Appendix B: Mathematical Details and Derivations

B.1 Bernoulli Model and Likelihood

Each segment is represented as (x_i, y_i), where $x_i \in R^{38}$ is the feature vector and $y_i \in \{0, 1\}$ is the case-level group label inherited at the segment level.

CourtShadow models the label as a Bernoulli random variable:

$$y_i \sim Bernoulli(p_i), \qquad p_i = P(y_i = 1 \mid x_i).$$

Assuming conditional independence given x_i and parameter vector θ, the likelihood of the entire dataset is:

$$L(\theta) = \prod_{i=1}^{n} P(y_i \mid x_i; \theta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i}.$$

Taking logs, the log-likelihood is:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} [y_i \log p_i + (1 - y_i) \log(1 - p_i)].$$

The model uses a logistic link to parameterize p_i.

B.2 Logistic Link and Negative Log-Likelihood

The logistic (sigmoid) function is:

$$\sigma(z) = \frac{1}{1 + e^{-z}}.$$

CourtShadow parameterizes the Bernoulli probability as:

$$p_i = \sigma(\theta^\top x_i) = \frac{1}{1 + e^{-\theta^\top x_i}}.$$

Substituting this into the log-likelihood, the *negative* log-likelihood (NLL) used as the loss is:

$$J(\theta) = -\ell(\theta) = -\sum_{i=1}^{n} [y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i))].$$

This is the standard cross-entropy loss for logistic regression.

B.3 Gradient of the Logistic Loss

Define $z_i = \theta^\top x_i$ and $p_i = \sigma(z_i)$. We compute the gradient of $J(\theta)$ with respect to θ.

First, observe:

$$\frac{\partial p_i}{\partial z_i} = \frac{\partial}{\partial z_i}\left(\frac{1}{1+e^{-z_i}}\right) = p_i(1-p_i).$$

By the chain rule,

$$\frac{\partial p_i}{\partial \theta} = \frac{\partial p_i}{\partial z_i}\frac{\partial z_i}{\partial \theta} = p_i(1-p_i)\,x_i.$$

Now differentiate the loss:

$$J(\theta) = -\sum_{i=1}^{n}[y_i \log p_i + (1-y_i)\log(1-p_i)].$$

Taking ∇_θ,

$$\nabla_\theta J(\theta) = -\sum_{i=1}^{n}\left[\frac{y_i}{p_i}\frac{\partial p_i}{\partial \theta} - \frac{1-y_i}{1-p_i}\frac{\partial p_i}{\partial \theta}\right].$$

Substitute $\frac{\partial p_i}{\partial \theta} = p_i(1-p_i)x_i$:

$$\nabla_\theta J(\theta) = -\sum_{i=1}^{n}\left[\frac{y_i}{p_i}p_i(1-p_i)x_i - \frac{1-y_i}{1-p_i}p_i(1-p_i)x_i\right].$$

Simplify:

$$\nabla_\theta J(\theta) = -\sum_{i=1}^{n}\left[y_i(1-p_i)x_i - (1-y_i)p_i x_i\right].$$

Rearrange the terms inside:

$$y_i(1-p_i) - (1-y_i)p_i = y_i - y_i p_i - p_i + y_i p_i = y_i - p_i.$$

So:

$$\nabla_\theta J(\theta) = -\sum_{i=1}^{n}(y_i - p_i)x_i = \sum_{i=1}^{n}(p_i - y_i)x_i.$$

This is the gradient used by gradient-based optimizers.

B.4 L2-Regularized Objective

To reduce overfitting on a small dataset and keep weights in a numerically stable regime, CourtShadow adds an L2 penalty term:

$$\Omega(\theta) = \lambda \sum_j \theta_j^2.$$

The regularized objective becomes:

$$J_{L2}(\theta) = J(\theta) + \Omega(\theta) = -\sum_{i=1}^{n} [y_i \log p_i + (1 - y_i) \log(1 - p_i)] + \lambda \sum_j \theta_j^2.$$

Differentiating the penalty term:

$$\nabla_\theta \Omega(\theta) = 2\lambda\, \theta.$$

Thus the gradient of the regularized loss is:

$$\nabla_\theta J_{L2}(\theta) = \sum_{i=1}^{n} (p_i - y_i) x_i + 2\lambda\, \theta.$$

In practice, many numerical packages absorb the factor of 2 into λ, but the idea is the same: larger weights incur larger penalties, shrinking coefficients and improving generalization on held-out data.

B.5 Feature Scaling and Its Effect

Continuous features are standardized using

$$x' = \frac{x - \mu}{\sigma},$$

where μ and σ are computed on the training set. Substituting x' into $\theta^\top x$ yields:

$$\theta^\top x' = \sum_j \theta_j \frac{x_j - \mu_j}{\sigma_j}.$$

This has two benefits:

1. Features on different scales (e.g., token counts vs. rates) contribute comparably to the decision boundary.

2. The magnitude of θ_j becomes more interpretable: it represents the effect of a one-standard-deviation change in feature j.

Binary topic indicators are left unscaled to preserve their direct "on/off" interpretation.

B.6 Case-Level Aggregation

Segment-level probabilities p_j are aggregated to form a case-level *Linguistic Environment Score* (LES):

$$\bar{p}_{case} = \frac{1}{m} \sum_{j=1}^{m} p_j,$$

where m is the number of segments in a case.

From a statistical perspective, $\bar{p}_{case}$ approximates the expected probability that a randomly sampled segment from that case is classified as Group 1 by the model. This aggregation reduces within-case noise (e.g., one unusually harsh turn) and focuses on the overall environment.

B.7 Linear Contributions and Feature Families

Because logistic regression is linear in feature space, we can write the log-odds for a segment as:

$$\theta^\top x = \sum_{k=1}^{d} \theta_k x_k.$$

If we partition the indices $1, \ldots, d$ into disjoint families (structure, framing, pronouns, topics), the total log-odds decomposes as:

$$\theta^\top x = \underbrace{\sum_{k \in structure} \theta_k x_k}_{structure} + \underbrace{\sum_{k \in framing} \theta_k x_k}_{framing} + \underbrace{\sum_{k \in pronouns} \theta_k x_k}_{pronouns} + \underbrace{\sum_{k \in topics} \theta_k x_k}_{topics}.$$

This decomposition is used in the website's interpretability plots to show how each family pushes a segment or case toward Group A or Group B. Case-level family contributions are obtained by averaging these family sums across all segments in a case.

B.8 ROC AUC and Calibration

The ROC area under the curve (AUC) is defined as:

$$AUC = P(s_{pos} > s_{neg}),$$

where s_{pos} and s_{neg} are scores for randomly chosen positive (Group 1) and negative (Group 0) examples. In practice, AUC is computed by ranking cases by $\bar{p}_{case}$ and computing the fraction of correctly ordered positive–negative pairs.

Calibration is evaluated by binning predicted probabilities and comparing them to empirical frequencies:

$$calibration(bin) = E[Y \mid \hat{p} \in bin],$$

where $\hat{p}$ is the model's predicted probability. If the model is well-calibrated, the reliability curve (empirical vs. predicted) lies near the diagonal. In CourtShadow, these diagnostics help confirm that the logistic probabilities are meaningful as *degrees of belief* about Group A environments, not merely ranking scores.