JMLR: Workshop and Conference Proceedings vol 65:1–21, 2017

Fast Rates for Empirical Risk Minimization of Strict Saddle Problems

Alon Gonen ALONGNN@CS.HUJI.AC.IL and **Shai Shalev-Shwartz** SHAIS@CS.HUJI.AC.IL
School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel

Abstract

We derive bounds on the sample complexity of empirical risk minimization (ERM) in the context of minimizing non-convex risks that admit the strict saddle property. Recent progress in non-convex optimization has yielded efficient algorithms for minimizing such functions. Our results imply that these efficient algorithms are statistically stable and also generalize well. In particular, we derive fast rates which resemble the bounds that are often attained in the strongly convex setting. We specify our bounds to Principal Component Analysis and Independent Component Analysis. Our results and techniques may pave the way for statistical analyses of additional strict saddle problems.

1. Introduction

Stability analysis is a central tool in statistical learning theory (Bousquet and Elisseeff (2002)). Roughly speaking, a learning algorithm is stable if a slight change in the input of the algorithm does not change its output much. It was shown (Shalev-Shwartz et al. (2010); Mukherjee et al. (2006)) that stability characterizes learnability, and in particular, stability is equivalent to the estimation error of empirical risk minimization.

Stability analysis has been mostly carried out in the context of convex risk minimization. More concretely, some form of strong convexity is often assumed (e.g., exp-concavity in Koren and Levy (2015); Gonen and Shalev-Shwartz (2016)). The crux of the technique is to show that minima of two similar strongly convex (and Lipschitz/smooth) functions must be close ((Shalev-Shwartz and Ben-David, 2014, Section 13.3)).

In this paper we address the non-convex setting while restricting our attention to recently studied "nice" non-convex problems. Namely, we will consider non-convex functions which satisfy the strict saddle property (a.k.a. ridable or $\mathcal{X}$-functions, see Sun et al. (2015)). Roughly speaking, a strict saddle function has no spurious local minimum and its saddle points are strict, in the sense that second-order information suffices for identifying a descent direction. We also assume that the restriction of the function to a certain neighborhood of each of its minima is strongly convex.

Many important non-convex problems such as Principal Component Analysis (PCA), complete dictionary recovery (Sun et al. (2015)), tensor decomposition, ICA (Ge et al. (2015), Anandkumar et al. (2016)) and matrix completion (Ge et al. (2016); Bhojanapalli et al. (2016)) are strict-saddle. Furthermore, there exist efficient empirical risk minimizers (ERM) for these problems (e.g., SGD and Cubic Regularization, see Section 9).

2. Our contribution

We consider the problem of minimizing a risk of the form

$$F(w) = \mathbb{E}_{z\sim\mathcal{D}}[f(w,z)] \tag{1}$$

where for every $z \in \mathcal{Z}$, $f(\cdot, z)$ is a twice continuously differentiable *loss* function defined over the closed set $\mathcal{W} \subseteq \mathbb{R}^d$. Given an i.i.d. sample $S = (z_1, \ldots, z_d) \sim \mathcal{D}^n$, the output of an ERM algorithm is[1]

$$\hat{w} \in \arg\min_{w \in \mathcal{W}} \left\{ \hat{F}(w) = \frac{1}{n} \sum_{i=1}^{n} f_{z_i}(w) \right\} , \tag{2}$$

The sample complexity of ERM is the minimal size of a sample S for which $\mathbb{E}[F(\hat{w})] - \min_{w \in \mathcal{W}} F(w^\star) \leq \epsilon$.[2] We make the following assumptions on the loss functions:

(A1) For each z, $f(\cdot, z)$ is ρ-Lipschitz.
(A2) For each z, $f(\cdot, z)$ is twice continuously differentiable and

$$(\forall w \in \mathcal{W})\ (\forall i \in [d]) \quad |\lambda_i(\nabla^2 f(w, z))| \leq \beta_1 \ .$$

(A3) For each z, the Hessian of $f(\cdot, z)$ is β_2-Lipschitz.

While for each example of strict saddle objective listed above one may construct a dedicated sample complexity analysis, the goal of this paper is to provide a systematic unified approach, which emphasizes the geometric structure of the objective.

We distinguish between two cases. First, we consider the case where the empirical risk is strict saddle (with high probability) and prove stability and sample complexity bounds that depend solely on the strict saddle parameters of the empirical risk and the Lipschitz constants. In particular, the bound is dimensionality independent.

Theorem 1 *Let $\epsilon \in (0, 1)$. Suppose that that the empirical risk is (α, γ, τ)-strict saddle with high probability (see Section 3.2). Then the sample complexity of every ERM hypothesis is at most* $\max\left\{ \frac{\beta_1}{\gamma}, \frac{\rho}{\tau}, \frac{2\rho^2}{\alpha\epsilon} \right\}$.

In some applications it may be easier to prove that F itself is strict saddle. Under the additional assumption that $\mathcal{W}$ is bounded, we are able to prove the next theorem.

Theorem 2 *Suppose that F (Equation (1)) is (α, γ, τ)-strict saddle. The sample complexity is at most* $\tilde{O}\left(d \left(\frac{\rho}{\tau^2} + \frac{\beta_1}{\gamma^2} + \frac{\beta_1}{\alpha\epsilon} \right) \right)$.[3]

Remark 3 *The proof of this theorem actually reveals something stronger. Suppose we do not require all local minima of F to be optimally global and consider the family of empirical risk local minimizers. The same upper bound on the number of samples stated in Theorem 2 also suffices for ensuring that the value, $F(\hat{w})$, associated with the output of any such algorithm is ϵ-close to the value of some local minimum of F.*

We note that our bounds scale with $1/\epsilon$. In the literature, such bounds are often referred to as *fast rates*, because standard concentration bounds typically scale with $1/\epsilon^2$ (e.g., standard VC-dimension bounds in the agnostic setting ((Shalev-Shwartz and Ben-David, 2014, Theorem 6.8)).

1. We always assume the existence of a minima.
2. Alternatively, given ϵ and $\delta \in (0, 1)$, we ask for the minimal size of a sample S for which $F(\hat{w}) - \min_{w \in \mathcal{W}} F(w^\star) \leq \epsilon$ with probability at least $1 - \delta$.
3. The $\tilde{O}$ notation hides polylogarithmic dependencies.

2.1. Applications

2.1.1. PCA

In Section 6 we apply Theorem 1 to a stochastic formulation of Principal Component Analysis (PCA). Our goal is to approximately recover the leading eigenvector of the correlation matrix $\mathbb{E}[xx^\top]$, where x is drawn according to some unknown distribution $\mathcal{D}$ with bounded support. The standard measure of success is given by the non-convex objective $\min_{\|w\|=1} -w^\top \mathbb{E}[xx^\top]w$. It is known that the sample complexity of ERM for this problem is $\Omega(1/\epsilon^2)$ (Blanchard et al. (2007); Gonen et al. (2016)).

