Margin-Based Convergence Results

- **Margin:**
  \[ F = \{ f : X \to R \} \text{ is a class of functions } f \in F \]
  \[ S = \{ x_1, \ldots, x_m \}, x_i \in X, \text{ is a sample} \]
  \[ y_1, \ldots, y_m \text{ are the labels } y_i \in \{-1, 1\} \]
  \[ m_S(f) = \min_i \gamma_i = \min_i y_i f(x_i) \]

- **Theorem 1:** Let \( F = \{ f : X \to [-M, M] \} \), hypothesis class \( H = \{ sign(f) \mid f \in F \} \), \( d = Fat_F(\frac{2}{\gamma}) \), and fix any \( \gamma > 0 \).
  
  For all distributions \( D \), parameters \( \epsilon, \delta \):
  
  if we take IID sample with \( m \) examples and find hypothesis \( h = \text{sign}(f) \) with \( m_S(f) \geq \gamma \),
  then with probability at least \( 1 - \delta \),
  \[ \epsilon_{err}(h) \leq \frac{2}{m} \left( \log \frac{4}{\delta} + d \log \left( \frac{16eM}{d\gamma} \right) \log \left( \frac{128M^2}{\gamma^2} \right) \right) \].

- **Notation:**
  \[ f \in F, h = \text{sign}(f), S = \{ x_1, \ldots, x_m \} \]
  \[ err(f) = \Pr_{(x,y) \in D}[h(x) \neq y] \]
  \[ err_S(f) = \frac{1}{m} | \{ x_i | h(x_i) \neq y_i \} | \]

- **Probability Experiment 1:** Draw IID sample \( S_1 \) of \( m \) examples.
- **Event A:** \( \exists \) hypothesis \( h = \text{sign}(f) \) s.t. \( err_{S_1}(f) = 0, \ epsilon(f) \geq \epsilon, m_{S_1}(f) \geq \gamma \).

- **Probability Experiment 2:** Draw two IID samples \( S_1, S_2 \) each of \( m \) examples.
- **Event B:** \( \exists \) hypothesis \( h = \text{sign}(f) \) s.t. \( err_{S_1}(f) = 0, err_{S_2}(f) \geq \frac{m}{\gamma}, m_{S_1}(f) \geq \gamma \).

- **Claim 1:** If \( m \geq \frac{8}{\epsilon} \ln 2 \), then
  \[ Pr[A] \leq 2Pr[B] \].

- **Fix G a \( \frac{1}{2} \)-cover of \( F \) w.r.t. \( S_1, S_2 \).**
- **Let num_S(g, \gamma) be the number of examples in sample \( S \) for which \( g \) has margin \( < \gamma \).**
- **Event C:** \( \exists g \in G \) s.t. \( err_{S_1}(g) = 0, \]
  \[ \text{num}_{S_2}(g, \gamma) \geq \frac{m}{\gamma}, m_{S_1}(g) \geq \gamma \].

- **Claim 2:** Event B implies event C and therefore \( Pr[A] \leq 2Pr[C] \).
• Probability Experiment 3:
  - Draw IID sample $S$ of $2m$ examples.
  - Randomly permute the order of examples.
  - Let $S_1$ be first $m$ examples and $S_2$ the rest.
• Fact: same distribution as in experiment 2.

• Probability Experiment 4:
  (parameterized by sample $S$ of $2m$ examples)
  - Randomly permute the order of examples.
  - Let $S_1$ be first $m$ examples and $S_2$ the rest.

• Event $C'(S)$: same as $C$ but for fixed $S$ and experiment 4.

• Event $C''(S,g)$: same as $C'(S)$ but for fixed $g \in G$.

• Claim 3: $Pr[C] \leq \max_S Pr[C'(S)]$.

• Claim 4: $Pr[C'(S)] \leq N(F,2m,\frac{\gamma}{2})\left(\frac{4m}{\gamma}\right)$.

• Proof of Theorem 1:
  $Pr[A] \leq 2N(F,2m,\frac{\gamma}{2})\left(\frac{4m}{\gamma}\right) \leq \ldots \leq \delta$
  To get the last bound we apply Theorem 0 with $\gamma \leftarrow \frac{2}{\gamma}$, $m \leftarrow 2m$, $(b-a) \leftarrow 2M$, $d \leftarrow \text{Fat}_F(\frac{\gamma}{2})$.

