보조정리 1 — 연속 구간 정리

내림차순으로 정렬된 실수열

$$s_1 \geq s_2 \geq \cdots \geq s_n$$

을 k개의 그룹 $G = G_1, \ldots, G_k$ 으로 분할한다고 하자.
목적 함수는

$$L(G) = \sum_{g \in G} \mathrm{Var}(g), \qquad \mathrm{Var}(g) = \frac{1}{|g|} \sum_{x \in g} (x - \mu_g)^2, \quad \mu_g = \frac{1}{|g|} \sum_{x \in g} x.$$

> **정리**
> $L(G)$를 최소화하는 최적 분할 $G^{\star}$의 각 그룹은 반드시 **연속한 인덱스 구간**이다.

증명 — 교환(Exchange) 논법

1. 모순 가정

최적이지만 비연속 그룹을 포함하는 분할을 $G^{\star}$라 가정한다.
따라서

$$i < j < k, \qquad s_i, , s_k \in G_1, s_j \in G_2, \qquad G_1, G_2 \in G^{\star}$$

인 (i, j, k)가 존재한다.

2. 표기

- $m = |G_1|, n = |G_2|$
- $\mu_1 = \mu_{G_1}, \mu_2 = \mu_{G_2}$
- $V_1 = \mathrm{Var}(G_1), V_2 = \mathrm{Var}(G_2)$
- $x = s_j$

3. 보조정리 (분산 갱신 식)

(a) 원소 추가

집합 A의 크기를 $|A| = m$, 평균을 μ_A, 분산을 V_A라 하자.
새 원소 x를 추가하여 $A^+ = A \cup x$ 를 만들면 $|A^+| = m + 1$ 이고 새로운 평균은

$$\mu_{A^+};=;\frac{m\mu_A + x}{m+1}.$$

모든 원소에 대한 제곱 편차 합(총 제곱오차, SSE)은

\$\$
₩text{SSE}*{A^+}
= ₩sum*{y₩in A^+} (y-₩mu_{A^+})^2
= ₩sum_{y₩in A}(y-₩mu_{A^+})^2 + (x-₩mu_{A^+})^2
\$\$

평균 이동량 $\mu_{A^+}-\mu_A=\dfrac{x-\mu_A}{m+1}$ 을 이용하여
$₩sum_{y₩in A}(y-₩mu_{A^+})^2 = ₩text{SSE}A + m(₩mu_A-₩mu{A^+})^2$ 임을 계산하면

$$
₩text{SSE}{A^+}
= mV_A + m₩Bigl(₩tfrac{x-₩mu_A}{m+1}₩Bigr)^2 + (x-₩mu{A^+})^2
= mV_A + ₩frac{(x-₩mu_A)^2}{m+1}.
$$

따라서

$$V_{A^+}=\frac{\mathrm{SSE}_{A^+}}{|A^+|}=\frac{m}{m+1}V_A;+;\frac{(x-\mu_A)^2}{m+1}.$$

(b) 원소 제거

집합 B의 크기를 $|B|=n(!\geq 2)$, 평균을 μ_B, 분산을 V_B라 하고
원소 $x\in B$를 제거하여 $B^-=B\setminus x$를 만들면 $|B^-|=n-1$ 이다.
새 평균은

$$\mu_{B^-};=;\frac{n\mu_B-x}{n-1}.$$

마찬가지로 SSE_{B^-}를 전개하면

$$
₩text{SSE}_{B^-}
= ₩text{SSE}B - (x-₩mu_B)^2 - (n-1)(₩mu_B-₩mu{B^-})^2,
$$

여기서 $\mu_B-\mu_{B^-}=\dfrac{x-\mu_B}{,n-1,}$ 이므로

$$\mathrm{SSE}_{B^-}=nV_B-(x-\mu_B)^2-\frac{(x-\mu_B)^2}{,n-1,}=nV_B-\frac{n}{,n-1,}(x-\mu_B)^2.$$

따라서

$$V_{B^-}=\frac{\mathrm{SSE}_{B^-}}{|B^-|}=\frac{n}{,n-1,}V_B;-;\frac{(x-\mu_B)^2}{,n-1,}.$$

이 두 식을 이용하면 교환 분할 후 목적 함수 변화량 ΔL 을 식 (1)로 바로 계산할 수 있다.

4. s_j를 G_1로 옮긴 분할 G'

$$G_1'=G_1\cup x,\qquad G_2'=G_2\setminus x,\qquad G'=G^\star\setminus G_1,G_2\cup G_1',G_2'.$$

전체 목적 함수 변화량은

$$\Delta L=\bigl[V_{G_1'}-V_1\bigr]+\bigl[V_{G_2'}-V_2\bigr]=\frac{(x-\mu_1)^2-V_1}{m+1}+\frac{V_2-(x-\mu_2)^2}{n-1} \tag{1}$$

5. 부등식 평가

내림차순이므로 μ_1은 구간 $[s_k, s_i]$ 안에 있다 ⇒

$$(x-\mu_1)^2 \le \max((s_i-\mu_1)^2, (s_k-\mu_1)^2) \le V_1. \tag{2}$$

따라서 첫째 항 ≤ 0.

한편

$$V_2 = \frac{1}{n}\big[(x-\mu_2)^2 + (\text{나머지 제곱합})\big] \ge \frac{(x-\mu_2)^2}{n} \Longrightarrow V_2 - (x-\mu_2)^2 \le -\frac{n-1}{n}(x-\mu_2)^2 < 0 \tag{3}$$

즉 둘째 항도 < 0이다.
(1)–(3)으로부터

$$\Delta L < 0$$

6. 모순 도출

$\Delta L < 0$이면 G'가 $G^\star$보다 더 작은 목적 값을 갖는다.
이는 $G^\star$를 **최적** 이라 가정한 것과 모순.

$n = 1$인 경우(즉 $G_2 = x$)에는 s_i 또는 s_k를 G_2로 옮기는 대칭 교환을 취해도 같은 방식으로 $\Delta L < 0$임을 확인할 수 있다.

7. 결론

따라서 최적 분할에 비연속 그룹이 존재할 수 없다.
즉, **최적 분할은 각 그룹이 연속한 인덱스 구간으로 이루어진다. ■**