Offline Change Point Detection in Bernoulli Data: Probability that the Estimated Change Point Equals the True One

Guneesh Vats

March 14, 2025

Problem Statement and Overview

We have a sequence of Bernoulli observations (each taking values 0 or 1). These observations come from two distinct Bernoulli distributions:

- For time indices $i = 1, 2, \ldots, \tau$, observations X_i are i.i.d. from Bernoulli(p_1).
- For time indices $i = \tau + 1, \ldots, n$, observations X_i are i.i.d. from Bernoulli(p_2),

where $p_1 \neq p_2$. The integer $\tau \in \{1, 2, \ldots, n-1\}$ is an unknown "true" change point. We observe $X_1, X_2, \ldots, X_n$ in an *offline* setting (i.e., we have all the data from 1 to n at once) and we want to estimate τ by some procedure, typically the maximum likelihood estimator (MLE). Our question is:

$$\text{What is } \mathbb{P}\big(\hat{\tau} = \tau\big)?$$

that is, *the probability that our detected change point equals the actual, true change point.*

In what follows, we give a step-by-step derivation of the main theoretical result. We will break down each step and indicate whenever a known result from the literature is invoked, explaining how that result is used in this derivation. We also provide references at the end.

1 Step 1: Notation and Setup

1.1 Observation Window and Indices

- Let n be the total number of observations, with $n \in \mathbb{N}$ and $n \geq 2$.

- We define the *true change point* as an integer τ satisfying $1 \leq \tau \leq n-1$.
- The observations are $\{X_i\}_{i=1}^n$, each $X_i \in \{0, 1\}$.

1.2 Bernoulli Model

The data generating process (DGP) is:

$$X_1, \ldots, X_\tau \overset{\text{iid}}{\sim} \text{Bernoulli}(p_1), \quad X_{\tau+1}, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p_2), \quad p_1 \neq p_2.$$

All X_i are independent, but their distribution changes at time τ. The assumption $p_1 \neq p_2$ ensures there is a genuine change to be detected.

- **Significance of $p_1 \neq p_2$:** This guarantees identifiability of a change. If $p_1 = p_2$, there would be no change and the problem is ill-defined.
- **Parameter labeling:**
  - p_1: Bernoulli parameter before the change,
  - p_2: Bernoulli parameter after the change,
  - τ: the true change point.

2 Step 2: Likelihood and Log-Likelihood Functions

2.1 Case A: Known Parameters p_1 and p_2

Suppose for simplicity we know the true values of p_1 and p_2. For a *candidate* change point $k \in \{0, 1, \ldots, n\}$, the likelihood of the entire data sequence under that hypothesis is:

$$L(k) \;=\; \prod_{i=1}^{k} p_1^{X_i}(1-p_1)^{1-X_i} \;\times\; \prod_{i=k+1}^{n} p_2^{X_i}(1-p_2)^{1-X_i}. \tag{1}$$

Taking logarithms yields the *log-likelihood*:

$$\ell(k) \;=\; \ln L(k) \;=\; \sum_{i=1}^{k}\big[X_i \ln p_1 + (1-X_i)\ln(1-p_1)\big] \;+\; \sum_{i=k+1}^{n}\big[X_i \ln p_2 + (1-X_i)\ln(1-p_2)\big]. \tag{2}$$

Significance and Usage

- **Why log-likelihood?** The log-likelihood is easier to manipulate than the product form. In change-point detection research, comparing log-likelihoods under different candidate change points is standard (see, for example, [1], Section 2.1).

- **How is this used here?** We will define our estimator $\hat{\tau}$ as the argument that maximizes $\ell(k)$ over all possible k. This is the *maximum likelihood* principle.

2.2 Case B: Unknown Parameters p_1 and p_2

If p_1 and p_2 are not known, for each candidate k we estimate them via maximum likelihood *within* each segment:

$$\hat{p}_{1,k} \;=\; \frac{1}{k}\sum_{i=1}^{k} X_i, \quad \hat{p}_{2,k} \;=\; \frac{1}{n-k}\sum_{i=k+1}^{n} X_i.$$

Then

$$\ell(k) \;=\; \sum_{i=1}^{k}\Big[X_i \ln \hat{p}_{1,k} + (1-X_i)\ln(1-\hat{p}_{1,k})\Big] + \sum_{i=k+1}^{n}\Big[X_i \ln \hat{p}_{2,k} + (1-X_i)\ln(1-\hat{p}_{2,k})\Big].$$

- **Significance and Usage:** In real applications, p_1 and p_2 are typically unknown and must be estimated from data. The maximum likelihood approach still holds, but the expressions are more involved. The logic, however, is the same: define $\ell(k)$ via the best-fit Bernoulli parameters for each segment, then choose k to maximize that $\ell(k)$.

3 Step 3: The Change-Point Estimator

We define the **maximum likelihood estimator** (MLE) of the change point as

$$\hat{\tau} \;=\; \arg\max_{k\in\{0,1,\ldots,n\}} \ell(k). \tag{3}$$

(In practice, one usually restricts k to $\{1,\ldots,n-1\}$ to ensure a change is indeed within the interior of the sequence.)

