Optimization for Machine Learning 50.579

Instructor: Ioannis Panageas
.

Scribed by: Sai Ganesh Nagarajan
Rishabh Bhardwaj

Introduction to Non-convex Optimization-Gradient Descent Avoids Saddle Points.

Lecture: 4[b] & 5[a] Week: 5 & 6

1 Introduction

Before diving into defining non-convex optimization, let us have a look at linear dynamical systems, which will be helpful in understanding problems in non-convex optimization.

1.1 Linear Dynamical Systems

A linear dynamical system (LDS) can be described by the following iterative equation:

$$x_{t+1} = Ax_t \tag{1}$$

for all $t = 0, 1, 2, \ldots$. Here, x_t is the n-dimensional state vector and A is the linear operator (a $n \times n$ constant matrix).

Now, one can simplify this evolution with respect to the initial condition x_0 as follows:

$$x_{t+1} = Ax_t \implies x_t = A^t x_0 \tag{2}$$

Hence, the properties of the state vector in the limit as $t \to \infty$ is completely characterized by x_0 and the eigenvalues of A.

Definition 1.1 *Let A be a symmetric $n \times n$ matrix with eigenvalues $\lambda_1, ..., \lambda_n$. The spectral norm of A is denoted as $\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$. Moreover, it holds that $\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$.*

We start with the following claim:

Lemma 1.1 *Let A be a symmetric matrix of size $n \times n$ and assume $\|A\|_2 < 1$. Then for all $x_0 \in \mathbb{R}^n$.*

$$\lim_{t \to \infty} x_t = 0 \tag{3}$$

Proof: Since A is $n \times n$ real symmetric matrix, it will have n linearly independent eigenvectors (not neccessary n distinct eigenvalues). Hence, the eigenvectors span the whole $\mathbb{R}^n$. Let $v_1, v_2, ..., v_n$ are these eigenvectors with eigenvalues $\lambda_1, ..., \lambda_n$. We can express-

$$x_0 = \sum_{k=1}^{n} c_k v_k \tag{4}$$

i.e., a linear combination of the eigenvectors. Thus state vector after t steps can be written as-

$$x_t = A^t x_0 = \sum_{k=1}^{n} c_k \lambda_k^t v_k \tag{5}$$

Since $\|A\|_2 < 1$, largest eigenvalue of $(A^T A)$ or (A^2) (as A is symmetric) is < 1. It follows that $|\lambda_k| < 1$ for all k, that is $\lim_{t\to\infty} \lambda_k^t = 0$. Hence

$$\lim_{t\to\infty} x_t = \lim_{t\to\infty} A^t x_0 = \lim_{t\to\infty} \sum_{k=1}^{n} c_k \lambda_k^t v_k = \sum_{k=1}^{n} c_k \lim_{t\to\infty} \lambda_k^t v_k = 0 \tag{6}$$

■

Remark 1.2 *The same holds for a non-symmetric matrix.*

Let spectral radius $\rho(A) := \max\{|\lambda_1|, \ldots, |\lambda_n|\}$ *and this is bounded by* $\|A\|_2$. *The power sequence (in eq. (3)) will converge to zero if* $\rho(A) < 1$. *We can use Jordan normal form decomposition, such that* $A = VJV^{-1}$, *where* V *is non-singular and* J *is a block diagonal matrix. Then* $A^k = VJ^kV^{-1}$. *Finally, to analyze* J^k *as* k *tends to infinity, we use the fact the spectral radius of* A *is* < 1. [1]

Remark 1.3 *If* A *has some eigenvalues greater than one then the behavior of* x_t *will depend on* x_0.

Lemma 1.2 *Let* A *be a symmetric matrix of size* $n \times n$ *and assume* $v_1, ..., v_k$ *are eigenvectors with respective eigenvalues* $\lambda_1, ..., \lambda_k$ *where* $|\lambda_i| < 1$. *Assume* $x_0 \in span(v_1, ..., v_k)$. *Then*

$$\lim_{t\to\infty} x_t = 0 \tag{7}$$

Proof: The components in x_0 corresponding to the indices with eigenvalues greater than 1 is zero. Then the analysis reduces to the previous case. ■

Remark 1.4 *What if* $x_0 \perp v_j \neq 0$ *(x_0 has a component in direction of v_j), where v_j has corresponding eigenvalue* $|\lambda_j| > 1$. *This implies a non-zero coefficient of* v_j *in expressing* x_0 *in terms of* $v_1, ..., v_k$. *The trajectory of state vector* x_t *will diverge as* t *increases.*

2 Why do we care?

This section is about how to use theory of LDS in optimization. We start by considering a quadratic function.