Better bounds can be achieved under eigengap assumptions: there exists a gap, denoted $G_{1,2}$, between the two leading eigenvalues of $\mathbb{E}[xx^\top]$. We can use the matrix Bernstein inequality to show that given an i.i.d. sample of size $n = \Omega(\log(d/\delta)/G_{1,2}^2)$, with probability at least $1-\delta$, a gap of the same order also appears in the empirical correlation matrix. We then show that if such a gap exists, then the empirical risk is strict-saddle, where the parameters are inversely proportional to $G_{1,2}$. This allows us to deduce a bound of order $1/(n \cdot G_{1,2})$ on the stability and the generalization error. We summarize the above in the next theorem.

Theorem 4 *The sample complexity of PCA is* $\tilde{O}\left(\frac{1}{G_{1,2}^2} + \frac{1}{\epsilon \cdot G_{1,2}}\right)$.

This bound is superior to the general $\tilde{O}(1/\epsilon^2)$ bound if $\epsilon = o(G_{1,2})$. One can claim that establishing the strict-saddle parameters of the empirical risk already requires statistical tools which usually already yield generalization bounds. Indeed, in the above example, one can use the matrix Bernstein inequality to show that $\tilde{O}(1/\epsilon_2^2)$ examples suffice in order to ensure that the expected distance between the true correlation matrix and the empirical correlation matrix (in operator norm) is at most ϵ. It is then straightforward to establish the standard $1/\epsilon^2$ bound on the generalization error. However, here we rely on Bernstein inequality only in order to ensure that the gap in $\mathbb{E}[xx^\top]$ appears also in the empirical correlation matrix. Consequently, we are able to prove a better bound (in a wide regime).

2.1.2. ICA

In Section 6 we apply Theorem 2 to a stochastic formulation of Independent Component Analysis (ICA). Let A be an orthonormal linear transformation. Suppose that x is uniform on $\{\pm 1\}^d$ and let $y = Ax$. Our goal is to recover the matrix A using the observations y. As was shown in Ge et al. (2015), this problem can be reduced to tensor decomposition. Moreover, the latter can be formulated as a strict saddle objective of the form (1), which can be efficiently minimized using SGD.

Theorem 5 *The sample complexity of ICA as formulated above is* $\tilde{O}\left(\mathrm{poly}(d) + \frac{d^{5/2}}{\epsilon}\right)$.

This result is meaningful in the regime where d is small and we are interested in a high accuracy solution.

2.2. Our approach

As we discussed above, most of the literature on stability analysis presumes some notion of strong convexity. Strict saddle objectives resemble strongly convex functions in the following sense: it is

provided that the restriction of the objective to a small neighborhood around any local minimum is strongly convex. However, there are several major differences. First, as opposed to strongly convex functions, there may exists several minima. More importantly, there are regions of the domain where the function is non-convex.

Our analysis essentially reduces to the strongly convex setting by excluding the other scenarios listed in Definition 8. Namely, we provide bounds on how many examples are needed in order to ensure that a minimizer corresponding to a slight change in the input must be in a strongly convex region around a local minimum $w^\star$. There is one more subtlety we need to tackle; we are not guaranteed that the minimizer of the (unmodified) empirical risk coincides with $w^\star$. However, as we shall see, since we deal with average stability and since all local minima are global, we may assume that this is the case w.l.o.g.

3. Preliminaries

3.1. Stability and generalization error

Definition 6 *Let $(z_1, \ldots, z_n) \sim \mathcal{D}^n$ and let $\hat{w}$ be an ERM (see Equation (2)). For every $i \in [n]$, let $\hat{w}_i \in \arg\min_w \frac{1}{n-1} \sum_{j \neq i} f_j(w)$ and let $\Delta_i = f_i(\hat{w}_i) - f_i(\hat{w})$.[4] We say that the ERM algorithm is on average stable with stability rate $\epsilon_{stab} : \mathbb{N} \to \mathbb{R}_{>0}$ if*

$$\Delta := \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \Delta_i\right] \leq \epsilon_{stab}(n) \ .$$

Here and in the sequel, the expectation is taken both over the randomness of the algorithm and the draw of $(z_1, \ldots, z_n)$.

For $(z_1, \ldots, z_n) \sim \mathcal{D}^n$, we define the generalization error of ERM by $\epsilon_{\text{gen}}(n) = \mathbb{E}[\hat{F}(\hat{w}) - F(\hat{w})]$. The next lemma relates the stability rate to the generalization error (see (Shalev-Shwartz and Ben-David, 2014, Theorem 13.2)).

Lemma 7 *For every n,*

$$\mathbb{E}_{S \sim \mathcal{D}^{n-1}}[L(\hat{w}) - L(w^\star)] \leq \mathbb{E}_{S \sim \mathcal{D}^n}[\Delta(S)]$$

Therefore, for every n, $\epsilon_{gen}(n) = \epsilon_{stab}(n)$.

3.2. Strict saddle functions

Due to their similarity to local extrema, saddle points raise a fundamental challenge to optimization algorithms. Intuitively, the easier saddle points are those for which second-order inrormation reveals a clear descent direction. The following definition due to Sun et al. (2015); Ge et al. (2015) captures this idea.

Definition 8 *A twice continuously differentiable function $\hat{F} : \mathbb{R}^d \to \mathbb{R}$ is called (α, γ, τ)-strict saddle, if it has no spurious local minimum, and for any point $x \in \mathbb{R}^d$ at least one of the following conditions holds:*

4. We do not assume uniqueness. The definition applies to any arbitrary rule for picking minimizers.

1. $\|\nabla \hat{F}(w))\| \geq \tau$
2. $\lambda_{\min}(\nabla^2 \hat{F}(w)) \leq -\gamma$
3. *There exists $\nu > 0$ and a local minimum $w^\star$ with $\|w - w^\star\| \leq \nu$, such that the restriction of $\hat{F}$ to 2ν-neighborhood of $w^\star$ is α-strongly convex.*[5]

Remark 9 *The requirement that every local minimum is globally optimal can be relaxed. Namely, for a desired accuracy $\epsilon > 0$, we may require that every local minimum is $\epsilon/2$-optimal. Extending our analysis to handle this case is straightforward.*

While Ge et al. (2015); Sun et al. (2015) also require a lower bound on the magnitude of ν (which appears in the last condition), it turns out that this quantity does not play any role in our analysis.

4. Stability Bounds for Strict Saddle Empirical Risks: Unconstrained Setting

In this section we consider the unconstrained setting (i.e., $\mathcal{W} = \mathbb{R}^d$). Our main result (Theorem 1) follows from the following theorem.

Theorem 10 *Let $\delta \in (0,1)$. Suppose that that the empirical risk $\hat{F}$ is (α, γ, τ)-strict saddle (Definition 8) with probability at least $1-\delta$. If $n > \max\left\{\frac{\rho}{\tau}, \frac{\beta_1}{\gamma}\right\}$, then with probability at least $1-\delta$, the expected generalization error and stability rate of ERM are bounded by*

$$\epsilon_{gen}(n) = \epsilon_{stab}(n) \leq \frac{2\rho^2}{\alpha n} .$$

The proof reduces to the strongly convex case by bounding the number of examples that are needed in order to exclude the first two scenarios listed in Definition 8. Throughout the rest of this section we assume that $\hat{F}$ is (α, γ, τ)-strict saddle.

