The kl inequality

Yevgeny Seldin

Target

- Derive an inequality that is tighter than Hoeffding's

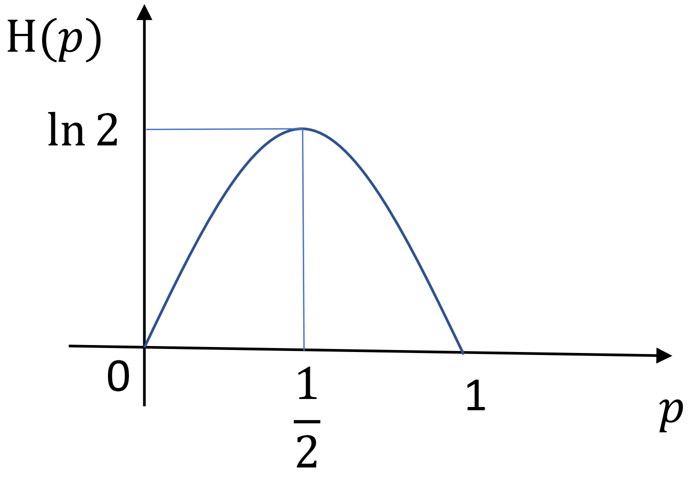
Basics in Information Theory

- Entropy of a distribution p

$$\mathrm{H}(p) = -\sum_{x} p(x) \ln p(x)$$

- Binary entropy
  - Entropy of a Bernoulli distribution $(1-p, p)$

$$\mathrm{H}(p) = \underbrace{-(1-p)\ln(1-p)}_{x=0} \underbrace{-p \ln p}_{x=1}$$



The method of types

0 0 0 0	1 0 0 0	1 1 0 0	0 1 1 1	1 1 1 1
	0 1 0 0	1 0 1 0	1 0 1 1	
	0 0 1 0	1 0 0 1	1 1 0 1	
	0 0 0 1	0 1 1 0	1 1 1 0	
		0 1 0 1		
		0 0 1 1		
$\binom{4}{0} = 1$	$\binom{4}{1} = 4$	$\binom{4}{2} = 6$	$\binom{4}{3} = 4$	$\binom{4}{4} = 1$

- All sequences within the same type have the same probability
- The probability of a type is the number of sequences times the probability of an individual sequence
- The probability of observing (the empirical error) $\hat{p}_n = \frac{k}{n}$ is the probability of observing the type $\frac{k}{n}$

$$\mathbb{P}\left(\hat{p}_n = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mathbb{P}\left(\hat{p}_n = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k}$$

Bound on the binomial coefficients

- Lemma: for $1 \leq k \leq n-1$

$$\frac{1}{2}\sqrt{\frac{n}{2k(n-k)}} \leq \binom{n}{k} e^{-n\mathrm{H}\left(\frac{k}{n}\right)} \leq \frac{1}{2}\sqrt{\frac{n}{k(n-k)}}$$

- Proof:
  - $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
  - Stirling's approximation of the factorial:
    - $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leq n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}$
  - Three lines of technical derivation (see lecture notes)

- Message: $e^{n\mathrm{H}\left(\frac{k}{n}\right)}$ is a good approximation of $\binom{n}{k}$ and $\binom{n}{k} e^{-n\mathrm{H}\left(\frac{k}{n}\right)}$ is "small"
- Note: for $k = 0$ and $k = n$ we have $\binom{n}{k} e^{-n\mathrm{H}\left(\frac{k}{n}\right)} = 1$, so the message is also valid

$$\mathrm{H}(p) = -\sum_x p(x) \ln p(x)$$

Kullback-Leibler (KL) divergence / relative entropy

- "Distance" between probability distributions p and q
- $\mathrm{KL}(p||q) = \sum_x p(x) \ln \frac{p(x)}{q(x)} = \mathbb{E}_{X\sim p}\left[\ln \frac{p(X)}{q(X)}\right] = \mathbb{E}_{X\sim p}\left[\ln \frac{1}{q(X)}\right] - \mathrm{H}(p)$
- Properties:
  - $\mathrm{KL}(p||p) = 0$
  - $\mathrm{KL}(p||q)$ is convex in the pair (p, q)
    - $\mathrm{KL}(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) \leq \lambda KL(p_1||q_1) + (1-\lambda)KL(p_2||q_2)$
  - Asymmetry: $\mathrm{KL}(p||q) \neq \mathrm{KL}(q||p)$
- Binary kl:

$$\mathrm{kl}(p||q) = \mathrm{KL}\big((1-p, p)||(1-q, q)\big) = (1-p)\ln\frac{1-p}{1-q} + p\ln\frac{p}{q}$$

