10-725/36-725: Convex Optimization
Prerequisite Topics

February 3, 2015

This is meant to be a brief, informal refresher of some topics that will form building blocks in this course. The content of the first two sections of this document is mainly taken from Appendix A of B & V, with some supplemental information where needed. See the end for a list of potentially helpful resources you can consult for further information.

1 Real Analysis and Calculus

1.1 Properties of Functions

Limits You should be comfortable with the notion of limits, not necessarily because you will have to evaluate them, but because they are key to understanding other attributes of functions. Informally, $\lim_{x \to a} f(x)$ is the value that f approaches as x approaches the value a.

Continuity A function $f(x)$ is continuous at a particular point x' if, as a sequence $x_1, x_2, ...$ approaches x', the value $f(x_1), f(x_2), ...$ approaches $f(x')$. In limit notation: $\lim_{i \to \infty} f(x_i) = f(\lim_{i \to \infty} x_i)$. f is continuous if it is continuous at all points $x' \in \text{dom} f$.

Differentiability A function $f : \mathbb{R}^n \to \mathbb{R}$ is considered differentiable at $x \in \text{int dom} f$ if there exists a vector $\bigtriangledown f(x)$ that satisfies the following limit:

$$\lim_{z \in \text{dom} f, z \neq x, z \to x} \frac{||f(z) - f(x) - Df(x)(z - x)||_2}{||z - x||_2} = 0$$

We refer to $\bigtriangledown f(x)$ as the derivative of f, and it is the transpose of the gradient.

Smoothness f is smooth if the derivatives of f are continuous over all of $\text{dom} f$. We can describe smoothness of a certain order if the derivatives of f are continuous up to a certain derivative. It is also reasonable to talk about smoothness over a particular interval of the domain of f.

Lipschitz A function f is Lipschitz with Lipschitz constant L if $||f(x) - f(y)|| \leq L||x - y||$ $\forall x, y \in \text{dom} f$. If we refer to a function f as Lipschitz, we are making a stronger statement about the continuity of f. A Lipschitz function is not only continuous, but it does not change value very rapidly, either. This is obviously not unrelated to the smoothness of f, but a function can be Lipschitz but not smooth.

Taylor Expansion The first order Taylor expansion of a function gives us an easy way to form a linear approximation to that function:

$$f(y) \approx f(x) + \nabla f(x)(y - x)$$

And equivalent form that is often useful is the following:

$$f(y) = f(x) + \int_0^1 \nabla f(t(x - y) + y)(y - x)\, dt.$$

For a quadratic approximation, we add another term:

$$f(y) \approx f(x) + \nabla f(x)(y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x)$$

Often when doing convergence analysis we will upper bound the Hessian and use the quadratic approximation to understand how well a technique does as a function of iterations.

1.2 Sets

Interior The interior int$C$ of the set C is the set of all points $x \in C$ for which $\exists \epsilon > 0$ s.t. $\{y || y - x||_2 \leq \epsilon\} \subseteq C$.

Closure The closure cl$C$ of a set C is the set of all x such that $\forall \epsilon > 0\ \exists y \in C$ s.t. $||x - y||_2 \leq \epsilon$. The closure only makes sense for closed sets (see below), and can be considered the union of the interior of C and the boundary of C.

Boundary The boundary is the set of points bd$C$ for which the following is true: $\forall \epsilon\ \exists y \in C, z \notin C$ s.t. $||y - x||_2 \leq \epsilon$ and $||z - x||_2 \leq \epsilon$.

Complement The complement of the set $C \subseteq \mathbb{R}^n$ is denoted by $\mathbb{R}^n \setminus C$. It is the set of all points not in C

Open vs Closed A set C is open if int$C = C$. A set is closed if its complement is open.

Equality You'll notice that above we used a notion of equality for sets. To show formally that sets A and B are equal, you must show $A \subseteq B$ *and* $B \subseteq A$.

1.3 Norms

See B & V for a much more detailed treatment of this topic. I am going to list the most common norms so that you are aware of the notation we will be using in this class:

ℓ_0 $||x||_0$ is the number of nonzero elements in x. We often want to minimize this, but it is non-convex (and actually, not a real norm), so we approximate it (you could say we relax it) to other norms (e.g. ℓ_1).

