Information measures and geometry of the Zeta distributions and related distributions

Frank Nielsen

Sony Computer Science Laboratories Inc.

Tokyo, Japan

https://franknielsen.github.io/

Sony CSL

[arXiv:2104.10548]

Zipf's distributions

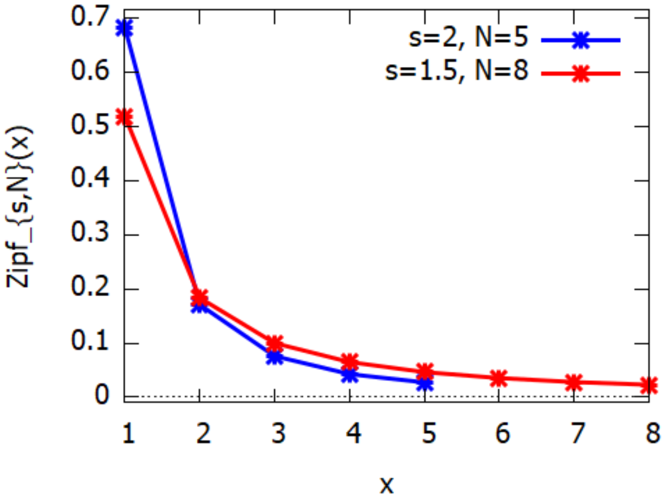
- A **discrete power law** with probability mass function

$$p_{s,N}(x) \propto \frac{1}{x^s}, \quad x \in \mathcal{X} = \{1, \ldots, N\}, s > 0$$

George K. Zipf
1902-1950
American linguist

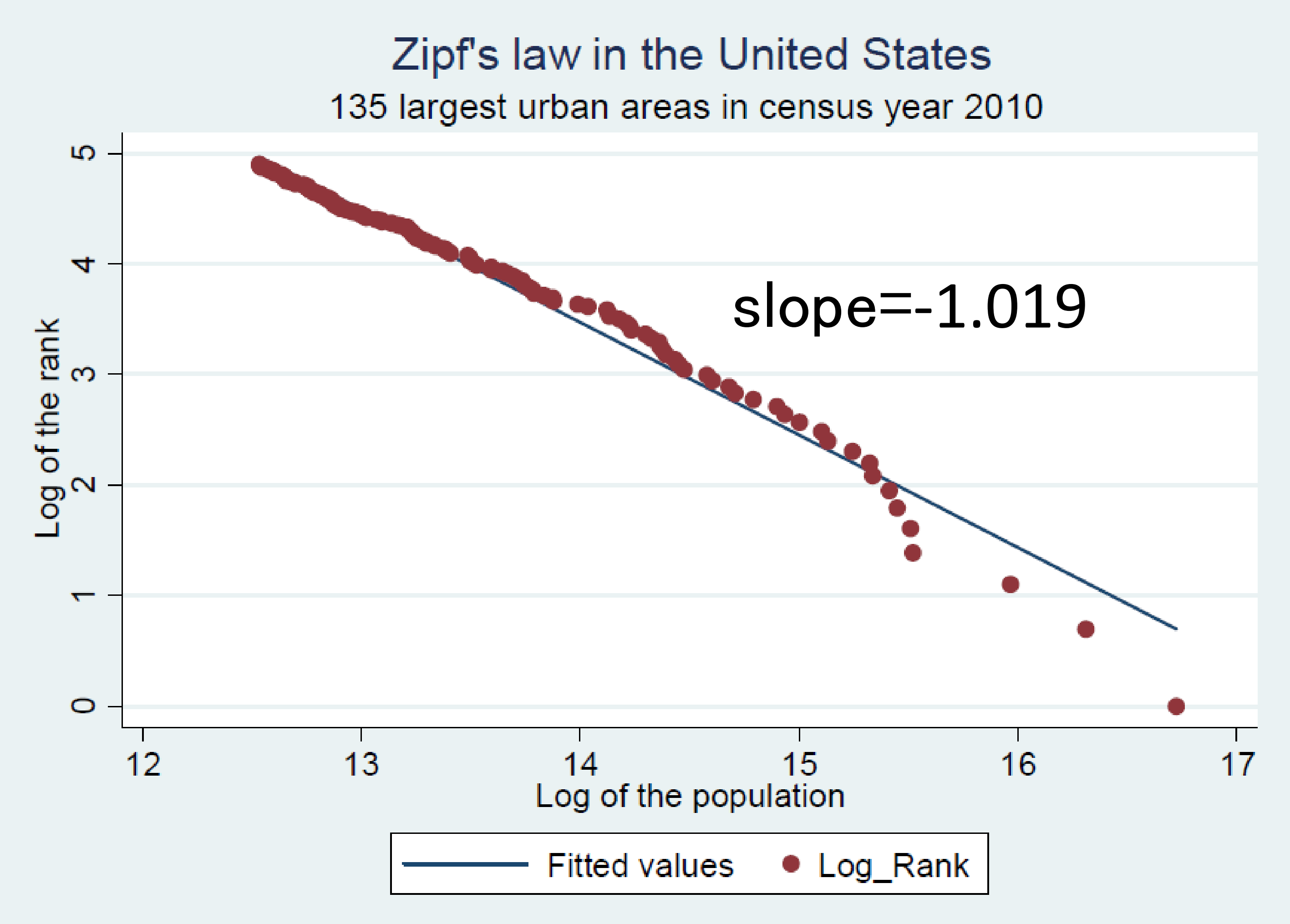
- Normalizers are **generalized harmonic numbers** :

$$H_{s,N} = \sum_{x \in \{1,\ldots,N\}} \frac{1}{x^s} \qquad p_{s,N}(x) = \frac{1}{x^s} \frac{1}{H_{s,N}}$$



- Examples: Empirical distributions of word frequencies sorted by their rankings in natural languages, population sizes of cities, etc.