Lemma 11 *Let $n > \rho/\tau$ and $(z_1, \ldots, z_n) \in \mathcal{Z}^n$. Then for any $i \in [n]$, $\|\nabla \hat{F}(\hat{w}_i)\| \leq \tau$.*

Proof Since $\hat{w}_i$ minimizes $\frac{1}{n}\sum_{j \neq i} f_j(w)$, we have that

$$\hat{g}_{-i} := \frac{1}{n}\sum_{j \neq i} \nabla f_j(\hat{w}_i) = 0 .$$

Therefore, using the triangle inequality and the Lipschitzness of each f_i, we obtain

$$\|\nabla \hat{F}(\hat{w}_i)\| \leq \|\hat{g}_{-i}\| + \frac{1}{n}\|\nabla f_i(\hat{w}_i)\| \leq 0 + \rho/n < \tau .$$

■

The proof of the next lemma has the same flavor.

Lemma 12 *Let $n > \beta_1/\gamma$ and $(z_1, \ldots, z_n) \in \mathcal{Z}^n$. Then for any $i \in [n]$, $\lambda_{\min}(\nabla^2 \hat{F}(\hat{w}_i)) > -\gamma$.*

5. That is, for all w in this neighborhood, $\nabla^2 F(w) \succeq \alpha I$

Proof By second-order conditions, $\hat{H}_{-i} := \frac{1}{n}\sum_{j\neq i}\nabla^2 f_j(\hat{w}_i)$ is positive semidefinite. Therefore, for all nonzero $v \in \mathbb{R}^d$

$$\frac{v^\top \nabla^2 \hat{F}(\hat{w}_i)v}{v^\top v} = \frac{v^\top \hat{H}_{-i}v}{v^\top v} + \frac{1}{n}\frac{v^\top \nabla^2 f_i(\hat{w}_i)v}{v^\top v} \geq 0 - \beta_1/n > -\gamma\,.$$

■

It follows that for $n > \max\{\rho/\tau, \beta_1/\gamma\}$, we only need to consider the third scenario listed in Definition 8.

Lemma 13 *For $n > \max\{\rho/\tau, \beta_1/\gamma\}$. Then,*

$$\epsilon_{gen}(n) = \epsilon_{stab}(n) = \frac{2\rho^2}{\alpha n}\,.$$

Proof Let $(z_1, \ldots, z_n) \in \mathcal{Z}^n$ for $n > \max\{\rho/\tau, \beta_1/\gamma\}$ and fix some $i \in [n]$. According to the previous two lemmas, $\hat{w}_i$ lies in a neighborhood around a local minimum $\bar{w}$ such that the restriction of $\hat{F}$ to this neighborhood is strongly convex. The crucial part is that since all the local minima are global, for the sake of upper bounding the stability we may assume w.l.o.g. that $\hat{w} = \bar{w}$. Indeed, the stability looks at the empirical risk of $\hat{w}$, which is equal to the empirical risk of $\bar{w}$ (here we can also allow an approximation error of order ϵ, see Remark 9). From here the proof follows along the lines of the standard proof in the Lipschitz and strongly convex case (e.g., see (Gonen and Shalev-Shwartz, 2016, Lemma 3)). We provide the details for completeness.

Fix some $i \in [n]$. By elementary properties of strongly convex functions, we have

$$\hat{F}(\hat{w}_i) - \hat{F}(\hat{w}) \geq \frac{\alpha}{2}\|\hat{w}_i - \hat{w}\|^2$$

On the other hand, since $\hat{w}_i$ minimizes the loss $w \in \mathcal{W} \mapsto \frac{1}{n}\sum_{j\neq i} f_j(w)$, the suboptimality of $\hat{w}_i$ w.r.t. the objective $\hat{F}$ is controlled by its suboptimality w.r.t. f_i, i.e.

$$\hat{F}(\hat{w}_i) - \hat{F}(\hat{w}) \leq \frac{1}{n}\Delta_i$$

Using Lipschitzness of f_i, we have

$$\Delta_i \leq \rho\|\hat{w}_i - \hat{w}\|$$

Combining the above, we obtain

$$\Delta_i^2 \leq \rho^2\|\hat{w}_i - \hat{w}\|^2 \leq \frac{2\rho^2}{\alpha}(\hat{F}(\hat{w}_i) - \hat{F}(\hat{w})) \leq \frac{2\rho^2}{\alpha n}\Delta_i$$

Dividing by Δ_i (we can assume w.l.o.g. that $\Delta_i > 0$) we conclude the proof. ■

This concludes the proof of Theorem 10.

5. Stability Bounds for Strict Saddle Empirical Risks: Constrained Setting

We now consider the case where $\mathcal{W}$ is described using equality constraints:

$$\mathcal{W} = \{w \in \mathbb{R}^d : c_i(w) = 0,\ i = 1, \ldots, m\}\,,$$

where for each i, $c_i(w)$ is twice continuously differentiable.

5.1. First and second-order conditions

In this part we recall basic facts on first and second-order conditions in the constrained setting (see for example Borwein and Lewis (2010)). We introduce the Lagrangian $\hat{L} : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$:

$$\hat{L}(w, \lambda) = \hat{F}(w) + \sum_{i=1}^{m} \lambda_i c_i(w) \; .$$

We call a vector $\lambda \in \mathbb{R}^m$ a Lagrange multiplier for $w \in \mathcal{W}$ if w is a critical point of $\hat{L}(\cdot, \lambda)$. A vector $w \in \mathcal{W}$ satisfies the linear independence constraint qualification (LICQ) condition if the set $\{\nabla c_i(w) : i \in [m]\}$ is linearly independent.

Theorem 14 ***(KKT conditions)*** *If $w \in \mathcal{W}$ is a local minimum of $\hat{F}$ and LICQ holds at w, then there exists a Lagrange multiplier λ for w.*

Note that λ can be found analytically using

$$\lambda(w) = -(C(w))^{\dagger} \nabla \hat{F}(w) \; ,$$

where C is the matrix whose columns are $\nabla c_1(w), \ldots, \nabla c_m(w)$. In the sequel we often use the notation

$$\hat{L}(w) = \hat{L}(w, \lambda(w)), \; \nabla \hat{L}(w) = \nabla_w \hat{L}(w, \lambda(w)), \; \nabla^2 \hat{L}(w) = \nabla^2_{ww} \hat{L}(w, \lambda(w)) \; .$$

The tangent space at any point $w \in \mathcal{W}$ is defined by $\mathcal{T}(w) = \{v \in \mathbb{R}^d : \; (\forall i \in [m]) \; v^\top \nabla c_i(w) = 0\}$. Following this notation, we observe that $\nabla \hat{L}(w)$ is simply the projection of $\nabla \hat{L}(w)$ onto the tangent space $\mathcal{T}(w)$. In particular, Theorem 14 provides conditions under which this projection vanishes. The next theorem extends the standard second-order conditions to our setting.

Theorem 15 ***(Second-order necessary conditions)*** *If $w \in \mathcal{W}$ is a local minimum of $\hat{L}$ and the set $\{\nabla c_i(w) : i \in [m]\}$ is linearly independent, then for all $v \in \mathcal{T}(w)$,*

$$v^\top \nabla^2 \hat{L}(w) v \geq 0 \; .$$

5.2. Strict saddle property in the constrained setting

We now provide a definition of the strict saddle property in the constrained setting.

