DeLTA Seminar, University of Copenhagen, December 1, 2025

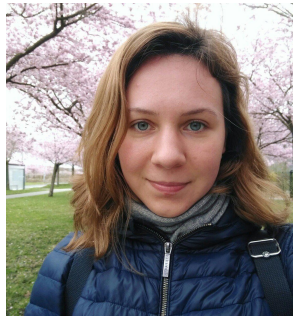
Nearly Minimax Discrete Distribution Estimation in Kullback-Leibler Divergence with High Probability

Tim van Erven

University of Amsterdam

Joint work with:

Dirk van der Hoeven



Julia Olkhovskaya

Also exists awesome concurrent paper by Mourtada (2025)!

Discrete Distribution Estimation

- Probability mass function p^* on **domain of size K**, i.e. on $\{1, \ldots, K\}$
- Sample S: independent draws $X_1, \ldots, X_n$ from p^*
- Goal: estimate p^* from S in Kullback-Leibler divergence:

$$\mathrm{KL}(p^* \| \hat{p}) = \sum_{i=1}^{K} p^*(i) \ln \frac{p^*(i)}{\hat{p}(i)}$$

Discrete Distribution Estimation

- Probability mass function p^* on **domain of size K**, i.e. on $\{1, \ldots, K\}$
- Sample S: independent draws $X_1, \ldots, X_n$ from p^*
- Goal: estimate p^* from S in Kullback-Leibler divergence:

$$\mathrm{KL}(p^* \| \hat{p}) = \sum_{i=1}^{K} p^*(i) \ln \frac{p^*(i)}{\hat{p}(i)} = \underbrace{\mathbb{E}[\ell(X, \hat{p})] - \mathbb{E}[\ell(X, p^*)]}_{\substack{\text{excess risk for} \\ \text{log loss } \ell(x,p) = -\ln p(x)}}$$

Discrete Distribution Estimation

- Probability mass function p^* on **domain of size K**, i.e. on $\{1, \ldots, K\}$
- Sample S: independent draws $X_1, \ldots, X_n$ from p^*
- Goal: estimate p^* from S in Kullback-Leibler divergence:

$$\mathrm{KL}(p^* \| \hat{p}) = \sum_{i=1}^{K} p^*(i) \ln \frac{p^*(i)}{\hat{p}(i)} = \underbrace{\mathbb{E}[\ell(X, \hat{p})] - \mathbb{E}[\ell(X, p^*)]}_{\substack{\text{excess risk for} \\ \text{log loss } \ell(x,p) = -\ln p(x)}}$$

Minimax rate in probability:

Smallest $r_n^*(\delta)$ for which there exists $\hat{p}$ that guarantees, for all possible p^*,

$$\mathbb{P}_{p^*}\Big(\mathrm{KL}(p^* \| \hat{p}) \leq r_n^*(\delta) \Big) \geq 1 - \delta.$$

Discrete Distribution Estimation

- Probability mass function p^* on **domain of size K**, i.e. on $\{1, \ldots, K\}$
- Sample S: independent draws $X_1, \ldots, X_n$ from p^*
- Goal: estimate p^* from S in Kullback-Leibler divergence:

$$\mathrm{KL}(p^* \| \hat{p}) = \sum_{i=1}^{K} p^*(i) \ln \frac{p^*(i)}{\hat{p}(i)} = \underbrace{\mathbb{E}[\ell(X, \hat{p})] - \mathbb{E}[\ell(X, p^*)]}_{\substack{\text{excess risk for} \\ \text{log loss } \ell(x,p) = -\ln p(x)}}$$

Minimax rate in probability:
Smallest $r_n^*(\delta)$ for which there exists $\hat{p}$ that guarantees, for all possible p^*,

$$\mathbb{P}_{p^*}\Big(\mathrm{KL}(p^* \| \hat{p}) \le r_n^*(\delta) \Big) \ge 1 - \delta.$$

Interest:

- Simplest setting with unbounded losses
- Cannot get optimal rate with existing techniques

Maximum Likelihood Estimator

Maximum likelihood estimator:

$$\hat{p}_{\mathrm{ml}}(i) = \frac{n_i}{n}, \qquad \text{where } n_i = |\{j \in [n] : X_j = i\}|$$

1. Achieves the minimax rate for smaller distances:

- $\hat{p}_{\mathrm{ml}}$ achieves the minimax rate $\sqrt{\frac{K+\ln(1/\delta)}{n}}$ for total variation distance V and Hellinger distance H (Canonne, 2020)

Maximum Likelihood Estimator

Maximum likelihood estimator:

$$\hat{p}_{\mathrm{ml}}(i) = \frac{n_i}{n}, \qquad \text{where } n_i = |\{j \in [n] : X_j = i\}|$$

1. Achieves the minimax rate for smaller distances:

- $\hat{p}_{\mathrm{ml}}$ achieves the minimax rate $\sqrt{\frac{K+\ln(1/\delta)}{n}}$ for total variation distance V and Hellinger distance H (Canonne, 2020)
- $V^2(p,q) \lesssim H^2(p,q) \leq \mathrm{KL}(p\|q)$, so minimax rate for KL is at least

$$r_n^*(\delta) \gtrsim \frac{K + \ln(1/\delta)}{n}$$

Maximum Likelihood Estimator

Maximum likelihood estimator:

$$\hat{p}_{\mathrm{ml}}(i) = \frac{n_i}{n}, \qquad \text{where } n_i = |\{j \in [n] : X_j = i\}|$$

$$r_n^*(\delta) \gtrsim \frac{K + \ln(1/\delta)}{n}$$

2. Matches this lower bound asymptotically as $n \to \infty$:

$$n\,\mathrm{KL}(p^* \| \hat{p}_{\mathrm{ml}}) \rightsquigarrow \tfrac{1}{2} Y, \quad \text{“converges in distribution”}$$

$$Y \lesssim M - 1 + \ln(1/\delta) \qquad \text{w.p.} \geq 1 - \delta$$

where

- Y has a χ^2 distribution with $M - 1$ degrees of freedom
- $M = |\{i : p^*(i) > 0\}| \leq K$ is size of support of p^*

Maximum Likelihood Estimator

Maximum likelihood estimator:

$$\hat{p}_{\mathrm{ml}}(i) = \frac{n_i}{n}, \qquad \text{where } n_i = |\{j \in [n] : X_j = i\}|$$

$$r_n^*(\delta) \gtrsim \frac{K + \ln(1/\delta)}{n}$$

3. . . . but fails catastrophically in minimax sense:

For any n, exist p^* such that
$\mathrm{KL}(p^* \| \hat{p}_{\mathrm{ml}}) = \infty$ with p^*-probability $> \delta$.

