Matrix Methods in Machine Learning Lecture Notes

Rebekah Dix

November 11, 2018

Contents

1 Elements of Machine Learning

1. Collect data
2. Preprocessing: changing data to simplify subsequent operations without losing relevant information.
3. Feature extraction: reduce raw data by extracting features or properties relevant to the model.
4. Generate training samples: a large collection of examples we can use to learn the model.
5. Loss function: To learn the model, we choose a loss function (i.e. a measure of how well a model fits the data)
6. Learn the model: Search over a collection of candidate models or model parameters to find one that minimizes the loss on training data.
7. Characterize generalization error (the error of our predictions on new data that was not used for training).

2 Linear Algebra Review

2.1 Products

Inner products:

$$\langle x, w \rangle = \sum_{j=1}^{p} w_j x_j = x^T w = w^T x \tag{1}$$

Thus this inner product is a weighted sum of the elements of x.

Matrix-vector multiplication:

$$Xw = \begin{bmatrix} —x_1^T— \\ —x_2^T— \\ \vdots \\ —x_n^T— \end{bmatrix} w = \begin{bmatrix} x_1^T w \\ x_2^T w \\ \vdots \\ x_n^T w \end{bmatrix} \tag{2}$$

Matrix-matrix multiplication:

Example 1. Let $X \in \mathbb{R}^{n \times p}$, n movies, p people. $T \in \mathbb{R}^{n \times r}$, and $W \in \mathbb{R}^{r \times p}$. We can think of T as the taste profiles of r representative customers and W as the weights on each representative profile (there will be one set of weights for each customer). Suppose we have two representative taste profiles (i.e. an action lover and a romance lover). Then w will be a 2-vector containing the weights of on the two representative taste profiles. Then

Tw is the expected preferences of a customer who weights the representative taste profiles of T with the weights given in w.

Now we can think about the full matrix product $X = TW$

$$X = TW \implies X_{ij} = \langle i\text{th row of T}, j\text{th column of W} \rangle \tag{3}$$

- The jth column of X is a weighted sum of the columns of T, where the jth column of W tells us the weights.
$$x_j = Tw_j \tag{4}$$
That is, the tastes (preferences) of the jth customer.
- The ith row of X is $x_i^T = t_i^T W$ where t_i^T is the ith row of T. This gives us how much each customer likes movie i.

Inner product representation:

$$TW = \begin{bmatrix} —t_1^T— \\ —t_2^T— \\ \vdots \\ —t_n^T— \end{bmatrix} \begin{bmatrix} | & | & & | \\ w_1 & w_2 & \dots & w_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} t_1^T w_1 & t_1^T w_2 & \dots & t_1^T w_p \\ t_2^T w_1 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ t_n^T w_1 & & & t_n^T w_p \end{bmatrix} \tag{5}$$

Outer Product Representation:

$$TW = \begin{bmatrix} | & | & & | \\ T_1 & T_2 & \dots & T_r \\ | & | & & | \end{bmatrix} \begin{bmatrix} —w_1^T— \\ —w_2^T— \\ \vdots \\ —w_r^T— \end{bmatrix} = \sum_{k=1}^{r} T_k w_k^T \tag{6}$$

(the sum of rank 1 matrices. TW has rank r if and only if the columns of T are rows of W are linearly independent). In this representation, we can think about T_k as the kth representative taste profile and w_k^T as the kth row of W, or the affinity of each customer with the kth representative profile.

2.2 Linear Independence

Definition 1. *(Linear Independence) Vectors $v_1, v_2, \dots, v_n \in \mathbb{R}^p$ are linearly independent vectors if and only if*

$$\sum_{j=1}^{n} \alpha_j v_j = 0 \iff \alpha_j = 0, j = 1, \dots, n \tag{7}$$

Definition 2. *(Matrix rank) The rank of a matrix is the maximum number of linearly independent columns. The rank of a matrix is less than the smallest dimension of the matrix.*

3 Linear Systems and Vector Norms

Example 2. (Condition on $rank(A)$ for existence of exact solution)

Consider the linear system of equations $Ax = b$. This means that b is a weighted sum of the columns of A. Suppose A is full rank. Now consider the matrix $\begin{bmatrix} A & b \end{bmatrix}$. If the rank of $\begin{bmatrix} A & b \end{bmatrix}$ were *greater* than the rank of A (since the number of columns of the matrix increased by 1 and A is assumed full rank, this would imply the rank is $rank(A) + 1$), this would mean that b could not be written as a linear combination of the columns of A, and that the system would not have an exact solution. Therefore, we must have that $rank(\begin{bmatrix} A & b \end{bmatrix}) = rank(A)$ in order for the system $Ax = b$ to have an exact solution.

To see how the definition of linear independence applies here, observe that $Ax = b \implies Ax - b = 0$. Therefore

$$\begin{bmatrix} A & b \end{bmatrix} \begin{bmatrix} x \\ -1 \end{bmatrix} = 0 \tag{8}$$

Thus, if $Ax = b$ has an exact solution, then $\begin{bmatrix} A & b \end{bmatrix}$ does not have linearly independent columns.

Example 3. (Condition on $rank(A)$ for more than one exact solution)

If the system of linear equations $Ax = b$ has more than one exact solution, then there is at least one non zero vector w for which $x + w$ is also a solution. That is, $A(x + w) = b$. If x is an exact solution, then $Ax = b$. This implies $Aw = 0$. Therefore, the columns of A are linearly dependent. Thus, if $rank(A) < dim(x)$, then there will be more than one exact solution.

Example 4. (Apply the above conditions) Let

$$A = \begin{bmatrix} 1 & -2 \\ -1 & 2 \\ -2 & 4 \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ -2 \\ -4 \end{bmatrix} \tag{9}$$

We want to solve $Ax = b$.

- This system has an exact solution, since $rank(A) = rank(\begin{bmatrix} A & b \end{bmatrix})$. This follows since the columns of A are linearly dependent, so it has rank 1, and b is a multiple of the columns of A, so the rank of $\begin{bmatrix} A & b \end{bmatrix}$ is also 1.
- Note that $1 = rank(A) < dim(x) = 2$. Therefore this system does not have a unique solution.

