Submitted to the Annals of Statistics
arXiv: arXiv:1607.06534

THE LANDSCAPE OF EMPIRICAL RISK FOR NON-CONVEX LOSSES

By Song Mei[*,‡], Yu Bai[‡] and Andrea Montanari[†,‡]

Stanford University[‡]

Most high-dimensional estimation methods propose to minimize a cost function (empirical risk) that is a sum of losses associated to each data point (each example). In this paper we focus on the case of non-convex losses. Classical empirical process theory implies uniform convergence of the empirical (or sample) risk to the population risk. While –under additional assumptions– uniform convergence implies consistency of the resulting M-estimator, it does not ensure that the latter can be computed efficiently.

In order to capture the complexity of computing M-estimators, we study the landscape of the empirical risk, namely its stationary points and their properties. We establish uniform convergence of the gradient and Hessian of the empirical risk to their population counterparts, as soon as the number of samples becomes larger than the number of unknown parameters (modulo logarithmic factors). Consequently, good properties of the population risk can be carried to the empirical risk, and we are able to establish one-to-one correspondence of their stationary points. We demonstrate that in several problems such as non-convex binary classification, robust regression, and Gaussian mixture model, this result implies a complete characterization of the landscape of the empirical risk, and of the convergence properties of descent algorithms.

We extend our analysis to the very high-dimensional setting in which the number of parameters exceeds the number of samples, and provide a characterization of the empirical risk landscape under a nearly information-theoretically minimal condition. Namely, if the number of samples exceeds the sparsity of the parameters vector (modulo logarithmic factors), then a suitable uniform convergence result holds. We apply this result to non-convex binary classification and robust regression in very high-dimension.

1. Introduction .

M-estimation is arguably the most popular approach to high-dimensional estimation. Given data-points $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_n\}$, $\boldsymbol{z}_i \in \mathbb{R}^d$,

[*] Supported by Office of Technology Licensing Stanford Graduate Fellowship

[†] Supported by the NSF grant CCF-1319979

MSC 2010 subject classifications: Primary 62F10, 62J02; secondary 62H30

Keywords and phrases: non-convex optimization, empirical risk minimization, landscape, uniform convergence

we estimate a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$ via

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta} \in \Theta_{n,p}} \widehat{R}_n(\boldsymbol{\theta})\,, \tag{1.1}$$

$$\widehat{R}_n(\boldsymbol{\theta}) \equiv \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}; \boldsymbol{z}_i)\,. \tag{1.2}$$

Here $\ell : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ is a loss function, and $\Theta_{n,p}$ is a constraint set. Prominent examples of this general framework include maximum likelihood (ML) estimation [Fis22] and empirical risk minimization [Vap98].

Once the objective (1.1) is formed, it remains to define a computationally efficient scheme to approximate it. Gradient descent is the most frequently applied idea. Assuming –for the moment– $\Theta_{n,p} = \mathbb{R}^p$, this takes the form

$$\hat{\boldsymbol{\theta}}_n(k+1) = \hat{\boldsymbol{\theta}}_n(k) - h_k\, \nabla \widehat{R}_n(\hat{\boldsymbol{\theta}}_n(k))\,. \tag{1.3}$$

While a large number of variants and refinements have been developed over the years (projected gradient, accelerated gradient [Nes13b], stochastic gradient [RM51], distributed gradient [TBA84], and so on), these share many of the strengths and weaknesses of the elementary iteration (1.3).

If gradient descent is adopted, the only freedom is in the choice of the loss function $\ell(\,\cdot\,;\,\cdot\,)$. Convexity has been a major guiding principle in this respect. If the function $\ell(\,\cdot\,; \boldsymbol{z}) : \mathbb{R}^p \to \mathbb{R}$ is convex, then the empirical risk $\widehat{R}_n(\,\cdot\,)$ is convex as well and hence gradient descent is globally convergent to an M-estimator (the latter is unique under strict convexity). Also, strong convexity of $\widehat{R}_n(\,\cdot\,)$ can be used to prove optimal statistical guarantees for the M-estimator $\hat{\boldsymbol{\theta}}_n$. This line of thought can be traced back as far as Fisher's argument for the asymptotic efficiency of maximum likelihood estimators [Fis22, Fis25], and originated many beautiful contributions. In recent years, a flourishing line of research addresses the very high-dimensional regime $p \gg n$, by leveraging on suitable restricted strong convexity assumptions [CT05, CT07, BRT09, NRWY12].

Despite these successes, many problems of practical interest call for non-convex loss functions. Let us briefly mention a few examples of non-convex M-estimators that are often preferred by practitioners to their convex counterparts. We will revisit these examples in Section 4.

In binary linear classification we are given n pairs $\boldsymbol{z}_1 = (y_1, \boldsymbol{x}_1), \dots, \boldsymbol{z}_n = (y_n, \boldsymbol{x}_n)$ with $y_i \in \{0,1\}$, $\boldsymbol{x}_i \in \mathbb{R}^d$, and would like to learn a model of the form $\mathbb{P}(Y_i = 1 | \boldsymbol{X}_i = \boldsymbol{x}) = \sigma(\langle \boldsymbol{\theta}_0, \boldsymbol{x}\rangle)$ with $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ a parameter vector and $\sigma : \mathbb{R} \to [0,1]$ a threshold function. The non-linear square loss $\ell(\boldsymbol{\theta}; y, \boldsymbol{x}) =$

$(y-\sigma(\langle\boldsymbol{\theta},\boldsymbol{x}\rangle))^2$ is commonly used in practice

$$\widehat{R}_n(\boldsymbol{\theta}) \equiv \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \sigma(\langle\boldsymbol{\theta},\boldsymbol{x}_i\rangle)\Big)^2 . \tag{1.4}$$

Several empirical studies [CDT+09, WL12, NS13] demonstrate superior robustness and classification accuracy of non-convex losses in contrast to convex losses (e.g. hinge or logistic loss). The same loss function is commonly used used in neural-network models [LBH15].

A similar scenario arises in robust regression. In this case, we are given n pairs $\boldsymbol{z}_1 = (y_1,\boldsymbol{x}_1)$,. . . ,$\boldsymbol{z}_n = (y_n,\boldsymbol{x}_n)$ with $y_i \in \mathbb{R}$, $\boldsymbol{x}_i \in \mathbb{R}^d$, and we assume the linear model $y_i = \langle\boldsymbol{\theta}_0,\boldsymbol{x}_i\rangle + \varepsilon_i$, where the noise terms ε_i are i.i.d. with mean zero. Since Huber's seminal work [Hub73], M-estimators are the method of choice for this problem:

$$\widehat{R}_n(\boldsymbol{\theta}) \equiv \frac{1}{n}\sum_{i=1}^{n}\rho\big(y_i - \langle\boldsymbol{\theta},\boldsymbol{x}_i\rangle\big) . \tag{1.5}$$

Robustness naturally suggests to investigate the use of a non-convex function $\rho:\mathbb{R}\to\mathbb{R}$, either bounded or increasing slowly at infinity.

Finally, missing data problems famously lead to non-convex optimization formulations. Consider for instance a mixture-of-Gaussians problems in which we are given data points $\boldsymbol{z}_1,\dots,\boldsymbol{z}_n \in \mathbb{R}^d$, $\boldsymbol{z}_i \sim_{iid} \sum_{a=1}^{k} p_a \mathsf{N}(\boldsymbol{\theta}_a, \mathrm{I}_{d\times d})$ (for the sake of simplicity we assume identity covariance and known proportions). The maximum-likelihood problem requires to minimize[1]

$$\widehat{R}_n(\boldsymbol{\theta}) \equiv -\frac{1}{n}\sum_{i=1}^{n}\log\left(\sum_{a=1}^{k} p_a\,\phi_d\big(\boldsymbol{z}_i - \boldsymbol{\theta}_a\big)\right) , \tag{1.6}$$

with respect to the cluster centers $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_k) \in \mathbb{R}^{d\times k}$. Other examples include low-rank matrix completion [KOM09], phase retrieval [SQW16], tensor estimation problems [MR14], and so on.

M-estimation with non-convex loss functions $\ell(\,\cdot\,;\boldsymbol{z}) : \mathbb{R}^p \to \mathbb{R}$ is far less understood than in the convex case. Empirical process theory guarantees uniform convergence of the sample risk $\widehat{R}_n(\,\cdot\,)$ to the population risk $R(\boldsymbol{\theta}) \equiv \mathbb{E}[\widehat{R}_n(\boldsymbol{\theta})]$ [BLM13]. However, this does not provide a computationally practical scheme, since gradient descent can get stuck in stationary points that are not global minimizers.

[1] Here and below $\phi_d(\boldsymbol{x}) \equiv \exp\{-\|\boldsymbol{x}\|_2^2/2\}/(2\pi)^{d/2}$ denotes the d-dimensional standard Gaussian density.

In this paper, we present several general results on non-convex M-estimation and apply them to develop new analysis in each of the three problems mentioned above. We next overview our main results and the paper's organization, referring to Section 2 for a discussion of related work.

Uniform convergence of gradient and Hessian. We prove that, under technical conditions on the loss function $\ell(\,\cdot\,;\,\cdot\,)$, $\sup_{\boldsymbol{\theta}} \|\nabla \widehat{R}_n(\boldsymbol{\theta}) - \nabla R(\boldsymbol{\theta})\|_2 \lesssim \sqrt{p(\log n)/n}$ and $\sup_{\boldsymbol{\theta}} \|\nabla^2 \widehat{R}_n(\boldsymbol{\theta}) - \nabla^2 R(\boldsymbol{\theta})\|_{\rm op} \lesssim \sqrt{p(\log n)/n}$ (we use $\lesssim$ to hide constant factors). We refer to Section 3.1 for formal statements.

These results complement the classical analysis that implies uniform convergence of the risk itself, but allow us to control the behavior of stationary points. Note that they guarantee uniform convergence of the gradient and Hessian provided $n, p \to \infty$ with $p(\log p)/n \to 0$. Apart from logarithmic factors, this is the optimal condition.

(In this paper we will refer to the asymptotics $n, p \to \infty$ with n roughly of the same order as p as *high-dimensional regime*[2], to contrast it with the low-dimensional analysis for $n \gg p$. We will refer to the asymptotics $n \ll p$ under sparsity assumptions as *very high-dimensional regime*.)

Topology of the empirical risk. As an immediate consequence of the previous result, the structure of the empirical risk function $\boldsymbol{\theta} \mapsto \widehat{R}_n(\boldsymbol{\theta})$ is –in many cases– surprisingly simple. Recall that a Morse function is a twice differentiable function whose stationary points are non-degenerate (i.e. have an invertible Hessian). In particular, stationary points are isolated, and have a well-defined index. Assume that the population risk $R(\boldsymbol{\theta})$ is *strongly Morse* (i.e., at any stationary point $\boldsymbol{\theta}$, all the eigenvalues of the Hessian are bounded away from zero $|\lambda_i(\nabla^2 R(\boldsymbol{\theta}))| \geq \delta$). Then, for $n \gtrsim p \log p$, the stationary points of the empirical risk $\widehat{R}_n(\boldsymbol{\theta})$ are in one-to-one correspondence with those of the population risk and have the same index (minima correspond to minima, saddles to saddles, and so on). Weaker conditions ensure this correspondence for local minima alone.

Very high-dimensional regime. We then extend the above picture to the case in which the number of parameters p exceeds the number of samples n, under the assumption that the true parameter vector $\boldsymbol{\theta}_0$ is s_0-sparse. This setting is relevant to a large number of applications, ranging from genomics [PZB+10] to signal processing [Don06]. In order to promote sparse estimates, we study the following ℓ_1-regularized non-

[2] The specific asymptotics $n, p \to \infty$ with n/p converging to a constant is also known as 'Kolmogorov asymptotics' [Ser13].

convex problem, cf. Section 3.3:

$$\begin{aligned} &\text{minimize} \quad && \widehat{R}_n(\boldsymbol{\theta}) + \lambda_n \|\boldsymbol{\theta}\|_1 \,, \\ &\text{subject to} \quad && \|\boldsymbol{\theta}\|_2 \le r \,. \end{aligned} \tag{1.7}$$

We introduce a *generalized gradient linearity* condition on the loss function $\ell(\,\cdot\,,\,\cdot\,)$ and prove that – under this condition– the above problem has a unique local minimum for $n \gtrsim s_0 \log p$. Again this is a nearly optimal scaling since no consistent estimation is possible when $n \lesssim s_0$.