Definition 2.1 *Let* A *be a square matrix of size* $n \times n$. *A function* f *has quadratic form if*

$$f(x) = x^T A x \tag{8}$$

[1] `https://en.wikipedia.org/wiki/Spectral_radius#Power_sequence`

Remark 2.2 *Observe* $x^TAx = x^TA^Tx = \frac{1}{2}x^T(A + A^T)x$. *This shows we can rewrite* $f(x) = \frac{1}{2}x^T(A + A^T)x$, *where* $A + A^T$ *is symmetric. This helps us in further analysis as we can formulate the problem considering* A *as a symmetric matrix i.e.* $A := \frac{1}{2}(A + A^T)$.

Gradient of f(x) with the symmetric matrix A can be written as Ax. Hence Gradient decent for quadratic functions is a linear dynamical system possessing the state transition-

$$\begin{aligned} x_{t+1} &= x_t - \epsilon A x_t \\ &= (I - \epsilon A)x_t \end{aligned} \tag{9}$$

2.1 Building intuition through Quadratic

Let's study gradient descent (GD) for quadratic in more details.

Lemma 2.1 *Let* A *be a symmetric matrix of size* $n \times n$ *and* L *be the maximum eigenvalue of* A *(in absolute value). Set* $\epsilon < \frac{1}{L}$. *Suppose* $x = 0$ *is a strict local minimum, then GD converges to it for all* x_0.

Proof: Since 0 is a strict local minimum, the quadratic form equation-(8) is always positive for any value of $x > 0$.
Claim-1: First, we observe that A is positive definite.

$$Av = \lambda v \quad (v \text{ is a non-zero eigenvector}) \tag{10}$$
$$v^TAv = \lambda v^Tv \tag{11}$$
$$\Rightarrow \lambda||v||^2 > 0 \qquad (\text{L.H.S is non-negative}) \tag{12}$$
$$\Rightarrow \quad \lambda > 0 \qquad (||v||^2 > 0) \tag{13}$$

where $||v||^2$ is length or L_2 norm of the vector. A has all eigenvalues greater than zero and thus positive definite.
Claim-2: The matrix ϵA has eigenvalues in the interval $(0, 1)$ (as $\epsilon Av = [\epsilon\lambda]v$, and $0 < \epsilon\lambda < 1$ which is an eigenvalue for ϵA).
Claim-3: $(I - \epsilon A)$ has eigenvalues in the interval $(0, 1)$. Consider v satisfies $Av = \lambda v$. $(I - \epsilon A)v = (1 - \epsilon\lambda)v$. Hence, an eigenvalue of $(I - \epsilon A)$ is $(1 - \epsilon\lambda) \in (0, 1)$.

Therefore, using claim-1,2,3 and Lemma-1.2, we can show-

$$\lim_{t\to\infty} x_t = \lim_{t\to\infty} (I - \epsilon A)^t x_0 = 0 \tag{14}$$

■

Remark 2.3 *If* A *has negative eigenvalues, it means* $x = 0$ *is a saddle point and not local minimum!*