ℓ_p $||x||_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$, where $p \geq 1$. Some common examples:

- $||x||_1 = \sum_{i=1}^n |x_i|$
- $||x||_2 = \sqrt{\sum_{i=1}^n x_i^2}$

- $||x||_\infty = \max_i |x_i|$

Spectral/Operator Norm $||X||_{op} = \sigma_1(X)$, the largest singular value of X.

Trace Norm $||X||_{tr} = \sum_{i=1}^{r} \sigma_r(X)$, the sum of all the singular values of X.

1.4 Linear/Affine Functions

In this course, a linear function will be a function $f(x) = a^T x$. Affine functions are linear functions with a nonzero intercept term: $g(x) = a^T x + b$.

1.5 Derivatives of Functions

See B & V for some nice examples. Consider the following for a function $f : \mathbb{R}^n \to \mathbb{R}$:

Gradient The i^th element of $\bigtriangledown f$ is the partial derivative of f w.r.t. the i^th dimension of the input x: $\bigtriangledown f(x)_i = \frac{\partial f(x)}{\partial x_i}$

Chain Rule Let $h(x) = g(f(x))$ for $g : \mathbb{R} \to \mathbb{R}$. We have: $\bigtriangledown h(x) = g'(f(x)) \bigtriangledown f(x)$

Hessian In the world of optimization, we denote the Hessian matrix as $\bigtriangledown^2 f(x) \in \mathbb{R}^{n \times n}$ (some of you have maybe seen this symbol used as the Laplace operator in other courses). The ij^th entry of the Hessian is given by: $\bigtriangledown f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$

Matrix Differentials In general we will not be using these too much in class. The major differentials you need to know are:

- $\partial X^T X = X$
- $\frac{\partial}{\partial X} tr(XA) = A^T$

2 Linear Algebra

2.1 Matrix Subspaces

Row Space The row space of a matrix A is the subspace spanned of the rows of A.

Column Space The column space of a matrix A is the subspace spanned of the columns of A.

Null Space The null space of a matrix A is the set of all x such that $Ax = 0$.

Rank rank$A$ is the number of linearly independent columns in A (or, equivalently, the number of linearly independent rows). A matrix $A \in \mathbb{R}^{m \times n}$ is full rank if rank$A = \min\{m, n\}$. Recall that if A is square and full rank, it is invertible.

2.2 Orthogonal Subspaces

Two subspaces $S_1, S_2 \in \mathbb{R}^n$ are orthogonal if $s_1^T s_2 = 0 \; \forall \; s_1 \in S_1, s_2 \in S_2$.

2.3 Decomposition

Eigen Decomposition If $A \in S^n$, the set of real, symmetric, $n \times n$ matrices, then A can be factored:

$$A = Q\Lambda Q^T$$

Here Q is an orthogonal matrix, which means that $Q^TQ = I$. $\Lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$, where the eigenvalues λ_i are ordered by decreasing value. Some useful facts about A that we can ascertain from the eigen decomposition:

- $|A| = \prod_{i=1}^n \lambda_i$
- $\text{tr}A = \sum_{i=1}^n \lambda_i$
- A is invertible iff (if and only if) all its eigenvalues are nonzero. Then $A^{-1} = Q\Lambda^{-1}Q^T$ (note that I have used the fact that for orthogonal Q, $Q^{-1} = Q^T$
- A is positive semidefinite if all its eigenvalues are nonnegative.

Singular Value Decomposition Any matrix $A \in \mathbb{R}^{m\times n}$ with rank r can be factored as:

$$A = U\Sigma V^T$$

Here $U \in \mathbb{R}^{m\times r}$ has the property that $U^TU = I$ and $V \in \mathbb{R}^{n\times r}$ likewise satisfies $V^TV = I$. $\Sigma = \text{diag}(\sigma_1, \sigma_2, ..., \sigma_r)$ where the singular values σ_i are ordered by decreasing value. Some useful facts that we can learn using this decomposition:

- The SVD of A has the following implication for the eigendecomposition of A^TA:

$$A^TA = [VW] \begin{bmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{bmatrix} [VW]^T$$

  W is the matrix such that $[VW]$ is orthogonal.
- The *condition number* of A (an important concept for us in this course) is $\text{cond}A = \frac{\sigma_1}{\sigma_r}$

Pseudoinverse The SVD of a singular matrix A yields the pseudoinverse $A^\dagger = V\Sigma^{-1}U^T$.

