcal{R}(\boldsymbol{\theta}^*) + \xi$. By Assumption A and B, and by conveniently setting $\lambda_n = \alpha \varepsilon_{n,\delta}$ for $\alpha \geq 1$:

$$\begin{aligned}\mathcal{L}(\widehat{\boldsymbol{\theta}}_n) - \mathcal{L}(\boldsymbol{\theta}^*) &\leq \widehat{\mathcal{L}}_n(\widehat{\boldsymbol{\theta}}_n) - \widehat{\mathcal{L}}_n(\boldsymbol{\theta}^*) + \varepsilon_{n,\delta} c(\widehat{\boldsymbol{\theta}}_n) + \varepsilon_{n,\delta} c(\boldsymbol{\theta}^*) \\ &\leq -\lambda_n \mathcal{R}(\widehat{\boldsymbol{\theta}}_n) + \lambda_n \mathcal{R}(\boldsymbol{\theta}^*) + \varepsilon_{n,\delta} c(\widehat{\boldsymbol{\theta}}_n) + \varepsilon_{n,\delta} c(\boldsymbol{\theta}^*) + \xi \\ &\leq -\lambda_n r(c(\widehat{\boldsymbol{\theta}}_n)) + \lambda_n \mathcal{R}(\boldsymbol{\theta}^*) + \varepsilon_{n,\delta} c(\widehat{\boldsymbol{\theta}}_n) + \varepsilon_{n,\delta} c(\boldsymbol{\theta}^*) + \xi \\ &= \varepsilon_{n,\delta}(-\alpha r(c(\widehat{\boldsymbol{\theta}}_n)) + c(\widehat{\boldsymbol{\theta}}_n)) + \varepsilon_{n,\delta}(\alpha \mathcal{R}(\boldsymbol{\theta}^*) + c(\boldsymbol{\theta}^*)) + \xi \\ &\leq \varepsilon_{n,\delta}(-r(c(\widehat{\boldsymbol{\theta}}_n)) + c(\widehat{\boldsymbol{\theta}}_n)) + \varepsilon_{n,\delta}(\alpha \mathcal{R}(\boldsymbol{\theta}^*) + c(\boldsymbol{\theta}^*)) + \xi \\ &\leq \varepsilon_{n,\delta}(\alpha \mathcal{R}(\boldsymbol{\theta}^*) + c(\boldsymbol{\theta}^*)) + \xi\end{aligned}$$

□

A.2. Proof of Theorem 2

Proof. By Assumption C for $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_n$ and by Theorem 1, we have $b(\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|_\infty) \leq \mathcal{L}(\widehat{\boldsymbol{\theta}}_n) - \mathcal{L}(\boldsymbol{\theta}^*) \leq \varepsilon_{n,\delta}(\alpha \mathcal{R}(\boldsymbol{\theta}^*) + c(\boldsymbol{\theta}^*)) + \xi$. Since the function b is nondecreasing, its inverse function $b^\dagger$ (as defined in eq.(8)) exists and we prove our claim. □

A.3. Proof of Theorem 3

Proof. For clarity, we remove the dependence of $\widehat{\boldsymbol{\theta}}_n$ with respect to the number of samples n. That is, $\widehat{\boldsymbol{\theta}} \equiv \widehat{\boldsymbol{\theta}}_n$. By Theorem 2, we have $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \leq b^\dagger(\varepsilon_{n,\delta}(\alpha \mathcal{R}(\boldsymbol{\theta}^*) + c(\boldsymbol{\theta}^*)) + \xi) \equiv \tau$. Therefore, for all i we have $|\widehat{\theta}_i - \theta_i^*| \leq \tau$. Next, we analyze the three possible cases.

Case 1: $\theta_i^* = 0 \Rightarrow \widetilde{\theta}_i = 0$. Assume that $i \notin \mathcal{S}(\boldsymbol{\theta}^*)$. Since $\theta_i^* = 0$, we have $|\widehat{\theta}_i - \theta_i^*| = |\widehat{\theta}_i| \leq \tau$. Therefore, $\widetilde{\theta}_i = h_i(\widehat{\boldsymbol{\theta}}, \tau) = \widehat{\theta}_i 1[|\widehat{\theta}_i| > \tau] = 0$.

Case 2: $\theta_i^* > 2\tau \Rightarrow \widetilde{\theta}_i > \tau$. Assume that $i \in \mathcal{S}(\boldsymbol{\theta}^*)$ and $\theta_i^* > 2\tau$. Since $|\widehat{\theta}_i - \theta_i^*| \leq \tau$, we have $\widehat{\theta}_i \geq \theta_i^* - \tau > 2\tau - \tau = \tau$. Therefore, $\widetilde{\theta}_i = h_i(\widehat{\boldsymbol{\theta}}, \tau) = \widehat{\theta}_i 1[|\widehat{\theta}_i| > \tau] = \widehat{\theta}_i > \tau$.

