

# Course Notes for EE227C (Spring 2018): Convex Optimization and Approximation

Instructor: Moritz Hardt

Email: [hardt+ee227c@berkeley.edu](mailto:hardt+ee227c@berkeley.edu)

Graduate Instructor: Max Simchowitz

Email: [msimchow+ee227c@berkeley.edu](mailto:msimchow+ee227c@berkeley.edu)

April 14, 2018

## 18 Lecture 18: Saddle points

This lecture formalizes and shows the following intuitive statement for nonconvex optimization:

**Gradient descent almost never converges to (strict) saddle points.**

The result was shown in [LSJR16]. Let's start with some definitions.

**Definition 18.1** (Stationary point). We call  $x^*$  a stationary point if the gradient vanishes at  $x^*$ , i.e.,  $\nabla f(x^*) = 0$ .

We can further classify stationary points into different categories. One important category are saddle points.

**Definition 18.2** (Saddle point). A stationary point  $x^*$  is a *saddle point* if for all  $\epsilon > 0$ , there are points  $x, y \in B(x^*; \epsilon)$  s.t.  $f(x) \leq f(x^*) \leq f(y)$ .

**Definition 18.3** (Strict saddle point). For a twice continuously differentiable function  $f \in C^2$ , a saddle point  $x^*$  is a *strict saddle point* if the Hessian at that point is not positive semidefinite, i.e.  $\lambda_{\min}(\nabla^2 f(x^*)) < 0$ , where  $\lambda_{\min}$  denotes the smallest eigenvalue.

## 18.1 Dynamical systems perspective

It'll be helpful to think of the trajectory defined by gradient descent as a dynamical system. To do so, we view each gradient descent update as an operator. For a fixed step size  $\eta$ , let

$$g(x) = x - \eta \nabla f(x)$$

so the notation for the result of iteration  $t$  from our previous discussion of gradient descent carries over as  $x_t = g^t(x_0) = g(g(\dots g(x_0)))$ , where  $g$  is applied  $t$  times on the initial point  $x_0$ . We call  $g$  the gradient map. Note that  $x^*$  is stationary iff. it is a fixed point of the gradient map i.e.  $g(x^*) = x^*$ . Also note that  $Dg(x) = I - \eta \nabla^2 f(x)$  (Jacobian of  $g$ ), a fact that will become important later. Now we formalize a notion of the set of "attractors" of  $x^*$ .

**Definition 18.4.** The global stable set of  $x^*$ , is defined as

$$W^S(x^*) = \{x \in \mathbb{R}^n : \lim_t g^t(x) = x^*\}$$

In words, this is the set of points that will eventually converge to  $x^*$ .

With this definition out of the way, we can state the main claim formally as follows.

**Theorem 18.5.** Assume  $f \in C^2$  and is  $\beta$ -smooth. Also assume that the step size  $\eta < 1/\beta$ . Then for all strict saddle points  $x^*$ , its set of attractors  $W^S(x^*)$  has Lebesgue measure 0.

**Remark 18.6.** In fact, it could be proven with additional technicalities that the Lebesgue measure of  $\bigcup_{\text{strict saddle points } x^*} W^S(x^*)$  is also 0. This is just another way to say that gradient descent almost surely converges to local minima.

**Remark 18.7.** By definition, this also holds true to any probability measure absolutely continuous w.r.t. the Lebesgue measure (e.g. any continuous probability distribution). That is,

$$\mathbb{P}(\lim_t x_t = x^*) = 0$$

However, the theorem above is only an asymptotic statement. Nonasymptotically, even with fairly natural random initialization schemes and non-pathological functions, gradient descent can be significantly slowed down by saddle points. The most recent result [DJL<sup>+</sup>17] is that gradient descent takes exponential time to escape saddle points (even though the theorem above says that they do escape eventually). We won't prove this result in this lecture.

## 18.2 The case of quadratics

Before the proof, let's go over two examples that will make the proof more intuitive:

**Example 18.8.**  $f(x) = \frac{1}{2}x^T Hx$  where  $H$  is an  $n$ -by- $n$  matrix, symmetric but not positive semidefinite. For convenience, assume 0 is not an eigenvalue of  $H$ . So 0 is the only stationary point and the only strict saddle point for this problem.