Reason:

- If $n_i = 0$, but $p^*(i) > 0$, then $\mathrm{KL}(p^* \| \hat{p}_{\mathrm{ml}}) = \infty$
- For any n, exist p^* with i such that $p^*(i) > 0$ but $\mathbb{P}_{p^*}\big(n_i = 0\big) > \delta$

Natural Fix: Add-γ Estimator

Add-γ estimator:

$$\hat{p}_\gamma(i) = \frac{n_i + \gamma}{n + K\gamma}, \qquad \text{where } n_i = |\{j \in [n] : X_j = i\}|$$

$\gamma = 1$: "Laplace estimator"

$\gamma = 1/2$: "Krichevsky-Trofimov estimator"

Natural Fix: Add-γ Estimator

Add-γ estimator:

$$\hat{p}_\gamma(i) = \frac{n_i + \gamma}{n + K\gamma}, \qquad \text{where } n_i = |\{j \in [n] : X_j = i\}|$$

Achieves the minimax rate **in expectation**:

1. Variant with data-dependent choice of γ achieves rate:

$$\min_{\hat{p}} \max_{p^*} \mathbb{E}_{p^*}\left[\mathrm{KL}(p^* \| \hat{p})\right] = \frac{K-1}{2n} + o\left(\frac{1}{n}\right) \qquad \text{(Braess \& Sauer, 2004)}$$

Natural Fix: Add-γ Estimator

Add-γ estimator:

$$\hat{p}_\gamma(i) = \frac{n_i + \gamma}{n + K\gamma}, \qquad \text{where } n_i = |\{j \in [n] : X_j = i\}|$$

Achieves the minimax rate **in expectation**:

1. Variant with data-dependent choice of γ achieves rate:

$$\min_{\hat{p}} \max_{p^*} \mathbb{E}_{p^*}\left[\mathrm{KL}(p^* \| \hat{p})\right] = \frac{K-1}{2n} + o\left(\frac{1}{n}\right) \qquad \text{(Braess \& Sauer, 2004)}$$

2. For $\gamma = 1$ achieves this rate up to factor of 2:

$$\max_{p^*} \mathbb{E}_{p^*}\left[\mathrm{KL}(p^* \| \hat{p}_1)\right] \leq \frac{K-1}{n} \qquad \text{(Catoni, 1997; Forster \& Warmuth, 2002)}$$

Minimax Rate with High Probability

Upper Bounds:

Author	Rate[1]	Estimator
Canonne et al. (2023)	$\frac{K}{n} + \frac{\sqrt{K \ln(K/\delta)^5}}{n}$	Laplace $\hat{p}_1$
v.d. Hoeven et al. (2023)	$\frac{K+\ln(n)\ln(1/\delta)}{n}$	Surrogate loss + online-to-batch (OTB)
Ours	$\frac{K \ln\ln(K)+\ln(K)\ln(1/\delta)}{n}$	OTB with suffix-averaging for $\hat{p}_\gamma$

[1] Up to constant factors

Minimax Rate with High Probability

Upper Bounds:

Author	Rate[1]	Estimator
Canonne et al. (2023)	$\frac{K}{n} + \frac{\sqrt{K \ln(K/\delta)^5}}{n}$	Laplace $\hat{p}_1$
v.d. Hoeven et al. (2023)	$\frac{K+\ln(n)\ln(1/\delta)}{n}$	Surrogate loss + online-to-batch (OTB)
Ours	$\frac{K \ln\ln(K)+\ln(K)\ln(1/\delta)}{n}$	OTB with suffix-averaging for $\hat{p}_\gamma$

Lower Bounds:

Author	Rate[1]	Technique
Canonne (2020)	$\frac{K+\ln(1/\delta)}{n}$	Assouad's method for total variation V
Polyanskiy & Wu (2025)	$\frac{K+\ln(1/\delta)}{n}$	Reduction to testing + Fano for V
Ours	$\frac{K+\ln(K)\ln(1/\delta)}{n}$	1. Ad hoc; 2. Reduction to "weak testing"

[1] Up to constant factors

Minimax Rate with High Probability

Upper Bounds:

Author	Rate[1]	Estimator
Canonne et al. (2023)	$\frac{K}{n} + \frac{\sqrt{K \ln(K/\delta)^5}}{n}$	Laplace $\hat{p}_1$
v.d. Hoeven et al. (2023)	$\frac{K+\ln(n)\ln(1/\delta)}{n}$	Surrogate loss + online-to-batch (OTB)
Ours	$\frac{K \ln\ln(K)+\ln(K)\ln(1/\delta)}{n}$	OTB with suffix-averaging for $\hat{p}_\gamma$

Lower Bounds:

Author	Rate[1]	Technique
Canonne (2020)	$\frac{K+\ln(1/\delta)}{n}$	Assouad's method for total variation V
Polyanskiy & Wu (2025)	$\frac{K+\ln(1/\delta)}{n}$	Reduction to testing + Fano for V
Ours	$\frac{K+\ln(K)\ln(1/\delta)}{n}$	1. Ad hoc; 2. Reduction to "weak testing"

[1] Up to constant factors

Minimax Rate with High Probability

Upper Bounds:

Author	Rate[1]	Estimator
Canonne et al. (2023)	$\frac{K}{n} + \frac{\sqrt{K \ln(K/\delta)^5}}{n}$	Laplace $\hat{p}_1$
v.d. Hoeven et al. (2023)	$\frac{K+\ln(n)\ln(1/\delta)}{n}$	Surrogate loss + online-to-batch (OTB)
Ours	$\frac{K\ln\ln(K)+\ln(K)\ln(1/\delta)}{n}$	OTB with suffix-averaging for $\hat{p}_\gamma$
Mourtada (2025)	$\frac{K+\ln(K)\ln(1/\delta)}{n}$	$\hat{p}_\gamma$ with $\gamma = 1 \vee \frac{\ln(1/\delta)}{K}$

Lower Bounds:

Author	Rate[1]	Technique
Canonne (2020)	$\frac{K+\ln(1/\delta)}{n}$	Assouad's method for total variation V
Polyanskiy & Wu (2025)	$\frac{K+\ln(1/\delta)}{n}$	Reduction to testing + Fano for V
Ours	$\frac{K+\ln(K)\ln(1/\delta)}{n}$	1. Ad hoc; 2. Reduction to "weak testing"
Mourtada (2025)	$\frac{K+\ln(K)\ln(1/\delta)}{n}$	Ad hoc

[1] Up to constant factors

Minimax Rate with High Probability

Upper Bounds:

Author	Rate[1]	Estimator
Canonne et al. (2023)	$\frac{K}{n} + \frac{\sqrt{K \ln(K/\delta)^5}}{n}$	Laplace $\hat{p}_1$
v.d. Hoeven et al. (2023)	$\frac{K+\ln(n)\ln(1/\delta)}{n}$	Surrogate loss + online-to-batch (OTB)
Ours	$\frac{K\ln\ln(K)+\ln(K)\ln(1/\delta)}{n}$	**OTB with suffix-averaging for $\hat{p}_\gamma$**
Mourtada (2025)	$\frac{K+\ln(K)\ln(1/\delta)}{n}$	**$\hat{p}_\gamma$ with $\gamma = 1 \vee \frac{\ln(1/\delta)}{K}$**

Lower Bounds:

Author	Rate[1]	Technique
Canonne (2020)	$\frac{K+\ln(1/\delta)}{n}$	Assouad's method for total variation V
Polyanskiy & Wu (2025)	$\frac{K+\ln(1/\delta)}{n}$	Reduction to testing + Fano for V
Ours	$\frac{K+\ln(K)\ln(1/\delta)}{n}$	1. Ad hoc; 2. Reduction to "weak testing"
Mourtada (2025)	$\frac{K+\ln(K)\ln(1/\delta)}{n}$	Ad hoc

[1] Up to constant factors

Minimax Rate with High Probability

Upper Bounds:

Author	Rate[1]	Estimator
Canonne et al. (2023)	$\frac{K}{n} + \frac{\sqrt{K \ln(K/\delta)^5}}{n}$	Laplace $\hat{p}_1$
v.d. Hoeven et al. (2023)	$\frac{K+\ln(n)\ln(1/\delta)}{n}$	Surrogate loss + online-to-batch (OTB)
Ours	$\frac{K\ln\ln(K)+\ln(K)\ln(1/\delta)}{n}$	OTB with suffix-averaging for $\hat{p}_\gamma$
Mourtada (2025)	$\frac{K+\ln(K)\ln(1/\delta)}{n}$	$\hat{p}_\gamma$ with $\gamma = 1 \vee \frac{\ln(1/\delta)}{K}$

Lower Bounds:

Author	Rate[1]	Technique
Canonne (2020)	$\frac{K+\ln(1/\delta)}{n}$	Assouad's method for total variation V
Polyanskiy & Wu (2025)	$\frac{K+\ln(1/\delta)}{n}$	Reduction to testing + Fano for V
Ours	$\frac{K+\ln(K)\ln(1/\delta)}{n}$	1. Ad hoc; **2. Reduction to "weak testing"**
Mourtada (2025)	$\frac{K+\ln(K)\ln(1/\delta)}{n}$	Ad hoc

[1] Up to constant factors

Upper Bound via Online-to-Batch Conversion

$$J = \max\left\{3, |\{i : n_i(n/2) < 32\ln(K/\delta)|\}\right\} \le K$$
$$\tilde{J} = \max\left\{3, |\{i : np^*(i) < 32\ln(K/\delta)\}|\right\} \le K$$

$$p_{t+1}(i) = \frac{n_i(t) + \gamma_i}{t + \sum_{j=1}^K \gamma_j}, \qquad p^{\mathrm{OTB}} = \frac{2}{n}\sum_{t=n/2+1}^{n} p_t$$

$$\gamma_i = \begin{cases} 0 & \text{if } n_i(n/2) \ge 32\ln(4K/\delta) \\ \max\{1, \frac{\ln(K/\delta)}{J}\} & \text{otherwise} \end{cases}$$

Theorem

If $n \ge \max\{\ln(K/\delta), K\}$, *then, with probability* $\ge 1-\delta$,

$$\mathrm{KL}(p^* \| p^{\mathrm{OTB}}) \lesssim \frac{K + \tilde{J}\ln\ln(\tilde{J}) + \ln(J)\ln(1/\delta)}{n}.$$

Proof Outline I

On an event $\mathcal{V}$ of probability $\geq 1-\delta$: online-to-batch + Freedman's concentration inequality for martingales + many bounds gives

$$\mathrm{KL}(p^* \| p^{\mathrm{OTB}}) \lesssim \frac{1}{n}\Big(\mathcal{R}_T + \sum_{i=1}^{K} \gamma_i + \ln(\beta)\ln(1/\delta)\Big),$$

where the regret on the second half of the data is

$$\mathcal{R}_T = \sum_{t=n/2+1}^{n} \log \frac{p^+(X_t)}{p_t(X_t)}$$

for $p^+(i) = \frac{np^*(i)+\gamma_i}{n+\sum_{j=1}^{K}\gamma_j}$ and $\ln\beta$ bounds the loss range:

$$\max_i \Big|\ln \frac{p^+(i)}{p_t(i)}\Big| \leq \ln\beta.$$

Proof Outline II

$$\mathcal{R}_T = \sum_{t=n/2+1}^{n} \log \frac{p^+(X_t)}{p_t(X_t)}$$

$$p_{t+1} = \arg\min_{p \in \Delta} \sum_{s=n/2+1}^{t} -\ln p(X_s) + R(p)$$

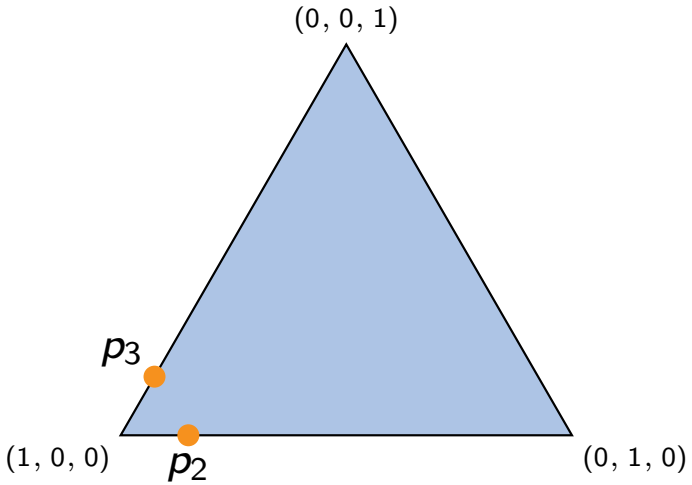
$$R(p) = \sum_{s=1}^{n/2} -\ln p(X_s) + \sum_{i=1}^{K} -\gamma_i \ln p(i)$$

Also on the event $\mathcal{V}$: FTRL-analysis + lots of bounding gives

$$\mathcal{R}_T \leq \sum_{i=1}^{K} \gamma_i \ln \frac{p_{n/2}(i)}{p^+(i)}$$
$$+ 2\sum_{i=1}^{K} \gamma_i + \frac{n}{2} \mathrm{KL}(\bar{p}_{n/2+1}(i) \| p^*(i)) + \sum_{i=1}^{K} \ln \left(\frac{n_i + \gamma_i}{n_i(n/2) + \gamma_i} \right),$$

with $\bar{p}_{n/2+1}(i) = \frac{n_i(n/2)}{n/2}$. Complete the proof by concentration inequalities.

Ad Hoc Lower Bound

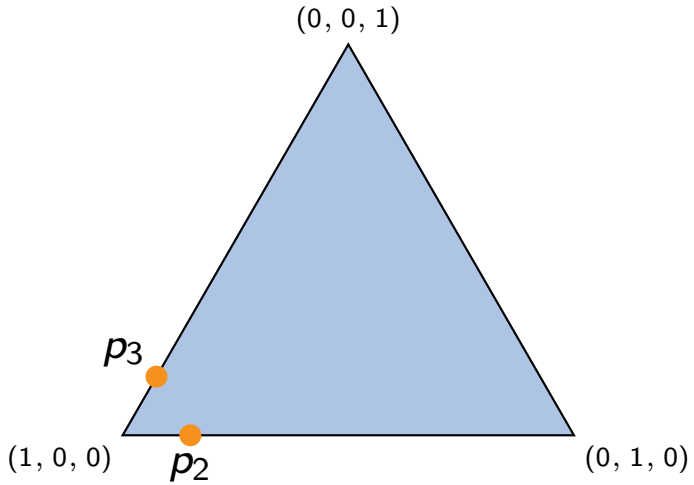


For $j = 2, \ldots, K$:

$$p_j(i) = \begin{cases} 1 - \frac{\alpha}{n} & \text{if } i = 1 \\ \frac{\alpha}{n} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Assume $\alpha := \frac{2}{3} \ln(1/\delta) \leq \frac{n}{2}$.