Definition 2.4 *Critical points and Saddle points:*

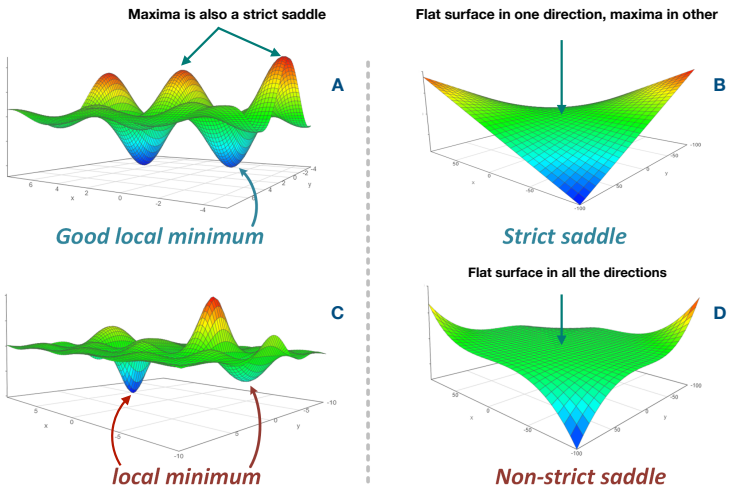

Figure 1: An illustration of possible saddle points. (A) Shows local maxima as strict saddle points; (B) Shows a strict saddle point as $\lambda_1 < 0$ and $\lambda_2 > 0$; (C) shows local minima; (D) Shows non-strict saddle point as $\lambda_1 = 0$ and $\lambda_2 = 0$.[3]

1. *A point x^* is a critical point of f if $\nabla f(x^*) = 0$.*
2. *A critical point x^* of f is a saddle point if for all neighborhoods $\mathcal{U}$ around x^* there are $y, z \in \mathcal{U}$ such that $f(z) \leq f(x^*) \leq f(y)$.*
3. *A critical point x^* of f is a strict saddle if $\lambda_{min}(\nabla^2 f(x^*) < 0)$ (minimum eigenvalue of Hessian is negative).*

Therefore if a matrix A has negative eigenvalues, then $x = 0$ is a strict saddle point. Refer to Figure 1

Note on Hessians-Hessian of an equation is symmetric, its eigenvalues are real numbers. If the Hessian at a given point has all positive eignenvalues, is is known to be positive-definite matrix (concave up). If all of them are negative, it is negative-definite matrix (concave down). If either eigenvalue is 0, then more analysis is needed. If eigenvalues are mixed of positive and negative, it is a saddle point.

Remark 2.5 *When do we converge to a saddle point? Only if x_0 belongs to the span of the eigenvalues that are less than one of $(I - \epsilon A)$.*

Claim 2.2 *(GD for Quadratic) Let A be an invertible symmetric matrix of size $n \times n$ and L be the maximum eignenvalue of A (in absolute value). Set $\epsilon < \frac{1}{L}$. Let $v_1, ..., v_k$ are eigenvectors that*

[3]The figure is edited, the original can be found at:- `https://www.offconvex.org/2018/11/07/optimization-beyond-landscape/`

correspond to eigenvalues greater than zero and $v_{k+1}, ..., v_n$ be the eigenvectors that correspond to eigenvalues smaller than zero. Then-

$$\lim_{t\to\infty} x_t = 0 \quad \textit{iff} \quad x_0 \in span(v_1, ..., v_k) \tag{15}$$

Proof: The eigenvectors that correspond to negative eigenvalues for A are eigenvectors greater than one for $(I - \epsilon A)$. Denote $E^s = span(v_1, ..., v_k)$ and $E^u = span(v_{k+1}, ..., v_n)$. GD converges to $x = 0$ only if $x_0 \in E^s$ by Lemma-2.1. ■

Remark 2.6 *How likely it is that $x_0 \in E^s$ if $k < n$? We will see that in the following section.*

3 Gradient Descent Avoids Strict Saddles

Theorem 3.1 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function, L-smooth and 0 be a strict saddle point $\epsilon < \frac{1}{L}$. For any continuous distribution D, if we sample initialization x_0 from D, GD converges to 0 with probability zero.*

Proof: Since GD is a non-linear dynamical system-

$$x_{t+1} = x_t - \epsilon \nabla f(x_t) \tag{16}$$

If we linearize it, we get-

$$x_{t+1} = (I - \epsilon \nabla^2 f(0))x_t + error(t) \tag{17}$$

with error(t) = $\mathcal{O}(\|x_t\|_2^2)$, so if we start close to zero, it should be negligible. ■

Let g: $\mathcal{X} \to \mathcal{X}$ is optimization algorithm where $x_k = g(x_{k-1}) = g^k(x_0)$, where k-fold composition is g^k. Let expression $g(x) = x - \alpha \nabla f(x)$ is gradient decent with step size α. A fixed point $x^* \in \mathcal{X}$ if $g(x^*) = x^*$. Note that all critical points of f are fixed points of gradient descent g and vice-versa.