Applications. Given a particular M-estimation problem with a suitable statistical model, we combine the above results with an analysis of the population risk $R(\boldsymbol{\theta})$ to derive precise characterizations of the empirical risk. In Section 4 we demonstrate that this program can be carried out by studying the three problems outlined below:

1. *Binary linear classification.* We prove that, for[3] $n \gtrsim d \log d$, the empirical risk has a unique local minimum, that is also the global minimum. Further, gradient descent converges exponentially to this minimizer: $\|\hat{\boldsymbol{\theta}}_n(k) - \hat{\boldsymbol{\theta}}_n\|_2 \le C\|\hat{\boldsymbol{\theta}}(0) - \hat{\boldsymbol{\theta}}_n\|_2\,(1 - h/C)^k$, and enjoys nearly optimal estimation error guarantees: $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 \le C\sqrt{(d \log n)/n}$. If the true parameter $\boldsymbol{\theta}_0$ is s_0-sparse, for $n \gtrsim s_0 \log d$, the ℓ_1-regularized empirical risk has a unique local minimum, that is also the global minimum. The minimizer enjoys nearly optimal estimation error guarantees: $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 \le C\sqrt{(s_0 \log n)/n}$.
2. *Robust regression.* We establish similar results for the robust regression model, under technical assumptions on the loss function $\rho : \mathbb{R} \to \mathbb{R}$ and on the distribution of the noise ε_i. Namely, we prove that the empirical risk has a unique local minimum, that can be found efficiently via gradient descent, provided $n \gtrsim d \log d$. If the true parameter $\boldsymbol{\theta}_0$ is s_0-sparse, for $n \gtrsim s_0 \log d$, the ℓ_1-regularized empirical risk has a unique local minimum.
3. *Mixture of Gaussians.* We consider the special case of two Gaussians with equal proportions, i.e. $k = 2$ with $p_1 = p_2 = 1/2$. We prove that, for $n \gtrsim d \log d$, the empirical risk has two global minima that are related by exchange of the two Gaussian components $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$ and $(\hat{\boldsymbol{\theta}}_2, \hat{\boldsymbol{\theta}}_1)$, connected via saddle points. The trust region algorithm converges to one of these two minima when initial-

[3] Recall that, in this case, the number of parameters p is equal to the ambient dimension d.

ized at random. Also the two minima are within nearly optimal statistical errors from the true centers.

1.1. *Notations.* We use normal font for scalars (e.g. $a, b, c \dots$) and boldface for vectors ($\boldsymbol{x}, \boldsymbol{w}, \dots$). We will typically reserve capital letters for random variables (and capital bold for random vectors). Given $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^m$, their standard scalar product is denoted by $\langle \boldsymbol{u}, \boldsymbol{v} \rangle \equiv \sum_{i=1}^m u_i v_i$. The ℓ_p norm of a vector is –as usual– indicated by $\|\boldsymbol{x}\|_p$. The $m \times m$ identity matrix is denoted by $\mathrm{I}_{m\times m}$.

Given a matrix $\boldsymbol{M} \in \mathbb{R}^{m\times m}$, we denote by $\lambda_i(\boldsymbol{M})$, $i \in \{1, \dots, m\}$ its eigenvalues in decreasing order, and by $\|\boldsymbol{M}\|_{\rm op} = \max\{\lambda_1(\boldsymbol{M}), -\lambda_m(\boldsymbol{M})\}$ its operator norm. Finally, we shall occasionally consider third order tensors $\boldsymbol{T} \in \mathbb{R}^{m\times m\times m}$. In this case the operator (or injective) norm is defined as $\|\boldsymbol{T}\|_{\rm op} = \max\{|\langle \boldsymbol{T}, \boldsymbol{x}^{\otimes 3}\rangle| \ : \ \|\boldsymbol{x}\|_2 = 1\}$, where $\langle \boldsymbol{T}, \boldsymbol{x}^{\otimes 3}\rangle = \sum_{i,j,k} T_{ijk} x_i x_j x_k$.

We let $\mathsf{B}_q^d(\boldsymbol{a}, \rho) \equiv \{\boldsymbol{x} \in \mathbb{R}^d : \ \|\boldsymbol{x} - \boldsymbol{a}\|_q \le \rho\}$ be the ℓ_q ball in $\mathbb{R}^d$ with center $\boldsymbol{a}$ and radius ρ. We will often omit the dimension superscript d when clear from the context, the subscript q when $q = 2$, and the center $\boldsymbol{a}$ when $\boldsymbol{a} = \boldsymbol{0}$. In particular $\mathsf{B}(\rho)$ is the euclidean ball of radius ρ. For any set $D \subset \mathbb{R}^d$, we let ∂D be the boundary of the set.

We will generally use upper case letters for random variables and lower case for deterministic values (unless the latter are matrices).

2. Related literature.

While developing a theory on non-convex M-estimators is an outstanding challenge, several important facts are by now well understood thanks to a stream of beautiful works. We will provide a necessarily incomplete summary in the next paragraphs.

Uniform convergence of the empirical risk. Let $R(\boldsymbol{\theta}) = \mathbb{E}\widehat{R}_n(\boldsymbol{\theta})$ denote the population risk. Under mild conditions on the loss function ℓ and on the sample size, it is known that with high probability

$$\sup_{\boldsymbol{\theta} \in \Theta_{n,p}} \big|\widehat{R}_n(\boldsymbol{\theta}) - R(\boldsymbol{\theta})\big| \le \varepsilon_n \,, \tag{2.1}$$

for some small $\varepsilon_n \to 0$ [VdG00, BLM13]. This immediately implies guarantees for the M-estimator $\hat{\boldsymbol{\theta}}$ in ℓ-loss (or prediction error). Under additional conditions on the population risk $R(\boldsymbol{\theta})$, bounds in estimation error can be derived as well.

For general non-convex losses, uniform convergence results of the form (2.1) do not preclude the existence of multiple local minima of the sample risk $\widehat{R}_n(\boldsymbol{\theta})$. Hence, this theory does not provide –by itself– computationally practical methods to compute $\hat{\boldsymbol{\theta}}$.

Algorithmic convergence to the 'statistical neighborhood'. In general, gradient descent and other local optimization procedures are expected to converge to local minima of the empirical risk $\widehat{R}_n(\boldsymbol{\theta})$. In several cases, it is proved that every local minimizer $\hat{\boldsymbol{\theta}}^{\text{loc}}$ is 'statistically good'. More precisely, the estimation error (e.g. the ℓ_2 error $\|\hat{\boldsymbol{\theta}}^{\text{loc}} - \boldsymbol{\theta}_0\|_2$) is within a constant from the minimax rate for the problem at hand. Also, gradient descent converges to such a neighborhood of the true $\boldsymbol{\theta}_0$ within a small number of iterations. Results of this type have been proved, among others, for linear regression with noisy covariates [LW12], generalized linear models with non-convex regularizers [LW13], robust regression [LM13], and sparse regression [YWL$^+$15].

While these results are very important, they are not completely satisfactory. For instance, one natural question is whether the statistical error might be improved by finding a better local minimum. If, for instance, the estimation error could be improved by a factor 2 by finding a better local minimum, it would be worth in many applications to restart gradient descent at multiple initializations. Also, since convergence to a fixed point is not guaranteed, these approaches come without a clear stopping criterion. Finally these proofs make use of the restricted strong convexity (RSC) assumption introduced [NRWY12, LW13], but do not provide any general tool to establish this condition. In contrast, we prove uniform convergence results that can be used to ensure a condition similar to RSC.

To the best of our knowledge, the only proof of unique local minimum of the regularized empirical risk is obtained in a recent paper by Po-Ling Loh [Loh15]. This works assumes the linear regression model $y_i = \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle + \varepsilon_i$, and establishes uniqueness for penalized regression with a certain class of bounded regularizers. This result is comparable to our Theorem 8, see Section 4.4, which uses ℓ_1 regularization instead. Note that, in [Loh15], the sample size is required to scale quadratically in the sparsity: $n \gtrsim s_0^2$. Our proof technique is substantially different from the one of [Loh15], and we only require $n \gtrsim s_0 \log d$.

Hybrid optimization methods. It is often difficult to ensure global convergence to a minimizer of the sample risk $\widehat{R}_n(\,\cdot\,)$ or even to a statistical neighborhood of the true parameters. Several papers develop two-stage procedures to overcome this problem. The first stage constructs a smart initialization $\hat{\boldsymbol{\theta}}(0)$ that is within a certain large neighborhood of the true parameters. Spectral methods are often used to implement this step. In the second stage, the estimate is refined by gradient descent (or another local procedure) initialized at $\hat{\boldsymbol{\theta}}(0)$. This general approach was studied in a number of problems including matrix completion [KOM09], phase retrieval [CC15], tensor decomposition [AGJ15].

In some cases, the local optimization stage is only proved to converge to a statistical neighborhood of $\boldsymbol{\theta}_0$, and hence this style of analysis shares the shortcomings emphasized in the previous paragraph. In others, it is proven to converge to a single point. Further, in practice, the smart initialization is often not needed, and descent algorithms converge from random initialization as well. Finally, as mentioned above, these analyses are typically carried on in a case-by-case manner.

3. Uniform convergence results. In this section we develop our key tools, that are uniform convergence results on the gradient and Hessian of the empirical risk. We also establish some of the direct implications of our results. Throughout, the data consists of the i.i.d. random variables $\{\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n\}$. We will use $\{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n\}$ if we want to refer to the corresponding realization. The empirical risk is defines by Eq. (1.2) and the corresponding population risk is $R(\boldsymbol{\theta}) = \mathbb{E}\widehat{R}_n(\boldsymbol{\theta}) = \mathbb{E}\ell(\boldsymbol{\theta}; \boldsymbol{Z})$. The true parameter vector $\boldsymbol{\theta}_0$ satisfies the condition $\nabla R(\boldsymbol{\theta}_0) = \mathbb{E}[\nabla \ell(\boldsymbol{\theta}_0; \boldsymbol{Z})] = \mathbf{0}$.

We consider two regimes, a *high dimensional regime* in which the number of parameters p is allowed to diverge roughly in proportion with the number of samples n, and a *very high-dimensional regime* in which the true parameters' vector $\boldsymbol{\theta}_0$ is sparse and the number of parameters p can be much larger than n. We treat these two cases separately because the theory is simpler and more general in the first regime.

3.1. *High-dimensional regime.* In order to avoid technical complications, we will limit optimization to a bounded set, i.e. we will let $\Theta_{n,p} = \mathsf{B}^p(r) \equiv \{\boldsymbol{\theta} \in \mathbb{R}^p,\ \|\boldsymbol{\theta}\|_2 \le r\}$ to be the Euclidean ball in p dimensions.

We begin by stating our assumptions. Assumptions 1 and 2 below quantify the amount of statistical noise in the gradient and Hessian of the loss function.

Assumption 1 (Gradient statistical noise). *The gradient of the loss is τ^2-sub-Gaussian. Namely, for any $\boldsymbol{\lambda} \in \mathbb{R}^p$, and $\boldsymbol{\theta} \in \mathsf{B}^p(r)$*

$$\mathbb{E}\Big\{ \exp\Big(\langle \boldsymbol{\lambda}, \nabla \ell(\boldsymbol{\theta}; \boldsymbol{Z}) - \mathbb{E}[\nabla \ell(\boldsymbol{\theta}; \boldsymbol{Z})]\rangle\Big)\Big\} \le \exp\left(\frac{\tau^2 \|\boldsymbol{\lambda}\|_2^2}{2}\right). \tag{3.1}$$

Assumption 2 (Hessian statistical noise). *The Hessian of the loss, evaluated on a unit vector, is τ^2-sub-exponential. Namely, for any $\boldsymbol{\lambda} \in \mathsf{B}^p(1)$,* *and* $\boldsymbol{\theta} \in \mathsf{B}^p(r)$

$$\mathcal{Z}_{\boldsymbol{\lambda},\boldsymbol{\theta}} \equiv \langle \boldsymbol{\lambda}, \nabla^2 \ell(\boldsymbol{\theta}; \boldsymbol{Z}) \boldsymbol{\lambda} \rangle \,, \tag{3.2}$$

$$\mathbb{E}\left\{ \exp\left(\frac{1}{\tau^2} \big| \mathcal{Z}_{\boldsymbol{\lambda},\boldsymbol{\theta}} - \mathbb{E}\mathcal{Z}_{\boldsymbol{\lambda},\boldsymbol{\theta}} \big| \right) \right\} \le 2\,. \tag{3.3}$$

Our third assumption requires the Hessian of the loss to be a Lipschitz function of the vector of parameters $\boldsymbol{\theta}$.

