MS&E 213 / CS 269O : Chapter 1
Introduction to "Introduction to Optimization Theory"

*

By Aaron Sidford (sidford@stanford.edu)

April 29, 2017

1 What is this course about?

The central problem we consider throughout this course is as follows. We have a function $f : \mathbb{R}^n \to \mathbb{R}$, which we refer to as our *objective function*, and we wish to minimize it, i.e. we wish to solve the following.

$$\min_{x \in \mathbb{R}^n} f(x) \,.$$

This is known as an *unconstrained minimization problem* and it is one of the most well studied problems in optimization.

It is also incredibly prevalent. It is hard to think of cases where we have access to a lot of data and then do not want to optimize over that data set, finding the best thing about it in some way. Maybe we have purchase history for some website and we want to know what is the most popular or we want to fit a linear model too it. Maybe we have the map of the world and we want to know the minimum path from one point to another. Maybe we have a bunch of constraints on how we can spend our time, utilities for what we do, and want to find the best schedule. Optimization broadly concerns the problem of finding such optimal solutions and each of these can be easily modeled as such an unconstrained minimization problem.

Note, that most optimization problems we might think of can be phrased as an unconstrained minimization problems even if that is not the typical way we think of it. For example if we have a set $S \subseteq \mathbb{R}^n$, which we refer to as our *constraint set*, and wish to solve the following, *constrained minimization problem*

$$\min_{x \in S} f(x)$$

we can always define $g : \mathbb{R}^n \to \mathbb{R}$ by

$$g(x) = \begin{cases} f(x) & x \in S \\ \infty & x \notin S \end{cases}$$

and then solve the equivalent unconstrained minimization problem

$$\min_{x \in \mathbb{R}^n} g(x) \,.$$

*These notes are a work in progress. They are not necessarily a subset or superset of the in-class material and there may also be occasional *TODO* comments which demarcate material I am thinking of adding in the future. These notes will converge to a superset of the class material that is TODO-free. Your feedback is welcome and highly encouraged. If anything is unclear, you find a bug or typo, or if you would find it particularly helpful for anything to be expanded upon, please do not hesitate to post a question on the Piazza or contact me directly at sidford@stanford.edu.

Consequently, many problems can be written as unconstrained minimization problems, though this may obfuscate some salient properties of the original problem.

The focus of this class is how to solve these optimization problems *efficiently* while making *minimal assumptions* about f (and possibly S). The key question we will ask in the class is under a certain set of mild assumptions on S what algorithm should we design to solve the problem as quickly as possible. Note that there is a lot freedom in these definitions. There are many assumptions we could make on how we get access to f, what we mean by efficiency, and what we assume about f. In the remainder of this section, I will clarify the way in which we will think of optimization in this class and what we will try to show throughout the course.

1.1 How do we access f?

In order to reduce the assumptions we make about f, throughout this course we will typically assume we only have fairly restrictive access to f. For much of this class, rather than assuming we have our objective function f explicitly, we will assume we can only access f through invocation of some procedure that gives us only limited information of f at a certain point. Formally we will call this procedure an *oracle* and we will refer to an invocation of the procedure as a *query* to the oracle. When we only access f through an oracle we will say we only have *oracle access* to f.

Two standard oracle assumptions we encounter throughout the class are the following.

Definition 1 (Value Oracle)**.** A *value oracle*, also known as an *evaluation oracle* or a *0-th order oracle* for a function $f : \mathbb{R}^n \to \mathbb{R}$ is a procedure that when queried with a point $x \in \mathbb{R}^n$ outputs the value of f at x, i.e. $f(x)$.

Definition 2 (Gradient Oracle)**.** A *gradient oracle* or a *first order oracle* for a function $f : \mathbb{R}^n \to \mathbb{R}$ is a procedure that when queried with a point $x \in \mathbb{R}^n$ outputs the gradient of f at x, i.e. $\bigtriangledown f(x) \in \mathbb{R}^n$ where $[\bigtriangledown f(x)]_i = \frac{\partial}{\partial x_i} f(x)$.

Note that there are many oracles one can encounter in different situations. For example, a natural oracle in many settings is the following:

Definition 3 (Stochastic Gradient Oracle)**.** A *stochastic gradient oracle* for a function $f \in \mathbb{R}^n$ is a procedure that when given a query point $x \in \mathbb{R}^n$ outputs a random vector $e \in \mathbb{R}^n$ such that $\mathbb{E}e = \bigtriangledown f(x)$.