Theorem 3.2 ***(Stable Manifold Theorem)*** *Let x^* be a fixed point of a local diffeomorphism g. Let E_s be the span of the eigenvectors of $Dg(x^*)$ corresponding to eigenvalues of magnitude less than or equal to one. Then there is an embedded disk W^{cs}_{loc} tangent to E_s at x^* called local stable center manifold. Moreover, there is a neighborhood B of x^*, such that $g(W^{cs}_{loc}) \cap B \subsetneq W^{cs}_{loc}$*

Unfolding the theorem: Before getting into the proof, we will define some terms. A function is a diffeomorphism A function is diffeomorphism if it has an inverse f^{-1} and both f and f^{-1} are smooth. A slight weaker term is local diffeomorphism if f^{-1} can be identified in small area around all $x \in \mathcal{X}$, however, one single f^{-1} can not be identified for the whole space.

Essentially, if g is diffeomorphism, we can trace back initial iterate x_0 using g. However, in local diffeomorphism, we can trace back in small region around x_k.

Let, $x_{t+1} = \phi(x_t)$.

- Locally in the neighborhood of 0, it suffices to analyze the first derivative of $\phi, D\phi$.

- All the trajectories that converge to 0 (reach a neighborhood of 0 and remain there forever, must lie in some set W^{cs}_{loc} of dimension as E^s.

(*We encourage reader to refer article*[4] for detailed explanation.)

Proof of Theorem-3.1: A sufficient condition for diffeomorphism is when the Jacobian is invertible. Jacobian of GD is just

$$I - \epsilon \nabla^2 f(x) \tag{18}$$

the eigenvalues of which are greater than zero with choice of ϵ and L-smoothness.

Now since 0 is a strict saddle, $\nabla^2 f(0)$ has a negative eigenvalue, hence E^u has dimension greater than one or equivalently E^s has dimension less than n.

Hence W^{cs}_{loc} has dimension less than n (measure zero set!).

The set of initial points x_0 so that GD converges to 0 is (assume ϕ is the update rule of GD).

$$\cup_{t=0}^{\infty} \phi^{-1}(W^{cs}_{loc}) \tag{19}$$

Claim 3.3 *(Measure zero to measure zero). Let g be a diffeomorphism and S is a measure zero set . Then $g(S)$ is also measure zero.*

Therefore each $\phi^{-t}(W^{cs}_{loc})$ is measure zero and thus the union.

■

Since the set of initial conditions that converge to 0 is of measure zero, any continuous distribution will not start from that set with probability one.

Remark 3.1 *Note that Theorem-3.1 is generally true for the unconstrained setting as we have an example for the constrained setting where GD converges to a saddle point with positive probability.*

Constrained GD-Convergence to Saddle point This example appeared in [2].

Consider the following optimization problem:

$$\min_{x,y} f(x,y) \triangleq -xy \exp(-(x^2+y^2)) + \frac{y^2}{2} \;\; s.t \;\; x + y \leq 0 \tag{20}$$

Note that $(0,0)$ is a saddle point for the above function f, since $\nabla f(0,0) = (0,0)$ and the Hessian $\nabla^2 f(0,0) = \begin{pmatrix} 0 & -1 \\ -1 & 1 \end{pmatrix}$. Suppose the initial condition starts in the box shown in Figure 2, if one does gradient descent, as the vector field of the negative gradient may eventually take the point outside the feasible set and then one has to project it onto the line $x+y=0$. With a small constant step size and continuing this projected gradient descent, we can see from Figure 2 that this leads to the saddle point $(0,0)$.

[4] `http://noahgolmant.com/avoiding-saddle-points.html` for detailed version.

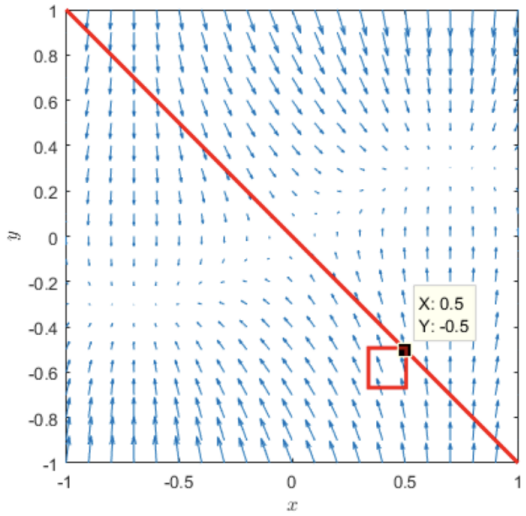

Figure 2: Negative gradient flow for $f(.)$ [2]

GD with Vanishing Step Sizes Going back to the unconstrained setting, it is known that GD even with *vanishing* step sizes avoids (strict) saddle points and we can state the following theorem due to [4].