Assumption 3 (Hessian regularity). *The Hessian of the population risk is bounded at one point. Namely, there exists* $\theta_* \in \mathsf{B}^p(r)$ *and* H *such that* $\|\nabla^2 R(\boldsymbol{\theta}_*)\|_{\rm op} \le H$.

Further, the Hessian of the loss function is Lipschitz continuous with integrable Lipschitz constant. Namely, there exists J_* *(potentially diverging polynomially in* p*) such that*

$$J(\boldsymbol{z}) \equiv \sup_{\boldsymbol{\theta}_1 \ne \boldsymbol{\theta}_2 \in \mathsf{B}^p(r)} \frac{\big\| \nabla^2 \ell(\boldsymbol{\theta}_1; \boldsymbol{z}) - \nabla^2 \ell(\boldsymbol{\theta}_2; \boldsymbol{z}) \big\|_{\rm op}}{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2}\,, \tag{3.4}$$

$$\mathbb{E}\big\{ J(\boldsymbol{Z}) \big\} \le J_*\,. \tag{3.5}$$

Further, there exists a constant c_h *such that* $H \le \tau^2 p^{c_h}$, $J_* \le \tau^3 p^{c_h}$.

Remark 1. *The constant* J_* *serves as a third derivative control of the loss function, and controls the discretization error in proving the uniform convergence of the Hessian. The sample size will depend on* H *and* J_* *logarithmically, which is why we assume* H *and* J_* *to grow at most polynomially in dimension* p.

Remark 2. *Note that* $\nabla \ell$ *has the same units*[4] *as* $1/r$, *and* $\nabla^2 \ell$ *has the same units as* $1/r^2$. *Thus,* τ *has the same units as* $1/r$, H *has the same units as* τ^2, *and* J_* *has the same units as* τ^3. *This is the reason why we bound* H *and* J_* *in the form as in Assumption 3. In this way,* $(r \cdot \tau)$ *and* c_h *are dimensionless.*

Discrete loss functions (e.g. the $0-1$ loss) are common within the statistical learning literature, but do not satisfy the above assumption because the gradient and Hessian are not defined everywhere. Note however that these can be well approximated by differentiable losses, with little –if any– practical difference.

We are now in position to state our uniform convergence result.

[4] By this we mean that the two quantities behave in the same way under a rescaling of the parameters $\boldsymbol{\theta}$.

Theorem 1. *Under Assumptions 1, 2, and 3 stated above, there exists a universal constant C_0, such that letting $C = C_0 \cdot (c_h \vee \log(r\tau/\delta) \vee 1)$, the following hold:*

(a) The sample gradient converges uniformly to the population gradient in Euclidean norm. Namely, if $n \geq Cp\log p$, we have

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\mathsf{B}^p(r)} \big\|\nabla\widehat{R}_n(\boldsymbol{\theta}) - \nabla R(\boldsymbol{\theta})\big\|_2 \leq \tau\sqrt{\frac{Cp\log n}{n}}\right) \geq 1-\delta\,. \tag{3.6}$$

(b) The sample Hessian converges uniformly to the population Hessian in operator norm. Namely, if $n \geq Cp\log p$, we have

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\mathsf{B}^p(r)} \big\|\nabla^2\widehat{R}_n(\boldsymbol{\theta}) - \nabla^2 R(\boldsymbol{\theta})\big\|_{\rm op} \leq \tau^2\sqrt{\frac{Cp\log n}{n}}\right) \geq 1-\delta\,. \tag{3.7}$$

3.2. *Topology of the empirical risk.* Theorem 1 immediately implies that the structure of stationary points of the sample risk $\widehat{R}_n(\,\cdot\,)$ must reflect that of the population risk. In order to formalize this intuition, we will discuss its implications for a class of functions that we will call *strongly Morse function.* We will then consider a broader set of functions known as *strict saddle.*

3.2.1. *Strongly Morse functions.* Given a differentiable function $F : \mathsf{B}^d(r) \to \mathbb{R}$, we say that $\boldsymbol{x}$ in the interior of the ball $\mathsf{B}^d(r)$ is critical (or stationary) if $\nabla F(\boldsymbol{x}) = 0$.

Recall that a twice differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ is said to be a *Morse function* if all its critical points are non-degenerate, i.e. have an invertible Hessian. In other words $\nabla F(\boldsymbol{x}) = 0$ implies $\lambda_i(\nabla^2 F(\boldsymbol{x})) \neq 0$ for all $i \in \{1, \dots, d\}$. Morse functions behave well under differentiable reparametrizations, and hence play a central role in differential topology: we refer to [GP10] for a readable introduction to this area, and to [Mil63, Mil97, DFN12] for additional background. The supplementary material contains a brief introduction to the most important notions we use in the proofs.

One key feature of a Morse function F is that, for any $\boldsymbol{x}_0 \in \mathbb{R}^d$, there exists a neighborhood $\mathsf{B}(\boldsymbol{x}_0, \varepsilon)$ of $\boldsymbol{x}_0$ such that, within $\mathsf{B}(\boldsymbol{x}_0, \varepsilon)$, $F(\boldsymbol{x})$ is qualitatively well described by its second order Taylor expansion at $\boldsymbol{x}_0$. In particular, if $\boldsymbol{x}_0$ is a critical point of F, then there exists a small neighborhood of $\boldsymbol{x}_0$ that does not contain any other critical point[5]: all critical points are isolated.

[5] Since F is twice differentiable, by Taylor expansion there exists $\delta(\varepsilon)$ (with $\delta(\varepsilon) \to 0$ as $\varepsilon \to 0$) such that $\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{x}_0) - \nabla^2 F(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0)\|_2 \leq \|\boldsymbol{x} - \boldsymbol{x}_0\|_2\delta(\varepsilon)$ for all $\boldsymbol{x} \in \mathsf{B}(\boldsymbol{x}_0, \varepsilon)$. For $\boldsymbol{x}_0$ a critical point, assume by contradiction that $\boldsymbol{x}$ is another critical

As a consequence, a morse function can only have a finite number of critical point in a compact domain $K \subseteq \mathbb{R}^d$. If this wasn't the case, i.e. if the set of critical points $S \subseteq K$ was infinite, it would have an accumulation point $\boldsymbol{x}_* \in K$. By continuity of the gradient, $\boldsymbol{x}_*$ would be itself a critical point, and have infinitely many other critical points in any neighborhood, thus leading to a contradiction.

The index of a non-degenerate critical point $\boldsymbol{x}_0$ of a twice differentiable function F is the number of negative eigenvalues of the Hessian $\nabla F(\boldsymbol{x}_0)$: we will denote this integer by $\text{Ind}_{\boldsymbol{x}_0}(F)$. Morse Lemma characterizes completely the behavior of F in a neighborhood of $\boldsymbol{x}_0$. Namely, if $\text{Ind}_{\boldsymbol{x}_0}(F) = k$, there exists differentiable coordinates[6] $\varphi_1(\boldsymbol{x}), \dots, \varphi_d(\boldsymbol{x})$ defined on $\mathsf{B}(\boldsymbol{x}_0, \varepsilon)$ such that $F(\boldsymbol{x}) = F(\boldsymbol{x}_0) + \sum_{i=1}^{d-k} \varphi_i(\boldsymbol{x})^2 - \sum_{i=d-k+1}^{d} \varphi_i(\boldsymbol{x})^2$. In other words, all critical points with the same index look alike, modulo differentiable changes of coordinates.

Our next definition provides a quantitative version of the notion of Morse functions. We focus on the case in which F has a bounded domain (a Euclidean ball) because this is the relevant setting for our applications.

DEFINITION 1. *We say that a twice differentiable function $F : \mathsf{B}^d(r) \to \mathbb{R}$ is* (ε, η)-strongly Morse *if $\|\nabla F(\boldsymbol{x})\|_2 > \varepsilon$ for $\|\boldsymbol{x}\|_2 = r$ and, for any $\boldsymbol{x} \in \mathbb{R}^d$, $\|\boldsymbol{x}\|_2 < r$, the following holds*

$$\big\|\nabla F(\boldsymbol{x})\big\|_2 \le \varepsilon \quad \Rightarrow \quad \min_{i \in [d]} \big|\lambda_i\big(\nabla^2 F(\boldsymbol{x})\big)\big| \ge \eta\,. \tag{3.8}$$

The next theorem implies that if the population risk $R(\,\cdot\,)$ is strongly Morse, then the empirical risk retains, with high probability, the same topological structure.

THEOREM 2. *Under Assumptions 1, 2, and 3, let $n \ge 4Cp \log n \cdot ((\tau^2/\varepsilon^2) \vee (\tau^4/\eta^2))$, where $C = C(\tau^2, \delta, r, c_h)$ is as in the statement of Theorem 1. Then the following happens with probability at least $1 - \delta$.*

If the population risk $R : \boldsymbol{\theta} \to R(\boldsymbol{\theta})$ is (ε, η)-strongly Morse in $\mathsf{B}^p(r)$, then the sample risk $\widehat{R}_n : \boldsymbol{\theta} \mapsto \widehat{R}_n(\boldsymbol{\theta})$ is $(\varepsilon/2, \eta/2)$-strongly Morse in $\mathsf{B}^p(r)$. Further there is a one-to-one correspondence between the set of critical points of $R(\,\cdot\,)$, $\mathcal{C} = \{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(k)}\}$ and the set of critical points of $\widehat{R}_n(\,\cdot\,)$, $\mathcal{C}_n = \{\hat{\boldsymbol{\theta}}_n^{(1)}, \dots, \hat{\boldsymbol{\theta}}_n^{(k)}\}$ such that (letting $\hat{\boldsymbol{\theta}}_n^{(j)}$ be the point in correspondence with $\boldsymbol{\theta}^{(j)}$, for any $j \in [k]$)

point in $\mathsf{B}(\boldsymbol{x}_0, \varepsilon)$. Then we have, for $\boldsymbol{H}_0 = \nabla^2 F(\boldsymbol{x}_0)$, $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{x}_0$, $\|\boldsymbol{H}_0 \boldsymbol{v}\|_2 \le \|\boldsymbol{v}\|_2 \delta(\varepsilon)$. But $\|\boldsymbol{H}_0 \boldsymbol{v}\|_2^2 = \langle \boldsymbol{v}, \boldsymbol{H}_0^2 \boldsymbol{v} \rangle \ge \min_{i \le d} |\lambda_i(\boldsymbol{H}_0)|^2 \cdot \|\boldsymbol{v}\|_2^2$, which gives a contradiction if we choose ε so that $\delta^2(\varepsilon) < \min_{i \le d} |\lambda_i(\boldsymbol{H}_0)|^2$.

[6] This means that the map $\boldsymbol{x} \mapsto (\varphi_1(\boldsymbol{x}), \dots, \varphi_d(\boldsymbol{x}_0))$ is a diffeomorphism.

(a) *The index of* $\hat{\boldsymbol{\theta}}_n^{(j)}$ *coincides with the index of* $\boldsymbol{\theta}^{(j)}$*. (In particular, local minima correspond to local minima, and saddles to saddles.)*

(b) *If we further let* $L = \sup_{\boldsymbol{\theta} \in \mathsf{B}^p(r)} \|\nabla^3 R(\boldsymbol{\theta})\|_{\rm op}$*, and assume* $n \geq 4Cp \log n / \eta_*^2$ *where* $\eta_*^2 = (\varepsilon^2/\tau^2) \wedge (\eta^2/\tau^4) \wedge (\eta^4/(L^2\tau^2))$*, we have, for each* $j \in \{1, \dots, k\}$*,*

$$\|\hat{\boldsymbol{\theta}}_n^{(j)} - \boldsymbol{\theta}^{(j)}\|_2 \leq \frac{2\tau}{\eta}\sqrt{\frac{Cp\log n}{n}}\,. \tag{3.9}$$

3.2.2. *Strict saddle functions.* The strong Morse assumption imposes conditions on all the eigenvalues of the Hessian $\nabla^2 R(\boldsymbol{\theta})$ at near-critical points, and implies a detailed characterization of the empirical risk. In some applications only weaker properties can be established for the population risk. These can nevertheless be very useful and Theorem 1 can be used to transfer them to the empirical risk. A useful general notion is the one of *strict saddle* functions, first introduced in [GHJY15].

