Fitting the Residuals

$$
\begin{aligned}
(\beta_m, \gamma_m) &= \arg\min_{\beta,\gamma} \sum_{i=1}^{N} L\Big(y_i, f_{m-1}(x_i) + \beta b(x_i;\gamma)\Big) \\
(\beta_m, \gamma_m) &= \arg\min_{\beta,\gamma} \sum_{i=1}^{N} \Big(\underbrace{y_i - f_{m-1}(x_i)}_{\text{Residual } r_i} - \beta b(x_i;\gamma)\Big)^2 \\
(\beta_m, \gamma_m) &= \arg\min_{\beta,\gamma} \sum_{i=1}^{N} \Big(r_i - \beta b(x_i;\gamma)\Big)^2 \qquad (1)
\end{aligned}
$$

This corresponds to fitting the "delta" model / weak learner to the residuals r_i between the observed values and the current model.

Generalization to Other Losses

Residuals and Loss Functions:

- For the **Mean Squared Error (MSE)**, residuals are the difference between observed and predicted values:

$$r_i = y_i - f_{m-1}(x_i)$$

- For a general loss function $L(y, f(x))$, residuals are replaced by **pseudo-residuals**, defined as:

$$r_i = -\frac{\partial L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)}$$

For the MSE $L = 1/2(y - f_{m-1}(x))^2$, the pseudo-residuals are the residuals themselves.

Key Idea:

- Instead of fitting to standard residuals, the model minimizes the **pseudo-residuals** derived from the gradient of the loss function.
- This allows Gradient Boosting to handle various loss functions, making it adaptable to different problems.

Example: Logistic Loss for Classification

The logistic loss is given by:

$$L(y, f(x)) = \log(1 + e^{-yf(x)})$$

Its gradient (pseudo-residual) is:

$$r_i = -\frac{\partial L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)} = \frac{-y_i}{1 + e^{y_i f_{m-1}(x_i)}}$$

- The algorithm fits the base learner to these pseudo-residuals at each stage.
- This allows for iterative improvement of the classification model.

General Framework for Boosting

The boosting framework can be summarized as:

1. Compute pseudo-residuals:

$$r_i = -\frac{\partial L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)}$$

2. Fit the base learner to the pseudo-residuals:

$$(\beta_m, \gamma_m) = \arg\min_{\beta,\gamma} \sum_{i=1}^{N} \Big(r_i - \beta b(x_i; \gamma) \Big)^2$$

3. Update the model:

$$f_m(x) = f_{m-1}(x) + \beta_m b(x; \gamma_m)$$

This framework applies to a variety of loss functions, making boosting a versatile method.