Stochastic Convex Optimization **Fall 2021**

Lecture 1: Introduction

Lecturer: Roi Livni *Scribe:*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1.1 Learning

Learning is often described generally as the following problem: Given a data $z \sim D$ sampled from some unknown distribution D, choose a parameter w (or a *model*) that minimizes the *true risk* function:

$$F(w) = \mathop{\mathbb{E}}_{z\sim D}[f(w,z)].$$

For example, a classical problem in learning is *binary prediction.* In binary prediction we assume that $z = (x, y)$ is a tuple of an element x coming from some domain (usually a vector in $\mathbb{R}^n$) and a boolean label $y = \{0, 1\}$.

The model w comes from a family of functions $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$, namely every $w \in \mathcal{H}$ is a function such that $w(x) \in \{0, 1\}$.

Finally, the loss function f measures the *prediction error* of w and equals:

$$f(w,z) = \begin{cases} 0 & w(x) = y \\ 1 & w(x) \neq y \end{cases}.$$

The problem of binary classification is a classical problem in learning that is extensively studied (and is also the main focus of the course Introduction to Computational Learning Theory).

Throughout the course we will see other learning problems that fall into the setting described above.

The course here focuses on the special setting of learning where we add one further assumption on the function f, and we assume f is a *convex* function.

1.2 What are convex learning problems?

A function $f : \mathcal{W} \to \mathbb{R}$ over a domain $\mathcal{W} \subseteq \mathbb{R}^d$ is called convex if for every $0 \leq \lambda \leq 1$ and $x, y \in \mathcal{W}$:

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$$

If $-f$ is a convex function, then f is called concave.

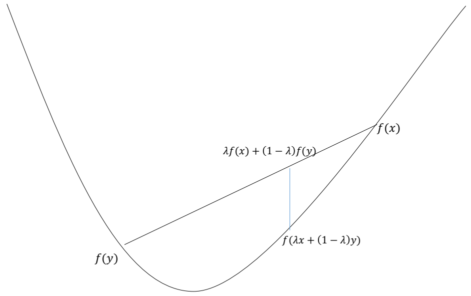

$f(x)$
$\lambda f(x) + (1-\lambda)f(y)$
$f(\lambda x + (1-\lambda)y)$
$f(y)$

Figure 1.1: Convex Function: illustration

Example 1.1. *The following functions are convex*

1. ***Linear:*** *The function $f(w) = w^\top x$ is convex (and also concave).*
2. *ℓ_2 **norm squared:*** *The function $\|w\|^2 = \sum_{i=1}^n w(i)^2$ is a convex function.*
3. *ℓ_1 **norm:*** *The function $\|w\|_1 = \sum |w(i)|$ is known to be convex, in particular the absolute value function is convex.*
4. *ℓ_p **norm:*** *The function $\|w\|_p = \sqrt[p]{\sum |w(i)|^p}$.*
5. ***Logistic loss:*** *The function $f_\alpha(x) = \ln(1 + \exp(\alpha x))$ is convex for any α.*

Using the following facts, one can construct more complex convex functions:

Fact 1.1.

If f is convex and $\alpha \geq 0$ then $\alpha \cdot f$ is convex.

If f, g are convex then $\max(f, g)$ *is convex.*

If f is convex and A is a linear operator then $f(A \cdot x)$ is also convex.

Also, if f, g are convex then so is $f + g$, in particular any function of the form

$$\hat{F}_m = \frac{1}{m} \sum_{i=1}^{m} f_i(w),$$

where f_i are convex, is convex.

More generally, given a distribution D over convex functions we have that the following function is convex

$$F(w) = \mathop{\mathbb{E}}_{f \sim D}[f(w)].$$

These can all be easily verified from the definition.