Note that this is the oracle that is a decent abstraction for many algorithms currently used for deep learning. It is also a reasonable model in many cases where we have an overwhelming amount of data, want to optimize something about the whole, and can only sub-sample pieces. We may encounter this oracle model more later in the course.

1.2 Why the oracle model?

This is a nice abstraction for working with many problems. In many cases exactly specifying the objective function can be quite complicated and perhaps we really only have access to f through a restricted oracle. For example, we might be trying to set prices and f is the total utility we get when we try selling items at those prices. Here, we can possibly only evaluate f by actually setting those prices and seeing how the world responds.

Moreover in some cases maybe the data may be so large or noisy that we cannot interact with the data directly. Maybe we only have stochastic or adversarial noisy access, in this case this model is natrual.

Even in cases where we have all the information about f, i.e. we are doing something simple like regression, $\min_x \frac{1}{2}\|\mathbf{A}x - b\|_2^2$ it is good to understand when having a simple procedure just based on the gradient (e.g. ability to apply $\mathbf{A}$ to vectors) and when are exploiting deeper algebraic structure of $\mathbf{A}$.

In short, oracles provide a nice way to extract the salient information about a problem we wish to use for algorithm design. They also provide a good lens to think about the computational complexity of optimization. Whereas for an explicitly given function, if optimizing it is difficult, i.e. takes a lot of time, proving this is currently typically outside the realm of what we know how to do computational complexity theory. However, if we assume oracle access to the input, in many cases it is possible to obtain provable lower bounds on how many oracle calls are required to obtain certain accuracies on optimization problems. Thus we may hope for tight bounds on optimization problems under oracle models; ultimately this helps inform us as to what an algorithm can or should be exploiting or not.

1.3 What do we want to do?

Once we have specified our oracle model we wish to design algorithms to minimize our objective function f as efficiently as possible or prove that more efficient minimization is impossible. There are two key questions we need to address here (1) what exactly do we mean by *minimize* and (2) what do we mean by *efficiency*.

There is a serious choice to be made in answering each of these questions. The issue of (1) is important as for almost all the problems we see in this course we will rarely be computing an exact minimum or minimizer to our objective function (as we will rarely know it exactly). Instead, we will be computing an algorithm that is approximately minimum or optimal in some sense. This is somewhat essential as in the oracle model it is rare to know you are at the exact minimum or optimal value, in some sense the input may just be too noisy. One common notion we will use to parameterize our distance from suboptimality is the following.

Definition 4 (ϵ-suboptimal point)**.** For $f : \mathbb{R}^n \to \mathbb{R}$ we call $x \in \mathbb{R}^n$ an ϵ-suboptimal point for some $\epsilon > 0$ if

$$f(x) - f_* \leq \epsilon \text{ where } f_* \stackrel{\text{def}}{=} \inf_{x \in \mathbb{R}^n} f(x)\,.$$

Note that this is a strong notion of suboptimality as it is global. Although we will achieve this in many cases in other cases we may only obtain local suboptimality perhaps as measured by low moments. For example, we may try to compute an ϵ-critical point.

Definition 5 (ϵ-critical point)**.** For $f : \mathbb{R}^n \to \mathbb{R}$ we call $x \in \mathbb{R}^n$ an ϵ-critical point for some $\epsilon > 0$ if $\| \bigtriangledown f(x)\|_2 \leq \epsilon$ where $\|x\|_2 \stackrel{\text{def}}{=} \sqrt{\sum_{i \in [n]} x_i^2}$.[1]

Note that a critical point might not be maximal or minimal apriori but if we have an algorithm that decreases the function value and finds a critical point, this may be a decent proxy for a local minima, that is a point where no direction makes progress if we move in that direction infinitesimally.

Next, (2), what do we mean by efficiency? The running time of our algorithms are going to have essentially two components (1) how many times did we invoke our oracle, i.e. what is the oracle complexity of our routine, and (2) what was the computational complexity of computation we did with the results of the oracle. Note that both of these are important and it is not always clear which is more critical. For example, in the case of setting prices, the oracle calls could be incredibly expensive, however as the size of data gets large, paying running times of $O(n^4)$ can be prohibitively expensive. We will attempt to minimize each and present our running time in terms of both pieces. However, our emphasis will typically be on oracle complexity.