Ad Hoc Lower Bound



For $j = 2, \ldots, K$:

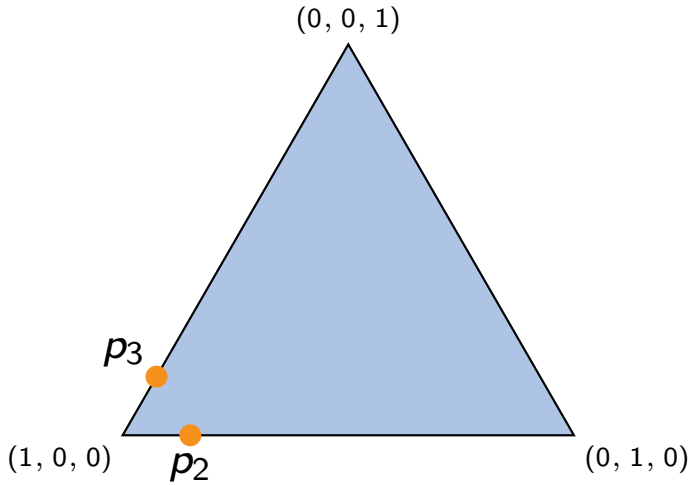
$$p_j(i) = \begin{cases} 1 - \frac{\alpha}{n} & \text{if } i = 1 \\ \frac{\alpha}{n} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Assume $\alpha := \frac{2}{3}\ln(1/\delta) \leq \frac{n}{2}$.

Lower bound:

- α chosen s.t. $\mathbb{P}(n_1 = n) = (1 - \frac{\alpha}{n})^n > \delta$ under all $p_2, \ldots, p_K$
- So, if $n_1 = n$, we cannot distinguish between $p_2, \ldots, p_K$

Ad Hoc Lower Bound



For $j = 2, \ldots, K$:

$$p_j(i) = \begin{cases} 1 - \frac{\alpha}{n} & \text{if } i = 1 \\ \frac{\alpha}{n} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Assume $\alpha := \frac{2}{3}\ln(1/\delta) \leq \frac{n}{2}$.

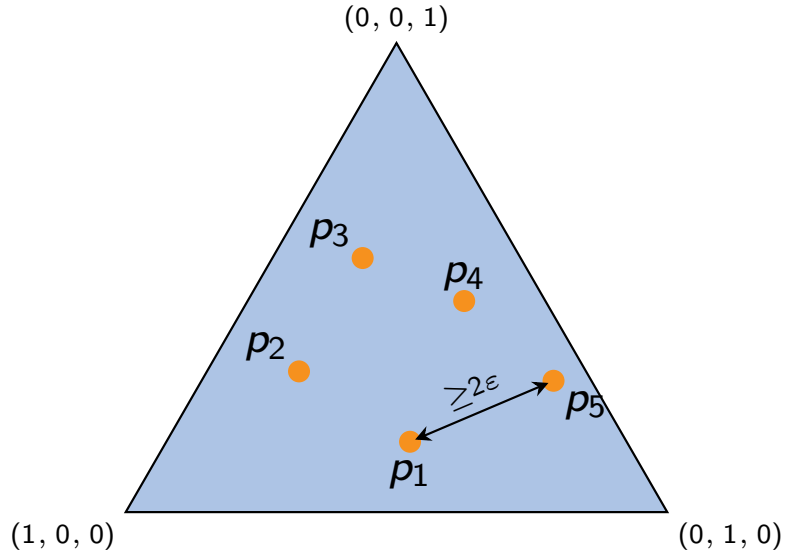
Lower bound:

- α chosen s.t. $\mathbb{P}(n_1 = n) = (1 - \frac{\alpha}{n})^n > \delta$ under all $p_2, \ldots, p_K$
- So, if $n_1 = n$, we cannot distinguish between $p_2, \ldots, p_K$
- Hence, letting $\bar{p} = \frac{1}{K-1}\sum_j p_j$, the minimax rate is at least

$$\min_p \max_j \mathrm{KL}(p_j \| p) \geq \min_p \frac{1}{K-1}\sum_{j=2}^{K} \mathrm{KL}(p_j \| p) = \frac{1}{K-1}\sum_{j=2}^{K} \mathrm{KL}(p_j \| \bar{p})$$

$$= \frac{\alpha}{n}\ln(K-1) = \frac{2\ln(K-1)\ln(1/\delta)}{3n}$$

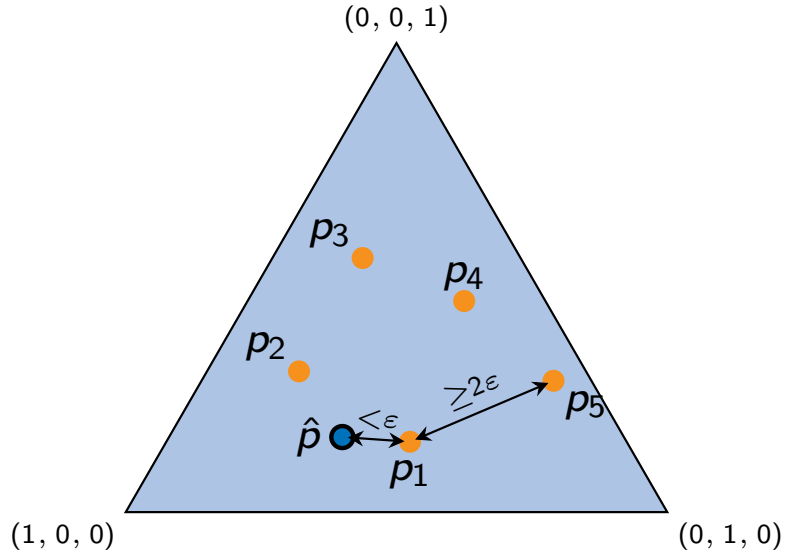
Lower Bound by Reduction to Hypothesis Testing



Reduction to Hypothesis Testing: (Tsybakov, 2009)

- Choose $p_1, \ldots, p_M$ such that $\mathrm{d}(p_j, p_k) \geq 2\epsilon$ for $j \neq k$