Definition 16 *A twice continuously differentiable function $\hat{F} : \mathcal{W} \to \mathbb{R}$ with constrains $c_i(w)$ and associated Lagrangian L is called (α, γ, τ)-strict saddle if it has no spurious local minimum, and for any point $w \in \mathcal{W}$ at least one of the following conditions holds:*

1. $\|\nabla \hat{L}(w)\| \geq \tau$
2. *There exists a unit vector $v \in \mathcal{T}(w)$ s.t. $v^\top \nabla^2 L(w) v \leq -\gamma$*
3. *There exists a local minimum $w^\star$ such that*

$$\frac{\|\nabla L(w)\|^2}{2\alpha} \geq \hat{L}(w) - \hat{L}(w^\star) \geq \frac{\alpha}{2} \|w - w^\star\|^2$$

While our last condition is slightly different from its counterparts in Ge et al. (2015); Sun et al. (2015), we argue that it is often easier to establish the condition stated here (e.g., see Appendix B).[6]

6. Actually, it seems that our condition is also required in the proof of Ge et al. (2015)[Lemma 34] (see equation 121).

5.3. Analysis in the constrained setting

Throughout the section we prove that Theorem 1 holds also in the constrained setting. We assume that $\mathcal{W}$ is described using m equality constraints of the form $c_i(w) = 0$ and that the LICQ holds for all $w \in \mathcal{W}$.

As in the constrained setting, we first bound the number of examples that are needed in order to exclude the two first scenarios listed in Definition 16.

Lemma 17 *Let $n > \rho/\tau$ and $(z_1, \ldots, z_m) \in \mathcal{Z}^n$. Then for any $i \in [n]$, $\|\nabla \hat{L}(\hat{w}_i)\| \leq \tau$.*

Proof Since $\hat{w}_i$ minimizes the risk w.r.t. $\frac{1}{n}\sum_{j \neq i} f_j(w)$, we have that

$$\tilde{g}_{-i} = \frac{1}{n}\sum_{j \neq i} \nabla f_j(\hat{w}_i) - \sum_{s=1}^{m} \lambda_s(\hat{w}_i)\nabla c_s(w) = 0 \,.$$

Therefore, using the triangle inequality, we obtain

$$\|\nabla \hat{L}(\hat{w}_i)\| \leq \|\tilde{g}_{-i}\| + \frac{1}{n}\|\nabla f_i(\hat{w}_i)\| \leq \rho/n < \tau \,.$$

■

Lemma 18 *Let $n > \beta_1/\gamma$ and $(z_1, \ldots, z_m) \in \mathcal{Z}^n$. Then for any $i \in [n]$ and $v \in \mathcal{T}(\hat{w}_i)$ $v^\top \nabla^2(\hat{L}(\hat{w}_i))v \geq -\gamma$.*

Proof By second-order conditions, when restricted to $\mathcal{T}(\hat{w}_i)$, $\tilde{H}_{-i} := \frac{1}{n}\sum_{j \neq i} \nabla^2 f_j(\hat{w}_i) + \sum_{s=1}^{m} \lambda_s(w)\nabla^2 c_s(w)$ is positive semidefinite. Therefore, for every (nonzero) $v \in \mathcal{T}(\hat{w}_i)$,

$$\frac{v^\top \nabla^2 \hat{L}(\hat{w}_i) v}{v^\top v} \geq \frac{v^\top \tilde{H}_{-i} v}{v^\top v} + \frac{1}{n}\frac{v^\top \nabla^2 f_i(\hat{w}_i) v}{v^\top v} \geq 0 - \beta_1/n > -\gamma \,.$$

■

It follows that for $n > \max\{\rho/\tau, \beta_1/\gamma\}$, we only need to consider the third scenario listed in Definition 16. The proof of the next lemma is almost identical to the proof of Lemma 13 and is therefore given in the appendix (Appendix C).

Lemma 19 *For $n > \max\{\rho/\tau, \beta_1/\gamma\}$ we have:*

$$\epsilon_{gen}(n) = \epsilon_{stab}(n) \leq \frac{2\rho^2}{\alpha n} \,.$$

6. Application to PCA

Consider the following stochastic formulation of PCA. Let $\mathcal{D}$ be a distribution over $\mathcal{Z} \subseteq \mathbb{R}^d$. We are interested in minimizing the objective

$$F(w) = \frac{1}{2}\mathbb{E}_{z \sim \mathcal{D}}[\|z - ww^\top z\|^2]$$

over all possible unit vectors $w \in \mathbb{R}^d$. We assume for simplicity that $\mathcal{Z}$ is contained in the Euclidean unit ball. It is well known that the minimum is the leading eigenvector of the positive definite matrix $\mathbb{E}[zz^\top]$. As we shall see, this problem becomes strict saddle once we make the following standard assumption:

(A4) There is a positive gap, denoted $G_{1,2}$, between the two leading eigenvalues of $\mathbb{E}[xx^\top]$.

Given a sample $(z_1, \ldots, z_n) \sim \mathcal{D}^n$, let us denote by $A = \frac{1}{n}\sum_{i=1}^n z_i z_i^\top$. The empirical risk is given by

$$\bar{F}(w) = \frac{1}{2n}\sum_{i=1}^n \|z_i - ww^\top z_i\|^2$$

One can easily see that an equivalent objective is given by

$$\hat{F}(w) = -\frac{1}{2}w^\top A w \,.$$

Hence, the empirical risk admits exactly two (local and global) minima, namely u and $-u$, where u is the leading eigenvector of A.

We now would like to show that for sufficiently large n, the empirical risk is strict saddle. The first step should be to translate our eigengap assumption on $\mathbb{E}[zz^\top]$ to a similar assumption on A. The following lemma, which follows from a simple application of the Matrix Bernstein inequality (Tropp (2015)[Section 1.6.3]), shows that for sufficiently large n, the eigengap between the two leading eigenvalues of A is $\Omega(G_{1,2})$.

Lemma 20 *Let $\delta \in (0,1)$. For $n = \Omega\left(\frac{\log(d/\delta)}{G_{1,2}^2}\right)$, we have that with probability at least $1-\delta$,*

$$\|A - \mathbb{E}[xx^\top]\| \le G_{1,2}/2 =: G$$

It follows that with probability at least $1-\delta$, the gap between the leading eigenvalues of A is at least G.

The following theorem implies Theorem 4.

Theorem 21 *For any $\delta \in (0,1)$, if the sample size n is $\Omega\left(\frac{\log(d/\delta)}{G_{1,2}^2}\right)$, then with probability at least $1-\delta$, the PCA objective satisfies the conditions in Definition 16 with $\tau, \gamma, \alpha \in \Omega(G_{1,2})$. Consequently, for any $n = \Omega\left(\frac{\log(d/\delta)}{G_{1,2}^2}\right)$,*

$$\epsilon_{gen}(n) = \epsilon_{stab}(n) \le \frac{4}{n \cdot G_{1,2}} \,.$$

Proof (idea) Critical points of the Lagrangian correspond to eigenvectors of A (where we refer to the zero vector as an eigenvector as well). We show that if the gradient at some point w is small, then w either belongs to a strongly convex region around the leading eigenvector or to a strict saddle neighborhood of another eigenvector (or 0). ■

The proof is given in Appendix A.