1.4 What do the algorithms look like?

Since we just have oracle access, just have local information, most algorithms we will see are essentially greedy local methods. They run some simple iterative procedure: essentially, query the oracle, make an

[1] Note, other norms could be considered here, but we will focus on the ℓ_2 norm in the majority of the course.

improvement and repeat. Such algorithms are known as *iterative methods* and despite the simplicity of the procedure efficient iterative methods and their analysis can be quite complicated. This comes from the fact that sometimes to get the best performance somewhat complicated performance measures need to be used (and finding the best thing to do with all the information available can be complex).

This class can in fact be viewed, in part, as a introduction to the theory of iterative methods. These are powerful tools in designing fast algorithms for all sorts of problems (both continuous and discrete). They have been the workhorse in practice behind many machine learning algorithms and they have recently used to get numerous breakthrough results in the theory of computation. This is also an incredibly active area of research that this course should equip you to go into.

1.5 What do the problems look like?

So far we have summarized what the course will be about. We will introduce some broad class of objective functions f, specify an oracle model for accessing f, design an iterative method for optimizing f, and characterize the performance of the method in terms of the number of oracle calls (typically the number of iterations of the algorithm) and the total computational cost (typically the work per iteration). By doing this we will have a good sense of the computational boundaries for continuous optimization.

What we haven't specified is what exactly these classes of optimization problems will look like. In the remainder of this chapter we will consider some basic classes of f and apply this lens. This will give us a feel for what kind of assumptions are natural. We will conclude these notes with a summary of the main problem classes we will consider in the future.

2 General Unconstrained Optimization

Let us start with the most general class of unconstrained function minimization. Suppose we have an objective function $f : \mathbb{R}^n \to \mathbb{R}$ and we wish to minimize it. Furthermore, suppose we just have an evaluation oracle for f, i.e. for any $x \in \mathbb{R}^n$ we can compute $f(x)$ in time $O(\mathcal{T}_{EO})$. Lastly, suppose we are even promised that the minimizer of f lies in $[0, 1]$ and that $f(x) \in [0, 1]$ for all $x \in [0, 1]$. What is the oracle complexity?

Unfortunately this problem seems quite hard. It doesn't seem like information about any point gives you information about any other point in the domain. We can actually make this formal. First, consider the function f such that for some $z \in [0, 1]$ we have

$$f(x) = \begin{cases} 1 & \text{if } x \neq z \\ 0 & \text{if } x = z \end{cases}$$

Now, if we just have an evaluation oracle when we query points, unless we query the f at z we won't be able to tell if its the minimum. Even worse, if we just have an evaluation oracle and the oracle can be adversarial (i.e. doesn't have to precommit to the function), it could just keep picking z to be a point other than the point we picked. In other words there is no number of oracle queries we can make to compute an ϵ-suboptimal point for $\epsilon < 1$. In other words, fully general optimization, even in one dimension seems fairly hopeless. (We could try to find a critical point, but the function could easily be non-differentiable.)

3 Continuous Functions

The problem in the previous example was that the value of the function at a point told us very little information about the function. It didn't even give us any local information about the function. To get around this issue we could try assuming just a little more structure on f, we could assume that f does not

change too quickly, or more precisely that f is continuous. To get provable guarantees we could quantify this and make the following common assumption that f is continuous.

Definition 6. We say a function $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to a norm $\|\cdot\|$ if for all $x, y \in \mathbb{R}^n$ we have

$$|f(x) - f(y)| \leq L \cdot \|x - y\|$$

Recall the following definition of norm:

Definition 7. Recall that $\|x\|$ for $x \in \mathbb{R}^n$ is called a norm on $\mathbb{R}^n$ if the following hold, (1) (absolute homogeneity) $\forall \alpha \in \mathbb{R}$ and $x \in \mathbb{R}^n$ it holds $\|\alpha x\| = |\alpha| \cdot \|x\|$ (2) (triangle inequality) $\forall x, y \in \mathbb{R}^n$ it holds $\|x + y\| \leq \|x\| + \|y\|$, (3) $\|x\| = 0$ if and only $x = \vec{0}$.

Typically in this class we will take $\|\cdot\|$ to be the Euclidean norm, that is $\|x\| = \sqrt{\sum_{i \in [n]} x_i^2}$. If we don't say the norm, we will assume it to be the Euclidean norm. Occasionally we will work with other Schatten p-norms, i.e. $\|x\|_p = (\sum_{i \in [n]} |x_i|^p)^{1/p}$ with $\|x\|_\infty = \max_{i \in [n]} |x_i|$. Note that all norms are equivalent up to ... and thus if we use a norm ever to define a limit it doesn't matter, but if we use it in the definition of Lipschitz, it does matter.

