Gradients for the Loss!

Towards Provable Gradient Algorithms for Non-Convex Problems

Fotis Iliopoulos, Vrettos Moulous, Vaishaal Shankar, Max Simchowitz
EE227 BT
Department of EECSS

Contents

Introduction

Stochastic Gradient Descent (SGD) is perhaps one of the most practical and theoretically well motivated techniques for optimizing convex objectives for which noisy estimates of gradient can be computed efficiently. Surprisingly, variants of SGD appears to perform surprisingly well in highly non-convex problems, including but not limited to neural networks Ngiam et al. (2011), Bayesian models Hoffman et al. (2013), and Matrix Factorization Lee and Seung (2001). Motivated by the success of gradient based methods, recent literature has attempted to establish provable convergence rates for SGD and SGD-inspired local search algorithms on a variety of non-convex problems, most notably Sparse Coding Arora et al. (2015), PCA Shamir (2015b), Matrix Completion/Regression Chen and Wainwright (2015), Phase Retrieval Candes et al. (2015), and Tensor Decomposition Anandkumar et al. (2014).

While many of the above algorithms can converge to optimal or near-optimal solutions at a linear rate, they have fairly sensitive basins of convergence, and often rely upon involved spectral initializations to perform correctly. Furthermore, the proofs of convergence of these algorithms rely on seemingly ad-hoc computations of gradients, verifying roughly that if $f(w)$ is our objective, and $w^* \in \arg\min_w f(w)$, then as long as $\|w - w^*\|$ is sufficiently small, $\nabla f(w)$ is well correlated with the direction $w - w^*$. Thus, if the step size η is small enough, we can use standard techniques from convex optimization to show step-wise contraction (see, for example, the arguments in Arora et al. (2015)):

$$\|w - \eta \nabla f(w) - w^*\|^2 \leq \|w - w^*\|^2 \tag{1}$$

The aim of this project is will be to focus on recent advances in SGD-algorithms in problems for which the *global geometry* of the problem is understood, and for which global convergence can be established. Starting with PCA, moving to optimizing functions of semi-definite matrices, and concluding with the striking "Saddle Property" from Ge et al. (2015), this survey will attempt to highlight the common features of these settings which enables gradient-based local search to converge to global optima. We will devote most of our attention to the various proof techniques which have been used to establish this convergence, noting their similarities to contractivity arguments in classical proofs of convergence for convex problems. Finally, our project will complement the literature review with experiments comparing the performance of algorithms proposed in the literature on real and synthetic datasets.

Chapter 1

Convergence of Stochastic Gradient for PCA

1.1 SGD for Rank 1 Stochastic PCA

We consider a simple stochastic setting, where we have data $x_1, x_2, \ldots \in \mathbb{R}^d$ that are assumed to be drawn i.i.d. from an uknown underlying distribution. Our goal will be to find a direction of approximately maximal variance (or equivalently, finding an approximate leading eigenvector of the covariance matrix $\mathbb{E}[xx^\top]$), i.e., solving the optimization problem:

$$\max_{w:\|w\|_2=1} w^\top \mathbb{E}[xx^\top] w \ . \tag{1.1}$$

Note that the optimization problem is not convex, but it's "almost there", in the sense that if we knew the the underlying distribution then it is reduced to an eigendecomposition (also see subsections 1.3.6 and 1.3.5). As a matter of fact, the following simple method solves the problem: Approximate the covariance matrix by the empirical covariance matrix $\frac{1}{T}\sum_{i=1}^{T} x_i x_i^\top$, where T is a sufficiently large number of samples so that the sum of samples concentrates around its expectation. Then, compute its leading eigenvector by an eigen-decomposition. Using concentration of measure arguments, it's not hard to see that this method gives an $\mathcal{O}(\sqrt{\frac{1}{T}})$ approximate optimal solution in time $\mathcal{O}(Td^2 + d^3)$.

While polynomial time, this method is not suitable for large scale application where the size d might be huge. To cope with the problem in such cases, people often use Stochastic Gradient Descent (SGD). This corresponds to initializing a vector w_0, and then, at each iteration t, perform stochastic gradient step with respect to $x_t x_t^\top$, followed by a projection on the unit sphere.

$$w_t = (I + \eta_t x_t x_t^\top) w_{t-1}, \ \ w_t := w_t / \|w_t\| \ ,$$

where η_t is the step size parameter, to be suitably tuned (and which it may change over time).

As we have already explained, the goal of this project is to understand the features of certain non-convex problems which make them (both provably and practically) amenable

to gradient-based local search convergence to global optima. For the case of PCA, our understanding (based on studying literature such as Shamir (2015b) and Sa et al. (2015)) of why SGD works comes down the following high-level argument[1]:

Informally, one could say that Principal Component Analysis is a convex optimization problem in "expectation". That means is relatively easy to see, both intuitively and formally, that since our samples are i.i.d. coming from an (unknown) but nevertheless *fixed* distribution, applying an algorithm that follows the gradient in expectation outputs a vector w that is expected to maximize the objective function.

Having established the above, and in the context of "worst-case" analysis, one needs to establish some concentration of measure results as in why with reasonable probability the actual output of the algorithm is close to the expected one.

In the following subsections we briefly review the papers of Shamir (2015b) and Sa et al. (2015), we present experiments where we run variants of SGD for PCA in real data and comparisons between them (as well as with the theoretical results), and finally, in the last subsection, we give a more formal explanation of how PCA can be seen as a convex optimization problem, shedding further light on why we expect a gradient-based local search algorithm to work.

1.1.1 A brief overview of Shamir's result

In this subsection we briefly overview Shamir's result Shamir (2015b) on the convergence of SGD for PCA. As a matter of fact, in Shamir (2015b) a slightly more general setting is considered:

We are given access to a stream of i.i.d. of positive semidefinite matrices $\tilde{A}_t$, where $\mathbb{E}[\tilde{A}_t] = A$ is also a positive semidefinite matrix (e.g. $x_t x_t^\top$ in the PCA case) and we study the problem of solving

$$\min_{w \in \mathbb{R}^d : \|w\|_2 = 1} -w^\top A w \ . \tag{1.2}$$

using SGD. Notice that the gradient of Eq. (1.2) at a point w equals $2Aw$, with an unbiased stochastic estimate being $2\tilde{A}_t w$. Therefore, applying SGD to Eq. (1.2) reduces to the following:

Algorithm 1: Online SGD for PCA

Initialize: $w_0 \overset{unif}{\sim} \mathcal{S}^{d-1}$
for $t = 1, 2, \ldots, T$ **do**
  $w_t = (I + \eta \tilde{A}_t) w_{t-1}$
return $w = \frac{w_T}{\|w_T\|}$

Remark 1. It is not hard to verify that performing one last projection to the unit sphere at the end is mathematically equivalent to the original SGD algorithm. Furthermore, in

[1] We note though that given "eigengap" assupmtions (i.e., when the leading eigenvalue is significantly larger than the second one) one could mobilise different arguments for convergence. However, we focus on literature Shamir (2015b), Sa et al. (2015) where such an assumption is not needed (in the expense of worse running time bounds).

practice, we may normalize w_t at each step to ovoid numerical flow. Note that this does not alter the direction of w_t, and hence has no bearing on the convergence analysis.

Informally, the result Shamir (2015b) is that if we run the above algorithm for T steps doing a random initialization, then we get an $\tilde{\mathcal{O}}(\frac{\sqrt{d}}{T})$ - optimal solution with probability $\Omega(1/d)$. On the other hand, if we initialise by picking a unit norm vector by performing a single of approximate power iteration, then the dependence on d is replaced by a dependence of the numerical rank of the matrix, i.e., $n_A = \frac{\|A\|_F^2}{\|A\|}$, which is bounded by d in most application can be considered as a constant. The latter means that we initialise from $\frac{\tilde{A}w}{\|\tilde{A}w\|}$, where $\tilde{A} = \frac{1}{T_0}\sum_{t=1}^{T_0}\tilde{A}_t$ and w is sampled from a standard Gaussian distribution on $\mathbb{R}^d$. Formally, the following Theorem is proven:

Theorem 1.1.1. *Suppose that:*

- *For some leading eigenvector* v *of* A, $\frac{1}{(v^\top w_0)^2} \leq p$ *for some* p *(assumed to be* ≥ 7 *for simplicity)*
- *For some* $b \geq 1$, *both* $\frac{\|\tilde{A}_t\|_2}{\|A\|_2}$ *and* $\frac{\|\tilde{A}_t - A\|_2}{\|A\|_2}$ *are at most* b *with probability* 1.

If we run the algorithm above for T *iterations with* $\eta = \frac{1}{b\sqrt{pT}}$ *(assumed to be* ≤ 1*), then with probability at least* $\frac{1}{cp}$, *the returned* w *satisfies*

$$1 - \frac{w^\top A w}{\|A\|} \leq c' \frac{\log(T) b \sqrt{p}}{\sqrt{T}},$$

where c, c' *are positive numerical constants.*

Following are the two main corollary of Theorem 1.1.1 which bound the running time of SGD given the appropriate initialisation (and whose proof we compeletely omit since there outside the purpose of the project).

Corollary. *If* w_0 *is initialised chosen uniformly at random from the unit sphere, then Theorem 1.1.1 applies with* $p = \mathcal{O}(d)$, *and the returned* w *satisfies, with probability at least* $\Omega(1/d)$,

$$1 - \frac{w^\top A w}{\|A\|_2} \leq \mathcal{O}\left(\frac{\log(T) b \sqrt{d}}{\sqrt{T}}\right) .$$

Corollary. *If* w_0 *is initialised with a single approximate power iteration, then Theorem 1.1.1 applies with* $p = \mathcal{O}(\log(d) n_A)$, *and the returned* w *satisfies, with probability at least* $\Omega(1/n_A \log(d))$,

$$1 - \frac{w^\top A w}{\|A\|_2} \leq \mathcal{O}\left(\frac{\log(T) b \sqrt{\log(d) n_A}}{\sqrt{T}}\right) .$$

Proof Sketch of Theorem 1.1.1

We now give a very brief proof sketch of Theorem 1.1.1, presenting only the most important lemmata (without proof), while also giving a high-level explanation.

To simplify things, we will assume that we work in a coordinate system where A is diagonal, $A = \text{diag}(s_1, \ldots, s_d)$, where $s_1 \geq s_2 \ldots \geq s_d \geq 0$, and s_1 is the eigenvalue corresponding to v. This is without loss of generality, since the algorithm and the theorem conditions are invariant to the choice of the coordinate system. Moreover, throughout the proof it is assumed that $\|A\|_2 = s_1 = 1$, since the objective function in the theorem is invariant to $\|A\|_2$.

Let $\epsilon \in (0, 1)$ be a parameter to be determined later. Our goal will be to lower bound the probability of the objective function (which under the assumption $\|A\| = 1$, equals $1 - \mathrm{w}^\top A \mathrm{w}$) being suboptimal by at most ϵ. Letting $V_T = w_T^\top((1-\epsilon)I - A)w_T$, this can be written as $\Pr[V_T \leq 0]$.

The proof now is composed of a three parts. At first, it is proved that choosing ϵ and the step size η appropriately, then $\mathbb{E}[V_T] \leq -\tilde{\Omega}\left((1+\eta)^{2T}\frac{\epsilon}{p}\right)$. Thus, all it remains is to prove a concentration result, namely that V_T is not much larger than its expectation, since this would imply that $\Pr[V_T \leq 0]$ is indeed large. Unfortunately, it is not clear how one can prove such a concentration result so Shamir (2015b) follows a less straightforward approach. In particular, it turns out that it is possible to prove that V_T is not much *smaller* than its expected value: More precisely, that $V_T \geq -\tilde{\mathcal{O}}\left((1+\eta)^{2T}\epsilon\right)$ with high probability. Then, it is shown that given such a high-probability *lower bound* on V_T, and a bound on its expectation, we can produce an *upper bound* on V_T which holds with probability $\tilde{\Omega}(1/p)$, hence leading the result stated in the theorem.

For completeness, we state the basic lemmata that formally implement the above strategy as well as some comments on their proof.

Lemma 1.1.2 (Convergence in Expectation)**.** *If* $\eta = \frac{1}{b}\sqrt{\frac{1}{pT}}$ *and* $\epsilon = c\frac{\log(T)b\sqrt{p}}{\sqrt{T}} \leq 1$ *for some sufficiently large constant* c*, then it holds that*

$$\mathbb{E}[V_T] \leq -(1+\eta)^{2T}\frac{\epsilon}{4p} \ .$$

Intuitively, the proof of Lemma 1.1.2 is based on the fact that the problem is convex "in expectation", so following the gradient in expectation should work on average, and can be shown using standard linear algebra techniques (although it is rather painful).

Lemma 1.1.3 (High probability lower bound)**.** *Suppose that* $\tilde{\mathcal{A}}_t$ *is positive semidefinite for all* t*, and* $\Pr[\|\tilde{A}\| \leq b] = 1$*. Then, for any* $\delta \in (0, 1)$*, we have with probability at least* $1 - \delta$ *that*

$$V_T > -\exp\left(\eta b\sqrt{T\log(1/\delta)} + (b^2+3)T\eta^2\right)(1+\eta)^{2T}\epsilon \ .$$

The proof of Lemma 1.1.3 is based on a standard martingales argument that uses the Mc Diarmind inequality. Specifically, the author notices that:

$$V_T = w_T^\top((1-\epsilon)I - A)w_T \geq -\epsilon\|w_T\|^2 \ ,$$

and thus that is sufficient to prove that

$$\|w_T\|^2 < \exp(\eta b\sqrt{T\log(1/\delta}) + (b^2+3)T\eta^2)(1+\eta)^{2T} \quad . \tag{1.3}$$

To do so, he uses a telescoping argument, to show that:

$$\log(\|w_T\|^2) \leq \sum_{t=0}^{T-1}\left(\frac{\|I+\eta\tilde{A}_t w_t\|^2}{\|w_t\|^2} - 1\right) \quad .$$

Now, he considers the Doob's Martingale for that sum, and using Mc Diarmind's inequality he establishes (1.3).

Lemma 1.1.4 (Upper Bound from expectation and lower bound)**.** *Let X be a non-negative random variable such that for some α, $\beta \in [0,1]$, we have $\mathbb{E}[X] \geq \alpha$, and for any $\delta \in (0,1]$,*

$$\Pr\left[X \geq \exp(\beta\sqrt{\log(1/\delta)}\right] \leq \delta \quad .$$

Then,

$$\Pr[X > \frac{\alpha}{2}] \geq \frac{\alpha - \exp(-\frac{2}{\beta^2})}{14}.$$

We note that Lemma 1.1.4 is the main technical reason that we get convergence with low probability (something that is believed to be an artifact of the proof).

1.1.2 Rank 1 Stochastic PCA Experiments

1.2 The Alecton Algorithm for Higher Rank PCA

To recover the top r singular vectors of a PSD matrix M from an algorithm which recovers its top eigenvector, we can use a deflation algorithm (see **?**); that is, if $\tilde{v}$ is an approximation to the top eigenvector vector v of M, we then compute the top singular vector of $\tilde{M} := (I - \tilde{v}^T)M$. Using offline algorithms like the Power method or Lanczos methodsLanczos (1950), we can compute v to extremely high precision, and so deflation works quite well. Unfortunately, the algorithm in Shamir (2015b) converges at a sublinear rate of $O(1/\sqrt{T})$, due in part to the stochastic setting of the problem. Consequently, deflation techniques to estimate the top r singular vectors of a PSD will perform rather poorly.

Thus, it makes sense instead to consider an algorithm to compute the top r eigenvectors simultaneously, so as not to suffer any instability from deflation techniques. To this end, we introduce the Alection algorithm from Sa et al. (2015), which uses a gradient descent style algorithm to compute the top r singular vectors. Moreover, algorithm converges with constant probability from a random initialization.

We are given noisy observations $\tilde{A} \succeq 0$ of a true matrix $A := \mathbb{E}[\tilde{A}] \succ 0$, and our goal is to find a rank $p \leq n$ approximation $\hat{A}$ to A. Encoding the rank constraints explicitly, we want the solution to the following stochastic problem

$$\min_{Y \in \mathbb{R}^{n\times}} \mathbb{E}[\|\tilde{A} - YY^T\|_F] \tag{1.4}$$

At each time step k, we let $\tilde{A}_k$ be our noisy sample of A, and run alecton, we run essentially the matrix analogue of the algorithm in Shamir (2015b); I.e.

$$Y_k = (I + \eta \tilde{A}_k)Y_k \tag{1.5}$$

In sum, our algorithm is as follows:

Algorithm 2: Online SGD for Rank-K Eigenspace Estimation (Alecton)

Initialize: $(Y_0)_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$
for $t = 1, 2, \ldots, T$ **do**
  $Y_t = (I + \eta \tilde{A}_t)Y_{t-1}$
return $\hat{Y} \leftarrow Y_T(Y_T^T Y_T)^{-1/2}$

1.2.1 Algorithm Intuition

Another way to derive this algorithm comes from stochastic gradient descent on a Riemannian manifold. THe key intuition is that the map $\psi(Y) = YY^T$ has unitary symmetry. More precisely, let $\mathcal{O}_p = \{U \in \mathbb{R}^{p\times p} : U^T U = I\}$ denote the the p-dimensional orthogonal group. Then, $\psi(YU) = \psi(Y)$ for all $U \in \mathcal{O}_p$. Hence, instead of optimizing over Y, we can optimize over an equivalent space where all points $Y_1, Y_2 \in \mathbb{R}^{n\times p}$ for which $Y_1 = UY_2$ for $U \in \mathcal{O}_p$ are regarded as the same point. This indentification is know as the quotient manifold $\mathcal{M} := \mathbb{R}^{n\times p}/O_p$.

In the flat euclidean space $\mathbb{R}^{n\times p}$, we have the standard inner product $\langle U, V\rangle = \text{trace}(U^T V) = \sum_{i,j} U_{ij}V_{ij}$, which is indentical at everypoint. When we quotient out by $\mathcal{O}_p$, we introduce curvature into the manifold. This corresponds to a Riemmanian metric which varies at different points in the manifold. For $\mathcal{M}$, the induced inner product is

$$\langle U, V\rangle_Y := \text{trace}(UYY^T V^T) = \text{trace}(U(VYY^T)) := \sum_{i,j} U_{ij} G_{(i,j),(i,j)}(Y) V_{ij} \tag{1.6}$$

where $G_{(i,j),(i,j)}(Y)$ corresponds to the linear map $\mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ induced by right multiplication by $Y^T Y$. $G_{(i,j),(i,j)}(Y)$ can be thought of as the Riemanian metric.