Zipf's law: Human populations in cities



log–rank versus log–size plot

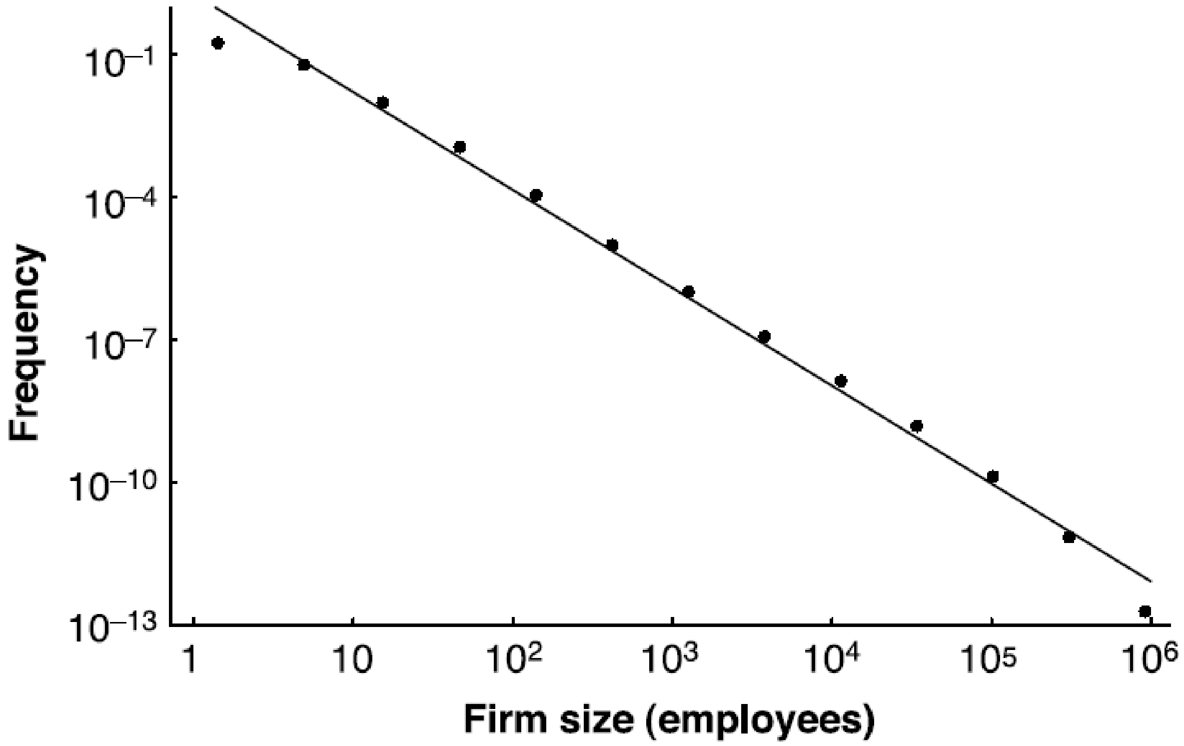
$$p_{s,N}(x) = \frac{1}{x^s}\frac{1}{H_{s,N}}$$

in log–log expression: linear

$$\log p_{s,N}(x) = -\log H_{s,N} - s\log x$$

Arshad, Sidra, Shougeng Hu, and Badar Nadeem Ashraf. "Zipf's law and city size distribution: A survey of the literature and future research agenda." Physica A: Statistical mechanics and its applications 492 (2018): 75-92.

Zipf's law: US firm sizes, surname frequencies, etc.



Reports

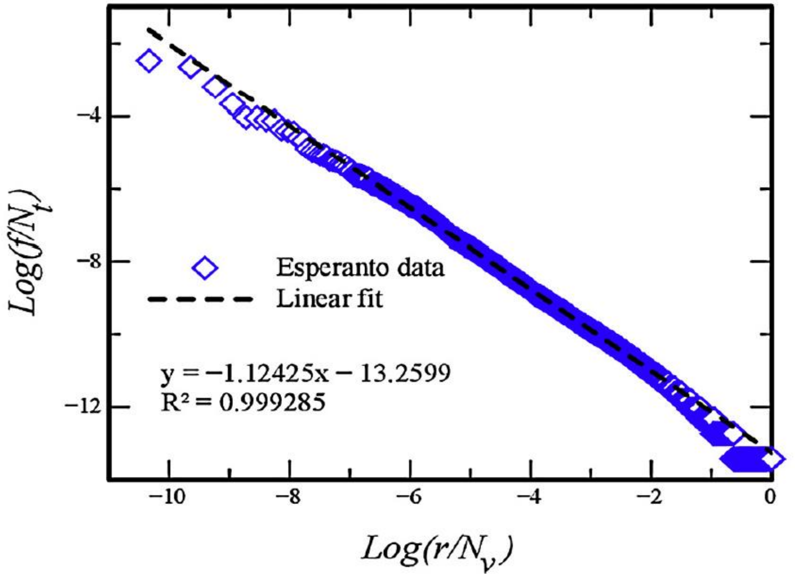
Fig. 1. Histogram of U.S. firm sizes, by employees. Data are for 1997 from the U.S. Census Bureau, tabulated in bins having width increasing in powers of three (*30*). The solid line is the OLS regression line through the data, and it has a slope of 2.059 (SE = 0.054; adjusted R^2 = 0.992), meaning that α = 1.059; maximum likelihood and nonparametric methods yield similar results. The data are slightly concave to the origin in log-log coordinates, reflecting finite size cutoffs at the limits of very small and very large firms.

Remark: Estimators by **line fitting methods** in log-log coordinates...

Axtell, Robert L. "Zipf distribution of US firm sizes." *Science* 293.5536 (2001): 1818-1820.
Fox, Wendy R., and Gabriel W. Lasker. "The distribution of surname frequencies." International Statistical Review (1983): 81-87.