Theorem 3.4 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function which is $L-$smooth and x^* be a strict saddle point and ϵ_t is of the order $\Omega(\frac{1}{t})$ (*vanishing*). For any continuous distribution D, if we sample the initial condition x_0 from D, GD converges to x^* with probability zero.*

The "proof" below is mainly shown for the quadratic case to build intuition. Readers can refer to [4] for the full proof.

Proof of Theorem-3.4: Let $f(x) = \frac{1}{2}x^T Ax$ be the quadratic function where the matrix A is symmetric. Now GD update boils down to:

$$x_{t+1} = x_t - \epsilon_t A x_t = (I - \epsilon_t A)x_t \tag{21}$$

By recursively applying the above update, we get the following equation:

$$x_{t+1} = \Pi_{z=0}^{t}(I - \epsilon_z A)x_0 \tag{22}$$

Since A is symmetric, we can write $A = P^T \Delta P$, where Δ is a diagonal matrix with real entries and $P^T P = I$, where P is an orthogonal matrix.

Therefore by re-writing $I - \epsilon_z A = P^T(I - \epsilon_z \Delta)P$ and then applying telescopic multiplication and using the fact that $P^T P = I$, we get the following:

$$x_{t+1} = P^T \Pi_{z=0}^{t}(I - \epsilon_z \Delta)P x_0 \tag{23}$$

Note that $\Pi_{z=0}^{t}(I - \epsilon_z \Delta) = \Delta'$ where Δ' is a diagonal matrix with the i^{th} entry being $\Pi_{z=0}^{t}(1 - \epsilon_z \lambda_i)$.

Thus, the eigenvalues as t tends to infinity can be written as: $\exp(\sum_{z=0}^{\infty} \ln(1-\epsilon_z\lambda_i)) \approx \exp(-\lambda_i \sum_{z=0}^{\infty} \epsilon_z)$, as ϵ_z is very small.

Now assume that, there exists $\lambda_i < 0$ and as long as $\sum_{z=0}^{\infty} \epsilon_z = \infty$ for gradient descent to converge to the saddle point, we require $Px_0 \perp e_i$ (i.e, the Px_0 should not have any component along e_i) and this means that these starting points which converge to the saddle point form a measure zero set in $\mathbb{R}^n$. ■

4 Gradient Descent Efficiently Avoids Strict Saddle Points

So far we have seen qualitative results about convergence of GD to it's stationary points and that unconstrained GD avoids strict saddle points. However, to obtain some convergence guarantees we require some weaker notions such as an approximate first order stationary points and approximate second order stationary points. We introduce some of these definitions below.

However, we require an additional assumption on Hessian smoothness as shown below:

Assumption 4.1 *We assume that the twice differentiable functions have ρ-Hessian Lipschitzness, formally:*

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq \rho \|x-y\|_2 \ \ \forall x, y \tag{24}$$

Definition 4.1 *Approximate first/second order stationary points:*

1. *A point x^* is an ϵ-first order stationary point (or critical point) of f if $\|\nabla f(x^*)\|_2 \leq \epsilon$.*
2. *A point x^* is an ϵ-strict saddle point of f if it is an ϵ- first order stationary point and $\lambda_{\min}(\nabla^2 f(x^*)) \leq -\sqrt{\rho\epsilon}$.*
3. *The ϵ-first order points that are not ϵ-strict saddles are ϵ- second order stationary points.*

4.1 Convergence to First Order Stationarity

First we look at convergence rates for ϵ-first order stationary points.