- **Significance:** This $\hat{\tau}$ is the fundamental output of the offline (batch) detection. If $\hat{\tau} = \tau$, we have perfectly detected the true change location.

- **How it leads to result:** Our goal is to compute or bound $\mathbb{P}(\hat{\tau} = \tau)$. We first express this event in terms of $\ell(k)$.

4 Step 4: Probability That $\hat{\tau}$ Equals the True τ

By definition,

$$\{\hat{\tau} = \tau\} \;=\; \Big\{\ell(\tau) \;\geq\; \ell(k) \text{ for all } k\Big\}.$$

Hence

$$\mathbb{P}(\hat{\tau} = \tau) \;=\; \mathbb{P}\Big(\ell(\tau) \;\geq\; \ell(k) \text{ for all } k\Big).$$

4.1 Log-Likelihood Differences

Define a log-likelihood difference

$$D(\tau, k) \;=\; \ell(\tau) \;-\; \ell(k).$$

Then

$$\{\hat{\tau} = \tau\} \;=\; \bigcap_{k \neq \tau} \{\, D(\tau, k) \;\geq\; 0\}, \quad \text{thus} \quad \mathbb{P}(\hat{\tau} = \tau) \;=\; \mathbb{P}\Big(D(\tau, k) \;\geq\; 0 \text{ for all } k \neq \tau\Big).$$

- **Significance:** This is the formal event of correct detection. Studying $D(\tau, k)$ for each k amounts to comparing all candidate segmentations to the true segmentation.

- **Utility of this difference approach:** It is common in the change-point literature (e.g. [2], Section 1.2) to look at log-likelihood ratio (LLR) *differences* between candidate solutions. This clarifies the geometry of the probability event.

5 Step 5: Exact Finite-n Expression and Its Complexity

The probability of correct detection can be written explicitly (in principle) as:

$$\mathbb{P}(\hat{\tau} = \tau) \;=\; \sum_{x_1, \ldots, x_n \in \{0,1\}} \mathbf{1}\Big\{\ell(\tau; x_1, \ldots, x_n) \;\geq\; \ell(k; x_1, \ldots, x_n)\ \forall k\Big\} \times \prod_{i=1}^{n} f_{X_i}(x_i),$$

where $f_{X_i}(x_i)$ is the *true* Bernoulli pmf for X_i, i.e.:

$$f_{X_i}(x_i) \;=\; \begin{cases} p_1 & \text{if } x_i = 1 \text{ and } i \le \tau, \\ 1 - p_1 & \text{if } x_i = 0 \text{ and } i \le \tau, \\ p_2 & \text{if } x_i = 1 \text{ and } i > \tau, \\ 1 - p_2 & \text{if } x_i = 0 \text{ and } i > \tau. \end{cases}$$

- **Significance:** This is the direct, brute-force approach. It sums over all 2^n possible binary data sequences, checking if τ is the MLE.
- **Why it is not simplified further:** For large n, 2^n is enormous; this sum has no closed-form solution in general.
- **Reference usage:** Some authors (e.g. [1], eq. (2.1.10)) show how the maximum-likelihood principle implies a partition of the sample space, but do not simplify the exact probability expression for general n. Instead, they typically resort to *asymptotic* results or numerical approximations.

6 Step 6: Asymptotic Consistency (Key Theoretical Result)

Main Theorem (Consistency): Under mild regularity conditions (e.g. $p_1 \neq p_2$ and τ not too close to the boundaries), the MLE $\hat{\tau}$ is *consistent*, meaning that

$$\lim_{n \to \infty} \mathbb{P}\big(\hat{\tau} = \tau\big) \;=\; 1.$$

Equivalently, $\hat{\tau}$ converges in probability to τ. If τ scales with n (e.g. $\tau = \lfloor \theta n \rfloor$ for some $\theta \in (0, 1)$), one can prove $\hat{\tau}/n \to \tau/n$ in probability.

6.1 Connection to Literature

- **Basseville and Nikiforov (1993) [1], Chapter 2:** They give a thorough treatment of *offline* change detection, showing that *log-likelihood ratio* type methods are consistent provided the Kullback–Leibler divergence between the distributions before and after the change is non-zero (which it is here, since $p_1 \neq p_2$).

  $$D_{\mathrm{KL}}\big(\text{Bernoulli}(p_1) \,\|\, \text{Bernoulli}(p_2)\big) \;>\; 0,$$

  ensures that the probability of detecting the correct change point tends to 1 as $n \to \infty$.

- **Csörgő and Horváth (1997) [2], Chapter 1:** They prove limit theorems specifically for change-point estimators in i.i.d. data. A key statement (e.g. their Theorem 1.2.3) is that the MLE for a single change point is strongly consistent in the sense that $|\hat{\tau} - \tau| = O(\ln(n))$ almost surely, under suitable conditions. This immediately implies

$$\mathbb{P}\big(\hat{\tau} = \tau\big) \;\to\; 1 \quad \text{as } n \to \infty \quad \text{if } \tau \text{ is not too close to } n.$$

- **How exactly is the result used here?** We apply the known theorem *directly* to our Bernoulli model. Because the model is straightforward and $p_1 \neq p_2$, the divergence is positive, so the MLE must locate the change point consistently.