Zipf's distributions: A rank-frequency distribution

	Language	(l, h)	ζ	N_t	N_v		Language	(l, h)	ζ	N_t	N_v
1	Achuar	$(-7,-1)$	1.124	174103	19140	51	K'iche'	$(-7,-2)$	0.985	485530	82102
2	Afrikaans	$(-6,-2)$	1.191	797099	16599	52	Korean	$(-5,-3)$	1.442	1273350	2753
3	Aguaruna	$(-6,-2)$	1.055	148528	24809	53	Latin	$(-6,-2)$	1.065	535036	46673
4	Akawaio	$(-6,-2)$	1.248	189996	8614	54	Latvian	$(-6,-2)$	1.002	132005	15532
5	Albanian	$(-8,-2)$	1.091	749487	32555	55	Lithuanian	$(-7,-2)$	1.015	468280	47277
6	Amharic	$(-6,-2)$	0.952	99858	26315	56	Lukpa	$(-5,-1)$	1.417	218489	7091
7	Amuzgo	$(-6,-2)$	1.175	199786	13948	57	Malagasy	$(-7,-2)$	1.161	712101	29933
8	Arabic	$(-6,-2)$	1.030	435126	51530	58	Malayalam	$(-8,-2)$	0.971	403012	79708
9	Armenian	$(-6,-2)$	1.010	114425	13945	59	Mam	$(-5,-1)$	1.409	216854	11036
10	Aukan	$(-5,-2)$	1.289	306705	1937	60	Manx	$(-5,-1)$	1.246	93290	4444
11	Barasana	$(-6,-2)$	1.188	224639	17511	61	Maori	$(-6,-2)$	1.430	978523	9212
12	Basque	$(-6,-2)$	0.978	132258	17116	62	Marathi	$(-6,-2)$	1.152	636857	48513
13	Bulgarian	$(-6,-2)$	1.122	645279	29922	63	Myanmar	$(-8,-3)$	0.810	298870	136337
14	Cabécar	$(-6,-2)$	1.277	199200	8250	64	Nahuatl	$(-6,-2)$	1.166	175709	15517
15	Cakchiquel	$(-6,-2)$	1.297	314530	8354	65	Nepali	$(-6,-2)$	1.189	666249	48711
16	Campa	$(-7,-3)$	1.001	122296	25869	66	Norwegian	$(-6,-2)$	1.225	720146	19162
17	Camsá	$(-6,-2)$	1.132	193216	19886	67	Ojibwa	$(-6,-3)$	1.104	142132	35739
18	Cebuano	$(-6,-2)$	1.233	864613	27473	68	Paite	$(-8,-4)$	1.031	763335	30659
19	Chamorro	$(-6,-2)$	1.106	133171	11191	69	Polish	$(-6,-2)$	1.105	604716	41953
20	Cherokee	$(-7,-2)$	1.005	116610	24103	70	Portuguese	$(-6,-2)$	1.196	698662	29257
21	Chinantec	$(-6,-2)$	1.290	304955	11731	71	Potawatomi	$(-7,-2)$	0.994	31090	10487
22	Chinese	$(-5,-3)$	0.792	1031010	1699	72	Q'eqchi'	$(-6,-2)$	1.266	917383	22123
23	Coptic	$(-8,-2)$	0.969	122956	22050	73	Quichua	$(-6,-2)$	1.022	116854	15164
24	Croatian	$(-7,-2)$	1.065	574549	49727	74	Romani	$(-4,-1)$	1.411	183270	8097
25	Czech	$(-7,-2)$	1.112	597212	41163	75	Romanian	$(-6,-2)$	1.137	703832	21632
26	Danish	$(-6,-2)$	1.158	654968	26290	76	Russian	$(-6,-2)$	1.116	561097	43349
27	Dinka	$(-6,-2)$	1.196	193493	5809	77	Serbian	$(-6,-2)$	1.138	585460	34345
28	English	$(-6,-2)$	1.258	789631	12702	78	Shuar	$(-7,-2)$	1.025	132294	24527
29	Esperanto	$(-8,-2)$	1.124	678105	30723	79	Slovak	$(-6,-2)$	1.135	620851	42087
30	Estonian	$(-6,-2)$	1.070	153549	16381	80	Slovene	$(-7,-2)$	1.112	647280	40130
31	Ewe	$(-6,-2)$	1.189	220394	10481	81	Somali	$(-6,-2)$	1.145	727664	38844
32	Farsi	$(-10,-4)$	0.981	671353	33252	82	Spanish	$(-7,-2)$	1.158	715950	28146
33	Finnish	$(-8,-2)$	0.997	542551	54863	83	Swahili	$(-5,-2)$	1.204	139028	17721
34	French	$(-7,-2)$	1.161	723763	24716	84	Swedish	$(-6,-2)$	1.208	732789	23764
35	Gaelic	$(-4,-1)$	1.153	15505	2173	85	Syriac	$(-6,-2)$	1.010	109720	13013
36	Galela	$(-6,-2)$	1.252	285292	12010	86	Tachelhit	$(-6,-2)$	0.969	104791	23572
37	German	$(-7,-2)$	1.191	696960	21120	87	Tagalog	$(-6,-2)$	1.248	825302	24257
38	Greek	$(-6,-2)$	1.130	706111	32588	88	Tamajaq	$(-7,-2)$	1.031	54462	9399
39	Gujarati	$(-5,-1)$	1.271	197942	13790	89	Telugu	$(-9,-2)$	0.944	442228	95094
40	Haitian	$(-6,-2)$	1.426	919031	8082	90	Thai	$(-9,-5)$	0.765	111534	97640
41	Hebrew	$(-8,-3)$	0.938	415150	42964	91	Turkish	$(-7,-2)$	0.974	449906	58944
42	Hindi	$(-8,-3)$	1.081	790569	18790	92	Ukranian	$(-8,-2)$	1.006	132251	15883
43	Hungarian	$(-8,-2)$	0.996	598036	64361	93	Uma	$(-6,-2)$	1.157	192555	11998
44	Icelandic	$(-7,-2)$	1.138	666402	35404	94	Uspanteco	$(-6,-2)$	1.268	225698	8646
45	Indonesian	$(-6,-2)$	1.143	630508	19775	95	Vietnamese	$(-5,-2)$	1.232	837080	7061
46	Italian	$(-6,-2)$	1.160	668894	31728	96	Wolaytta	$(-6,-2)$	1.015	126840	16924
47	Jakalteko	$(-7,-1)$	1.297	218570	12209	97	Wolof	$(-5,-1)$	1.399	171471	7234
48	Japanese	$(-10,-7)$	0.974	31089	30785	98	Xhosa	$(-7,-2)$	1.043	443036	82199
49	Kabyle	$(-5,-2)$	1.414	271538	7241	99	Zarma	$(-6,-2)$	1.311	810724	9865
50	Kannada	$(-7,-2)$	1.078	154789	16764	100	Zulu	$(-6,-2)$	0.961	96555	25303