7. Sample Complexity Bounds for Strict Saddle Expected Risks

In some cases it may be easier to establish the strict saddle property of the expected risk (Equation (1)). We now assume that F is (α, τ, γ)-strict saddle. We consider the constrained setting and denote the Lagrangian of F by L. We add the following boundedness assumption:

(A4) The set $\mathcal{W}$ is contained in $\{w : \|w\| \leq B\}$.

The proof of Theorem 2 is given in Appendix C. Below we give the main idea.

Proof (idea) of Theorem 2 We use Matrix Bernstein inequality together with covering to show that with high probability, points with large gradient do not form minima of $\hat{F}$. Similar argument shows that strict saddle points of F do not become minima of $\hat{L}$. Then, we can restrict ourselves to strongly convex regions of F and show that any w with $F(w) - \min_{w' \in \mathcal{W}} F(w') > \epsilon$ can not be a minimum of $\hat{F}$. ■

8. Application to ICA Through Tensor Decomposition

A p-order tensor is a p-dimensional array. Here we focus on 4-order tensors. For a tensor $T \in \mathbb{R}^{d^4}$ and indices $i_1, \ldots, i_4 \in [d]$, we denote the $(i_1, \ldots, i_4)$-th entry of T by $T_{i_1,\ldots,i_4}$. Every d-dimensional vector a induces a rank-one 4-order tensor, denoted $a^{\otimes 4}$, where $a^{\otimes 4}_{i_1,i_2,i_3,i_4}$ is $a_{i_1} a_{i_2} a_{i_3} a_{i_4}$. We can present the tensor T using a multilinear form. Given vectors $u, v, z, w \in \mathbb{R}^d$, we define

$$T(u, v, z, w) = \sum_{i_1,i_2,i_3,i_4} T_{i_1,\ldots,i_4} u_{i_1} v_{i_2} z_{i_3} w_{i_4}$$

The tensor T has an orthogonal decomposition if it can be written as

$$T = \sum_{i=1}^{d} a_i^{\otimes 4} \,. \tag{3}$$

In case that such decomposition exists, it is unique up to a permutation of the a_i's and sign flips. A central problem in machine learning is to compute the tensor decomposition of a given tensor T (Anandkumar et al. (2014)). While we have exponentially many equivalent solutions, the average of two solutions does not form a solution. Hence, any reasonable formulation of this problem must be non-convex. Luckily, as was shown in Ge et al. (2015), there exists a strict saddle formulation of this problem.

For simplicity, we consider the problem of finding one component (one can proceed and find all the components using deflation). Consider the following objective:

$$\max_{\|u\|=1} T(u, u, u, u) \,. \tag{4}$$

Lemma 22 *(Ge et al. (2015)) Suppose that T admits a Tensor decomposition as in (3). The only local minima of (4) are $\pm a_i$. Furthermore, the objective (4) is (α, γ, τ)-strict saddle with $\alpha = \Omega(1), \gamma = 7/d$ and $\tau = 1/\mathrm{poly}(d)$. Last, for $p = 1, 2, 3$, the magnitude of the p-th order derivative of this objective is $O(\sqrt{d})$.*

Although our definition of strict saddle functions in the constrained setting is slightly different from its counterpart in Ge et al. (2015), it is not hard to show that Lemma 22 still holds (see Appendix B).

In applications, we often have access to T only through a stochastic oracle. Following Ge et al. (2015), we consider the following formulation of ICA. Let A be an orthonormal linear transformation. Suppose that x is uniform on $\{\pm 1\}^d$ and denote by $y = Ax$. Our goal is to recover the matrix A using the observations y. It turns out that ICA reduces to tensor decomposition. Namely, define $Z \in \mathbb{R}^{d^4}$ by

$$(\forall i \in [d])\ \ Z(i,i,i,i) = 3,\ \ (\forall i \neq j)\ \ Z(i,i,j,j) = Z(i,j,j,i) = Z(i,j,i,j) = 1\ ,$$

where all other entries of Z are zero.

Lemma 23 *The expectation $\frac{1}{2}\mathbb{E}[Z - y^{\otimes 4}]$ is equal to T, where the vectors participating in the decomposition of T correspond to columns of A.*

Following the lemma, we can rewrite (4) as the following expected risk:

$$\max_{\|u\|=1} \mathbb{E}\left[\frac{1}{2}\left(Z - y^{\otimes 4}\right)\right](u,u,u,u)\ . \tag{5}$$

Furthermore, as was shown in Ge et al. (2015), one can efficiently compute a stochastic gradient and use SGD to optimize this objective. Using Lemma 22 and Theorem 2, we conclude that the sample complexity of extracting a single column of A is $\tilde{O}\left(\mathrm{poly}(d) + \frac{d^{3/2}}{\epsilon}\right)$. The sample complexity of extracting all the columns is $\tilde{O}\left(\mathrm{poly}(d) + \frac{d^{5/2}}{\epsilon}\right)$.

9. Related Work

9.1. Efficient ERM for Strict Saddle Functions

There is a growing interest in developing efficient algorithms for minimization of strict saddle functions. We mention two central approaches. Intuitively, one can escape from a saddle point by moving in the direction of the eigenvector corresponding to the minimal eigenvalue. This intuition has been made precise by Nesterov and Polyak (Nesterov and Polyak (2006)). More surprisingly, in Ge et al. (2015) it was shown that a variant of SGD also converges to a local minimum. Recent improvements in terms of runtime are given in Agarwal et al. (2016); Levy (2016).

9.2. Stability of SGD

Recently, Hardt et al. (2015) analyzed the stability of the SGD algorithm both in a convex and non-convex setting. As we mentioned above, in our setting, SGD forms an empirical risk minimizer. Our bounds on the stability rate of SGD in this setting improve over the (more general) bounds of Hardt et al. (2015). In particular, our bounds imply that SGD can be trained for arbitrarily long time.

9.3. Generalization Bounds using SGD

It is known that one can obtain generalization bounds directly using SGD (Shalev-Shwartz and Ben-David (2014)[Chapter 14]). Hence, the time complexity bound of Ge et al. (2015) translates into identical sample complexity bound. However, their bounds, which scale with $1/\epsilon^4$, are inferior to our bounds when high accuracy is desired.

9.4. Fast rates for PCA

Generalization bounds for stochastic PCA have been studied in Bousquet and Elisseeff (2002); Gonen et al. (2016). Both works prove an upper bound of $1/\sqrt{n}$ on the generalization error in the general case. The latter work (which also considers the challenge of partial information) establishes a matching lower bound. The former work also considers the case of a positive eigengap between the leading eigenvalues of $\mathbb{E}[xx^\top]$[7] and establishes fast rates similar to our bounds using Local Rademacher complexities. We believe that these techniques are much more involved than our techniques and lack any geometric interpretation.

Acknowledgments

We thank Kfir Levy for bringing Remark 3 into our attention. We also thank Nati Srebro for helpful discussions.

7. More generally, these works consider the task of approximating the k leading eigenvectors. It is not hard to extend our results to this task as well.

References

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding local minima for nonconvex optimization in linear time. *arXiv preprint arXiv:1611.01146*, 2016.

