Outline

- Weak Learnability = Linear separability
  - Follows directly from Von-Neumann’s minimax theorem

- Relaxations
  - The equivalence yields a family of relaxations to the separability assumption
  - Proof technique: Fenchel duality & Infimal Convolution

- Boosting Algorithms
  - A primal-dual algorithm
  - Applicable to entire family of relaxations
  - Rate of convergence analysis
Input:

- \( m \) training examples \((x_1, y_1), \ldots, (x_m, y_m)\)
- \( n \) base hypotheses \( h_1, \ldots, h_n \)
Input:

- $m$ training examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$
- $n$ base hypotheses $h_1, \ldots, h_n$

$$A = \begin{pmatrix} y_1 h_1(\mathbf{x}_1) & \cdots & y_1 h_n(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ y_m h_1(\mathbf{x}_m) & \cdots & y_m h_n(\mathbf{x}_m) \end{pmatrix}$$
Boosting

Input:

- $m$ training examples $(x_1, y_1), \ldots, (x_m, y_m)$
- $n$ base hypotheses $h_1, \ldots, h_n$

$$A = \begin{pmatrix}
  y_1 h_1(x_1) & \cdots & y_1 h_n(x_1) \\
  \vdots & \ddots & \vdots \\
  y_m h_1(x_m) & \cdots & y_m h_n(x_m)
\end{pmatrix}$$

Output:

- ‘strong’ hypothesis $H_w(\cdot) = \sum_{i=1}^{n} w_i h_i(\cdot)$
**Weak Learnability**

**Definition: $\gamma$-weak-learnability**

A matrix $A$ is $\gamma$-weak-learnable if

$$\gamma = \min_{d \in \mathbb{S}^m} \max_{j \in [n]} |(d^\dagger A)_j|.$$
Schapire: Weak learnability implies separability
Schapire: Weak learnability implies separability
**Schapire**: Weak learnability implies separability

- **Weak learnability**: Convex hull of rows of A does not contain the origin
**Schapire:** Weak learnability implies separability

- **Weak learnability:** Convex hull of rows of $A$ does not contain the origin
- **Separability:** Exists hyperplane that goes through origin s.t. all rows of $A$ resides in one side
Weak-to-strong learnability

**Schapire**: Weak learnability implies separability

- **Weak learnability**: Convex hull of rows of $A$ does not contain the origin
- **Separability**: Exists hyperplane that goes through origin s.t. all rows of $A$ resides in one side

Quantification?
Definition: separability with $\ell_1$ margin $\gamma$

A matrix $A$ is linearly separable with $\ell_1$ margin $\gamma$ if

$$
\gamma = \max_{\mathbf{w} \in \mathbb{B}_1^n} \min_{i \in [m]} (A\mathbf{w})_i
$$

unit $\ell_1$ ball (weights over features)

'margin' of $i^{th}$ example

$$(A\mathbf{w})_i = y_i \sum_j w_j h_j(x_i)$$
The following properties are equivalent:

- matrix $A$ is $\gamma$-weak-learnable
- matrix $A$ is linearly separable with $\ell_1$ margin of $\gamma$

In other words,

$$\max_{\mathbf{w} \in \mathbb{B}_1^n} \min_{i \in [m]} (A\mathbf{w})_i = \min_{\mathbf{d} \in \mathbb{S}^m} \max_{j \in [n]} |(d^\dagger A)_j|$$

**Proof:** Equivalence follows from Von-Neumann’s minimax theorem
\[
\max_{\mathbf{w} \in \mathbb{B}_1^n} \min_{i \in [m]} (A\mathbf{w})_i = \min_{\mathbf{d} \in \mathbb{S}_m^n} \max_{j \in [n]} |(d^\dagger A)_j|
\]
\[
\max_{w \in \mathbb{B}_1^n} \min_{i \in [m]} (A w)_i = \min_{d \in \mathbb{S}^m} \max_{j \in [n]} |(d^\top A)_j|
\]
Relaxations -- Main Idea

\[
\max_{w \in B_1^n} \min_{i \in [m]} (Aw)_i = \min_{d \in S^m} \max_{j \in [n]} |(d^\top A)_j|
\]

\[
\max_{w \in B_1^n} \min_{i \in [m]} (AW)_i = \min_{d \in S^m \cap C} \max_{j \in [n]} |(d^\top A)_j|
\]