$e^{-n\mathrm{kl}\left(\frac{k}{n}||p\right)}$ governs the probability of type $\frac{k}{n}$

$$\mathbb{P}\left(\hat{p}_n = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} e^{n\left(\frac{k}{n}\ln p + \frac{n-k}{n}\ln(1-p)\right)}$$

$$= \binom{n}{k} e^{-n\mathrm{H}\left(\frac{k}{n}\right)} e^{n\mathrm{H}\left(\frac{k}{n}\right)} e^{n\left(\frac{k}{n}\ln p + \frac{n-k}{n}\ln(1-p)\right)}$$

$$= \underbrace{\binom{n}{k} e^{-n\mathrm{H}\left(\frac{k}{n}\right)}}_{"small"} e^{-n\mathrm{kl}\left(\frac{k}{n}||p\right)}$$

$$\mathrm{KL}(p||q) = \mathbb{E}_{X\sim p}\left[\ln\frac{1}{q(X)}\right] - \mathrm{H}(p)$$

$$\frac{1}{2}\sqrt{\frac{n}{2k(n-k)}} \leq \binom{n}{k} e^{-n\mathrm{H}\left(\frac{k}{n}\right)} \leq \frac{1}{2}\sqrt{\frac{n}{k(n-k)}}$$

- Message:
  - $\mathbb{P}\left(\hat{p}_n = \frac{k}{n}\right) \approx e^{-n\mathrm{kl}\left(\frac{k}{n}||p\right)}$
  - $e^{-n\mathrm{kl}\left(\frac{k}{n}||p\right)}$ governs the probability of observing type $\frac{k}{n}$ when sampling from p

The $\mathbf{kl}$ lemma

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \mathbb{P}(X = x)$$

$$\hat{p}_n \in \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\right\}$$

$$\mathbb{P}\left(\hat{p}_n = \frac{k}{n}\right) = \underbrace{\binom{n}{k} e^{-n\mathrm{H}\left(\frac{k}{n}\right)}}_{\text{"}small\text{"}} e^{-n\mathrm{kl}\left(\frac{k}{n}||p\right)}$$

- Lemma: $\mathbb{E}\left[e^{n\mathrm{kl}(\hat{p}_n||p)}\right] \leq 2\sqrt{n}$

- Proof:

$$\mathbb{E}\left[e^{n\mathrm{kl}(\hat{p}_n||p)}\right]$$

$$= \sum_{k=0}^{n} \mathbb{P}\left(\hat{p}_n = \frac{k}{n}\right) e^{n\mathrm{kl}\left(\frac{k}{n}||p\right)}$$

$$= \sum_{k=0}^{n} \underbrace{\binom{n}{k} e^{-n\mathrm{H}\left(\frac{k}{n}\right)}}_{\text{"}small\text{"}} \underbrace{e^{-n\mathrm{kl}\left(\frac{k}{n}||p\right)} e^{n\mathrm{kl}\left(\frac{k}{n}||p\right)}}_{=1}$$

$$\leq 2\sqrt{n}$$

The $\mathbf{kl}$ lemma is tight

- Lemma: $\mathbb{E}\left[e^{n\mathrm{kl}(\hat{p}_n||p)}\right] \leq 2\sqrt{n}$

- Lemma: for $p \in (0,1)$: $\mathbb{E}\left[e^{n\mathrm{kl}(\hat{p}_n||p)}\right] \geq \sqrt{n}$

- Proof:

$$\mathbb{E}\left[e^{n\mathrm{kl}(\hat{p}_n||p)}\right] = \sum_{k=0}^{n} \mathbb{P}\left(\hat{p}_n = \frac{k}{n}\right) e^{n\mathrm{kl}\left(\frac{k}{n}||p\right)}$$

$$= \sum_{k=0}^{n} \underbrace{\binom{n}{k} e^{-n\mathrm{H}\left(\frac{k}{n}\right)}}_{"small"} \underbrace{e^{-n\mathrm{kl}\left(\frac{k}{n}||p\right)} e^{n\mathrm{kl}\left(\frac{k}{n}||p\right)}}_{=1}$$

$$\geq \sqrt{n}$$

Markov:

$$\mathbb{P}(X \geq \varepsilon) \leq \frac{\mathbb{E}[X]}{\varepsilon}$$

The kl lemma:

$$\mathbb{E}\left[e^{n\mathrm{kl}(\hat{p}_n||p)}\right] \leq 2\sqrt{n}$$

The $\mathbf{kl}$ inequality via the $\mathbf{kl}$ lemma

- Theorem: $\mathbb{P}\left(\mathrm{kl}(\hat{p}_n||p) \geq \frac{\ln\frac{2\sqrt{n}}{\delta}}{n}\right) \leq \delta$