Anima Anandkumar, Yuan Deng, Rong Ge, and Hossein Mobah. Homotopy method for tensor principal component analysis. *arXiv preprint arXiv:1610.09322*, 2016.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15 (1):2773–2832, 2014.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221*, 2016.

Gilles Blanchard, Olivier Bousquet, and Laurent Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2-3):259–294, 2007.

Jonathan M Borwein and Adrian S Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media, 2010.

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

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points-online stochastic gradient descent for tensor decomposition. In *Proceedings of The 29th Conference on Learning Theory*, pages 797–842, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

Alon Gonen and Shai Shalev-Shwartz. Average stability is invariant to data preconditioning. implications to exp-concave empirical risk minimization. *arXiv preprint arXiv:1601.04011*, 2016.

Alon Gonen, Dan Rosenbaum, Yonina C Eldar, and Shai Shalev-Shwartz. Subspace learning with partial information. *Journal of Machine Learning Research*, 17(52):1–21, 2016.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

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

Kfir Y Levy. The power of normalization: Faster evasion of saddle points. *arXiv preprint arXiv:1611.04831*, 2016.

Jiří Matoušek. *Lectures on discrete geometry*, volume 108. Springer New York, 2002.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006.

Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

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

Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

Appendix A. PCA Is Strict Saddle: Complete Proof

This section is devoted to the proof of Theorem 21. Let us start with some basic calculations. The gradient and the Hessian of $\hat{F}(w)$ are given by

$$\nabla \hat{F}(w) = -Aw, \;\; \nabla^2 \hat{F}(w) = -A \, .$$

It is apparent that both the domain and the the objective are not convex. The following lemma is immediate.

Lemma 24 *The restriction of $\hat{F}$ to the unit sphere in $\mathbb{R}^d$ is 1-Lipschitz and 1-smooth.*

Letting $c(w) = \frac{1}{2}(\|w\|^2 - 1)$, the Lagrangian is given by

$$\hat{L}(w, \lambda) = \hat{F}(w) + \lambda c(w) = -\frac{1}{2} w^\top A w + \frac{\lambda}{2}(\|w\|^2 - 1) \, .$$

It follows that

$$\lambda(w) = w^\top A w \, .$$

Therefore, the gradient and the Hessian of $\hat{L}(w)$ are given by

$$\nabla \hat{L}(w) = (\lambda(w) I - A) w, \;\; \nabla^2 \hat{L}(w) = (\lambda(w) I - A)$$

Note also that LICQ trivially holds at any point $w \in \mathcal{W}$.

Proof (of Theorem 21) Let w be a unit vector in $\mathbb{R}^d$ and suppose that $\|\nabla f(w)\| \leq \tau = cG$ for some constant $c \in (0, 1/32)$. We show that w satisfies either the second or the third condition in Definition 16.

First step (setup):
Let $w = \sum_{i=1}^d \alpha_i u_i$ be the decomposition of w according to the eigenbasis of A. Note that by the optimality of u_1, $\lambda \leq \lambda_1$. Also, by assumption

$$\tau^2 \geq \|(\lambda I - A) w\|^2 = w^\top \sum_{i=1}^d (\lambda - \lambda_i)^2 u_i u_i^\top w = \sum_{i=1}^d \alpha_i^2 (\lambda - \lambda_i)^2 \, . \tag{6}$$

Second step (bounding the mass of distant eigenvalues):
Note that $\|\alpha\|^2 = 1$, hence the vector $\alpha^2 = (\alpha_1^2, \ldots, \alpha_d^2)$ can be seen as a probability vector. We next apply Markov's inequality in order to bound the mass of eigenvalues located far from λ. For every $t = 0, 1, \ldots$, define

$$I_t = \{i \in [d] : |\lambda - \lambda_i| \leq 2^t \tau\} \, .$$

We claim that for every t,

$$\sum_{i \notin I_t} \alpha_i^2 \leq 2^{-2t} \, . \tag{7}$$

Indeed, for $t = 0$ the bound is trivial and for $t \geq 1$ we apply (6) to otbain

$$\tau^2 \geq \sum_{i \notin I_t} \alpha_i^2 (\lambda - \lambda_i)^2 \geq 2^{2t} \tau^2 \sum_{i \notin I_t} \alpha_i^2 \, .$$

By rearranging, we conclude the claim.

Third step (the strongly convex case):
Consider the case where $1 \in I_4$. It follows that

$$\lambda_1 - 16cG = \lambda_1 - 2^4\tau \leq \lambda = \sum_{i=1}^{d} \alpha_i^2 \lambda_i \leq \alpha_1^2 \lambda_1 + \sum_{i=2}^{d} \alpha_i^2 (\lambda_1 - G) = \lambda_1 - G \sum_{i=2}^{d} \alpha_i^2 \,,$$

where the last equality uses the fact that $\sum_{i=1}^{d} \alpha_i^2 = 1$. Hence, $\sum_{i=2}^{d} \alpha_i^2 \leq 16c$, so

$$\alpha_1^2 \geq (1 - 16c) \geq 1/2 \Rightarrow \sum_{i \geq 2} \alpha_i^2 \leq 1/2 \,. \tag{8}$$

We now show that $\hat{F}(w) - \hat{F}(u_1) \geq \frac{G}{4}\|w - u_1\|^2$. First we calculate the distance between w and u_1:

$$\|w - u_1\|^2 = (\alpha_1 - 1)^2 + \sum_{i \geq 2} \alpha_i^2 = \sum_{i=1}^{d} \alpha_i^2 + 1 - 2\alpha_1 = 2(1 - \alpha_1) \,. \tag{9}$$

Since w and u_1 are feasible, $\hat{F}(w) = \hat{L}(w)$ and $\hat{F}(u_1) = \hat{L}(u_1)$. Since $\hat{L}$ is quadratic and u_1 is optimal (hence $\nabla \hat{L}(u_1) = 0$), we have

$$\hat{F}(w) = \hat{F}(u_1) + \langle \nabla \hat{L}(u_1), w - u_1 \rangle + \frac{1}{2}(w - u_1)^\top \nabla^2 \hat{L}(u_1)(w - u_1) = \hat{L}(u_1) + \frac{1}{2}(w - u_1)^\top \nabla^2 \hat{L}(u)(w - u_1)$$

It is left to bound the quadratic term from below. Since $0 \leq \lambda_1 - \lambda \leq 16cG$ for $c \in (0, 1/32)$,

$$\lambda_1 - \lambda \leq G/2 \Rightarrow (\forall i \geq 2) \quad \lambda - \lambda_i \geq G/2 \,. \tag{10}$$

Therefore,

$$\begin{aligned}
\frac{1}{2}(w - u_1)^\top \nabla^2 \hat{L}(u)(w - u_1) &= (\alpha_1 - 1)^2(\lambda_1 - \lambda_1) + \sum_{i \geq 2} \alpha_i^2 (\lambda_1 - \lambda_i) \\
&\geq G \sum_{i \geq 2} \alpha_i^2 \geq G(-(\alpha_1 - 1)^2 + \sum_{i \geq 2} \alpha_i^2) \\
&= \frac{G}{2}(\sum_{i=1}^{d} \alpha_i^2 - 2\alpha_1^2 + 2\alpha_1 - 1) = \frac{G}{2}(2\alpha_1 - 2\alpha_1^2) \\
&= \frac{G}{2} 2\alpha_1(1 - \alpha_1) \underbrace{=}_{(9)} \frac{G}{2}\alpha_1 \|w - u_1\|^2 \\
&\underbrace{\geq}_{(8)} \frac{G}{4}\|w - u_1\|^2 \,,
\end{aligned}$$