In general, languages have a power law range

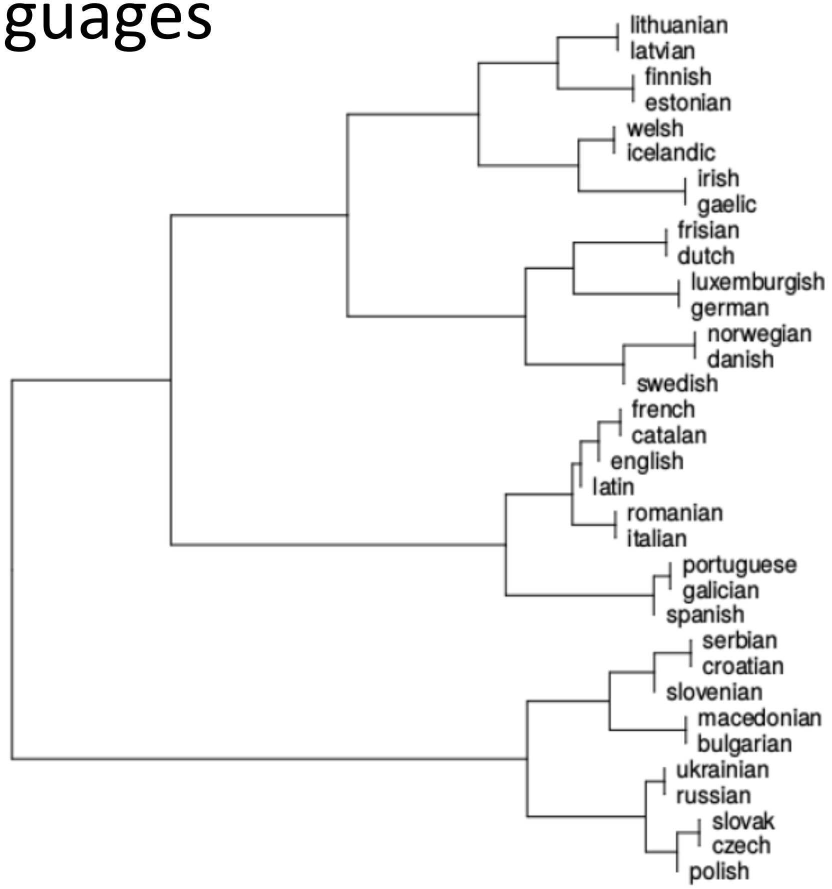
Natural language	θ	N
English	1.258	12702
French	1.161	24716
Japanese	0.774	30785
Danish	1.158	26290
Chinese	0.792	1699
Finnish	0.997	54863
...	...	...

- Ali Mehri and Maryam Jamaati. Variation of Zipf's exponent in one hundred live languages: A study of the Holy Bible translations. Physics Letters A, 381(31):2470{2477, 2017.

- Ferrer i Cancho, Ramon. "The variation of Zipf's law in human language." The European Physical Journal B-Condensed Matter and Complex Systems 44.2 (2005): 249-257.

Clustering Zipf's distributions

- Agglomerative hierarchical clustering yields a **dendrogram** for finding similarities between human languages



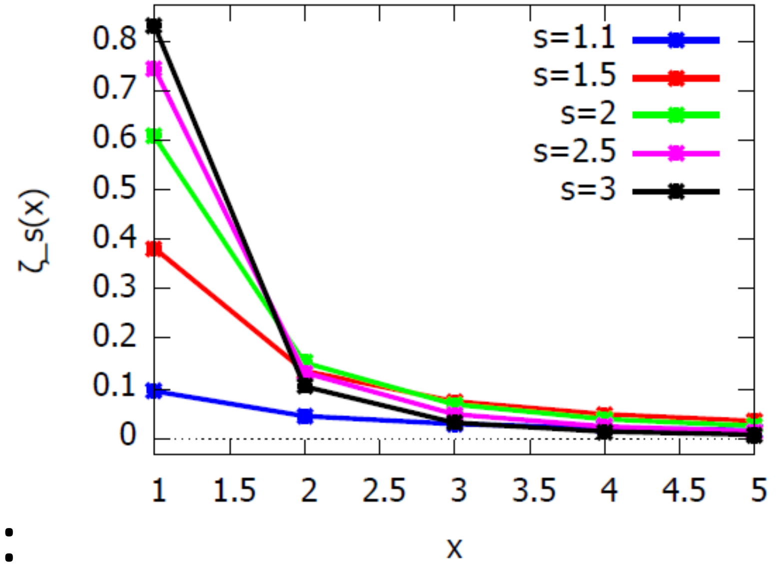
Delta-Eder distance
Kullback-Leibler divergence

Pablo Gamallo, Jose Ramom Pichel, and Inaki Alegria. Measuring language distance of isolated european languages. Information, 11(4):181, 2020.