What does this mean for gradient descent? Let's start off in Euclidean Space: Let $\tilde{f}(Y) := \|YY^T - \tilde{A}\|_F^2$. we can write

$$\begin{aligned} \mathbb{E}[\tilde{f}(Y_k)] &= \text{trace}(Y_k Y_k^T Y_k Y_k^T) - 2\,\text{trace}(Y_k \tilde{A} Y_k^T) + \|\tilde{A}\|_F^2 & (1.7)\\ \nabla \tilde{f}(Y) &= 4(Y_k Y_k^T Y_k - A_k Y_k) & (1.8) \end{aligned}$$

Standard gradient descent with step size α_k thus corresponds to the updates

$$Y_k - \alpha_k \nabla \tilde{f}(Y_k) = Y_k - 4\alpha_k(Y_k Y_k^T Y_k - A_k Y_k) \tag{1.9}$$

On a general manifold, we reweight by a by the Riemmanian metric and take updates of the form

$$Y_k - \alpha_k G_{Y_k}^{-1} \nabla \tilde{f}(Y_k) \tag{1.10}$$

in this case, G_{Y_k} is the map which right multiplies by $Y_k^T Y_k$, and so $G_{Y_k}^{-1}$ is the linear map which multiplies by $(Y_k^T Y_k)^{-1}$. To avoid running into singularies, we introduce a step size parameter $\eta = \alpha_k/4$, and consider a "flattened out" metric which corresponds to multiplication by $(I + \eta Y_k^T Y_k)$. This yields updates

$$\begin{aligned} Y_{k+1} &= Y_K - \eta(Y_k Y_k^T Y_K - \tilde{A}_k)(I + \eta Y_k^T Y_k)^{-1} && (1.11) \\ &= (I + \eta \tilde{A}_k) Y_k (I + \eta Y_k^T Y_k)^{-1} && (1.12) \end{aligned}$$

In practice, we would never want to invert $(I + \eta Y_k^T Y_k)^{-1}$ at every iteration. However, if we only want to estimate the span of the top p-eignvectors of A, then we can get rid of the multiplication by $(I + \eta Y_k^T Y_k)^{-1}$, since it doesn't change the columnspace of Y_{k+1}. Hence, we try to learn the top columnspace of A by

$$Y_{k+1} = (I + \eta \tilde{A}_k) Y_k \tag{1.13}$$

Alternately, we can think of the updates as "turning" the updates of Y_k in the direction of the dominant eigenspace of A.

1.2.2 Precise Setup

Let $A = \mathbb{E}[\tilde{A}]$. We denote the top eigenvectors of A by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, and singular vectors corresponding to the eigenvalues $u_1, \ldots, u_n$. Our goal is to recover a Y whose column span is as close to $u_1, \ldots, u_p$ as possible. In practice, if there is a very small eigengap, or if eigenvalues occur with multiplicity, we will need to recover a Y whose colmmn span is close to $u_1 \ldots, u_q$ for some $q \geq p$. We let $\Delta := \lambda_q - \lambda_{q-1}$ denote the eigengap. Throughout, we fix an $\epsilon > 0$. We will also make the following assumptions

1. Bounded Variance. There are constants $\sigma_a, \sigma_r > 0$ such that

$$\mathbb{E}[y^T \tilde{A} W \tilde{A}^T y)] \leq \sigma_a^2 \tilde{W} \|y\|^2 \tag{1.14}$$

   For any $W \succ 0$ be a matrix that commutes with A. (To estimate the radial component, we need $\mathbb{E}[(y^T \tilde{A} y)^2] \leq \sigma_r^2 \|y\|^4$)

2. Rank: Either $p = 1$, or $\text{rank}\tilde{A} = 1$. This indispensible to the proof.

3. Step Size. Rescale the step size to be independent of problem parameters by letting $\gamma = \frac{2n\sigma_a^2 p^2 (p+\epsilon}{\Delta \epsilon} \eta$. We require that η is chosen so that $\gamma \leq 1$.

To measure the success of the algorithm, we need a notion of "successful eignspace estimation", which we give as follows:

Definition 1.2.1. For a fixed $\epsilon > 0$, we say that *sucess* occurs at time k if, for all $z \in \mathbb{R}^p$,

$$\frac{\|UY^k z\|}{\|Y^k\|} \geq (1 - \epsilon) \tag{1.15}$$

We say success has occured by time t if success has occured for some timestep $k \leq t$.

We actually found numerous flaws in the analysis in Sa et al. (2015), and were actually able to improve their resuls using eleemntary techniques. We present these results as follows:

Theorem 1.2.1. *[Alecton High Probability]*

$$\Pr(F_t) \leq Z_p(\gamma) + \frac{np^2}{\gamma q \epsilon} \exp(-\frac{\Delta^2 \gamma \epsilon t}{4n^2 p^2 (p+\epsilon)}) \tag{1.16}$$

where $Z_p(\gamma) = 2(1 - \mathbb{E}[\det(I + \gamma p^{-1}(R^T R)^{-1})^{-1}]$, *where* $R \in \mathbb{R}^{p \times p}$, *and* $R_{i,j} \overset{iid}{\sim} \mathcal{N}(0,1)$. *In particular, given* $\delta > 0$

$$\Pr(F_t) \leq Z_p(\gamma) + \delta \tag{1.17}$$

as long as $t \geq \frac{4n^2 p^2 (p+\epsilon}{\Delta^2 \gamma \epsilon} \log(np^2 \gamma q \epsilon / \delta)$.

1.3 Sketching the Proof of Theorem 1.2.1

1.3.1 Martingales

One helpful technique is the use of Martingales and Stopping Times. The idea is that, with non-zero possibly, our algorithm can fail, or reach a fixed point, or make painfully small progress. Using martingales, we will be able to control this probability explicitly. First, some preliminaries are in order:

Definition 1.3.1 (Martingales, Sub- and Supermartingales, and Stopping Times)**.** Given a filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_k \subset \ldots$:

1. We say X_k is an adapted process if X_k is $\mathcal{F}_k$ measurable
2. We say that X_k is a martingale if $\mathbb{E}[X_{k+1}|\mathcal{F}_k] = X_k$.
3. We say X_k is a submartingale if $\mathbb{E}[X_{k+1}|\mathcal{F}_k] \geq X_k$, and X_k is a supermartingale if $\mathbb{E}[X_{k+1}|\mathcal{F}_k] \leq X_k$.

We say that T is a stopping time if the event $\{T \leq k\} \in \mathcal{F}_k$

Theorem 1.3.1 (Optional Stopping Theorem)**.** *If* X_k *is a martingale (or sub or supermartingale) for a filtration* $\mathcal{F}_k$, *and* T *as stopping time with respect to the filtration* $\mathcal{F}_k$, *then* $x_{\min(k,T)}$ *is a martingale (resp, sub- resp. supermartingale) with respect to* $\mathcal{F}_k$. *In particular, if* X_k *is bounded almost surely, or if* T *is bounded almost surely, then this implies* $\mathbb{E}[x_0] \leq \mathbb{E}[x_T]$ *for bounded submartingales and* $\mathbb{E}[x_0] \geq \mathbb{E}[x_T]$ *for bounded super martingales.*

1.3.2 Preliminaries

For the purposes of the analysis, it will be useful to work with an easier to work with proxy for the ratio in the above display. This motivates the definition of the sequence τ_k, defined as follows

1. Let $U \in \mathbb{R}^{n \times p}$ be projection matrix onto q singular vectors.

2. Let $W = I\frac{q\gamma}{np^2} + (1 - \frac{q\gamma}{np^2})U$. Since $\gamma \leq 1$, $q\gamma/np^2 \leq q/np^2 \ll 1$ in most cases. Hence, $W = \frac{q\gamma}{np^2}(I - U) + U$, so W is the sum of the projection U, plus a small rescaled projection on $U^{\perp}$.

3. Next, we consider $\tau_k = \frac{\det(Y_k^T U Y_k)}{\det(Y_k^T W Y_k)} = \frac{\det(Y_k^T U Y_k)}{\det(Y_k^T U Y_k + \frac{q\gamma}{np^2} Y_k^T (I-U) Y_k}$. Here τ represents the ratio of how much Y_k lies in the subspace spanned by U, to how much Y_k crosses over into $U^{\perp}$.

Lemma 1.3.2 makes precise how τ_k helps measures convergence of the ration $\frac{\|UY^k z\|}{\|Y^k\|}$ to 1:

Lemma 1.3.2 (τ_k measure success)**.** *If success does not occur at time step k, then $\tau_k \leq 1 - \frac{\gamma q \epsilon}{np^2}$.*

For the analysis, we need the notion of a catastrophic failure; to avoid drama, we'll call this "deep failure".

Definition 1.3.2 (Deep Failure)**.** We define the deep failure event f_k as the event when $t_k \leq 1/2$. We define total failure at time t as the event that either deep failure has occured, that, or success has not yet occured. We define the stopping time T as the first time at which success or deep failure occur.

The idea is that if deep failure occurs, Y lies far from the dominant eigenspace of Y, and it will take a long time for Y to return to a reasonable estimate of U.

Crucially, the following lemma shows that, as long as success or failure have not occured at time k (that is $k < T$), τ_k increases in expectation. Rearranging the result show in fact that, as long as success or failure have not occured, τ_k decreases geometrically:

Lemma 1.3.3. *[Sufficient increase] For any time k at which success has not yet*

$$\mathbb{E}[\tau_{k+1}|\mathcal{F}_k] \geq \tau_k(1 + \eta\Delta(1 - \tau_k)) \tag{1.18}$$

In particular, until success has occured, τ_{k+1} is a subartingale. Written more suggestively,

$$\begin{aligned} \mathbb{E}[1 - \tau_k|\mathcal{F}_j] &\leq (1 - \tau_k)(1 - \eta\Delta\tau_k) \\ &\leq (1 - \tau_k)(1 - \eta\Delta\frac{1}{2}) \end{aligned} \tag{1.19}$$

where the last inequality holds as long as failure has not yet occured.

Proof. Demonstration of the recurrence in Equation 1.18 is technical and ommited. The first inequality in equation 1.19 follows from rearranging Equation 1.18, and the second follows since if deep failure doesnt occur at time k. $\tau_k \geq 1/2$. □

1.3.3 Main Arguments and Completing the proof

As a consequence, we have the following control on the failure. This result is extremely generic, and relies upon the principal strategy in this paper: Build a martingale Z_t, find a good stopping time T, apply the optional stopping theorem to show that Z_T is a sub- or super martingale, and apply markov:

Lemma 1.3.4. *The probability that extreme failure occurs before the success event is*

$$\Pr(f_T) \leq 2(1 - \mathbb{E}[t_0]) := Z_p(\gamma) \tag{1.20}$$

Proof. Let T be defined as in the lemma. For $k < T$, Lemma 1.3.3 holds, so that τ_k is a submartingale. Moreover, $\tau_k \in [1/2, 1]$ since we have not deeply failed. Hence, τ_k is bounded almost surely, so by the optional stopping theorem, we have

$$\begin{aligned} \mathbb{E}[\tau_0] &\leq \mathbb{E}[\tau_t] \\ &\leq \mathbb{E}[\tau_T | f_T]\Pr(f_t) + \mathbb{E}[\tau_t | \neg f_t]\Pr(\neg f_t) \end{aligned}$$

If, under deep failure at time f_T, $\tau_T \leq 1/2$. On the other hand, $\tau \leq 1$ almost surely. Thus,

$$\mathbb{E}[\tau_0] \leq \frac{1}{2}\Pr(f_t) + 1(1 - \Pr(f_t))$$

and rearranging implies that $\Pr(f_t) \leq 2(1 - \mathbb{E}[\tau_0])$ □

Remark. Note that the preceeding theorem needed only the fact that, until T, τ_k is a supermartingale in $[1/2, 1]$ almost surely. So, we can think of the result as saying that the chance a bounded super martinagle hits its lower bound is bounded above by how close to the upper bound the martingale starts at.

The upshot of this argument is to "push" the probability of failure from any possible failing during the run time of the main algorithm, into a failure event depending on the initialization.

Concluding the Proof of Theorem 1.2.1

We now prove Theorem 1.2.1. We remark that the proof we present is original, simpler, and attains a better control on the probability of failure than the main theorem of Sa et al. (2015). We have already contacted the authors of the aforemention work, and presented them with this alternate argument.

Again, let T be the stopping time at which either deep failure or success occur first. First, we have

$$\Pr(\mathbb{1}(k < T)(1 - \tau_k) < \frac{\gamma q \epsilon}{np^2}) = \Pr(k \geq T) \tag{1.21}$$

since, the only way that $\mathbb{1}(k < T)(1 - \tau_k) \leq \frac{\gamma q \epsilon}{np^2}$ is if either

1. $\mathbb{1}(k < t) = 0$, so that $k \geq T$
2. or if $1 - \tau_k < \frac{\gamma q \epsilon}{np^2}$. By Lemma 1.3.2, this occurs only if success occurs at time k, and thus $k \geq T$ as well.

Thus, by Markov's Inequality,

$$\begin{aligned} \Pr(k \geq T) = &\Pr(\mathbb{1}(k < T)(1 - \tau_k) < \frac{\gamma q \epsilon}{np^2}) \\ \geq &1 - \frac{np^2 \mathbb{E}[\mathbb{1}(k < T)(1 - \tau_k)]}{\gamma q \epsilon} \end{aligned} \tag{1.22}$$

We now upper bound the expectation in the above display. Since $1-\tau_k$ is nonnegative, we can use the tower property to write

$$\begin{aligned}
\mathbb{E}[\mathbb{1}(k<T)(1-\tau_k)] &\leq \mathbb{E}[\mathbb{1}(k-1<T)(1-\tau_k)] && (1.23)\\
&= \mathbb{E}[\mathbb{E}[\mathbb{1}(k-1<T)(1-\tau_k)\big|\mathcal{F}_{k-1}]] && (1.24)\\
&= \mathbb{E}[\mathbb{1}(k-1<T)\mathbb{E}[(1-\tau_k)\big|\mathcal{F}_{k-1}]] && (1.25)
\end{aligned}$$

where the last step follows since $\mathbb{1}(k-1<T)$ is $\mathcal{F}_{k-1}$ measurable; that is, it does not dependent on information that has not yet been determined by time $k-1$. When $k-1<T$, $k-1$ is neither a success nor a deep failure step. Thus, by the recusion in Lemma 1.3.3

$$\mathbb{1}(k-1<T)\mathbb{E}[(1-\tau_k)\big|\mathcal{F}_{k-1}] \leq (1-\frac{1}{2}\eta\Delta)(1-\tau_{k-1}) \tag{1.26}$$

Applying this argument recursively, we have

$$\begin{aligned}
\mathbb{E}[\mathbb{1}(k<T)(1-\tau_k)] &\leq (1-\frac{1}{2}\eta\Delta)\mathbb{E}[(1-\tau_{k-1})] && (1.27)\\
&\leq (1-\frac{1}{2}\eta\Delta)\mathbb{E}[\mathbb{1}(k-1<T)(1-\tau_{k-1})] && (1.28)\\
&\leq (1-\frac{1}{2}\eta\Delta)^k\mathbb{E}[\mathbb{1}(0<T)(1-\tau_0)] && (1.29)\\
&\leq (1-\frac{1}{2}\eta\Delta)^k && (1.30)
\end{aligned}$$

where the last step follows since $(1-\tau_0)\leq 1$. Putting it all together,

$$\Pr(k\geq T) \geq 1-\frac{np^2}{\gamma q\epsilon}(1-\frac{1}{2}\eta\Delta)^k \geq 1-\frac{np^2e^{-k\eta\Delta/2}}{\gamma q\epsilon} \tag{1.31}$$

Let F_k denote the event that success has not occured by time k, and let f_k denote the event that failure has occured at time k. If $k\geq T$, either success or failure has occured by time k. So if $k\geq T$ and if success hasn't occured by time k, then deep failure must occur at T (T is the first time either success or deep failure occur). Thus, $\Pr(k\geq T)\leq \Pr(f_T)+\Pr(\neg F_k)$. Rearraging,

$$\Pr(\neg F_k) \geq \Pr(k\geq T)-\Pr(f_T) \geq 1-\frac{np^2e^{-k\eta\Delta/2}}{\gamma q\epsilon}-\Pr(f_T) \tag{1.32}$$

Hence

$$\Pr(F_k) \leq \Pr(f_T)+\frac{np^2e^{-k\eta\Delta/2}}{\gamma q\epsilon} \tag{1.33}$$

Theorem 1.2.1 now follows by writing η in terms of γ, and using Lemma 1.3.4 to bound $\Pr(f_T)$.

1.3.4 Experiments for Alecton

1.3.5 Toy Convexity in Two Dimensions

To build our intuition, consider the problem of PCA in two dimensions, $\min -x^T Ax$. where $A \succ 0$. By changing our coordinate system, we may assume that $A = \text{diag}(\alpha, \beta)$ is diagonal with $\alpha \geq \beta \geq 0$. Since $x \in \S^1$, we may parameterize $x(\theta) = (\cos\theta, \sin\theta)$. Hence, as a function of θ, our objective reads

$$\min_\theta f(\theta) = -\alpha \cos^2\theta - \beta \sin^2\theta \tag{1.34}$$

It turns out that this parameterization of x is "natural" for the problem, in the sence that $x = x(\theta)$ traces out an *geodesic arc of unit length*, that is, the path $x(\theta), \theta \in [\theta_1, \theta_2]$ is the shortest path from θ_1, θ_2, see Absil et al. (2009). This motivates a generalized definition of convexity, known as "geodesic convexity" (again, see Absil et al. (2009)). More generally, given a Riemannian manifold $\mathcal{M}$, we see $f : \mathcal{M} \to \mathbb{R}$ is a "geodesically convex" if, given a geodesic curve $\gamma(t) : \mathbb{R} \to \mathcal{M}$, the composition $f(\gamma(t))$ is convex. In Euclidean space, the shortest between two points is precisely the line segment between them, so geodesics are just straight lines. Thus, a function in Euclidean space is convex iff it is convex when restricted to any line segment; which is precisely the classical definition of convexity.