- Proof

$$\mathbb{P}\left(\mathrm{kl}(\hat{p}_n||p) \geq \frac{\ln\frac{2\sqrt{n}}{\delta}}{n}\right) = \mathbb{P}\left(n\mathrm{kl}(\hat{p}_n||p) \geq \ln\frac{2\sqrt{n}}{\delta}\right)$$

$$\underbrace{=}_{\substack{\text{Chernoff's} \\ \text{bounding} \\ \text{technique}}} \mathbb{P}\left(e^{n\mathrm{kl}(\hat{p}_n||p)} \geq \frac{2\sqrt{n}}{\delta}\right)$$

$$\underbrace{\leq}_{\text{Markov}} \frac{\delta}{2\sqrt{n}}\mathbb{E}\left[e^{n\mathrm{kl}(\hat{p}_n||p)}\right]$$

$$\underbrace{\leq}_{\text{kl lemma}} \delta$$

The kl inequality

- Theorem: $\mathbb{P}\left(\mathrm{kl}(\hat{p}_n||p) \geq \frac{\ln\frac{1}{\delta}}{n}\right) \leq \delta$

- Earlier: $\mathbb{P}\left(\mathrm{kl}(\hat{p}_n||p) \geq \frac{\ln\frac{2\sqrt{n}}{\delta}}{n}\right) \leq \delta$

- Proof:
  - Based on direct derivation (not via the kl lemma); omitted

- The direct derivation is incompatible with PAC-Bayesian analysis
  - There we will need to go via the kl lemma and pay $\ln 2\sqrt{n}$

Relaxations & comparison to Hoeffding

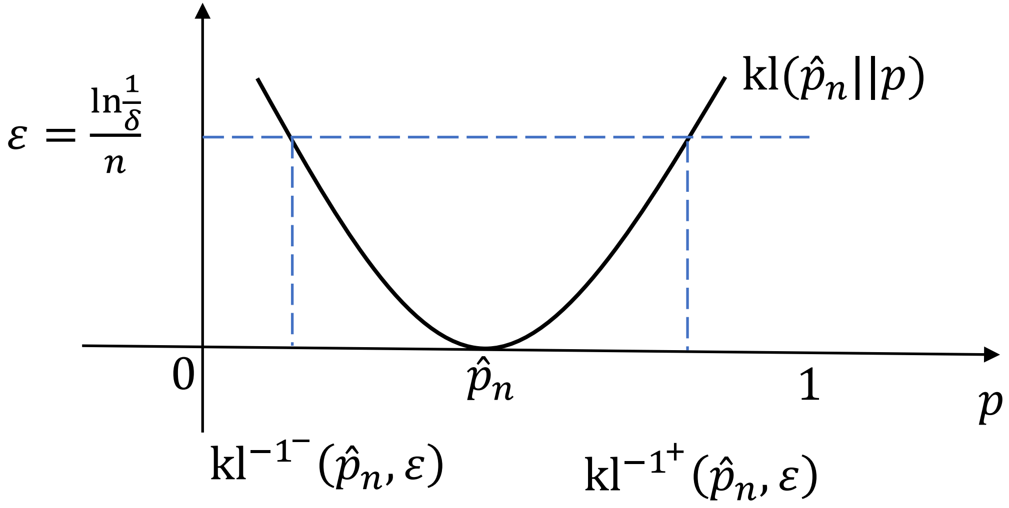
- The kl inequality: $\mathbb{P}\left(\mathrm{kl}(\hat{p}_n||p) \leq \frac{\ln\frac{1}{\delta}}{n}\right) \geq 1-\delta$

- Pinsker's inequality: $\mathrm{kl}(\hat{p}_n||p) \geq 2(p-\hat{p}_n)^2$
- Corollary: $\mathbb{P}\left(p \leq \hat{p}_n + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \geq 1-\delta$

- Hoeffding:
- $\mathbb{P}\left(p \leq \hat{p}_n + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \geq 1-\delta$