Case 3: $\theta_i^* < -2\tau \Rightarrow \widetilde{\theta}_i < -\tau$. Assume that $i \in \mathcal{S}(\boldsymbol{\theta}^*)$ and $\theta_i^* < -2\tau$. Since $|\widehat{\theta}_i - \theta_i^*| \leq \tau$, we have $\widehat{\theta}_i \leq \theta_i^* + \tau < -2\tau + \tau = -\tau$. Therefore, $\widetilde{\theta}_i = h_i(\widehat{\boldsymbol{\theta}}, \tau) = \widehat{\theta}_i 1[|\widehat{\theta}_i| > \tau] = \widehat{\theta}_i < -\tau$. □

A.4. Proof of Claim i

Proof. Let $\|\cdot\|_*$ be the dual norm of $\|\cdot\|$. Note that $\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}) = -\langle \widehat{\mathbf{T}}_n - \mathbf{T}, \boldsymbol{\theta} \rangle$. By the generalized Cauchy-Schwarz inequality, we have:

$$\begin{aligned}(\forall \boldsymbol{\theta})\ |\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})| &= |\langle \widehat{\mathbf{T}}_n - \mathbf{T}, \boldsymbol{\theta} \rangle| \\ &\leq \|\widehat{\mathbf{T}}_n - \mathbf{T}\|_* \|\boldsymbol{\theta}\| \\ &\leq \varepsilon_{n,\delta} \|\boldsymbol{\theta}\|\end{aligned}$$

□

A.5. Proof of Claim ii

Proof. Let $\|\cdot\|_*$ be the dual norm of $\|\cdot\|$. Note that $\widehat{\mathcal{L}}_n(\boldsymbol{\theta})-\mathcal{L}(\boldsymbol{\theta}) = -(\frac{1}{n}\sum_i t(y^{(i)})\langle \mathbf{x}^{(i)}, \boldsymbol{\theta}\rangle - \frac{1}{n}\sum_i \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)]\langle \mathbf{x}^{(i)}, \boldsymbol{\theta}\rangle)$. By the generalized Cauchy-Schwarz inequality, we have:

$$\begin{aligned}(\forall \boldsymbol{\theta})\ |\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})| &= |\tfrac{1}{n}\textstyle\sum_i t(y^{(i)})\langle \mathbf{x}^{(i)}, \boldsymbol{\theta}\rangle - \tfrac{1}{n}\sum_i \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)]\langle \mathbf{x}^{(i)}, \boldsymbol{\theta}\rangle| \\ &= |\langle \tfrac{1}{n}\textstyle\sum_i (t(y^{(i)}) - \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)])\mathbf{x}^{(i)}, \boldsymbol{\theta}\rangle| \\ &\le \|\tfrac{1}{n}\textstyle\sum_i (t(y^{(i)}) - \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)])\mathbf{x}^{(i)}\|_*\|\boldsymbol{\theta}\| \\ &\le \varepsilon_{n,\delta}\|\boldsymbol{\theta}\|\end{aligned}$$

□

A.6. Proof of Claim iii

Proof. Let $\|\cdot\|_*$ be the dual norm of $\|\cdot\|$. Note that $\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}) = -(\frac{1}{n}\sum_{ij} t(x_{ij})\theta_{ij} - \frac{1}{n}\sum_{ij}\mathbb{E}_{x\sim\mathcal{D}_{ij}}[t(x)]\theta_{ij})$. By the generalized Cauchy-Schwarz inequality, we have:

$$\begin{aligned}(\forall \boldsymbol{\theta})\ |\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})| &= |\tfrac{1}{n}\textstyle\sum_{ij} t(x_{ij})\theta_{ij} - \tfrac{1}{n}\sum_{ij}\mathbb{E}_{x\sim\mathcal{D}_{ij}}[t(x)]\theta_{ij}| \\ &= |\tfrac{1}{n}\textstyle\sum_{ij} (t(x_{ij}) - \mathbb{E}_{x\sim\mathcal{D}_{ij}}[t(x)])\theta_{ij}| \\ &\le \|\tfrac{1}{n}(t(x_{11}) - \mathbb{E}_{x\sim\mathcal{D}_{11}}[t(x)], \ldots, t(x_{n_1n_2}) - \mathbb{E}_{x\sim\mathcal{D}_{n_1n_2}}[t(x)])\|_*\|\boldsymbol{\theta}\| \\ &\le \varepsilon_{n,\delta}\|\boldsymbol{\theta}\|\end{aligned}$$

□

A.7. Proof of Claim iv

Proof. Let K be the Lipschitz constant of f. Without loss of generality, assume that $f(0) = 0$ (this can be accomplished by adding a constant factor to f). Note that $\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{ij} f(x_{ij}\theta_{ij}) - \frac{1}{n}\sum_{ij}\mathbb{E}_{x\sim\mathcal{D}_{ij}}[f(x\theta_{ij})]$. Recall that $x_{ij}\in\{-1,+1\}$. We have:

