ECE 901: Large-scale Machine Learning and Optimization **Spring 2018**

Lecture 3 — 01/30

Lecturer: Dimitris Papailiopoulos *Scribe: Sourish Sinha, Christopher Vandamme*

Note: These lecture notes are still rough, and have only have been mildly proofread.

3.1 From Statistical Bounds to Optimization

The main topics of this section are computational aspects of Empirical Risk Minimization (ERM).

3.1.1 Concentration Bounds

We start by assuming that the optimal risk $\hat{R}[\cdot]$ of a hypothesis class , h, is bounded by the some value value ϵ as

$$\hat{R}[h] < R[h] + \epsilon, \forall h \in H \tag{3.1}$$

with probability $1 - \delta$, where H is the collection of all possible hypothesis classes. Emperical Risk Minimization is aimed at finding the hypothesis class that minimizes the risk as

$$\hat{h}^* = argmin_{h \in H} \hat{R}[h] \tag{3.2}$$

i.e. the optimal hypothesis class. What we really want is the "true" performance of $\hat{h}^*$. The true performance is represented by the expectation of the loss function as

$$R[\hat{h}^*] = E[l(\hat{h}^*(x), y)] \tag{3.3}$$

If we enforce the concentration, $\forall h \in H$, then we can obtain

$$\begin{aligned} R[\hat{h}^*] &= R[\hat{h}^*] + (\hat{R}[\hat{h}^*] - R[\hat{h}^*]) \\ &= \hat{R}[\hat{h}^*] + (R[\hat{h}^*] - \hat{R}[\hat{h}^*]) \\ &\leq \hat{R}[\hat{h}^*] + \epsilon \end{aligned} \tag{3.4}$$

The last part of Eq (3.4) is with probability $1-\delta$. We can show that if the ER concentrates then

$$R[\hat{h}^*] \leq \hat{R}[\hat{h}^*] + \epsilon \tag{3.5}$$

Furthermore, we can relate $R[\hat{h}^*]$ to the best predictor in H, starting with

$$\hat{h}^* = argmin_{h \in H} R[h] \tag{3.6}$$

We can show that

$$\begin{aligned} R[\hat{h}^*] &= \hat{R}[\hat{h}^*] + \epsilon \\ &\leq \hat{R}[h^*] + \epsilon \\ &\leq R[h^*] + [\hat{R}[h^*] - R[h^*]] + \epsilon \\ &\leq \hat{R}[\hat{h}^*] + 2\epsilon \end{aligned} \tag{3.7}$$

In conclusion, assuming concentration, we can then argue about the best possible predictor using the performance of the ERM. The previous analysis provides a brief overview of why ERM in general is a good approach. The next section discusses some forms of ERM found in different applications.

3.1.2 ERM Examples

This section discusses some common structural forms or Empirical Risk Minimization.

Regression

1. Linear Regression
   Linear regression has the following structure

$$min||Xw - y||^2 = min\frac{1}{2}\sum_{i=1}^{n}(y_i - wx_i)^2 + \lambda R(w) \tag{3.8}$$

   where $R(w)$ can have the form of $||w||_2^2$ relating to penalization of the 2-norm of weight vectors. This is termed ridge regression in machine learning, known as Tikhonov Regularization in many other fields. This relates to a Gaussian prior placed onto the weights. This is often implemented in order to restrain the size of the weight vector. An alternative regularization is to restrain the 1-norm of the weight vector having form of $||w||_1$ which is termed Least Absolute Selection and Shrinkage Operator (LASSO).

2. Nonlinear Regression (e.g. Neural Network)
   Nonlinear regression is the natural extension of Linear Regression in which the cost

function evaluation does not have a close form linear algebra solution. The general form of a nonlinear regression problem with regularization is

$$min\frac{1}{n}\sum_{i=1}^{n}(y_i - h(x_i, w))^2 + \lambda R(w) \tag{3.9}$$

Common examples of nonlinear regression are nonlinear least squares and neural networks.

Classification

1. Binary
   The binary classification with a logistic loss function can be represented as

$$min\frac{1}{n}\sum_{i=1}^{n} log(1 + e^{-y_i x_i^T w}) \tag{3.10}$$

   The binary classification with a 0-1 function can be represented as

$$min\frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i x_i^T w) \tag{3.11}$$

2. Multi-Class (for 1-sample)
   The mulit-class classification with cross-entropy loss is represented as

$$min\frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i x_i^T w) \tag{3.12}$$

3.1.3 ERM Optimization Task

The ERM optimization problem is tasked within finding the parameters that minimizes the sum of loss functions with the general form

$$min_w = \frac{1}{n}\sum_{i=1}^{n} l_i + \lambda R(w) \tag{3.13}$$

The two major questions associated with this problem are; is it solvable and if so how fast can it be solved. Knowing the form of the loss function, type of regularization function and underlying structure of the problem can help establish information about the "solvability" and "scalability" of the optimization task at hand. In general ERM is NP-Hard which asserts that polynomial time algorithm is unlikely.

