Notation

Sanjay Lall and Stephen Boyd

EE104
Stanford University
Basic mathematical notation
we follow the (standard) notation in

*Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares* (VMLS) by Boyd & Vandenberghe, with a few differences noted below
Vectors and matrices

- we denote a (column) vector using \((a, b, c)\), or in vertical form
  \[
  \begin{bmatrix}
  a \\
  b \\
  c
  \end{bmatrix}
  \]

- matrices are generally denoted using capitals, e.g., \(X\), with entries \(X_{ij}\)

- some standard sets:
  - \(a \in \mathbb{R}\) means \(a\) is a scalar (number)
  - \(x \in \mathbb{R}^n\) means \(x\) is an \(n\)-vector
  - \(Z \in \mathbb{R}^{p \times q}\) means \(Z\) is a \(p \times q\) matrix

- transpose of a matrix is \(Z^\top\)

- if \(u\) is a (column) vector, \(u^\top\) is a row vector

- inner product of vectors \(a\) and \(b\) is \(a^\top b\)

- \(1\) is the vector with all entries one
Vector norms

for a vector $x \in \mathbb{R}^n$ there are several common norms

- $\|x\|_2 = \left( x_1^2 + x_2^2 + \cdots + x_n^2 \right)^{1/2}$ is called the 2-norm of a vector or Euclidean norm
  - the most common norm, and so often written without the subscript as $\|x\|$.
  - in VMLS, $\|x\|_2$ is written without the subscript.
- $\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|$ is called the 1-norm.
- $\|x\|_\infty = \max\{|x_1|, |x_2|, \ldots, |x_n|\}$ is called the $\infty$-norm.
- all members of the $p$-norm family, defined as $\|x\|_p = \left( |x_1|^p + \cdots + |x_d|^p \right)^{1/p}$ for $p \geq 1$.
- $\|a - b\|_p$ is the $p$-norm distance between vectors $a$ and $b$. 


Matrix norms

for a matrix $X \in \mathbb{R}^{m \times n}$, there are several common norms

- we use the *Frobenius norm*, denoted $\|X\|_F$

\[
\|X\|_F = \left( \sum_{i,j} X_{ij}^2 \right)^{1/2}
\]

(in VMLS this is denoted without the subscript as $\|X\|$)

- $\|X\|_2$ is the *spectral norm* or *2-norm*, which we won’t use in this course

- $\|X\|_1 = \sum_{i,j} |X_{ij}|$ is the *1-norm*
Course specific notation
Feature mapping

- \( u \): original independent variable or input (not necessarily a number or vector)
- \( v \): original dependent variable or output (not necessarily a number or vector)
- \( x = \phi(u) \)
  - \( x \) is the feature vector in \( \mathbb{R}^d \)
  - \( \phi \) is the feature mapping or embedding
- \( y = \psi(v) \)
  - \( y \) is the target or output vector in \( \mathbb{R}^m \)
  - \( \psi \) is the output feature mapping
Data sets

- $x^1, \ldots, x^n$ and $y^1, \ldots, y^n$ is a data set of $n$ examples
- $x^i, y^i$ is the $i$th data pair
- $n$ is the number of examples or samples
- associated data matrices

\[
X = \begin{bmatrix}
(x^1)^T \\
\vdots \\
(x^n)^T
\end{bmatrix} \in \mathbb{R}^{n \times d}, \quad Y = \begin{bmatrix}
(y^1)^T \\
\vdots \\
(y^n)^T
\end{bmatrix} \in \mathbb{R}^{n \times m}
\]

- rows are feature and target vectors, transposed
Predictors

- $g_\theta : \mathbb{R}^d \to \mathbb{R}^m$ is a predictor
- $\hat{y} = g_\theta(x)$ is the prediction of $y$, given $x$
- $\theta \in \mathbb{R}^p$ is a vector of parameters in the predictor
- choosing $\theta$ based on some data is called training or fitting the predictor
Empirical risk minimization

- given data set \(x^i, y^i, i = 1, \ldots, n\)
- prediction of \(y^i\), given \(x^i\), is \(\hat{y}^i = g_\theta(x^i)\)
- loss on \(i\)th data pair is \(\ell(\hat{y}^i, y^i)\)
- empirical risk is average loss over data set, \(\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\hat{y}^i, y^i)\)
- empirical risk minimization (ERM): choose \(\theta\) to minimize \(\mathcal{L}(\theta)\)
- regularized ERM: choose \(\theta\) to minimize \(\mathcal{L}(\theta) + r(\theta)\)
- \(r\) is regularizer function, which measures sensitivity of \(g_\theta\)