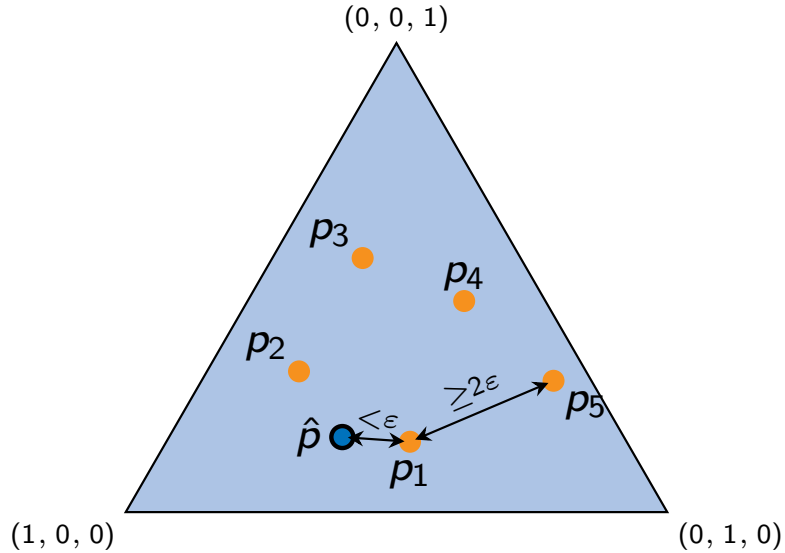
Lower Bound by Reduction to Hypothesis Testing



Reduction to Hypothesis Testing: (Tsybakov, 2009)

- Choose $p_1, \ldots, p_M$ such that $\mathrm{d}(p_j, p_k) \geq 2\epsilon$ for $j \neq k$
- Suppose exists estimator $\hat{p}$ such that $\mathrm{d}(p_j, \hat{p}) < \varepsilon$ with p_j-probability $\geq 1 - \delta$ for all j

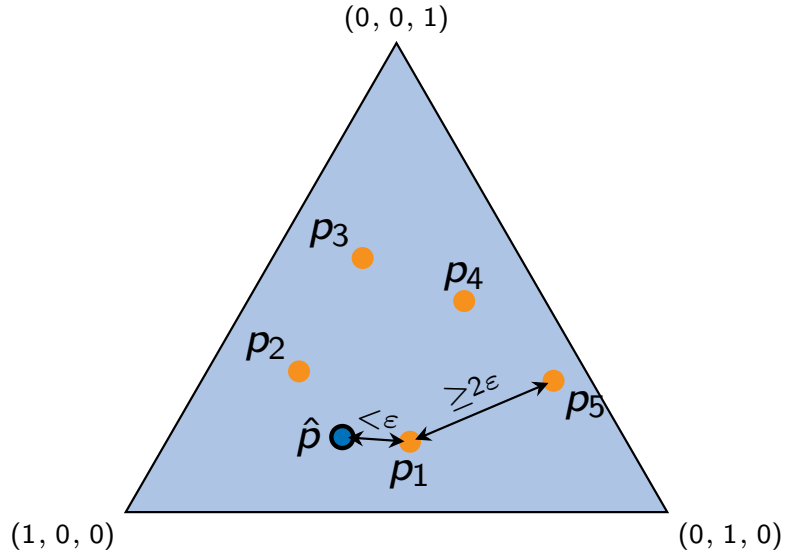
Lower Bound by Reduction to Hypothesis Testing



Reduction to Hypothesis Testing: (Tsybakov, 2009)

- Choose $p_1, \ldots, p_M$ such that $\mathrm{d}(p_j, p_k) \geq 2\epsilon$ for $j \neq k$
- Suppose exists estimator $\hat{p}$ such that $\mathrm{d}(p_j, \hat{p}) < \varepsilon$ with p_j-probability $\geq 1 - \delta$ for all j
- Then can identify p_j with hypothesis test $\hat{j} = \arg\min_j \mathrm{d}(p_j, \hat{p})$

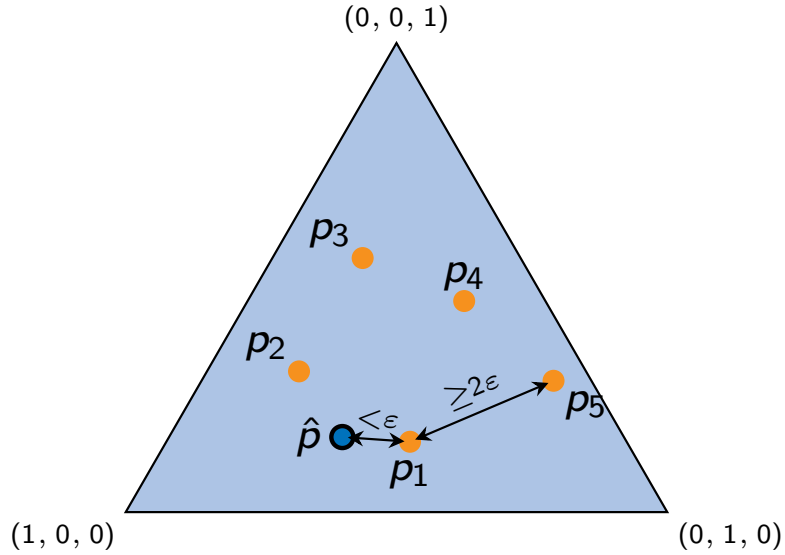
Lower Bound by Reduction to Hypothesis Testing



Reduction to Hypothesis Testing: (Tsybakov, 2009)

- Choose $p_1, \ldots, p_M$ such that $\mathrm{d}(p_j, p_k) \geq 2\epsilon$ for $j \neq k$
- Suppose exists estimator $\hat{p}$ such that $\mathrm{d}(p_j, \hat{p}) < \varepsilon$ with p_j-probability $\geq 1 - \delta$ for all j
- Then can identify p_j with hypothesis test $\hat{j} = \arg\min_j \mathrm{d}(p_j, \hat{p})$
- Show that no such hypothesis test exists for $p_1, \ldots, p_M$, e.g. using Fano's inequality
- Hence no estimator $\hat{p}$ can achieve ε accuracy

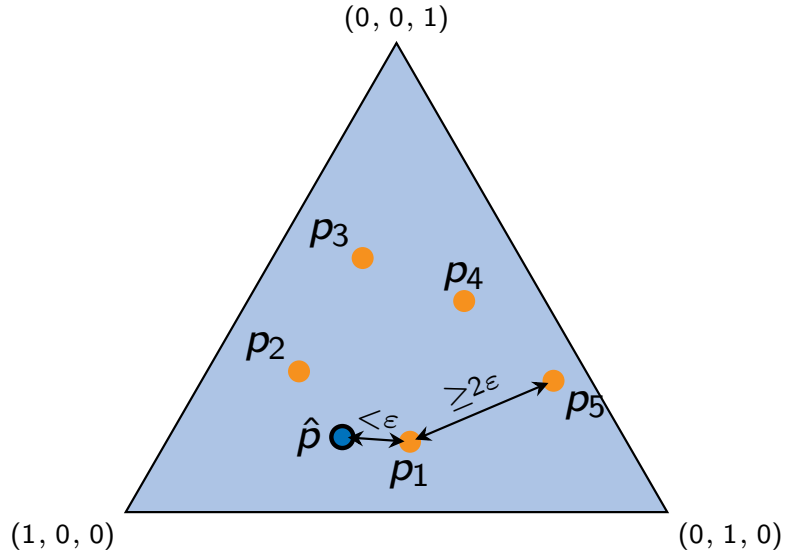
Lower Bound by Reduction to Hypothesis Testing