Zeta distributions

- Discrete power law distributions on **natural numbers**

$$p_s(x) \propto \frac{1}{x^s}, \qquad x \in \mathcal{X} = \mathbb{N} = \{1, 2, \ldots\}$$



- Normalizing function is the **real Riemann zeta function**:

  ⇨ many fast approximation algorithms to numerically calculate zeta

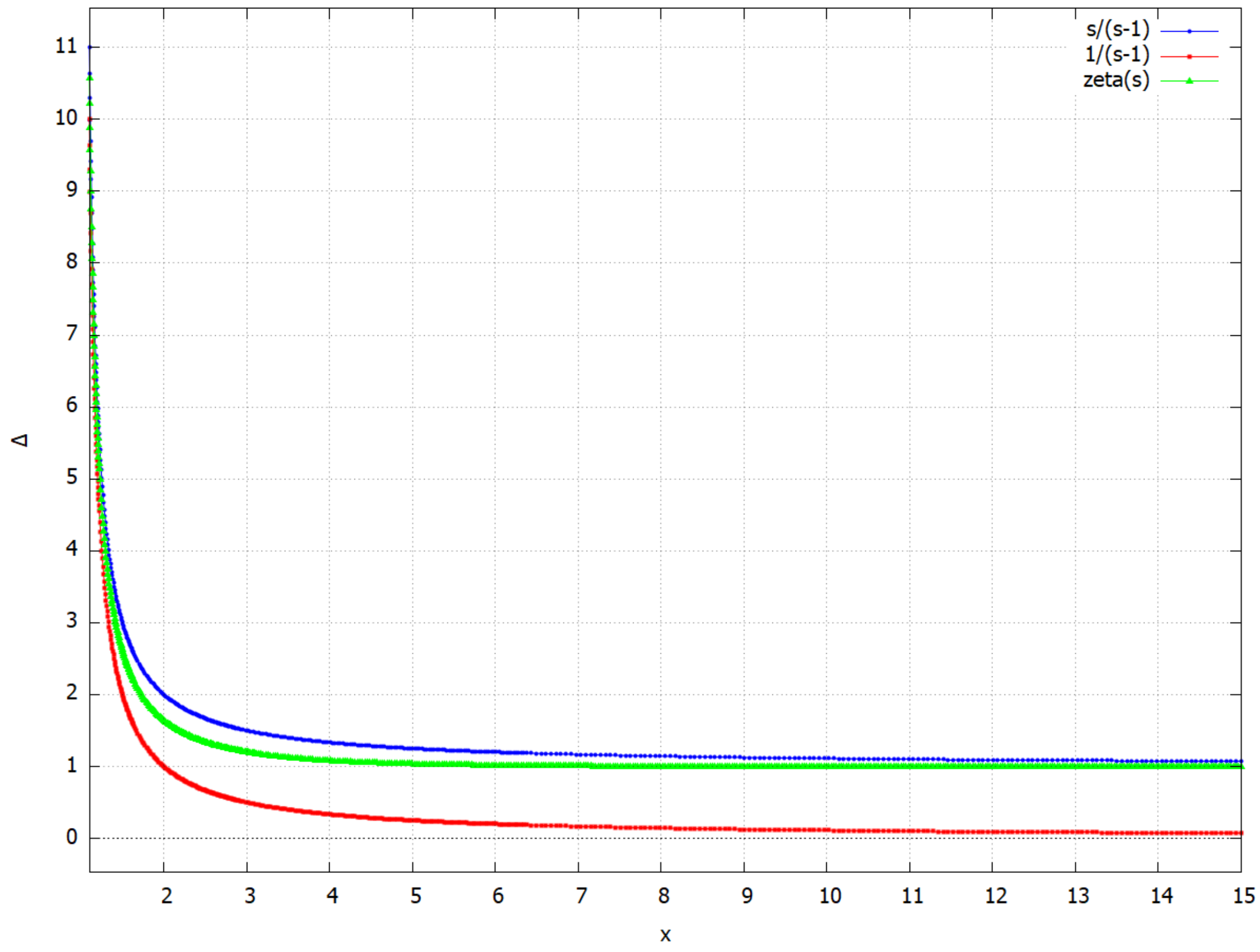
$$\zeta(s) = \sum_{i=1}^{\infty} \frac{1}{i^s}, \qquad s > 1$$

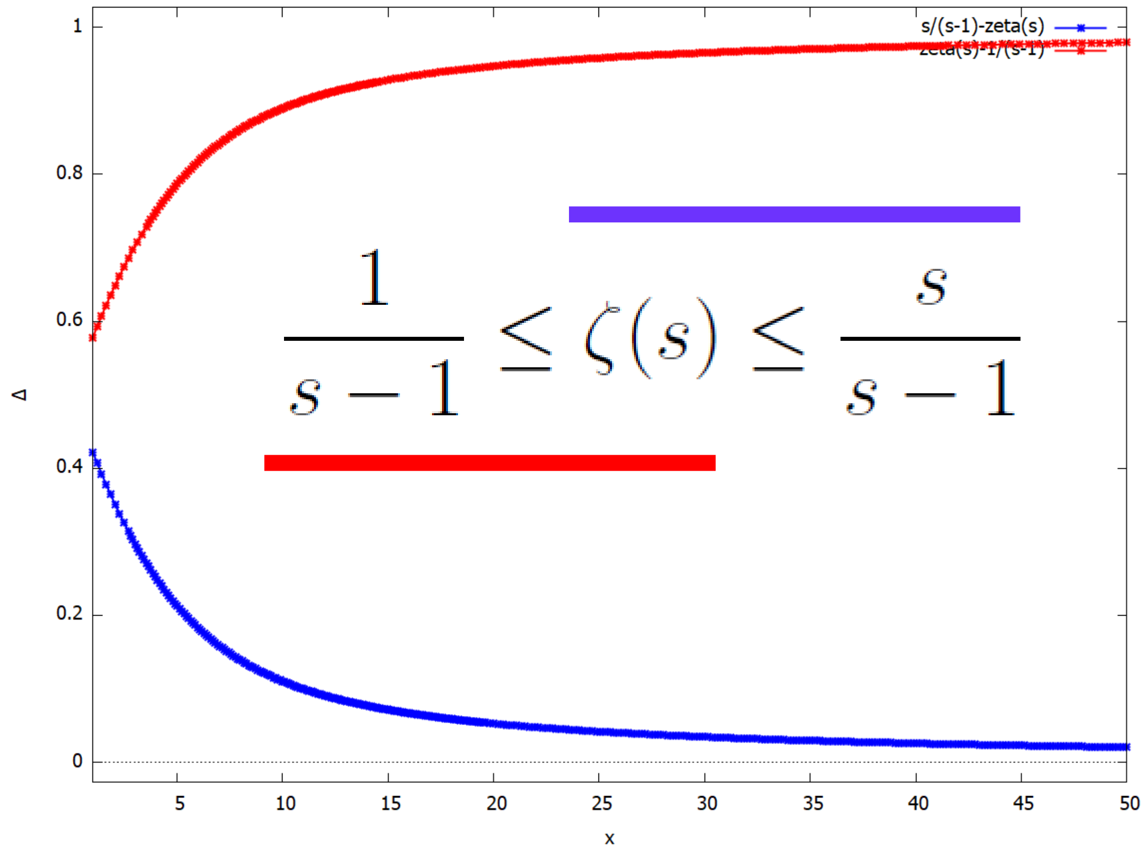
Probability mass function of a zeta distribution:

$$p_s(x) = \frac{1}{\zeta(s)\, x^s}$$

- Because of infinite summation, we need **s>1** for convergence. This contrasts with Zipf's distributions are defined for s>0.