- Refined Pinsker's inequality: for $p > \hat{p}_n$, $\mathrm{kl}(\hat{p}_n||p) \geq \frac{(p-\hat{p}_n)^2}{2p}$
- Corollary: $\mathbb{P}\left(p \leq \hat{p}_n + \underbrace{\sqrt{\frac{2\hat{p}_n \ln\frac{n+1}{\delta}}{n}}}_{\to 0 \text{ for } \hat{p}_n \to 0} + \frac{2\ln\frac{n+1}{\delta}}{n}\right) \geq 1-\delta$
  - "Fast convergence rates" (at the rate of $\frac{1}{n}$ rather than $\frac{1}{\sqrt{n}}$)
  - (Significantly) tighter bound for $\hat{p}_n \ll \frac{1}{8}$
  - The kl inequality is even tighter

Inversion of kl



- $\mathbb{P}\left(\mathrm{kl}(\hat{p}_n||p) \leq \frac{\ln\frac{1}{\delta}}{n}\right) \geq 1 - \delta$

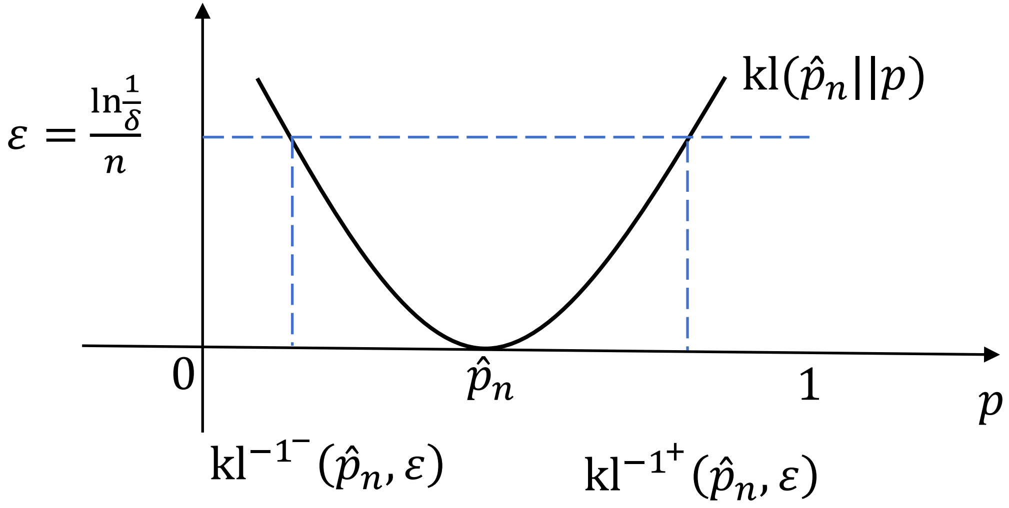
- Corollary:

$$\mathbb{P}\left(\mathrm{kl}^{-1^-}\left(\hat{p}_n, \frac{\ln\frac{1}{\delta}}{n}\right) \leq p \leq \mathrm{kl}^{-1^+}\left(\hat{p}_n, \frac{\ln\frac{1}{\delta}}{n}\right)\right) \geq 1 - \delta$$

$\mathrm{kl}^{-1^+}(\hat{p}_n, \varepsilon) = \max\{p\colon \mathrm{kl}(\hat{p}_n||p) \leq \varepsilon\}$; $\mathrm{kl}^{-1^-}(\hat{p}_n, \varepsilon) = \min\{p\colon \mathrm{kl}(\hat{p}_n||p) \leq \varepsilon\}$

- Inversion of kl:
  - $\mathrm{kl}(\hat{p}_n||p)$ is convex in p
  - $\mathrm{kl}(\hat{p}_n||\hat{p}_n) = 0$ is the minimum
  - $p \in [0,1]$
  - Use binary search on each side of $\hat{p}_n$

