First, we can prove that the training error goes down. If we write the error at time \( t \) as \( \frac{1}{2} - \gamma_t \),

\[
\hat{R}(h) \leq \exp \left\{ -2 \sum_t \gamma_t^2 \right\} \tag{1}
\]

- If \( \forall t : \gamma_t \geq \gamma > 0 \), then \( \hat{R}(h) \leq \exp \left\{ -2 \gamma^2 T \right\} \)

**AdaBoost**: do not need \( \gamma \) or \( T \) a priori
Training Error Proof: Preliminaries

Repeatedly expand the definition of the distribution.

\[
D_{t+1}(i) = \frac{D_t(i) \exp \{-\alpha_t y_i h_t(x_i)\}}{Z_t}
\]

\[
D_{t-1}(i) \exp \{-\alpha_{t-1} y_i h_{t-1}(x_i)\} \exp \{-\alpha_t y_i h_t(x_i)\}
\]

\[
= \frac{Z_{t-1} Z_t}{m \prod_{s=1}^{t} Z_s}
\]

\[
\exp \{-y_i \sum_{s=1}^{t} \alpha_s h_s(x_i)\}
\]

(2)

(3)

(4)
Training Error Intuition

- On round $t$ weight of examples incorrectly classified by $h_t$ is increased
- If $x_i$ incorrectly classified by $H_T$, then $x_i$ wrong on (weighted) majority of $h_t$'s
  - If $x_i$ incorrectly classified by $H_T$, then $x_i$ must have large weight under $D_T$
  - But there can’t be many of them, since total weight $\leq 1$
Training Error Proof: It’s all about the Normalizers

\[ \hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} 1 \left[ y_i g(x_i) \leq 0 \right] \]  

Definition of training error
Training Error Proof: It’s all about the Normalizers

\[ \hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} 1 \left[ y_i g(x_i) \leq 0 \right] \]

\[ \leq \frac{1}{m} \sum_{i=1}^{m} \exp \left\{-y_i g(x_i)\right\} \]  

\[ 1 \left[ u \leq 0 \right] \leq \exp -u \text{ is true for all real } u. \]
Training Error Proof: It’s all about the Normalizers

\[ \hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} 1 \left[ y_i g(x_i) \leq 0 \right] \]  

(5)

\[ \leq \frac{1}{m} \sum_{i=1}^{m} \exp \left\{ -y_i g(x_i) \right\} \]  

(6)

(7)

Final distribution \( D_{t+1}(i) \)

\[ D_{t+1}(i) = \frac{\exp \left\{ -y_i \sum_{s=1}^{t} \alpha_s h_s(x_i) \right\}}{m \prod_{s=1}^{t} Z_s} \]  

(8)
Training Error Proof: It’s all about the Normalizers

\[ \hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} 1 \left[ y_i g(x_i) \leq 0 \right] \]  

\[ \leq \frac{1}{m} \sum_{i=1}^{m} \exp \left\{ -y_i g(x_i) \right\} \]  

\[ = \frac{1}{m} \sum_{i=1}^{m} \left( m \prod_{t=1}^{T} Z_t \right) D_{T+1}(i) \]  

\( m \)'s cancel, \( D \) is a distribution
Training Error Proof: It’s all about the Normalizers

\[ R(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1} \left[ y_i g(x_i) \leq 0 \right] \]  \hspace{1cm} (5)

\[ \leq \frac{1}{m} \sum_{i=1}^{m} \exp \left\{ -y_i g(x_i) \right\} \]  \hspace{1cm} (6)

\[ = \frac{1}{m} \sum_{i=1}^{m} \left[ m \prod_{t=1}^{T} Z_t \right] D_{T+1}(i) \]  \hspace{1cm} (7)

\[ = \prod_{t=1}^{T} Z_t \]  \hspace{1cm} (8)
Training Error Proof: Weak Learner Errors