3.1 Vector Norms

Definition 3. *(Vector Norm) A vector norm is a function $\|\cdot\|$ mapping from $\mathbb{R}^n \to \mathbb{R}$ with the following properties.*

1. $\|x\| \geq 0$ *for all* $x \in \mathbb{R}^n$.

2. $\|x\| = 0$ *if and only if* $x = 0$.

3. $\|\alpha x\| = |\alpha| \, \|x\|$ *for all* $\alpha \in \mathbb{R}$, $x \in \mathbb{R}^n$.

4. $\|x + y\| \leq \|x\| + \|y\|$ *for all* $x, y \in \mathbb{R}^n$.

Helpful fact: $\|x\|_{q'} \leq \|x\|_q$ if $1 \leq q \leq q' \leq \infty$.

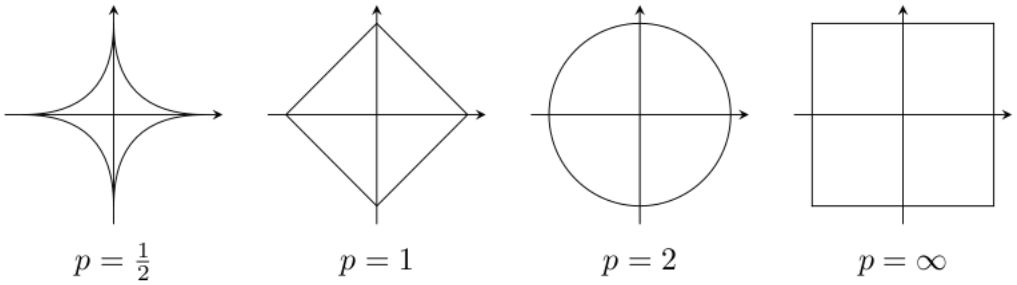

Figure 1: The l_p norm in $\mathbb{R}^2$

4 Least Squares

We are given:

1. Vector of labels $y \in \mathbb{R}^n$
2. Matrix of features $X \in \mathbb{R}^{n \times p}$

We want to find:

1. Vector of weights $w \in \mathbb{R}^p$

Assumptions:

1. $n \geq p$, and $rank(X) = p$.

If $y = Xw$, then we have a system of n linear equations, where the ith equation is

$$y_i = w_1 x_{i1} + w_2 x_{i2} + \cdots + w_p x_{ip} = \sum_{j=1}^{p} w_j x_{ij} = \langle w, x_{\cdot i} \rangle \tag{10}$$

where $x_{\cdot i}$ is the ith row of X.

In general, $y \neq Xw$ for any w. We define a residual $r_i = y_i - \langle w, x_{\cdot i} \rangle$. Our goal is then to find w $\sum_{i=1}^{n} |r_i|^2$ (the sum of square residuals/errors).

Why should we minimize the sum of square errors?

1. Magnifies the effect of large errors

2. Allows us to compute derivatives

3. Simple geometric interpretation

4. Coincides with modeling $y = Xw + \epsilon$, where ϵ is Gaussian noise

4.1 Geometric Approach

We know $\hat{r} = y - X\hat{w}$ is orthogonal to the span of the columns of X. Thus $x_i^T \hat{r} = 0$, or $X^T \hat{r} = 0$. This implies $X^T(y - X\hat{w}) = 0$. Thus $\hat{w}$ is a solution to the linear system of equations

$$X^T X\hat{w} = X^T y \tag{11}$$

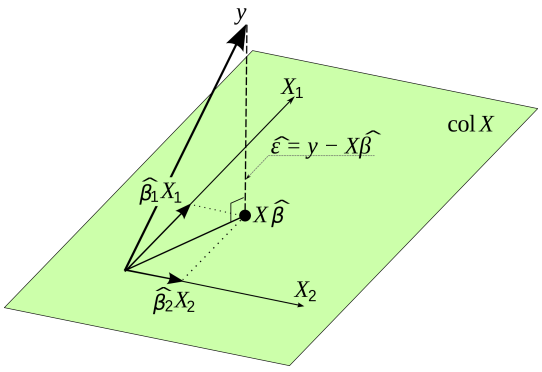

Figure 2: Geometry of LS in $\mathbb{R}^2$

Observations:

- The question we're trying to answer: What is the point in $col(X)$ that has the shortest distance to y? In $\mathbb{R}^2$, what are the weights β_1 and β_2 such that $\beta_1 x_1 + \beta_2 x_2$ has the shortest distance to y?

- $colX$ is the space of all vectors that can be written as $\alpha x_1 + \beta x_2$ for some $\alpha, \beta \in \mathbb{R}$, that is the span of the columns of X. y may not lie in this space.

- The residual vector will form a right angle with $colX$, because any other angle would correspond to a longer distance.

4.2 Vector Calculus Approach

4.2.1 Review of Vector Calculus

Let w be a p-vector and let f be a function of w that maps $\mathbb{R}^p$ to $\mathbb{R}$. Then the gradient of f with respect to w is

$$\nabla_w f(w) = \begin{pmatrix} \frac{\partial f(w)}{\partial w_1} \\ \vdots \\ \frac{\partial f(w)}{\partial w_p} \end{pmatrix} \tag{12}$$

Example 5. (Gradient of an Inner Product) Let $f(w) = \langle a, w \rangle = w^T a = \sum_{i=1}^n w_i a_i$. Then

$$\nabla_w w^T a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = a \tag{13}$$

Example 6. (Gradient of an Inner Product, Squared) Let $f(w) = \|w\|^2 = w^T w = w_1^2 + \cdots + w_p^2$. Then

$$\nabla_w w^T w = \begin{pmatrix} 2w_1 \\ 2w_2 \\ \vdots \\ 2w_p \end{pmatrix} = 2w \tag{14}$$

(This is a special case of the Quadratic Form discussed below, where $w^T Q w$, and $Q = I$)

Example 7. (Gradient of a Quadratic Form) Let $x \in \mathbb{R}^n$ and $f(x) = x^T Q x$, where Q is symmetric (if Q isn't symmetric we could replace Q with $\frac{1}{2}(Q + Q^T)$). Then

$$\begin{aligned} f(x) &= x^T Q x \\ &= \sum_{i=1}^n \sum_{j=1}^n x_i Q_{ij} x_j \end{aligned}$$

Therefore

$$[\nabla_x f]_k = \frac{df}{dx_k} = \begin{cases} 2Q_{kk}x_k & i = j = k \\ Q_{kj}x_j & i = k, i \neq j \\ Q_{ik}x_i & j = k, j \neq i \end{cases} \tag{15}$$

Therefore

$$\nabla_x f = (Q + Q^T)x \tag{16}$$

If Q is symmetric, then this equals $2Qx$.