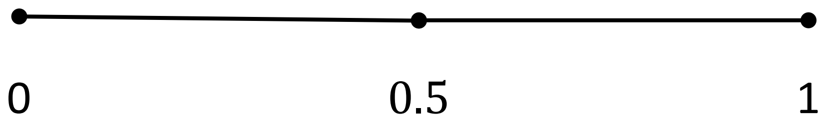
Summary



- kl lemma: $\mathbb{E}\left[e^{n\mathrm{kl}(\hat{p}_n||p)}\right] \leq 2\sqrt{n}$
- kl inequality: $\mathbb{P}\left(\mathrm{kl}(\hat{p}_n||p) \leq \frac{\ln\frac{1}{\delta}}{n}\right) \geq 1-\delta$
- Pinsker's relaxation: $\mathbb{P}\left(p \leq \hat{p}_n + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \geq 1-\delta$
- Refined Pinsker's relaxation: $\mathbb{P}\left(p \leq \hat{p}_n + \sqrt{\frac{2\hat{p}_n \ln\frac{1}{\delta}}{n}} + \frac{2\ln\frac{1}{\delta}}{n}\right) \geq 1-\delta$
  - "Fast rate"
- Direct inversion: $\mathbb{P}\left(\mathrm{kl}^{-1^-}\left(\hat{p}_n, \frac{\ln\frac{1}{\delta}}{n}\right) \leq p \leq \mathrm{kl}^{-1^+}\left(\hat{p}_n, \frac{\ln\frac{1}{\delta}}{n}\right)\right) \geq 1-\delta$
  - Use binary search

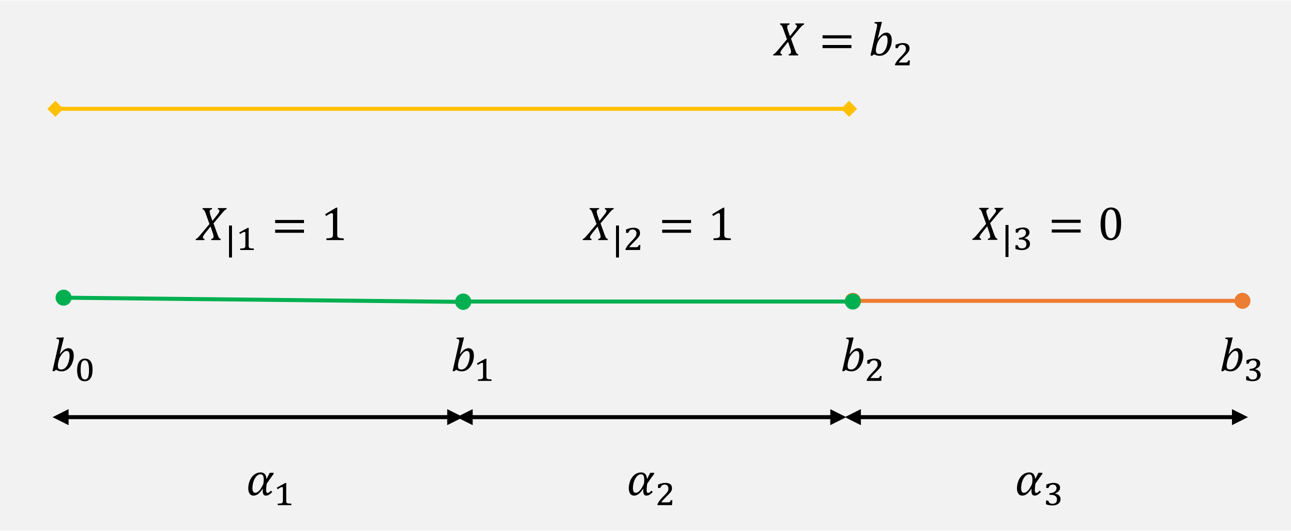
Split-kl inequality

- Motivation: the kl inequality is “blind” to the variance
  - $\mathbb{P}\left(\mathrm{kl}(\hat{p}_n||p) \leq \frac{\ln\frac{1}{\delta}}{n}\right) \geq 1 - \delta$



Split-kl inequality

- Solution for discrete random variables $X \in \{b_0, b_1, \dots, b_K\}$:
  - Representation as a superposition of Bernoulli random variables
  - $\alpha_j = b_j - b_{j-1}$
  - $X_{|j} = \mathbb{I}(X \geq b_j)$
    - "progress bar"
    - $X_{|j}$ is Bernoulli
  - $X = b_0 + \sum_{j=1}^{K} \alpha_j X_{|j}$



  - For $X_1, \dots, X_n$ let $\hat{p}_{|j} = \frac{1}{n} \sum_{i=1}^{n} X_{i|j}$
  - $\mathbb{E}[X] = p = b_0 + \sum_{j=1}^{K} \alpha_j \mathbb{E}\left[X_{|j}\right] = b_0 + \sum_{j=1}^{K} \alpha_j p_{|j}$
  - Apply kl inequality to bound the distance between $\hat{p}_{|j}$ and $p_{|j}$ for all j and take a union bound
  - Details in the lecture notes