L08 – Week 4
Intro to Non-convex Optimization: GD avoids saddle points

CS 295 Optimization for Machine Learning

Ioannis Panageas

Linear Dynamical Systems

Definition (Linear Dynamical Systems). *Let A be a symmetric matrix of size $n \times n$.*

$$x_{t+1} = Ax_t.$$

One can show that

$$x_t = A^t x_0.$$

- Vector 0 is a fixed point. Does x_t converge to 0?

Linear Dynamical Systems

Definition (Linear Dynamical Systems). *Let A be a symmetric matrix of size $n \times n$.*

$$x_{t+1} = Ax_t.$$

One can show that

$$x_t = A^t x_0.$$

- Vector 0 is a fixed point. Does x_t converge to 0?

Depends on the eigenvalues of A!

Linear Dynamical Systems

Lemma (Linear Dynamical Systems)**.** *Let A be a symmetric matrix of size $n \times n$ and assume that $\|A\|_2 < 1$. Then for all $\boldsymbol{x}_0 \in \mathbb{R}^n$*

$$\lim_{t \to \infty} x_t = 0.$$

Proof. Since A is symmetric, it has eigenvalues whose eigenvectors span the whole $\mathbb{R}^n$. Let $v_1, ..., v_n$ these eigenvectors with eigenvalues $\lambda_1, ..., \lambda_n$

Linear Dynamical Systems

Lemma (Linear Dynamical Systems). *Let A be a symmetric matrix of size $n \times n$ and assume that $\|A\|_2 < 1$. Then for all $\boldsymbol{x}_0 \in \mathbb{R}^n$*

$$\lim_{t\to\infty} x_t = 0.$$

Proof. Since A is symmetric, it has eigenvalues whose eigenvectors span the whole $\mathbb{R}^n$. Let $v_1, \ldots, v_n$ these eigenvectors with eigenvalues $\lambda_1, \ldots, \lambda_n$

Express $x_0 = \sum_{k=1}^{n} c_k v_k$ (as a linear combination of the eigenvectors).

Therefore $A^t x_0 = \sum_{k=1}^{n} c_k \lambda_k^t v_k$.

Linear Dynamical Systems

Lemma (Linear Dynamical Systems)**.** *Let A be a symmetric matrix of size $n \times n$ and assume that $\|A\|_2 < 1$. Then for all $\boldsymbol{x}_0 \in \mathbb{R}^n$*

$$\lim_{t\to\infty} x_t = 0.$$

Proof. Since A is symmetric, it has eigenvalues whose eigenvectors span the whole $\mathbb{R}^n$. Let $v_1, ..., v_n$ these eigenvectors with eigenvalues $\lambda_1, ..., \lambda_n$

Express $x_0 = \sum_{k=1}^n c_k v_k$ (as a linear combination of the eigenvectors).

Therefore $A^t x_0 = \sum_{k=1}^n c_k \lambda_k^t v_k$.

Since $\|A\|_2 < 1$, it follows that $\lambda_k < 1$ for all k, that is $\lim_{t\to\infty} \lambda_k^t = 0$.

Linear Dynamical Systems

Lemma (Linear Dynamical Systems)**.** *Let A be a symmetric matrix of size $n \times n$ and assume that $\|A\|_2 < 1$. Then for all $\boldsymbol{x}_0 \in \mathbb{R}^n$*

$$\lim_{t \to \infty} x_t = 0.$$

Proof. Since A is symmetric, it has eigenvalues whose eigenvectors span the whole $\mathbb{R}^n$. Let $v_1, \ldots, v_n$ these eigenvectors with eigenvalues $\lambda_1, \ldots, \lambda_n$

Express $x_0 =$ ...ectors).

Same holds if A not symmetric (use spectral radius and Jordan decomposition)!

Therefore $A^t x_0 = \sum_{k=1}^n c_k \lambda_k^t v_k$.

Since $\|A\|_2 < 1$, it follows that $\lambda_k < 1$ for all k, that is $\lim_{t \to \infty} \lambda_k^t = 0$.

