KL Divergence (Relative Entropy) / lower bound

Theorem 14.1: KL Divergence (Relative Entropy)

1. Measure Space

$$(\Omega, \mathcal{F})$$

- Ω : sample space (all possible outcomes)
- $\mathcal{F}$: collection of events (σ-algebra)

$\Rightarrow$ **A space on which probabilities can be defined**

2. Probability Measures P, Q

- P, Q : probability measures on the same $(\Omega, \mathcal{F})$
- Not numbers, not functions
- Measures assigning probabilities to events

3. Definition of KL Divergence

$$D(P,Q) = \begin{cases} \displaystyle\int \log\left(\frac{dP}{dQ}(\omega)\right) dP(\omega), & \text{if } P \ll Q, \\ \infty, & \text{otherwise.} \end{cases}$$

Meaning

- How different P is from Q when viewed under Q
- Expected log-likelihood ratio under P

4. Absolute Continuity ($P \ll Q$)

$$Q(A) = 0 \;\Rightarrow\; P(A) = 0$$

- Events impossible under Q are also impossible under P
- Without this, the log-ratio explodes and $D(P, Q) = \infty$

5. Common Dominating Measure λ

A universal choice:

$$\boxed{\lambda = P + Q}$$

Then

$$P \ll \lambda, \qquad Q \ll \lambda$$

6. Radon–Nikodym Derivatives (Densities)

$$p = \frac{dP}{d\lambda}, \qquad q = \frac{dQ}{d\lambda}$$

Key Property

$$P(A) = \int_A p \, d\lambda, \qquad Q(A) = \int_A q \, d\lambda$$

7. Density Form of KL Divergence

By the chain rule:

$$\frac{dP}{dQ} = \frac{dP/d\lambda}{dQ/d\lambda} = \frac{p}{q}$$

Hence,

$$\boxed{D(P, Q) = \int p \log \frac{p}{q} \, d\lambda}$$

(Books often omit $d\lambda$.)

8. Basic Properties of KL Divergence

- $D(P, Q) \geq 0$
- $D(P, Q) = 0 \iff P = Q$ (almost everywhere)
- Not symmetric
- No triangle inequality
  $\Rightarrow$ **Not a metric**

9. Common KL Formulas

(1) Normal Distributions (Same Variance)

$$P = \mathcal{N}(\mu_1, \sigma^2), \quad Q = \mathcal{N}(\mu_2, \sigma^2)$$

$$\boxed{D(P, Q) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}}$$

(2) Bernoulli Distributions

$$P = \mathrm{Bern}(p), \quad Q = \mathrm{Bern}(q)$$

$$\boxed{D(P, Q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}}$$

Final One-Line Summary

> **KL divergence is the expectation (under P) of the log density ratio between P and Q on events where P typically occurs.**

Theorem 14.2: Bretagnolle–Huber Inequality

Goal

For any measurable set $A \in \mathcal{F}$,

$$\boxed{P(A) + Q(A^c) \;\geq\; \frac{1}{2} \exp\big(-D(P, Q)\big)}$$

Overall Strategy (One-Line Summary)

1. Lower bound $P(A) + Q(A^c)$ by $\displaystyle\int p \wedge q$
2. Lower bound $\displaystyle\int p \wedge q$ by $\displaystyle\frac{1}{2}\left(\int \sqrt{pq}\right)^2$
3. Lower bound $\displaystyle\left(\int \sqrt{pq}\right)^2$ by $\exp\big(-D(P, Q)\big)$

Chaining these inequalities yields the result.