Relaxed weak-learnability
Relaxations -- Theorem

Hard margin

\[
\max_{w \in \mathbb{B}^n_1} \min_{i \in [m]} (Aw)_i = \min_{d \in \mathbb{S}^m} \max_{j \in [n]} |(d^\top A)_j|
\]

Soft margin

\[
\max_{w \in \mathbb{B}^n_1} \gamma - \nu \| \gamma - Aw \|_* = \min_{d \in \mathbb{S}^m \cap C} \max_{j \in [n]} |(d^\top A)_j|
\]

\[C = \{ w : \| w \| \leq \nu \}\]
Relaxations -- Theorem

**Hard Margin**

$$\max_{w \in \mathbb{B}_1^n} \min_{i \in [m]} (Aw)_i = \min_{d \in S^m} \max_{j \in [n]} \left| (d^\dagger A)_j \right|$$

**Soft Margin**

$$\max_{w \in \mathbb{B}_1^n} \gamma - \nu \| [\gamma - Aw]_+ \|_* = \min_{d \in S^m \cap C} \max_{j \in [n]} \left| (d^\dagger A)_j \right|$$

$$C = \{ w : \| w \| \leq \nu \}$$
If $C = \{ w : \|w\|_\infty \leq \frac{1}{k} \}$ soft margin is:

$$\max_{\gamma} \gamma - \frac{1}{k} \| [\gamma - A w]_+ \|_1 = \text{AvgMin}_k (A w)$$
If $C = \{ w : \| w \|_\infty \leq \frac{1}{k} \}$ soft margin is:

$$\max_{\gamma} \gamma - \frac{1}{k} \| [\gamma - A w]_+ \|_1 = \text{AvgMin}_k(A w)$$
Proof Technique

\[
\min_{d \in S^m \cap C} \max_i |(d^t A)_i| = \min_d f(d) + g(d^t A) = \max_{w} -f^*(-A w) - g^*(w)
\]

**supp**($S^m \cap C$) \quad || \cdot || \infty

Fenchel Duality

- In our case:
  - \( g^*(\cdot) = \text{supp}(B^n_1) \)
  - The tricky part is to show that
    \[
    f^*(\theta) = - \max_{\gamma \in \mathbb{R}} (\gamma - \nu \| [\gamma + \theta]_+ \|_*)
    \]
  - We show that using **infimal convolution** theory
    \[
    f_1^* + f_2^* = (f_1 \otimes_{\text{inf}} f_2)^*
    \]
**A Boosting Algorithm**

**Initialize:** \( w_1 = 0, \quad \beta = \frac{\epsilon}{2 \log(m)} \)

**For** \( t = 1, 2, \ldots, T \)

\[
 d_t = \arg\min_{d \in \mathbb{S}^m \cap C} D_{KL}(d, \hat{d}) \quad \text{where} \quad \hat{d}_{t,i} \propto \exp\left(-\frac{1}{\beta} (A w_t)_i \right)
\]

\[
 j_t \in \arg\max_j |(d_t^\dagger A)_j|
\]

\[
 \eta_t = \max \left\{ 0, \min \left\{ 1, \frac{\beta d_t^\dagger A(e_j - w_t)}{\|A(e_j - w_t)\|_\infty^2} \right\} \right\}
\]

\[
 w_{t+1} = (1 - \eta_t) w_t + \eta_t e_j
\]
A Boosting Algorithm

**Initialize:** \( w_1 = 0, \quad \beta = \frac{\epsilon}{2 \log(m)} \)

For \( t = 1, 2, \ldots, T \)

