CS 7545: Machine Learning Theory **Fall 2019**

Lecture 1: Course Introduction and Linear Algebra Review

Lecturer: Jacob Abernethy *Scribes: Alejandro Carderera and Reuben Tate*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

1.1 Course Introduction

1.1.1 Basic Course Information

Instructor Jacob Abernethy (prof@gatech.edu)
TA's:

- Zihao Hu (zihaohu@gatech.edu)
- Bhuvesh Kumar (bhuvesh@gatech.edu)

Location: Weber SST III (Lecture Hall 2)
Office Hours: TBA

1.1.2 Pre-requisites

This course expects that students are coming into the course with some basic knowledge of Advanced Linear Algebra, Graduate Level Probability and Statistics, and Convex Optimization/Analysis. Students that don't know at least at least 2 of the 3 topics above will likely struggle with the course.

1.1.3 Course Outline

The following is a brief overview of the topics that will be discussed in the course.

- Basis (review pre-reqs)
- Online Learning
  - Useful when data arrives sequentially
  - Robust to adversarial settings
  - Applications:
    - Sequential optimization algorithms, Stochastic Gradient Descent (SGD)
    - Solving zero-sum games
    - Reinforcement learning, "bandits"
- Statistical Learning Theory
  - Focusing on generalization guarantees when data points are iid.
  - Bounding/Controlling the difference between training/testing error due to bias-variance tradeoff (See Figure 1.1)
  - Vapnik-Chervonenkis (VC) Theory
  - Uniform Deviation

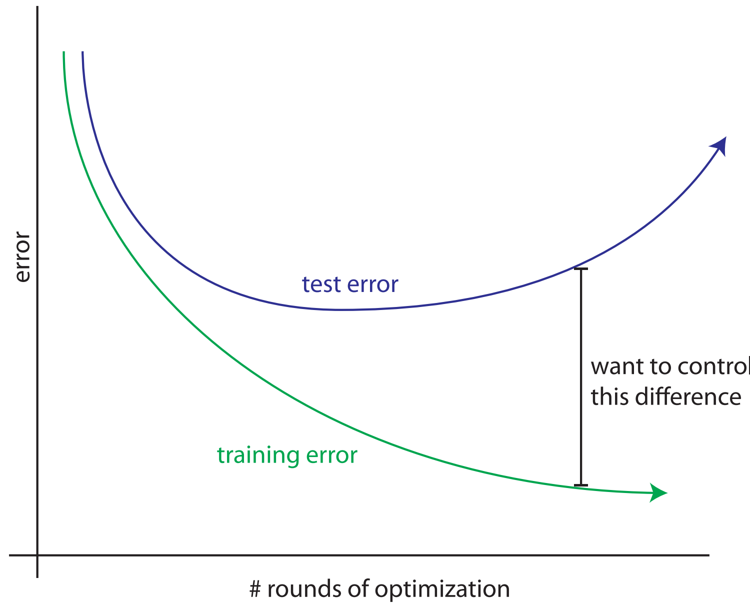

Figure 1.1: Observe that as we do more iterations, the training error decreases; however, at some point, the test error stops decreasing and begins to increase instead.

  - Sauer's Lemma

- Current theoretical results currently work well in practice (i.e. in deep learning). In other words, they seem to do better than what we can prove.
- Old Page: mltheory.github.io

1.1.4 Course Logistics

- Grade Breakdown
  - Homework: 50% (5 homeworks)
  - Final: 40%
  - Scribe: 10%
- Scribe Notes: due one week after lecture. They will be graded but there will be an opportunity to revise.
- Textbook: (will follow lightly) Foundations of Machine Learning by Mohri et al., 2nd edition.
- Use Piazza for homework/content discussions and also for any questions regarding policies.
- Only email the professor if it's a personal issue.
- Homework solutions need to be in LaTeX.
- Homework 1 will be released soon to give students an idea of what to expect regarding the homework in the course.
- A list of things this course *doesn't* cover is graphical models, PyTorch, or TensorFlow.
- Homework in the course will have no/minimal programming.

1.2 Linear Algebra Review

The following conventions will be used throughout the course.

- Matrices in $\mathbb{R}^{n\times m}$ will be denoted as: M.
- A vector in $\mathbb{R}^n$ will be denoted as: $\vec{x}$.
- To refer to the $i-th$ element in a vector we use x_i.

1.2.1 Positive Semidefinite (PSD) and Positive Definite (PD) matrices.

Definition 1.1 (Positive Semidefinite (PSD) Matrix) *A symmetric matrix $M \in \mathbb{R}^{n\times n}$ is said to be* Positive Semidefinite *(PSD), also denoted as $M \succeq 0$, if and only if:*

$$\vec{x}^\top M\vec{x} \geq 0,$$

for all $\vec{x} \in \mathbb{R}^n$. This is equivalent to all the eigenvalues of M being non-negative (greater then or equal to zero).

Definition 1.2 (Positive Definite (PD) Matrix) *A symmetric matrix $M \in \mathbb{R}^{n\times n}$ is said to be* Positive Definite *(PD), also denoted as $M \succ 0$, if and only if:*

$$\vec{x}^\top M\vec{x} > 0,$$

for all $\vec{x} \in \mathbb{R}^n \setminus \vec{0}$. This is equivalent to all the eigenvalues of M being positive (greater than zero).