DEFINITION 2. *We say that a twice differentiable function* $F : \mathsf{B}^d(r) \to \mathbb{R}$ *is* (ε, η)-strict saddle *if* $\|\nabla F(\boldsymbol{x})\|_2 > \varepsilon$ *for* $\|\boldsymbol{x}\|_2 = r$ *and, for any* $\boldsymbol{x} \in \mathbb{R}^d$*,* $\|\boldsymbol{x}\|_2 < r$*, the following holds*

$$\big\|\nabla F(\boldsymbol{x})\big\|_2 \leq \varepsilon \quad \Rightarrow \quad \big|\lambda_{\min}\big(\nabla^2 F(\boldsymbol{x})\big)\big| \geq \eta\,, \tag{3.10}$$

where $\lambda_{\min}(\boldsymbol{M}) = \min_{i \leq d} \lambda_i(\boldsymbol{M})$ *is the minimum eigenvalue of matrix* $\boldsymbol{M} \in \mathbb{R}^{d \times d}$*.*

This definition is completely analogous to the one of strongly Morse functions: the only difference is that we are only imposing a condition on the smallest eigenvalue of the Hessian. In particular, strongly Morse function are a subclass of strict saddle functions.

In strict saddle functions, near critical points are either strongly convex, or have significant negative direction of the Hessian (and hence can be escaped by optimization algorithms). Our definition might seem to impose weaker conditions than the original one in [GHJY15], which additionally requires existence of local minima close to convex points. However, Lemma 8 in the supplemental file Supplement A implies that the two definitions are equivalent.

Notice that any local minimum of a strict saddle function is a non-degenerate critical point. Hence, by the same argument in the previous section, local minima are isolated, and there can be finitely many of them in any compact domain. Also, since $\|\nabla F(\boldsymbol{x})\|_2 > \varepsilon$ on the boundary, all local minima are in the interior of $\mathsf{B}^d(r)$.

THEOREM 3. *Under Assumptions 1, 2, and 3, let $n \geq 4Cp \log n \cdot((\tau^2/\varepsilon^2) \vee (\tau^4/\eta^2))$, where $C = C(\tau^2, \delta, r, c_h)$ is as in the statement of Theorem 1. Then the following happens with probability at least $1 - \delta$.*

If the population risk $R : \boldsymbol{\theta} \to R(\boldsymbol{\theta})$ is (ε, η)-strict saddle in $\mathsf{B}^p(r)$, then the sample risk $\widehat{R}_n : \boldsymbol{\theta} \mapsto \widehat{R}_n(\boldsymbol{\theta})$ is $(\varepsilon/2, \eta/2)$-strict saddle in $\mathsf{B}^p(r)$. Further there is a one-to-one correspondence between the set of local minima of $R(\,\cdot\,)$, $\mathcal{C} = \{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(k)}\}$ and the set of local minima of $\widehat{R}_n(\,\cdot\,)$, $\mathcal{C}_n = \{\hat{\boldsymbol{\theta}}_n^{(1)}, \dots, \hat{\boldsymbol{\theta}}_n^{(k)}\}$ such that (letting $\hat{\boldsymbol{\theta}}_n^{(j)}$ be the local minimum in correspondence with $\boldsymbol{\theta}^{(j)}$, for any $j \in [k]$), for each $j \in \{1, \dots, k\}$,

$$\|\hat{\boldsymbol{\theta}}_n^{(j)} - \boldsymbol{\theta}^{(j)}\|_2 \leq \frac{2\tau}{\eta}\sqrt{\frac{Cp\log n}{n}}\,. \tag{3.11}$$

3.3. *Very high-dimensional regime.* In the very-high dimensional regime $n \ll p$, we will solve the ℓ_1-penalized risk minimization problem

$$\begin{aligned} \text{minimize} \quad & \widehat{R}_n(\boldsymbol{\theta}) + \lambda_n \|\boldsymbol{\theta}\|_1\,, \\ \text{subject to} \quad & \|\boldsymbol{\theta}\|_2 \leq r\,. \end{aligned} \tag{3.12}$$

We need some additional assumptions. It is fairly straightforward to check them in specific cases, see e.g. Section 4.1. The first assumption is mainly technical, and not overly restrictive: it requires the loss function to have almost surely bounded gradient, in a suitable sense.

ASSUMPTION 4 (Gradient bounds). *There exists a constant T_* such that $\boldsymbol{Z}$-almost surely, for all $\boldsymbol{\theta} \in \mathsf{B}_2^p(r)$,*

$$\big\|\nabla \ell(\boldsymbol{\theta}; \boldsymbol{Z})\big\|_\infty \leq T_*\,. \tag{3.13}$$

Our key structural assumption is stated next. It requires the gradient of the loss function to depend on the parameters only through a linear function of $\boldsymbol{\theta}$, possibly dependent on the feature vector $\boldsymbol{z}$. Note that $\boldsymbol{\theta}_0$ is regarded here as fixed, and hence omitted from the arguments.

ASSUMPTION 5 (Generalized gradient linearity). *There exist functions $g : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$, $(t, \boldsymbol{z}) \mapsto g(t; \boldsymbol{z})$ and $\boldsymbol{\psi} : \mathbb{R}^d \to \mathbb{R}^p$, $\boldsymbol{z} \mapsto \boldsymbol{\psi}_2(\boldsymbol{z})$, such that*

$$\langle \nabla \ell(\boldsymbol{\theta}; \boldsymbol{z}), \boldsymbol{\theta} - \boldsymbol{\theta}_0 \rangle = g(\langle \boldsymbol{\theta} - \boldsymbol{\theta}_0, \boldsymbol{\psi}(\boldsymbol{z}) \rangle; \boldsymbol{z})\,. \tag{3.14}$$

In addition, $g(t; \boldsymbol{z})$ is L_*-Lipschitz to its first argument, $g(0; \boldsymbol{z}) = 0$, and $\boldsymbol{\psi}(\boldsymbol{Z})$ is mean-zero and τ^2-sub-Gaussian.

As an example, in the case of binary linear classification and robust regression, the data is given as a pair $\boldsymbol{z} = (y, \boldsymbol{x})$, and there exists a function $f(t; \boldsymbol{z})$ such that $\nabla \ell(\boldsymbol{\theta}; \boldsymbol{z}) = f(\langle \boldsymbol{\theta} - \boldsymbol{\theta}_0, \boldsymbol{x}\rangle; \boldsymbol{z})\boldsymbol{x}$. Assumption 5 is satisfied with $g(t; \boldsymbol{z}) = t\, f(t; \boldsymbol{z})$ provided the latter is Lipschitz as a function of $t \in \mathbb{R}$.

Theorem 4. *Under Assumptions 2, 3, 4 and 5 stated above, there exists a constant C_1 that depends on (r, τ^2, c_h, δ), and a universal constant C_0 such that letting $C_2 = C_0 \cdot (c_h \vee \log(r\tau/\delta) \vee 1)$, the following hold:*

(a) *The sample directional gradient converges uniformly to the population directional gradient, along the direction $(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$. Namely, we have*

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in \mathsf{B}_2^p(r)\setminus\{\mathbf{0}\}} \frac{\left|\langle \nabla \widehat{R}_n(\boldsymbol{\theta}) - \nabla R(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\theta}_0\rangle\right|}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_1} \le (T_* + L_*\tau)\sqrt{\frac{C_1 \log(np)}{n}}\right) \ge 1 - \delta\,. \tag{3.15}$$

(b) *The sample restricted Hessian converges uniformly to the population restricted Hessian in the set $\mathsf{B}_2^p(r) \cap \mathsf{B}_0^p(s_0)$ for any $s_0 \le p$. Namely, as $n \ge C_2 s_0 \log(np)$ we have*

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in \mathsf{B}_2^p(r)\cap \mathsf{B}_0^p(s_0), \boldsymbol{v}\in \mathsf{B}_2^p(1)\cap \mathsf{B}_0^p(s_0)} \left|\left\langle \boldsymbol{v}, \left(\nabla^2 \widehat{R}_n(\boldsymbol{\theta}) - \nabla^2 R(\boldsymbol{\theta})\right)\boldsymbol{v}\right\rangle\right| \le \tau^2\sqrt{\frac{C_2 s_0 \log(np)}{n}}\right) \ge 1 - \delta\,. \tag{3.16}$$

4. Applications.

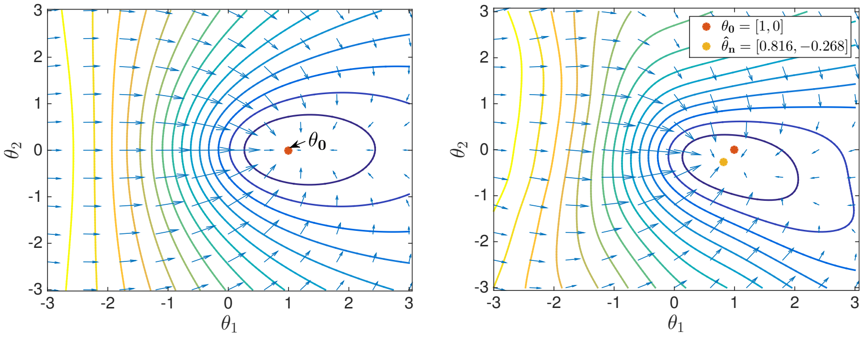
4.1. *Binary linear classification: High dimensional regime.* As mentioned in the introduction, in this case we are given n pairs $\boldsymbol{z}_1 = (y_1, \boldsymbol{x}_1), \dots, \boldsymbol{z}_n = (y_n, \boldsymbol{x}_n)$ with $y_i \in \{0, 1\}$, $\boldsymbol{x}_i \in \mathbb{R}^d$, whereby $\mathbb{P}(Y_i = 1|\boldsymbol{X}_i = \boldsymbol{x}) = \sigma(\langle \boldsymbol{\theta}_0, \boldsymbol{x}\rangle)$ (hence $p = d$ in this case). We estimate $\boldsymbol{\theta}_0$ by minimizing the non-linear square loss (1.4), which we copy here for the reader's convenience:

$$\begin{aligned} &\text{minimize} \quad \widehat{R}_n(\boldsymbol{\theta}) \equiv \frac{1}{n}\sum_{i=1}^n \Big(y_i - \sigma(\langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle)\Big)^2, \\ &\text{subject to} \quad \|\boldsymbol{\theta}\|_2 \le r\,. \end{aligned} \tag{4.1}$$

This can be regarded as a smooth version of the $0 - 1$ loss.

We collect below the technical assumptions on this model.

Assumption 6 (Binary linear classification). (a) *The activation $z \mapsto \sigma(z)$ is three times differentiable with $\sigma'(z) > 0$ for all z, and has*



(a) (b)

Fig 1: Binary linear classification: (a) Population risk for $d = 2$. (b) A realization of the empirical risk for $d = 2$, and $n/d = 20$.

bounded first, second and third derivatives. Namely, for some constant $L_\sigma > 0$*:*

$$\max\left\{\|\sigma'\|_\infty, \|\sigma''\|_\infty, \|\sigma'''\|_\infty\right\} \le L_\sigma\,. \tag{4.2}$$

(b) *The feature vector* $\boldsymbol{X}$ *has zero mean and is* τ^2*-sub-Gaussian, that is* $\mathbb{E}[e^{\langle \boldsymbol{\lambda}, \boldsymbol{X}\rangle}] \le e^{\frac{\tau^2\|\boldsymbol{\lambda}\|_2^2}{2}}$ *for all* $\boldsymbol{\lambda} \in \mathbb{R}^d$.

(c) *The feature vector* $\boldsymbol{X}$ *spans all directions in* $\mathbb{R}^d$*, that is,* $\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}] \succeq \underline{\gamma}\tau^2 \mathrm{I}_{d\times d}$ *for some* $0 < \underline{\gamma} < 1$.

Assumption 6.(a) is satisfied by many classical activation functions, a prominent example being the logistic (or sigmoid) function $\sigma_L(z) = (1 + e^{-z})^{-1}$.

Our main results on binary linear classification are summarized in the theorem below.