\[
d_t = \arg\min_{d \in S^m \cap C} D_{\text{KL}}(d, \hat{d}) \quad \text{where} \quad \hat{d}_{t,i} \propto \exp\left(-\frac{1}{\beta} (A w_t)_i \right)
\]

\[
j_t \in \arg\max_j |(d_t^\dagger A)_j|
\]

\[
\eta_t = \max \left\{ 0, \min \left\{ 1, \frac{\beta d_t^\dagger A(e^{j,t} - w_t)}{\|A(e^{j,t} - w_t)\|_\infty^2} \right\} \right\}
\]

\[
w_{t+1} = (1 - \eta_t)w_t + \eta_t e^{j,t}
\]
A Boosting Algorithm

**Initialize:** $w_1 = 0$, $\beta = \frac{\epsilon}{2 \log(m)}$

For $t = 1, 2, \ldots, T$

$$d_t = \operatorname{argmin}_{d \in \mathbb{S}^m \cap C} D_{\text{KL}}(d, \hat{d})$$

where $\hat{d}_{t,i} \propto \exp\left(-\frac{1}{\beta} (A w_t)_i\right)$

$$j_t \in \operatorname{argmax}_j |(d_t^\dagger A)_j|$$

$$\eta_t = \max\left\{0, \min\left\{1, \frac{\beta d_t^\dagger A(e^{j_t} - w_t)}{\|A(e^{j_t} - w_t)\|_\infty^2}\right\}\right\}$$

$$w_{t+1} = (1 - \eta_t)w_t + \eta_t e^{j_t}$$
A Boosting Algorithm

**INITIALIZE:** \( w_1 = 0, \quad \beta = \frac{\epsilon}{2 \log(m)} \)

**FOR** \( t = 1, 2, \ldots, T \)

\[
\begin{align*}
    d_t &= \text{argmin}_{\hat{d} \in S^m \cap C} D_{KL}(d, \hat{d}) \quad \text{where} \quad \hat{d}_{t,i} \propto \exp \left( -\frac{1}{\beta} (A w_t)_i \right) \\
    j_t &= \text{arg max}_j |(d_t^\dagger A)_j| \\
    \eta_t &= \max \left\{ 0, \min \left\{ 1, \frac{\beta d_t^\dagger A(e^{j_t} - w_t)}{\|A(e^{j_t} - w_t)\|_\infty^2} \right\} \right\} \\
    w_{t+1} &= (1 - \eta_t) w_t + \eta_t e^{j_t}
\end{align*}
\]
A Boosting Algorithm

**Initialize:** \( w_1 = 0, \ \beta = \frac{\epsilon}{2 \log(m)} \)

For \( t = 1, 2, \ldots, T \)

\[
d_t = \arg\min_{\hat{d} \in S^m \cap C} D_{\text{KL}}(d, \hat{d}) \quad \text{where} \quad \hat{d}_{t,i} \propto \exp\left(-\frac{1}{\beta}(A w_t)_i\right)
\]

\[
j_t \in \arg\max_j |(d_t^\dagger A)_j|
\]

\[
\eta_t = \max \left\{ 0, \min \left\{ 1, \frac{\beta d_t^\dagger A(e_j - w_t)}{\|A(e_j - w_t)\|_\infty^2} \right\} \right\}
\]

\[
w_{t+1} = (1 - \eta_t)w_t + \eta_t e_j
\]
A Boosting Algorithm

**INITIALIZE:** \( w_1 = 0, \quad \beta = \frac{\epsilon}{2 \log(m)} \)

**FOR** \( t = 1, 2, \ldots, T \)

\[ d_t \leftarrow \arg \min_{d \in S^m \cap C} D_{\text{KL}}(d, \hat{d}) \text{ where } \hat{d}_{t,i} \propto \exp \left( -\frac{1}{\beta} (A w_t)_i \right) \]