Reduction to Hypothesis Testing: **(our variant for KL, b.c. not a metric)**

- Choose $p_1, \ldots, p_M$ such that $\mathrm{KL}\left(p_j \| \frac{p_j + p_k}{2}\right) \geq \epsilon$ for $j \neq k$
- Suppose exists estimator $\hat{p}$ such that $\mathrm{KL}(p_j \| \hat{p}) < \varepsilon$ with p_j-probability $\geq 1 - \delta$ for all j
- Then can identify p_j with hypothesis test $\hat{j} = \arg\min_j \mathrm{KL}(p_j \| \hat{p})$
- Show that no such hypothesis test exists for $p_1, \ldots, p_M$, e.g. using Fano's inequality
- Hence no estimator $\hat{p}$ can achieve ε accuracy

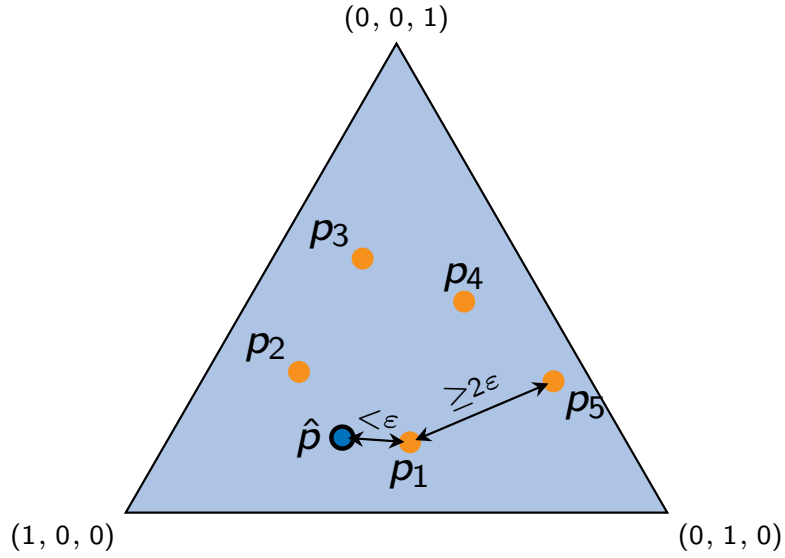
Lower Bound by Reduction to Hypothesis Testing



Reduction to Hypothesis Testing: **(our variant for KL, b.c. not a metric)**

- Choose $p_1, \ldots, p_M$ such that $\mathrm{KL}\left(p_j \| \frac{p_j + p_k}{2}\right) \geq \epsilon$ for $j \neq k$
- Suppose exists estimator $\hat{p}$ such that $\mathrm{KL}(p_j \| \hat{p}) < \varepsilon$ with p_j-probability $\geq 1 - \delta$ for all j
- Then can identify p_j with hypothesis test $\hat{j} = \arg\min_j \mathrm{KL}(p_j \| \hat{p})$
- Show that no such hypothesis test exists for $p_1, \ldots, p_M$, e.g. using Fano's inequality
- Hence no estimator $\hat{p}$ can achieve ε accuracy

Lower Bound by Reduction to Hypothesis Testing



Reduction to Hypothesis Testing: **(our variant for KL, b.c. not a metric)**

- Choose $p_1, \ldots, p_M$ such that $\mathrm{KL}\left(p_j \| \frac{p_j+p_k}{2}\right) \geq \epsilon$ for $j \neq k$

For same $p_2, \ldots, p_K$ as in ad hoc lower bound, first step immediately requires that

$$\varepsilon \leq \mathrm{KL}\left(p_j \| \frac{p_j + p_k}{2}\right) = \frac{\alpha \ln(2)}{n} = \frac{2 \ln(2) \ln(1/\delta)}{3n},$$

which **misses the** $\ln(K)$ **factor**, so standard reduction does not work!

New Reduction to "Weak Hypothesis Testing"

Lemma

Let $P_1, \ldots, P_m$ be any probability distributions for which **weak hypothesis testing** *is not possible with confidence at $1 - \delta$. Then the minimax rate is at least*

$$r_n^*(\delta) \geq \min_P \max_j \mathrm{KL}(P_j \| P).$$

- Same lower bound as in the ad hoc argument
- **No conditions on separation** between distributions!

New Reduction to "Weak Hypothesis Testing"

Lemma

Let $P_1, \ldots, P_m$ be any probability distributions for which **weak hypothesis testing** *is not possible with confidence at $1 - \delta$. Then the minimax rate is at least*

$$r_n^*(\delta) \geq \min_P \max_j \mathrm{KL}(P_j \| P).$$

- Same lower bound as in the ad hoc argument
- **No conditions on separation** between distributions!

Weak hypothesis testing:

- Normal test correct if identifies true hypothesis
- Weak test correct if identifies **any false hypothesis**

New Reduction to "Weak Hypothesis Testing"

Lemma

Let $P_1, \ldots, P_m$ be any probability distributions for which **weak hypothesis testing** *is not possible with confidence at $1 - \delta$. Then the minimax rate is at least*

$$r_n^*(\delta) \geq \min_P \max_j \mathrm{KL}(P_j \| P).$$

- Same lower bound as in the ad hoc argument
- **No conditions on separation** between distributions!

Weak hypothesis testing:

- Normal test correct if identifies true hypothesis
- Weak test correct if identifies **any false hypothesis**
- **Still not possible** with confidence $1 - \delta$ for $p_2, \ldots, p_K$ from ad hoc bound, because all distributions give probability $> \delta$ to event $n_1 = n$.

New Reduction to "Weak Hypothesis Testing"

Lemma

Let $P_1, \ldots, P_m$ be any probability distributions for which **weak hypothesis testing** *is not possible with confidence at $1 - \delta$. Then the minimax rate is at least*

$$r_n^*(\delta) \geq \min_P \max_j \mathrm{KL}(P_j \| P).$$

- Same lower bound as in the ad hoc argument
- **No conditions on separation** between distributions!

Proof sketch:

- Define weak test $\hat{j} = \arg\max_j \mathrm{KL}(P_j \| \hat{P})$ (furthest away)
- Then weak test incorrect ($\hat{j} = j^*$) implies $\mathrm{KL}(P_{j^*} \| \hat{P}) = \max_j \mathrm{KL}(P_j \| \hat{P}) \geq \min_P \max_j \mathrm{KL}(P_j \| P)$

References I

- D. van der Hoeven, J. Olkhovskaia, T. van Erven. **Nearly Minimax Discrete Distribution Estimation in Kullback-Leibler Divergence with High Probability**, ArXiv:2507.17316 preprint, 2025.