Linear Dynamical Systems

- What if A has eigenvalues greater than one as well?

The behavior of x_t depends on x_0!

Linear Dynamical Systems

- What if A has eigenvalues greater than one as well?

The behavior of x_t depends on x_0!

Lemma (Linear Dynamical Systems). *Let A be a symmetric matrix of size $n \times n$ and assume that $v_1, \ldots, v_k$ are eigenvectors with eigenvalues less than one. Assume that $x_0 \in span(v_1, \ldots, v_k)$. Then*

$$\lim_{t \to \infty} x_t = 0.$$

- Remark: Proof exactly the same as before. What if $x_0 \perp v_j \neq 0$ with v_j an eigenvector with eigenvalue greater than one?

Linear Dynamical Systems

- What if A has eigenvalues greater than one as well?

The behavior of x_t depends on x_0!

Lemma (Linear Dynamical Systems). *Let A be a symmetric matrix of size $n \times n$ and assume that $v_1, \ldots, v_k$ are eigenvectors with eigenvalues less than one. Assume that $x_0 \in span(v_1, \ldots, v_k)$. Then*

$$\lim_{t \to \infty} x_t = 0.$$

- Remark: Proof exactly the same as before. What if $x_0 \perp v_j \neq 0$ with v_j an eigenvector with eigenvalue greater than one?

Trajectory diverges!

Why do we care?

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

$$f(x) = x^T A x.$$

- Remark: We may assume that A is symmetric. Why?

Why do we care?

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

$$f(x) = x^T A x.$$

- Remark: We may assume that A is symmetric. Why?

Observe that $f(x) = \frac{1}{2}x^T A x + \frac{1}{2}x^T A^T x = \frac{1}{2}x^T (A + A^T)x.$

Why do we care?

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

$$f(x) = x^T Ax.$$

- Remark: We may assume that A is symmetric. Why?

Observe that $f(x) = \frac{1}{2}x^T Ax + \frac{1}{2}x^T A^T x = \frac{1}{2}x^T (A + A^T)x$.

$\boxed{A + A^T \text{ is symmetric!}}$

Why do we care?

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

$$f(x) = x^T A x.$$

- Remark: We may assume that A is symmetric. Why?

Observe that $f(x) = \frac{1}{2}x^T A x + \frac{1}{2}x^T A^T x = \frac{1}{2}x^T (A + A^T) x.$

$A + A^T$ is symmetric!

Fact (GD for Quadratic). *Let $f(x) = \frac{1}{2}x^T A x$. GD boils down to:*

$$x_{t+1} = x_t - \epsilon A x_t = (I - \epsilon A) x_t.$$

Why do we care?

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

$$f(x) = x^T A x.$$

- Remark: We may assume that A is symmetric. Why?

Observe that $f(x) = \frac{1}{2}x^T A x + \frac{1}{2}x^T A^T x = \frac{1}{2}x^T(A + A^T)x$.

$A + A^T$ is symmetric!

Fa

Linear Dynamical System!

:

$$x_{t+1} = x_t - \epsilon A x_t = (I - \epsilon A)x_t.$$

Building intuition through Quadratic

Lemma (GD for Quadratic). *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.*

Building intuition through Quadratic

Lemma (GD for Quadratic)**.** *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, we have that A is positive definite.

Building intuition through Quadratic

Lemma (GD for Quadratic)**.** *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, we have that A is positive definite.

- ϵA has eigenvalues in the interval $(0, 1)$.
- $\Rightarrow I - \epsilon A$ has eigenvalues in the interval $(0, 1)$.

Building intuition through Quadratic

Lemma (GD for Quadratic). *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, we have that A is positive definite.

- ϵA has eigenvalues in the interval $(0, 1)$.
- $\Rightarrow I - \epsilon A$ has eigenvalues in the interval $(0, 1)$.

$$\boxed{\text{Therefore } \lim_t x_t = \lim_t (I - \epsilon A)^t x_0 = 0.}$$

- Remark: What if A has negative eigenvalues?