We can calculate  $g(x) = x - \eta Hx = (I - \eta H)x$  and  $g^t(x) = (I - \eta H)^t x$ . And we know that  $\lambda_i(I - \eta H) = 1 - \eta \lambda_i(H)$ , where  $\lambda_i$  for  $i = 1 \dots n$  could denote any one of the eigenvalues. So in order for  $\lim_t g^t(x) = \lim_t (1 - \eta \lambda_i(H))^t x$  to converge to 0, we just need  $\lim_t (1 - \eta \lambda_i(H))^t$  to converge to 0, that is,  $|1 - \eta \lambda_i(H)| < 1$ . This implies that

$$W^S(0) = \text{span} \left\{ u \mid Hu = \lambda u, 0 < \lambda < \frac{\eta}{2} \right\}$$

i.e. the set of eigenvectors for the positive eigenvalues smaller than  $\frac{\eta}{2}$ . Since  $\eta$  can be arbitrarily large, we just consider the larger set of eigenvectors for the positive eigenvalues. By our assumption on  $H$ , this set has dimension lower than  $n$ , thus has measure 0.

**Example 18.9.** Consider the function  $f(x, y) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$  with corresponding gradient update

$$g(x, y) = \begin{bmatrix} (1 - \eta)x \\ (1 + \eta)y - \eta y^3 \end{bmatrix},$$

and Hessian

$$\nabla^2 f(x, y) = \begin{bmatrix} 1 & 0 \\ 0 & 3y^2 - 1 \end{bmatrix}.$$

We can see that  $(0, -1)$  and  $(0, 1)$  are the local minima, and  $(0, 0)$  is the only strict saddle point. Similar to in the previous example,  $W^S(0)$  is a low-dimensional subspace.

### 18.3 The general case

We conclude this lecture with a proof of the main theorem.

*Proof of Theorem 18.5.* First define the local stable set of  $x^*$  as

$$W_\epsilon^S(x^*) = \{x \in B(x^*; \epsilon) : g^t(x) \in B(x^*; \epsilon) \forall t\}$$

Intuitively, this describes the subset of  $B(x^*; \epsilon)$  that stays in  $B(x^*; \epsilon)$  under arbitrarily many gradient maps. This establishes a notion of locality that matters for gradient descent convergence, instead of  $B(x^*; \epsilon)$  which has positive measure.

Now we state a simplified version of the stable manifold theorem without proof: For a diffeomorphism  $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ , if  $x^*$  is a fixed point of  $g$ , then for all  $\epsilon$  small enough,  $W_\epsilon^S(x^*)$  is a submanifold of dimension equal to the number of eigenvalues of the  $Dg(x^*)$  that are  $\leq 1$ . A diffeomorphism, roughly speaking, is a differentiable isomorphism. In fact, since differentiability is assumed for  $g$ , we will focus on the isomorphism.

Let  $x^*$  be a strict saddle point. Once we have proven the fact that  $g$  is a diffeomorphism (using the assumption that  $\eta < 1/\beta$ ), we can apply the stable manifold theorem since  $x^*$  is a fixed point of  $g$ . Because  $\nabla^2 f(x^*)$  must have an eigenvalue  $< 0$ ,  $Dg$  must have an eigenvalue  $> 1$ , so the dimension of  $W_\epsilon^S(x^*)$  is less than  $n$  and  $W_\epsilon^S(x^*)$  has measure 0.

If  $g^t(x)$  converges to  $x^*$ , there must exist  $T$  large enough s.t.  $g^T(x) \in W_\epsilon^S(x^*)$ . So  $W^S(x^*) \subseteq \bigcup_{t \geq 0} g^{-t}(W_\epsilon^S(x^*))$ . For each  $t$ ,  $g^t$  is in particular an isomorphism (as a composition of isomorphisms), and so is  $g^{-t}$ . Therefore  $g^{-t}(W_\epsilon^S(x^*))$  has the same cardinality as  $W_\epsilon^S(x^*)$  and has measure 0. Because the union is over a countable set, the union also has measure 0, thus its subset  $W^S(x^*)$  ends up with measure 0 and we have the desired result.

Finally we show that  $g$  is bijective to establish the isomorphism (since it is assumed to be smooth). It is injective because, assuming  $g(x) = g(y)$ , by smoothness,

$$\|x - y\| = \|g(x) + \eta \nabla f(x) - g(y) - \eta \nabla f(x)\| = \eta \|\nabla f(x) - \nabla f(y)\| \leq \eta \beta \|x - y\|$$

Because  $\eta \beta < 1$ , we must have  $\|x - y\| = 0$ . To prove that  $g$  is surjective, we construct an inverse function

$$h(y) = \operatorname{argmin}_x \frac{1}{2} \|x - y\|^2 - \eta f(x)$$

a.k.a. the proximal update. For  $\eta < 1/\beta$ ,  $h$  is strongly convex, and by the KKT condition,  $y = h(y) - \nabla f(h(y)) = g(h(y))$ . This completes the proof. ■

## References

- [DJL<sup>+</sup>17] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Póczos. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pages 1067–1077, 2017.
- [LSJR16] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *arXiv preprint arXiv:1602.04915*, 2016.