Bounding the real zeta function ζ(s)





$$\frac{1}{s-1} \leq \zeta(s) \leq \frac{s}{s-1}$$

Interesting range in practice is in (1,small s) not asymptotic s→∞

Ferrer and Fernández. "Power laws and the golden number." Problems of general, germanic and slavic linguistics, 518-523.

Zeta distributions: A discrete exponential family

- **Canonical decomposition** of exponential families:

$$p_s(x) = \frac{1}{\zeta(s)\, x^s} \qquad\qquad p_s(x) = \exp\left(\theta(s)t(x) - F(\theta(s))\right)$$

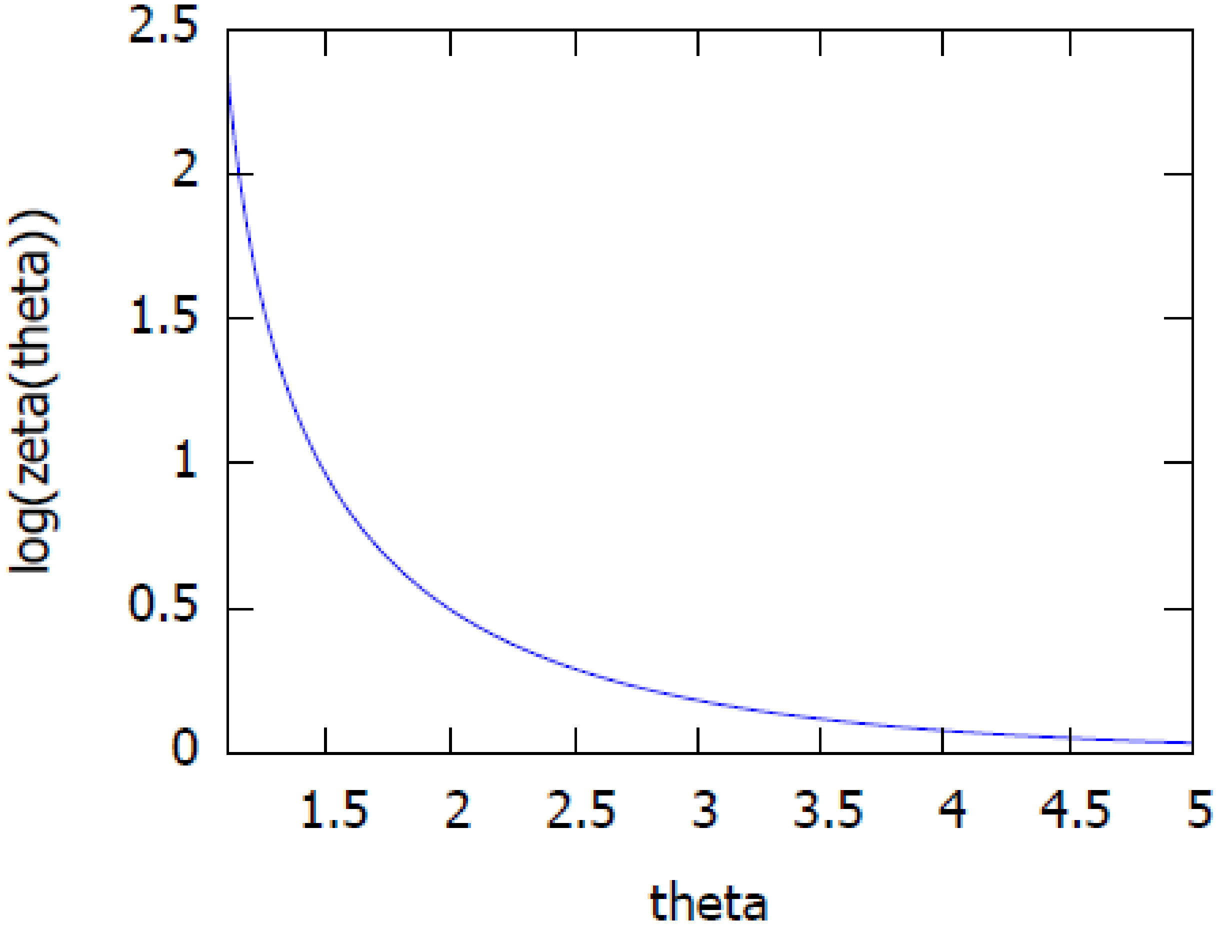
- **Natural parameter** and **sufficient statistic**:

$$\theta = s \qquad\qquad t(x) = -\log x$$

- **Log-normalizer** (also called cumulant function): $F(\theta) = \log \zeta(\theta)$
- **Fisher information**:

$$I(s) = -E_{p_s}[(\log p_s(x))''] = F''(s) = \underline{(\log \zeta(s))''} = \frac{\zeta(s)\zeta''(s) - \zeta'(s)^2}{\zeta^2(s)}.$$