Building intuition through Quadratic

Lemma (GD for Quadratic). *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, we have that A is positive definite.

- ϵA has eigenvalues in the interval $(0, 1)$.
- $\Rightarrow I - \epsilon A$ has eigenvalues in the interval $(0, 1)$.

Therefore $\lim_t x_t = \lim_t (I - \epsilon A)^t x_0 = 0$.

- Remark: What if A has negative eigenvalues?

Then $x = 0$ is not a local minimum! It is a saddle point!

Definitions

Definition (Critical and Saddle points). *We provide the following definitions:*

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

Definitions

Definition (Critical and Saddle points). *We provide the following definitions:*

- *A point x^* is a critical or first-order stationary point of f if $\nabla f(x^*) = 0$.*
- *A critical point x^* of f is a saddle point if for all neighborhoods U around x^* there are $y, z \in U$ such that $f(z) \le f(x^*) \le f(y)$.*
- *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 in the previous question, if A has negative eigenvalues, then $x = 0$ is a strict saddle point.

Building intuition through Quadratic (cont.)

- Question: But if it is a saddle point, when do we converge to it?

Building intuition through Quadratic (cont.)

- Question: But if it is a saddle point, when do we converge to it?

- Answer: Only if x_0 belongs to the span of the eigenvalues that are less than one of $I - \epsilon A$.

Claim (GD for Quadratic). *Let A be an invertible symmetric matrix of size $n \times n$ and L be the maximum eigenvalue of A (in absolute value). Set $\epsilon < \frac{1}{L}$. Let $v_1, \ldots, v_k$ are eigenvectors that correspond to eigenvalues greater than zero and $v_{k+1}, \ldots, v_n$ be the eigenvectors that correspond to eigenvalues smaller than zero. Then*

$$\lim_t x_t = 0 \textit{ iff } x_0 \in \textit{span}(v_1, \ldots, v_k).$$

Building intuition through Quadratic (cont.)

- Question: But if it is a saddle point, when do we converge to it?

- Answer: Only if x_0 belongs to the span of the eigenvalues that are less than one of $I - \epsilon A$.

Claim (GD for Quadratic). *Let A be an invertible symmetric matrix of size $n \times n$ and L be the maximum eigenvalue of A (in absolute value). Set $\epsilon < \frac{1}{L}$. Let $v_1, ..., v_k$ are eigenvectors that correspond to eigenvalues greater than zero and $v_{k+1}, ..., v_n$ be the eigenvectors that correspond to eigenvalues smaller than zero. Then*

$$\lim_t x_t = 0 \text{ iff } x_0 \in span(v_1, ..., v_k).$$

Proof. The eigenvectors that correspond to negative eigenvalues for A, are eigenvectors with eigenvalues greater than one for $I - \epsilon A$...

- Denote $E^s = \text{span}(v_1, ..., v_k)$
- Denote $E^u = \text{span}(v_{k+1}, ..., v_n)$.

Building intuition through Quadratic (cont.)

- Conclusion: GD converges to $x = 0$ only if $x_0 \in E^s$.

- But how likely it is that $x_0 \in E^s$ if $k < n$?

Building intuition through Quadratic (cont.)

- Conclusion: GD converges to $x = 0$ only if $x_0 \in E^s$.
- But how likely it is that $x_0 \in E^s$ if $k < n$?

Very unlikely!

Lemma (GD for Quadratic). *Let A be a symmetric invertible matrix of maximum eigenvalue in absolute value L such that E^s has dimension $k < n$ (i.e., $x = 0$ is a strict saddle for function $f(x) = \frac{1}{2}x^T Ax$). We set $\epsilon < 1/L$. For any continuous distribution D, if we sample initialization x_0 from D, GD converges to $x = 0$ with probability zero.*

Convergence to first order stationarity

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

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

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

Proof. Recall

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

Convergence to first order stationarity

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

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

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

Proof. Recall

$$f(x - \frac{1}{L}\nabla f(x)) - f(x) \leq -\frac{1}{2L}\left\|\nabla f(x)\right\|_2^2.$$

Assume that $\|\nabla f(x_t)\|_2 > \epsilon$ for $t = 1, ..., T$. We get that

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

Convergence to first order stationarity

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

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

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

Proof. Recall

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

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

Convergence to first order stationarity

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

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

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

Proof. Recall

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

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

Contradiction!