Exercise 1.1. *Prove fact 1.1.*

Another notion that is usefull is *a convex set* defined as follows

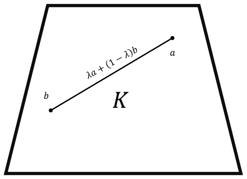
Definition 1.1. *A set $K \subseteq \mathbb{R}^n$ is called convex if for every $x, y \in K$ and $0 \leq \lambda \leq 1$:*

$$\lambda x + (1 - \lambda) y \in K.$$

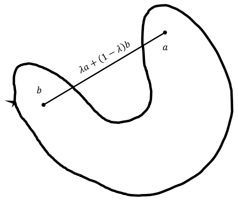
Example 1.2. *The following are convex sets*

1. *The positive cone: $\{x : x(i) \geq 0, \quad i = 1, \ldots, n\} \subseteq \mathbb{R}^n$ is convex.*
2. *The Euclidean ball: $\{x : \|x\| \leq 1\}$ is a convex set.*
3. *The unit–cube $\{x : \max\{x_i\} \leq 1\}$.*
4. *In general, if f is convex then the empigraph set $\{x : f(x) \leq a\}$ is a convex set.*

Exercise 1.2. *Prove that all examples in example 1.2 are indeed convex.*



(a) A convex set



(b) Not a convex set

A convex problem is then any problem of the form:

$$\begin{aligned} \text{minimize} \quad & f(w) \\ \text{subject to} \quad & w \in \mathcal{W} \end{aligned}$$

is called a convex program if f is convex and $\mathcal{W}$ is a convex set.

Equivalently, a concave program is a problem of the form

$$\begin{aligned} \text{maximize} \quad & g(w) \\ \text{subject to} \quad & w \in \mathcal{W} \end{aligned}$$

when g is concave and $\mathcal{W}$ is convex.

1.3 Examples of convex problems in learning

1.3.1 Linear Regression

Arguably, the simplest example of a convex learning problem is known as *linear regression.* In linear regression, we assume tuples (x, y) are drawn from some unknown distribution D, where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}$.

Our task is to find some *linear regressor*, namely we want to a vector w such that

$$w \cdot x \approx y.$$

To identify w one often wishes to solve the following *least square problem:*

$$\text{minimize,} \quad \mathop{\mathbb{E}}_{(x,y)\sim D}[(w\cdot x - y)^2] \tag{1.1}$$

Exercise 1.3. *Suppose that (x, y) are drawn from a distribution D, where for some $w^\star$:*

$$y = w^\star \cdot x + \xi,$$

where ξ is a random variable, independent of x and $\mathbb{E}[\xi] = 0$. Show that $w^\star$ is the minimizer of eq. (1.1). Is $w^\star$ the unique *minimizer?*

In general, trying to minimize the true risk function is infeasible as we don't know the distribution D. In practice, then, we often turn to minimize the *empirical risk* which is defined by sampling i.i.d examples from the distribution D: $\{(x_i, y_i)\}_{i=1}^m$, and define:

$$\hat{F}_m(w) = \frac{1}{m}\sum_{i=1}^m (w\cdot x_i - y_i)^2.$$

1.3.2 Logistic Regression

In the setting of classification, we observe a set of examples $\{(x_i, y_i)\} \subseteq \mathbb{R}^n \times \{-1, 1\}$, and we want to learn a predictor $f_w(x) = w \cdot x$ so that $\text{sign}(w \cdot x_i) = y_i$. A popular method to compute w is to consider as a loss $f(w, z_i) = \log(1 + \exp(-y_i x_i \cdot w))$. Again considering the empirical risk, this leads to the optimization problem:

$$\text{minimize,} \ \frac{1}{m}\sum_{i=1}^m \log(1 + \exp(-y_i x_i \cdot w))$$

The intuition is that when $w \cdot x_i$ agrees with the sign of y_i, as we take w to larger, the loss vanishes to zero, but when $w \cdot x_i$ disagree with the sign of y_i the loss increases. Thus, solving this problem encourages a solution w that correctly classifies y given x.

1.3.3 Support Vector Machines

Another, similar convex problem known as support vector machine (SVM) is obtained by taking

$$f(w, z_i) = \max(0, 1 - y_i w \cdot x_i).$$

The above loss is also called *hinge loss.* In SVM we consider then the problem

$$\text{minimize} \quad \frac{1}{m}\sum_{i=1}^{m} \max(0, 1 - y_i w \cdot x_i) + \frac{\lambda}{2}\|w\|_2^2$$

Note the second term that penalizes the *norm* of the solution. This term is referred to as *regularization.* And the above objective is comprised of two terms: *the empirical risk* (which we have discussed so far) and a *regularization penalty* that intuitively penalizes the solution for being of too large norm.

Understanding the role of regularization and its importance for learning will be another important element in this course.