4.2.2 Application to Least Squares

Let $f(w) = \|y - Xw\|_2^2$. Then the least squares problem is

$$\hat{w} = \arg\min_w f(w) \tag{17}$$

We can expand $f(w)$ as

$$\begin{aligned} f(w) &= (y - Xw)^T(y - Xw) \\ &= y^T y - y^T Xw - w^T X^T y + w^T X^T Xw \\ &= y^T y - 2w^T X^T y + w^T X^T Xw \end{aligned}$$

Then

$$\nabla_w f(w) = -2X^T y + 2X^T Xw$$

At an optimum we have that $\hat{w}$ solves $X^T y = X^T Xw$. Then if $(X^T X)^{-1}$ exists, we have that

$$\hat{w} = (X^T X)^{-1} X^T y \tag{18}$$

Theorem 1. *(Sufficient Condition for Existence/Uniqueness of LS Solution) If the columns of X are linearly independent, then $X^T X$ is non-singular, and there exists a unique least squares solution $\hat{w} = (X^T X)^{-1} X^T y$.*

Proof. □

4.3 Positive Definite Matrices

Definition 4 (Positive Definite, pd)**.** *A matrix Q ($n \times n$) is positive definite (written $Q \succ 0$) if $x^T Qx > 0$ for all $x \in \mathbb{R}^n$, $x \neq 0$.*

Definition 5 (Positive Semi-Definite, psd)**.** *A matrix Q ($n \times n$) is positive semi-definite (written $Q \succeq 0$) if $x^T Qx \geq 0$ for all $x \in \mathbb{R}^n$, $x \neq 0$.*

Properties of Positive Definite matrices:

1. If $P \succ 0$ and $Q \succ 0$, then $P + Q \succ 0$.
2. If $P \succ 0$ and $\alpha > 0$, then $\alpha P \succ 0$.
3. For any matrix A, $A^T A \succeq 0$ and $AA^T \succeq 0$. Further, if the columns of A are linearly independent, then $A^T A \succ 0$.
4. If $A \succ 0$, then A^{-1} exists.
5. Notation: $A \succ B$ means $A - B \succ 0$.

Example 8. Let

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \tag{19}$$

Then

$$X^T X = \begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix} \tag{20}$$

Consider the vector $a = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Then $a^T X^T X a = 0$. Therefore $X^T X$ is not positive definite.

4.4 Subspaces

Definition 6. *(Subspace) A set of points $S \subseteq \mathbb{R}^n$ is a subspace if*

1. *$0 \in S$ (S contains the origin)*
2. *If $x, y \in S$, then $x + y \in S$*
3. *If $x \in S$, $\alpha \in \mathbb{R}$, then $\alpha x \in S$.*

4.5 Least Squares with Orthonormal Basis for Subspace

Suppose are given a training sample $\{x_i, y_i\}_{i=1}^n$, $x_i \in \mathbb{R}^p$ and $y \in \mathbb{R}$. If the columns of X (the data matrix) are linearly dependent, then $X^T X$ is not invertible. It is then impossible to tell which features are significant predictors of y.

Given $X = \begin{bmatrix} x_1 \dots x_p \end{bmatrix}$, the following are options to represent the corresponding subspace spanned by the columns of X:

1. Use X, but then LS can be hard to interpret.
2. Use an orthonormal basis for the subspace.

4.5.1 Orthogonal Matrices and Orthonormal Basis

Definition 7. *(Orthonormal basis for X) An orthonormal basis for the columns of X is a collection of vectors $\{u_1, \dots, u_r\}$ such that the span of the columns of X equals the span of $\{u_1, \dots, u_r\}$. That is, $span(\{x_1, \dots, x_p\}) = span(\{u_1, \dots, u_r\})$. Furthermore,*

$$u_i^T u_j = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} \tag{21}$$

That is, the u vectors are orthogonal and have norm 1.

Observations:

- The rank r of the subspace must satisfy $r \leq \min(n, p)$. r is the number of linearly independent columns of X.
- We can place the basis vectors into a basis matrix $U \in \mathbb{R}^{n \times r}$.

Claim 1. *(Properties of orthogonal (basis) matrices) Let $U \in \mathbb{R}^{n \times r}$ be an orthogonal (basis) matrix.*

1. $U^T U = I$
2. *If U and V are both orthogonal, then UV is also orthogonal.*
3. *U is length preserving: $\|Uv\|_2 = \|v\|_2$ for $v \in \mathbb{R}^n$.*

Proof. We prove each item as follows:

1. We can easily see this from the inner product interpretation of matrix multiplication.
2. $(UV)^T UV = V^T U^T UV = V^T V = I$.
3. $\|Uv\|_2^2 = (Uv)^T Uv = v^T U^T Uv = v^T v = \|v\|_2^2$.

□

4.5.2 Back to LS

Suppose U is an orthonormal basis matrix for our data matrix X. Then, the least-squares problem is

$$\hat{v} = \arg\min_v \|y - Uv\|_2^2 \tag{22}$$

We need $\hat{v}$ to satisfy $U^T U\hat{v} = U^T y$. Thus, $\hat{v} = U^T y$.

4.5.3 Gram-Schmidt Orthogonalization Algorithm

How can we take X and get an orthonormal basis U?