Step 1: $P(A) + Q(A^c) \geq \int p \wedge q$

Decompose:

$$\int p \wedge q = \int_A p \wedge q + \int_{A^c} p \wedge q$$

On A:

$$p \wedge q \leq p$$

On A^c:

$$p \wedge q \leq q$$

Thus,

$$\int p \wedge q \leq \int_A p + \int_{A^c} q = P(A) + Q(A^c)$$

Hence,

$$\boxed{P(A) + Q(A^c) \ \geq \ \int p \wedge q}$$

Step 2: $\int p \wedge q \geq \frac{1}{2} \left(\int \sqrt{pq} \right)^2$

Key identity:

$$pq = (p \wedge q)(p \vee q)$$

Therefore,

$$\int \sqrt{pq} = \int \sqrt{(p \wedge q)(p \vee q)}$$

Apply Cauchy–Schwarz:

$$\left(\int \sqrt{pq} \right)^2 \leq \left(\int p \wedge q \right) \left(\int p \vee q \right)$$

Since

$$p \wedge q + p \vee q = p + q,$$

we have

$$\int p \vee q = \int (p + q) - \int p \wedge q = 2 - \int p \wedge q \leq 2$$

Thus,

$$\left(\int \sqrt{pq} \right)^2 \leq 2 \int p \wedge q$$

Rearranging,

$$\boxed{\int p \wedge q \ \geq \ \frac{1}{2} \left(\int \sqrt{pq} \right)^2}$$

Step 3: $\left(\int \sqrt{pq}\right)^2 \geq \exp\big(-D(P,Q)\big)$

Rewrite:

$$\left(\int \sqrt{pq}\right)^2 = \exp\left(2\log \int \sqrt{pq}\right)$$

Since

$$\sqrt{pq} = p\sqrt{\frac{q}{p}} \quad (p>0),$$

we obtain

$$\exp\left(2\log \int p\sqrt{\frac{q}{p}}\right)$$

Because log is concave and $p\,d\lambda$ is a probability measure, Jensen's inequality gives

$$\log\left(\int p\sqrt{\frac{q}{p}}\right) \geq \int p\log\left(\sqrt{\frac{q}{p}}\right)$$

Multiplying by 2 and exponentiating,

$$\exp\left(2\log \int p\sqrt{\frac{q}{p}}\right) \geq \exp\left(2\int p\log\sqrt{\frac{q}{p}}\right)$$

Since

$$\log\sqrt{\frac{q}{p}} = \frac{1}{2}\log\frac{q}{p},$$

we have

$$2\int p\log\sqrt{\frac{q}{p}} = \int p\log\frac{q}{p} = -\int p\log\frac{p}{q}$$

By definition of KL divergence,

$$D(P,Q) = \int p\log\frac{p}{q}$$

Therefore,

$$\boxed{\left(\int \sqrt{pq}\right)^2 \;\geq\; \exp\big(-D(P,Q)\big)}$$

Final Step: Chaining All Inequalities

Combining all steps,

$$P(A)+Q(A^c) \;\geq\; \int p\wedge q \;\geq\; \frac{1}{2}\left(\int \sqrt{pq}\right)^2 \;\geq\; \frac{1}{2}\exp\big(-D(P,Q)\big)$$

Hence,

$$\boxed{P(A)+Q(A^c) \;\geq\; \frac{1}{2}\exp\big(-D(P,Q)\big)}$$

Remark

A common mistake is reversing the final inequality. The correct direction is $\geq$, not $\leq$.

Interpretation

- $P(A)$: error under distribution P
- $Q(A^c)$: error under distribution Q
- Their sum represents the fundamental limitation of simultaneously distinguishing P and Q

If $D(P, Q)$ is small, no decision rule can perform well under both distributions.

Lemma 15.1: Divergence Decomposition

Statement (Goal)

Consider two bandit environments

$$\nu = (P_1, \ldots, P_k), \qquad \nu' = (P'_1, \ldots, P'_k),$$

and a fixed policy π run for n rounds.

Let $\mathbb{P}_\nu$ and $\mathbb{P}_{\nu'}$ denote the induced distributions over the entire interaction trajectory. Then,

$$\boxed{D(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{i=1}^{k} \mathbb{E}_\nu[T_i(n)]\, D(P_i, P'_i)}$$

1. Probability Space

A trajectory is given by

$$\omega = (a_1, x_1, \ldots, a_n, x_n),$$

where

- a_t is the arm selected at time t,
- x_t is the observed reward.

The policy π is fixed:

- arm selection probabilities are identical under ν and ν',
- only the reward distributions differ.

