Practical Learning Algorithm, Optimization and Duality

Based on lecture by Prof. Liwei Wang (Peking University)

October 21, 2025

1 Review

1.1 Generalization (Worst-Case Analysis)

The goal is to bound the probability of a "bad event." This event is defined as the existence of any function f in our hypothesis class $\mathcal{F}$ for which the true error is much larger than the empirical (training) error.

$$\mathbb{P}\left(\exists f \in \mathcal{F},\ \mathbb{P}(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[y_i \neq f(x_i)\right] \geq \epsilon\right)$$

Where:

- $\mathbb{P}(Y \neq f(X))$ is the true generalization error.
- $\hat{\mathbb{P}}(Y \neq f(X)) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[y_i \neq f(x_i)\right]$ is the empirical error on the training set.
- ϵ is the tolerance.

The lecture's goal is to find an upper bound for this probability.

1.2 VC Dimension (VC-dim)

The VC dimension of a hypothesis class $\mathcal{F}$, denoted $VC(\mathcal{F})$, is defined as d if:

1. **There exists** a set of d points that $\mathcal{F}$ can shatter:

$$\exists x_1, \ldots, x_d \quad \text{s.t.} \quad |\{(f(x_1), \ldots, f(x_d)) : f \in \mathcal{F}\}| = 2^d$$

2. **For all** sets of $d+1$ points, $\mathcal{F}$ cannot shatter them:

$$\text{and } \forall x_1, \ldots, x_{d+1}, \quad |\{(f(x_1), \ldots, f(x_{d+1})) : f \in \mathcal{F}\}| < 2^{d+1}$$

1.3 The Generalization Bound Theorem

The probability of the "bad event" is bounded by:

$$\mathbb{P}\left(\exists f \in \mathcal{F},\ \mathbb{P}(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[y_i \neq f(x_i)\right] \geq \epsilon\right) \leq 2\left(\frac{2e}{d}\right)^d n^d \cdot \exp\left(-\frac{1}{2}n\epsilon^2\right)$$

This is derived by combining Hoeffding's inequality (using a symmetrization trick) and Sauer's Lemma (which bounds the growth function $m_F(2n)$ by a polynomial in n, i.e., $O(n^d)$).

Finally, by setting the "bad event" probability bound equal to δ (our desired confidence level) and solving for ϵ. This gives a direct bound on the generalization error itself.

Theorem 1. Let $VC(\mathcal{F}) = d$. Then $\forall \delta > 0$, with probability at least $1 - \delta$:

$$\mathbb{P}(Y \neq f(X)) \leq \hat{\mathbb{P}}(Y \neq f(X)) + \sqrt{\frac{2}{n}\left(d \log \frac{2en}{d} + \log \frac{1}{\delta} + \log 2\right)}$$

This holds simultaneously for all $f \in \mathcal{F}$.

This theorem provides the "worst-case" generalization guarantee. It shows that (with high probability) the true error is close to the training error, and the "gap" (ϵ) between them is controlled by the VC dimension d and the number of samples n.

2 Practical Algorithm: The Linear Classifier

We now move from theoretical bounds to practical algorithms. We focus on the linear classifier.

- **Input Space:** $X \subseteq \mathbb{R}^d$
- **Output Space:** $y = \{\pm 1\}$
- **Hypothesis Class ($\mathcal{F}$):** $\mathcal{F} = \{f(x) = \text{sign}(w^\top x + b) \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}$

Given a training set $S = \{(x_i, y_i)\}_{i=1}^n$, our first task is to determine if a solution exists.

2.1 Feasibility and Linear Separability

A dataset S is **linearly separable** if there exist parameters w, b that fully separate the training data. This can be expressed as two conditions:

$$w^\top x_i + b \geq 0 \quad \text{for all } i \text{ such that } y_i = +1$$

$$w^\top x_i + b < 0 \quad \text{for all } i \text{ such that } y_i = -1$$

These can be elegantly combined into a single inequality:

$$y_i(w^\top x_i + b) \geq 0 \quad \text{for all } i \in [n]$$

To determine if such a w and b exist, we can formulate a **Linear Programming (LP)** problem. We try to find the maximum t that satisfies the constraints:

$$\begin{aligned} \max_{w,b,t} \quad & t \\ \text{s.t.} \quad & y_i(w^\top x_i + b) \geq t, \quad i = 1, \ldots, n. \end{aligned}$$

This LP can be solved efficiently (e.g., using Interior-Point Methods). If the optimal solution $t^* > 0$, the data is linearly separable.