1. Input $X = \begin{bmatrix} x_1 \dots x_p \end{bmatrix} \in \mathbb{R}^{n \times p}$
   Output: $U = \begin{bmatrix} u_1 \dots u_r \end{bmatrix} \in \mathbb{R}^{n \times r}$
   where $r = rank(X) \leq \min(n, p)$
2. Initialize $u_1 = \frac{x_1}{\|x_1\|_2}$
3. For $j = 2, 3, \dots, p$
   $x_j' =$ all the components of x_j not represented by $u_1, \dots, u_{j-1}$.

$$x_j' = x_j - \sum_{i=1}^{j-1} (u_i^T x_j) u_i \tag{23}$$

here $(u_i^T x_j)$ is the least squares weight for u_i.

$$u_j = \begin{cases} \frac{x_j'}{\|x_j'\|_2} & x_j' \neq 0 \\ 0 & x_j' = 0 \end{cases} \tag{24}$$

Next, by construction, each column of U, u_i, is in $span(\{x_1, \ldots, x_p\})$. Therefore we can write

$$u_i = \alpha_{i1} x_1 + \alpha_{i2} x_2 + \cdots + \alpha_{ip} x_p \tag{25}$$

where the $\alpha_{ij} \in \mathbb{R}$. We can write this in matrix form as

$$U = XA \tag{26}$$

where X is $n \times p$ and A is $p \times r$, and the ith column of A is

$$a_i = \begin{bmatrix} \alpha_{i1} \\ \alpha_{i2} \\ \vdots \\ \alpha_{ip} \end{bmatrix} \tag{27}$$

Thus, $u_i = Xa_i$.

Now, suppose $w \in \mathbb{R}^p$ is the vector of weights we found using LS, and as above, v is our vector of weights founding using LS with an orthonormal basis matrix. We have two equations for the predicated label $\hat{y}$

$$\begin{aligned} \hat{y} &= w_1 x_2 + w_2 x_2 + \cdots + w_p x_p \\ &= v_1 u_1 + v_2 u_2 + \cdots + v_r u_r \\ &= v_1 X a_1 + v_2 X a_2 + \cdots + v_r X a_r \\ &= v_1(\alpha_{11} x_1 + \alpha_{12} x_2 + \cdots + \alpha_{1p} x_p) \\ &\quad \vdots \\ &\quad + v_r(\alpha_{r1} x_1 + \alpha_{r2} x_2 + \cdots + \alpha_{rp} x_p) \\ &= x_1(v_1 \alpha_{11} + \cdots + v_r \alpha_{r1}) \\ &\quad \vdots \\ &\quad + x_p(v_1 \alpha_{1p} + \cdots + v_r \alpha_{rp}) \end{aligned}$$

Notice that

$$\begin{aligned} w_1 &= v_1 \alpha_{11} + \cdots + v_r \alpha_{r1} \\ &\vdots \\ w_p &= v_1 \alpha_{1p} + \cdots + v_r \alpha_{rp} \end{aligned}$$

Therefore

$$\hat{y} = XAv = Xw \tag{28}$$

so that $Av = w$.

In sum, given a new sample $x_{new} \in \mathbb{R}^p$, we have two ways to predict label y_{new}:

1. $\hat{y}_{new} = \langle x_{new}, w \rangle$
2. Using an orthonormal basis U, we know that $U = XA$. Therefore, $u_{new}^T = x_{new}^T A$. Equivalently, $u_{new} = Ax_{new}$. Then $y_{new} = \langle u_{new}, v \rangle$.

If the columns of X are linearly independent ($r = p$), we can calculate using LS (recalling $u_i = Xa_i$)

$$a_i = (X^T X)^{-1} X^T u_i \tag{29}$$

Theorem 2. *Let $X \in \mathbb{R}^{n \times p}$, $n \geq p$, be full rank (the p columns of X are linearly independent) and $y \in \mathbb{R}^n$. Let $u_1, \ldots, u_p$ be orthonormal basis vectors such that $span(\{x_1, \ldots, x_p\}) = span(\{u_1, \ldots, u_p\})$. Then $\hat{y} = X\hat{w}$ where $\hat{w} = \arg\min_w \|y - Xw\|_2^2$ is given by $\hat{y} = UU^T y$, where $U = \begin{bmatrix} u_1 & u_2 & \ldots & u_p \end{bmatrix}$.*

Proof.

$$\hat{y} = X\hat{w} = X(X^T X)^{-1} X^T y \tag{30}$$

where $P_x = X(X^T X)^{-1} X^T$ is a projection matrix. Since $span(\{x_1, \ldots, x_p\}) = span(\{u_1, \ldots, u_p\})$, we must have that

$$P_x y = P_u y \tag{31}$$

which implies $P_x = P_u$. Thus

$$P_x = P_u = U(U^T U)^{-1} U^T = UU^T \tag{32}$$

Finally

$$\hat{y} = P_x y = P_u y = UU^T y \tag{33}$$

□

5 Least Squares Classification

We are given a training sample $\{x_i, y_i\}_{i=1}^n$, $x_i \in \mathbb{R}^p$ and $y \in \mathbb{R}$ (or $y \in \{+1, -1\}$).

Definition 8. *(Linear Predictor) We have a linear predictor if each label is a linear combination of the features i.e. we can find weights $\{w_i\}_{i=1}^p$ such that*

$$y_i = w_1 x_{i1} + w_2 x_{i2} + \ldots w_p x_{ip} \tag{34}$$

In words, this says the label for observation i is a linear combination of the features for example i.

The steps to complete least squares classification in this environment are as follows:

1. Build a data matrix or feature matrix and label vector

$$X = \begin{bmatrix} —x_1^T— \\ —x_2^T— \\ \vdots \\ —x_n^T— \end{bmatrix} = \begin{bmatrix} x_1^T & 1 \\ x_2^T & 1 \\ \vdots & \\ x_n^T & 1 \end{bmatrix} \in \mathbb{R}^{n \times p}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{35}$$

   The linear model is then $\hat{y} = Xw$.

2. Solve a least squares optimization problem

$$\hat{w} = \arg\min_w \|y - Xw\|_2^2 = \arg\min_w \sum_{i=1}^{n} (y_i - x_i^T w)^2 \tag{36}$$

   (this last equality makes it clear that we are minimizing the sum of squared residuals). If the columns of X are linearly independent, then $X^T X$ is positive definite. Therefore $X^T X$ is invertible. In sum, if $X^T X$ is positive definite, then there exists a unique LS solution

$$\hat{w} = (X^T X)^{-1} X^T y \tag{37}$$

   The predicted labels are

$$\begin{aligned} \hat{y} &= Xw \\ &= X(X^T X)^{-1} X^T y \end{aligned}$$

3. Validate with test/hold out data

6 Tikhonov Regularization/Ridge Regression

We are given $X \in \mathbb{R}^{n \times p}$ (n training samples, p features) and $y \in \mathbb{R}^n$ (n labels). Our model is $y \approx Xw$, which means $y_i \approx x_i^T w$ for some $w \in \mathbb{R}^p$.