$$\begin{aligned}(\forall \boldsymbol{\theta})\ |\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})| &= |\tfrac{1}{n}\textstyle\sum_{ij} (f(x_{ij}\theta_{ij}) - \mathbb{E}_{x\sim\mathcal{D}_{ij}}[f(x\theta_{ij})])| \\ &= |\tfrac{1}{n}\textstyle\sum_{ij} (1[x_{ij}=+1]f(\theta_{ij}) + 1[x_{ij}=-1]f(-\theta_{ij}) - \mathbb{P}_{x\sim\mathcal{D}_{ij}}[x=+1]f(\theta_{ij}) - \mathbb{P}_{x\sim\mathcal{D}_{ij}}[x=-1]f(-\theta_{ij}))| \\ &= |\tfrac{1}{n}\textstyle\sum_{ij} ((1[x_{ij}=+1] - \mathbb{P}_{x\sim\mathcal{D}_{ij}}[x=+1])f(\theta_{ij}) + (1[x_{ij}=-1] - \mathbb{P}_{x\sim\mathcal{D}_{ij}}[x=-1])f(-\theta_{ij}))| \\ &\le \tfrac{1}{n}\textstyle\sum_{ij} (|1[x_{ij}=+1] - \mathbb{P}_{x\sim\mathcal{D}_{ij}}[x=+1]|\ |f(\theta_{ij})| + |1[x_{ij}=-1] - \mathbb{P}_{x\sim\mathcal{D}_{ij}}[x=-1]|\ |f(-\theta_{ij})|) \\ &\le \tfrac{1}{n}\textstyle\sum_{ij} (K|\theta_{ij}| + K|\theta_{ij}|) \\ &= \tfrac{2K}{n}\|\boldsymbol{\theta}\|_1\end{aligned}$$

□

A.8. Proof of Claim v

Proof. First, we represent the function $\theta : \mathcal{X}\to\mathbb{R}$ by using the infinitely dimensional orthonormal basis. That is, $\theta(\mathbf{x}) = \sum_{j=1}^{\infty}\nu_j^{(\theta)}\psi_j(\mathbf{x}) = \langle \boldsymbol{\nu}^{(\theta)}, \boldsymbol{\psi}(\mathbf{x})\rangle$, where $\boldsymbol{\nu}^{(\theta)} = (\nu_1^{(\theta)}, \ldots, \nu_\infty^{(\theta)})$. In the latter, the superindex (θ) allows for associating the infinitely dimensional coefficient vector $\boldsymbol{\nu}$ with the original function θ. Then, we define the norm of the function θ with respect to the infinitely dimensional orthonormal basis. That is, $\|\theta\| = \|\boldsymbol{\nu}^{(\theta)}\|$.

Let $\|\cdot\|_*$ be the dual norm of $\|\cdot\|$. Note that $\widehat{\mathcal{L}}_n(\theta) - \mathcal{L}(\theta) = -(\frac{1}{n}\sum_i t(y^{(i)})\theta(\mathbf{x}^{(i)}) - \frac{1}{n}\sum_i \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)]\theta(\mathbf{x}^{(i)}))$. By the generalized Cauchy-Schwarz inequality, we have:

$$\begin{aligned}(\forall \theta)\ |\widehat{\mathcal{L}}_n(\theta) - \mathcal{L}(\theta)| &= |\tfrac{1}{n}\textstyle\sum_i t(y^{(i)})\theta(\mathbf{x}^{(i)}) - \tfrac{1}{n}\sum_i \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)]\theta(\mathbf{x}^{(i)})| \\ &= |\tfrac{1}{n}\textstyle\sum_i t(y^{(i)})\langle \boldsymbol{\psi}(\mathbf{x}^{(i)}), \boldsymbol{\nu}^{(\theta)}\rangle - \tfrac{1}{n}\sum_i \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)]\langle \boldsymbol{\psi}(\mathbf{x}^{(i)}), \boldsymbol{\nu}^{(\theta)}\rangle| \\ &= |\langle \tfrac{1}{n}\textstyle\sum_i (t(y^{(i)}) - \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)])\boldsymbol{\psi}(\mathbf{x}^{(i)}), \boldsymbol{\nu}^{(\theta)}\rangle| \\ &\le \|\tfrac{1}{n}\textstyle\sum_i (t(y^{(i)}) - \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)])\boldsymbol{\psi}(\mathbf{x}^{(i)})\|_*\|\boldsymbol{\nu}^{(\theta)}\| \\ &\le \varepsilon_{n,\delta}\|\theta\|\end{aligned}$$

□

A.9. Proof of Claim vi

Proof. Let $\|\cdot\|_*$ be the dual norm of $\|\cdot\|$. Let $\mathcal{C}^{(j,\boldsymbol{\theta})} = \{\mathbf{x} \in \mathcal{X} \mid j = \arg\min_k -\langle \mathbf{t}(\mathbf{x}), \boldsymbol{\theta}^{(k)} \rangle + \log \mathcal{Z}(\boldsymbol{\theta}^{(k)})\}$. Note that $\mathcal{C}^{(1,\boldsymbol{\theta})}, \ldots, \mathcal{C}^{(\infty,\boldsymbol{\theta})}$ define a partition of $\mathcal{X}$. We can rewrite the empirical loss as follows $\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) = \sum_j \frac{1}{n} \sum_i 1[\mathbf{x}^{(i)} \in \mathcal{C}^{(j,\boldsymbol{\theta})}](-\langle \mathbf{t}(\mathbf{x}^{(i)}), \boldsymbol{\theta}^{(j)} \rangle + \log \mathcal{Z}(\boldsymbol{\theta}^{(j)}))$. Similarly, the expected loss can be written as $\mathcal{L}(\boldsymbol{\theta}) = \sum_j \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[1[\mathbf{x} \in \mathcal{C}^{(j,\boldsymbol{\theta})}](-\langle \mathbf{t}(\mathbf{x}), \boldsymbol{\theta}^{(j)} \rangle + \log \mathcal{Z}(\boldsymbol{\theta}^{(j)}))]$.