Theorem 5. *Under Assumption 6, further assume* $\|\boldsymbol{\theta}_0\|_2 \le r/3$*. There exist positive constants* C_1*,* C_2 *and* $h_{\max}$ *depending on parameters* $(L_\sigma, r, \tau^2, \underline{\gamma}, \delta)$ *and the activation function* $\sigma(\cdot)$*, but independent of* n *and* d*, such that, if* $n \ge C_1 d \log d$*, the following hold with probability at least* $1 - \delta$*:*

(a) *The empirical risk function* $\boldsymbol{\theta} \mapsto \widehat{R}_n(\boldsymbol{\theta})$ *has a unique local minimizer in* $\mathsf{B}^d(\mathbf{0}, r)$*, that is the global minimizer* $\hat{\boldsymbol{\theta}}_n$*.*

(b) *Gradient descent with fixed step size* $h_k = h \le h_{\max}$ *converges exponentially fast to the global minimizer, for any initialization* $\boldsymbol{\theta}_s \in \mathsf{B}^d(\boldsymbol{\theta}_0, 2r/3)$*:* $\|\hat{\boldsymbol{\theta}}_n(k) - \hat{\boldsymbol{\theta}}_n\|_2 \le C_1\|\boldsymbol{\theta}_s - \hat{\boldsymbol{\theta}}_n\|_2\,(1 - h/C_1)^k$.

(*c*) *We have* $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 \le C_2\sqrt{(d\log n)/n}$.

The proof of this theorem can be found in section E.1 in the supplemental file Supplement A, and is based on the following two-step strategy. First, we study the population risk $R(\boldsymbol{\theta})$, and establish its qualitative properties using analysis. In particular, our results imply that $R(\boldsymbol{\theta})$ is strongly Morse in the domain $\mathsf{B}^d(\mathbf{0}, r)$ (but we prove that an even stronger characterization). Second, we use our uniform convergence result (Theorem 1) to prove that the same properties carry over to the sample risk $\widehat{R}_n(\boldsymbol{\theta})$. Figure 1 presents a small numerical example that illustrates how the qualitative features of the population risk apply to the empirical risk as well.

A few remarks are in order. First of all, the convergence rate of gradient descent (at point (*b*)) is *independent of the dimension* d *and number of samples* n. In other words, $O(\log(1/\varepsilon))$ iterations are sufficient to converge within distance ε from the global minimizer. Classical theory of empirical risk minimization only concerns the statistical properties of the optimum, but does not provide efficient algorithms.

Next, note that our condition on the sample size n is nearly optimal. Indeed, it is information-theoretically impossible to estimate $\boldsymbol{\theta}_0$ from less than $n < d$ binary samples. Finally, the convergence rate at point (*c*) also nearly matches the optimal (parametric) rate $\sqrt{d/n}$.

4.2. *Binary linear classification: Very high-dimensional regime.* As in the previous section, we are given n pairs $\boldsymbol{z}_1 = (y_1, \boldsymbol{x}_1), \dots, \boldsymbol{z}_n = (y_n, \boldsymbol{x}_n)$ with $y_i \in \{0,1\}$, $\boldsymbol{x}_i \in \mathbb{R}^d$, and $\mathbb{P}(Y_i = 1|\boldsymbol{X}_i = \boldsymbol{x}) = \sigma(\langle\boldsymbol{\theta}_0, \boldsymbol{x}\rangle)$. However $\boldsymbol{\theta}_0$ is assumed to be sparse, and the number of samples n is allowed to be much smaller than the ambient dimension $d = p$. We adopt again the non-linear square loss (1.4), but now use a ℓ_2-constrained ℓ_1-regularized risk minimization, as per Eq. (3.12), which we rewrite here explicitly for the reader's ease

$$\begin{aligned} &\text{minimize} \quad \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \sigma(\langle\boldsymbol{\theta}, \boldsymbol{x}_i\rangle)\Big)^2 + \lambda_n\|\boldsymbol{\theta}\|_1\,, \\ &\text{subject to} \quad \|\boldsymbol{\theta}\|_2 \le r\,. \end{aligned} \tag{4.3}$$

The very high-dimensional regime $d \gg n$ is of interest in many contexts. In machine learning, the number of parameters p can increase when a large number of additional features are added to the model (for instance, nonlinear functions of an original set of features). In signal processing, $\boldsymbol{\theta}_0$ represents an unknown signal, of which we measure noisy random linear projections $\langle\boldsymbol{x}_i, \boldsymbol{\theta}_0\rangle$, $i \in [n]$, quantized to *one single bit*. This scenario is relevant to group testing [AS12] and analog-to-digital conversion [LWYB11, LB12], and has been studied under the name of 'one-bit compressed sensing'; see [PV13a] and references therein.

In the very high-dimensional regime we need additional assumptions on the distribution of $\boldsymbol{X}$ as well as the activation function σ.

Assumption 7 (Fast-decaying activation). *The activation function σ satisfy $\sup_{t\in\mathbb{R}}\{|\sigma'(t)t|, |\sigma''(t)t|\} \leq C_\sigma$ for some absolute constant C_σ.*

Assumption 8 (Continuous and bounded features). *The feature vector $\boldsymbol{X}$ has a density $p(\,\cdot\,)$ in $\mathbb{R}^d$, that is, $\mathbb{P}(\boldsymbol{X} \in A) = \int_A p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ for all Borel sets $A \subseteq \mathbb{R}^d$. In addition, the feature vector is bounded: $\|\boldsymbol{X}\|_\infty \leq M\tau$, and $|\langle \boldsymbol{X}, \boldsymbol{\theta}_0/\|\boldsymbol{\theta}_0\|_2\rangle| \leq M\tau$ almost surely, with $\boldsymbol{\theta}_0$ the ground truth parameter. Here M is a dimensionless constant greater than 1.*

Remark 3. *Assumption 7 holds popular examples of activation functions, such as the logistic $\sigma_L(z) = (1+e^{-z})^{-1}$ or probit $\sigma_P(z) = \Phi(z)$.*

Also note that Assumption 8 requires $|\langle \boldsymbol{X}, \boldsymbol{\theta}_0\rangle|/\|\boldsymbol{\theta}_0\|_2 \leq M\tau$ to hold only when $\boldsymbol{\theta}_0$ is the fixed ground truth parameter *and not uniformly over all s_0-sparse vectors. For unbounded sub-Gaussian feature vectors, this assumption does not hold directly. However, for any dataset $\{\boldsymbol{X}_i\}_{i=1}^n$ with $\boldsymbol{X}_i$ independent τ^2-sub-Gaussian, with high probability $\sup_{i\in[n]}\{\|\boldsymbol{X}_i\|_\infty, |\langle \boldsymbol{X}_i, \boldsymbol{\theta}_0/\|\boldsymbol{\theta}_0\|_2\rangle|\} \leq C\sqrt{\log(nd)}\tau$. The next theorem can then be supplemented by a truncation argument at level $M = C\sqrt{\log(nd)}$, leading to the same conclusions with an additional $\log(nd)$ factor in the error bound.*

In the statement of the following theorem, for convenience, we will also assume $n \leq d^{100}$. This is a technical assumption so that we can bound $\log(nd) \leq 101\log(d)$. And since we are considering the very high dimensional regime, it is not meaningful to discuss $n > d^{100}$.

Theorem 6. *Under Assumptions 6, 7 and 8, further assume $\|\boldsymbol{\theta}_0\|_0 \leq s_0$, $\|\boldsymbol{\theta}_0\|_2 \leq r/2$, and $n \leq d^{100}$. Then there exist constants C_n, C_λ, C_s, and ε_0 depending on $(L_\sigma, C_\sigma, r, \tau^2, \underline{\gamma}, \delta)$ and the activation function $\sigma(\cdot)$, but independent of n, d, s_0, and M, such that as $n \geq C_n\, s_0 \log d$ and $\lambda_n \geq C_\lambda M\sqrt{(\log d)/n}$, the following hold with probability at least $1-\delta$:*

- (a) *Any stationary point of problem (4.3) is in $\mathsf{B}_2^d(\boldsymbol{\theta}_0, C_s\sqrt{(M^2 s_0\log d)/n + s_0\lambda_n^2})$.*
- (b) *As long as n is large enough such that $n \geq C_n\, s_0\, \log^2 d$ and $C_s\sqrt{(M^2 s_0\log d)/n + s_0\lambda_n^2} \leq \varepsilon_0$, the problem has a unique local minimizer $\hat{\boldsymbol{\theta}}_n$ which is also the global minimizer.*

As in the previous section, our proof makes a crucial use of the sparse uniform convergence result, Theorem 4, together with an analysis of the population risk.

REMARK 4. *Let us emphasize that Theorem 6 leaves open the existence of a fast algorithm to find the global optimizer $\hat{\boldsymbol{\theta}}_n$. However [Nes13a, Theorem 3] implies that, by running k steps of projected gradient descent, we can find an estimate $\hat{\boldsymbol{\theta}}_n(k)$ which has a subgradient of order $O(1/k)$. While we expect this sequence to converge to $\hat{\boldsymbol{\theta}}_n$, we defer this question to future work.*

Theorem 6 establishes a nearly optimal upper bound on the ℓ_2 estimation error $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2$. Indeed this error is within a logarithmic factor from the error achieved by an oracle estimator that is given the exact support of $\boldsymbol{\theta}_0$. For comparison, [PV13a, PV13b] proves $\|\hat{\boldsymbol{\theta}}_n^{\text{LP}} - \boldsymbol{\theta}_0\|_2 \lesssim (s_0/n)^{1/4}(\log p/s_0)^{1/4}$ for a linear programming formulation, under the more restrictive assumption of Gaussian feature vectors $\boldsymbol{x}_i \sim \mathsf{N}(\mathbf{0}, \mathrm{I}_{d\times d})$. This analysis was generalized in [ALPV14] to feature vectors with i.i.d. entries, although with the same estimation error bound. The optimal rate $\|\hat{\boldsymbol{\theta}}_n^{\text{cvx}} - \boldsymbol{\theta}_0\|_2 \lesssim (s_0/n)\log(p/s_0)$ was obtained only recently in [PVY14], again for standard Gaussian feature vectors.

Let us finally emphasize that the estimator defined here uses a bounded loss function and is potentially more robust to outliers than other approaches that use a convex loss (e.g. logistic loss).

4.3. *Robust regression: High-dimensional regime.* In robust regression we are given data $\boldsymbol{z}_1 = (y_1, \boldsymbol{x}_1), \dots, \boldsymbol{z}_n = (y_n, \boldsymbol{x}_n)$ with $y_i \in \mathbb{R}$, $\boldsymbol{x}_i \in \mathbb{R}^d$, and we assume the linear model $y_i = \langle \boldsymbol{\theta}_0, \boldsymbol{x}_i \rangle + \varepsilon_i$, where the noise terms ε_i are i.i.d. with mean zero. Also in this case we have $p = d$. We use the loss (1.5), which we copy here for the reader's convenience:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{n}\sum_{i=1}^{n} \rho\big(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle\big)\,, \\ &\text{subject to} \quad \|\boldsymbol{\theta}\|_2 \le r\,. \end{aligned} \tag{4.4}$$

Classical choices for loss function $t \mapsto \rho(t)$ are the Huber loss [Hub73] which is convex with $\rho_{\text{Huber}}(t) = |t| - \text{const.}$ for t large enough, and Tukey's bisquare loss, which is bounded and defined as

$$\rho_{\text{Tukey}}(t) = \begin{cases} 1 - \big(1 - (t/t_0)^2\big)^3 & \text{for } |t| \le t_0, \\ 1 & \text{for } |t| \ge t_0. \end{cases} \tag{4.5}$$

It is common to define the associated score function as $\psi(t) = \rho'(t)$.

We next formulate our assumptions.

Assumption 9 (Robust regression). *(a) The score function $z \mapsto \psi(z)$ is twice differentiable and odd in z with $\psi(z) \geq 0$ for all $z \geq 0$, and has bounded zero, first, and second derivatives. Namely, for some constant $L_\psi > 0$:*

$$\max\Big\{\|\psi\|_\infty, \|\psi'\|_\infty, \|\psi''\|_\infty\Big\} \leq L_\psi\,. \tag{4.6}$$

(b) The noise ε has a symmetric distribution, i.e. is such that ε is distributed as $-\varepsilon$. Further, defining $g(z) \equiv \mathbb{E}_\varepsilon\{\psi(z+\varepsilon)\}$ we have $g(z) > 0$ for all $z > 0$, as well as $g'(0) > 0$.

(c) The feature vector $\boldsymbol{X}$ has zero mean and is τ^2-sub-Gaussian, that is $\mathbb{E}[e^{\langle \boldsymbol{\lambda}, \boldsymbol{X}\rangle}] \leq e^{\frac{\tau^2\|\boldsymbol{\lambda}\|_2^2}{2}}$ for all $\boldsymbol{\lambda} \in \mathbb{R}^d$.