The LS problem is

$$\hat{w}_{LS} = \arg\min_w \|y - Xw\|_2^2 = \arg\min_w \sum_{i=1}^{n} (y_i - x_i^T w)^2 \tag{38}$$

There are two cases

1. If X is full rank (i.e. the columns of X are linearly independent), then $\hat{w}_{LS}$ is unique and

$$\hat{w}_{LS} = (X^T X)^{-1} X^T y \tag{39}$$

2. If X is not full rank, then $X^T X$ is not invertible. $\hat{w}_{LS}$ is not unique; there are infinitely many solutions.

6.1 Tikhonov Regularization Derivation

In this second case (and it can also be useful in the first), we can define a new objective

$$\hat{w} = \arg\min_{w} \|y - Xw\|_2^2 + \lambda \|w\|_2^2 \tag{40}$$

where $\|y - Xw\|_2^2$ measures the fit to the data, $\lambda > 0$ is a regularization parameter or tuning parameter, and $\|w\|_2^2$ is a regularizer. $\|w\|_2^2$ measures the energy in w.

Observations about this problem:

1. $\hat{w}$ is unique even when no unique least square solution exists
2. Even when X is full rank, X^TX can be badly behaved, and regularization adjusts for this.

6.1.1 Derivation with Vector Calculus

Let $f(w) = \|y - Xw\|_2^2 + \lambda \|w\|_2^2$. Then

$$\begin{aligned} f(w) &= y^Ty - 2w^TX^Ty + w^TX^TXw + \lambda w^Tw \\ &= y^Ty - 2w^TX^Ty + w^T(X^TX + \lambda I)w \end{aligned}$$

Then

$$\nabla_w f(w) = -2X^Ty + 2(X^TX + \lambda I) \tag{41}$$

If $(X^TX + \lambda I)$ is invertible, then $\hat{w} = (X^TX + \lambda I)^{-1}X^Ty$. BUT, $(X^TX + \lambda I)$ is *always* invertible. Recall that if a matrix is positive definite, then it is invertible. We can show that $(X^TX + \lambda I)$ is indeed positive definite and hence invertible. To see this, fix $0 \neq a \in \mathbb{R}^n$, then

$$\begin{aligned} a^T(X^TX + \lambda I)a &= a^TX^TXa + \lambda a^Ta \\ &= \|Xa\|_2^2 + \lambda \|a\|_2^2 \end{aligned}$$

Now note that $\|Xa\|_2^2 \geq 0$ (it could be 0 if X is not full rank and a is in the null space of X – this is what causes troubles with LS) but $\lambda \|a\|_2^2 > 0$. Therefore,$(X^TX + \lambda I)$ is positive definite.

6.1.2 Alternative Derivation

Note that for vectors a, b,

$$\|a\|_2^2 + \|b\|_2^2 = \left\| \begin{bmatrix} a \\ b \end{bmatrix} \right\|_2^2 \tag{42}$$

Therefore,

$$\begin{aligned} f(w) &= \|y - Xw\|_2^2 + \lambda \|w\|_2^2 \\ &= \|y - Xw\|_2^2 + \left\|\sqrt{\lambda} w\right\|_2^2 \\ &= \left\| \begin{bmatrix} y - Xw \\ \sqrt{\lambda} w \end{bmatrix} \right\|_2^2 \\ &= \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} Xw \\ \sqrt{\lambda} w \end{bmatrix} \right\|_2^2 \\ &= \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix} w \right\|_2^2 \\ &= \|\tilde{y} - \tilde{X} w\|_2^2 \end{aligned}$$

We can solve this problem with LS, so that

$$\hat{w} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X} \tilde{y} \tag{43}$$

where

$$\tilde{X}^T \tilde{X} = X^T X + \lambda I \tag{44}$$

and

$$\tilde{X} \tilde{y} = X^T y \tag{45}$$

Thus this is equivalent to the derivation above.

7 Singular Value Decomposition

Theorem 3. *Every matrix $X \in \mathbb{R}^{n \times p}$ can be factorized as*

$$X = U \Sigma V^T \tag{46}$$

where

1. *$U \in \mathbb{R}^{n \times n}$ is orthogonal ($U^T U = U U^T = I$) are the left singular vectors of X*
2. *$\Sigma \in \mathbb{R}^{n \times p}$ is diagonal and contains the singular values of X*
3. *$V \in \mathbb{R}^{p \times p}$ is orthogonal ($V^T V = V V^T = I$) and contains the right singular vectors of X*

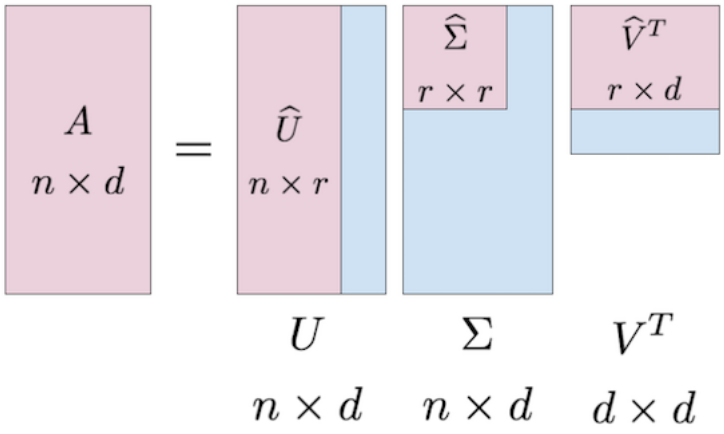

Figure 3: SVD

7.1 Interpretation of SVD

1. U is an orthonormal basis for the columns of X
2. ΣV^T are the basis coefficients

Example 9. (Netflix) Let $X \in \mathbb{R}^{n \times p}$ be a matrix (full rank) where the columns are taste profiles of customers and the rows are single movie ratings across customers.

1. The ith column of U is a basis vector in $\mathbb{R}^n$ and is the ith representative customer taste profile (i.e. vector of normalized movie ratings).
2. The jth column of V^T (the jth row of V) is the relative importance of each representative taste profile to predicting customer j's preferences.
3. The ith row of V^T (the ith column of V) is the vector of users' affinities to the ith representative profile.