We now show study the convexity of Equation 1.34 in this angular parameterization. Let $e_1 = (1, 0)$ and $e_2 = (0, 1)$, so that $\pm e_1$ are the top eigenvectors of A. We show that, for all x which a 45-degree angle with e_1, the objective is convex. Indeed, we have

$$f'(\theta) = (2\alpha \sin\theta\cos\theta - 2\beta\alpha \sin\theta\cos\theta) = 2(\alpha - \beta)\sin 2\theta \tag{1.35}$$

and Hessian $f''(\theta)$ is given by

$$\begin{aligned} f''(\theta) &= \frac{d}{d\theta}(\alpha - \beta)\sin 2\theta & (1.36)\\ &= 2(\alpha - \beta)\cos 2\theta & (1.37) \end{aligned}$$

In particular, for $\theta \in [-\pi/4, \pi/4] \cup [3\pi/4, 5\pi/4]$, $f''(\theta) \geq 0$, so that $f(\theta)$ is convex.

But what can we say above global convexity? From Equation 1.35, we see that for any $\theta \in (-\pi/2, \pi/2)$, $f'(\theta)$ points in the direction of $\theta = 0$, that is $x = e_1$, whereas for $\theta \in (\pi/2, 3\pi/2)$, $f'(\theta)$ points in the direction of $\theta = -\pi$, that is, $x = -e_1$. Hence, up to a scaling factor $\eta(\theta)$, we have that f obeys rescaled subgradient-like inequalities of the form

$$f(\theta) - f(0) \leq \eta(\theta) f'(\theta)\theta \quad \forall \theta \in (-\pi/2, \pi/2) \tag{1.38}$$

and

$$f(\theta) - f(\pi) \leq \eta(\theta) f'(\theta)(\theta - \pi) \quad \forall \theta \in (-\pi/2, \pi/2) \tag{1.39}$$

In other words, up to a varying scaling $\eta(\eta)$, the derivatives $f'(\theta)$ of f behave as if f is convex, restricted to the left and right halfs of the circle $(-\pi/2, \pi/2)$ and $(\pi/2, 3\pi/2)$. We remark that, we can take this scaling to be $\eta(\theta) = O(\frac{1}{\alpha - \beta} \cdot \csc\theta)$, which behaves like

$$\frac{1}{(\alpha - \beta)\min_{j \text{ odd}} |j\pi/2 - 2\theta|} \tag{1.40}$$

when θ is near $\pm\pi/2$, that is $x(\theta) \approx \pm e_2$. In other words, the amount of "help" the step size η needs to provide grows as one over the distance to the fixed points $x_2 = \pm e_2$.

1.3.6 Convexity Results in Higher Dimensions

In Shamir (2015a) the author wonders how non-convex is really the optimization problem for PCA we are tackling. As we have already seen, it is more than clear that it has a nice structure (for one, it can be solved in polynomial time) but perhaps we can actually prove that the problem is convex in some domain (at least near optimal points). To discuss this question, the author considers the negative Rayleigh quotient function:

$$F_A(w) = -\frac{w^\top A w}{\|w\|^2} = \frac{1}{n}\sum_{i=1}^{n}\left(-\frac{\left(w^\top x_i\right)^2}{\|w\|^2}\right)$$

, where $A = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$.

Unfortunately, the following Theorem shows that the function is *not* convex almost everywhere:

Theorem 1. Consider the matrix $A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. The Hessian of F_A is not positive semidefinite for all but a measure-zero set.

Proof. The determinant of the Hessian of F_A equals $-\frac{4w_1^2 w_2^2}{(w_1^2+w_2^2)^4}$, which is always non-positive, and strictly negative for w for which $w_1 w_2 \neq 0$ (which holds for all but a measure-zero set of $\mathbb{R}^d$). Since the determinant of a positive semidefinite matrix is always non-negative, this implies that the Hessian isn't positive semidefinite for any such w. □

While the Theorem indeed implies that we are cannot use convex optimization tools as-is on the function F_A, even if we're close to an optimum, it does not preclude the possibility of having convexity on a more constrained, lower-dimensional set. As a matter of fact, the author shows the following: Given some point w_0 close enough to an optimum, we can explicitly construct a simple convex set, such that:

1. The set includes an optimal point of F_A
2. The function F_A is $\mathcal{O}(1)$-smooth and λ-strong convex in that set, where λ is the (eigen-)gap between the first and second eigenvalue.

Before we continue, we mention that while this potentially suggests a two-stage approach for solving streaming PCA, i.e., using some existing algorithm to find w_0 and then switch to a convex optimization algorithm, the fact that we will need w_0 to be *very* close to the optimum, namely $\|v_1 - w_0\| \leq \mathcal{O}(\lambda)$, does not leave much space for improvement. In any case though, this construction is at least theoretically interesting since it sheds some further light on the nice structure of streaming PCA.

We now describe the construction and the related theorems (without proof). Given a unit vector w_0, let:

$$\begin{aligned} H_{w_0} &= \left\{w^\top w_0 = 1\right\} \\ B_{w_0} &= \{w : \|w - w_0\| \leq r\} \end{aligned}$$

be the hyperplane tangent to w_0 and the Euclidean ball of radius r centered at w_0, respectively. The convex set we use, given such a w_0, is simply the intersection of the two, $H_{w_0} \cap B_{w_0}(r)$, where r is a sufficiently small number. Now the following theorem establishes that if w_0 is $\mathcal{O}(\lambda)$-close to an optimal point and we choose the radius of $B_{w_0}(r)$ appropriately, then $H_{w_0} \cap B_{w_0}(r)$ contains an optimal point, and the function F_A is indeed λ-strongly convex and smooth on that set.

Theorem 2. For any positive semidefinite A with spectral norm 1 (for simplicity), eigengap λ and a leading eigenvector v_1, and any unit vector w_0 such that $\|w_0 - v_1\| \leq \lambda/44$, the function $F_A(w)$ is 20-smooth and λ-strongly convex on the convex set $H_{w_0} \cap B_{w_0}(\frac{\lambda}{22})$, which contains a global optimum of F_A.

Finally, in Shamir (2015a) it's also problem that the convexity property is lost if w_0 is significantly further away from v_1.

Theorem 3. For any $\lambda, \epsilon \in (0, \frac{1}{2})$, there exists a positive semidefinite matrix A with spectral norm 1, eigengap λ, and leading eigenvector v_1, as well as a unit vector w_0 for which $\|v_1 - w_0\| \leq \sqrt{2(1+\epsilon)\lambda}$, such that F_A is not convex in any neighborhood of w_0 on H_{w_0}.

1.3.7 Experimental Results

For our experiment we compared the two streaming PCA algorithms (Shamir's and Alecton) vs Power Iteration. We ran our experiments on the MNIST digit dataset and the Labeled Faces dataset.

Power Iteration (green) vs SGD (blue)

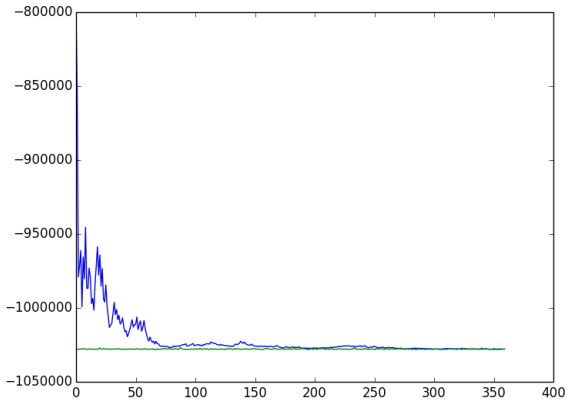


Figure 1.1: This is the convergence of SGD for the leading eigenvector for the MNIST dataset

Face

Figure 1.2: We attempt to generate a low rank approximation of this face

Figure 1.3: As expected Shamir's deflation based rank K approximation (left) does much worse than power iteration (right)

Figure 1.4: Alecton algorithm does a lot better than Shamir's

Chapter 2

Dropping Convexity for Faster Semi-definite Optimization

2.1 Introduction: Back to Rank r-PCA

In the general principal component analysis, one approximates a possibly full rank matrix $M \in \mathbb{R}^{m\times n}$ with singular value decomposition $M = U\Sigma V^T$ with the rank r-approximation $M_r \in \mathbb{R}^{m\times n}$ given by setting all but the r-largest singular values in Σ to 0. Unless otherwise specified, we will restrict out attention to the PSD-PCA, that is, the setting where M is square, symmetric, and positive semidefinite; extensions to the rectangular setting will be disccused briefly at the end of this section. For example, M might be a data covariance matrix

$$M = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T \tag{2.1}$$

and PCA might attempt to extract the r directions along while the data $\{x_i\}$ exhibit the most variance. Numerous works have been written which study the statistical advantages of performing rank PCA, particularly are drawn from a mostly low-rank distribution, plus isotropic noise.

Consistent with the theme of this course, we will instead investigate PCA as an optimization problem; one which is decidely non-convex, but both shares many of the same appealing regularity conditions as convexity, and admits efficient first order algorithms in the same spirit as gradient descent. Formally, PSD-PCA can be expressed as

$$\min_X \|X - M\|_F^2 : \quad \text{rank}(X) \le r \quad \tilde{M} \succ 0 \tag{2.2}$$

While the objective function $\|X - M\|_F^2$ is convex, the rank constraints on $\tilde{M}$ are certainly nonconvex. For example, the matrices $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ are both rank one, but their average is $\frac{1}{2}I_2$, which is rank 2.

One can, however, perform a variable substitution which renders the constraint set convex, but at the cost of sacrificing the convexity of the objective. We note that, if

$X \in \mathbb{R}^{n\times n}$ is a rank at most k and PSD, then there exists a matrix $U \in \mathbb{R}^{n\times r}$ for which $X = UU^T$. On the other, if $U \in \mathbb{R}^{n\times k}$, then UU^T is PSD and has rank at most r. With this in mind, we may rewrite the problem in Equation 2.2 as an unconstrained problem:

$$\min_{U\in\mathbb{R}^{n\times r}} \|M - UU^T\|_F^2 \tag{2.3}$$

While our measure of distance is the Frobenius norm, the objective is decidely nonconvex in U. One way to see this is to consider the solution set. Indeed, if $O \in \mathbb{R}^{r\times r}$ is in the orthogonal group $\mathcal{O}(r) := \{O \in \mathbb{R}^{r\times r} : O^TO = I\}$. satisfies $O^TO = I$, then $(UO)(UO)^T = UOO^TU = UU^T = X$. Thus, if U^* is optimal for Equation 2.3, then so are all $\{U^*O : O \in \mathcal{O}(r)\}$. If the objective were convex, then for any $O_1, O_2 \in \mathcal{O}(r)$, then $\alpha O_1U^* + (1-\alpha)O_2U^* = (\alpha O_1 + (1-\alpha)O_2)U^*$ would be optimal as well. Taking $\alpha = 1/2$, $O_1 = I$ and $O_2 = -I$, this would imply that $U^* = 0$ would be an optimal rank r approximation of any PSD M, which is patently absurd.

2.2 From PCA to General Rank-Constrained Optimization

More generally, we may attempt a generalization of equation 2.3, where we replace that objective $\|M - X\|_F^2$ with an arbitrary convex function $f : \mathbb{R}^{n\times n} \to \mathbb{R}$. Rather than enforcing the rank constraints on X, we may instead consider the non-convex program

$$\min_{U\in\mathbb{R}^{n\times r}} f\left(UU^\top\right), \text{ where } r \le n \tag{2.4}$$

Suppose that $X^* := \arg\min_{X\succ 0} f(X)$, and let X_r^* be the rank r-SVD of X. Surprisingly, Bhojanapalli et al. (2015) demonstrate that, under suitable convexity and smoothness conditions on f, a simple gradient descent algorithm recovers a rank r solution X whose objective value $f(X)$ is no more than than $f(X_r^*)$. In particular, if the optimal $X^* = X_r^*$ is exactly rank, and f is strongly convex, then gradient descent recovers X^* exactly.

The advantages are twofold: first, by encoding the low rank constrain explicity, we can expect substantially improved run times, since we replace n^2 parameters in a matrix $X \succ 0$ with only nr paremeters in the matrix U. This is particularly crucial in large scale data problems, where the data dimension n greatly exceeds the low rank structure r. Secondly, when r is comparable to n, the non-convex formulation has the advantage of avoiding costly projection steps to low rank matrices, which rely on SVD computations.

But even for the case where we pick r to be comparable to n it makes sense to consider the non-convex formulation (2.4), because it automatically encodes the PSD constraint, and in Bhojanapalli et al. (2015) is being proved that a simple gradient descent converges (fast) to optimal (or near optimal) solutions. On the other hand in order to solve the original convex problem (2.4) the standard approach (projected gradient descent) involves enforcing the PSD constraint at every iteration by calculating a significant number of eigenvalues and eigenvectors and only keeping the rank one components of the matrix which correspond to non-negative eigenvalues. This means that every iteration is computationally heavy, and that's why in Bhojanapalli et al. (2015) they propose the non-convex formulation (2.4).

In the rest of the review we will study why the simple update rule:

$$U^+ = U - \eta\nabla f\left(UU^\top\right)\cdot U$$

under the proper choice of the step size η and the initial point U^0, converges linearly to a neighborhood of U_r^*, where $X^* = U^*(U^*)^\top$ is the optimal solution to problem (**??**).

Formally, our algorithm is

Algorithm 3: Factored Gradient Descent (FGD)

Initialize: Initial estimate X_0 (see Section 2.4), step size η, rank r
$U_0 = \text{Cholesky}(X_r^0)$
for $k = 1, 2, \ldots, K$ **do**
  $U_k = U - \eta \nabla f(U_{k-1}U_{k-1}^T)U_{k-1}$
return $X_K := U_K U_K^T$

2.2.1 Notation and Assumptions

Let $\mathbb{S}_+^n$ denote the PSD cone, and let $f : \mathbb{S}_+^n \to \mathbb{R}$ be a convex function. Again, we let $X^* = U^*(U^*)^T \in \mathbb{R}^{n\times n}$ denote an optimal solution to Equation[] over the PSD cone, and let $X_r^* \in \mathbb{R}^{n\times n}$ be the r-SVD of X^*, and $U_r^* \in \mathbb{R}^{n\times r}$ be such that $X_r^* = U_r^*(U_r^*)^T$. We will let $\nabla f(\cdot)$ denote the gradient of f with respect to the argument $X \in \mathbb{S}_+^n$, and $\nabla_U f(UU^T)$ to denote the gradient of f with respect to its factorized form. Note that

$$\nabla_U f(UU^T) = 2\nabla f(U)U \tag{2.5}$$

by the chain rule.

As in the rank r PCA setting, we name rotational symmetry: namely, $X_r^* = \tilde{U}_r^*(\tilde{U}_r^*)^T$ for any $\tilde{U}^* = U^*O$, where $O \in \mathcal{O}(r)$. To respect this symmetry, we define a distance function as follows

Definition 2.2.1. Let $U, V \in \mathbb{R}^{n\times r}$. We define $\text{dist}(()U, V) := \min_{R\in\mathcal{O}(r)} \|U - VR\|_F$, where again $\mathcal{O}(r)$ is the set of $r \times r$ matrices for which $R^T R = I$.

Since $\|\cdot\|_F$ is unitarily invariant, we see that equivalently, $\text{dist}(()U, V) = \min_{R\in\mathcal{O}(r)} \|UR - V\|_F = \min_{R_1,R_2\in\mathcal{O}(r)} \|UR_1 - VR_2\|_F$.

Convexity Assumptions

For Algorithm [] to succeed, we will need f to obey the same convexity and smoothness assumptions that are made in the convex optimization literature. We assume that $f : \mathbb{S}_+^n$ is M-smooth, that is

$$\|\nabla f(X) - \nabla f(Y)_F \leq M\|X - Y\| \tag{2.6}$$

for all $X, Y \in \mathcal{S}_+^n$. Under the smoothness assumption alone, Bhojanapalli et al. (2015) demonstrate that, if $X^* \approx X_r^*$, and if the initialization is suitable, then we shall see that Algorithm [] converges to an approximate solution to Equation at the rate of $\frac{1}{k}$, where k is the number of iterations. Note that this is analogous to the rates of standard gradient descent for smooth, convex functions, though accelerated methods can improve the rate to $\frac{1}{k^2}$. cite

Just as in the standard convex analysis literature, achieving linear convergence rates will require that to impose the additional assumption of strong convexity. It turns out that many objectives of interest for matrix problems (such as RIP matrix sensing, as we shall see in Section []) are not strongly convex over the set of all PSD matrices, but we can we can get away with a more limited definition of restricted strong convexity, defined as follows

Definition 2.2.2. We say that a f is (m, r)-restircted strongly convex if, for any rank r $X, Y \in \mathbb{S}^n_+$, it holds that

$$f(Y) \geq f(X) + \langle \nabla f(X), X - Y \rangle + \frac{m}{2}\|X - Y\|_r^2 \tag{2.7}$$

When f is also M smooth, we designated the condition number of f by $\kappa := M/m$.

Note that when $r = n$, we recover m-strong convexity over all matrix $X, Y \in \mathbb{S}^n_+$. The best way to think about restricted strong convexity is that f is convex, *and then some*: that is, the gradient approximations at X underestimate $f(Y)$ by a factor which grows with the square distance between X and Y.

Assumptions beyond convexity

Unfortunately, we can't get simply get away with the f being nice and convex; indeed, the reparameterization $f(UU^T)$ is a nonconvex one, and suffers from fixed points and saddle points. To see this, suppose that $f(UU^T) := \|UU^T - X^*\|_F^2$. Then,

$$\nabla_U f(UU^T) = 4(UU^T - X^*)U \tag{2.8}$$

In particular, if $X^* = U^*(U^*)^T$, and $U = U^*_{r-1}$, then we have that

$$\nabla_U f(UU^T) = 4(U^*_{r-1}U^*_{r-1} - U^*(U^*)^T)U^*_{r-1} = 0 \tag{2.9}$$

Since the columns of U lie in the span of the first $r-1$ eigendirections of X^*, whilst the rows of $U^*_{r-1}U^*_{r-1} - U^*(U^*)^T$ lie in span of the the remaining $n-r$ eigendirections of X^*. In other words, rank deficient estimates can yield fixed points. Moreover, if the true solution X^* is not well approximated by its low rank projection X^*_r, then it is not clear how to show that factorized gradient descent will produce reasonable solutions; in fact, if $X^* \not\approx X^*_r$, then the low rank approximation is perhaps mispecified, and some other dimension reduction strategy might be better advised.