So let's go back to our example. Suppose we have $f : \mathbb{R} \to \mathbb{R}$ and suppose that we know that f obtains its minimum on $[0, 1]$ and we know that $f \in [0, 1]$. Now further suppose the f is L-Lipschitz. This tells us that as we move a point by δ that the value of the function can change by at most $L|\delta|$. In other words, if we look at the lines of slope L and $-L$ through a point the value of the function is always between these two lines.

Now with this assumption, how many oracle queries do we need to find an ϵ-sub-optimal point?

Lemma 8. *If $f : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz such that $f_* = f(x_*)$ for some $x_* \in [0, 1]$ and $f(x) \in [0, 1]$ for all $x \in [0, 1]$ then for all $\epsilon \in (0, 1)$ there is an algorithm that finds an ϵ-suboptimal point with $O(\frac{L}{\epsilon})$ evaluation oracle queries.*

Proof. We query the evaluation oracle at points i/k for all $i \in [k]$ and return the point with minimum value. Note that $|x_* - \frac{i}{k}| \leq \frac{1}{k}$ for some $i \in [k]$ and therefore $|f(i/k) - f_*| \leq \frac{L}{k}$. Consequently

$$f(i/k) = f_* + f(i/k) - f_* \leq f_* + \frac{L}{k}\,.$$

Picking $k = L/\epsilon$ then yields the result. □

Now a natural question to ask is, is this tight? Can we do better? In particular, what is the right ass Below we show that in the adversarial oracle setting (i.e. where the oracle just needs to be consistent with some underlying function) we show that we cannot. The reason, is simply that we can construct an L-smooth function with these properties, where the only the only optimum is different in a very small region.

First we construct a function that is 1 everywhere except a small region where all the ϵ-suboptimal points lie.

Lemma 9. *For all $z \in \mathbb{R}$, $\epsilon, L \geq 0$ the function $f_z : \mathbb{R} \to \mathbb{R}$ defined for all $x \in \mathbb{R}$ as follows*

$$f_{z,\epsilon,L}(x) = \begin{cases} 1 & \text{if } |x - z| \geq \frac{2\epsilon}{L}\,. \\ 1 - 2\epsilon + L\,|x - z| & \text{otherwise} \end{cases}$$

is L-Lipshitz and x is ϵ-optimal if and only if $|x - z| \leq \frac{\epsilon}{L}$.

Proof. Note, that we can write this function alternatively as $f_z(x) = 1 - \epsilon + \min\{\epsilon, L|x - z|\}$. Furthermore, note that clearly as we change x by δ it can change by at most $L\delta$ since the function is either constant or changing locally at rate L by the $|x - z|$ and that when the term in the min changes the function has the same value. Finally, note that the function obtains its minimum value of $1 - 2\epsilon$ at z and then grows in value as we move monotonically away from z and it grows by ϵ only at distance $\frac{\epsilon}{L}$.

Using this we show that we can for any interval of length $\frac{8\epsilon}{L}$ we can construct two L-Lipschitz functions which both have value 1 everywhere outside the interval and have non-overlapping intervals containing the ϵ-suboptimal points. □

Lemma 10. *For any $c \in \mathbb{R}$ and $\epsilon, L \geq 0$ the functions $f_{c+\frac{2\epsilon}{L},\epsilon,L}$ and $f_{c-\frac{2\epsilon}{L},\epsilon,L}$ each have value 1 outside the interval $[c - \frac{4\epsilon}{L}, c + \frac{4\epsilon}{L}]$ and note that the sets of ϵ-suboptimal points for each function are disjoint.*

Proof. Each of the functions has value that is not 1 on an interval of length $\frac{4\epsilon}{L}$ and these intervals overlap only on the boundary at which point they have the same value. □

Using this we see that if we ever do not query a point in an interval of length $\frac{8\epsilon}{L}$ we cannot ensure that we have always found an ϵ-suboptimal point and from that we can prove our desired lower bound.