Theorem 4.2 *For any $\epsilon > 0$, assume that the differentiable function f is L-smooth and let $\alpha = \frac{1}{L}$. Moreover, let $f(x^*)$ be the global minimum of f. Then the gradient descent algorithm*

$$x_{t+1} = x_t - \alpha \nabla f(x_t) \tag{25}$$

will visit an ϵ-first order stationary point at least once in $T := \frac{2L(f(x_0)-f(x^*))}{\epsilon^2}$ iterations.

Proof: We begin with a property of GD on L-smooth functions. Recall that:

$$f(x - \frac{1}{L}\nabla f(x)) - f(x) \leq -\frac{1}{2L}\|\nabla f(x)\|_2^2 \tag{26}$$

Now assume that $\|\nabla f(x_t)\|_2 > \epsilon$ for $t = 1, 2, \ldots, T$ and we will try to arrive at a contradiction.

We get that:

$$f(x_T) - f(x_{T-1}) + f(x_{T-1}) - f(x_{T-2}) + \ldots + f(x_1) - f(x_0) < -\frac{\epsilon^2 T}{2L} \tag{27}$$

Then,

$$f(x^*) - f(x_0) \leq f(x_T) - f(x_0) < -\frac{\epsilon^2 T}{2L} = f(x^*) - f(x_0) \tag{28}$$

Hence, there is a contradiction. ■

We end by introducing a variant of Gradient Descent by introducing controlled perturbation in order to efficiently escape saddle points.

4.2 Perturbed Gradient Descent

Definition 4.2 (Perturbed Gradient Descent (PGD)) *Let $f : \mathbb{R}^d \to \mathbb{R}$, be a twice differentiable function, then PGD is defined by the following algorithm:*

Initialization x_0, stepsize η, perturbation radius r;

for $t = 1, 2, 3, \ldots, T$ **do**

$\quad x_{t+1} = x_t - \eta\left(\nabla f(x_t) + \xi_t\right)$ *with* $\xi_t \sim \mathcal{N}\left(0, \frac{r^2}{d} I\right)$;

end

Algorithm 1: Perturbed Gradient Descent

Now we can state the following theorem about PGD:

Theorem 4.3 *Let f be a twice differentiable L-smooth function with Hessian ρ-Lipschitz. For any $\epsilon, \delta > 0$, set $\eta = \Theta(\frac{1}{L})$ and $r = \Theta\left(\frac{\epsilon}{\log^4 \frac{d}{\delta\epsilon}}\right)$. PGD will visit an ϵ-second order stationary point at least once with probability at least $1 - \delta$ in at most $T = \mathcal{O}\left(\frac{L(f(x_0) - f(x^*))}{\epsilon^2} \log^4 \frac{d}{\rho\epsilon\delta}\right)$ iterations.*

Remark 4.3 *Firstly, not that the radius is strictly less than ϵ. Moreover, this can work for any noise (or perturbation) which is spherical, non-degenerate and has exponential tails (Gaussian-like).*

Proof Outline:

1. When the current iterate is not an ϵ-second order stationary point, it must either have a **large gradient** (case 1) or the **Hessian has a strictly negative eigenvalue** (case 2).
2. We can show that both cases yield a significant decrease in the function value in a controlled number of iterations.
3. Since the decrease cannot be more than $f(x_0) - f(x^*)$ (the global minimum is bounded), we reach a contradiction.

■

First we show a lemma which deals with case 1, when the current iterate has a large gradient.

Lemma 4.4 *Assume that f is twice differentiable, L-smooth function with $\eta = \frac{1}{L}$. Then it holds with probability $1-\delta$*

$$f(x_{t+1}) - f(x_t) \leq -\frac{\|\nabla f(x_t)\|_2^2}{2L} + \mathcal{O}\left(\frac{r^2}{d}\log\left(\frac{1}{\delta}\right)\right) \tag{29}$$

Proof: Using L-smoothness we have:

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|_2^2 \\ &= -\frac{1}{L}\nabla f(x_t)^T \nabla f(x_t) - \frac{1}{L}\xi_t^T \nabla f(x_t) + \frac{L}{2}\frac{1}{L^2}\|\nabla f(x_t) + \xi_t\|_2^2 \\ &= -\frac{1}{2L}\|\nabla f(x_t)\|_2^2 + \frac{1}{2L}\|\xi_t\|_2^2 \end{aligned} \tag{30}$$

