Theoretical Bounds for Multi-Task Output Features

ICF Estimation Project · 2025-12-06

This document establishes theoretical bounds and expected ranges for multi-task learning outputs beyond ICF prediction: Language Detection, Era Classification, Temporal ICF Prediction, and Text Reduction.

0.1. Notation

- **Accuracy**: Classification accuracy
- **L_CE**: Cross-entropy loss
- **L_MSE**: Mean squared error loss
- **ρ**: Spearman rank correlation coefficient
- **Regret**: Embedding regret (1 - cosine similarity)
- **W_2**: Wasserstein-2 distance (for optimal transport formulation)
- **I(S; Embedding)**: Mutual information between selected words S and embedding
- **H(Embedding)**: Shannon entropy of embedding distribution
- **N**: Number of classes (languages, eras)
- **k**: Number of words to keep (text reduction budget)

0.2. Language Detection

0.2.1. Task Description

Predict the language of a word from character patterns. This is a multi-class classification task (typically 10+ languages).

0.2.2. Theoretical Bound

$$\text{Acc}_{\text{language}} = \frac{1}{|D|} \sum_{(x,y) \in D} 1[\hat{y}(x) = y]$$

where `hat(y)(x)` is the predicted language and `y` is the true language.

0.2.3. Expected Range

Expected Range for Character-Level Models:

- **Best case:** Accuracy ∈ [0.70, 0.85] (character patterns are language-specific)
- **Good:** Accuracy ∈ [0.60, 0.70]
- **Acceptable:** Accuracy ∈ [0.50, 0.60]
- **Poor:** Accuracy < 0.50 (worse than random for balanced classes)

0.2.4. Mathematical Foundation

- For `N` languages with balanced classes, random baseline: Accuracy_random = 1/N
- For `N = 10` languages: Accuracy_random = 0.10
- Character patterns (n-grams, character frequency) are language-specific
- Expected accuracy: Accuracy ∈ [0.60, 0.85] depending on language similarity
- Upper bound: Accuracy ≤ 1 - H(Language|Characters) / H(Language) (Fano's inequality)

0.3. Era Classification

0.3.1. Task Description

Predict the historical era when a word was commonly used (e.g., 1800s, 1900s, 2000s). This is a multi-class classification task (typically 3-5 eras).

0.3.2. Expected Range

Expected Range for Character-Level Models:
- **Best case:** `0.50 - 0.70` accuracy (some temporal patterns exist)
- **Good:** `0.40 - 0.50` accuracy
- **Acceptable:** `0.30 - 0.40` accuracy
- **Poor:** `< 0.30` accuracy (worse than random for 3-5 classes)

0.3.3. Mathematical Foundation
- For N eras with balanced classes, random baseline = 1/N
- For 5 eras: random = 0.20
- Character patterns change over time (spelling reforms, new words)
- But changes are subtle and may not be captured by character-level models
- Expected accuracy: `0.30 - 0.60` (lower than language detection)

0.4. Temporal ICF Prediction

0.4.1. Task Description
Predict ICF scores across multiple decades (e.g., 1800, 1900, 2000). This is a regression task with temporal consistency constraints.

0.4.2. Expected Range

Expected Range for Character-Level Models:
- **Best case:** `0.15 - 0.18` Spearman per decade (similar to ICF)
- **Good:** `0.12 - 0.15` Spearman per decade
- **Acceptable:** `0.10 - 0.12` Spearman per decade
- **Poor:** `< 0.10` Spearman per decade

0.4.3. Mathematical Foundation
- Each decade has its own ICF distribution
- Character patterns ⟶ ICF is still indirect (same bound as ICF)
- Temporal consistency helps: predictions should be smooth across decades
- Expected Spearman: `0.12 - 0.18` (similar to ICF, but may be slightly lower due to temporal complexity)

0.5. Text Reduction (Embedding Regret Minimization)

0.5.1. Task Description
Minimize embedding regret when reducing text by selecting a subset of words that preserves the original embedding as much as possible. This is a **ranking + embedding similarity** task that can be **disjoint from ICF prediction** (doesn't require ICF scores, but can use them as a heuristic).

Key Insight: The task is to find the minimal "path" of embedding regret - i.e., select words such that the embedding of the reduced text is as close as possible to the original embedding.

0.5.2. Theoretical Bound

Bound: Depends on embedding quality, word selection strategy, and whether ICF is used

Expected Range for Character-Level Models:
- **Best case:** `0.05 - 0.15` regret (cosine distance)
- **Good:** `0.15 - 0.30` regret
- **Acceptable:** `0.30 - 0.50` regret
- **Poor:** `> 0.50` regret

0.5.3. Mathematical Foundation

Regret = `1 - cosine_similarity(original_embedding, reduced_embedding)`

Path Regret: Cumulative embedding change along the reduction path

- Perfect reduction (keeping all important words): regret ≈ 0.0
- Random reduction: regret ≈ 0.5 - 0.7
- **ICF-based reduction:** regret ≈ 0.15 - 0.30 (if ICF scores are accurate)
- **Direct embedding-based reduction** (disjoint from ICF): regret ≈ 0.10 - 0.25 (potentially better, as it directly optimizes embedding similarity)
- Expected regret: `0.15 - 0.30` for ICF-based, `0.10 - 0.25` for direct embedding-based

0.5.4. Connection to ICF

Option 1 (Coupled): Text reduction uses ICF scores to rank words (rare words = important)
- Better ICF prediction ⟶ better text reduction
- Multi-task learning can improve both by learning shared features

Option 2 (Disjoint): Text reduction directly optimizes embedding similarity without ICF
- Can be trained independently of ICF
- May perform better (direct optimization vs proxy via ICF)
- Still benefits from shared character-level features in multi-task setup

0.6. Summary Table

Task	Metric	Best Case	Good	Acceptable	Poor
Language Detection	Accuracy	0.70-0.85	0.60-0.70	0.50-0.60	< 0.50
Language Detection	Classification Loss	0.2-0.5	0.5-0.7	0.7-1.0	> 1.0
Era Classification	Accuracy	0.50-0.70	0.40-0.50	0.30-0.40	< 0.30
Era Classification	Classification Loss	0.5-1.0	1.0-1.5	1.5-2.0	> 2.0
Temporal ICF	Spearman	0.15-0.18	0.12-0.15	0.10-0.12	< 0.10
Temporal ICF	Consistency Loss	0.0-0.05	0.05-0.10	0.10-0.20	> 0.20
Text Reduction	Regret	0.05-0.15	0.15-0.30	0.30-0.50	> 0.50
Text Reduction	Path Regret	0.10-0.20	0.20-0.40	0.40-0.60	> 0.60
Text Reduction	Ranking Loss	0.0-0.1	0.1-0.2	0.2-0.3	> 0.3

0.7. Practical Guidelines

0.7.1. Multi-Task Balance

Monitor task weight ratios:
- ICF should dominate (primary task)
- Auxiliary tasks should help, not hurt
- If auxiliary task accuracy is poor, reduce its weight

0.7.2. Convergence Indicators

- **Language accuracy:** Should increase to 0.60-0.70
- **Era accuracy:** Should increase to 0.40-0.50
- **Temporal Spearman:** Should increase to 0.12-0.15
- **Text reduction regret:** Should decrease to 0.15-0.30

0.7.3. Warning Signs

- **Language accuracy < 0.50:** Model not learning language features
- **Era accuracy < 0.30:** Model not learning era features
- **Temporal Spearman < 0.10:** Temporal prediction failing

- **Text reduction regret > 0.50:** Reduction not preserving meaning
- **One task dominates:** AMOO not balancing tasks