3.2 Families of functions:

Let's try and put some structure to this:

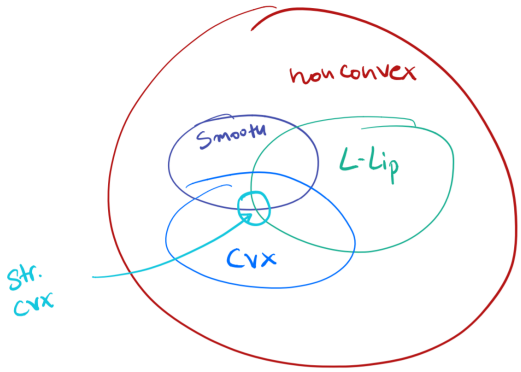

Figure 3.1: Illustrative overlap of the different families of functions, showing nonconvex, smooth, L-Lipschitz and strongly convex functions

3.2.1 Convex functions:

Convex functions operating on vectors take a form wherein they are upper bounded by some convex combination of those vectors. The following is a general way of expressing a convex function $\mathbf{f}$:

$$\boxed{\mathbf{f}(\mathbf{a}\vec{\mathbf{x}} + (\mathbf{1}-\mathbf{a})\vec{\mathbf{y}}) \leq \mathbf{a}\mathbf{f}(\vec{\mathbf{x}}) + (\mathbf{1}-\mathbf{a})\mathbf{f}(\vec{\mathbf{y}})} \tag{3.14}$$

wherein $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$ are 2 vectors, and $\mathbf{a} \in [0, 1]$. This is clearly illustrated when we consider the plot of the convex function $\mathbf{f}$ below:

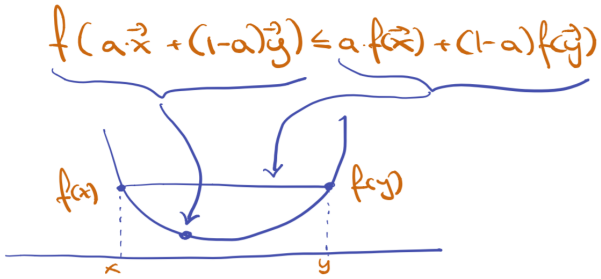

Figure 3.2: Graph of the convex function $\mathbf{f}$ evaluated at $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$

Note that any single point on the function $\mathbf{f}$, given by $\mathbf{f}(\mathbf{a}\overrightarrow{\mathbf{x}} + (\mathbf{1}-\mathbf{a})\overrightarrow{\mathbf{y}})$ always lies below the line described by $\mathbf{af}(\overrightarrow{\mathbf{x}}) + (\mathbf{1}-\mathbf{a})\mathbf{f}(\overrightarrow{\mathbf{y}})$, which gives us the inequality referenced in equation 3.14.

This fact of convexity makes our lives easy. For instance, in the case of convex functions $\min_{\mathbf{x}} \mathbf{f}(\mathbf{x})$ is always solvable in polynomial time.

Note that an important property of convex functions is that every local minimum is always equivalent to the global minimum!

3.2.2 *L*-Lipschitz functions:

L-Lipschitz functions are those that don't 'change fast'. In other words, a given function $\mathbf{f}$ is *L*-Lipschitz if the following holds:

$$\boxed{|\mathbf{f}(\overrightarrow{\mathbf{x}}) - \mathbf{f}(\overrightarrow{\mathbf{y}})| \leq L \cdot ||\overrightarrow{\mathbf{x}} - \overrightarrow{\mathbf{y}}|| \quad \forall \ \overrightarrow{\mathbf{x}}, \overrightarrow{\mathbf{y}}} \tag{3.15}$$

wherein L is the Lipschitz constant. This indicates that the difference in value obtained by evaluating the function at 2 different points $\overrightarrow{\mathbf{x}}$ and $\overrightarrow{\mathbf{y}}$ is upper bounded by the distance between the points. So for points that are closely spaced, the functional value does not change too much between them.

3.2.3 β-smooth functions:

β-smooth functions are functions with gradients that don't change fast. In other words, a given function $\mathbf{f}$ is β-smooth if the following holds:

$$\boxed{||\nabla\mathbf{f}(\overrightarrow{\mathbf{x}}) - \nabla\mathbf{f}(\overrightarrow{\mathbf{y}})|| \leq \beta \cdot ||\overrightarrow{\mathbf{x}} - \overrightarrow{\mathbf{y}}|| \quad \forall \ \overrightarrow{\mathbf{x}}, \overrightarrow{\mathbf{y}}} \tag{3.16}$$

This is somewhat analogous to the previous Lipschitz case. Here, the difference in the slopes of the function evaluated at 2 different points $\overrightarrow{\mathbf{x}}$ and $\overrightarrow{\mathbf{y}}$ is upper bounded by the distance between the points. Again, for points that are closely spaced, there is not much variation in the gradients.