Gradient Descent Avoids strict saddles

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

Proof. GD is a dynamical system (but not linear).

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

Gradient Descent Avoids strict saddles

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

Proof. GD is a dynamical system (but not linear).

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

If you linearize it you get

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

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

Gradient Descent Avoids strict saddles

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

Proof. GD is a dynamical system (but not linear).

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

If you linearize it you get

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

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

Pause...

Gradient Descent Avoids strict saddles

Assume you are given a dynamical system $x_{t+1} = \phi(x_t)$.

Theorem (Stable Manifold Theorem). *Let 0 be a fixed point for the C^r local diffeomorphism $\phi : U \to E$, where U is a neighborhood of 0 in the Banach space E. Suppose that $E = E_s \oplus E_u$, where E_s is the span of the eigenvectors corresponding to eigenvalues less than or equal to 1 of $D\phi(0)$, and E_u is the span of the eigenvectors corresponding to eigenvalues greater than 1 of $D\phi(0)$. Then there exists a C^r embedded disk W^{cs}_{loc} that is tangent to E_s at 0 called the* local stable center manifold. *Moreover, there exists a neighborhood of 0, B, such that $\phi(W^{cs}_{loc}) \cap B \subset W^{cs}_{loc}$, and $\cap_{k=0}^{\infty} \phi^{-k}(B) \subset W^{cs}_{loc}$.*

Everybody please remain calm. The theorem above just says:

- Locally in the neighborhood of 0, it suffices to analyze the first derivative of ϕ, $D\phi$.
- All the trajectories that converge to 0 (reach a neighborhood of 0 and remain there forever, must lie in some set W_{loc}^{cs} of dimension as E^s.

Gradient Descent Avoids strict saddles

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

Proof cont. A sufficient condition for diffeomorphism is when the Jacobian derivative is invertible. Jacobian of GD is just

$$I - \epsilon \nabla^2 f(x)$$

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

Gradient Descent Avoids strict saddles

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

Proof cont. A sufficient condition for diffeomorphism is when the Jacobian derivative is invertible. Jacobian of GD is just

$$I - \epsilon \nabla^2 f(x)$$

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

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.

Gradient Descent Avoids strict saddles

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

Proof cont. A sufficient condition for diffeomorphism is when the Jacobian derivative is invertible. Jacobian of GD is just

$$I - \epsilon \nabla^2 f(x)$$

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

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!).

Gradient Descent Avoids strict saddles

Proof cont. So if x_t converges to 0, there exists a time T such that $x_T \in W_{loc}^{cs}$ which is a measure zero set.

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

$$\cup_{t=0}^{\infty} \phi^{-t}(W_{loc}^{cs}).$$

Gradient Descent Avoids strict saddles

Proof cont. So if x_t converges to 0, there exists a time T such that $x_T \in W^{cs}_{loc}$ which is a measure zero set.

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

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

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

Gradient Descent Avoids strict saddles

Proof cont. So if x_t converges to 0, there exists a time T such that $x_T \in W^{cs}_{loc}$ which is a measure zero set.

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

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

Claim (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.

Gradient Descent Avoids strict saddles

Proof cont. So if x_t converges to 0, there exists a time T such that $x_T \in W_{loc}^{cs}$ which is a measure zero set.

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

$$\cup_{t=0}^{\infty} \phi^{-t}(W_{loc}^{cs}).$$

Claim (**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 $\phi^{-t}(W_{loc}^{cs})$ is measure zero and thus the union.

$\boxed{\text{Since the set of initial conditions that converge to 0 is of measure zero, any continous distribution will not start from that set with probability one.}}$

Conclusion

- Introduction to Non-convex Optimization.
  - Gradient Descent avoids strict saddles!
- Next lecture we will talk about more about non-convex optimization.