(d) The feature vector $\boldsymbol{X}$ spans all directions in $\mathbb{R}^d$, that is, $\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}] \succeq \underline{\gamma}\tau^2\mathrm{I}_{d\times d}$ for some $0 < \underline{\gamma} < 1$.

Note that the condition $g(z) \equiv \mathbb{E}_\varepsilon\{\psi(z+\varepsilon)\} > 0$ for all $z > 0$ and $g'(0) > 0$ are quite mild, and holds –for instance– if the noise has a density that is strictly positive for all ε, and decreasing for $\varepsilon > 0$.

Theorem 7. *Under Assumption 9, further assume $\|\boldsymbol{\theta}_0\|_2 \leq r/3$. Then there exist positive constants C_1, C_2 and $h_{\max}$ depending on parameters $(L_\psi, r, \tau^2, \underline{\gamma}, \delta)$, the loss function $\rho(\cdot)$, and the law of noise $\mathbb{P}_\varepsilon$ but independent of n and d, such that as $n \geq C_1 d\log d$, the robust regression estimator satisfies the following with probability at least $1-\delta$:*

(a) The empirical risk function $\boldsymbol{w} \mapsto \widehat{R}_n(\boldsymbol{\theta})$ has a unique local minimizer in $\mathsf{B}^d(r)$, that is the global minimizer $\hat{\boldsymbol{\theta}}_n$.

(b) Gradient descent with fixed step size $h_k = h \leq h_{\max}$ converges exponentially fast to the global minimizer, for any initialization $\boldsymbol{\theta}_s \in \mathsf{B}^d(\boldsymbol{\theta}_0, 2r/3)$: $\|\hat{\boldsymbol{\theta}}_n(k) - \hat{\boldsymbol{\theta}}_n\|_2 \leq C_1\|\boldsymbol{\theta}_s - \hat{\boldsymbol{\theta}}_n\|_2\,(1-h/C_1)^k$.

(c) We have $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 \leq C_2\sqrt{(d\log n)/n}$.

The proof follows the same two steps strategy as for the binary classification problem. In particular, we obtain a precise characterization of the population risk, which (in particular) is strongly Morse.

4.4. *Robust regression: Very high-dimensional regime.* As in the previous section, we are given n pairs $\boldsymbol{z}_1 = (y_1, \boldsymbol{x}_1), \dots, \boldsymbol{z}_n = (y_n, \boldsymbol{x}_n)$ with $y_i \in \mathbb{R}$,

$\boldsymbol{x}_i \in \mathbb{R}^d$, and we assume the linear model $y_i = \langle \boldsymbol{\theta}_0, \boldsymbol{x}_i \rangle + \varepsilon_i$, where the noise terms ε_i are i.i.d. with mean zero. However $\boldsymbol{\theta}_0$ is assumed to be sparse, while the number of samples n is much smaller than the ambient dimension $d = p$. We adopt again the loss (1.5), but now use a ℓ_2-constrained ℓ_1-regularized risk minimization, as per Eq. (3.12), which we rewrite here explicitly for the reader's ease

$$\begin{aligned} &\text{minimize} \quad \frac{1}{n}\sum_{i=1}^{n} \rho\big(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle\big) + \lambda_n \|\boldsymbol{\theta}\|_1 \,, \\ &\text{subject to} \quad \|\boldsymbol{\theta}\|_2 \le r \,. \end{aligned} \tag{4.7}$$

Like the case of very high dimensional binary classification, we also need continuous and bounded feature assumptions, i.e. Assumption 8, and need a fast decaying assumption on $\psi = \rho'$.

ASSUMPTION 10 (Fast-decaying score function). *The score function ψ satisfies $\sup_{t\in\mathbb{R}}\{|\psi(t)t|\} \le C_\psi$ for some absolute constant C_ψ.*

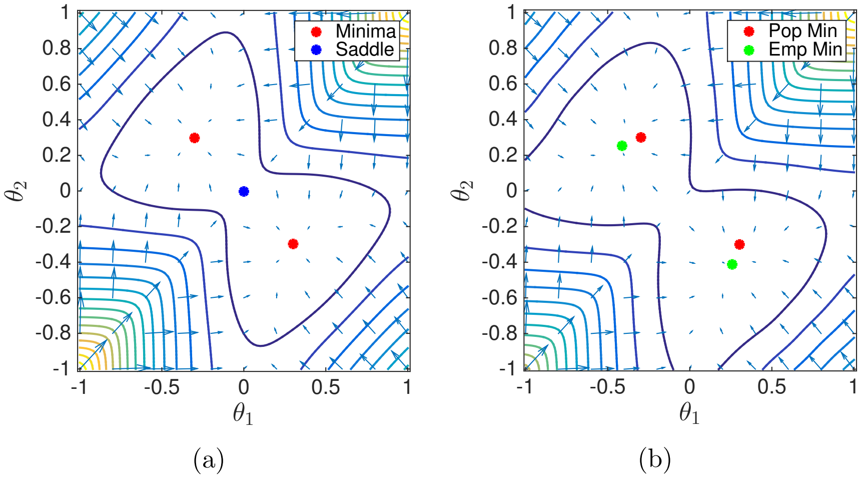
THEOREM 8. *Under Assumptions 6, 8 and 10, further assume $\|\boldsymbol{\theta}_0\|_0 \le s_0$, $\|\boldsymbol{\theta}_0\|_2 \le r/2$, and $n \le d^{100}$. Then there exist constants C_n, C_λ, C_s, and ε_0 depending on $(L_\psi, C_\psi, r, \tau^2, \underline{\gamma}, \delta)$, the loss function ρ, and the law of noise $\mathbb{P}_\varepsilon$, but independent of n, d, s_0 and M, such that as $n \ge C_n\, s_0 \log d$ and $\lambda_n \ge C_\lambda M \sqrt{(\log d)/n}$, the following hold with probability at least $1 - \delta$:*

(a) *Any stationary point of problem (4.7) is in $\mathsf{B}_2^d(\boldsymbol{\theta}_0, C_s\sqrt{(M^2 s_0 \log d)/n + s_0\lambda_n^2})$.*
(b) *As long as n is large enough such that $n \ge C_n\, s_0\, \log^2 d$ and $C_s\sqrt{(M^2 s_0 \log d)/n + s_0\lambda_n^2} \le \varepsilon_0$, the problem has a unique local minimizer $\hat{\boldsymbol{\theta}}_n$ which is also the global minimizer.*

The proof of this theorem is almost the same as the proof of Theorem 6. We will omit the proof to avoid redundancies.

4.5. *Gaussian mixture model.* In the applications considered so far, the population risk has a unique stationary point which is also the global minimum. We used our uniform convergence theorems to prove that the empirical risk has the same property and hence can be optimized efficiently.

In order to illustrate our approach on an example with multiple local minima, we consider clustering within a simple Gaussian mixture model. We are given data points $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n \in \mathbb{R}^d$, with $\boldsymbol{z}_i$ drawn from a mixture of two Gaussians, in equal proportions, $\boldsymbol{z}_i \sim (1/2)\mathsf{N}(\boldsymbol{\theta}_{0,1}, \mathrm{I}_{d\times d}) + (1/2)\mathsf{N}(\boldsymbol{\theta}_{0,2}, \mathrm{I}_{d\times d})$. Define the separation parameter $D = \|\boldsymbol{\theta}_{0,2} - \boldsymbol{\theta}_{0,1}\|_2/2$. We want to estimate



(a) (b)

Fig 2: Gaussian mixture model: (a) Population risk for $d = 1$. (b) A realization of the empirical risk for $d = 1$, and $n = 30$.

the centers $\boldsymbol{\theta}_{0,1}$, $\boldsymbol{\theta}_{0,2}$ by solving the maximum likelihood problem (here $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathbb{R}^{2d}$)

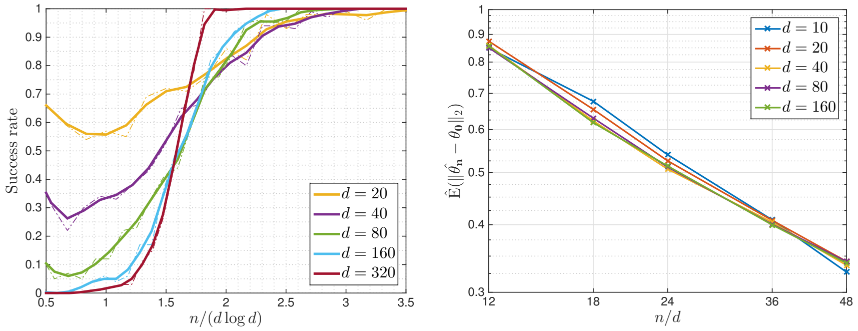
$$\text{minimize} \quad \widehat{R}_n(\boldsymbol{\theta}) \equiv -\frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{a=1}^{2} \phi_d(\boldsymbol{z}_i - \boldsymbol{\theta}_a)\right). \tag{4.8}$$

In this case, the population risk has at least two global minima related by the symmetry under exchange of the two components: $\boldsymbol{\theta}_+ = (\boldsymbol{\theta}_{0,1}, \boldsymbol{\theta}_{0,2})$ and $\boldsymbol{\theta}_- = (\boldsymbol{\theta}_{0,2}, \boldsymbol{\theta}_{0,1})$, as well as a saddle point $\boldsymbol{\theta}_s = ((\boldsymbol{\theta}_{0,1} + \boldsymbol{\theta}_{0,2})/2, (\boldsymbol{\theta}_{0,1} + \boldsymbol{\theta}_{0,2})/2)$. This is a common phenomenon: symmetries lead to multiple minima of the risk function. In a recent paper, Xu, Hsu and Maleki [XHM16] prove that these are the only critical points. A related analysis was carried out by Daskalakis, Tzamos, Christos and Zampetakis [DTZ16] in order to study the behavior of the EM algorithm. See Figure 2 for an illustration.

Theorem 9. *Let $\widehat{R}_n(\boldsymbol{\theta})$ be the empirical risk for an equal-proportion mixture of two Gaussians. Then there exist constants C_1, C_2, and C_3 depending on (D, δ) but independent of n and d, such that as $n \geq C_1 \, d \log d$, the following holds with probability at least $1 - \delta$:*

(a) *In side $\mathsf{B}^{2d}(\boldsymbol{\theta}_s, C_2)$, the empirical risk has exactly two local minima $\hat{\boldsymbol{\theta}}_+$, $\hat{\boldsymbol{\theta}}_-$ related by an exchange of the two classes.*

(b) *For any initialization $\hat{\boldsymbol{\theta}}_0 \in \mathsf{B}^{2d}(\boldsymbol{\theta}_s, C_2)$, the trust region algorithm will converge to one of the local minima.*



(a)

(b)

Fig 3: Binary linear classification, high dimensional: (a) Success rate versus $n/(d \log d)$ for several ambient dimensions d, with $\|\boldsymbol{\theta}_0\|_2 = 3$ (dashed lines are empirical averages, continuous lines are a smoothed version); (b) Estimation error $\widehat{\mathrm{E}}[\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2]$ versus n/d, for $\|\boldsymbol{\theta}_0\|_2 = 1$.

(c) *The local minima satisfy*

$$(4.9) \qquad \|\hat{\boldsymbol{\theta}}_+ - \boldsymbol{\theta}_+\|_2 \le C_3 \sqrt{\frac{d \log n}{n}}, \quad \|\hat{\boldsymbol{\theta}}_- - \boldsymbol{\theta}_-\|_2 \le C_3 \sqrt{\frac{d \log n}{n}}.$$

As in previous examples, we obtain a precise characterization of the population risk, building on [XHM16], and then transfer the result to empirical risk using our uniform convergence results. Our analysis implies –in particular– that the population risk is strict saddle.

5. Numerical experiments. We carried out extensive numerical experiments in order to verify how accurate is our theory. Sections 5.1 to 5.3 present simulations for the non-convex binary classification and robust regression models studied in Section 4. Sections 5.4 present illustrations with real data. We will present simulations for the Gaussian mixture model (Section H.1) and binary classification using the Australian credit dataset (Section H.2) in the supplemental file Supplement A.

5.1. *Binary linear classification: high-dimensional regime .* Figures 3a, 3b, 4a, 4b report our results for the non-convex binary classification model of Section 4.1.