7.2 Low-Rank Approximation

Theorem 4. *(Subspace Approximation) If $X \in \mathbb{R}^{p \times n}$ has rank $r > k$, then*

$$\min_{Z \,:\, rank(Z)=k} \|X - Z\|_F^2 \tag{47}$$

is given by $Z = X_k = U_k \Sigma_k V_k^T$ and

$$\|X - X_k\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2 \tag{48}$$

8 Power Iteration and Page Rank

8.1 SVD: Connection to Eigenvalues/vectors

Suppose $X = U\Sigma V^T \in \mathbb{R}^{p\times n}$. Then

$$\begin{aligned} A &:= X^T X \\ &= V\Sigma U^T U\Sigma V^T \\ &= V\Sigma^2 V^T \\ &= V\Lambda V^T \end{aligned}$$

Thus V is a matrix of the eigenvectors of A and $\Lambda = \Sigma^2$ contains its eigenvalues.

Power iteration gives a method to find the 1st right singular vector.

9 Matrix Completion

9.1 Iterative Singular Value Thresholding

Algorithm 1 Iterative Singular Value Thresholding

Require: $\hat{X} = zeros(n, p)$
Require: $\hat{X}_\Omega = X_\Omega$ ▷ fill in obs. entries
Require: Threshold or r
  for $k \leftarrow 0, 1, \ldots$ **do**
    $[U, S, V] \leftarrow svd(\hat{X})$
    $\hat{S} \leftarrow S \geq$ threshold ▷ if threshold, keep sing vals $\geq$
    $\hat{S} \leftarrow S(1:r, 1:r)$ ▷ if rank, keep r sing vals
    $\hat{X} \leftarrow U\hat{S}V^T$
    $\hat{X}_\Omega \leftarrow X_\Omega$ ▷ fill in obs. entries
  If converged: $\left\|\hat{X} - \hat{X}_{old}\right\| < \epsilon$, stop.
  end for

10 Iterative Solvers

Let $\tau > 0$ be step size.

Algorithm 2 Landweber Iteration

Require: $w^{(0)}$
  for $k \leftarrow 0, 1, \ldots$ **do**
    $w^{(k+1)} \leftarrow w^{(k)} - \tau X^T(Xw^{(k)} - y)$
  end for

10.1 Gradient Descent/Landweber Iteration

This algorithm takes a step in the direction of negative gradient of each iteration of the objective function $f(w) = \|Xw - y\|_2^2$. Notice that

$$\begin{aligned} f(w) &= \|Xw - y\|_2^2 \\ \nabla_w f(w) &= \nabla_w \quad (Xw - y)^T(Xw - y) \\ &= \nabla_w \quad w^T X^T Xy - 2w^T X^T y + y^T y \\ &= 2X^T Xw - 2X^T y \\ &= 2X^T(Xw - y) \end{aligned}$$

Thus the new iterate equals the old iterate plus a step in direction of negative gradient.

Claim 2. *(Convergence of Landweber Iteration)*

Proof. We want to show that

$$\left\| Xw^{(k+1)} - y \right\|_2^2 \leq \left\| Xw^{(k)} - y \right\|_2^2 \tag{49}$$

Recall that the iteration is given by $w^{(k+1)} = w^{(k)} - \tau X^T(Xw^{(k)} - y)$. Then

$$\begin{aligned} \left\| Xw^{(k+1)} - y \right\|_2^2 &= \left\| X(w^{(k)} - \tau X^T(Xw^{(k)} - y)) - y \right\|_2^2 \\ &= \left\| Xw^{(k)} - y - \tau XX^T(Xw^{(k)} - y) \right\|_2^2 \\ &= \left\| Xw^{(k)} - y \right\|_2^2 + \tau^2 \left\| XX^T(Xw^{(k)} - y) \right\|_2^2 - 2\tau(Xw^{(k)} - y)^T XX^T(Xw^{(k)} - y) \end{aligned}$$

Now observe that

$$\left\| XX^T(Xw^{(k)} - y) \right\|_2^2 \leq \|X\|_{op}^2 \left\| X^T(Xw^{(k)} - y) \right\|_2^2 \tag{50}$$

Therefore

$$\begin{aligned}\left\|Xw^{(k+1)}-y\right\|_2^2 &\leq \left\|Xw^{(k)}-y\right\|_2^2+\tau\left(\tau\left\|X\right\|_{op}^2\left\|X^T(Xw^{(k)}-y)\right\|_2^2-2\left\|X^T(Xw^{(k)}-y)\right\|_2^2\right)\\ &= \left\|Xw^{(k)}-y\right\|_2^2+\tau\left\|X^T(Xw^{(k)}-y)\right\|_2^2(\tau\left\|X\right\|_{op}^2-2)\end{aligned}$$

Thus, if $(\tau\left\|X\right\|_{op}^2-2)<0$, then $\left\|Xw^{(k+1)}-y\right\|_2^2 \leq \left\|Xw^{(k)}-y\right\|_2^2$. Therefore, for convergence we require that

$$0<\tau<\frac{2}{\left\|X\right\|_{op}^2} \tag{51}$$

in order to ensure convergence. Under this condition, then

$$w^{(k)} \to (X^TX)^{-1}X^Ty \tag{52}$$

□

The above proof made use of the following claim.

Claim 3. *(Bound on 2-norm of matrix-vector product) Let X be matrix and w a vector (conformable). Then*

$$\left\|Xw\right\|_2 \leq \left\|X\right\|_{op}\left\|w\right\|_2 \tag{53}$$

Proof. Recall: **The 2-norm (here, operator norm) of a matrix is it's largest singular value.**

$$\left\|X\right\|_{op} = \left\|X\right\|_2 = \sigma_{max} \tag{54}$$

Suppose $X = U\Sigma V^T$. Then

$$\begin{aligned}\left\|Xw\right\|_2 &= \left\|U\Sigma V^Tw\right\|_2\\ &= \left\|\Sigma V^Tw\right\|_2 && (U \text{ orthonormal, preserves norms})\\ &= \left(\sum_i\left(\sigma_i(V^Tw)_i\right)^2\right)^{\frac{1}{2}}\\ &\leq \sigma_{max}\left(\sum_i(V^Tw)_i^2\right)^{\frac{1}{2}}\\ &= \sigma_{max}\left\|V^Tw\right\|_2 && (\text{definition of norm})\\ &= \sigma_{max}\left\|w\right\|_2 && (V, V^T \text{ orthonormal, preserves norms})\end{aligned}$$