In the final step, we know that with probability $1-\delta$, the norm squared of the Gaussian noise vector is bounded by: $\mathcal{O}\left(\frac{r^2}{d}\log\left(\frac{1}{\delta}\right)\right)$. Also by the choice of r (the noise term is strictly less than ϵ^2) and the fact the the gradient is large, we have that the whole term is of the order $\Theta(\epsilon^2)$, as the magnitude of the gradient dominates the magnitude of the noise term. ■

Finally, we end with giving the intuition behind case 2 and it can be shown that the following lemma holds:

Lemma 4.5 *Assume f is twice differentiable, L-smooth and ρ-Hessian Lipschitz. Moreover, assume that $\|\nabla f^2(x_0)\|_2 \leq \epsilon$ and also $\lambda_{\min}\left(\|\nabla f^2(x_0)\|\right) \leq -\sqrt{\rho\epsilon}$. Assume that we run PGD from x_0, then*

$$\Pr[f(x_t) - f(x_0) \leq -\frac{t'}{2}] \geq 1 - \frac{L\sqrt{d}}{\sqrt{\rho\epsilon}}\exp\left(-\Theta\left(\log^4\left(\frac{d}{\rho\epsilon}\right)\right)\right) \tag{31}$$

for $t = \frac{L}{\sqrt{\rho\epsilon}}\Theta\left(\log^4\left(\frac{d}{\rho\epsilon}\right)\right)$ *and* $t' = \frac{\epsilon^2}{\sqrt{\rho\epsilon}}\Theta\left(\log^4\left(\frac{d}{\rho\epsilon}\right)\right)$

Proof Outline: The proof looks at characterizing the region around a strict saddle point [3]. Suppose, we are not at an ϵ-second order stationary point, then we have to worry about the case when the Hessian has a strictly negative eigenvalue.

The idea is to try and characterize the volume of the region which does not lead to a significant decrease in the function value (called the **stuck region**) and show that this volume is *tiny* and a random perturbation due to PGD will take us out of the stuck region after some number of iterations with high probability. Let us suppose we are at the point x_0 and we run PGD. Now, let x be the result of the PGD. If e_1 be the component of the negative eigenvalue and suppose the remaining eigenvalues are positive then the **stuck region** is characterized when $x - x_0$ has a *small* component along e_1 (since moving along e_1 will lead to a decrease in function value). This small band around x_0 can be seen in Figure 3.

By bounding the above volume, one can show the probability that the function value decreases after t steps by $-\frac{t'}{2}$ is given in the statement.

Finally, since $f(x_0) - f(x^*)$ is bounded and t is of the order $\Theta\left(t'\epsilon^2\right)$, after $\Theta\left(\frac{f(x_0) - f(x^*)}{\epsilon^2}\right)$, we have to visit an ϵ-second order stationary point, as we reach a contradiction otherwise. ■

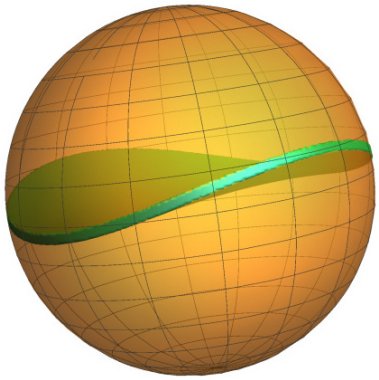

Figure 3: The green band enclosed in the sphere (which represents the perturbation ball) captures the **stuck region**. [3]

References

[1] Lee, Jason D., Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht *"First-order methods almost always avoid strict saddle points."* Mathematical programming. 2019 Jul 1;176(1-2):311-37.

[2] Nouiehed, Maher, Jason D. Lee, and Meisam Razaviyayn. *"Convergence to second-order stationarity for constrained non-convex optimization."* arXiv preprint arXiv:1810.02024 (2018).

[3] Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., Jordan, M. I. *How to escape saddle points efficiently.* In Proceedings of the 34th International Conference on Machine Learning 2019-Volume 70 (pp. 1724-1732).

[4] Panageas, Ioannis, Piliouras, Georgios and Wang, Xiao. *First-order methods almost always avoid saddle points: The case of vanishing step-sizes.* In Advances in Neural Information Processing Systems 2019 (pp. 6471-6480).