Single Weak Learner

\[ Z_t = \sum_{i=1}^{m} D_t(i) \exp\{-\alpha_t y_i h_t(x_i)\} \]  \hspace{1cm} (9)

\[ = \]  \hspace{1cm} (10)

\[ = \]  \hspace{1cm} (11)

\[ = \]  \hspace{1cm} (12)
Training Error Proof: Weak Learner Errors

Single Weak Learner

\[
Z_t = \sum_{i=1}^{m} D_t(i) \exp \{-\alpha_t y_i h_t(x_i)\} \tag{9}
\]

\[
= \sum_{i:\text{right}} D_t(i) \exp \{-\alpha_t\} + \sum_{i:\text{wrong}} D_t(i) \exp \{\alpha_t\} \tag{10}
\]

\[
= \tag{11}
\]

\[
= \tag{12}
\]
Training Error Proof: Weak Learner Errors

Single Weak Learner

\[ Z_t = \sum_{i=1}^{m} D_t(i) \exp \{-\alpha_t y_i h_t(x_i)\} \quad (9) \]

\[ = \sum_{i:\text{right}} D_t(i) \exp \{-\alpha_t\} + \sum_{i:\text{wrong}} D_t(i) \exp \{\alpha_t\} \quad (10) \]

\[ = (1 - \epsilon_t) \exp \{-\alpha_t\} + \epsilon_t \exp \{\alpha_t\} \quad (11) \]

\[ = \quad (12) \]
Training Error Proof: Weak Learner Errors

Single Weak Learner

\[ Z_t = \sum_{i=1}^{m} D_t(i) \exp \{-\alpha_t y_i h_t(x_i)\} \]  
\[ = \sum_{i:\text{right}} D_t(i) \exp \{-\alpha_t\} + \sum_{i:\text{wrong}} D_t(i) \exp \{\alpha_t\} \]  
\[ = (1 - \epsilon_t) \exp \{-\alpha_t\} + \epsilon_t \exp \{\alpha_t\} \]  
\[ = (1 - \epsilon_t) \sqrt{\frac{\epsilon_t}{1 - \epsilon_t}} + \epsilon_t \sqrt{\frac{1 - \epsilon_t}{\epsilon_t}} \]
Training Error Proof: Weak Learner Errors

Single Weak Learner

\[ Z_t = (1 - \epsilon_t) \sqrt{\frac{\epsilon_t}{1 - \epsilon_t}} + \epsilon_t \sqrt{\frac{1 - \epsilon_t}{\epsilon_t}} \]  \quad (9)

Normalization Product

\[
\prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} 2 \sqrt{\epsilon_t (1 - \epsilon_t)} = \sqrt{1 - 4 \left( \frac{1}{2} - \epsilon_t \right)^2} \]
\[
\prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} 2 \sqrt{\epsilon_t (1 - \epsilon_t)} = \sqrt{1 - 4 \left( \frac{1}{2} - \epsilon_t \right)^2} \]  \quad (10)

\[
\prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} 2 \sqrt{\epsilon_t (1 - \epsilon_t)} = \sqrt{1 - 4 \left( \frac{1}{2} - \epsilon_t \right)^2} \]  \quad (11)
Training Error Proof: Weak Learner Errors

Normalization Product

\[
\prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} 2\sqrt{\epsilon_t(1-\epsilon_t)} = \sqrt{1 - 4\left(\frac{1}{2} - \epsilon_t\right)^2} \tag{9}
\]

\[
\leq \prod_{t=1}^{T} \exp\left\{ -2\left(\frac{1}{2} - \epsilon_t\right)^2 \right\} \tag{10}
\]

(11)
Training Error Proof: Weak Learner Errors

Normalization Product

\[
\prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} 2 \sqrt{\epsilon_t(1-\epsilon_t)} = \sqrt{1 - 4 \left( \frac{1}{2} - \epsilon_t \right)^2}
\]

(9)

\[
\leq \prod_{t=1}^{T} \exp \left\{ -2 \left( \frac{1}{2} - \epsilon_t \right)^2 \right\}
\]

(10)

\[
= \exp \left\{ -2 \sum_{t=1}^{T} \left( \frac{1}{2} - \epsilon_t \right)^2 \right\}
\]

(11)
### Generalization

#### VC Dimension

\[ \leq 2(d + 1)(T + 1) \log [(T + 1)e] \]

#### Margin-based Analysis

AdaBoost maximizes a linear program maximizes an \( L_1 \) margin, and the weak learnability assumption requires data to be linearly separable with margin \( 2\gamma \).