With these observations in mind, we present to assumptions central to the analysis, one for the sublinear rates for smooth FGD,

Assumption 1. *[Assumptions for Smooth Gradient Descent] We assume that U_0 is sufficiently close to U^*_r, in the sense that*

$$\operatorname{dist}(U_0, U^*_r) \leq \rho\sigma_r(U^*_r) \quad \textit{where} \quad \rho = \frac{1}{100}\frac{\sigma_r(X^*)}{\sigma_1(X^*}) \tag{2.10}$$

Further, we require that X^* is close to X^*_r, in the sense that

$$\|X^* - X^*_r\|_F \leq \sigma_r(X^*)\rho \tag{2.11}$$

where ρ is as defined above

and slightly stronger assumptions that will enable us to establish linear converge rates under the strong convexity assumptions:

Assumption 2. *[Assumptions for Smooth Gradient Descent] We assume that U_0 is sufficiently close to U_r^*, in the sense that*

$$\text{dist}(U_0, U_r^*) \leq \rho' \sigma_r(U_r^*) \quad \textit{where} \quad \rho' = \frac{1}{100\kappa} \frac{\sigma_r(X^*)}{\sigma_1(X^*}) = \rho/\kappa \tag{2.12}$$

Further, we require that X^* is close to X_r^*, in the sense that

$$\|X^* - X_r^*\|_F \leq \frac{\rho'}{2\sqrt{\kappa}} \sigma_r(X^*) \tag{2.13}$$

where ρ' is as defined above.

2.2.2 Results

We are now ready to state our result. Even though their proofs make use of some technical linear algebraic arguments, the moral is quite similar: with a close enough initialization, depending crucially on the conditioning of the problem (i..e, condition number of X^* and approximation quality of $X^* \approx X_r^*$), then the optimization problem essentially looks convex. We shall also need an appropriately chosen step size, the selection of which is discussed in further detail in Section 2.2.3:

Definition 2.2.3. We define our step size η by

$$\eta = (16(M\|X_0\|_2 + \|\nabla f(X_0)\|))^{-1} \tag{2.14}$$

We then have the following results. In the smooth case,

Theorem 2.2.1. *Let $X^* = U^*(U^*)^T$ denote a global minimizer of f over the PSD cone, and let $X_0 = U_0 U_0^T$ be the initial solution. Then if $f(X^0) > f(X_r^*)$, and Assumption 1 is true, then the k-th iterate of FGD with step size η chosen as in Definition 2.2.3 satisfies*

$$f(X^*) - f(X^*) \leq \frac{6\text{dist}(U_0, U_r^*)^2}{k\eta^* + 6\frac{\text{dist}(U_0, U_r^*)^2}{f(X_0) - f(X_r)}} \tag{2.15}$$

where $\eta^* = (16(M\|X^*\|_2 + \|\nabla f(X^*)\|))^{-1}$

In the strongly convex case, we have

Theorem 2.2.2. *Let U_0 be the current iterate of FGD, which satisfies Assumption 2 hold; that is $\text{dist}(U_0, U^*) \leq \rho' \sigma_r(U_r^*)$. Then, if η is chosen as in Equation 2.2.3, the next iterate $U_1 = U_0 - \eta \nabla f(U_0 U_0) U_0$ satisfies the constraction*

$$\text{dist}(U_1, U_r^*)^2 \leq \alpha \text{dist}(U, U_r *) + \beta \|X^* - X_r^*\|^2 \tag{2.16}$$

for

$$\alpha = 1 - \frac{m\sigma_r(X^*)}{208(M\|X^*\|_2 + \|\nabla f(X^*)\|_2)} \quad \textit{and} \quad \beta = \frac{M}{24(M\|X^*\|_2 + \|\nabla f(X^*)\|_2)} \tag{2.17}$$

Furthermore, U_1 satisfies Assumption 2, that is, $\text{dist}(U_1, U^*) \leq \rho' \sigma_r(U_r^*)$

We remark that, when $X^* = X_r^*$ is exactly rank r, then this contraction implies linear convergence to the unique optimum X^*. We also find it relevant to comment on the dependenc of the contraction on the conditioning of X^*

Remark 2. Observe that the convergence rate factor a depends both on the condition number $k = \frac{M}{n}$ of f, as well on the condition number $\tau(X_r^*) = \frac{\sigma_1(X^*)}{\sigma_r(X^*}$ of X_r^*, and on $\|\nabla f(X^*)\|_2$. Because of the convergence rate of classical gradient descent, we already expected the dependence on the condition number k of f. The dependence on $\tau(X_r^*)$ comes from the fact that, in the factored form, the condition number of the hessian of f $\nabla_U^2 f(UU^T)$ incurs a on the condition number of $\tau(X_r^*)$. For further discussion, see Appendix E in Bhojanapalli et al. (2015). We remark that this new dependence on condition number is analogous to the way that switching to a factored formulation changes the gradients from $\nabla f(X)$, to $\nabla f(UU^T)U$, which may possibly introduce saddle points (see Equation 2.9).

2.2.3 A Note on Step Size Selection

Some caution should be given when selecting the step size η. In particular we must make sure that when we are close to X^*, we do not " overshoot" the optimum X^*.

In order to get some intuition on how small the step size should be, let's consider the special case where $r = 1$ and f is a separable function with $f(X) = \sum_{ij} f_{ij}(X_{ij})$. Define $g(u) = f(uu^\top)$ and then it is easy to see that:

$$\nabla g(u) = \nabla f(uu^\top) \cdot u$$

$$\nabla^2 g(u) = mat\left(diag(\nabla^2 f(uu^\top) \cdot vec(uu^\top)\right) + \nabla f(uu^\top)$$

where $mat : \mathbb{R}^{n^2} \to \mathbb{R}^{n\times n}, vec : \mathbb{R}^{n\times n} \to \mathbb{R}^{n^2}$ and $diag : \mathbb{R}^{n^2\times n^2} \to \mathbb{R}^{n^2\times n^2}$ are the matricization, vectorization and diagonalization operations respectively.

Now assuming that u is close to the optimum, standard convex optimization suggests that η should be chosen so that $\eta < \frac{1}{\nabla^2 g(.)}$. If we further assume that f is M-smooth, i.e. $\|\nabla^2 f(uu^\top)\| \leq M$, and that the initial point X^0 is constant relative distance apart to X^*, we can deduce that we should pick η so that:

$$\eta < \frac{1}{\nabla^2 g(.)} \propto \frac{1}{M\|X^0\|_2 + \|\nabla f(X^0)\|_2}$$

2.3 Proof Sketch of Results

In this section, we sketch the proofs of Theorems 2.2.1 and 2.2.2. The arguments mirror standard proof of converge in gradieny descent, when use that that the gradients of convex functions point towards the optimal solutions: formally, if $f : \mathbb{R}^d \to \mathbb{R}$ is convex over a convex set $\mathcal{W}$, and $w^* \in \arg\min_{w\in\mathcal{W}} f(w)$, then

$$\forall w \in \mathcal{W} : \langle \nabla f(w), w - w^* \rangle \geq 0 \tag{2.18}$$

and, if f is m-strongly convex, we can replace the zero on the right hand side with a $\frac{m}{2}\|w - w^*\|^2$.

In the interest of brevity, we shall only handle the strong convex case, which is more more straighforward, as the strong convexity gives us a bit of "wiggle-room". The key step in the proof is the descent lemma, which shows that the factorized gradient $\nabla f(U_0)U_0$ is well correlated with the difference between between U and a well-chosen representative U_r^* of the set $\{\tilde{U}_r^* : \tilde{U}_r^*(\tilde{U}_r^*)^T = X_r^*\}$. Formally, let

$$R_U := \arg\min_{R\in\mathcal{O}(r)} \|U - U_r^* R\|_F \quad \text{and} \quad \Delta = U - U_r^* R_U \tag{2.19}$$

Then,

Lemma 2.3.1. *Let f be M-smooth and (m,r)-restricted strongly convex. Then, under Assumption 2.2.2, then there is an $R_U \in \mathcal{O}(r)$ for which*

$$\langle \nabla f(X_0)U_0, U - U_r^* R_U\rangle \geq \frac{2}{3}\eta\|\nabla f(UU^T)U\|_F^2 + \frac{3m}{20}\sigma_r(X^*)\| - \frac{M}{4}\|X^* - X_r^*\|_F^2 \tag{2.20}$$

Proof. For ease of notation, let $X = X_0$ and $U = U_0$. Adding and subtracting $\frac{1}{2}X_r^*$, we have

$$\begin{aligned}
\langle \nabla f(X)U, U - U_R^* R_u\rangle &= \langle f(X), X - U_r^* R_U U^T\rangle \\
&= \frac{1}{2}\langle f(X), X - X_r^*\rangle + \frac{1}{2}\langle \nabla f(X), X + X_r^* - U_r^* R_U U^T \\
&= \frac{1}{2}\langle f(X), X - X_r^*\rangle + \frac{1}{2}\langle \nabla f(X), \Delta\Delta^T
\end{aligned}$$

The strategy is to the strong convexity of f to give a large lower bound on $\langle f(X), X - X_r^*$. On the other hand, the term $\frac{1}{2}\langle \nabla f(X), \Delta\Delta^T$ can be possibly negative, but we show that, in a sufficiently small radius of X_r^*, this term scales like $O(\|\Delta\|^2)$, and is dominated by $\langle f(X), X - X_r^*$. Formally, if we define Q_U to be the orthgoprojector onto the column space of U, and define the intermediate step size

$$\hat{\eta}^{-1} = 16\left(M\|X\|_2 + \|\nabla f(X)Q_U\|_2\right) \tag{2.21}$$

then we have the following estimates

Lemma 2.3.2.

$$\begin{aligned}
\langle \nabla f(X), X - X_r^*\rangle &\geq \frac{18\hat{\eta}}{10}\|\nabla f(X)U\|_F^2 + \frac{m-M}{2}\|X - X_r^*\|_F^2 \\
\langle \nabla f(X), \Delta\Delta^T\rangle &\geq -\frac{2\hat{\eta}}{25}\|\nabla f(X)U\|_F^2 - \left(\frac{m\sigma_r(X^*)}{20} + M\|X^* - X_r^*\|_F^2\right)\|\Delta\|^2
\end{aligned} \tag{2.22}$$

Proof Intutition of Lemma 2.3.2. If we inspect the altered step size $\hat{\eta}$, we note that it transforms η to a quantity which denotes on the projection of $\nabla f(X)$ onto the column space of U, via $\|Q_U\nabla f(X)\|_2$, instead of simply $\|\nabla f(X)\|$. Hence, our goal is to obtain lower bounds which depend on these projected quantities.

For the second estimate, we introduce the column space of U and U^*. Letting Q_M denote the projection onto the column space of M, we have

$$\begin{aligned}
\rangle \nabla f(X), \Delta\Delta^T\rangle &= \rangle Q_\Delta \nabla f(X), \Delta\Delta^T\rangle \\
&\geq -|\text{trace}\left(Q_\Delta \nabla f(X)\Delta\Delta^T\right)| \\
&\geq -\|Q_\Delta \nabla f(X)_2\|_2\|\Delta\|_F^2 \\
&\geq -\left(\|Q_U\nabla f(X)_2\|_2 + \|Q_{U_r^*}\nabla f(X)_2\|_2\right)\|\Delta\|_F^2
\end{aligned}$$

where the second to last step follows from the Von Neuman Trace inequality Mirsky (1975), and the last step follows since the columnspace of Δ lies in the union of column spaces of of U and U_r^*. The final idea is to control $\|Q_U \nabla f(X)_2\|_2$, and then use the initialization assumptions so show that $Q_U \approx Q_{U_r^*}$; note that this explains the dependence of the initialization on the condition number of U_r^*, since we have $\|Q_U - Q_{U_r^*}\|_2 \lesssim \frac{\sigma_1(U_r^*)}{\sigma_r(U_r^*)}\|U - U_r^*\|$ (see Bhatia (2013))

For the first estimate, we need to construct a pseudo iteration $\hat{U} := U - \hat{\eta}\nabla f(X)U$, and $\hat{X} = \hat{U}\hat{U}^T$ which provides a larger lower bound than that from using the normal step size η, so that this lower bound can counter the possibly large negative term from the second line in Equation 2.22.

From smoothness of f and the fact that $X^* \in \arg\min_X f(X)$, we have

$$\begin{aligned} f(X) &\geq f(\hat{X}) - \langle \nabla f(X), \hat{X} - X\rangle - \frac{M}{2}\|\hat{X} - X\|_F^2 \qquad (2.23) \\ &\geq f(X^*) - \langle \nabla f(X), \hat{X} - X\rangle - \frac{M}{2}\|\hat{X} - X\|_F^2 \qquad (2.24) \end{aligned}$$

Further, since X_r^* is feasible, we have by smoothness, we also have that $f(X_r^*) \leq f(X^*) + \langle \nabla f(X^*), X_r^* - X^*\rangle + \frac{M}{2}\|X_r^* - X^*\|_F^2$. Now, the paper claims that $\nabla f(X^*), X_r^* - X^*\rangle = 0$, since $\nabla f(X^*), X^*\rangle = 0$ by the KKT conditions. The argument is that, if $\nabla f(X^*)$ is orthogonal to X^*, then it is also orthogonal to $X_r^* - X^*$, whose columns space is the bottom $n - r$ eigenvectors of X^*. Unfortunately, this argument is not entirely accurate, since $\nabla f(X^*)$ may not be PSD, and so $\nabla f(X^*), X_r^* - X^*\rangle$ may not be 0. However, we will assume that $\nabla f(X^*), X_r^* - X^*\rangle = 0$ to continue with the proof.

By combining the above observations with (m, r)-restricted strong convexity, which shows that $f(X_r^*) \geq f(X) + \langle \nabla f(x), X_r^* - X\rangle + \frac{m}{2}\|X_r^* - X\|_F^2$, we have

$$\langle \nabla f(X), X - X_r^*\rangle \geq \langle \nabla f(X), X - \hat{X}\rangle - \frac{M}{2}\|\hat{X} - X\|_F^2 + \frac{m - M}{2}\|X_r^* - X\|_F^2 \qquad (2.25)$$

The proof now follows from a lot of algebraic manipulation, but the key inside to ease the algebra is to compate the pseudo iterate $\hat{X}$ to X by noting that,

$$\hat{X} = \hat{U}\hat{U}^T = X - \hat{\eta}\nabla f(X)X\Lambda - X\Lambda^T\hat{\eta}\nabla f(X)X \qquad (2.26)$$

where $\Lambda = I - \frac{\hat{\eta}}{2}Q_U Q_U^T \nabla f(X) \in \mathbb{R}^{n\times n}$. By our conditions on the step size of $\hat{\eta}$, and Λ is very well conditioned:

$$\frac{31}{32}I \preceq \Lambda \preceq \frac{33}{32}\Lambda \qquad (2.27)$$

so that, morally, $\Lambda \approx I$. □

Concluding the proof of our Descent Lemma 2.3.1 is even more algebra, but relies crucially on the assumption that $\|X^* - X_r^*\|$ is small, and that $\hat{\eta} = \Omega(\eta)$, which relies heavily on the fact that X is well initialized, and $\|X - X_r^*\|$ is also not too large. □

Concluding the proof of Theorem 2.2.2 is yet even more algebra: Again, let $U = U_0$ and $X = X_0$. We can begin by decomposing the distance between U_1 and U_r^* using the

polarization identity

$$\begin{aligned} \operatorname{dist}(U_1, U_r^*) &\leq \|U_1 - U_r^* R_U\|_F^2 = \|(U_1 - U) - (U_r^* R_U - U)\|_F^2 && (2.28) \\ &\leq \|U_1 - U\|_F^2 + \|U - U_r^* R_U\|_F^2 - 2\langle U_1 - U, U_r^* R_U - U\rangle && (2.29) \end{aligned}$$

Essentially, we can then use Lemma 2.3.1 to show that the decrease from $\langle U_1 - U, U_r^* R_U - U\rangle$ outweighs the contribution of the term $\|U_1 - U\|_F^2 + \|U - U_r^* R_U\|_F^2$.

2.4 Initialization

Because the problem is non-convex in order to make sure that gradient decent converges we must start from a decent initial point. One way to find such a point is to use one of the standard convex algorithms to do some iterations for the convex problem (**??**), and as soon as we have a U^0 which is within constant error to U^* we can switch to the gradient decent described above in order to get the high precision solution.

In this section we present another way to compute an initial point for a general M-smooth and m-strongly convex function f. We will argue that a scale of $\mathcal{P}_+(-\nabla f(0))$, where $\mathcal{P}_+$ is the projection on the PSD cone operator, yields a good enough initial point. Note that by the strong convexity and smoothness of f, we have that:

$$\left\| \frac{1}{M}\mathcal{P}_+(-\nabla f(0)) - X^* \right\|_F \leq 2\left(1 - \frac{m}{M}\right)\|X^*\|_F$$

this means that $\frac{1}{M}\mathcal{P}_+(-\nabla f(0))$ could serve as a descent initialization, but M is not always easy to compute exactly. Instead, one can use the inequality:

$$m \leq \|\nabla f(0) - \nabla f(e_1 e_1^\top)\|_F \leq M$$

to claim that the point $X^0 = \frac{1}{\|\nabla f(0) - \nabla f(e_1 e_1^\top)\|_F}\mathcal{P}_+(-\nabla f(0))$ is a descent initial point as well.

2.4.1 Test case: Matrix sensing problem

In this section we study the following variation of the matrix sensing problem:

$$\min_{X \in \mathbb{R}^{n\times n}} \frac{1}{2}\|b - \mathcal{A}(X)\|_2^2 \ : \ \operatorname{rank} X \leq r, \ X \succeq 0$$

where $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^p$ is the linear sensing mechanism, such that the i-th entry of $\mathcal{A}(X)$ is given by $\langle A_i, X\rangle$, for $A_i \in \mathbb{R}^{n\times n}$ sub-Gaussian independent measurements.

The Hessian of f is given by $\mathcal{A}^\star\mathcal{A}$, so restricted strong convexity will hold if:

$$\|\mathcal{A}(Z)\|_2^2 \geq c\|Z\|_F^2$$

for a restricted set of directions Z, where $c > 0$ is a small constant.

Instead of strong convexity we will examine the stricter notion of restricted isometry property (RIP).

Definition 2.4.1. A liner map $\mathcal{A}$ satisfies the r-RIP with constant δ_r, if

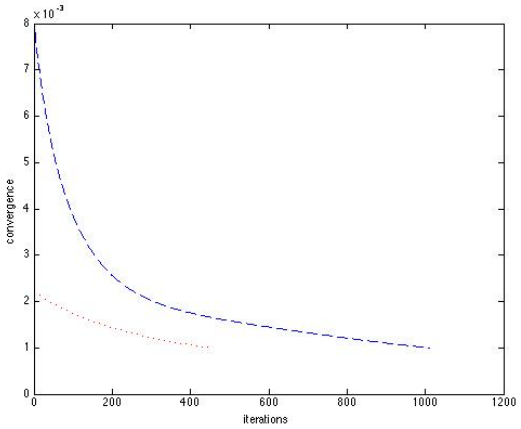
$$(1-\delta_r)\|X\|_F^2 \leq \|\mathcal{A}(X)\|_2^2 \leq (1+\delta_r)\|X\|_F^2$$

It turns out that linear maps that satisfy RIP for low rank matrices, also satisfy the restricted strong convexity.