To motivate regularization, we look at this problem of classification.

In classification we would often search for a classifier w such that

$$\operatorname{sgn}(w \cdot x) = y.$$

(then our model is given by $\operatorname{sgn}(w \cdot x)$). Our task, then, is to find a classifier w such that

$$y(w \cdot x) > 0,$$

namely the sign of $w \cdot x$ predicts y. Our optimization objective is comprised only of finite data and we are concerned that a solution w that predicts well on the data, will not exhibit good performance of the true risk (this problem is known as *overfitting*).

Therefore, intuitively, one solution to this problem is to try and find not just *some predictor* but a *well-behaved* predictor. This is often referred to as *inductive-bias* of the algorithm, where the algorithm chooses a certain structured solution from all those who optimize a certain empirical risk problem.

For example, we might want to find a predictor where the data is *far-away* from the *decision boundary.* This task can be formalized as follows (see fig. 6.1, for illustration)

$$\text{minimize } \|w\|^2$$
$$\text{s.t. } y_i w \cdot x_i \geq 1 \text{ }, \forall i = 1, \ldots, m$$

For some small enough λ this problem is equivalent to the objective

$$\text{minimize} \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m} \max(0, 1 - y_i w \cdot x_i)$$

Indeed, when λ is sufficiently small, the solution will obtain zero loss on the second term (hence $yw \cdot x \geq 1$), and will minimize the norm of $\|w\|^2$ from all the solutions who obtain loss zero on the second term.

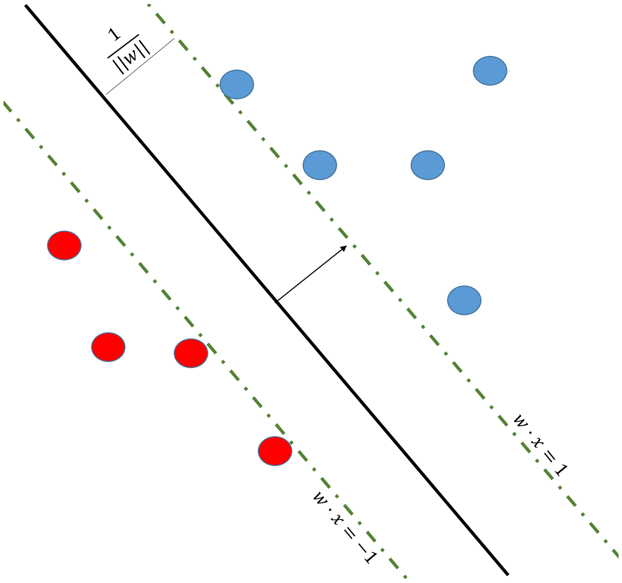


Figure 1.3: illustration of the SVM solution: The solid line represents the decision boundary of a linear classifier. This is the set where $\{x : w \cdot x = 0\}$, comprised of the vectors of the form: $\alpha \cdot w^{\perp}$. The dashed lines represent the level lines where $\{x : w \cdot x = 1\}$ and $\{x : w \cdot x = -1\}$ respectively, the area between is termed *margin.* The norm of w controls the width of the margin. In SVM we aim to classify the data so that this margin will be as large as possible.

1.4 Scope of this course

As discussed this course is concerned with convex learning problems. Fundamental problems that we will care about will be:

- How do we optimize convex optimization problems of different sorts?
- What is the relation between the empirical risk (which can often be minimized) and the true risk (which we often don't know due to lack of knowledge of the distribution D)?
- What is the role of regularization?
- What is the importance of the optimization algorithm, and its inductive bias? for example, does a large margin solution better than an arbitrary solution?

1.4.1 Why convexity?

As discussed, there are many important examples of convex learning problems. Often solved in practice.

But in general, and in many cases, many of the problems we aim to solve in learning are inherently non-convex. Nevertheless, the setting of convex learning turns out to be a very rich and important research framework that exhibits some highly non-trivial and important phenomena.

The key developments in this course will help us provide tools to prove generalization properties for different optimization algorithms. And in turn, to analyze the role of the algorithms, and their inductive bias.

These developments will help us to explore the relation of different notions such as stability, regularization and uniform convergence to generalization, with an attempt to figure out what is the underlying principle that induces generalization in SCO, which will lead to discussions on the implications to learning theory beyond convex functions.

References