Lemma 11. *If for all functions $f : \mathbb{R} \to \mathbb{R}$ that are L-Lipschitz and with $f_* = f(x_*)$ for $x_* \in [0, 1]$ and $f(x) \in [0, 1]$ for all $x \in [0, 1]$ if an algorithm always computes an ϵ-suboptimal point with k queries to an evaluation oracle for f then $k \geq \frac{1}{8} \cdot \frac{L}{\epsilon} - 1$.*

Proof. Let $0 \leq x_1 \leq x_2 \leq ... \leq x_k \leq 1$ be the k' points queried by the algorithm (not necessarily in order) that lied between 0 and 1. Note that there are $k' + 1$ intervals between the x_i, 0, and 1 and therefore at least one of these intervals has length $1/(k' + 1)$ which is at least $1/(k + 1)$. Consequently, if the oracle had answered 1 everywhere the algorithm would not have been able to identify the ϵ-optimal region if $1/(k + 1) < \frac{8\epsilon}{L}$ consequently $1/(k + 1) \geq \frac{8\epsilon}{L}$ and the result follows. □

A natural question to ask then is, what happens if we are in higher dimensions?

Lemma 12. *If $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz such that $f_* = f(x_*)$ for some $x_* \in [0, 1]^n$ and $f(x) \in [0, 1]$ for all $x \in [0, 1]^n$ then for all $\epsilon \in (0, 1)$ there is an algorithm that finds an ϵ-suboptimal point with $O((\frac{L}{\epsilon})^n)$ evaluation oracle queries.*

Proof. We query the evaluation oracle at points $x \in \mathbb{R}^n$ for all vectors where $x_i = i/k$ for $i \in [k]$ and return the point with the minimum value. Note that there are k^n such points. Note that for all $i \in [n]$ it is the case that $|x_*(i) - \frac{j}{k}| \leq \frac{1}{k}$ for some $j \in [k]$ and therefore for some point q that was queried we have that $\|q - x_*\|_\infty \leq \frac{1}{k}$ and therefore

$$f(q) - f_* = |f(q) - f(x_*)| \leq L\|q - x_*\|_\infty \leq \frac{L}{k}.$$

Consequently, $f(q) \leq f_* + \frac{L}{k}$ and setting $k = L/\epsilon$ then yields the result. □

Using similar tricks as before we can show that an algorithm need $(\Omega(L/\epsilon))^n$ queries to solve this problem for some ϵ. Consequently, while global minimization for Lipschitz functions is doable, it comes at an exponential cost in dimension and we only get a weak, i.e. polynomial dependence on the error ϵ.

4 What Else?

What will we do in the rest of the course? Now that we understand the game a bit we can better explain the structure of the course. In the next few lectures we will restrict our attention to the minimization of *smooth functions* that is functions where the gradient doesn't change to quickly, i.e. the gradient is Lipschitz. We will show that under this assumption we may not be able to compute global optimum efficiently, but we can compute critical points fairly sufficiently.

With the structure of smooth functions in hand we will add the assumption that our function is convex, i.e. that $f(\alpha x + (1-\alpha)y) \leq \alpha \cdot f(x) + (1-\alpha) \cdot f(y)$, that is the function lies underneath the curve between any two points. These assumptions will be enough that we can provably find global optimum of our function. We will consider different algorithms that better exploit the exact convexity and smoothness assumptions we make. We will also discuss extensions of this to other norms and other oracle models which may be useful in various settings.

Next, we study the structure of convex functions that are continuous, i.e. Lipshitz, but not necessarily smooth. Here the structure is a little more complicated and we will consider two classes of algorithms. First we will consider algorithms with a great dependence on the desired ϵ-accuracy but a poor (i.e. polynomial) dependence on dimension, these are known as cutting plane method and they underly the theory of optimization and are critical for many of the best guarantees we have for solving both combinatorial and continuous problems. Then we will consider algorithms with a worse dependence on ϵ but a better dependence on dimension, the algorithms will be called mirror descent or subgradient descent and is used in various online, streaming, stochastic, and practical settings to get good performance.

With a firm understanding of smooth and non-smooth convex optimization we will cover some more advanced topics in iterative methods like acceleration and variance reduction.

Finally, we will conclude the course by studying higher-order methods, that is iterative methods which compute not just the gradient but the Hessian of a function, i.e. Newton's method. Moreover, we will show how to use such methods to outperform cutting plane methods for broad classes of more structured optimization. In particular, we will introduce interior point methods which provide the fastest known algorithms for problems like linear programming and semidefinite programming.

References