We will assume that the objective function satisfies RIP. The condition number of the objective depends on the RIP constant of the linear map $\mathcal{A}$, in particular one can show that $k \propto \frac{1+\delta}{1-\delta}$. For δ sufficiently small, and n sufficiently large, $k \approx 1$, which with high probability is the case for $\mathcal{A}$ drawn from a sub-Gaussian distribution.

For our experimental setup we synthesize the absolute truth $X^* \in \mathbb{S}_+^n$, as a rank $r = 4$ positive semidefinite matrix of dimension $n \times n = 64 \times 64$. Using a random Gaussian linear operator $\mathcal{A}$ we sub-sample X^* by observing $m = 512$ entries according to $y = \mathcal{A}(X^*)$. The stopping criterion for factored gradient descent is: $\frac{\|U^+(U^+)^\top - UU^\top\|_F}{U^+(U^+)^\top} < 10^{-3}$. Then we run the algorithm for both a random initial point and the point specified at 2.4, and we plot the convergence, in terms of the ration $\frac{\|X-X^*\|_f}{\|X^*\|_F}$, for both approaches. As we can see the initialization proposed at 2.4 converges faster to the optimal point.

Figure 2.1: convergence for a random initialization (blue) vs the proposed initialization (red)



Chapter 3

Noisy Gradients in the Presence of Saddle Points

In this chapter, we extend our scope beyond non-convex matrix factorization problems. Like the previous chapters, we imagine that we have a function $f : \mathbb{R}^d \to \mathbb{R}$ for which f is twice continuously differentiable, though not necessarily convex. The point of this chapter is to investigate very general conditions under which we can compute *locally optimal* solutions to the problem

$$\min f(w) : w \in \mathbb{R}^d \tag{3.1}$$

Note that we are content with only *locally optimal* solutions, since finding globally optimal solutions are known to be hard Ausiello et al. (2012). In fact, even computing local optima is computationally daunting in the general case: indeed, there are no polynomial algorithms which can compute local optima in a polynomial number of steps, and in the the discrete setting where $f : \{0, 1\}^d \to \mathbb{R}$, it is known that computing local minima is PLS complete Johnson and Yannakaki (1988).

One of the primary difficulties is that, without convexity, first order gradient information can be quite deficient. Indeed, f might have saddles points w_0 for which $\nabla f(w_0) = 0$, but $\nabla^2 f(w_0)$ is neither positive nor semidefinite. Thus a naive gradient algorithm would get "stuck" at w_0, even though w is not a local optimum. Recall that such difficulties cannot arise in convex problems, since $\nabla f(w_0) = 0$ implies that w is globally optimal.

In this chapter, we review recent work by Ge et al. (2015) which shows that stochastic gradient descent with added noise can overcome the presence of saddle points in non-convex objectives. Before stating the regularity conditions which underpin this result, lets further develop the underlying intution. Again, suppose that w_0 is a saddle point; that is, $\nabla f(w_0) = 0$, but $\nabla^2 f(w_0)$ is neither positive nor semidefinite. By taking a Taylor expansition around $f(w_0)$, we have

$$f(w_0 + h) \approx \nabla f(w_0)^T h + h^T \nabla^2 f(w_0) h = h^T \nabla^2 f(w_0) h \tag{3.2}$$

Since $\nabla^2 f(w_0) \not\succeq 0$, then $\lambda_{\min}(\nabla^2 f(w_0)) < 0$, and so if h lies in the eigendirection corresponding to $\lambda_{\min}(\nabla^2 f(w_0)) < 0$, $f(w)$ decreases along the direction of $w_0 \to w_0 + h$. It is well known that trust region algorithms(Culot et al. (1992)) and Newton-lke algorithms

(Dauphin et al. (2014)) can use second-order information to leverage negative order information. However, in practical applications, Hessian information can be expensive to compute, and even numerically unstable. The surprising contribution from**?** is that if stochastic gradient descent has enough variance in every direction, is sufficient to "explore" the region around saddle points to find these decreasing eigendirections h, thereby "escaping" the saddle points, and ensuring convergence to local optima of f.

More precisely, we analyze the following algorithm. Let $\mathrm{SG}(w)$ be a noisy, unbiased oracle for gradients of f; that is $\mathbb{E}[\mathrm{SG}(w)] = \nabla f(w)$, and $\|\mathrm{SG}(w) - \nabla f(w)\| \leq Q$ for some absolute constant Q; we will call Q the *radius* of SG. The algorithm will follow by perturbing our gradients $\mathrm{SG}(w)$ with unbiased noise ξ such that $\mathbb{E}[\xi] = 0$.

Algorithm 4: Unconstrained Noisy SGD Ge et al. (2015)

Initialize w_0, η;
for $t = 1, 2, \ldots, T$ **do**
  $w_t = w_{t-1} - \eta(\mathrm{SG}(w_{t-1}) + \xi)$

We can simplify the analysis greatly by wrapping up the randomness in the stochastic oracle SG into ξ. Formally, by letting $\tilde{\xi} := \nabla f(w) - \mathrm{SG}(w) + \xi$, then our updates take steps with approximate gradients $\nabla f(w) + \tilde{\xi}$, where $\mathbb{E}[\tilde{\xi}]$ is unbiased. Shuffling notation, we may therefore regard $\mathrm{SG}(w) = \nabla f(w)$, and ξ as $\tilde{\xi}$. We will require that our noise is bounded, i.e. $\|\xi\| \leq Q$ almost surely, for some parameter $Q > 0$[1]

The crucial requirement is now that ξ varies sufficiently in every direction, and not too much in any one direction. For simplicity, we shall assume that ξ is isotropic, that is $\mathbb{E}[\xi\xi^T] = \sigma^2 I$. In practice, ξ also contains the randomness from the stochastic oracle, so we will actually have $\sigma_1 I \preceq \mathbb{E}[\xi\xi^T] \preceq \sigma_2 I$, and our bounds will end up depending on σ_1 and the condition number of the noise, σ_2/σ_1.

3.1 The Strict Saddle Property

In what follows, we assume that $f(w)$ is twice continuously differentiable. We call w a *stationary point* of f of $\nabla f(w) = 0$. f is said to have the strict saddle property if the following definition is satisfied:

Definition 3.1.1. Let $f(w)$ be a twice differentiable function. We call $f(w)$ a strict saddle if f if $\nabla^2 f(w) > 0$ whenever w is a local minimum, and $\lambda_{\min}(\nabla^2 f(w)) < 0$ whenever w is a stationary poin of f.

To derive actual rates, we shall need a robust, quantitative formulation of Definition 3.1.1:

Definition 3.1.2. Let $f(w)$ be a twice differentiable function. We say that $f(w)$ is an $(\alpha, \gamma, \epsilon, \delta)$ strict saddle if, for any $w \in \mathbb{R}^d$, *one* of the three conditions hold:

1. Strong First Order Information: $\|\nabla f(w)\| \geq \epsilon$

[1] this can be relaxed to $\|\xi\| \leq Q$, e.g when ξ is *iid* Gaussian

2. Strong Second Order Information: $\lambda_{\min}(\nabla^2 f(w)) \leq -\gamma$

3. Nearness to Local Optimum: there is a local minimum w^* such that $\|w - w^*\| \leq \delta$, and $\nabla^2 f(w') \succ \alpha I$ for all $w' : \|w' - w\| \leq 2\delta$ (that is, $f(\cdot)$ is α-strongly convex in a 2δ neighborhood of w^*).

Because our results rely heavily on first- and second-order information, we shall need to impose a further conditions on the smoothness of our gradients and Hessians:

Assumption 3. *We assume that f is β smooth, and that the Hessian of f is ρ-Lipschitz for some $\rho > 0$. That is,*

$$\|\nabla f(w) - \nabla f(w')\| \leq \beta \|w - w'\| \quad \text{and} \quad \|\nabla^2 f(w) - \nabla^2 f(w')\| \leq \rho \|w - w'\| \tag{3.3}$$

With these assumptions in place, we can state the main result of Ge et al. (2015)

Theorem 3.1.1. *Suppose that $f(w) : \mathbb{R}^d \to \mathbb{R}$ is an $(\alpha, \gamma, \epsilon, \delta)$-strict saddle, and $\mathrm{SG}(w)$ is a stochastic gradient oracle for r with radius Q. Moreover, suppose that $|f(w)| \leq B$, f is β-smooth, and its Hessian is ρ Lipschitz. Then, there is an $\eta_{\max} = \Theta^*(1)$ such that for any $\zeta > 0$, and any $\eta \leq \eta_{\max} / \max\{1, \log(1/\zeta\}$, then with probability $1 - \zeta$, Algorithm 4 outputs a point w which is $O^*(\sqrt{\eta \log(1/\eta\zeta)})$-close to a local minimum w^*, in time $t = O^*(\eta^{-2} \log(1/\zeta))$.*

Here, the notation O^*, Ω^*, Θ^* hides dependences which are polynomial in $\alpha, \gamma, \epsilon, \delta, \rho, B, Q, \beta, \sigma^2$ and d, but independent of η and ζ.

3.1.1 Sketch of the Proof

The key intuition behind the proof is to show that, unless the current iterate w is near a local minimum, then Algorithm 4 will make decrease the value of the objective f by $\Omega^*(\mathrm{poly}(\eta))\ O^*(\eta)$. As f is bounded above and below, it follows that after $\mathrm{poly}(\eta)$-steps, f must be in a region about the local minimum.

The simplest ingredient in the proof is to control the behavior of the SGD-algorithm when w is neither near a saddle, nor a local optimum. We will state the following lemma without proof, since it doesn't provide any deep insights into why the *strict saddle* condition proves so useful:

Lemma 3.1.2. *Under the assumptions of Theorem 3.1.1, then if $\|\nabla f(w_t)\| \geq \sqrt{2\eta\sigma^2\beta d}$, where $\sqrt{2\eta\sigma^2\beta d} < \epsilon$, then we have*

$$\mathbb{E}[f(w_{t+1}] \leq f(w_t) - \frac{\eta^2 \sigma^2 \beta d}{2} \tag{3.4}$$

The second ingredient states that local minima are "sticky", in the sense that if an iterate w_t lands in the basin of a local minimum, and our step size is chosen sufficiently small, then all future iterates will remain in that basin with high probability.

Lemma 3.1.3. *Given an iterate w_t which is $O^*(\sqrt{\eta}) < \delta$ close to aa local minimum w^*, then in $O^*(\eta^{-2} \log(1/\zeta))$ steps, all future iterates w_{t+i} will be $O^*(\sqrt{\eta \log(1/\zeta\eta)})$-close with probability at least $1 - \zeta/2$.*

Again, the proof mirrors standard arguments in the convex optimization literature Rakhlin et al. (2011), and we omit it in the interest of brevity. It is the last proof ingredient - quantifying how long it takes for iterates to "escape" the saddles - which captures the true novelty of Theorem 3.1.1:

Lemma 3.1.4. *Let w_t be a point where $\|\nabla f\| \leq \sqrt{2\eta\sigma^2\beta d}$ and $\lambda_{\min}(\nabla^2(w_t)) \leq -\gamma$, then after $T = T(w_t)$ steps,*

$$\mathbb{E}f(w_{T+t}) - f(w_t) \leq -\tilde{\Omega}(n) \tag{3.5}$$

Furthermore any w_t, $T(w_t) \leq O(\log d/\gamma\eta)$.

3.1.2 Proof Sketch of Lemma 3.1.4

The technique to Proof Lemma 3.1.4 is to build up a series of quadratic approximations to the iterates w_t. For ease of notation, let w_0 be our initial point for which $\|\nabla f\| \leq \sqrt{2\eta\sigma^2\beta d}$, and let $\mathcal{H} := \nabla^2 f(w_0)$,we define the second order approximation to f via

$$\tilde{f}(w) = f(w_0) + \nabla f(w_0)^T(w - w_0) + \frac{1}{2}(w - w_0)^T\mathcal{H}(w - w_0) \tag{3.6}$$

We also define $\tilde{w}_t$ to be iterates obtained by running Algorithm 4 on the function $\tilde{f}$. It will transpire that the approximate-iterates $\tilde{w}_t$ can be controlled using the Hessian information, and we can use the smoothness assumptions on ∇f and $\nabla^2 f$ to verify that the approximate-iterates do in fact approximate the true iterates w_t obtained by SGD on the true function f. One can verify the following recurence relations analytically

$$\begin{aligned} \nabla\tilde{f}(\tilde{w}_t) &= (1 - \eta\mathcal{H})^t\nabla f(w_0) - \eta\mathcal{H}\sum_{\tau=0}^{t-1}(1 - \eta\mathcal{H})^{t-\tau-1}\xi_t \\ \tilde{w}_t - w_0 &= -\eta\sum_{\tau=0}^{t-1}(1 - \eta\mathcal{H})^{\tau}\nabla f(w_0) - \eta\sum_{t=0}^{t-1}(1 - \eta\mathcal{H})^{t-\tau-1}\xi_t \end{aligned} \tag{3.7}$$

On the other hand, let T^* be any time T for which

$$\frac{d}{\eta\gamma_0} \leq \sum_{\tau=0}^{T-1}(1 + \eta\gamma_0)^{2\tau} < \frac{3d}{\eta\gamma_0} \tag{3.8}$$

Using the fact that the sum in the above display diverges, it is a technical and rather unilluminating excercise to show that such a T^* is guaranteed to exist; hence, we omit the proof of this fact in the interest of concision. Moreover, tedious applications of martingale concentration arguments, and of the smoothness properties of the gradients and Hessians of f, can be used to show that approximate iterates $\tilde{w}_t$ and gradients $\nabla\tilde{f}(\tilde{w}_t)$ uniformly approximate the true iterations w_t and gradients $\nabla f(w_t)$ for all times $0 \leq t \leq T^*$. Formally, we have the following lemmaGe et al. (2015), which we cite without proof:

Lemma 3.1.5. *Let w_0 be an initial point satisfying $\lambda_{\min}(\nabla^2 f(w_0)) \leq -\gamma$. Then for all times $0 \leq t \leq T^*$, it holds with probability at least $1 - \tilde{O}(\eta^2)$*

$$\|w_0 - \tilde{w}_t\| \leq O^*(\eta^{1/2}\log\frac{1}{\eta}), \quad \|w_t - \tilde{w}_t\| \leq O^*(\eta\log^2\tfrac{1}{\eta}), \quad \|\nabla f(w_t) - \nabla\tilde{f}(\tilde{w}_t)\| \leq O^*(\eta\log^2\frac{1}{\eta})$$

Again, we omit the proof in the interest of brevity. With the above result in place, we are now ready to sketch the proof of Lemma 3.1.4:

Proof of Lemma 3.1.4. The central idea in the proof of Lemma 3.1.4 is the Taylor expansion

$$f(w) \leq f(w_0) + \nabla f(w_0)^T (w - w_0) + \frac{1}{2}(w - w_0)^T \mathcal{H}(w_0)(w - w_0) + \frac{\rho}{6}\|w - w_0\|^2 \quad (3.9)$$

For ease of notation, set $\mathcal{H}(w_0) = \mathcal{H}$, $T = T^*$, and define $\delta_0 := \tilde{w}_T - w_0$, and $\delta = w_T - \tilde{w}_T$, so that $w_T - w_0 = \delta_0 + \delta_T$. Manipulating Equation 3.9, we have

$$f(w_T) - f(w_0) \quad \leq \quad [\nabla f(w_0)^T \delta_0 + \frac{1}{2}\delta_0^T \mathcal{H} \delta_0] + [f(w_0)^T \delta_T + \frac{1}{2}\delta_T^T \mathcal{H}\delta_T + \delta_T^T \mathcal{H}\delta_0 + \frac{\rho}{6}\|\delta_0 - \delta_T\|^3]$$

Next, let $\Lambda_0 := \nabla f(w_0)^T \delta_0 + \frac{1}{2}\delta_0^T \mathcal{H}\delta_0$ denote the first term, and $\Lambda_T := f(w_0)^T \delta_T + \frac{1}{2}\delta_T^T \mathcal{H}\delta_T + \delta_T^T \mathcal{H}\delta_0 + \frac{\rho}{6}\|\delta_0 - \delta_T\|^3$ denote the second term, so that $f(w_T) - f(w_0) \leq \Delta_0 + \Delta_T$. Thus, if $\mathcal{E}$ is the event

$$\{\forall t \leq T : \|\tilde{w}_t - w_0\| \leq O^*(\eta^{1/2} \log \frac{1}{\eta}), \|w_t - \tilde{w}_t\| \leq O^*(\eta \log^2 \frac{1}{\eta})\| \quad (3.10)$$

Lemma 3.1.5 ensures that $\Pr(\mathcal{E}) \geq 1 - O^*(\eta^2)$. Hence,

$$\begin{aligned} \mathbb{E}[f(w_T) - f(w_0)] &\leq \mathbb{E}[\Lambda_0] + \mathbb{E}[\Lambda_T] & (3.11) \\ &\leq \mathbb{E}[\Lambda_0] + \mathbb{E}[\Lambda_T \mathbb{1}(\mathcal{E})] + \mathbb{E}[\Lambda_T \mathbb{1}(\mathcal{E}^c)] & (3.12) \end{aligned}$$

Here we present a computation that $\mathbb{E}[\Lambda_0] \leq -\frac{\eta\sigma^2}{2}$, which is the crux of the proof, not only of Lemma 3.1.4, but really of Theorem 3.1.1. The error terms $\mathbb{E}[\Lambda_T \mathbb{1}(\mathcal{E})] + \mathbb{E}[\Lambda_T \mathbb{1}(\mathcal{E}^c)]$ are straightforward to control by $o^*(\eta)$ using Lemma 3.1.5, and we omit their proof.