However, this formulation has a flaw: it is **unbounded**. If (w, b) is a solution with margin t, then $(\alpha w, \alpha b)$ for any $\alpha > 0$ is also a solution with margin αt. As $\alpha \to \infty$, $t \to \infty$. This formulation can check for separability but cannot find a unique "best" classifier.

2.2 Finding the "Good" Classifier (Maximal Margin)

If the data is separable, we want to find a "good" classifier. The idea is to find the one that **maximizes the minimum distance** (the margin) between the data points and the decision boundary.

2.2.1 First Attempt: Non-Convex Formulation

To fix the unbounded issue, we can add a constraint to control the scale of w.

$$\begin{aligned} \max_{w,b,t} \quad & t \\ \text{s.t.} \quad & y_i(w^\top x_i + b) \geq t, \quad i = 1, \ldots, n \\ & \|w\| = 1 \end{aligned}$$

This formulation is meaningful, but the constraint $\|w\| = 1$ is **non-convex**, making the optimization problem difficult to solve efficiently.

2.2.2 The Solution: Reformulation via Scale Invariance

Our goal is to reformulate this into a **Convex Optimization** problem, which is efficiently solvable. We use the **scale invariance** of the problem.

Instead of fixing the norm of w, we fix the margin. Since we assume the data is separable, $t > 0$. We can scale w and b such that the margin $t = 1$. Start with the constraint: $y_i(w^\top x_i + b) \geq t$. Divide by t:

$$y_i \left(\left(\frac{w}{t} \right)^\top x_i + \frac{b}{t} \right) \geq 1$$

Let $w' = w/t$ and $b' = b/t$. Our original problem $\max t$ (with $\|w\| = 1$) is equivalent to:

$$\max \frac{1}{\|w'\|} \quad \text{s.t.} \quad y_i(w'^\top x_i + b') \geq 1$$

Maximizing $\frac{1}{\|w'\|}$ is equivalent to minimizing $\|w'\|$, which is in turn equivalent to minimizing $\frac{1}{2}\|w'\|^2$ (the $\frac{1}{2}$ and square are added for mathematical convenience, simplifying the derivative).

2.2.3 The Primal Problem (QP)

By dropping the primes, we arrive at the final **Primal Problem**:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & y_i(w^\top x_i + b) \geq 1, \quad i = 1, \ldots, n. \end{aligned}$$

This is a **Quadratic Program (QP)** because the objective is quadratic and the constraints are linear. A QP is a specific type of convex optimization problem and can be solved efficiently.

3 Preamble to Duality: Minimax Theory

To understand how to solve this QP, we first introduce the prerequisite knowledge of Lagrange Duality, which is built upon the Minimax Theorem.

3.1 The Two-Player Zero-Sum Game

We consider a two-player, zero-sum, one-shot matrix game.

- **Players:** Alice (chooses row i) and Bob (chooses column j).
- **Payoff Matrix:** A matrix M, where M_{ij} is the amount Alice pays to Bob.
- **Zero-Sum:** One player's gain is the other's loss, i.e., $M_{ij} + \tilde{M}_{ij} = 0$.
- **Rationality:** Alice wants to $\min M_{ij}$, Bob wants to $\max M_{ij}$.

3.2 Pure Strategy and the Minimax Lemma

A **Pure Strategy** is a deterministic choice of action. We analyze the two possible orders of play.

1. **Alice moves first (Minimax):**
   - Alice chooses a row i.
   - Bob observes i and rationally chooses j to maximize his payoff: $\max_j M_{ij}$.
   - Alice, knowing this, will choose i to minimize her maximum loss.
   - The value of the game is: $\min_i \max_j M_{ij}$.
2. **Bob moves first (Maximin):**
   - Bob chooses a column j.
   - Alice observes j and rationally chooses i to minimize her payment: $\min_i M_{ij}$.
   - Bob, knowing this, will choose j to maximize his minimum gain.
   - The value of the game is: $\max_j \min_i M_{ij}$.

Comparing these two values gives the fundamental **Minimax Lemma (Weak Duality)**:

$$\max_j \min_i M_{ij} \leq \min_i \max_j M_{ij}$$

This inequality implies that in a pure-strategy game, moving second is always advantageous (or equal).