By the generalized Cauchy-Schwarz inequality, we have:

$$\begin{aligned}
(\forall \boldsymbol{\theta})\ |\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})| &= |\sum_j (\tfrac{1}{n} \sum_i 1[\mathbf{x}^{(i)} \in \mathcal{C}^{(j,\boldsymbol{\theta})}] \langle \mathbf{t}(\mathbf{x}^{(i)}), \boldsymbol{\theta}^{(j)} \rangle - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[1[\mathbf{x} \in \mathcal{C}^{(j,\boldsymbol{\theta})}] \langle \mathbf{t}(\mathbf{x}), \boldsymbol{\theta}^{(j)} \rangle])| \\
&= |\sum_j \langle \tfrac{1}{n} \sum_i 1[\mathbf{x}^{(i)} \in \mathcal{C}^{(j,\boldsymbol{\theta})}] \mathbf{t}(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[1[\mathbf{x} \in \mathcal{C}^{(j,\boldsymbol{\theta})}] \mathbf{t}(\mathbf{x})], \boldsymbol{\theta}^{(j)} \rangle| \\
&\leq \sum_j |\langle \tfrac{1}{n} \sum_i 1[\mathbf{x}^{(i)} \in \mathcal{C}^{(j,\boldsymbol{\theta})}] \mathbf{t}(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[1[\mathbf{x} \in \mathcal{C}^{(j,\boldsymbol{\theta})}] \mathbf{t}(\mathbf{x})], \boldsymbol{\theta}^{(j)} \rangle| \\
&\leq \sum_j \|\tfrac{1}{n} \sum_i 1[\mathbf{x}^{(i)} \in \mathcal{C}^{(j,\boldsymbol{\theta})}] \mathbf{t}(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[1[\mathbf{x} \in \mathcal{C}^{(j,\boldsymbol{\theta})}] \mathbf{t}(\mathbf{x})]\|_* \|\boldsymbol{\theta}^{(j)}\|
\end{aligned}$$

By assumption, for all partitions $\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(\infty)}$ of $\mathcal{X}$, the dual norm fulfills $(\forall j)\ \|\frac{1}{n} \sum_i 1[\mathbf{x}^{(i)} \in \mathcal{X}^{(j)}] \mathbf{t}(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[1[\mathbf{x} \in \mathcal{X}^{(j)}] \mathbf{t}(\mathbf{x})]\|_* \leq \varepsilon_{n,\delta}$. Therefore:

$$(\forall \boldsymbol{\theta})\ |\widehat{\mathcal{L}}_n(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})| \leq \varepsilon_{n,\delta} \sum_j \|\boldsymbol{\theta}^{(j)}\|$$

□

A.10. Proof of Claim vii

Proof. By Pinsker's inequality and Theorem 5 of (Maurer, 2004) which assumes $n \geq 8$, we have:

$$\begin{aligned}
(\forall \theta)\ |\widehat{\mathcal{L}}_n(\theta) - \mathcal{L}(\theta)| &\leq \sqrt{\tfrac{1}{2}(\widehat{\mathcal{L}}_n(\theta) \log \tfrac{\widehat{\mathcal{L}}_n(\theta)}{\mathcal{L}(\theta)} + (1 - \widehat{\mathcal{L}}_n(\theta)) \log \tfrac{1-\widehat{\mathcal{L}}_n(\theta)}{1-\mathcal{L}(\theta)})} \\
&\leq \sqrt{\tfrac{1}{2n}(KL(\theta||\theta^{(0)}) + \log \tfrac{2\sqrt{n}}{\delta})} \\
&\leq \sqrt{\tfrac{1}{2n} \max(1, \log \tfrac{2\sqrt{n}}{\delta})} \sqrt{KL(\theta||\theta^{(0)}) + 1} \\
&\leq \sqrt{\tfrac{1}{2n} \log \tfrac{2\sqrt{n}}{\delta}} (KL(\theta||\theta^{(0)}) + 1)
\end{aligned}$$

□

A.11. Proof of Claim viii

Proof. The expected loss $\mathcal{L}$ is strongly convex with parameter ν if and only if:

$$(\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{H}, \mathbf{g} \in \tfrac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_1))\ \mathcal{L}(\boldsymbol{\theta}_2) - \mathcal{L}(\boldsymbol{\theta}_1) \geq \langle \mathbf{g}, \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1 \rangle + \tfrac{\nu}{2} \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_2^2$$

First, we set $\boldsymbol{\theta}_1 = \boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}$. Next, note that the subdifferential vanishes at $\boldsymbol{\theta}^*$, that is $\mathbf{0} \in \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}^*)$. Therefore:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*) &\geq \tfrac{\nu}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \\
&\geq \tfrac{\nu}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_\infty^2
\end{aligned}$$

which proves our first claim. Our second claim regarding twice continuously differentiable $\mathcal{L}$ is well-known in the calculus literature. □

A.12. Proof of Claim ix

Proof. Note that by the definition of $\mathcal{M}(\boldsymbol{\theta})$, we have:

$$(\forall \boldsymbol{\theta} \in \mathcal{H}, 0 \leq \gamma \leq \mathcal{M}(\boldsymbol{\theta}))\ \mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*) \geq \gamma(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 - \nu_1)$$