We consider i.i.d. predictors $\boldsymbol{X}_i \sim \mathsf{N}(\mathbf{0}, \mathrm{I}_{d\times d})$, and generate labels $Y_i \in \{0, 1\}$ with $\mathbb{P}(Y_i = 1|\boldsymbol{X}_i = \boldsymbol{x}) = \sigma(\langle \boldsymbol{\theta}_0, \boldsymbol{x}\rangle)$ where $\sigma(u) = \sigma_L(u) = (1 + e^{-u})^{-1}$ is the logistic activation. We perform gradient descent, cf. Eq (1.3) to minimize the empirical risk (1.4), with a minor revision in practice: we will project the points back into $\mathsf{B}^d(r)$ if the iteration points fall out of the ball, with $r = 3\|\boldsymbol{\theta}_0\|_2$. The step size is fixed to be $h = 1$.

In order to test the hypothesis that the landscape is simple (i.e. it has a unique local minimum), we run projected gradient descent starting from multiple random initializations $\boldsymbol{\theta}_s \sim \mathsf{N}(\mathbf{0}, \mathrm{I}_{d\times d}/d)$. If the landscape is simple, we expect the iterates $\hat{\boldsymbol{\theta}}_n(k)$ to converge to the same global minimizer with no dependence on the initialization. If the landscape is rough, projected gradient descent will converge to different points depending on the initialization. Given a maximum number of iterations $k_{\max}$, we define the following quantity, depending on the data $(\boldsymbol{Y}, \boldsymbol{X}) \equiv \{(Y_i, \boldsymbol{X}_i)\}_{1\le i\le n}$,

$$S_{\boldsymbol{Y},\boldsymbol{X}} = \sqrt{\mathrm{Tr}(\widehat{\mathrm{Var}}_{\mathrm{init}}(\hat{\boldsymbol{\theta}}_n(k_{\max})|\boldsymbol{Y}, \boldsymbol{X}))}\,, \tag{5.1}$$

where the variance is taken over the random initializations $\boldsymbol{\theta}_s$. In words, $S_{\boldsymbol{Y},\boldsymbol{X}}$ is the spread of the limit points of projected gradient descent, for the instance $(\boldsymbol{Y}, \boldsymbol{X})$. We then define the empirical success probability as

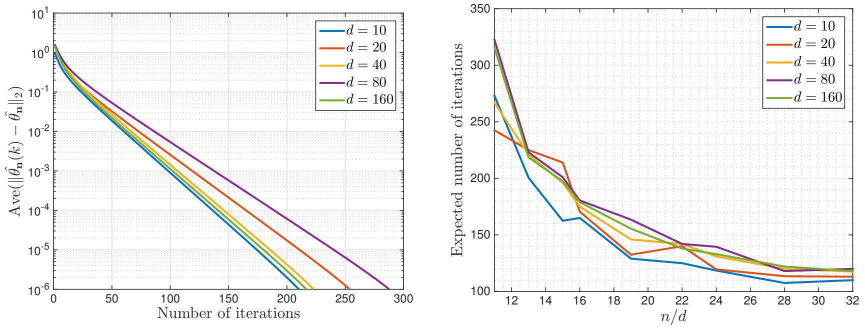
$$\widehat{\mathrm{P}}_{\mathrm{succ}} \equiv \widehat{\mathbb{P}}(S_{\boldsymbol{Y},\boldsymbol{X}} \le \varepsilon)\,. \tag{5.2}$$

In Figure 3a, we plot our results for the empirical success rate, for several values of n, d. In this experiment, we take $\|\boldsymbol{\theta}_0\|_2 = 3$. For each pair (n, d), we generate 100 instances $(Y_i, \boldsymbol{X}_i)$ and run projected gradient descent from 10 random initializations. We use $k_{\max} = 10^4$ iterations and tolerance $\varepsilon = 10^{-2}$ though results seem to be fairly insensitive to these parameters. For each dimension d, the success rate goes rapidly from 0 to 1 as the number of samples n crosses a threshold. We plot the success probability as function of the rescaled number of samples $n/(d \log d)$. On this scale, curves for different dimension cross each other, and become steeper as d increases. This is consistent with Theorem 5. This also suggests a sharp phase transition at $n_*(d)$ which is roughly of order $d \log d$. It is a fascinating open question whether a sharp threshold actually exists[7].

Figure 3b illustrates the behavior of the estimation error $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2$ achieved by gradient descent. In all the following experiments, we will take $\|\boldsymbol{\theta}_0\|_2 = 1$. We plot the estimation error (averaged over 100 random instances) $\widehat{\mathrm{E}}[\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2]$ versus n/d. Curves for different dimensions collapse, and are consistent with the optimal rate $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 = \Theta(\sqrt{d/n})$.

Figure 4a shows the convergence of gradient descent for several values of n and d, for fixed $n/d = 20$. Namely, we plot the distance from the global

[7] When convergence to a single global minimum fails, we observe that often projected gradient actually convergence to the boundary of $\mathsf{B}^d(r)$.



(a) (b)

Fig 4: Binary linear classification, high dimensional: (a) The convergence of the gradient descent algorithm. Here $\|\boldsymbol{\theta}_0\|_2 = 1$, $n/d = 20$. The y-axis is on a log-scale; (b) Minimum number of iterations needed to achieve average distance 10^{-4} from the global optimizer.

minimizer as a function of the number of iterations k, estimated using 100 realizations $(\boldsymbol{Y}, \boldsymbol{X})$. Since there is a small probability that gradient descent fails to find unique minimizer, we average the distance from the global minimizer over the results between the $(0.05, 0.95)$ quantiles of these 100 instances. Convergence to the global minimizer appears to be exponential as predicted by Theorem 5. Also, convergence is fairly independent of the dimension for fixed n/d.

Finally, Figure 4b shows the number of iterations needed to achieve the $\varepsilon = 10^{-4}$ optimization error. We run 100 instances, and we plot the expected number of iteration, by averaging the results between the $(0.05, 0.95)$ quantiles of these 100 instances. When n/d is small, the landscape is not very smooth, and convergence is slower. When n/d grows, the number of iterations decreases and converges to a constant. This is also predicted by Theorem 5: the landscape of empirical risk will be as smooth as the landscape of population risk, as $n \geq C\, d \log d$.

5.2. *Binary linear classification: very high-dimensional regime .* In Figures 5, 6a, 6b, we present our results on non-convex binary linear classification in the very high-dimensional regime. Data $(Y_i, \boldsymbol{X}_i)$ were generated as in the previous section, with $\boldsymbol{\theta}_0$ a vector k non-zero entries all of size $1/\sqrt{k}$. We use proximal gradient descent to solve problem (3.12) with $r = 10$.

In Figure 5, we use random initializations $\boldsymbol{\theta}_s \sim \mathsf{N}(\mathbf{0}, \mathrm{I}_{d\times d}/d)$, and plot the empirical standard deviation of the resulting iterates $\mathrm{std}(\hat{\boldsymbol{\theta}}_n(i)) = \mathrm{Tr}(\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_n(i)))^{1/2}$.

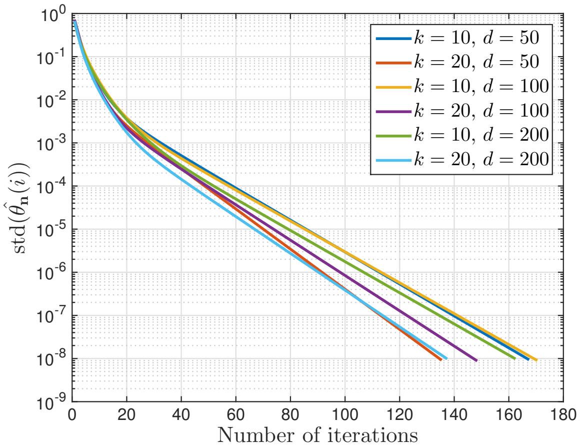


Fig 5: Binary linear classification, very high-dimensional. The standard deviation of each iteration point with respect to random initialization.

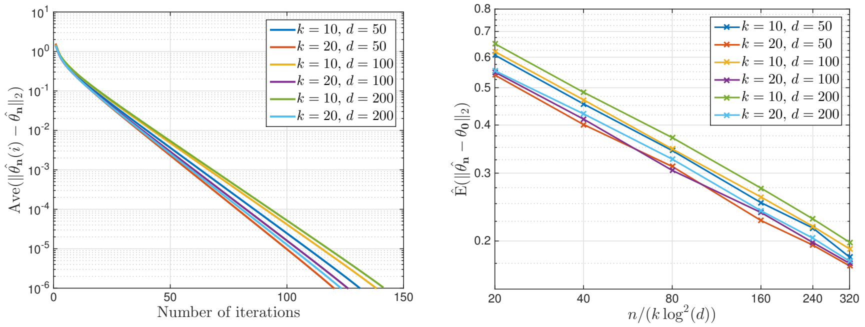
Note that the variance is taken over the random initializations, for a same realization of the data $(\boldsymbol{Y}, \boldsymbol{X})$, and hence captures smoothness (or roughness) of the empirical risk landscape. The standard deviation appears to converge exponentially fast to 0, confirming that indeed proximal gradient is converging to the unique local minimizer, as anticipated by Theorem 6.

In Figure 6a, we plot the expected distance from the global minimizer $\hat{\boldsymbol{\theta}}_n$ for each iterates. Proximal gradient appears to converge exponentially fast for $n \gg k \log^2(d)$.

5.3. *Robust linear regression .* In Figures 7, 8a, 8b we present simulations for robust regression. We generated random covariates $\boldsymbol{X}_i \sim \mathsf{N}(\boldsymbol{0}, \mathrm{I}_{d\times d})$ and responses $Y_i = \langle \boldsymbol{\theta}_0, \boldsymbol{X}_i \rangle + \varepsilon_i$, where $\|\boldsymbol{\theta}_0\|_2 = 1$. Again, we used projected gradient descent to solve the optimization problem (4.4) with $r = 10$. For the loss function we used Tukey's loss (4.5) with $t_0 = 4.685$.

In Fig. 7, we plot the standard deviation of the iterates $\mathrm{std}(\hat{\boldsymbol{\theta}}_n(i)) = \mathrm{Tr}(\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_n(i)))^{1/2}$ over random initializations $\boldsymbol{\theta}_s \sim \mathsf{N}(\boldsymbol{0}, 25\,\mathrm{I}_{d\times d}/d)$. In this case $\varepsilon_i \sim \mathsf{N}(0,1)$. Again, this standard deviation converges exponentially fast to 0 supporting the claim that proximal gradient descent converges to a unique global minimum irrespective of the initialization.

In Figures 8a, 8b we study the a contaminated model for the noise, namely $\varepsilon_i \sim (1-\delta)\mathsf{N}(0,1) + \delta\mathsf{N}(0,\sigma^2)$. In Figure 8a we plot the standard deviation of the estimates obtained with random initializations $\boldsymbol{\theta}_s \sim \mathsf{N}(\boldsymbol{0}, 25\,\mathrm{I}_{d\times d}/d)$, for $n = 480$, $d = 80$. Convergence rate remains exponential even for large contamination fraction. In Figure 8b we investigated the dependence of the



(a) (b)

Fig 6: Binary linear classification, very high-dimensional regime: (a) The convergence of proximal gradient descent. Here $\|\boldsymbol{\theta}_0\|_2 = 1$, and $n/(k\log^2(d)) = 20$, and $\lambda_n = 1/100 \cdot \sqrt{\log^2(d)/n}$. (b) Convergence of the statistical error.

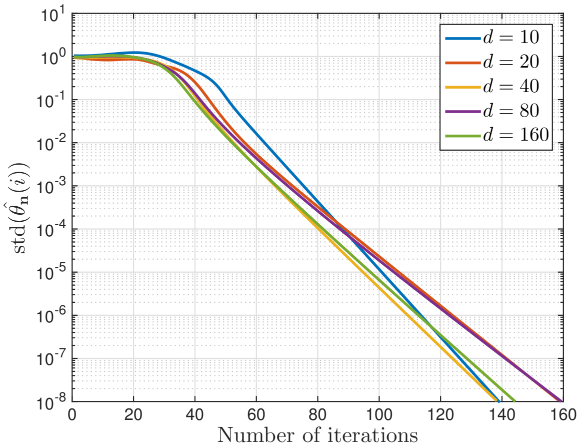
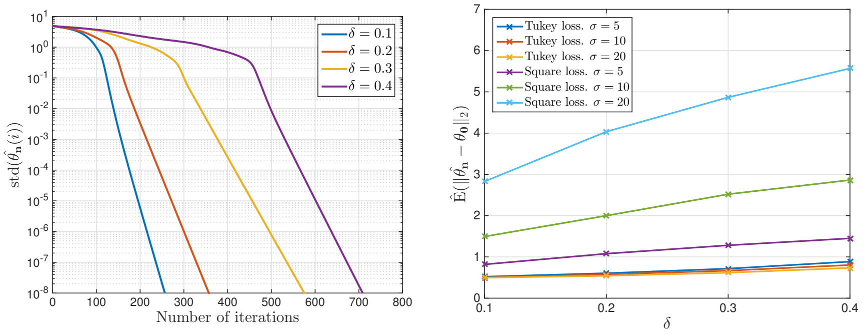


Fig 7: Robust regression. The standard deviation of each iteration point with respect to random initialization.