3.2.4 Strongly convex functions:

These are the best kind of functions. As can be seen from figure 3.1, strongly convex functions lie at the intersection of smooth, *L*-Lipschitz and convex functions, meaning they have the properties of all three. In general, a strongly convex function $\mathbf{f}$ may satisfy the following:

$$\boxed{\mathbf{f}(\overrightarrow{\mathbf{x}}) - \mathbf{f}(\overrightarrow{\mathbf{y}}) \leq \langle\nabla\mathbf{f}(\overrightarrow{\mathbf{x}}), \mathbf{x}-\mathbf{y}\rangle - \frac{\lambda}{\mathbf{2}}||\mathbf{x}-\mathbf{y}||^{\mathbf{2}}} \tag{3.17}$$

3.3 Examples of different functions:

3.3.1 Convex functions:

Convex functions include $||\mathbf{x}||^2$, $||\mathbf{x}||$, $\log(1+e^x)$ and $\max\{0, 1-x\}$.

In general, if a function $\mathbf{g}(\cdot)$ is convex, then $\mathbf{g}(\mathbf{w^T x}+\mathbf{b})$ is also convex. For instance, $\log(1+\exp(-y\langle \mathbf{w}, \mathbf{x}\rangle))$, $\mathbf{w^T x}-\mathbf{b^2}$ etc.

If $\mathbf{f_i}(\mathbf{x})$ are convex, then functions of the form $\sum_{\mathbf{i}} \mathbf{w_i f_i}$ and $\max_{\mathbf{i}} \mathbf{f_i}(\mathbf{x})$ are also convex.

3.3.2 *L*-Lipschitz functions:

The functions $|\mathbf{x}|$ and $\mathbf{f}(\mathbf{x}) = \log(1+e^x)$ are 1-Lipschitz functions.

Note that $\mathbf{x^2}$ is not Lipschitz, unless $|\mathbf{x}| \leq \rho$, in which case it is ρ-Lipschitz.

The function $\mathbf{f}(\mathbf{w}) = \mathbf{w^T x}+\mathbf{b}$ is $||\mathbf{x}||$-Lipschitz.

If a function is of the form $\mathbf{f}(\mathbf{x}) = \mathbf{g_1}(\mathbf{g_2}(\mathbf{x}))$, we may say the following: if $\mathbf{g_1}$ is L_1-Lipschitz, and $\mathbf{g_2}$ is L_2-Lipschitz, then this implies $\mathbf{f}$ is $L_1 L_2$-Lipschitz.

The function $\mathbf{g}(\mathbf{w^T x}+\mathbf{b})$ is $||\mathbf{x}||L_g$-Lipschitz, where L_g may be thought of as the Lipschitz constant of the activation function.

If $||\nabla \mathbf{f}(\mathbf{w})|| \leq L$, then this implies $\mathbf{f}$ is L-Lipschitz.

3.3.3 β-smooth functions:

The function $|\mathbf{x}|^2$ is 2-smooth.

The function $\log(1+e^x)$ is $\frac{1}{4}$-smooth.

If a function $\mathbf{g}$ is β-smooth, then $\mathbf{f}(\mathbf{w}) = \mathbf{g}(\mathbf{w^T x}+\mathbf{b})$ is $\beta_{\mathbf{g}}||\mathbf{x}||^2$-smooth, provided $\mathbf{g}$ is smooth with β_g.

The function $\mathbf{f}(\mathbf{w}) = \log(1+\exp(-y\langle \mathbf{w}, \mathbf{x}\rangle))$ is $\frac{||\mathbf{x}||^2}{4}$-smooth.

3.3.4 Strongly convex functions:

A function $\mathbf{f}$ is λ-strongly convex if $\mathbf{f}(\mathbf{w}) - \frac{\lambda}{\mathbf{2}}$ is convex.

For example, the function $\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{log}(\mathbf{1} + \exp(-\mathbf{y}\langle \mathbf{w}, \mathbf{x} \rangle)) + \frac{\lambda}{\mathbf{2}} ||\mathbf{w}||^{\mathbf{2}}$ is strongly convex.

3.4 Next time:

We will look at convergence bounds, which tell us how fast a function converges.

We will also look at questions like: Why is convexity useful? How do we exploit the structures of functions algorithmically? In addition, we will also start looking at gradient methods.