📓 Logistic Regression – Advanced Theoretical Analysis

Model:

$$P(Y = 1 \mid X = x) = \sigma(x^T \beta)$$

where

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

Log-likelihood:

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \log \sigma(x_i^T \beta) + (1 - y_i) \log(1 - \sigma(x_i^T \beta)) \right]$$

Negative log-likelihood (logistic loss):

$$J(\beta) = -\ell(\beta)$$

1 Prove Convexity of Logistic Loss

Consider one sample

$$\ell_i(\beta) = -\left[y_i \log \sigma(z_i) + (1 - y_i) \log(1 - \sigma(z_i))\right]$$

where $z_i = x_i^T \beta$.

Gradient:

$$\nabla J(\beta) = X^T(\sigma(X\beta) - y)$$

Hessian:

$$\nabla^2 J(\beta) = X^T W X$$

where:

$$W = \text{diag}(\sigma_i(1 - \sigma_i))$$

Since:

$$\sigma_i(1 - \sigma_i) > 0$$

W is diagonal positive definite.

For any vector v:

$$v^T X^T W X v = (Xv)^T W (Xv) \geq 0$$

Hence Hessian is PSD.

✅ Logistic loss is convex.

If X full rank ⇒ strictly convex.

2 Why Logistic Regression Does Not Overfit as Easily as High-Degree Polynomial Regression

Key reason:

- Logistic regression has **linear decision boundary**
- Parameter count small
- Convex optimization
- Often regularized

High-degree polynomial regression:

- Large hypothesis space

- High VC dimension
- Interpolates noise

Thus logistic regression has:

- Lower model capacity
- Better bias-variance balance

Overfitting arises from complexity, not just loss function.

3 Derive IRLS from Newton's Method

Newton update:

$$\beta_{new} = \beta_{old} - H^{-1}\nabla J(\beta)$$

We have:

$$\nabla J(\beta) = X^T(\sigma - y)$$

$$H = X^T W X$$

Plug in:

$$\beta_{new} = \beta - (X^T W X)^{-1} X^T(\sigma - y)$$

Rearrange:

Define working response:

$$z = X\beta + W^{-1}(y - \sigma)$$

Then update becomes:

$$\beta_{new} = (X^T W X)^{-1} X^T W z$$

This is **Weighted Least Squares**.

Hence:

Logistic regression = iteratively solving weighted linear regression.

This is IRLS (Iteratively Reweighted Least Squares).

4 Logistic Loss vs Hinge Loss (Mathematical Comparison)

Logistic Loss:

Logistic Loss:

$$\log(1 + e^{-yf(x)})$$

- Smooth
- Differentiable
- Probabilistic

Hinge Loss (SVM):

$$\max(0, 1 - yf(x))$$

- Non-smooth
- Piecewise linear
- Margin-based

Property	Logistic	Hinge
Smooth	Yes	No
Probabilistic	Yes	No
Margin focus	Soft	Hard margin
Optimization	Easier (Newton)	Needs subgradient

Logistic approximates hinge loss smoothly.

5 What Happens When Data is Perfectly Separable?

If exists β such that:

$$y_i x_i^T \beta > 0$$

Then:

Likelihood increases as:

$$||\beta|| \to \infty$$

Loss → 0.

Thus:

No finite optimum.

Weights diverge.

6 Why MLE Does Not Exist Under Perfect Separation

Likelihood:

$$L(\beta) = \prod \sigma(y_i x_i^T \beta)$$

If separable:

We can scale $\beta \to c\beta$.

As $c \to \infty$:

$$\sigma(y_i x_i^T c\beta) \to 1$$

Thus:

$$L(\beta) \to 1$$

No maximum at finite β.

Hence MLE does not exist.

Solution: Regularization.

7 Logistic Regression as Maximum Entropy Classifier

Principle:

Among all distributions satisfying:

Among all distributions satisfying:

$$\mathbb{E}[YX] = \text{observed}$$

Choose distribution maximizing entropy:

$$\max -\sum p(x,y)\log p(x,y)$$

Subject to constraints.

Using Lagrange multipliers yields:

$$P(Y=1|X) = \frac{\exp(\beta^T X)}{1+\exp(\beta^T X)}$$

Thus logistic regression = maximum entropy distribution under linear constraints.

8 Asymptotic Distribution of Estimator

Under regularity conditions:

MLE is asymptotically normal:

$$\sqrt{n}(\hat{\beta}-\beta) \to \mathcal{N}(0, I(\beta)^{-1})$$

Where Fisher Information:

$$I(\beta) = X^T W X$$

Thus:

$$\mathrm{Var}(\hat{\beta}) \approx (X^T W X)^{-1}$$

9 Generative vs Discriminative

Logistic Regression (Discriminative)

Models:

$$P(Y|X)$$

Naive Bayes (Generative)

Models:

$$P(X|Y), P(Y)$$

Then uses Bayes rule.

Comparison:

Aspect	Logistic	Naive Bayes
Model type	Discriminative	Generative
Bias	Low	Higher
Variance	Higher	Lower
Small data	Worse	Better
Asymptotic	Better	Suboptimal

Naive Bayes makes conditional independence assumption.

Logistic makes fewer distributional assumptions.

🔟 When Does Logistic Regression Fail?

1. Non-linear boundaries (unless features engineered)
2. Perfect separation (MLE diverges)
3. Severe multicollinearity
4. p >> n without regularization
5. Highly imbalanced data
6. Outliers with high leverage