(a) (b)

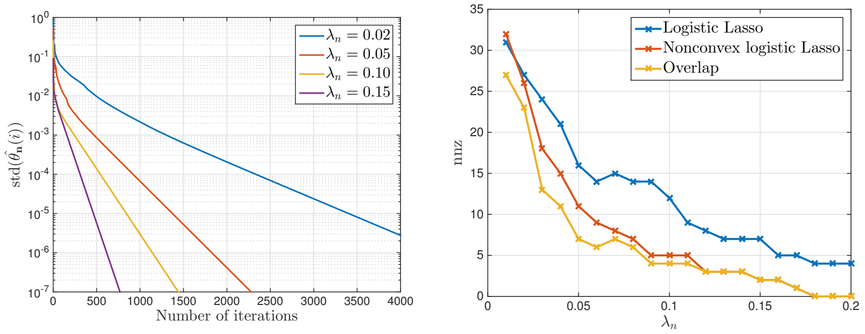
Fig 8: Robust regression: (a) The standard deviation of each iteration point with respect to random initialization, for different proportion of contamination. (b) The robustness of the global minimum between linear regression and Tukey regression.

estimation error on the contamination fraction, and the scale of outliers. Tukey's regression is fairly insensitive to outliers, while the least squares regression deteriorates as expected.

5.4. *Colon cancer data.* In Figures 9a, 9b we consider a gene-expression dataset from [ABN^{+}99]. The data set contains expression levels of of $2,000$ genes in 22 normal and 40 tumor colon tissues, hence $n = 62$ data points. Expression levels are normalized as in [ABN^{+}99] to have zero mean and unit standard deviation. We use the expression levels to form feature vectors $\boldsymbol{x}_i \in \mathbb{R}^d$, $d = 2000$ and encode the type of tissue using a binary label $y_i = 1$ (tumor) or $y_i = 0$ (no tissue).

We fit a model of the form $\mathbb{P}(Y_i = 1|\boldsymbol{X}_i = \boldsymbol{x}) = \sigma(\langle \boldsymbol{\theta}_0, \boldsymbol{x}\rangle)$ with $\sigma(u) = \sigma_L(u)$ the logistic function, by using the non-convex approach (4.3) and proximal gradient. We also used ℓ_1-regularized logistic regression, for comparison. Let us emphasize here that our focus here is not on the accuracy of the predictive model, but rather on showing that the non-convex approach is a viable alternative to the more standard regularized logistic regression.

In Figure 9a, we plot the standard deviation of the estimate $\hat{\boldsymbol{\theta}}_n(i)$, over random initializations $\boldsymbol{\theta}_s \sim \mathsf{N}(\mathbf{0}, \mathrm{I}_{d\times d}/d)$. The standard deviation decreases exponentially fast, suggesting that indeed the optimization problem has a unique local minimum. In Figure 9b we compare the model selected by the non-convex approach (4.3) to the one from ℓ_1-regularized logistic regression, and also plot the number of overlaps of their selected variables. Note that



(a) (b)

Fig 9: Colon cancer data: (a) The standard deviation of each iteration point with respect to random initialization, for different regularization parameter. (b) Number of non-zero elements of logistic Lasso and non-convex logistic Lasso.

most of the covariates selected by the non-convex regression method also appear in logistic regression. This suggests that the model produced by the non-convex approach is comparable to that produced by ℓ_1-regularized logistic regression.

SUPPLEMENTARY MATERIAL

Supplement A: Proofs and Simulations
(???). The supplement provides some technical background lemmas and gives all the proofs of the theorems, and additional numerical simulations.

References.

[ABN$^+$99] Uri Alon, Naama Barkai, Daniel A Notterman, Kurt Gish, Suzanne Ybarra, Daniel Mack, and Arnold J Levine, *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*, Proceedings of the National Academy of Sciences **96** (1999), no. 12, 6745–6750.

[AGJ15] Animashree Anandkumar, Rong Ge, and Majid Janzamin, *Learning overcomplete latent variable models through tensor methods*, Proceedings of the Conference on Learning Theory (COLT), Paris, France, 2015.

[ALPV14] Albert Ai, Alex Lapanowski, Yaniv Plan, and Roman Vershynin, *One-bit compressed sensing with non-gaussian measurements*, Linear Algebra and its Applications **441** (2014), 222–239.

[AS12] George K Atia and Venkatesh Saligrama, *Boolean compressed sensing and noisy group testing*, IEEE Transactions on Information Theory **58** (2012), no. 3, 1880–1901.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration inequalities: A nonasymptotic theory of independence*, OUP Oxford, 2013.

[BRT09] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov, *Simultaneous analysis of lasso and dantzig selector*, The Annals of Statistics (2009), 1705–1732.

[CC15] Yuxin Chen and Emmanuel Candes, *Solving random quadratic systems of equations is nearly as easy as solving linear systems*, Advances in Neural Information Processing Systems, 2015, pp. 739–747.

[CDT$^+$09] Olivier Chapelle, Chuong B Do, Choon H Teo, Quoc V Le, and Alex J Smola, *Tighter bounds for structured estimation*, Advances in neural information processing systems, 2009, pp. 281–288.

[CT05] Emmanuel J Candes and Terence Tao, *Decoding by linear programming*, IEEE transactions on information theory **51** (2005), no. 12, 4203–4215.

[CT07] Emmanuel Candes and Terence Tao, *The dantzig selector: Statistical estimation when p is much larger than n*, The Annals of Statistics (2007), 2313–2351.

[DFN12] Boris A Dubrovin, Anatolij Timofeevič Fomenko, and Sergeĭ Petrovich Novikov, *Modern geometrymethods and applications: Part ii: The geometry and topology of manifolds*, vol. 104, Springer, 2012.

[Don06] David L Donoho, *Compressed sensing*, IEEE Transactions on information theory **52** (2006), no. 4, 1289–1306.

[DTZ16] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis, *Ten steps of em suffice for mixtures of two gaussians*, arXiv:1609.00368 (2016).

[Fis22] Ronald Aylmer Fisher, *On the mathematical foundations of theoretical statistics*, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **222** (1922), 309–368.

[Fis25] ———, *Theory of statistical estimation*, Mathematical Proceedings of the Cambridge Philosophical Society, vol. 22, Cambridge Univ Press, 1925, pp. 700–725.

[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan, *Escaping from saddle pointsonline stochastic gradient for tensor decomposition*, Conference on Learning Theory, 2015, pp. 797–842.

[GP10] Victor Guillemin and Alan Pollack, *Differential topology*, vol. 370, American Mathematical Soc., 2010.

[Hub73] Peter J Huber, *Robust regression: asymptotics, conjectures and monte carlo*, The Annals of Statistics (1973), 799–821.

[KOM09] Raghunandan H Keshavan, Sewoong Oh, and Andrea Montanari, *Matrix completion from a few entries*, Information Theory, 2009. ISIT 2009. IEEE International Symposium on, IEEE, 2009, pp. 324–328.

[LB12] Jason N Laska and Richard G Baraniuk, *Regime change: Bit-depth versus measurement-rate in compressive sensing*, IEEE Transactions on Signal Processing **60** (2012), no. 7, 3496–3505.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, *Deep learning*, Nature **521** (2015), no. 7553, 436–444.

[LM13] Aurélie C Lozano and Nicolai Meinshausen, *Minimum distance estimation for robust high-dimensional regression*, arXiv:1307.3227 (2013).

[Loh15] Po-Ling Loh, *Statistical consistency and asymptotic normality for high-dimensional robust m-estimators*, arXiv:1501.00312 (2015).

[LW12] Po-Ling Loh and Martin J Wainwright, *High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity*, The Annals of Statistics (2012), 1637–1664.

[LW13] ______, *Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima*, Advances in Neural Information Processing Systems, 2013, pp. 476–484.

[LWYB11] Jason N Laska, Zaiwen Wen, Wotao Yin, and Richard G Baraniuk, *Trust, but verify: Fast and accurate signal recovery from 1-bit compressive measurements*, IEEE Transactions on Signal Processing **59** (2011), no. 11, 5289–5301.

[Mil63] John Milnor, *Morse theory*, vol. 51, Princeton University Press, 1963.

[Mil97] John Willard Milnor, *Topology from the differentiable viewpoint*, Princeton University Press, 1997.

[MR14] Andrea Montanari and Emile Richard, *A statistical model for tensor pca*, Advances in Neural Information Processing Systems, 2014, pp. 2897–2905.

[Nes13a] Yurii Nesterov, *Gradient methods for minimizing composite functions*, Mathematical Programming **140** (2013), no. 1, 125–161.

[Nes13b] ______, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer Science & Business Media, 2013.

[NRWY12] Sahand N Negahban, Padeep Ravikumar, Martin J Wainwright, and Bin Yu, *A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers*, Statistical science **27** (2012), no. 4, 538–557.

[NS13] Tan Nguyen and Scott Sanner, *Algorithms for direct 0–1 loss optimization in binary classification*, Proceedings of The 30th International Conference on Machine Learning, 2013, pp. 1085–1093.

[PV13a] Yaniv Plan and Roman Vershynin, *One-bit compressed sensing by linear programming*, Communications on Pure and Applied Mathematics **66** (2013), no. 8, 1275–1297.

[PV13b] ______, *Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach*, IEEE Transactions on Information Theory **59** (2013), no. 1, 482–494.

[PVY14] Yaniv Plan, Roman Vershynin, and Elena Yudovina, *High-dimensional estimation with geometric constraints*, arXiv preprint arXiv:1404.3749 (2014).

[PZB$^+$10] Jie Peng, Ji Zhu, Anna Bergamaschi, Wonshik Han, Dong-Young Noh, Jonathan R Pollack, and Pei Wang, *Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer*, The annals of applied statistics **4** (2010), no. 1, 53.

[RM51] Herbert Robbins and Sutton Monro, *A stochastic approximation method*, The annals of mathematical statistics (1951), 400–407.

[Ser13] V.I. Serdobolskii, *Multivariate statistical analysis: A high-dimensional approach*, vol. 41, Springer Science & Business Media, 2013.

[SQW16] Ju Sun, Qing Qu, and John Wright, *A geometric analysis of phase retrieval*, arXiv:1602.06664 (2016).

[TBA84] John N Tsitsiklis, Dimitri P Bertsekas, and Michael Athans, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, 1984 American Control Conference, 1984, pp. 484–489.

[Vap98] Vladimir Naumovich Vapnik, *Statistical learning theory*, vol. 1, Wiley New York, 1998.

[VdG00] Sara A Van de Geer, *Applications of empirical process theory*, vol. 91, Cambridge University Press Cambridge, 2000.

[WL12] Yichao Wu and Yufeng Liu, *Robust truncated hinge loss support vector machines*, Journal of the American Statistical Association (2012).

[XHM16] Ji Xu, Daniel Hsu, and Arian Maleki, *Global analysis of expectation maximization for mixtures of two gaussians*, arXiv preprint arXiv:1608.07630 (2016).

[YWL+15] Zhuoran Yang, Zhaoran Wang, Han Liu, Yonina C Eldar, and Tong Zhang, *Sparse nonlinear regression: Parameter estimation and asymptotic inference*, arXiv:1511.04514 (2015).

Song Mei
Stanford University
Huang Building 475 Via Ortega
Stanford, California 94305
USA
E-mail: songmei@stanford.edu

Yu Bai
Stanford University
390 Serra Mall
Stanford, California 94305
USA
E-mail: yub@stanford.edu

Andrea Montanari
Stanford University
350 Serra Mall
Stanford, California 94305
USA
E-mail: montanar@stanford.edu