We remark that, in the manuscript of Ge et al. (2015), this computation is greatly abgridged, and some of the key steps are ommited. Thus, the following calculation we present is one which is independent of Ge et al. (2015): For the first order term in $[\nabla f(w_0)^T \delta_0 = \nabla f(w_0)^T (\tilde{w}_T - w_0)$, define $\bar{H} := \sum_{\tau=0}^{t-1} \nabla f(w_0)^T (1 - \eta\mathcal{H})^\tau$. Then, we have

$$\begin{aligned} \mathbb{E}[\nabla f(w_0)^T (\tilde{w}_T - w_0)] &= -\eta \nabla f(w_0)^T \Sigma \nabla f(w_0) - \mathbb{E}[\eta \sum_{t=0}^{t-1} \nabla f(w_0)^T (1 - \eta\mathcal{H})^{t-\tau-1} \xi_t] \\ &= -\eta \nabla f(w_0)^T \bar{H} \nabla f(w_0) \end{aligned}$$

where we use the fact that $\mathbb{E}[\xi_t]$ are all unbiased. Similarly, using the fact that $\mathbb{E}[\xi_{t_1}] = 0$ and $\mathbb{E}[\xi_{t_1}\xi_{t_2}] = 0$ for $t_1 \neq t_2$, while $\mathbb{E}[\xi_t^2] = \sigma^2$,

$$\begin{aligned} \mathbb{E}[(\tilde{w}_T - w_0)^T \mathcal{H}(\tilde{w}_T - w_0)] &= \eta^2 \nabla f(w_0)^T \bar{H}\mathcal{H}\bar{H}\nabla f(w_0) \\ &+ \mathbb{E}[\eta^2 \sum_{\tau=0}^{t-1} \xi_\tau^T (1 - \eta\mathcal{H})^{t-\tau-1} \mathcal{H} (1 - \eta\mathcal{H})^{(\sqcup - \tau - \infty)} \xi_\tau] \\ &= \eta^2 \nabla f(w_0)^T \mathcal{H}\bar{H}^2 \nabla f(w_0) + \eta^2\sigma^2 \mathrm{tr}(\mathcal{H} \sum_{\tau=0}^{T-1} (1 - \eta\mathcal{H})^{2\tau}) \end{aligned}$$

where we used the fact that $\bar{H}, \mathcal{H}$ and all the terms $(1-\eta\mathcal{H})^\tau$ commute. Thus,

$$\mathbb{E}[\Lambda_0] = -\eta\nabla f(w_0)^T(\eta\bar{H} - \frac{\eta}{2}\mathcal{H}\bar{H}^2)\nabla f(w_0) + \frac{\eta^2\sigma^2}{2}\mathrm{tr}(\mathcal{H}\sum_{\tau=0}^{t-1}(1-\eta\mathcal{H})^{2\tau}) \tag{3.13}$$

Let $\lambda_1 \geq \cdots \geq \lambda_d$ denote the eigenvalues of $\mathcal{H}$. We will prove that $\bar{H} - \frac{\eta}{2}\mathcal{H}\bar{H}^2 \succ 0$ (note this quantity is symmetric, since $\mathcal{H}$ are symmetric and commute $\bar{H}$), which will imply that

$$\begin{aligned}
\mathbb{E}[\Lambda_0] &\leq \frac{\eta^2\sigma^2}{2}\mathrm{tr}(\mathcal{H}\sum_{\tau=0}^{t-1}(1-\eta\mathcal{H})^{2\tau}) \\
&= \frac{\eta^2\sigma^2}{2}\sum_{i=1}^{d}\lambda_i(1-\eta\lambda_i)^{2\tau} \\
&\leq \frac{\eta^2\sigma^2}{2}\sum_{i:\lambda_i>0}\lambda_i(1-\eta\lambda_i)^{2\tau} + \frac{\eta^2\sigma^2}{2}\gamma_0\sum_{\tau=0}^{t-1}(1-\eta\gamma_0)^{2\tau}
\end{aligned}$$

Now, for any $\lambda > 0$, we have

$$\lambda\sum_{\tau=0}^{t-1}(1-\eta\lambda)^{2\tau} \leq \lambda\sum_{\tau=0}^{\infty}(1-\eta\lambda) = \frac{\lambda}{1-(1-\eta\lambda)} = \frac{1}{\eta} \tag{3.14}$$

While, on the other hand, T was chosen (see Equation 3.8) so that $\gamma_0\sum_{\tau=0}^{t-1}(1-\eta\gamma_0)^{2\tau} \geq \frac{d}{\eta}$. Hence

$$\mathbb{E}[\Lambda_0] \leq \frac{\eta^2\sigma^2}{2}\left[\frac{d-1}{\eta} - \frac{d}{\eta}\right] \leq -\frac{\eta\sigma^2}{2} \tag{3.15}$$

as needed. It finally remains to verify

$$\bar{H} - \frac{\eta}{2}\bar{H}^2\mathcal{H} \succ 0 \tag{3.16}$$

Since $\bar{H}$ are $\mathcal{H}$ commute, and $\bar{H}$ is symmetric and PSD, it suffices to verify that $I - \frac{\eta}{2}\bar{H}\mathcal{H} \succ 0$, or equivalently, that $\frac{\eta}{2}\bar{H}\mathcal{H} \preceq I$. Diagonalizing, it suffices to show that

$$\frac{\eta\lambda_i}{2}\sum_{t=0}^{T-1}(1-\eta\lambda_i)^t \leq 1 \tag{3.17}$$

If $\lambda_i \leq 0$, this is obvious. On the otherhand, when $\lambda_i > 0$, the same argument in Equation 3.14 lets us control $\frac{\eta\lambda_i}{2}\sum_{t=0}^{T-1}(1-\eta\lambda_i)^t \leq \frac{\eta\lambda_i}{2\eta\lambda_i} \leq 1/2 \leq 1$, as needed.

□

3.2 Gradient Descent with Manifold Constraints

Recall that, in rank-one PCA, we attempted to optimize x^TAx subject to the smooth, though non-convex constraint $x^Tx = 1$. It will transpire that many of the problems of

interest that have the strict-saddle property have similar constraints (and the constraints often spherical). With this in mind, we present a generalization of the results on the previous section to smooth, though possibly non-affine *equality constraints*, which we shall refer to as "manifold constraints." Formally, we consider optimization problems of the form

$$\min_{w \in \mathbb{R}^d} f(w) \quad : c_i(w) = 0, i \in \{1, \dots, m\} \tag{3.18}$$

where the feasibility set is $\mathcal{W} = \bigcap_{i=1}^m \{w : c_i(w) = 0\}$. Thus, we use a projected variant of SGD As in the unconstrained setting, we will again assume, without loss of generality, that

Algorithm 5: Unconstrained Noisy SGD Ge et al. (2015)

Initialize w_0, η;
for $t = 1, 2, \dots, T$ **do**
  $u_t = w_{t-1} - \eta(\mathrm{SG}(w_{t-1}) + \xi)$
  $w_t = \Pi_{\mathcal{W}}(u_t)$

$\mathrm{SG}(w) = \nabla f(w)$, by shifting any of the excess noise from the stochastic gradient oracle into the added noise ξ. We remark that Algorithm 4 is essentially the same as the unconstrained case, but the key difference in the application of a projection global onto $\mathcal{W}$

$$\Pi_{\mathcal{W}}(w) := \arg\min_{u \in \mathcal{W}} \|u - w\|_2^2 \tag{3.19}$$

For affine, and more generally, convex sets $\mathcal{W}$, projecions are well-defined, unique, and have many favorable contraction properties (e.g., see Lemma 3.2.1). In the manifold case, however, we need to proceed with more care. The key idea will be to study affine-approximations to the manifold by defining the Tangent Space in the following section. We will then define appropriate generalizations of the Gradient and Hessian operators, which will let us approximate noisy gradient steps followed by projections as non-projected steps along the tangent space. With this language, we can state the appropriate generalization of the strict saddle. We will conclude the section by specializing our results to the projections on the sphere, which will prepare us for the applications in the two proceeding sections.

3.2.1 The Tangent and Normal Space

While there are numerous works and texts in the field of manifold optimization, it is often easier to work with the standard tools of Euclidean calculus. As a conseuqence, we transform the constrained optimization problem to one which is locally unconstrained problem. To do this, we consider a first-order approximation to the manifold $\mathcal{W}$ at w, which we call the tangent space:

Definition 3.2.1. We define the tangent space at w $\mathcal{T}(w) := \mathrm{span}(\{\nabla c_i(w)\})^{\perp}$, and the normal space $\mathcal{T}^c(w) = \mathrm{span}(\{\nabla c_i(w)\})$. We let $P_{\mathcal{T}(w)}$ and $P_{\mathcal{T}^c(w)}$ denote the orthogonal projectors on the Tangent and Normal Spaces, respectively.

For intuition, suppose that the $c_i(w)$ are affine functions $c_i(w) = C_i^T w + C_0$. Then $\mathcal{W}$ is the affine subspace defined by the itersection of the half planes $C_i^T w + C_0 = 0$. Then,

for any $w \in \mathcal{W}$, $\mathcal{T}(w) = \text{span}(\{C_i\})^\perp$. On the other hand, we know that for any $w \in \mathcal{W}$, we can express $\mathcal{W}$ as the translate of $\text{span}(\{C_i\})^\perp$ $\mathcal{W} = w + \text{span}(\{C_i\})^\perp$. In other words, when the constraints are affine,, $\mathcal{W}$ and $\mathcal{T}$ coincide, up to a translation.

Shortly, we'll make rigorous the claim the tangent space serves as a first order approximation to membership in the manifold for non-affine constraints, in the sense that if $w' = \Pi_{\mathcal{W}}(w - \eta v)$ for η smal enough, then $w' \approx w - \eta P_{\mathcal{T}(w)}(v)$, where $P_{\mathcal{T}(w)}$ is the linear projection onto $\mathcal{T}(w)$. In other words, we can regard $P_{\mathcal{T}(w)}(v)$ as the "component of v which stays on the manifold". The upshot is that, if we take a gradient step $w_t = \Pi_{\mathcal{W}}(w_{t-1} - \eta \nabla f)$, then we can approximate $w_t \approx w_{t-1} - \eta P_{\mathcal{T}(w)}(\nabla f(w))$, and thus regard the manifold gradient $\text{grad}(w) := P_{\mathcal{T}(w)}(\nabla f(w))$ as the direction of $\nabla f(w)$ that roughly "stays on the manifold".

It turns out that $P_{\mathcal{T}(w)}(\nabla f(w))$ has a very clean interpretation in terms of the Lagrangian function:

$$\mathcal{L}(w, \lambda) = f(w) - \sum_{i=1}^{m} \lambda_i c_i(w) \tag{3.20}$$

Now, if w^* is a local optimum, then KKT conditions tell us there is a dual variable λ^* for which

$$\nabla \mathcal{L}(w, \lambda^*) \quad = \quad \nabla f(w) - \sum_{i=1}^{m} \lambda_i^* \nabla c_i(w) = 0 \tag{3.21}$$

The intuition here is that, at a local minimum, the only way to decrease f is to move strictly off the manifold $\mathcal{W}$, and we can think of λ_i^* as the "dual-direction", in the basis give by $\{\nabla c_1(w), \ldots, \nabla c_m(w)\}$, which transports us back onto $\mathcal{W}$ [2].

In general, when w is not a local optimum, then ∇f may also have a component that also lies in $\mathcal{T}(w) = \text{span}(\{\nabla c_i(w)\})^\perp$, namely $P_{\mathcal{T}(w)}(\nabla f(w))$. As noted above, this is the component of ∇f which "remains on the manifold", and moving along this component should correspond to progress in our objective function. Writing the projection operator explicitly as a least squares problem, we get the equivalent defines

$$\begin{aligned} \text{grad}(w) &= P_{\mathcal{T}(w)}(\nabla f(w)) \\ &= \nabla f(w) - \sum_{i=1}^{m} \lambda^*(w)_i \nabla c_i(w) \\ &= \nabla_w \mathcal{L}(w, \lambda)\big|_{\lambda = \lambda^*(w)} \end{aligned} \tag{3.22}$$

where

$$\lambda^*(w) \in \arg\min_{\lambda \in \mathbb{R}^m} \|\nabla f(w) - \sum_i \lambda_i \nabla c_i(w)\| = \arg\min_{\lambda \in \mathbb{R}^m} \|\nabla_w L(w, \lambda)\| \tag{3.23}$$

We remark that $\text{grad}(w)$ is always well defined, and in the case where the $\nabla c_i(w)$ are all linear dependent, we the mapping $w \to \lambda^*(w) : \mathcal{W} \to \mathbb{R}^m$ is well-defined as well.

[2] In manifold optimization, the process of transporting back onto a manifold is known as retraction, and we defer the curious reader toAbsil et al. (2009) for further disccusion

3.2.2 Approximating SPGD with SGD in the Tangent Space

The analysis of projected gradient descent under convex constraints is rather straightforward to analyze: indeed, if a set $\mathcal{W}$ is convex, then projections are contracting

Lemma 3.2.1. *[Simplification of Lemma 3.1 in Bubeck (2014)] Let $w \in \mathcal{W}$ and $u \in \mathbb{R}^d$. Then, $\|\Pi_{\mathcal{W}}(u) - w\| \leq \|w - u\|^2$.*

In particular, if we take $w = w^*$ to be an optimum, w_t the current iterate, $u_{t+1} = w_t + \eta \nabla f(w)$ to be the $t+1$-th iterate, before the projection, and $w_{t+1} = \Pi_{\mathcal{W}}(u)$. Then, $\|w_{t+1} - w^*\|^2 \leq \|u_{t+1} - w^*\|$, so standard subgradient descent arguments can be easily modified to show a decrease in $f(w_{t+1}) - f(w_t)$ Bubeck (2014).

When projecting onto a possibly non-convex manifold, this contractive property no longer holds. Instead, the most we can say is that any possible *increase* in distance between u_{t+1} and w_{t+1} is an order of magnitude smaller than the *decrease* from $f(w_t)$ to $f(u_{t+1})$, at least in the direction of the tangent space. Thus, our strategy will be to show that w_{t+1} is by approximated

$$w_{t+1} \approx w_t - P_{\mathcal{T}(w)} \eta \nabla f(w_t) + o(\eta) \tag{3.24}$$

Before proceeding, we need to impose regularity assumptions on the manifold constraints $c_i(w)$ which roughly ensure that $\mathcal{W}$ is not too curved at any point. To this end, define the matrix $C(w) = [\nabla c_1(w), \ldots, \nabla c_n(w)]$, and note that if $\nabla c_i(w)$ has linearly indpendent columns, then

$$\lambda^*(w) = C(W)^\dagger \nabla f(w) \tag{3.25}$$

To ensure that $\lambda^*(w)$ is stable, we require that the columns of $C(w)$ are robustly independent at each $w \in \mathcal{W}$. Formally, we require

Definition 3.2.2. Define the matrix $C(w) = [\nabla c_1(w), \ldots, \nabla c_n(w)]$. We say that the constraints are α_c-robustly lineraly independent, or α_c-RLI, if, for all $w \in \mathcal{W}$ $\sigma_{\min}(C(w)) \geq \alpha_c$. We a say that $c_i(w)$ are β_i smooth, if, for all $w, w' \in \mathcal{W}$, $\|\nabla c_i(w) - \nabla c_i(w')\| \leq \beta_i \|w - w'\|$.

The α_c-RLI assumption immediately implies that $\|C(w_0)^T(w - w_0)\| = \|C(w_0)^T P_{\mathcal{T}^c(w_0)}(w - w_0)\|^2 \geq \alpha_c^2 \|w - w_0\|^2$. On the other hand, β_i-smoothness implies that, if $w, w_0 \in \mathcal{W}$, $|\nabla c_i(w_0)^T(w - w_0)| = |c_i(w) - c_i(w_0) - \nabla c_i(w_0)^T(w - w_0)| = \frac{\beta_i^2}{2}\|w - w_0\|^2$. Thus, $\|C(w_0)^T(w_0 - w)\|^2 \leq \sum_i (\nabla c_i(w_0)^T(w - w_0))^2 = \|w - w_0\|^2 \frac{\sum_i \beta_i^2}{4}$. In other words, restricted to directions between points $w_0, w \in \mathcal{W}$, $\|C(w_0)^T(w_0 - w)\|$ is on the order of $\|w_0 - w\|^2$, which is much less than $\|w_0 - w\|$ when w, w_0 are close by. Putting these facts together, and defining

$$R^2 := \frac{1}{\alpha_c^2} \sum_i \beta_i^2 \tag{3.26}$$

Then, we have

$$\|P_{\mathcal{T}^c(w_0)}(w - w_0)\|^2 \leq \frac{1}{4R}\|w - w_0\|^4 = \frac{1}{4R}\|P_{\mathcal{T}(w_0)}(w - w_0)\|^4 + \frac{1}{4R}\|P_{\mathcal{T}^c(w_0)}(w - w_0)\|^4$$

That is, to first order, the component of $w - w_0$ whie lies in the normal space is an order of magnitude smaller than the component which lies in the trangent space. This establishes the following lemma, which makes quantitative the intuition that the tangent space is, in fact, a first order approximation to the manifold $\mathcal{W}$ at w:

Lemma 3.2.2. *Assume that $\{c_i\}$ are β_i smooth are α_c-RLI, and set $R = \alpha_c^{-2}\sum_{i=1}^m \beta_i^2$. Then for all $w, w_0 \in \mathcal{W}$ we have that*

$$\|P_{\mathcal{T}^c(w_0)}(w - w_0)\| \leq \frac{1}{2R}\|w - w_0\|^2 \tag{3.27}$$

. Moreover, when $\|w - w_0\| \leq R$, we have

$$\|P_{\mathcal{T}^c(w_0)}(w - w_0)\| \leq \frac{\|P_{\mathcal{T}(w_0)}(w - w_0)\|^2}{2R} \tag{3.28}$$

The β_i-smoothness conditions also ensure that $\|\nabla c_i(w) - \nabla c_i(w)\| \leq \beta\|w - w_0\|$. This lets us show that if w and w_0 are in $\mathcal{W}$ are are sufficiently close, then their tangent spaces and normal spaces roughly coincide. Formally, we have

Lemma 3.2.3. *Assume that $\{c_i\}$ are β_i smooth are α_c-RLI, and set $R = \alpha_c^{-2}\sum_{i=1}^m \beta_i^2$. Then for all $\hat{v}_{\mathcal{T}} \in \mathcal{T}(w)$ and $\hat{v}_{\mathcal{T}^c} \in \mathcal{T}^c(w)$ such that $\|\hat{v}_{\mathcal{T}}\| = \|\hat{v}_{\mathcal{T}^c}\| = 1$, we have*

$$\|P_{\mathcal{T}^c(w_0)}(\hat{v}_{\mathcal{T}})\| \leq \frac{\|w - w_0\|}{R} \quad \textit{and} \quad \|P_{\mathcal{T}(w_0)}(\hat{v}_{\mathcal{T}^c})\| \leq \frac{\|w - w_0\|}{R} \tag{3.29}$$