We deduce that

$$\hat{F}(w) - \hat{F}(u_1) \geq \frac{G}{4}\|w - u_1\|^2 \,.$$

On the other hand,

$$\begin{aligned}\frac{1}{2}(w-u_1)^\top \nabla^2 \hat{L}(u)(w-u_1) &= \sum_{i\geq 2}\alpha_i^2(\lambda-\lambda_i+\lambda_1-\lambda) \underbrace{\leq}_{8,10} \sum_{i\geq 2}\alpha_i^2(\lambda-\lambda_i+\lambda_1-\lambda) \\ &\quad + \alpha_1^2(\lambda-\lambda_i) - \sum_{i\geq 2}\alpha_i^2(\lambda_1-\lambda) = \sum_{i\geq 2}\alpha_i^2(\lambda-\lambda_i) \\ &\underbrace{\leq}_{10} \sum_{i\geq 2}\alpha_i^2(\lambda-\lambda_i)^2/(G/2) \leq \sum_{i\geq 1}\alpha_i^2(\lambda-\lambda_i)^2/(G/2) \\ &= \frac{\|\nabla L(w)\|^2}{2(G/4)} \,.\end{aligned}$$

Fourth step (the strict saddle case):
Consider the case where $1 \notin I_4$. We construct a vector $v \in \mathcal{T}(w)$ such that $\frac{v^\top \nabla^2 \hat{L}(w) v}{\|v\|^2}$ is proportional to $-G$. Let

$$v = u_1 - \alpha_1 w$$

Note that v is perpendicular to w, hence $v \in \mathcal{T}(w)$. Also note that

$$v = (1-\alpha_1^2)u_1 - \alpha_1 \sum_{i\geq 2}\alpha_i u_i$$

Hence,

$$v^\top \nabla^2 \hat{L}(w) v = (1-\alpha_1^2)(\lambda-\lambda_1) + \sum_{i\geq 2}\alpha_i^2(\lambda-\lambda_i) \,.$$

We bound each of the terms in the RHS. Using (7) we upper bound α_1^2 by 2^{-8}. Since $\lambda \leq \lambda_1$, we have

$$(1-\alpha_1^2)(\lambda-\lambda_1) \leq -\frac{255}{256}\cdot 16\tau \leq -15\tau \,.$$

On the other hand, denoting $J_t = I_t \setminus \bigcup_{s=0}^{t-1} I_s$, we have

$$\begin{aligned}\sum_{j\geq 2}\alpha_j^2(\lambda-\lambda_j) \leq \sum_{j\geq 1}\alpha_j^2|\lambda-\lambda_j| &= \sum_{t=0}^{\infty}\sum_{j\in J_t}\alpha_j^2|\lambda-\lambda_i| \leq \sum_{t=0}^{\infty}\sum_{j\in J_t}\alpha_j^2 2^t\tau \\ &=\leq \tau\sum_{t=0}^{\infty}2^{-2t}2^t = 2\tau \,,\end{aligned}$$

where the last inequality follows from (7). Note also that $\|v\| \leq 2$. Overall, we obtain that

$$\frac{v^\top \nabla^2 \hat{L}(w) v}{\|v\|^2} \leq -13\tau/2 \leq -6cG \,.$$

■

Appendix B. ICA is Strict Saddle: Establishing Strong Convexity

Our notion of strong convexity in Definition 16 is slightly different from its counterpart in Ge et al. (2015). We now show that Lemma 22 holds using our definitions.

Let $w \in \mathcal{W}$. To simplify the presentation, we assume that $a_i = e_i$ for all i (alternatively, we could do a change of coordinates to w, which does not affect the structure of the problem). Denote

$$\tau_0 = (10d)^{-4},\ \tau = 4\tau_0^2,\ D = 2d\tau_0,\ I(w) = \{i \in [d] : |w_i| > \tau_0\}$$

Suppose that $\|\nabla L(w)\| \leq \tau$, where L is the Lagrangian associated with the expected risk F. It was shown in Ge et al. (2015) that if $|I(w)| \geq 2$, then w is a strict saddle point. Hence, it is left to consider the case where $|I(w)| = 1$. Assume w.l.o.g. that $I(w) = \{1\}$.

Lemma 25 *The suboptimality of w w.t.t. the minimum e_1 is bounded below by*

$$F(w) - F(e_1) \geq \frac{1}{4}\|w - e_1\|^2 .$$

Proof Since w is a unit vector,

$$1 \geq w_1^2 = 1 - \sum_{i \geq 2} w_i^2 \geq 1 - d\tau_0^2$$

The squared distance between w and the local minimum e_1 is at most

$$\|w - e_1\|^2 = (1 - w_1)^2 + \sum_{i \geq 2} w_i^2 \leq 2d\tau_0^2 \leq D^2 .$$

Let $c(w) = \frac{1}{2}(\|w\|^2 - 1)$. Since $c(w) = c(e_1) = 0$, using the 1-smoothness of c we obtain

$$0 = c(w) \leq c(e_1) + \nabla c(e_1)^\top (w - e_1) + \frac{1}{2}\|w - e_1\|^2 = e_1(w - e_1) .$$

Hence,

$$(1 - w_1)^2 = (e_1^\top (e_1 - w))^2 \leq \frac{1}{4}\|w - e_1\|^4 \leq \frac{1}{4}\|w - e_1\|^2 \tag{11}$$

As Ge et al. (2015) show, The Hessian of L at e_1 is a diagonal matrix with 4 on the diagonals except for the first diagonal entry whose value is -8. Since $F(w) = L(w)$ and $F(e_1) = L(e_1)$,

$$F(w) = F(w_1) + \underbrace{\nabla L(e_1)^\top}_{=0}(w - e_1) + \frac{1}{2}(w - e_1)^\top \nabla^2 L(w')(w - e_1)$$

for some w' that lies on the line between w and e_1. Note that

$$\begin{aligned}
&\frac{1}{2}(w - e_1)^\top \nabla^2 L(w')(w - e_1) \\
&= \frac{1}{2}(w - e_1)^\top \nabla^2 L(e_1)(w - e_1) + \frac{1}{2}(w - e_1)^\top (\nabla^2 L(w') - \nabla^2 L(e_1))(w - e_1) .
\end{aligned}$$

Using (11), we bound the first term in the RHS by

$$(w-e_1)^\top \nabla^2 L(e_1)(w-e_1) = -8(1-w_1)^2 + 4\sum_{i\geq 2} w_i^2 = 4((1-w_1)^2 + \sum_{i\geq 2} w_i^2) - 12(1-w_1)^2$$
$$\geq 4\|w-u_1\|^2 - 3\|w-u_1\|^2 = \|w-u_1\|^2$$