Other references:

Braess, Dietrich & Thomas Sauer (2004). "Bernstein polynomials and learning theory". In: *Journal of Approximation Theory* 128.2, pp. 187–206.

Canonne, Clément L. (2020). "A short note on learning discrete distributions". In: *arXiv preprint arXiv:2002.11457*.

Canonne, Clément L., Ziteng Sun, & Ananda Theertha Suresh (2023). "Concentration Bounds for Discrete Distribution Estimation in KL Divergence". In: *arXiv preprint arXiv:2302.06869*.

Catoni, Olivier (1997). "The mixture approach to universal model selection". In: *Preprints of École Normale Supérieure*.

Forster, Jürgen & Manfred K Warmuth (2002). "Relative expected instantaneous loss bounds". In: *Journal of Computer and System Sciences* 64.1, pp. 76–102.

Mourtada, Jaouad (2025). "Estimation of discrete distributions in relative entropy, and the deviations of the missing mass". In: *arXiv preprint arXiv:2504.21787*.

References II

Polyanskiy, Yury & Yihong Wu (2025). *Information Theory: From Coding to Learning*. Cambridge University Press.

Tsybakov, Alexandre B. (2009). *Introduction to Nonparametric Estimation*. Springer.

Van der Hoeven, Dirk, Nikita Zhivotovskiy, & Nicolò Cesa-Bianchi (2023). "High-Probability Risk Bounds via Sequential Predictors". In: *arXiv preprint arXiv:2308.07588*.

Van Erven, Tim & Peter Harremoës (2014). "Rényi Divergence and Kullback-Leibler Divergence". In: *IEEE Transactions on Information Theory* 60.7, pp. 3797–3820.

The Hierarchy of Divergence Measures

Rényi divergence $\mathrm{R}_\alpha(p\|q)$ is increasing in its order α,
and continuous in α (except if it jumps to ∞ at some $\alpha > 1$)

(van Erven & Harremoës, 2014)

The Hierarchy of Divergence Measures

Rényi divergence $\mathrm{R}_\alpha(p\|q)$ is increasing in its order α,
and continuous in α (except if it jumps to ∞ at some $\alpha > 1$)
(van Erven & Harremoës, 2014)

- V = Variational distance
- H = Hellinger distance
- KL = Kullback-Leibler divergence
- χ^2 = χ^2 divergence

The Hierarchy of Divergence Measures

Rényi divergence $\mathrm{R}_\alpha(p\|q)$ is increasing in its order α,
and continuous in α (except if it jumps to ∞ at some $\alpha > 1$)
(van Erven & Harremoës, 2014)

- V = Variational distance
- H = Hellinger distance
- KL = Kullback-Leibler divergence
- χ^2 = χ^2 divergence

The Hierarchy of Divergence Measures

Rényi divergence $\mathrm{R}_\alpha(p\|q)$ is increasing in its order α,
and continuous in α (except if it jumps to ∞ at some $\alpha > 1$)

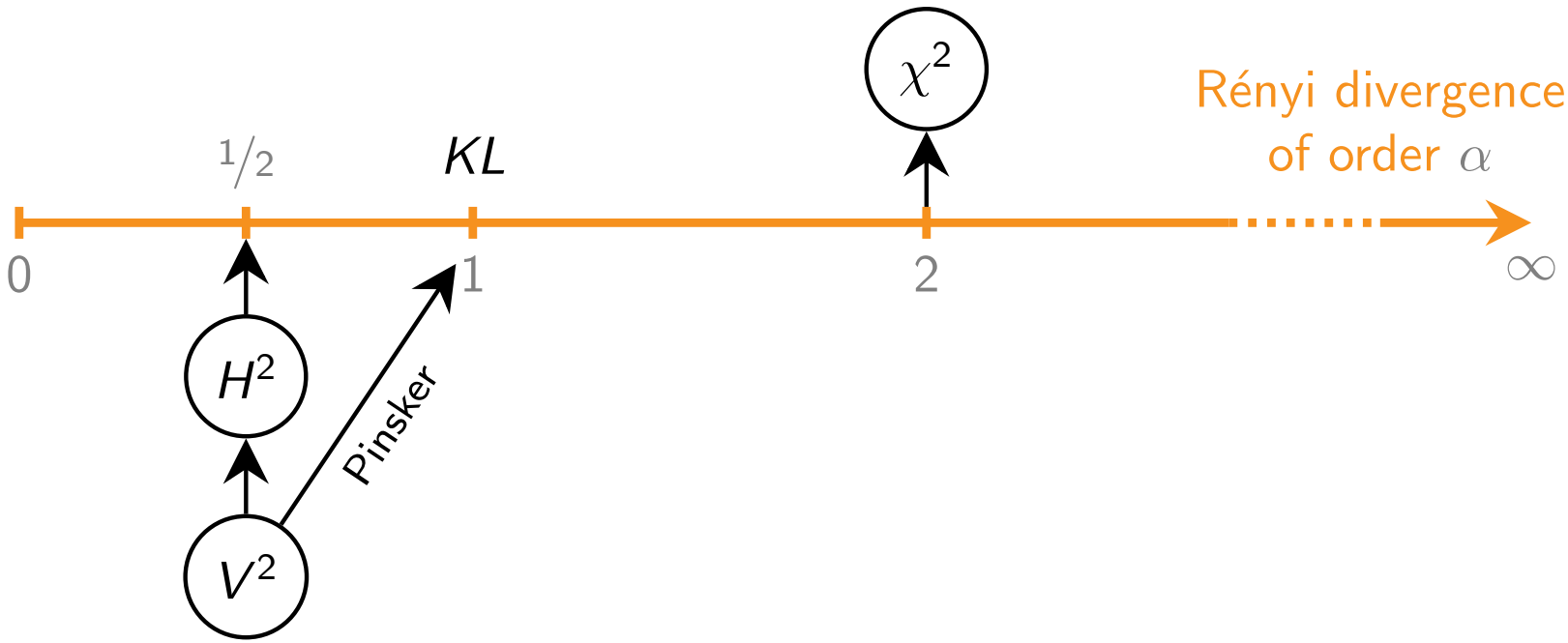
(van Erven & Harremoës, 2014)

- V = Variational distance
- H = Hellinger distance
- KL = Kullback-Leibler divergence
- $\chi^2 = \chi^2$ divergence

The Hierarchy of Divergence Measures

Rényi divergence $\mathrm{R}_\alpha(p\|q)$ is increasing in its order α,
and continuous in α (except if it jumps to ∞ at some $\alpha > 1$)

(van Erven & Harremoës, 2014)