2. Factorization of Path Distributions

Under environment ν,

$$\mathbb{P}_\nu(\omega) = \prod_{t=1}^{n} \pi(a_t \mid h_{t-1})\, P_{a_t}(x_t).$$

Under environment ν',

$$\mathbb{P}_{\nu'}(\omega) = \prod_{t=1}^{n} \pi(a_t \mid h_{t-1})\, P'_{a_t}(x_t).$$

The policy terms are identical in both distributions.

3. Radon–Nikodym Derivative

Canceling the policy terms yields

$$\log \frac{d\mathbb{P}_\nu}{d\mathbb{P}_{\nu'}}(\omega) = \sum_{t=1}^{n} \log \frac{P_{a_t}(x_t)}{P'_{a_t}(x_t)}.$$

This decomposition is the starting point of the proof.

4. Plug into the KL Definition

By definition,

$$D(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \mathbb{E}_\nu \left[\log \frac{d\mathbb{P}_\nu}{d\mathbb{P}_{\nu'}} \right].$$

Substituting the expression above and using linearity of expectation,

$$D(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{t=1}^{n} \mathbb{E}_\nu \left[\log \frac{P_{A_t}(X_t)}{P'_{A_t}(X_t)} \right].$$

5. Tower Property

Apply the tower property:

$$\mathbb{E}_\nu \left[\log \frac{P_{A_t}(X_t)}{P'_{A_t}(X_t)} \right] = \mathbb{E}_\nu \left[\mathbb{E}_\nu \left(\log \frac{P_{A_t}(X_t)}{P'_{A_t}(X_t)} \,\middle|\, A_t \right) \right].$$

Conditionally on $A_t = i$, we have $X_t \sim P_i$, hence

$$\mathbb{E}_\nu \left[\log \frac{P_{A_t}(X_t)}{P'_{A_t}(X_t)} \,\middle|\, A_t = i \right] = D(P_i, P'_i).$$

6. Grouping by Arms

Therefore,

$$\mathbb{E}_\nu\left[\log\frac{P_{A_t}(X_t)}{P'_{A_t}(X_t)}\right] = \sum_{i=1}^{k}\mathbb{P}_\nu(A_t=i)\,D(P_i,P'_i).$$

Summing over $t=1,\dots,n$,

$$D(\mathbb{P}_\nu,\mathbb{P}_{\nu'}) = \sum_{i=1}^{k}\left(\sum_{t=1}^{n}\mathbb{P}_\nu(A_t=i)\right)D(P_i,P'_i).$$

7. Appearance of $T_i(n)$

Define the number of pulls of arm i:

$$T_i(n) = \sum_{t=1}^{n}\mathbf{1}\{A_t=i\}.$$

Then,

$$\sum_{t=1}^{n}\mathbb{P}_\nu(A_t=i) = \mathbb{E}_\nu[T_i(n)].$$

Conclusion

$$\boxed{D(\mathbb{P}_\nu,\mathbb{P}_{\nu'}) = \sum_{i=1}^{k}\mathbb{E}_\nu[T_i(n)]\,D(P_i,P'_i)}$$

Interpretation

> The total information for distinguishing two environments equals the sum, over arms, of the expected number of observations of each arm multiplied by the divergence between its reward distributions.

Arms that are rarely pulled contribute little information, which is the key mechanism behind bandit lower bounds.

15.2 Minimax Regret Lower Bound for Gaussian Bandits

Final Goal

For any policy π,

$$\boxed{\inf_\pi\sup_\nu R_n(\pi,\nu) \;\geq\; \frac{1}{27}\sqrt{n(k-1)}}$$

1. One-Line Proof Strategy

Construct two environments that differ only on an arm the policy barely explores, and force regret via indistinguishability and hypothesis testing bounds.

2. Construction of Two Environments

(a) First Environment ν_μ

We consider a Gaussian bandit with variance 1 and mean vector

$$\mu = (\Delta, 0, 0, \ldots, 0).$$

Arm 1 is optimal.