With these ingredients in place, we can now prove the key lemma using a cute application of the KKT conditions:

Lemma 3.2.4. *Suppose that the constraint $\{c_i\}$ are β_i-smooth and are α_c-RLI. Then, if f is L-lipschitz and the noise bounded, then the iterates of ALgorithm 5 take the form*

$$w_t = w_{t-1} - \eta(\text{grad}(w_{t-}) + P_{\mathcal{T}(w_{t-1})}(\xi_{t-1})) + \text{err}_{t-1} \tag{3.30}$$

where err_{t-1} corrects for the projection, and $\|\text{err}_{t-1}\| \leq$

Proof. Define $v = \nabla f(w_{t-1}) + \xi$. Since $\text{grad}(w_{t-1}) = P_{\mathcal{T}(w_{t-1})}(\nabla f(w_{t-1})$, we want to show that

$$\begin{aligned} & \|w_t - (w_{t-1} - \eta P_{\mathcal{T}(w_{t-1})}v)\| && (3.31)\\ = \; & \|\Pi_{(\mathcal{W})}((w_{t-1} - \eta v) - (w_{t-1} - \eta P_{\mathcal{T}(w_{t-1})}v)\| && (3.32)\\ \leq \; & \frac{4\eta^2\|v\|}{R} && (3.33) \end{aligned}$$

To this end, it suffices to show that for any $w_0 \in \mathcal{W}$ and $v : \|v\| = 1$, then $w_1 = w_0 + \eta v$, and $w_2 = w_0 + \eta P_{\mathcal{T}(w_0)}$ satisfy

$$\|\Pi_{\mathcal{W}}(w_1) - w_2\| \leq \frac{4\eta^2}{R^2} \tag{3.34}$$

We can define $\Pi_{\mathcal{W}}(w_1)$ as the solution to the optimization problem

$$\min_u \|w_1 - u\|^2 : c_i(u), \quad i = 1, \ldots, m \tag{3.35}$$

Forming a Lagrangian $\mathcal{L}(u, \mu) = \|w_1 - u\|^2 + \sum_i \mu_i c_i(u)$, the first order KKT conditions imply that, at an optimal primal-dual solution (u^*, μ^*), we have

$$2(w_1 - u^*) + \sum_{i=1}^m \mu_i \nabla c_i(u^*) \tag{3.36}$$

In other words, $w_1 - u^* \in \text{span}(\{\nabla c_i(u^*)\}) = \mathcal{T}^c(u^*)$. Thus, an application of Lemma 3.2.3 yields

$$\|P_{\mathcal{T}(w_0)}(w_1 - w^*)\| \quad \leq \quad \frac{\|u^* - w_1\|\|w_0 - w^*\|}{R} \tag{3.37}$$

On the other hand, Lemma 3.2.2 shows that

$$\|P_{\mathcal{T}^c(w_0)}(u_1 - w_0)\| \leq \frac{\|u_1 - w_0\|^2}{2R} \tag{3.38}$$

Thus,

$$\begin{aligned}
\|u^* - w_2\| &\leq \|u^* - w_0 - P_{\mathcal{T}(w_0)}(u^* - w_0)\| + \|w_0 + P_{\mathcal{T}(w_0)}(u^* - w_0) - w_2\| \\
&\leq \|P_{\mathcal{T}^c(w_0)}(u^* - w_0)\| + \|P_{\mathcal{T}(w_0)}(w_0 - w_2)P_{\mathcal{T}(w_0)}(u^* - w_0)\| \\
&\leq \|P_{\mathcal{T}^c(w_0)}(u^* - w_0)\| + \|P_{\mathcal{T}(w_0)}(u^* - w_0)\| \\
&\leq \frac{\|u_1 - w_0\|^2}{2R} + \frac{\|u^* - w_1\|\|w_0 - w^*\|}{R}
\end{aligned} \tag{3.39}$$

The last line can now be upper bounded by $4\eta^2/R^2$: Indeed, $\|w_1 - u^*\| \leq \min_{u \in \mathcal{W}} \|u - w_1\|$, so we must have $\|w_1 - u^*\| \leq \|w_0 - w_1\| = \eta$, and, by the triangle inequality, $\|u^* - w_0\| \leq 2\eta$. $\square$

3.2.3 Hessians and Saddles on Manifolds

In order to make use of the strict saddle property, we need to define an appropriate generalization of the Hessian on a manifold. In Euclidean space, the Hessian is just the derivative operator applied coordinate-wise to the gradient. To this end, write

$$\text{grad} f(w) = \nabla f(w) - \sum_i \lambda_i^*(w) \nabla c_i(w) \tag{3.40}$$

If we treat $\lambda_i^* = \lambda_i^*(w)$ as constant with respect to w, then we can define $\text{hess} f(w)$ by differentiating $\nabla f(w) - \sum_i \lambda_i^* \nabla c_i(w)$. Thus, we set

$$\text{hess} f(w) \quad = \quad \nabla^2 f(w) - \sum_i \lambda_i^*(w) \nabla^2 c_i(w) \tag{3.41}$$

Equivalently, we can define hess directly from the Lagrangian, via

$$\text{hess} f(w) \quad = \quad \nabla^2_{ww} \mathcal{L}(w, \lambda^*) \tag{3.42}$$

Suppose that $\nabla^2 L(w, \lambda^*)$ is ρ_L lipschitz in the sense that $\|\nabla^2_{ww}\mathcal{L}(w_1, \lambda^*) - \nabla^2_{ww}\mathcal{L}(w_2, \lambda^*)\| \leq \rho_L \|w_1 - w_2\|$. Now, given feasible points w, w_0, Taylor expanding the Lagrangian about (w_0, λ^*) yields,

$$\mathcal{L}(w, \lambda^*) \quad \leq \quad \mathcal{L}(w_0, \lambda^*) + \nabla_w \mathcal{L}(w_0, \lambda^*) \tag{3.43}$$

$$+ \quad \frac{1}{2}(w - w_0)^T \nabla^w_{ww} \mathcal{L}(w_0, \lambda^*) + \frac{\rho_L}{6}\|w - w_0\|^2 \tag{3.44}$$

Using the definitions of hess and grad, and the fact that $\mathcal{L}(w, \lambda^*) = f(w)$ and $\mathcal{L}(w_0, \lambda^*) = f(w_0)$ for $w, w_0 \in \mathcal{W}$, this gives us

$$f(w) \leq f(w_0) + \text{grad}(w_0)^T(w - w_0) + \frac{1}{2}(w - w_0)^T \text{hess}(w_0)(w - w_0) + \frac{\rho_L}{6}\|w_0 - w\|^3 \quad (3.45)$$

The lagrangian formulation lets us motivate our choise of the hess operator with the second order KKT conditions:

Proposition 3.2.5 (Second Order KKT ConditionsNocedal and Wright (2006))**.**

Note that, when the constraints $c_i(w)$ are affine, hess$f(w)$ coincides exactly with the Euclidean Hessian $\nabla^2 f(w)$.

3.2.4 Strict Saddle

Naively, we might be tempted to define the strict saddle condition in terms of the eigenvalues of the hessian operator hess$f(w)$. But in view of Lemma 3.2.4, each gradient step resembles, to first order, a step in the direction of the tangent space $\mathcal{T}(w)$. Thus, if $w = \Pi_{\mathcal{W}}(w_0 - \eta v)$, then $w - w_0 \approx P_{\mathcal{T}(w_0)}(w - w_0)$. Thus, Equation 3.45 yields the following heuristic

$$\begin{aligned} f(w) &\approx f(w_0) + \text{grad}(w_0)^T P_{\mathcal{T}(w_0)}(w - w_0)(w - w_0) + \frac{1}{2}(P_{\mathcal{T}(w_0)}(w - w_0))^T \text{hess}(w_0) P_{\mathcal{T}(w_0)}(w - w_0) \\ &= f(w_0) + \text{grad}(w_0)^T (w - w_0)(w - w_0) + \frac{1}{2}(w - w_0)^T P_{\mathcal{T}(w_0)} \text{hess}(w_0) P_{\mathcal{T}(w_0)}(w - w_0) \end{aligned} \quad (3.46)$$

where the last line follows since $\text{grad}(w_0) \in \mathcal{T}(w_0)$. We can make the approximation in 3.46 quantitative, but we will leave it informal in the interest of brevity. The main take away, however, is that we need to have controls on grad$f$ instead of simply ∇f, and regularity conditions on the the hess operator *along the directions in the tangent space at* w_0. Indeed, when the constraints $c_i(w) = C_i^T w$ are linear, then $\mathcal{T}(w) = \mathcal{T} = \text{span}\{C_i^T\} = \mathcal{W}$ is a subspace, grad$f = \Pi_{\mathcal{W}}(\nabla f(w))$, hess$(w) = \nabla^2 f(w)$, and the restriction of hess(w) to $\mathcal{W}$ is precisely $\Pi_{\mathcal{W}}$hess$(w)\Pi_{\mathcal{W}}$. Thus, we define the manifold strict saddle condition as

Definition 3.2.3 (General Strict Saddle)**.** We say that a twice differentiable function $f(w)$ with twice differentiable constraints $c_i(w)$ is an $(\alpha, \gamma, \epsilon, \delta)$ strict saddle if, for any $w \in \mathcal{W}$, one of the following holds.

1. $\|\text{grad}(w)\| \geq \epsilon$.
2. The restriction of hess to $\mathcal{T}(w)$ has $\lambda_{\min} \leq -\gamma$. Formally,
$$\min_{v \in \mathcal{T}(w), \|v\|=1} v^T \text{hess}(w) v \leq -\gamma \quad (3.47)$$
3. There is a local mimum w^* such that $\|w - w^*\| \leq \delta$, and for all $w' : \|w^* - w'\| \leq 2\delta$, f looks α-strongly convex along directions $v \in \mathcal{T}(w)$. Formally,
$$\min_{v \in \mathcal{T}(w), \|v\|=1} v^T \text{hess}(w) v \leq \alpha \quad (3.48)$$

3.2.5 Example: *grad* and *hess* on the Sphere

Let's make this discussion more concrete by considering the case where the function f is arbitrary (though satisfies the usual regularity conditions), and $\mathcal{W} = \mathcal{S}^{d-1}$. We can enforce this constraint set with one constraint:

$$c(w) := (\frac{1}{2}\|w\|^2 - 1) \tag{3.49}$$

We then have that $\nabla c(w) = w$, which recovers the elementary fact that, for $w \in \S^{d-1}$, the normal space $\mathcal{T}^c(w)$ is the span of w, and the tangent space is its complement $\text{span}(w)^\perp$. Since $w \in \S^{d-1}$, then $P_{\mathcal{T}^c(w)} = ww^T$, while $P_{\mathcal{T}(w)} = I - ww^T$. Thus, our gradient is exactly

$$\text{grad} f(w) = (I - ww^T)\nabla f(w) \tag{3.50}$$

while, $\lambda^*(w)$ in our Lagrangian formulation is given by

$$\begin{aligned} \lambda^*(w) &= \min_\lambda \|\nabla f(w) - \lambda \nabla c(w)\| && (3.51)\\ &= \min_\lambda \|\nabla f(w) - \lambda w\| && (3.52)\\ &= w^T \nabla f(w) && (3.53) \end{aligned}$$

since $\|w\| = 1$. Since $\text{hess} c(w) = I$, our Hessian is

$$\text{hess} f(w) = \nabla^2 f(w) - (\langle w, \nabla f(w))I \tag{3.54}$$

Thus, our second order approximation for $w = \Pi_{\mathcal{S}^{d-1}}(w_0 + v)$ for $v \in \mathcal{T}(w)$ is

$$f(w) \approx f(w_0) + \text{grad} f(w)^T (I - ww^T)v + v(I - ww^T)\text{grad}^2 f(w)(I - ww^T)v \tag{3.55}$$

We now verify the smoothness and RLI properties. Since $\nabla c(w) = w$, we $\|\nabla c(w) - \nabla c(w')\| \leq \|w -' w\|$, so the constraint is 1-smooth. On the other hand, the matrix $C(w) = [\nabla c(w)]$ has spectral norm $\|\nabla c(w)\| = \|w\| = 1$, for $w \in \mathcal{W}$. Hence, $c(w)$ is also 1-RLI.

Finally, we remark that projections onto the surface of the sphere are easy to component, using the language of tangent spaces:

Claim 3.2.6. *Let $w \neq 0$. Then $\Pi_{\mathcal{S}^{d-1}}(w) = \frac{w}{\|w\|}$ is given by $u^* =$*

Proof. Letting $u^* := \Pi_{\mathcal{S}^{d-1}}(w)$, a KKT argument analogous to the one in Lemma[] shows that $u^* - w$ must lie in the normal space at u^*, which, for the spherical constraint, is precisely the span of u^*. Hence, $w \in \text{span}(u^*)$, so $u^* = tw$ for some $t \in \mathbb{R}$. Since $\|u^*\| = 1$, we must have $u^* = \pm\frac{w}{\|w\|}$, and its easy to check that $\|w/\|w\| - w\| \leq \|w/\|w\| + w\|$, whence $u^* = \frac{w}{\|w\|}$. □

3.2.6 Example: Rank-1 PCA is a Strict Saddle

To further our intuition for the saddle condition, we show that rank 1 PCA of a PSD matrix $A \succ 0$ is a strict saddle, and that the power method for PCA corresponds to projected gradient descent with step size η in the infinite step size limit $\eta \to \infty$.

More quantitatively, let γ be the gap between the first and second largest eigenvalues of A. Then, optimization problem is an $(\gamma, \gamma, 0, 0)$-strict saddle:

$$\min -\frac{1}{2} w^T A w : w \in \mathcal{S}^{d-1} \tag{3.56}$$

and every local optimum is a global optimum.

Remark. While this is $(\gamma, \gamma, 0, 0)$ saddle condition is obviously insuficient to show that Algorithm would succesfully converge to a local optimum, one could use standard subspace $(\Omega^*(\gamma), \Omega^*(\gamma), \epsilon, \delta)$ for ϵ and δ small enough, though inverse polynomial, functions of the non-zero eigenvalues of A. This would take a bit of work though, and in the interest of preserving clarity of exposition, we stick with establishing the cleaner point that rank one PCA $(\gamma, \gamma, 0, 0)$.

Again, we impose the spherical consraint $c(w) = \frac{1}{2}\|x\|^2 - 1$, and $f(w) = -\frac{1}{2} w^T A w$, so that $\nabla f(w) = -Aw$, and $\nabla^2 f(w) = -A$. Hence, by the computation of the spherical projection step in claim 3.2.6, gradient descent on $\mathcal{S}^{d-1}$ is precisely the lazy power method update with step size η

$$w_{t+1} \leftarrow \frac{w_t + \eta A w_t}{\|w_t + \eta A w_t\|} \tag{3.57}$$

Note that in the limit $\eta \to \infty$, $w_{t+1} = A w_t$. Next, we can compute

$$\text{grad} f(w) = (I - w w^T) A w \quad \text{and} \quad \text{hess} f(w) = -A + w^T A w I \tag{3.58}$$

Then, $\text{grad} f(w) = 0$ if $Aw \in \text{span}(w)$, i.e., w is an eigenvector of A. In other words, the only stationary points are the eigenvectors of A.

To see that $f(w)$ is in fact a strict saddle, let $w^* \in \arg\min_{w \in \mathcal{S}^{d-1}} f(w)$, $\bar{w}$ be a non-optimal stationary point, and again γ denote the gap between the largest and second largest eigenvalues. By the above discussion, w^* and $\bar{w}$ are eigenvectors of A. Thus, $(w^*)^T A w^* - \bar{w}^T A \bar{w} \geq \gamma$, and since A is symmetric, we must have that $w^* \perp \bar{w}$; that is, $P_{\bar{w}^\perp} w^* = w^*$. Hence,

$$\begin{aligned}
(w^*)^T P_{\bar{w}^\perp}^T \text{hess}(\bar{(}w))(w^*) P_{\bar{w}^\perp} &= (w^*)^T \text{hess}(\bar{w})(w^*) && (3.59)\\
&= -(w^*)^T A w^* + (w^*)^T I w^* \cdot \bar{w}^T A \bar{w} && (3.60)\\
&= -(w^*)^T A w^* + \bar{w}^T A \bar{w} && (3.61)\\
&\leq -\gamma && (3.62)\\
& && (3.63)
\end{aligned}$$

Moreover, at an optimum w^*, we have that for any $v \in \mathcal{T}(w^*)$ that

$$v^T \text{hess}(w^*) v = -v^T A v + w^* A w^* \geq \gamma \tag{3.64}$$

Thus, from the second order KKT conditions [], we see that the only local optima are the w^* which lie in the top eigenspace of A, and are thus global optima.

3.3 Application: Orthogonal Tensor Decomposition

As our first application, we consider the problem of computing the orthogonal decomposition of a symmetric four tensor, T. That is, we assume that T has the form

$$T = \sum_{i=1}^{m} a_i^{\otimes 4} \tag{3.65}$$

for orthogonal unit vectors $a_i \in \mathcal{S}^{d-1}$, such that $\langle a_i, a_j \rangle$. Here $a_i^{\otimes 4}$ denotes the tensor whose (i,j,k,l)-th entry is $a_i a_j a_k a_l$. Like PCA, this decomposition displays invariant to sign flips and permutations of the coordinates a_i, which renders the corresponding optimimization problem

$$\min_{\{a_i \in \mathcal{S}^{d-1}\}, \langle a_i, a_j \rangle = 0} \| \sum_{i=1}^{m} a_i \otimes^4 - T \|_F \tag{3.66}$$

non-convex[3]. Empirically, it turns out that stochastic gradient descent performs quite poorly on the objective in Equation 3.66. Instead of reconstructing directly, we can try to recover the orthgonal factors iteratively performing a tnesor analogue of the power method:

$$\max_{\|u\|^2 = 1} T(u,u,u,u) \tag{3.67}$$

followed by an *deflation* step, as inAnandkumar et al. (2014). It turns out that, when T is only approximated by an orthogonal decomposition, this deflation step is numerically unstable and highly sensitive to noise or model mispecification. As a more robust alternative, we can attempt to recover all the orthogonal vectors in the decomposition simultaneous, by recovering unit vecotrs which are have the least pairwise correlated under T, namely

$$\min_{u_i \in \S^{d-1}} \sum_{i \neq j} T(u_i, u_i, u_j, u_j) \tag{3.68}$$

where we understand T as a linear operator from $(\mathbb{R}^d)^4 \to R$ via

$$T(u^{(1)}, u^{(2)}, u^{(3)}, u^{(4)}) = \sum_{i_1, i_2, i_3, i_4} T_{i_1, i_j, i_3, i_4} u_{i_1}^{(1)} u_{i_2}^{(2)} u_{i_3}^{(3)} u_{i_4}^{(4)} \tag{3.69}$$

It turns out that, when T has an exact orthogonal decomposition, then the objectives in Equation 3.67 and 3.68 are a $(\Theta(1), \gamma, \epsilon, \delta)$-strict saddle, where $\gamma, \epsilon, \delta = 1/\text{poly}(d)$, and that all local optima are in fact global optima; that is, they correspond to vectors a_i for which the decomposition in Equation 3.65 is exact. Moreover, the corresponds on the u_i are either spheres, or cartesian product of spheres, and so the arguments in Section verify that they satisfy the requisite smoothness and RLI-conditions.