3 Canonical ML Problems

3.1 Linear Regression

Linear regression is the problem of finding $f : X \to Y$, where $X \in \mathbb{R}^{n\times p}$, Y is an n-dimensional vector of real values and f is a linear function. Canonically, we find f by finding the vector $\hat{\beta} \in \mathbb{R}^p$ that minimizes the *least squares objective*:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \, ||X\beta - Y||_2^2$$

For $Y \in \mathbb{R}^{n\times q}$, the multiple linear regression problem, we find a matrix $\hat{B}$ that such that:

$$\hat{B} = \underset{B}{\text{argmin}} \, ||XB - Y||_F^2$$

Note that in its basic form, the linear regression problem can be solved in closed form.

3.2 Logistic Regression

Logistic regression is the problem of finding $f : X \to Y$, where Y is an n-dimensional vector binary values, and f has the form $f(x) = \text{logit}(\beta^T x)$. The logit function is defined as $\text{logit}(\alpha) = \frac{1}{1+\exp(-\alpha)}$. We typically solve for β by maximizing the likelihood of the observed data, which results in the following optimization problem:

$$\hat{\beta} = \underset{\beta}{\text{argmax}} \sum_{i=1}^{n} [y_i \beta^T x_i - \log(1 + \exp(-y_i \beta^T x_i)]$$

3.3 Support Vector Machines

Like logistic regression, SVMs attempt to find a function that linearly separates two classes. In this case, the elements of Y are either 1 or -1. SVMs frame the problem as the following constrained optimization problem (in primal form):

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \frac{1}{2} ||\beta||_2^2$$

$$\text{s.t.} \qquad y_i(\beta^T x_i) \geq 1 \qquad \forall i = 1, ..., n$$

In its simplest form, the support vector machine seeks to find the hyperplane (parameterized by β) that separates the classes (encoded in the constraint) and does so in a way that creates the largest margin between the data points and the plane (encoded in the objective that is minimized).

3.4 Regularization/Penalization

Regularization (sometimes referred to as penalization) is a technique that can be applied to almost all machine learning problems. Most of the time, we regularize in an effort to simplify the learned function, often by forcing the parameters to be "small" (either in absolute size or in rank) and/or setting many of them to be zero. Regularization is also sometimes used to incorporate prior knowledge about the problem.

We incorporate regularization by adding either constraints or penalties to the existing optimization problem. This is easiest to see in the context of linear regression. Where previously we only had least squares loss, we can add penalties to create the following two variations:

Ridge Regression By adding an ℓ_2 penalty, our objective to minimize becomes:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \, ||X\beta - Y||_2^2 + \lambda ||\beta||_2$$

This will result in many elements of β being close to 0 (more so if λ is larger).

Lasso Regression By adding an ℓ_1 penalty, our objective to minimize becomes:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \, ||X\beta - Y||_2^2 + \lambda ||\beta||_1$$

This will result in many elements of β being 0 (more if λ is larger).

The first example is nice because it still can be solved in closed form. Notice however that the ℓ_1 penalty creates issues not only for a closed-form solution, but also for standard first-order methods, because it is not differentiable everywhere. We will study how to deal with this later in the course.

4 Further Resources

In addition to B & V, the following are good sources of information on these topics:

- Matrix Cookbook: `http://www.mit.edu/~wingated/stuff_i_use/matrix_cookbook.pdf`
- Linear Algebra Lectures by Zico Kolter: `http://www.cs.cmu.edu/~zkolter/course/linalg/index.html`
- Functional Analysis/Matrix Calculus Lectures by Aaditya Ramdas: `http://www.cs.cmu.edu/~aramdas/videos.html`