By the definition of ν_2, we have:

$$(\forall \boldsymbol{\theta} \in \mathcal{H})\ \nu_2 \leq \mathcal{M}(\boldsymbol{\theta})$$

By putting both statements together for $\gamma = \nu_2$, we have:

$$\begin{aligned}(\forall \boldsymbol{\theta} \in \mathcal{H})\ \mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*) &\geq \nu_2(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 - \nu_1) \\ &\geq \nu_2(\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_\infty - \nu_1)\end{aligned}$$

Since $(\forall \boldsymbol{\theta} \in \mathcal{H})\ \mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*) \geq 0$, we prove our claim. □

B. Discussion on Sparsistency

One way to prove sparsistency and sign consistency is to use the *primal-dual witness* method (Negahban & Wainwright, 2011; Obozinski et al., 2011; Ravikumar et al., 2008; Wainwright, 2009b; Wainwright et al., 2006). These results are specific to the given loss (linear regression (Negahban & Wainwright, 2011; Obozinski et al., 2011; Wainwright, 2009b), log-likelihood of Gaussian MRFs (Ravikumar et al., 2008), pseudo-likelihood of discrete MRFs (Wainwright et al., 2006)) as well as the specific regularizer (ℓ_1-norm (Ravikumar et al., 2008; Wainwright, 2009b; Wainwright et al., 2006), $\ell_{1,2}$-norm (Obozinski et al., 2011) and $\ell_{1,\infty}$-norm (Negahban & Wainwright, 2011)). Furthermore, due to nonuniqueness of the dual of the $\ell_{1,\infty}$-norm (Negahban & Wainwright, 2011), characterizing sign consistency by primal-dual arguments is difficult. In this paper, we prove sparsistency and sign consistency for general regularizers, besides the ℓ_1 and $\ell_{1,p}$ norms. Indeed, our results also hold for regularizers that are not norms.

Our approach is to perform thresholding of the empirical minimizer. In the context of ℓ_1-regularized linear regression, thresholding has been previously used for obtaining sparsistency and sign consistency (Meinshausen & Yu, 2009; Zhou, 2009). Note that the *primal-dual witness* method of (Negahban & Wainwright, 2011; Obozinski et al., 2011; Ravikumar et al., 2008; Wainwright, 2009b; Wainwright et al., 2006) applies only when mutual incoherence conditions hold. If such conditions are not met, sparsistency and sign consistency is not guaranteed, independently of the number of samples. In our two-step algorithm, the threshold decreases with respect to the amount of data samples. Potentially, the sparsity pattern of every true hypothesis can be recovered, even if mutual incoherence does not hold.

Seemingly contradictory results are shown in (Zhao & Yu, 2006) where mutual incoherence conditions are shown to be necessary and sufficient for ℓ_1-regularized linear regression. Note that here, we consider regularization followed by a thresholding step, which is not considered in (Zhao & Yu, 2006).

C. Dirty Multitask Prior

(Jalali et al., 2010) proposed a *dirty* multitask prior of the form $\mathcal{R}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}^{(1)}\|_1 + \|\boldsymbol{\theta}^{(2)}\|_{1,\infty}$ where $\boldsymbol{\theta} = \boldsymbol{\theta}^{(1)} + \boldsymbol{\theta}^{(2)}$. By the triangle inequality:

$$\begin{aligned}\|\boldsymbol{\theta}\|_{1,\infty} &= \|\boldsymbol{\theta}^{(1)} + \boldsymbol{\theta}^{(2)}\|_{1,\infty} \\ &\leq \|\boldsymbol{\theta}^{(1)}\|_{1,\infty} + \|\boldsymbol{\theta}^{(2)}\|_{1,\infty} \\ &\leq \|\boldsymbol{\theta}^{(1)}\|_1 + \|\boldsymbol{\theta}^{(2)}\|_{1,\infty} \\ &= \mathcal{R}(\boldsymbol{\theta})\end{aligned}$$

Thus, the *dirty* multitask prior fulfills Assumption B with $c(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_{1,\infty}$ and $r(z) = z$.

D. Specific Dimension-Dependent Rates $\varepsilon_{n,\delta}$

D.1. Claim i for the sub-Gaussian case and ℓ_1-norm

Let $\boldsymbol{\theta} \in \mathcal{H} = \mathbb{R}^p$. Let $\|\cdot\|_* = \|\cdot\|_\infty$ and $\|\cdot\| = \|\cdot\|_1$. Let $(\forall j)\ t_j(\mathbf{x})$ be sub-Gaussian with parameter σ. By the union bound, sub-Gaussianity and independence, we have $\mathbb{P}[(\exists j)\ |\frac{1}{n}\sum_i (t_j(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[t_j(\mathbf{x})])| > \varepsilon] \leq 2p\exp(-\frac{n\varepsilon^2}{2\sigma^2}) = \delta$. By solving for ε, we have $\varepsilon_{n,\delta} = \sigma\sqrt{2/n(\log p + \log 2/\delta)}$.