□

11 Regularized Regression

11.1 Proximal Gradient Algorithm

Algorithm 3 Proximal Gradient Algorithm: $\arg\min_w \|y - Xw\|_2^2 + \lambda r(w)$

Require: Initial $\boldsymbol{w}^{(0)}$
  for $k \leftarrow 0, 1, \ldots$ **do**
    $\boldsymbol{z}^{(k)} \leftarrow \boldsymbol{w}^{(k)} - \tau X^T(X\boldsymbol{w}^{(k)} - \boldsymbol{y})$ ▷ grad descent step
    $\boldsymbol{w}^{(k+1)} \leftarrow \arg\min_{\boldsymbol{w}} \left\|\boldsymbol{z}^k - \boldsymbol{w}\right\|_2^2 + \lambda\tau r(\boldsymbol{w})$ ▷ regularization step
  end for

11.2 LASSO (Least absolute selection and shrinkage operator)

LASSO solve the following problem

$$\hat{w}_L = \arg\min_w \|w\|_1 \text{ subject to } \|y - Xw\| < \epsilon \tag{55}$$

which is equivalent to

$$\hat{w}_L = \arg\min_w \|y - Xw\|_2^2 + \lambda \|w\|_1 \tag{56}$$

In the figure below, the rhombuses and circles show the locus of points for which the weight vector has a particular norm (in the L_1 and L_2 norms respectively). More precisely, they are $\{w : \|w\|_1 = \tau_1\}$ and $\{w : \|w\|_2 = \tau_2\}$. The red line is $\{w : y = Xw\}$.

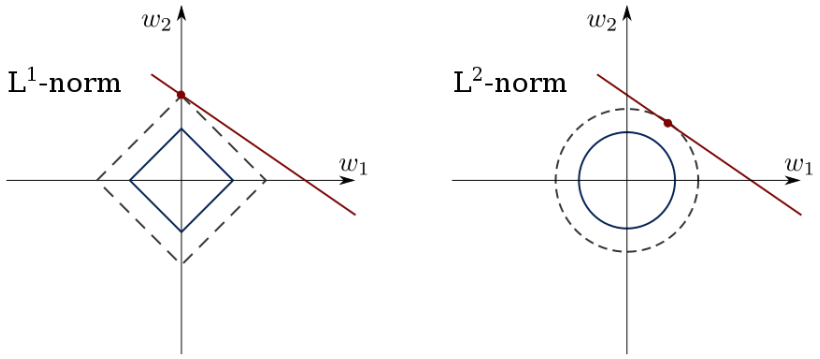

Figure 4: Weight vector with Lasso vs Ridge Regression

Example 10. ($r(w) = \|w\|_1 = \sum_i |w_i|$)

$$\begin{aligned}\hat{w} &= \arg\min_w \|z - w\|_2^2 + \lambda\tau \|w\|_1 \\ &= \arg\min_w \sum_i \Big((z_i - w_i)^2 + \lambda\tau |w_i| \Big) \qquad \text{(separable)}\end{aligned}$$

Thus we can solve a problem for each w_i.

$$\hat{w}_i = \arg\min_{w_i} (z_i - w_i)^2 + \lambda\tau |w_i| \tag{57}$$

where $\lambda, \tau > 0$.

1. Case 1: $z_i > 0$. Then $\hat{w}_i \geq 0$. Thus

$$\begin{aligned}\hat{w}_i &= \arg\min_{w_i} (z_i - w_i)^2 + \lambda\tau w_i \\ \frac{d}{dw_i}(obj) &= -2(z_i - w_i) + \lambda\tau \\ w_i &= z_i - \frac{\lambda\tau}{2}\end{aligned}$$

   Thus

   (a) If $z_i > \frac{\lambda\tau}{2}$, then $\hat{w}_i = z_i - \frac{\lambda\tau}{2}$.

   (b) If $z_i < \frac{\lambda\tau}{2}$, then $\hat{w}_i = 0$.

   In sum

$$\hat{w}_i = \left(z_i - \frac{\lambda\tau}{2} \right)_+ \tag{58}$$

2. Case 2: $z_i < 0$. Then $\hat{w}_i \leq 0$. Similarly,

$$w_i = z_i + \frac{\lambda\tau}{2} \tag{59}$$

   Thus

   (a) If $z_i < -\frac{\lambda\tau}{2}$, then $\hat{w}_i = z_i \frac{\lambda\tau}{2}$.

   (b) If $z_i > -\frac{\lambda\tau}{2}$, then $\hat{w}_i = 0$.

   In sum

$$\hat{w}_i = -\left(|z_i| - \frac{\lambda\tau}{2} \right)_+ \tag{60}$$

We can combine these two cases to get that

$$\hat{w}_i = -\left(|z_i| - \frac{\lambda\tau}{2}\right)_+ sign(z_i) \tag{61}$$

we we call the SoftThreshold$(z_i, \frac{\lambda\tau}{2})$. The figure below shows how $\hat{w}_i$ depends on z_i.

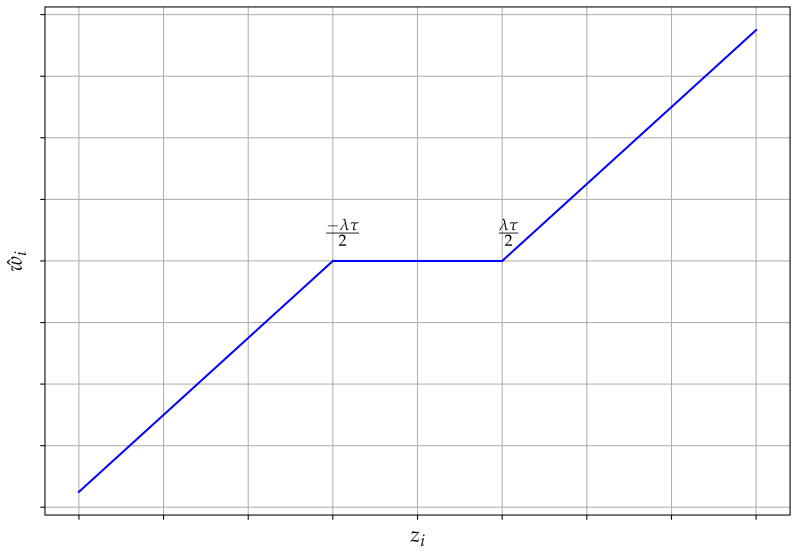

Figure 5: Soft Thresholding

12 Convexity and Support Vector Machines

12.1 Convexity

Example 11. (Convex function lies above tangent lines) Suppose $l(w)$ is a convex function. Then informally, $l(w) \geq$ tangent and w. More formally,

