

# Learnable Ensembling

## Theoretical Derivation (Single-Image Formulation)

Rose Abdulqadir Khairoalsendi

February 2026

## Abstract

This document presents the theoretical derivation of a **learnable ensembling mechanism** for object detection, formulated at the level of a **single image**. The derivation combines regression, classification, and constraint objectives into a unified optimization problem over ensemble weights.

## Contents

<b>1</b>	<b>Intuition Behind the Formulation</b>	<b>2</b>
<b>2</b>	<b>Notation</b>	<b>2</b>
<b>3</b>	<b>Ensemble Formulation</b>	<b>2</b>
<b>4</b>	<b>Ground-Truth Representation</b>	<b>3</b>
<b>5</b>	<b>Total Objective Function</b>	<b>3</b>
<b>6</b>	<b>Regression Loss (Localization)</b>	<b>3</b>
<b>7</b>	<b>Constraint Loss (Weight Normalization)</b>	<b>3</b>
<b>8</b>	<b>Classification Loss</b>	<b>3</b>
<b>9</b>	<b>Gradients with Respect to Ensemble Weights</b>	<b>4</b>
9.1	Regression Gradient . . . . .	4
9.2	Constraint Gradient . . . . .	4
9.3	Classification Gradient . . . . .	4
<b>10</b>	<b>Weight Update Rule</b>	<b>4</b>
<b>11</b>	<b>Key Characteristics of the Formulation</b>	<b>4</b>

# 1 Intuition Behind the Formulation

The core idea of this formulation is to treat ensemble weighting as a learnable optimization problem, rather than a fixed heuristic.

Each model contributes a prediction, but instead of averaging them uniformly, the ensemble learns how much to trust each model through the weight vector  $\Theta$ .

The regression loss encourages the weighted prediction to align with the ground-truth localization.

The classification loss ensures that models contributing confident class predictions are emphasized.

The constraint loss prevents trivial solutions by encouraging normalized weights.

By optimizing these objectives jointly, the ensemble:

- Adapts weights based on prediction quality
- Balances localization and classification performance
- Remains fully differentiable and end-to-end trainable

In effect, the ensemble learns which models matter most for a given prediction, rather than assuming all models are equally reliable.

## 2 Notation

Symbol	Description
$N$	Number of models / predictors in the ensemble
$\hat{Y} \in \mathbb{R}^{N \times 5}$	Matrix of model predictions (one row per model)
$\hat{Y}^T$	Transposed prediction matrix
$A \in \mathbb{R}^5$	Final ensembled prediction
$y \in \mathbb{R}^5$	Ground-truth target for a single image
$y_i \in \mathbb{R}^5$	Target associated with the $i$ -th predictor
$p_i$	Classification probability predicted by the $i$ -th model
$\theta_i$	Learnable weight assigned to the $i$ -th model
$\Theta \in \mathbb{R}^N$	Vector of all ensemble weights
$J_{\text{reg}}$	Regression (localization) loss
$J_{\text{cls}}$	Classification loss
$J_{\text{cons}}$	Constraint (normalization) loss
$J_\Theta$	Total objective function
$\alpha$	Learning rate

## 3 Ensemble Formulation

Let:

- $N$  be the **number of models / predictors**
- $\Theta \in \mathbb{R}^N$  be the vector of **ensemble weights**
- $\hat{Y} \in \mathbb{R}^{N \times 5}$  be the matrix of predicted outputs

The ensemble prediction is defined as:

$$A = \hat{Y}^T \Theta$$

where:

$$\hat{Y} \in \mathbb{R}^{N \times 5}, \quad \Theta \in \mathbb{R}^N, \quad A \in \mathbb{R}^5$$

## 4 Ground-Truth Representation

The ground-truth vector is defined as:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad y_i \in \mathbb{R}^5$$

Each  $y_i$  represents the target output associated with a single predictor.

## 5 Total Objective Function

For a single image, the total loss is defined as:

$$J_\Theta = J_{\text{reg}} + J_{\text{cons}} + J_{\text{cls}}$$

Each term captures a different aspect of the detection objective.

## 6 Regression Loss (Localization)

The regression loss penalizes deviation between the weighted prediction and the target:

$$J_{\text{reg}} = \frac{1}{2} \sum_{i=1}^N (\theta_i y_i - y)^2$$

This term corresponds to a **localization regression loss**.

## 7 Constraint Loss (Weight Normalization)

To enforce a normalization constraint on the ensemble weights:

$$J_{\text{cons}} = \frac{1}{2} \left( 1 - \sum_{i=1}^N \theta_i \right)^2$$

This term acts as a **mean-squared constraint loss**, encouraging the weights to sum to one.

## 8 Classification Loss

For classification, a binary cross-entropy-style loss is used:

$$J_{\text{cls}} = \sum_{i=1}^N [y \log(\theta_i p_i) + (1 - y) \log(1 - \theta_i p_i)]$$

where  $p_i$  denotes the class probability predicted by the  $i$ -th model.

## 9 Gradients with Respect to Ensemble Weights

### 9.1 Regression Gradient

$$\nabla_{\Theta} J_{\text{reg}} = \hat{Y}(\hat{Y}^T \Theta - Y)$$

### 9.2 Constraint Gradient

$$\nabla_{\Theta} J_{\text{cons}} = \Theta - \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

### 9.3 Classification Gradient

$$\nabla_{\Theta} J_{\text{cls}} = \begin{bmatrix} \frac{p_1 \theta_1 - y}{\theta_1(1 - \theta_1 p_1)} \\ \vdots \end{bmatrix}$$

## 10 Weight Update Rule

The ensemble weights are updated via gradient descent:

$$\Theta_{k+1} = \Theta_k - \alpha (\nabla_{\Theta} J_{\text{reg}} + \nabla_{\Theta} J_{\text{cls}} + \nabla_{\Theta} J_{\text{cons}})$$

where  $\alpha$  is the learning rate.

## 11 Key Characteristics of the Formulation

- Ensemble weights are **learned directly** via optimization
- Regression, classification, and normalization are **jointly enforced**
- Lowercase  $\theta_i$  denotes **individual model contributions**
- Uppercase  $\Theta$  denotes the **global ensemble parameter vector**
- Designed explicitly for **object detection outputs**