(b) Choosing a Rarely Pulled Arm

Define

$$i = \arg\min_{j>1} \mathbb{E}_\mu[T_j(n)].$$

This arm receives the least information under the policy. By averaging,

$$\mathbb{E}_\mu[T_i(n)] \leq \frac{n}{k-1}.$$

(c) Second Environment $\nu_{\mu'}$

Define a modified mean vector

$$\mu' = (\Delta, 0, \ldots, 2\Delta, \ldots, 0),$$

where only the i-th arm is changed. Now arm i is optimal.

3. Lower Bounding Regret via Bad Events

Key Idea

- Under ν_μ, pulling the optimal arm (arm 1) at most $n/2$ times incurs large regret.
- Under $\nu_{\mu'}$, pulling arm 1 more than $n/2$ times incurs large regret.

Formal Bounds

$$R_n(\pi, \nu_\mu) \;\geq\; \mathbb{P}_\mu\Big(T_1(n) \leq \frac{n}{2}\Big) \cdot \frac{n\Delta}{2},$$

$$R_n(\pi, \nu_{\mu'}) \;\geq\; \mathbb{P}_{\mu'}\Big(T_1(n) > \frac{n}{2}\Big) \cdot \frac{n\Delta}{2}.$$

4. Bretagnolle–Huber Inequality

Let

$$P = \mathbb{P}_\mu, \qquad Q = \mathbb{P}_{\mu'}, \qquad A = \{T_1(n) \le n/2\}.$$

By Theorem 14.2,

$$\mathbb{P}_\mu(A) + \mathbb{P}_{\mu'}(A^c) \;\ge\; \frac{1}{2}\exp\big(-D(\mathbb{P}_\mu, \mathbb{P}_{\mu'})\big).$$

5. Role of Lemma 15.1 (KL Decomposition)

Only arm i differs between ν_μ and $\nu_{\mu'}$, hence

$$D(\mathbb{P}_\mu, \mathbb{P}_{\mu'}) = \mathbb{E}_\mu[T_i(n)]\, D\big(\mathcal{N}(0,1), \mathcal{N}(2\Delta, 1)\big).$$

For Gaussians with equal variance,

$$D\big(\mathcal{N}(0,1), \mathcal{N}(2\Delta,1)\big) = 2\Delta^2.$$

Using $\mathbb{E}_\mu[T_i(n)] \le \frac{n}{k-1}$,

$$D(\mathbb{P}_\mu, \mathbb{P}_{\mu'}) \;\le\; \frac{2n\Delta^2}{k-1}.$$

6. Core Lower Bound

Combining the regret bounds with the Bretagnolle–Huber inequality,

$$R_n(\pi, \nu_\mu) + R_n(\pi, \nu_{\mu'}) \;\ge\; \frac{n\Delta}{4}\exp\left(-\frac{2n\Delta^2}{k-1}\right).$$

7. Choice of Δ

The parameter Δ is a design choice made by the analyst. We set

$$\Delta = \sqrt{\frac{k-1}{4n}}.$$

Then

$$\frac{2n\Delta^2}{k-1} = \frac{1}{2}, \qquad \exp\left(-\frac{2n\Delta^2}{k-1}\right) = e^{-1/2}.$$

8. From Two Environments to One

Using the elementary inequality $\max(a,b) \ge (a+b)/2$,

$$\max\big\{R_n(\pi, \nu_\mu),\; R_n(\pi, \nu_{\mu'})\big\} \;\ge\; \frac{n\Delta}{8}\exp\left(-\frac{2n\Delta^2}{k-1}\right).$$

Substituting the chosen Δ,

$$\max R_n \;\geq\; \frac{1}{16} e^{-1/2} \sqrt{n(k-1)}.$$

9. Explicit Constant

Since

$$\frac{1}{16} e^{-1/2} \approx 0.038 > \frac{1}{27},$$

we obtain the stated minimax lower bound

$$\boxed{\inf_{\pi} \sup_{\nu} R_n(\pi, \nu) \;\geq\; \frac{1}{27} \sqrt{n(k-1)}}.$$

One-Sentence Takeaway

Bandit lower bounds arise by attacking arms where the policy gathers too little information; KL divergence and hypothesis testing inequalities quantify the resulting indistinguishability and force regret.