[3]Here $\| \cdot \|_F$ denotes the tensor Frobenius norm, that is, the l_2 norm of the tensor viewed as a vector in $\mathbb{R}^{d^4}$

3.3.1 Experiments

1. Denerate orthogonal vectors $a_1, \ldots, a_m$ uniformly in $\mathbb{R}^d$
2. Set $T = \sum_{i=1}^m a_i^{\otimes 4}$, aso set $\tilde{T} = T + \epsilon * E$, where E is a uniform iid with entries of size $\mathcal{N}(0, \sqrt{m}/d^2)$ (this variance scaling is crucial!!!), ϵ is a small noise parameter say $\epsilon = 1/100$), and symmetrized so that $E_{i,j,k,l} = E_{\pi(i),\pi(j),\pi(k),\pi(l)}$ for any permutation (i, j, k, l). Maybe, if we have time, try a 3rd tensor which comes from a data set we want to do ICA on []
3. Run SGD on the objective in Equation 3.68

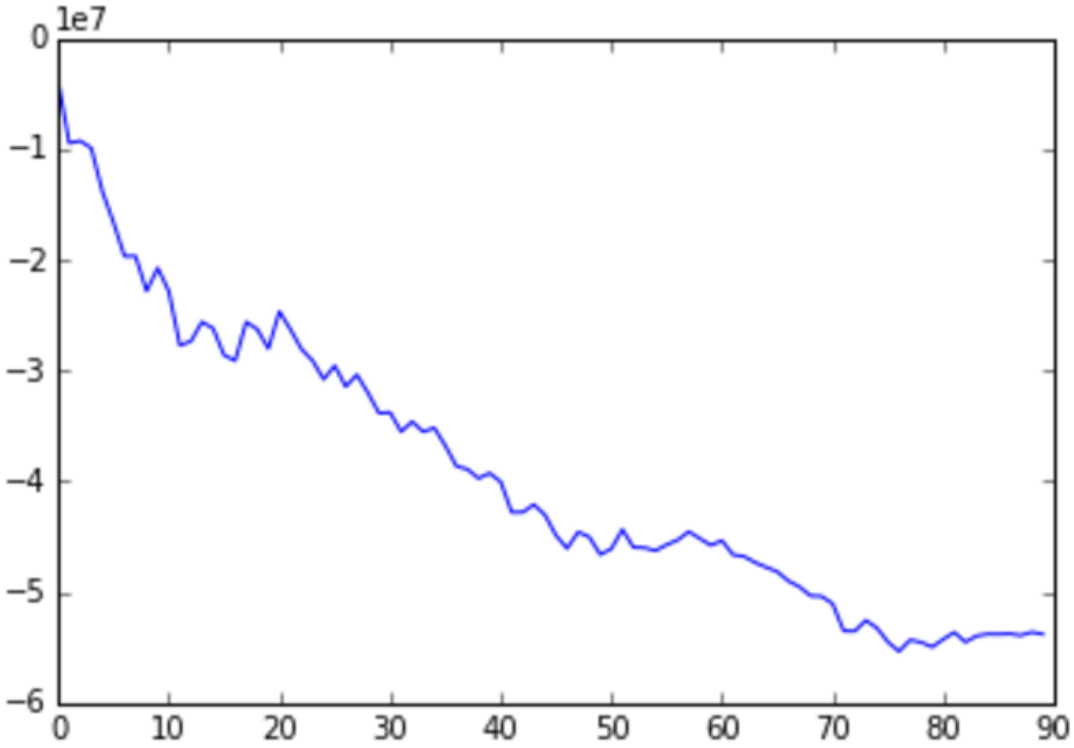


Figure 3.1: Convergence of SGD on tensor factorization objective, trying to recover canonical basis vectors

3.4 Application: Dictionary Learning

A fundamental unsupervised problem in machine learning is the sparse coding, or dictionary problem, where we given linear transformations $y^{(i)} = Ax^{(i)}$ of sparse vectors $x^{(i)}$ under a "dictionary matrix" A, and our goal is to recovery the dictionary and the samples which produced our observations. Aggregating our observations into matrix notation, we recover $Y = AX$, where $Y \in \mathbb{R}^{n \times p}$, $A \in \mathbb{R}^{n \times m}$, and $X \in \mathbb{R}^{m \times p}$ has sparse columns. Assume that

$m = n$ and A is full rank, or "complete", it follows that Y and X have the same row space. Conventionally, we assume that $x^{(i)}$ are iid samples from a sparse, subgaussian distribution: that is $x_j^{(i)} \overset{iid}{\sim} \text{BG}(\theta)$, where the parameter $\text{BG}(\theta)$ has the distribution of the product of a standard normal random variable, and a Bernoulli variable with parameter θ.

Under the generative assumptions described above, [] demonstrates that the rows of X are in fact the sparsest vectors in the common row space of X and Y, as long as $p = \Omega(n \log n)$. Hence, one might try to recover a row X, up to a scalar factor, by find the sparsest vectors in the row space of Y, namely by minimizing

$$\min_q \|q^T \hat{Y}\|_0 \quad s.t. \quad q \neq 0 \tag{3.70}$$

where $\|\cdot\|_0$ is the "0-norm", which counts the number of non-zero entries, and $\hat{Y}$ is a suitably preconditioned proxy for Y. Then here, $q^T \hat{Y}$ will correspond to a row of X. One can then recover A by solving the linearly system $AX = Y$

The 0 norm objective is non-convex and discontinuous, and difficult to optimize directly. Moreover, the constraint $q \neq 0$ introduces an indeterminacy in scaling, since if q is an optimal solution, so is αq for any $q \neq 0$. Previously, [] introduce relaxations of Equation 3.70 which replace the $\|\cdot\|_0$ with the 1-norm $\|\cdot\|_1$, and hold for sparsity of $\theta = O(\sqrt{n})$. Here, we present a novel relaxation from [] which is effective for sparsity of for any $\theta \in (0, 1/3)$. Formally, we minimize an analytic approximation to $\|q^T Y\|_1$

$$\min f(q; \hat{Y}) = \frac{1}{p} \sum_{k=1}^{p} h_\mu(q^T \hat{Y}_i) \quad s.t. \quad \|q\| = 1 \tag{3.71}$$

where $\hat{Y}$ is a (possibly preconditioned) proxy for Y, $\hat{Y}_i$ are its columns, and h_μ is an analytic proxy for the absolute value function:

$$h_\mu(z) = \mu \log \left(\frac{\exp(z/\mu) + \exp(-z/\mu)}{2} \right) := \mu \log \cosh(z/\mu) \tag{3.72}$$

Here, μ controls the "steepness" of the approximation, so that, for any $z \neq 0$, $\lim_{\mu \to 0} h_\mu(z) = |z|$.

Like PCA and the Tensor Decomposition problems described above, it turns out that the objective in Equation 3.71 is also a strict saddle, and that all of its local optima are in fact global. In the interest of brevity, we will only sketch the strict saddle property of f under a suitable parameterization of the sphere, to be described shortly. That f is a strict saddle over the manifold $\mathcal{S}^{n-1}$ requires more in depth exposition, and is far beyond the scope of this write up.

We also remark that**?**uggests optimizing over the objective in Equation 3.71 using a manifold variant of the Trust Region Algorithm [cite], described in detail in the monographAbsil et al. (2009). As the focus of this note is first order methods, inspired by SGD, we will abstain from describing their algorithm in detail, and simply remark that, like Algorithm [], the proof of its convergence also levels the strict saddle propery.

3.4.1 Dictionary Learning and the Strict Saddle

For ease of exposition, lets assume that our dictionary A is orthogonal; the general case follows from a preconditioning argument, but will unnecessarily complicate the exposition.

We will also assume that $\theta = \Omega(\sqrt{n}/\log n)$, since the case $\theta = o(\sqrt{n}/\log n)$ has been studied extensively in the pre-existing literature.

In order to get a feel for the geometry of the problem, lets assume that $A = I$ is known, and suppose that we are trying to recover the n-th row of X. In our simplified case, we have $Y = X$, so if our the parameter q we hope to recover is the vector is such that $q^T X$ is the n-th row of X; that is, q is just e_n, the n-th cannonical basis vector. Hence, it makes sense to reparameterize the sphere in terms of the equitorial directions about e_n, namely by

$$q(w) = (w, \sqrt{1 - \|w\|^2}) \quad : \quad \|w\| \leq 1 \tag{3.73}$$

Since our goal is to recover e_n (or, at least an approximation), we restrict ourselves to the open set $\Gamma = \{w : \|w\| \leq \frac{1-1/4n}{\}} = \{w : q_n(w) \geq \frac{1}{\sqrt{2n}}\}$. We remark that that the set Γ is appropriately sized, in the sense that if a starting vector q_0 is initialized uniformly on the sphere, q_0 will lie in Γ with constant probability.

Now, it turns out that, in this parametrization, our objective has the strict saddle property, and we can characterize the regions in which the Hessian has a large negative eigendirection, the gradient is large, and the function is strongly convex. Let's make this more formal. In the simplified case, note that our objective is simply

$$\min f(q; X) = \frac{1}{p} \sum_{k=1}^{p} h_\mu(q^T X_i) \quad s.t. \quad \|q\| = 1 \tag{3.74}$$

Now, if we define the composite parameterization,

$$g(w; X) = f(q(w); X) \tag{3.75}$$

then the following proposition decomstrates that g satisfies the strict saddle property:

Proposition 3.4.1. *If $A = I$, and $X_{i,j} \overset{iid}{\sim} BG(\theta)$, then as long as $\theta \in (0, 1/2)$, the steepness $\mu \leq C_1 \min\{\theta/n, n^{-5/4}\}$, a number of samples p satisfies $p \geq \frac{C}{\mu^2\theta^2} n^3 \log \frac{n}{\mu\theta}$, then the following hold with high probablity:*

$$\begin{aligned}
&\nabla^2 g(w; X) \succeq \frac{c\theta}{\mu} I && \forall w : \|w\| \leq \frac{\mu}{4\sqrt{n}} \\
&\frac{w^T \nabla g(w; X_0)}{\|w\}} \geq c\theta I && \forall w : \frac{\mu}{4\sqrt{n}} \leq \|w\| \leq \frac{1}{20\sqrt{5}} \\
&\frac{w^T \nabla^2 g(w; X) w}{\|w\|^2} \leq -c\theta && \forall w :\leq \frac{1}{20\sqrt{5}} \leq \sqrt{1 - 1/4n}
\end{aligned} \tag{3.76}$$

Moreover, $g(w; X)$ has exactly one local minimum on Γ, and it satisfies

$$\|w^* - 0\| = O\left(\min\{\frac{\mu}{\theta}\sqrt{\frac{n \log p}{p}}, \mu\}\right) \tag{3.77}$$

so that $q(w^*) \approx q(0) = e_n$. Here c, C_1, C_2 are universal numerical constants.

What is so interesting about Proposition is not simply that it establishes that the strict saddle property holds[4], but that it gives a precise description of the different regions in which different parts of the strict saddle take effect. More precisely:

1. Far away from the optimum, $\lambda_{\min}(g(w;X)) = -\Omega(\theta)$, and thus has a strict saddle
2. At a medium distance from the optimum, $g(w;X)$ has large gradients, that is $\|g(w;X)\| = \Omega(\theta)$. Note that the $\nabla g(w;X)$ has positive inner product with w, which means that g roughly increases along the direction of increasing $\|w\|$, or in other words, g roughly decreases as $w \to 0$.
3. Near the optimum, $\nabla^2 g(w;X) \succeq \frac{c\theta}{\mu}$, in other words, $g(w;X)$ is $\Omega(\theta/\mu)$-strongly convex.

Moreover, like PCA and Orthogonal Decompositions, proposition 3.4.1 establishes that the local minima are in fact global. We remark that, in order to recover the i-th row of X of an arbitrary orthogonal dictionary A, we simply consider the geometry of the function

$$g(w;X) = f(A^T P_{in} q(w); Y) = f(A^T P_{in} q(w); AX) \tag{3.78}$$

where $q(w)$ is the parameterization in Equation, P_{in} is any permutation matrix in $\mathbb{R}^n$ which swaps that i-th and n-th entries of a vector in $\mathbb{R}$, and we remark that $A^T = A^{-1}$ as A is orthogonal.

3.4.2 Experiments

Experiments:

1. Generate a dictionary $A \in \mathbb{R}^{n \times n}$ with random orthogonal columns, and sparse oberservations $x^{(1)}, \ldots, x^{(p)}$ for $p = O(n^4 \log^4 n)$, where $x_{i,j}^{(1)} \sim \mathrm{BG}(1/10)$.
2. Run SPGD on this objective to recover an approximate row q of X, measure distance between q and any other row of X, up to sign/scaling.

[4]We remark that the saddle property holds only over an open subset

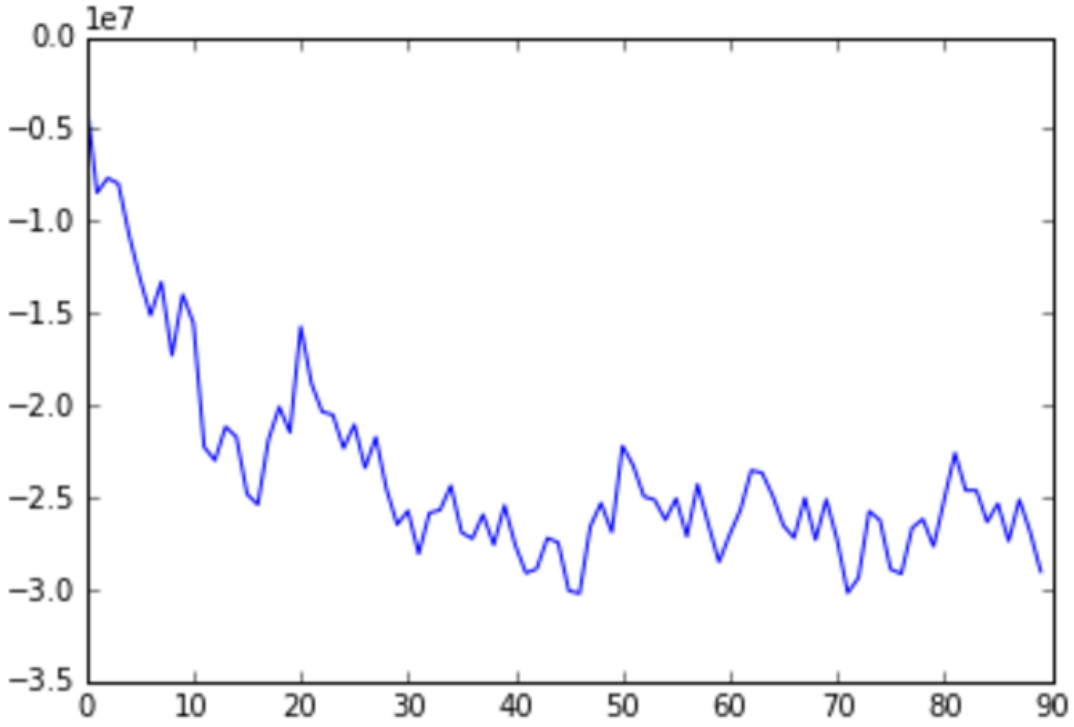


Figure 3.2: Dictionary Learning sgd convergence

Bibliography

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. Optimization algorithms on matrix manifolds. 2009.

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

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. *arXiv preprint arXiv:1503.00778*, 2015.

Giorgio Ausiello, Pierluigi Crescenzi, Giorgio Gambosi, Viggo Kann, Alberto Marchetti-Spaccamela, and Marco Protasi. Complexity and approximation: Combinatorial optimization problems and their approximability properties. 2012.

Rajendra Bhatia. Matrix analysis. 169, 2013.

Srinadh Bhojanapalli, Anastasios T. Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. *CoRR*, abs/1509.03917, 2015. URL http://arxiv.org/abs/1509.03917.

Sébastien Bubeck. Theory of convex optimization for machine learning. *arXiv preprint arXiv:1405.4980*, 2014.

Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *Information Theory, IEEE Transactions on*, 61(4):1985–2007, 2015.

Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

Patrick Culot, Georges Dive, Van Hen Nguyen, and Jean-Marie Ghuysen. A quasi-newton algorithm for first-order saddle-point location. *Theoretica Chimica Acta*, 82(3-4):189–205, 1992.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. pages 2933–2941, 2014.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. *CoRR*, abs/1503.02101, 2015. URL http://arxiv.org/abs/1503.02101.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Papdimitriou C.H. Johnson, D.S. and M. Yannakaki. How easy is local search? *Journal of computer and system sciences*, (37(1)):79–100, 1988.

Cornelius Lanczos. *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. United States Governm. Press Office, 1950.

Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

Leon Mirsky. A trace inequality of john von neumann. *Monatshefte für Mathematik*, 79(4): 303–306, 1975.

Jiquan Ngiam, Adam Coates, Ahbik Lahiri, Bobby Prochnow, Quoc V Le, and Andrew Y Ng. On optimization methods for deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 265–272, 2011.

Jorge Nocedal and Stephen Wright. Numerical optimization. 2006.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

Christopher De Sa, Christopher Re, and Kunle Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2332–2341, 2015. URL http://jmlr.org/proceedings/papers/v37/sa15.html.

Ohad Shamir. Fast stochastic algorithms for svd and pca: Convergence properties and convexity. *arXiv preprint arXiv:1507.08788*, 2015a.

Ohad Shamir. Convergence of stochastic gradient descent for pca. *arXiv preprint arXiv:1509.09002*, 2015b.