D.2. Claim i for the finite variance case and ℓ_1-norm

Let $\boldsymbol{\theta} \in \mathcal{H} = \mathbb{R}^p$. Let $\|\cdot\|_* = \|\cdot\|_\infty$ and $\|\cdot\| = \|\cdot\|_1$. Let $(\forall j)\ t_j(\mathbf{x})$ have variance at most σ^2. By the union bound and Chebyshev's inequality, we have $\mathbb{P}[(\exists j)\ |\frac{1}{n}\sum_i (t_j(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[t_j(\mathbf{x})])| > \varepsilon] \leq p\ \frac{\sigma^2}{n\varepsilon^2} = \delta$. By solving for ε, we have $\varepsilon_{n,\delta} = \sigma\sqrt{\frac{p}{n\delta}}$.

D.3. Claim ii for the sub-Gaussian case and ℓ_1-norm

Let $\boldsymbol{\theta} \in \mathcal{H} = \mathbb{R}^p$. Let $\|\cdot\|_* = \|\cdot\|_\infty$ and $\|\cdot\| = \|\cdot\|_1$. Let $(\forall \mathbf{x})\ \|\mathbf{x}\|_* \leq B$ and thus $(\forall ij)\ |x_j^{(i)}| \leq B$. Let $(\forall i \text{ and } y \sim \mathcal{D}_i)\ t(y)$ be sub-Gaussian with parameter σ. Therefore $(\forall i \text{ and } y \sim \mathcal{D}_i)\ t(y)x_j^{(i)}$ is sub-Gaussian with parameter σB. By the union bound, sub-Gaussianity and independence, we have $\mathbb{P}[(\exists j)\ |\frac{1}{n}\sum_i (t(y^{(i)}) - \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)])x_j^{(i)}| > \varepsilon] \leq 2p\exp(-\frac{n\varepsilon^2}{2(\sigma B)^2}) = \delta$. By solving for ε, we have $\varepsilon_{n,\delta} = \sigma B\sqrt{2/n(\log p + \log 2/\delta)}$.

D.4. Claim ii for the finite variance case and ℓ_1-norm

Let $\boldsymbol{\theta} \in \mathcal{H} = \mathbb{R}^p$. Let $\|\cdot\|_* = \|\cdot\|_\infty$ and $\|\cdot\| = \|\cdot\|_1$. Let $(\forall \mathbf{x})\ \|\mathbf{x}\|_* \leq B$ and thus $(\forall ij)\ |x_j^{(i)}| \leq B$. Let $(\forall i \text{ and } y \sim \mathcal{D}_i)\ t(y)$ have variance at most σ^2. Therefore $(\forall i \text{ and } y \sim \mathcal{D}_i)\ t(y)x_j^{(i)}$ has variance at most $(\sigma B)^2$. By the union bound and Chebyshev's inequality, we have $\mathbb{P}[(\exists j)\ |\frac{1}{n}\sum_i (t(y^{(i)}) - \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)])x_j^{(i)}| > \varepsilon] \leq p\ \frac{(\sigma B)^2}{n\varepsilon^2} = \delta$. By solving for ε, we have $\varepsilon_{n,\delta} = \sigma B\sqrt{\frac{p}{n\delta}}$.

D.5. Claim iii for the sub-Gaussian case and ℓ_1-norm

Recall $\boldsymbol{\theta} \in \mathcal{H} = \mathbb{R}^{n_1\times n_2}$ where $n = n_1 n_2$. Let $\|\cdot\|_* = \|\cdot\|_\infty$ and $\|\cdot\| = \|\cdot\|_1$. Let $(\forall ij \text{ and } x \sim \mathcal{D}_{ij})\ t(x)$ be sub-Gaussian with parameter σ. By the union bound, sub-Gaussianity and independence, we have $\mathbb{P}[(\exists ij)\ |t(x_{ij}) - \mathbb{E}_{x\sim\mathcal{D}_{ij}}[t(x)]| > n\varepsilon] \leq 2n\exp(-\frac{(n\varepsilon)^2}{2\sigma^2}) = \delta$. By solving for ε, we have $\varepsilon_{n,\delta} = \frac{\sigma}{n}\sqrt{2(\log n + \log 2/\delta)}$.

D.6. Claim iii for the finite variance case and ℓ_1-norm

Recall $\boldsymbol{\theta} \in \mathcal{H} = \mathbb{R}^{n_1\times n_2}$ where $n = n_1 n_2$. Let $\|\cdot\|_* = \|\cdot\|_\infty$ and $\|\cdot\| = \|\cdot\|_1$. Let $(\forall ij \text{ and } x \sim \mathcal{D}_{ij})\ t(x)$ have variance at most σ^2. By the union bound and Chebyshev's inequality, we have $\mathbb{P}[(\exists ij)\ |t(x_{ij}) - \mathbb{E}_{x\sim\mathcal{D}_{ij}}[t(x)]| > n\varepsilon] \leq n\ \frac{\sigma^2}{(n\varepsilon)^2} = \delta$. By solving for ε, we have $\varepsilon_{n,\delta} = \sigma/\sqrt{(n\delta)}$.