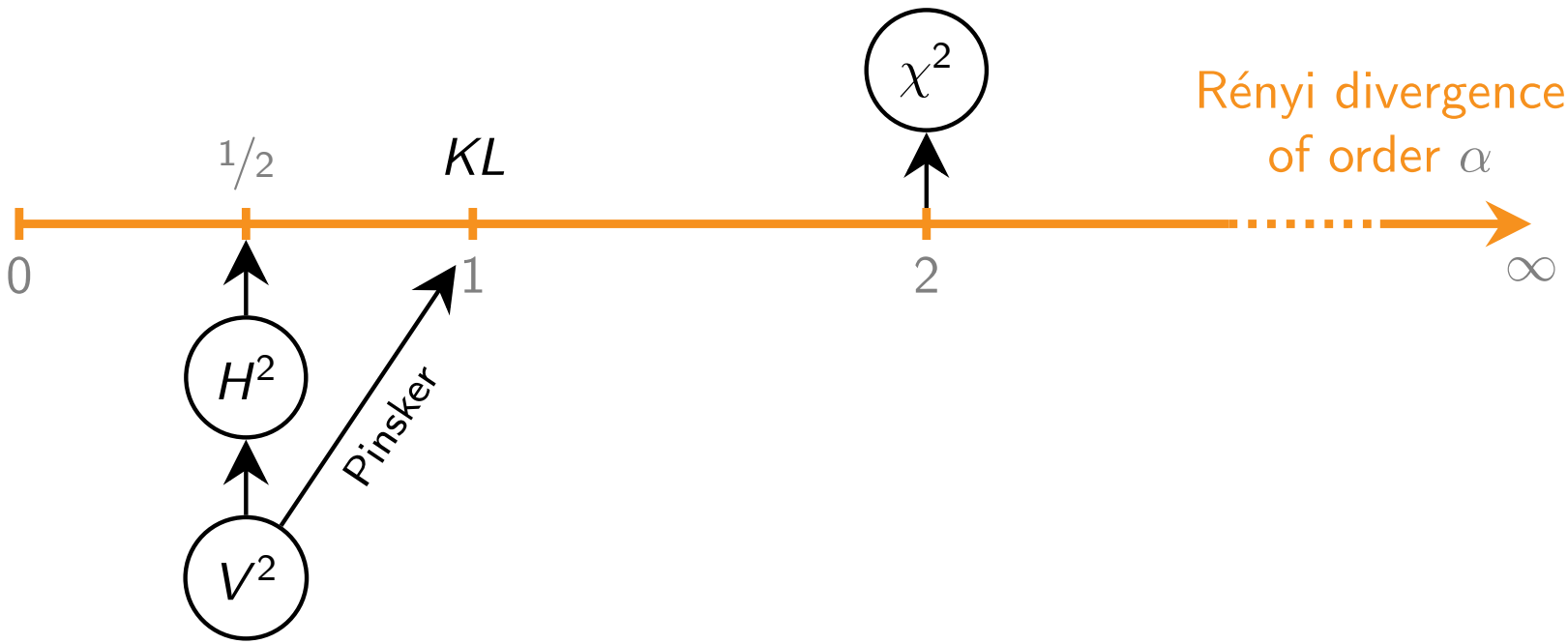


- V = Variational distance
- H = Hellinger distance
- KL = Kullback-Leibler divergence
- $\chi^2 = \chi^2$ divergence

The Hierarchy of Divergence Measures

Rényi divergence $\mathrm{R}_\alpha(p\|q)$ is increasing in its order α,
and continuous in α (except if it jumps to ∞ at some $\alpha > 1$)

(van Erven & Harremoës, 2014)

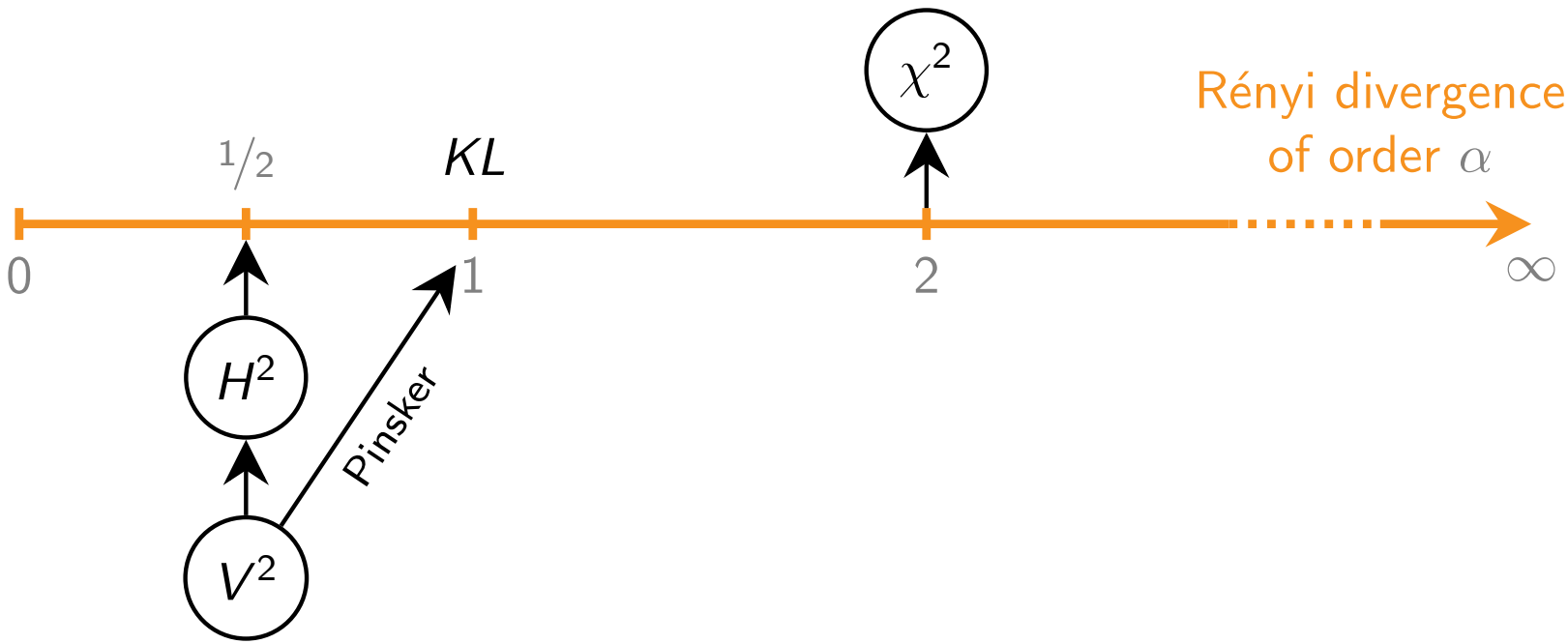


Remarks:

- $H^2 \approx R_{1/2}$ when either goes to 0
- $\chi^2 \approx R_2$ when either goes to 0
- $\frac{1}{4}V^2 \leq H^2 \leq R_{1/2} \leq KL$, but Pinsker gives tighter $\frac{1}{2}V^2 \leq KL$

The Hierarchy of Divergence Measures

Rényi divergence $\mathrm{R}_\alpha(p\|q)$ is increasing in its order α,
and continuous in α (except if it jumps to ∞ at some $\alpha > 1$)
(van Erven & Harremoës, 2014)



Used in this talk:

$$V^2 \lesssim H^2 \leq KL \leq \chi^2$$