Using the $O(\sqrt{d})$-Lipschitzness of the Hessian and the fact that $\|w'-e_1\| \leq D$, the second term is bounded by

$$(w-e_1)^\top(\nabla^2 L(w') - \nabla^2 L(e_1))(w-e_1) \leq \|w-e_1\|^2\|w'-e_1\|\sqrt{d} \leq \frac{1}{2}\|w-e_1\|^2 .$$

All in all,

$$F(w) - F(e_1) \geq \frac{1}{4}\|w-e_1\|^2 .$$

■

Lemma 26 *The suboptimality of w w.t.t. the minimum e_1 is bounded above by*

$$F(w) - F(e_1) \leq O(\|\nabla L(w)\|^2) .$$

Proof Using the previous lemma and the Lipschitzness of the Hessian, one can easily show that

$$F(e_1) \geq F(w) + \nabla L(w)^\top (e_1 - w) + \frac{c}{2}\|e_1 - w\|^2$$

for some constant $c \in (0,1)$. The RHS is at most

$$\min_{z\in\mathbb{R}^d} F(w) + \nabla L(w)^\top (z-w) + \frac{c}{2}\|z-w\|^2$$

The minimum is attained at $z = w - c^{-1}\nabla L(w)$. The desired inequality follows by substitution. ■

Appendix C. Omitted Proofs

Proof (of Lemma 19) According to the previous two lemmas, $\hat{w}_i$ lies in neighborhood around a local minimum $w^\star$ such that the restriction of $\hat{F}$ to this neighborhood is strongly convex. As in the unconstrained setting we may assume w.l.o.g. that $\hat{w} = w^\star$.

Fix some $i \in [n]$. By assumption

$$\hat{F}(\hat{w}_i) - \hat{F}(\hat{w}) = \hat{L}(\hat{w}_i) - \hat{L}(\hat{w}) \geq \frac{\alpha}{2}\|\hat{w}_i - \hat{w}\|^2$$

On the other hand, since $\hat{w}_i$ minimizes the loss $w \in \mathcal{W} \mapsto \frac{1}{n}\sum_{j\neq i} f_j(w)$, the suboptimality of $\hat{w}_i$ w.r.t. the objective $\hat{F}$ is controlled by its suboptimality w.r.t. f_i, i.e.

$$\hat{F}(\hat{w}_i) - \hat{F}(\hat{w}) \leq \frac{1}{n}\Delta_i$$

Using Lipschitzness of f_i, we have

$$\Delta_i \leq \rho \|\hat{w}_i - \hat{w}\|$$

Combining the above, we obtain

$$\Delta_i^2 \leq \rho^2 \|\hat{w}_i - \hat{w}\|^2 \leq \frac{2\rho^2}{\alpha}(\hat{F}(\hat{w}_i) - \hat{F}(\hat{w})) \leq \frac{2\rho^2}{\alpha n}\Delta_i$$

Dividing by Δ_i (we can assume w.l.o.g. that $\Delta_i > 0$) we conclude the proof. ■

Proof (of Lemma 20) The first part is a direct application of Bernstein inequality (Tropp (2015)[Section 1.6.3]). It is left to prove that if A, B are positive semidefinite and $\|A - B\| \leq \epsilon$, then for all i, $|\lambda_i(A) - \lambda_i(B)| \leq \epsilon$. Indeed,

$$\begin{aligned}
\lambda_i(B) &= \max_{\dim(V)=i} \min_{v \in V} \frac{v^\top B v}{v^\top v} \\
&= \max_{\dim(V)=i} \min_{v \in V} \frac{v^\top A v + v^\top (B-A) v}{v^\top v} \\
&\leq \max_{\dim(V)=i} \min_{v \in V} \frac{v^\top A v}{v^\top v} + \max_{v \in V} \frac{v^\top (B-A) v}{v^\top v} \\
&= \lambda_i(A) + \epsilon \,.
\end{aligned}$$

Analogous proof shows that $\lambda_i(A) \leq \lambda_i(B) + \epsilon$. ■

Proof (of Theorem 2) Recall that the Lagrangian of $\hat{F}$ is denoted by $\hat{L}$. We first show that with high probability, points with large gradient do not form minima of $\hat{F}$. Similar argument shows that strict saddle points of L do not become minima of $\hat{F}$. Then, we can restrict ourselves to strongly convex regions of L and show that any w with $F(w) - \min_{w' \in \mathcal{W}} F(w') > \epsilon$ can not be a minimum of $\hat{F}$.

Fix some point $w \in \mathcal{W}$ with $\|\nabla L(w)\| \geq \tau$. Using matrix Bernstein inequality, we deduce that if $n = \Omega(\rho \log(d/\delta)/\tau^2))$, then $\|\nabla \hat{L}(w)\| \geq \tau/2$. Also, using Property A2, we have that for any $u \in \mathcal{W}$ with $\|u - w\| \leq r_1 := \min\{\frac{\tau}{4\beta_1}, 1\}$, $\|\nabla \hat{L}(u)\| \geq \tau/4$. Since $\mathcal{W}$ is bounded we can cover $\mathcal{W}$ using $(4B/r_1)^d$ balls of radius r_1 (for example, see the proof of Matoušek (2002)[Lemma 13.11.1]). By applying the union bound we deduce that if $n = \Omega(d\rho \log(dB/(r_1\delta)/\tau^2)$, then with probability at least $1 - \delta$, all points w with $\|\nabla L(w)\| \geq \tau$ satisfy $\|\nabla \hat{L}(w)\| \geq \tau/4$.

We next fix some point $w \in \mathcal{W}$ for which there exists a unit vector $v \in \mathcal{T}(w)$ with $v^\top(\nabla^2 L(w))v \leq -\gamma$. Using matrix Bernstein inequality, we deduce that if $n = \Omega(\beta_1 \log(d/\delta)/\gamma^2)$, then $v^\top \nabla^2 L(w) v \leq -\gamma/2$. Also, using Property A3, we have that for any $u \in \mathcal{W}$ with $\|u - w\| \leq r_2 := \min\{\frac{\gamma}{4\beta_2}, 1\}$, there exists $v \in \mathcal{T}(u)$ with $v^\top \nabla^2 L(u) v$. Since $\mathcal{W}$ is bounded, we can cover $\mathcal{W}$ using $(4B/r_2)^d$ balls of radius r_2. By applying the union bound, we obtain that a sample of size $n = \Omega(d\beta_1 \log(dB/(r_2\delta))/\gamma^2)$ ensures that with probability at least $1 - \delta$, γ-strict saddle points of F are $\gamma/2$-strict saddle of $\hat{F}$.

In particular, using Theorem 14 and Theorem 15 we deduce that strict saddle points of F and points with large gradient do not form local minima of $\hat{L}$.

Consider now vectors $w \in \mathcal{W}$ which belong to a strongly convex region around some minimum of F, denoted $w^\star$. Suppose that $F(w) - F(w^\star) > \epsilon$. By strong convexity, $\|\nabla L(w)\|^2 \geq 2\alpha\epsilon$.

Using concentration and covering as above, we conclude that for $n = \Omega(\beta_1 \log(dB/(r_1\delta))/(\alpha\epsilon))$, then with probability at least $1-\delta$, $\|\nabla \hat{L}(w)\|^2 \geq \alpha\epsilon$, hence w is not a local minimum of $\hat{F}$. ■