D.7. Claim v for the sub-Gaussian case and ℓ_1-norm

Let $\mathbf{x} \in \mathcal{X} = \mathbb{R}^p$. Let $\|\cdot\|_* = \|\cdot\|_\infty$ and $\|\cdot\| = \|\cdot\|_1$. Let $(\forall \mathbf{x})\ \|\boldsymbol{\psi}(\mathbf{x})\|_* \leq B$ and thus $(\forall ij)\ |\psi_j(\mathbf{x}^{(i)})| \leq B$. Let $(\forall i \text{ and } y \sim \mathcal{D}_i)\ t(y)$ be sub-Gaussian with parameter σ. Therefore $(\forall i \text{ and } y \sim \mathcal{D}_i)\ t(y)\psi_j(\mathbf{x}^{(i)})$ is sub-Gaussian with parameter σB. The complexity of our nonparametric model grows with more samples. Let q_n be increasing with respect to the number of samples n. Assume that we have q_n orthonormal basis functions $\varphi_1, \ldots, \varphi_{q_n} : \mathbb{R} \to \mathbb{R}$. With these bases, we define $q_n p$ orthonormal basis functions of the form $\psi_j(\mathbf{x}) = \varphi_k(x_l)$ for $j = 1, \ldots, q_n p$, $k = 1, \ldots, q_n$, $l = 1, \ldots, p$. By the union bound, sub-Gaussianity and independence, we have $\mathbb{P}[(\exists j)\ |\frac{1}{n}\sum_i (t(y^{(i)}) - \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)])\psi_j(\mathbf{x}^{(i)})| > \varepsilon] \leq 2q_n p\exp(-\frac{n\varepsilon^2}{2(\sigma B)^2}) = \delta$. By solving for ε, we have $\varepsilon_{n,\delta} = \sigma B\sqrt{2/n(\log p + \log q_n + \log 2/\delta)}$.

In Table 1, we set $q_n = e^{n^{2\gamma}}$ for $\gamma \in (0; 1/2)$, although other settings are possible for obtaining a decreasing rate $\varepsilon_{n,\delta}$ with respect to n.

D.8. Claim v for the finite variance case and ℓ_1-norm

Let $\mathbf{x} \in \mathcal{X} = \mathbb{R}^p$. Let $\|\cdot\|_* = \|\cdot\|_\infty$ and $\|\cdot\| = \|\cdot\|_1$. Let $(\forall \mathbf{x})\ \|\boldsymbol{\psi}(\mathbf{x})\|_* \leq B$ and thus $(\forall ij)\ |\psi_j(\mathbf{x}^{(i)})| \leq B$. Let $(\forall i \text{ and } y \sim \mathcal{D}_i)\ t(y)$ have variance at most σ^2. Therefore $(\forall i \text{ and } y \sim \mathcal{D}_i)\ t(y)\psi_j(\mathbf{x}^{(i)})$ has variance at most $(\sigma B)^2$. The complexity of our nonparametric model grows with more samples. Let q_n be increasing with respect to the number of samples n. Assume that we have q_n orthonormal basis functions $\varphi_1, \ldots, \varphi_{q_n} : \mathbb{R} \to \mathbb{R}$. With these bases, we define $q_n p$ orthonormal basis functions of the form $\psi_j(\mathbf{x}) = \varphi_k(x_l)$ for $j = 1, \ldots, q_n p$, $k = 1, \ldots, q_n$, $l = 1, \ldots, p$. By the union bound and Chebyshev's inequality, we have $\mathbb{P}[(\exists j)\ |\frac{1}{n}\sum_i (t(y^{(i)}) - \mathbb{E}_{y\sim\mathcal{D}_i}[t(y)])\psi_j(\mathbf{x}^{(i)})| > \varepsilon] \leq q_n p\ \frac{(\sigma B)^2}{n\varepsilon^2} = \delta$. By

solving for ε, we have $\varepsilon_{n,\delta} = \sigma B \sqrt{\frac{q_n p}{n\delta}}$.

In Table 1, we set $q_n = n^{2\gamma}$ for $\gamma \in (0; 1/2)$, although other settings are possible for obtaining a decreasing rate $\varepsilon_{n,\delta}$ with respect to n.

D.9. Claim vi for the sub-Gaussian case and ℓ_1-norm

In order to allow for proper estimation of the parameters $\boldsymbol{\theta}^{(j)} \in \mathbb{R}^p$ of each cluster, we assume that the hypothesis class $\mathcal{H}$ allows only for clusters containing the same number of training samples. The complexity of our nonparametric model grows with more samples. Let q_n be increasing with respect to the number of samples n. Assume that we have q_n clusters with n/q_n samples each. In order to show that for all partitions $\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(\infty)}$ of $\mathcal{X}$, the dual norm fulfills $(\forall j)\ \|\frac{1}{n}\sum_i 1[\mathbf{x}^{(i)} \in \mathcal{X}^{(j)}]\mathbf{t}(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[1[\mathbf{x} \in \mathcal{X}^{(j)}]\mathbf{t}(\mathbf{x})]\|_* \leq \varepsilon_{n,\delta}$, we will show concentration for all subsets of $\{1, \ldots, n\}$ with size n/q_n. That is:

$$(\forall \mathcal{C} \subseteq \{1, \ldots, n\}, |\mathcal{C}| = n/q_n)\ \|\tfrac{1}{n}\textstyle\sum_i 1[\mathbf{x}^{(i)} \in \mathcal{C}]\mathbf{t}(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[1[\mathbf{x} \in \mathcal{C}]\mathbf{t}(\mathbf{x})]\|_* \leq \varepsilon_{n,\delta}$$

