Detailed Derivation of P_d for Changepoint Detection in Bernoulli Data

Introduction

This document derives the formula for P_d, the probability that the detected changepoint $\hat{t}$ is within a margin δ of the true changepoint t^*, in a Bernoulli-distributed sequence. The goal is to reference existing research to justify the assumptions and steps involved in this derivation.

Step 1: Problem Setup

We are given a sequence of T points from a Bernoulli distribution. At an unknown point t^*, the distribution changes from $\text{Bern}(p_1)$ to $\text{Bern}(p_2)$. The goal is to estimate t^* and compute P_d, the probability that the detected changepoint $\hat{t}$ lies within a margin δ around t^*.

The log-likelihood function for a sequence of Bernoulli variables is:

$$L(t^*) = \prod_{t=1}^{t^*} p_1^{y_t}(1-p_1)^{1-y_t} \prod_{t=t^*+1}^{T} p_2^{y_t}(1-p_2)^{1-y_t}$$

The log-likelihood is maximized with respect to t^* to obtain the estimate $\hat{t}$:

$$\log L(t^*) = \sum_{t=1}^{t^*} (y_t \log p_1 + (1-y_t)\log(1-p_1)) + \sum_{t=t^*+1}^{T} (y_t \log p_2 + (1-y_t)\log(1-p_2))$$

This method of estimating t^* using Maximum Likelihood Estimation (MLE) is supported by standard changepoint detection literature for Bernoulli-distributed data, such as Gichuhi's work [1].

Step 2: Asymptotic Normality of $\hat{t}$

For large sample sizes, the changepoint estimator $\hat{t}$ is asymptotically normally distributed around the true changepoint t^*. Specifically:

$$\hat{t} \sim \mathcal{N}(t^*, \sigma^2)$$

This result is a direct consequence of the properties of MLE for changepoint models, as discussed in the works of Bickel et al. [2] and Gichuhi [1]. Both references confirm that for large T, the detection error tends to follow a normal distribution.

Step 3: Deriving P_d

Given that $\hat{t}$ follows a normal distribution centered at t^*, we now compute P_d, the probability that $\hat{t}$ lies within δ of t^*:

$$P_d = P(t^* - \delta \leq \hat{t} \leq t^* + \delta)$$

This probability can be expressed as the integral of the normal distribution's probability density function (PDF):

$$P_d = \int_{t^*-\delta}^{t^*+\delta} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\hat{t} - t^*)^2}{2\sigma^2}\right) d\hat{t}$$

This result is supported by asymptotic theory, as shown in Bickel et al. [2], where the normal approximation for the MLE estimate is derived.

Step 4: Cumulative Distribution Function (CDF) of the Normal Distribution

The integral above simplifies to the difference between the cumulative distribution function (CDF) of the normal distribution evaluated at $\frac{\delta}{\sigma}$:

$$P_d = \Phi\left(\frac{\delta}{\sigma}\right) - \Phi\left(\frac{-\delta}{\sigma}\right)$$

Where $\Phi(\cdot)$ is the standard normal CDF, and σ is the standard deviation of the estimator. This final result holds for large sample sizes T, where the normal approximation is valid.

Step 5: Calculation of σ using the Fisher Information

The standard deviation σ can be computed using the Fisher Information. The Fisher Information $\mathcal{I}(t^*)$ is defined as the negative expected value of the second derivative of the log-likelihood function:

$$\mathcal{I}(t^*) = -E\left[\frac{\partial^2 \log L(t^*)}{\partial t^{*2}}\right]$$

For a Bernoulli sequence, we compute the second derivative of the log-likelihood:

- For the segment before the changepoint $t \leq t^*$:

$$\frac{\partial^2 \log L(t^*)}{\partial t^{*2}} = -\frac{t^*}{p_1(1-p_1)}$$

- For the segment after the changepoint $t > t^*$:

$$\frac{\partial^2 \log L(t^*)}{\partial t^{*2}} = -\frac{T-t^*}{p_2(1-p_2)}$$

The Fisher Information $\mathcal{I}(t^*)$ is the sum of these two terms:

$$\mathcal{I}(t^*) = \frac{t^*}{p_1(1-p_1)} + \frac{T-t^*}{p_2(1-p_2)}$$

Finally, the standard deviation σ is computed as the inverse of the square root of the Fisher Information:

$$\sigma = \frac{1}{\sqrt{\frac{t^*}{p_1(1-p_1)} + \frac{T-t^*}{p_2(1-p_2)}}}$$

This provides the standard deviation σ used in the calculation of P_d.

Conclusion

Thus, the final formula for P_d, the probability that the detected changepoint is within a margin δ of the true changepoint, is:

$$P_d = \Phi\left(\frac{\delta}{\sigma}\right) - \Phi\left(\frac{-\delta}{\sigma}\right)$$

Where σ is computed using the Fisher Information derived from the second derivative of the log-likelihood for the Bernoulli-distributed sequence.

References

[1] A. W. Gichuhi, *Nonparametric Changepoint Analysis for Bernoulli Random Variables*, Technische Universität Kaiserslautern, 2008.

[2] P. J. Bickel, et al., *Asymptotic Normality of Maximum Likelihood Estimators in Multiple Change-Point Models*, Project Euclid, 2010.