$$l(u) \geq l(w) + (u - w)^T \nabla l(w) \tag{62}$$

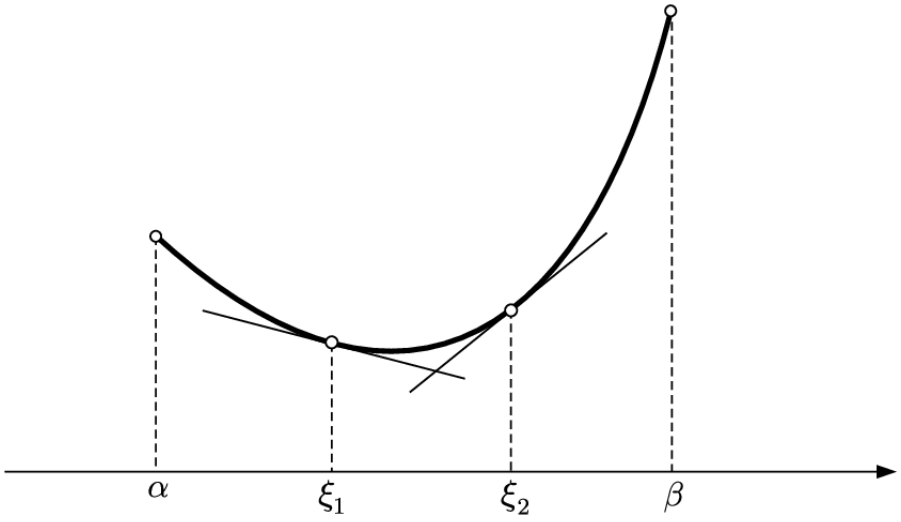

Figure 6: A convex function lies entirely above tangent lines

Definition 9. *(Subgradient) A subgradient v of a convex function l is as a vector satisfying*

$$l(u) \geq l(w) + (u - w)^T v \tag{63}$$

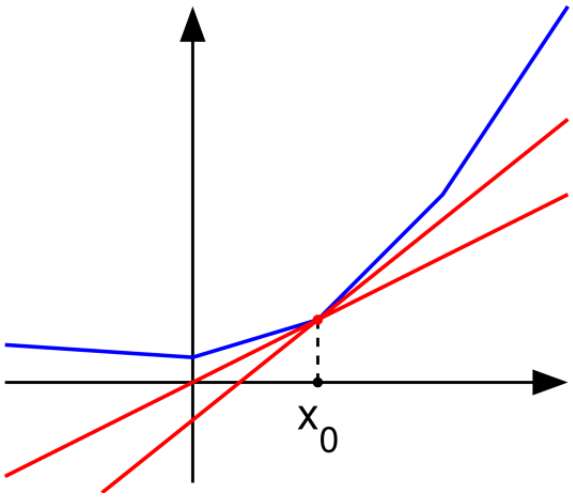

Figure 7: Subgradients

12.2 Support Vector Machines

When we use the classification rule $\hat{y} = sign(w^T x)$, our goal is thus to choose a w such that $\hat{y} = sign(w^T x)$ as often as possible. However, when we use least squares we do *not* actually minimize the number of mistakes. Minimizing the number of mistakes can be

written as minimizing the following sum of indicator variables:

$$\sum_{i=1}^{n} I_{y_i \neq sign(w^T x_i)} \tag{64}$$

LS actually minimizes

$$\sum_{i=1}^{n} (y_i - w^T x_i)^2 \tag{65}$$

We want to choose a convex function that mimics the ideal loss. We will use hinge loss, which is defined by

$$l(w) = \sum_{i=1}^{n} (1 - y_i x_i^T w)_+ \tag{66}$$

where

$$(a)_+ = \begin{cases} a & a > 0 \\ 0 & \text{otherwise} \end{cases} \tag{67}$$

Definition 10. *(Support Vector Machine) If we minimize*

$$\sum_{i=1}^{n} (1 - y_i x_i^T w)_+ + \lambda \|w\|_2^2 \tag{68}$$

this is called a support vector machine.

In the figure below, the black line is the ideal loss function (not convex). The green line is the squared loss function (LS), the blue line is hinge loss, and the red line is log loss.

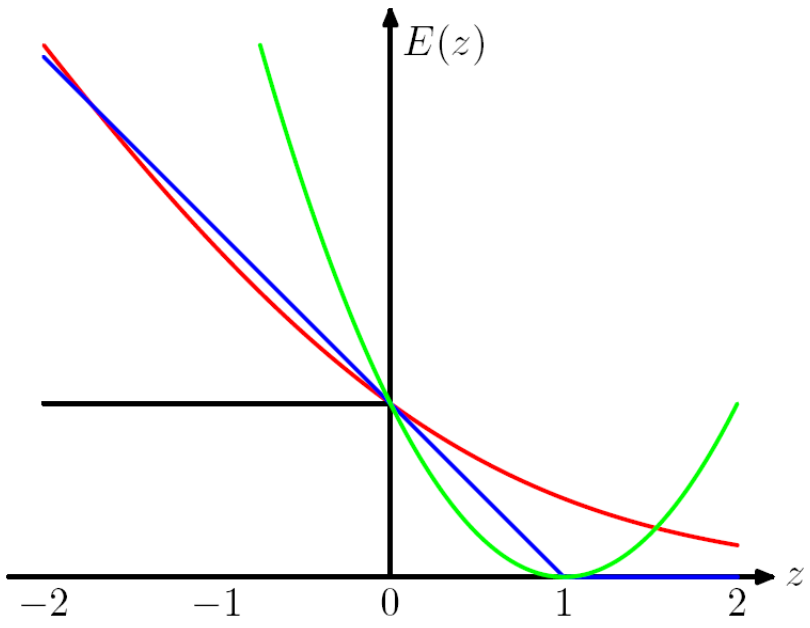

Figure 8: Loss Functions