Mathematical Explanation of the Meta Model

In this study, we employ **Gradient Boosting** as the meta-learner due to its superior efficiency in minimizing predictive error. The ensemble architecture utilizes **Random Forest**, **Extreme Gradient Boost (XGBoost)**, and **Gradient Descent** as base learners. We provide a rigorous mathematical proof for the convergence of this algorithm below.

Problem Formulation

Given a training dataset $\{(x_i, y_i)\}_{i=1}^n$, our objective is to find a function $F(x)$ that minimizes the **empirical risk**:

$$J = \sum_{i=1}^{n} L(y_i, F(x_i)) \tag{1}$$

where the loss function $L(y, F(x))$ is assumed to be differentiable and convex.

Gradient Boosting Algorithm

To minimize J, we perform functional gradient descent through the following steps:

1. **Initialization:** Initialize the model with a constant value:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \tag{2}$$

2. **Iterative Boosting (for $m = 1$ to M):**

   - **Compute Pseudo-Residuals:** Calculate the negative gradient of the loss with respect to the current model prediction:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \dots, n \tag{3}$$

   - **Fit Weak Base Learner:** Fit a weak learner $h_m(x)$ to the residuals r_{im} (using base learners like Random Forest or Gradient Descent).
   - **Line Search for Step Size:** Determine the optimal multiplier γ_m:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \tag{4}$$

   - **Update the Model:**

$$F_m(x) = F_{m-1}(x) + \nu \gamma_m h_m(x) \tag{5}$$

   where $\nu \in (0, 1]$ represents the **learning rate** (shrinkage).

Convergence Proof

Assumptions

1. **Convexity:** $L(y, F)$ is convex in F.
2. **Lipschitz Gradient:** The gradient of the loss is L-smooth, satisfying: $\|\nabla L(F_1) - \nabla L(F_2)\| \leq L_{lip}\|F_1 - F_2\|$.
3. **Weak Learner Coverage:** For all gradients, there exists a weak learner h_m and $\rho > 0$ such that: $\langle \nabla L(F_{m-1}), h_m \rangle \geq \rho \|\nabla L(F_{m-1})\|^2$.

Descent Lemma (Smoothness)

By the Lipschitz gradient property, the change in loss can be bounded as:

$$L(F_m) \leq L(F_{m-1}) + \langle \nabla L(F_{m-1}), (F_m - F_{m-1}) \rangle + \frac{L_{lip}}{2}\|F_m - F_{m-1}\|^2 \quad (6)$$

Monotonic Decrease

Substituting the update rule $F_m - F_{m-1} = \nu\gamma_m h_m$, we obtain:

$$L(F_m) \leq L(F_{m-1}) - c \cdot \|\nabla L(F_{m-1})\|^2 \quad (7)$$

For a sufficiently small constant $c > 0$, this ensures a **monotonic decrease** in the loss function at every iteration.

Convergence to Optimum

As $m \to \infty$, $\|\nabla L(F_m)\| \to 0$. If L is **strongly convex**, the convergence rate is exponential (linear convergence):

$$L(F_m) - L(F^*) \leq (1 - \kappa)^m (L(F_0) - L(F^*)) \quad (8)$$

where κ is a constant related to the condition number of the loss surface.