\[ j_t \in \arg \max_j |(d_t^\dagger A)_j| \]

\[ \eta_t = \max \left\{ 0, \min \left\{ 1, \beta \frac{d_t^\dagger A(e^{jt} - w_t)}{\|A(e^{jt} - w_t)\|_\infty^2} \right\} \right\} \]

\[ w_{t+1} = (1 - \eta_t) w_t + \eta_t e^{jt} \]
A Boosting Algorithm

**INITIALIZE:** \( w_1 = 0, \quad \beta = \frac{\epsilon}{2 \log(m)} \)

**FOR** \( t = 1, 2, \ldots, T \)

\[
d_t = \arg\min_{d \in \mathbb{S}^m \cap C} D_{KL}(d, \hat{d}) \quad \text{where} \quad \hat{d}_{t,i} \propto \exp\left( -\frac{1}{\beta} (A w_t)_i \right)
\]

\[
\eta_t = \max\left\{0, \min\left\{1, \frac{\beta d_t^\dagger A(e^{j_t} - w_t)}{\|A(e^{j_t} - w_t)\|_\infty^2} \right\} \right\}
\]

\[
w_{t+1} = (1 - \eta_t)w_t + \eta_t e^{j_t}
\]
A Boosting Algorithm

**INITIALIZE:** \( w_1 = 0, \quad \beta = \frac{\epsilon}{2 \log(m)} \)

**FOR** \( t = 1, 2, \ldots, T \)

\[
d_t = \arg\min_{\hat{d} \in \mathbb{S}^m \cap C} D_{KL}(d, \hat{d}) \quad \text{where} \quad \hat{d}_{t,i} \propto \exp\left(-\frac{1}{\beta}(A w_t)_i\right)
\]

\[
j_t \in \arg \max_j |(d_t^\dagger A)_j|
\]

\[
\eta_t = \max\left\{0, \min\left\{1, \frac{\beta d_t^\dagger A(e^{j_t} - w_t)}{\|A(e^{j_t} - w_t\|_2^2}\right\}\right\}
\]

\[
w_{t+1} = (1 - \eta_t)w_t + \eta_t e^{j_t}
\]
**Theorem**

- For any $m \times n$ matrix $A$ over $[-1, 1]$.
- For any relaxation set $C = \{d : \|d\| \leq \nu\}$.
- The number of iterations required by the algorithm to find an $\epsilon$-accurate solution is

$$T \leq O\left(\frac{\log(m)}{\epsilon^2}\right)$$

**Remarks:**

- Matches rate of AdaBoost$_*$ [RW05] and SoftBoost [WLR06].
- Also bounds the sparseness of solution.
Proof Technique

- **Step 1:** If loss function has $\beta$ Lipschitz continuous derivative:

  $$\epsilon_t - \epsilon_{t+1} \geq \eta \epsilon_t - \frac{2 \eta^2}{\beta} \implies \epsilon_t \leq \frac{8}{\beta(t+1)}$$

- Proof uses duality

- **Step 2:** Approximate any 'soft-margin' loss by 'nicely behaved' loss
  
  - Domain of conjugate of the loss is a subset of the simplex
  - Add a bit relative entropy
  - Use infimal convolution theorem
The most expensive operations are the Entropic projection on $C$ and the call to weak learner.

For $C = \{w : \|w\|_{\infty} \leq \nu\}$ projection can be performed in $O(m)$.

The trick: similar to median search.

Proof can be extended to approximated weak learners.
Summary

- Weak Learnability = Linear Separability
- Relaxing separability using relaxed weak learnability
- ‘Algorithmic’ relaxations

Current and Future Work

- Use equivalence for generalization bounds?
- Other intuitive relaxations
- Other algorithmic relaxations
- Relation between $L_1$ and sparsity in a more general setting