3.3 Mixed Strategy and von Neumann's Theorem

A **Mixed Strategy** is a randomized algorithm, where players choose a probability distribution over their actions.

- Alice chooses a probability distribution p over the rows.
- Bob chooses a probability distribution q over the columns.
- The expected payoff from Alice to Bob is $p^\top Mq$.

The game values are now defined over these distributions:

- Alice first (Minimax): $\min_p \max_q p^\top Mq$
- Bob first (Maximin): $\max_q \min_p p^\top Mq$

The weak duality lemma still holds for mixed strategies (as proven in the lecture):

$$\max_q \min_p p^\top Mq \leq \min_p \max_q p^\top Mq$$

However, the key result, **von Neumann's Minmax Theorem**, states that for mixed strategies in a two-player, zero-sum game, this inequality becomes an equality (**Strong Duality**):

$$\max_q \min_p p^\top Mq = \min_p \max_q p^\top Mq$$

(For proof, see Appendix)

This theorem establishes that a stable equilibrium value exists and removes the first/second-player advantage. This concept of switching the order of min and max is the foundation for applying Lagrange Duality to our optimization problem.

4 Appendix:

4.1 The VC-dimension of the linear classifier set $\mathcal{F} = \{\mathbf{sgn}(w^T x + b) \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}$

To prove that the VC-dimension of the linear classifier set $\mathcal{F} = \{\text{sgn}(w^T x + b) \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}$ is $d+1$, we need to prove it in two steps:

4.1.1 Step 1: Prove that $d+1$ points can be shattered

There exists $d+1$ points $x_1, x_2, \ldots, x_{d+1} \in \mathbb{R}^d$, construct the augmented vector $x_i' = \begin{pmatrix} 1 \\ x_i \end{pmatrix} \in \mathbb{R}^{d+1}$ (where $x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{id} \end{pmatrix}$), and let the augmented weight vector $w' = \begin{pmatrix} b \\ w \end{pmatrix} \in \mathbb{R}^{d+1}$ (where $w = \begin{pmatrix} w_1 \\ \vdots \\ w_d \end{pmatrix}$).

At this point, the classification function can be rewritten as:

$$\text{sgn}(w^T x + b) = \text{sgn}\left(w'^T x'\right)$$

Construct the matrix $X' = \begin{pmatrix} x_1' & x_2' & \ldots & x_{d+1}' \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}$. Since $x_1, \ldots, x_{d+1}$ are $d+1$ points in d-dimensional space, X' can be constructed to be full-rank (i.e., invertible).

For any label $y = (y_1, y_2, \ldots, y_{d+1}) \in \{-1, +1\}^{d+1}$, solve the linear system of equations:

$$w'^T X' = y$$

That is, $w' = yX'^{-1}$ (since X' is invertible, the solution exists). This shows that there exist w and b such that $\text{sgn}(w^T x_i + b) = y_i$ for all $i = 1, 2, \ldots, d+1$, so $d+1$ points can be shattered.

4.1.2 Step 2: Prove that $d+2$ points cannot be shattered

Consider $d+2$ points $x_1, x_2, \ldots, x_{d+2} \in \mathbb{R}^d$, whose augmented vectors are $x_1', x_2', \ldots, x_{d+2}' \in \mathbb{R}^{d+1}$.

Since any $d+2$ vectors in $\mathbb{R}^{d+1}$ must be linearly dependent, there exist coefficients $a_1, a_2, \ldots, a_{d+1} \in \mathbb{R}$ (not all zero) such that:

$$x_{d+2}' = \sum_{i=1}^{d+1} a_i x_i'$$

Suppose there exists $w' \in \mathbb{R}^{d+1}$ such that for $i = 1, 2, \ldots, d+1$, $\text{sgn}(w'^T x_i') = y_i$ ($y_i \in \{-1, +1\}$). Consider $w'^T x_{d+2}'$:

$$w'^T x_{d+2}' = \sum_{i=1}^{d+1} a_i w'^T x_i'$$

If for all $i \leq d+1$, a_i has the same sign as $w'^T x_i'$ (i.e., $a_i w'^T x_i' > 0$), then $w'^T x_{d+2}' > 0$, corresponding to the label $+1$; if there is a sign conflict, the label can also be deduced to be uniquely determined and cannot be freely assigned. Therefore, $d+2$ points cannot be shattered.

In conclusion, the VC-dimension of the linear classifier set $\mathcal{F}$ is $d+1$.