6.2 Finite-Sample Bounds

Even though an exact expression for $\mathbb{P}(\hat{\tau} = \tau)$ is complicated, large-deviation inequalities (such as Chernoff bounds) can be applied to show that *misplacing* the change point is exponentially unlikely for large n (as long as $p_1 \neq p_2$). For instance, one can often derive:

$$\mathbb{P}\big(|\hat{\tau} - \tau| > \delta\big) \;\leq\; C\, e^{-c\, n},$$

for some constants $C, c > 0$. By summing over possible values of τ or bounding them, one can deduce:

$$\mathbb{P}\big(\hat{\tau} = \tau\big) \;\geq\; 1 - C\, e^{-c\, n},$$

showing that this probability approaches 1 at an exponential rate.

6.3 Log-Likelihood Differences and Chernoff Bound Application

We define the log-likelihood difference:

$$D(\tau, k) = \ell(\tau) - \ell(k). \tag{4}$$

Correct detection of τ means that:

$$D(\tau, k) \geq 0, \quad \forall k \neq \tau. \tag{5}$$

Thus, the probability of incorrectly estimating τ is:

$$\mathbb{P}(\hat{\tau} \neq \tau) = \mathbb{P}\left(\exists k \neq \tau \text{ such that } D(\tau, k) < 0\right). \tag{6}$$

6.4 Chernoff Bound Application

Chernoff bounds provide an upper bound on the probability of deviations of sums of independent random variables. In our case:

1. $D(\tau, k)$ can be expressed as a sum of log-likelihood ratios, which behave like a sum of independent random variables.

2. Using the Chernoff bound:

$$\mathbb{P}(D(\tau, k) < 0) = \mathbb{P}\left(\sum_{i=1}^{n} Z_i \leq 0\right) \leq e^{-cn}, \tag{7}$$

   for some $c > 0$, where Z_i represents log-likelihood increments that follow a sub-Gaussian distribution.

3. Taking a union bound over all k leads to:

$$\mathbb{P}(|\hat{\tau} - \tau| > \delta) \leq Ce^{-cn}, \tag{8}$$

   where C accounts for the number of terms in the union bound.

7 Step 7: Conclusion

- **We established** that

$$\mathbb{P}\big(\hat{\tau} = \tau\big) \;=\; \mathbb{P}\Big(\ell(\tau) \geq \ell(k)\ \forall k\Big)$$

  cannot, in general, be simplified to a closed-form for finite n. However, it *does* admit either:

  1. A direct summation over $\{0, 1\}^n$ (impractical for large n), or
  2. Probabilistic bounds using concentration inequalities,
  3. Asymptotic results guaranteeing $\mathbb{P}(\hat{\tau} = \tau) \to 1$.

- **Hence the probability of detecting the correct change point** τ goes to 1 as $n \to \infty$, given $p_1 \neq p_2$.

- **Significance:** This final statement is crucial in offline change-point detection research: it shows that with enough observations, the MLE method almost surely recovers the true change point.

Summary of Step-by-Step Logical Flow:

1. *(Setup)* State the Bernoulli model with a single unknown change point τ.

2. *(Likelihood)* Derive the log-likelihood for any candidate k.

3. *(MLE)* Define $\hat{\tau}$ to maximize the log-likelihood over k.

4. *(Probability Event)* Note $\{\hat{\tau} = \tau\} = \{\ell(\tau) \geq \ell(k)\, \forall k\}$.

5. *(Exact Expression)* Observe that $\mathbb{P}(\hat{\tau} = \tau)$ can be written as a large sum over all possible binary sequences.

6. *(Consistency Theorem)* Invoke the known results (e.g. from [1], [2]) that MLE-based change-point estimators are consistent when $p_1 \neq p_2$.

7. *(Conclusion)* As $n \to \infty$, $\mathbb{P}(\hat{\tau} = \tau) \to 1$.

References

[1] M. Basseville and I. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, 1993. *Usage in this document:* We adopt their general treatment of change-point detection by log-likelihood ratio and their demonstration (Chapter 2) that a nonzero Kullback–Leibler divergence yields consistency of the estimator.

[2] M. Csörgő and L. Horváth. *Limit Theorems in Change-Point Analysis*. John Wiley & Sons, 1997. *Usage in this document:* They provide rigorous limit theorems (Theorem 1.2.3, Chapter 1, among others) showing that, under mild conditions, the maximum-likelihood change-point estimator is strongly consistent, implying $\mathbb{P}(\hat{\tau} = \tau) \to 1$.

[3] J. Bai. Common breaks in means and variances for panel data. *Journal of Econometrics*, 157(1):78–92, 2010. *Usage in this document:* While focusing on panel data, Bai provides insights on multiple structural breaks, illustrating the extension of classical single-break (change-point) asymptotics to more complex settings. The single-break Bernoulli case is a special instance of these more general frameworks.