  Alternatively solve for $\varepsilon$:
  \[
  \frac{\gamma n}{2} = \log \frac{2}{\delta} + \log N(F,2m,\frac{\gamma}{2}),
  \]
  \[
  \varepsilon = \frac{2}{m} \left( \log \frac{2}{\delta} + \log N(F,2m,\frac{\gamma}{2}) \right),
  \]
  and using the same values in Theorem 0:
  \[
  \varepsilon = \frac{2}{m} \left( \log \frac{4}{\delta} + d \log \left( \frac{16emM}{d\gamma} \right) \log \left( \frac{128mM^2}{\gamma^2} \right) \right).
  \]

• Theorem 2:
  For $X = \{ z \in R^n \mid \|z\| \leq M \}$
  $F = \{ f = w \cdot x \mid w \in R^n, \|w\| = 1 \}$
  we have $d = \text{Fat}_F(\gamma) \leq \left( \frac{M}{\gamma} \right)^2$.

• Corollary 3:
  (specialize theorem 1 for linear separators)
  \[
  \text{error}(h) \leq \frac{2}{m} \left( \log \frac{4}{\delta} + \frac{64M^2}{\gamma^2} \log \left( \frac{16em\gamma}{64M} \right) \log \left( \frac{512mM^2}{\gamma^2} \right) \right).
  \]

• Theorem 4:
  (stronger result w/o fixing $\gamma$)
  if algorithm finds hypothesis $h = \text{sign}(w \cdot x)$ with $m_S(w) = \gamma \geq \frac{M}{2\sqrt{m}}$
  then with probability at least $\delta$, $\text{error}(h) \leq \frac{2}{m} \left( \sqrt{m} + \log \frac{4}{\delta} + \frac{256M^2}{\gamma^2} \log \left( \frac{16em\gamma}{64M} \right) \log \left( \frac{512mM^2}{\gamma^2} \right) \right)$.

• Theorem 5:
  (variant of theorem 4)
  if algorithm finds hypothesis $h = \text{sign}(w \cdot x)$ with $m_S(w) = \gamma \geq \frac{M}{m}$
  then with probability at least $\delta$, $\text{error}(h) \leq \frac{2}{m} \left( \log(m) + \log \frac{4}{\delta} + \frac{256M^2}{\gamma^2} \log \left( \frac{16em\gamma}{64M} \right) \log \left( \frac{512mM^2}{\gamma^2} \right) \right)$.
• Theorem 6:
  (compression based convergence - lecture 8)
  \[ \text{error}(h) \leq \frac{1}{m-d}(d \ln\left(\frac{e^m d}{d}\right) + \ln\frac{1}{\delta}) \]
  Perceptron is a compression scheme with compression size \(d = \text{the mistake bound} \)
  \[ d = \frac{M^2}{\gamma^2} \]

• Theorem 7:
  (compression based bound for Perceptron)
  \[ \text{error}(h) \leq \frac{2}{m}\left(\frac{M^2}{\gamma^2} \ln\left(\frac{e^m \gamma^2}{M^2}\right) + \ln\frac{1}{\delta}\right) \]

• Compare this to the bound from theorem 5:
  \[ \frac{2}{m}\left(\log(m) + \log\frac{4}{\delta} + \frac{256M^2}{\gamma^2} \log\left(\frac{16em\gamma}{64M}\right) \log\left(\frac{512mM^2}{\gamma^2}\right)\right) \]

• Non-Separable Data:
  For \(w\) and \(\gamma\) and define \(\zeta = (\zeta_1, \ldots, \zeta_m)\)
  \[ \zeta_i = \max\{0, \gamma - y_i(w \cdot x_i - \theta)\} \]

Theorem 8:
  (convergence in non-separable case)
  if algorithm finds hypothesis \(w\) with slack vector \(\zeta\) w.r.t. \(\gamma\)
  then can use bound of corollary 3:
  \[ \text{error}(h) \leq \frac{2}{m}\left(\log\frac{4}{\delta} + \frac{64M^2}{\gamma^2} \log\left(\frac{16em\gamma}{64M}\right) \log\left(\frac{128mM^2}{\gamma^2}\right)\right) \]

with \(\hat{\gamma} = \frac{\gamma}{\sqrt{1+||\zeta||^2}}, \hat{M} = \sqrt{M^2 + 1}.\)