4.2 Proof of Minimax Lemma

4.2.1 Pure Strategy

We want to prove the "Weak Duality" lemma for pure strategies:

$$\max_j \min_i M_{ij} \leq \min_i \max_j M_{ij}$$

Proof. Let $v_{maximin} = \max_j \min_i M_{ij}$ and $v_{minimax} = \min_i \max_j M_{ij}$.

Let j^* be the column that achieves the maximin value: $v_{maximin} = \max_j \min_i M_{ij} = \min_i M_{ij^*}$

Let i^* be the row that achieves the minimax value: $v_{minimax} = \min_i \max_j M_{ij} = \max_j M_{i^*j}$

Now, consider the element $M_{i^*j^*}$.

1. By the definition of $v_{maximin}$, $v_{maximin}$ is the minimum value in column j^*. Therefore, $v_{maximin} \leq M_{i^*j^*}$ (since $M_{i^*j^*}$ is one of the elements in that column).

2. By the definition of $v_{minimax}$, $v_{minimax}$ is the maximum value in row i^*. Therefore, $v_{minimax} \geq M_{i^*j^*}$ (since $M_{i^*j^*}$ is one of the elements in that row).

Combining these two inequalities, we get:

$$v_{maximin} \leq M_{i^*j^*} \leq v_{minimax}$$

Thus, $\max_j \min_i M_{ij} \leq \min_i \max_j M_{ij}$. □

4.2.2 Mixed Strategy

We now prove the weak duality for mixed strategies, which states:

$$\max_q \min_p p^\top Mq \leq \min_p \max_q p^\top Mq$$

Proof. Let p_0 and q_0 be any arbitrary pair of probability distributions (strategies) for Alice and Bob, respectively.

1. By definition, $\max_q p^\top Mq$ gives the best possible payoff for Bob (max) against a fixed Alice strategy p. Therefore, for a fixed p_0, the payoff against an arbitrary q_0 can be no better than the payoff against the optimal q:
$$p_0^\top Mq_0 \leq \max_q p_0^\top Mq$$

2. Similarly, $\min_p p^\top Mq$ gives the best possible payoff for Alice (min) against a fixed Bob strategy q. Therefore, for a fixed q_0, the payoff against an arbitrary p_0 can be no worse (for Alice) than the payoff against the optimal p:
$$p_0^\top Mq_0 \geq \min_p p^\top Mq_0$$

Combining these two inequalities, we have for any p_0, q_0:

$$\min_p p^\top Mq_0 \leq p_0^\top Mq_0 \leq \max_q p_0^\top Mq$$

This directly implies that for any p_0, q_0:

$$\min_p p^\top Mq_0 \leq \max_q p_0^\top Mq$$

Let $L(q) = \min_p p^\top Mq$ and $R(p) = \max_q p^\top Mq$. The inequality above is $L(q_0) \leq R(p_0)$ for all p_0, q_0.

This means that any value of $L(q)$ is a lower bound for all values of $R(p)$. Therefore, the maximum possible value of $L(q)$ must be less than or equal to the minimum possible value of $R(p)$.

$$\max_{q_0} L(q_0) \leq \min_{p_0} R(p_0)$$

Substituting the definitions of $L(q)$ and $R(p)$ back, we get:

$$\max_q \min_p p^\top Mq \leq \min_p \max_q p^\top Mq$$

□

4.3 Proof of Sauer's Lemma

We want to prove the Sauer' s Lemma, i.e. $\forall 0 < d < n$:

$$\sum_{k=0}^{d} \binom{n}{k} \leq \left(\frac{en}{d}\right)^d$$

Proof.

$$\begin{aligned}
\sum_{k=0}^{d} \binom{n}{k} &= \left(\frac{n}{d}\right)^d \sum_{k=0}^{d} \left(\frac{d}{n}\right)^d \binom{n}{k} \\
&\leq \left(\frac{n}{d}\right)^d \sum_{k=0}^{d} \left(\frac{d}{n}\right)^k \binom{n}{k} \\
&\leq \left(\frac{n}{d}\right)^d \sum_{k=0}^{n} \left(\frac{d}{n}\right)^k \binom{n}{k} = \left(\frac{n}{d}\right)^d \left(\left(1+\frac{d}{n}\right)^{\frac{n}{d}}\right)^d \leq \left(\frac{en}{d}\right)^d
\end{aligned}$$

□