1.2.2 Norms.

Definition 1.3 (Norm) *A function $\|\cdot\| : \mathbb{R}^n \to [0,\infty)$ is a* norm *when the following are satisfied:*

- ***Identity of Indiscernibles***: $\|\vec{x}\| = 0$ *if and only if* $\vec{x} = \vec{0}$.
- ***Absolute Homogeneity***: $\|\alpha\vec{x}\| = |\alpha|\|\vec{x}\|$ *for all* $\vec{x} \in \mathbb{R}^n$ *and* $\alpha \in \mathbb{R}$.
- ***Triangle Inequality***: $\|\vec{x}+\vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$ *for all* $\vec{x},\vec{y} \in \mathbb{R}^n$.

Examples

- ℓ_2 norm: $\|\vec{x}\|_2 = \sqrt{\sum\limits_{i=1}^{n} x_i^2}$.
- ℓ_1 norm: $\|\vec{x}\|_1 = \sum_{i=1}^{n} |x_i|$.
- ℓ_∞ norm: $\|\vec{x}\|_\infty = \max\limits_{1\leq i\leq n} |x_i|$.
- ℓ_p ($p \in (1,\infty)$) norm: $\|\vec{x}\|_p = \sqrt[p]{\sum\limits_{i=1}^{n} |x_i|^p}$.
- M-norm ($M \in \mathbb{R}^{n\times n}$ is PD): $\|\vec{x}\|_M = \sqrt{\vec{x}^\top M\vec{x}}$.

Note that the ℓ_2, ℓ_1 and ℓ_∞ norms are special cases of the ℓ_p norm. The following example proof shows how ℓ_∞ satisfies the properties of norms in Definition 1.3.

Example Proof: In order to prove that the $\|\vec{x}\|_\infty$ is a valid norm, we must prove that the three norm properties are satisfied.

- **Identity of Indiscernibles**: $\|\vec{x}\|_\infty = \max_{1\leq i\leq n} |x_i| = 0$ if and only if $|x_i| = 0$ for all $i \in [1, n]$.
- **Absolute Homogeneity**: $\|\alpha\vec{x}\|_\infty = \max_{1\leq i\leq n} |\alpha x_i| = |\alpha| \max_{1\leq i\leq n} |x_i| = |\alpha|\|\vec{x}\|_\infty$.
- **Triangle Inequality**: $\|\vec{x}+\vec{y}\|_\infty = \max_{1\leq i\leq n} |x_i+y_i| \leq \max_{1\leq i\leq n} |x_i|+|y_i| \leq \max_{1\leq i,j\leq n} |x_i|+|y_j| = \|\vec{x}\|_\infty+\|\vec{y}\|_\infty$.

Definition 1.4 (Dual Norm) *Given a norm $\|\cdot\|$ we define the* dual norm $\|\cdot\|_*$ *as:*

$$\|\vec{y}\|_* = \sup_{\substack{x\in\mathbb{R}^n \\ \|\vec{x}\|=1}} \langle \vec{x}, \vec{y} \rangle ,$$

where $\langle\cdot,\cdot\rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ denotes the standard dot product.

Claim 1.5 *The ℓ_2 norm is self-dual (its dual norm is itself).*
Proof:

$$\|\vec{y}\|_{2,*} = \sup_{\substack{\vec{x}\in\mathbb{R}^n \\ \|\vec{x}\|_2=1}} \langle \vec{x}, \vec{y} \rangle = \sup_{\vec{x}\in\mathbb{R}^n} \left\langle \frac{\vec{x}}{\|\vec{x}\|_2}, \vec{y} \right\rangle = \frac{1}{\|\vec{y}\|_2} \langle \vec{y}, \vec{y} \rangle = \|\vec{y}\|_2,$$

where the key resides in the fact that the inner product is maximized by making $\vec{x}$ and $\vec{y}$ collinear and have the same direction. ■

(Exercise) Prove that the ℓ_p norm is dual to the ℓ_q norm when:

$$\frac{1}{p} + \frac{1}{q} = 1,$$

.
(Exercise) Prove that for a PSD matrix M the dual norm to the M-norm is the M^{-1}-norm, where M^{-1} denotes the inverse of M.

Theorem 1.6 (Hölders Inequality) *Let $\|\cdot\|$ and $\|\cdot\|_*$ be dual norms. Then for every $\vec{x}, \vec{y} \in \mathbb{R}^n$:*

$$\langle \vec{x}, \vec{y} \rangle \leq \|\vec{x}\|\|\vec{y}\|_*.$$

Proof:

$$\|\vec{y}\|_* = \sup_{\vec{z}\in\mathbb{R}^n \|\vec{z}=1\|} \langle \vec{x}, \vec{y} \rangle \geq \left\langle \frac{\vec{x}}{\|\vec{x}\|}, \vec{y} \right\rangle = \frac{1}{\|\vec{x}\|} \langle \vec{x}, \vec{y} \rangle .$$

Reordering the terms this leads to desired inequality. ■