Let $\|\cdot\|_* = \|\cdot\|_\infty$ and $\|\cdot\| = \|\cdot\|_1$. Let $(\forall j)\ t_j(\mathbf{x})$ be sub-Gaussian with parameter σ. By the union bound, sub-Gaussianity and independence, we have:

$$\begin{aligned}
\mathbb{P}[(\exists j, \mathcal{C})\ |\tfrac{1}{n/q_n}\textstyle\sum_i 1[\mathbf{x}^{(i)} \in \mathcal{C}]\mathbf{t}(\mathbf{x}^{(i)}) - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[1[\mathbf{x} \in \mathcal{C}]t_j(\mathbf{x})]| > \gamma] &\leq 2p\ \tbinom{n}{n/q_n} e^{-\frac{n/q_n\gamma^2}{2\sigma^2}} \\
&\leq 2p\ (q_n e)^{n/q_n} e^{-\frac{n/q_n\gamma^2}{2\sigma^2}} \\
&= \delta
\end{aligned}$$

By solving for γ, we have $\gamma = \sigma\sqrt{2(1 + \log q_n + \frac{q_n}{n}\log p + \frac{q_n}{n}\log 2/\delta)}$. Note that $\varepsilon_{n,\delta} = \gamma/q_n$ and by setting $q_n = \sqrt{n}$ we have:

$$\begin{aligned}
\varepsilon_{n,\delta} &= \sigma\sqrt{2(\tfrac{1+\log q_n}{q_n^2} + \tfrac{1}{nq_n}\log p + \tfrac{1}{nq_n}\log 2/\delta)} \\
&= \sigma\sqrt{2(\tfrac{1+\log\sqrt{n}}{n} + \tfrac{1}{n^{3/2}}\log p + \tfrac{1}{n^{3/2}}\log 2/\delta)}
\end{aligned}$$

D.10. Norm inequalities to extend results to other norms

- For the k-support norm $\|\cdot\|_k^{\text{sup}}$, we have $(\forall \boldsymbol{\theta} \in \mathbb{R}^p)\ \|\boldsymbol{\theta}\|_1 \leq \sqrt{k}\|\boldsymbol{\theta}\|_k^{\text{sup}}$.
- For the ℓ_2-norm $\|\cdot\|_2$, we have $(\forall \boldsymbol{\theta} \in \mathbb{R}^p)\ \|\boldsymbol{\theta}\|_1 \leq \sqrt{p}\|\boldsymbol{\theta}\|_2$.
- For the ℓ_∞-norm $\|\cdot\|_2$, we have $(\forall \boldsymbol{\theta} \in \mathbb{R}^p)\ \|\boldsymbol{\theta}\|_1 \leq p\|\boldsymbol{\theta}\|_\infty$.
- For the Frobenius norm $\|\cdot\|_{\mathfrak{F}}$, we have $(\forall \boldsymbol{\theta} \in \mathbb{R}^{\sqrt{p}\times\sqrt{p}})\ \|\boldsymbol{\theta}\|_1 \leq \sqrt{p}\|\boldsymbol{\theta}\|_{\mathfrak{F}}$.
- For the trace norm $\|\cdot\|_{\text{tr}}$, we have $(\forall \boldsymbol{\theta} \in \mathbb{R}^{\sqrt{p}\times\sqrt{p}})\ \|\boldsymbol{\theta}\|_1 \leq \sqrt{p}\|\boldsymbol{\theta}\|_{\text{tr}}$.
- For the $\ell_{1,2}$-norm $\|\cdot\|_{1,2}$, we have $(\forall \boldsymbol{\theta} \in \mathbb{R}^{\sqrt{p}\times\sqrt{p}})\ \|\boldsymbol{\theta}\|_1 \leq p^{1/4}\|\boldsymbol{\theta}\|_{1,2}$.
- For the $\ell_{1,\infty}$-norm $\|\cdot\|_{1,\infty}$, we have $(\forall \boldsymbol{\theta} \in \mathbb{R}^{\sqrt{p}\times\sqrt{p}})\ \|\boldsymbol{\theta}\|_1 \leq \sqrt{p}\|\boldsymbol{\theta}\|_{1,\infty}$.
- For the $\ell_{1,2}$-norm with overlapping groups $\|\cdot\|_{1,2}^{\text{ov}}$, we have $(\forall \boldsymbol{\theta} \in \mathbb{R}^p)\ \|\boldsymbol{\theta}\|_1 \leq \sqrt{g}\|\boldsymbol{\theta}\|_{1,2}^{\text{ov}}$ where g is the maximum group size. Additionally, we have $(\forall \boldsymbol{\theta} \in \mathbb{R}^p)\ \|\boldsymbol{\theta}\|_2 \leq \|\boldsymbol{\theta}\|_{1,2}^{\text{ov}}$.
- For the $\ell_{1,\infty}$-norm with overlapping groups $\|\cdot\|_{1,\infty}^{\text{ov}}$, we have $(\forall \boldsymbol{\theta} \in \mathbb{R}^p)\ \|\boldsymbol{\theta}\|_1 \leq g\|\boldsymbol{\theta}\|_{1,\infty}^{\text{ov}}$ where g is the maximum group size. Additionally, we have $(\forall \boldsymbol{\theta} \in \mathbb{R}^p)\ \|\boldsymbol{\theta}\|_{1,2}^{\text{ov}} \leq \sqrt{g}\|\boldsymbol{\theta}\|_{1,\infty}^{\text{ov}}$.