Log-normalizer F(θ) of Zeta distributions



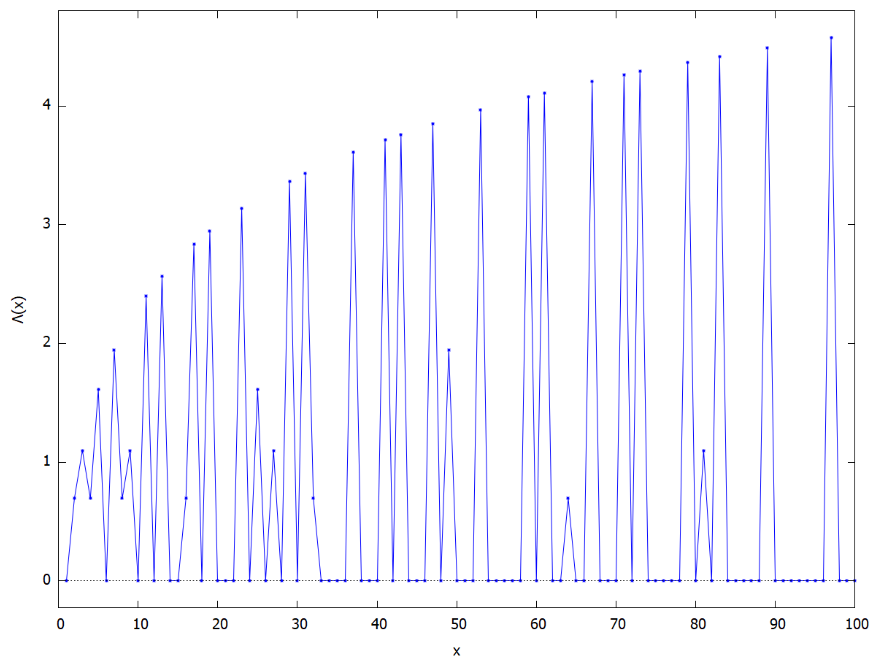
A strictly convex and real-analytic C^{∞} smooth function

Fisher information and von Mangoldt function

$$I(s) = -E_{p_s}[(\log p_s(x))''] = F''(s) = (\log \zeta(s))'' = \frac{\zeta(s)\zeta''(s) - \zeta'(s)^2}{\zeta^2(s)}.$$

$$I(s) = \sum_{i=1}^{\infty} \Lambda(i) \log(i) i^{-s}$$

$$\Lambda(i) = \begin{cases} \log p & \text{if } i = p^k \text{ for some prime } p \text{ and integer } k \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$



von Mangoldt function satisfies the identity:

$$\log(n) = \sum_{i|n} \Lambda(i)$$

Zeta distribution: A maximum entropy distribution

- Exponential families are distributions which **maximize entropy** subject to the moment constraints of their sufficient statistics.

$$\max_p \{H(p) \quad : \quad -E_p[\log x] = \eta \Rightarrow p = \mathrm{Zeta}(\theta)\}$$

- Shannon's entropy: $$H(p) = -\sum_{x\in\mathbb{N}} p(x)\log p(x)$$

- **Dual moment parameterization** of zeta distributions: $$\eta = F'(\theta) = \frac{\zeta'(s)}{\zeta(s)}$$

  Efficient numerical evaluation algorithms of F' but difficult to inverse!

Guiasu, Silviu. "An optimization problem related to the zeta-function." *Canadian Mathematical Bulletin* 29.1 (1986): 70-73.

Zeta distribution: Maximum likelihood estimator

- Given n identically and independently variates, estimate the Zeta moment parameter eta by **maximum likelihood**:

$$\max_s \frac{1}{n}\sum_{i=1}^{n} \log p_s(x_i) = \max_s -\frac{1}{n}\sum_{i=1}^{n} \log x_i - \log \zeta(s)$$

$$\hat{\eta} = \frac{\zeta'(\hat{\theta})}{\zeta(\hat{\theta})} = -\frac{1}{n}\sum_{i=1}^{n} \log x_i$$

- Better **quadratic distance estimator** which is also consistent:

QDE amounts to a line fitting procedure

Louis G Doray and Andrew Luong. Quadratic distance estimators for the zeta family. Insurance: Mathematics and Economics, 16(3):255{260, 1995.

Zeta distribution: Cramér–Rao lower bound

- The variance of any unbiased estimator is bounded by its inverse Fisher information (CRLB):

$$X = (X_1, \ldots, X_n) \sim_{\text{iid}} \text{Zeta}(s)$$

- Fisher information of iid random vector is additive: $I_X = nI(s)$

$$I(s) = -E_{p_s}[(\log p_s(x))''] = (\log \zeta(s))'' = \frac{\zeta(s)\zeta''(s) - \zeta'(s)^2}{\zeta^2(s)}$$

$$\text{Var}[\hat{s}] \geq \frac{1}{n} I^{-1}(s) = \frac{\zeta^2(s)}{n(\zeta(s)\zeta''(s) - \zeta'(s)^2)}.$$

Variance of unbiased estimator **matches** the CRLB only for exponential families

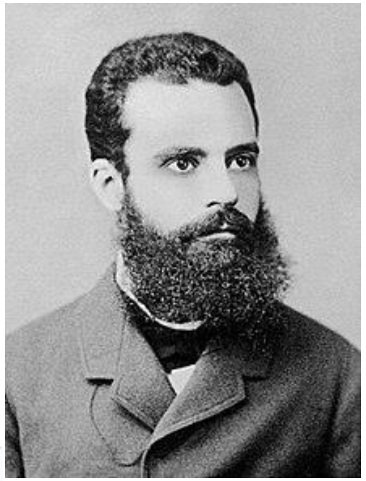
in accordance with Section 2 of **Louis G Doray and Andrew Luong. Quadratic distance estimators for the zeta family. Insurance: Mathematics and Economics, 16(3):255{260, 1995.**

Pareto distributions: Continuous exponential family

- Power law distribution with **probability density function**:

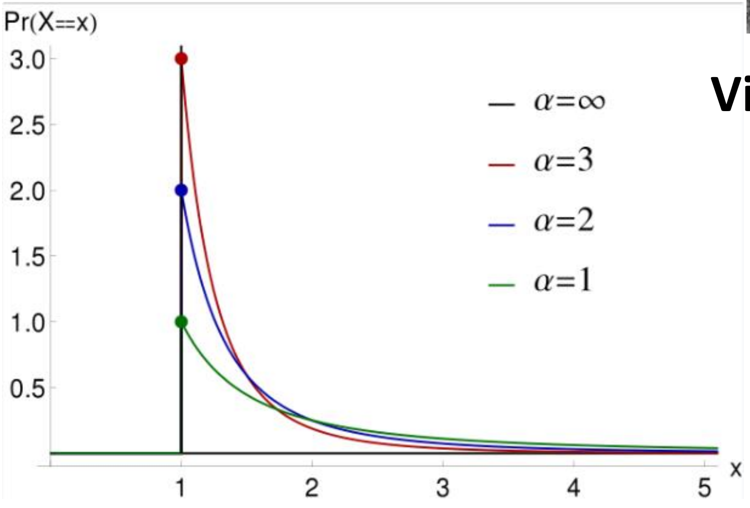
  Shape parameter s in (0,∞)

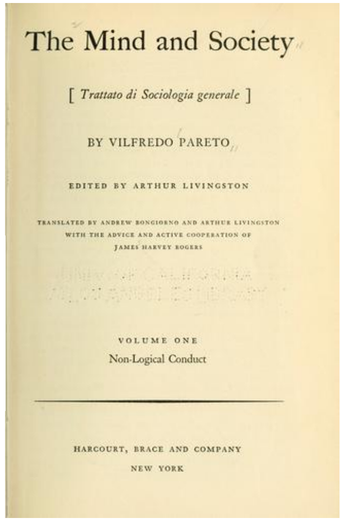
$$q_s(x) = \frac{s-1}{x^s}, \quad x > 1, s > 0$$



Vilfredo F. D. Pareto (1848-1923) Sociologist

- **Continuous exponential family** (of order 1):
  - Support (1,∞)
  - Natural parameter θ=s
  - sufficient statistics t(x)=-log(x)
  - log-normalizer F(θ)=-log(θ-1)
  - Differential entropy $-F^*(\eta)=1+1/(s-1)-\log(s-1)$
  - Fisher information $I(\theta)=1/(\theta-1)^2$




The Mind and Society

[*Trattato di Sociologia generale*]

BY VILFREDO PARETO

EDITED BY ARTHUR LIVINGSTON

TRANSLATED BY ANDREW BONGIORNO AND ARTHUR LIVINGSTON
WITH THE ADVICE AND ACTIVE COOPERATION OF
JAMES HARVEY ROGERS

VOLUME ONE
Non-Logical Conduct

HARCOURT, BRACE AND COMPANY
NEW YORK

- Two-parameter Pareto distributions:

$$\left\{ q_{s,a}(x) = (s-1)\frac{a^{s-1}}{x^s}, s > 1, a > 0 \right\}$$

Positive-curvature Fisher-Rao manifold

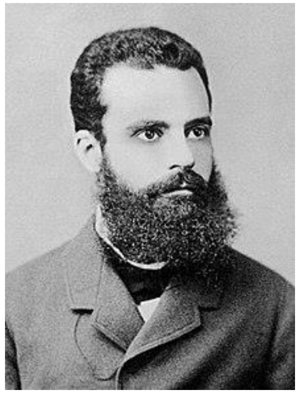
$$q_s(x) = q_{s,1}(x)$$

Power laws: Discrete versus continuous distributions

	Zeta distribution	Pareto distribution
	Exponential family $\exp(\theta t(x) - F(\theta))$	
	Discrete EF	Continuous EF
PMF/PDF	$p_s(x) = \frac{1}{x^s \zeta(s)}, \quad \zeta(s) = \sum_{i=1}^{\infty} \frac{1}{i^s}$	$q_s(x) = \frac{s-1}{x^s}$
Support $\mathcal{X}$	$\mathbb{N} = \{1, 2, \ldots\}$	$(1, \infty)$
Natural parameter θ	$s \in \Theta = (1, \infty)$	$s \in \Theta = (0, \infty)$
Cumulant $F(\theta)$	$\log \zeta(\theta)$	$-\log(\theta - 1)$
Sufficient statistic $t(x)$	$-\log x$	$-\log x$
Fisher information	$\sum_{i=0}^{\infty} \Lambda(i) \log(i) i^{-s}$	$\frac{1}{(s-1)^2}$

Power law distributions have long tail distributions

George Kingsley Zipf
1902-1950
American linguist



Vilfredo F. D. Pareto
(1848-1923)
Sociologist

Skewed Bhattacharrya coefficients and related statistical divergences

Bhattacharrya similarity coefficients: $I_\alpha[p_1, p_2] := \sum_{i=1}^{\infty} p_1(x)^\alpha p_2(x)^{1-\alpha}, \quad \alpha \in (0,1)$

α-divergences: $D_\alpha[p_{s_1} : p_{s_2}] := \frac{1}{\alpha(1-\alpha)} \left(1 - I_\alpha[p_{s_1} : p_{s_2}]\right)$

When the densities belong to the same exponential family:

get a **Jensen divergence** induced by log-normalizer: $I_\alpha[p_{s_1} : p_{s_2}] = \exp\left(-J_{F,\alpha}(s_1 : s_2)\right)$

$$J_{F,\alpha}(s_1 : s_2) \quad := \quad \alpha F(s_1) + (1-\alpha)F(s_2) - F(\alpha s_1 + (1-\alpha)s_2) \geq 0,$$

$$\boxed{D_\alpha[p_{s_1} : p_{s_2}] = \frac{1}{\alpha(1-\alpha)} \left(1 - \frac{\zeta(\alpha s_1 + (1-\alpha)s_2)}{\zeta(s_1)^\alpha \zeta(s_2)^{1-\alpha}}\right), \quad \alpha \in (0,1)}$$

α=1/2 yields the **squared Hellinger divergence**

Limit cases of α-divergences: α→±1 Forward and reverse Kullback-Leibler divergences

- **Kullback-Leibler divergence** : $$D_{\mathrm{KL}}[p_{s_1} : p_{s_2}] := \sum_{i=1}^{\infty} p_{s_1}(i) \log \frac{p_{s_1}(i)}{p_{s_2}(i)}$$
- Limit case α→1: $$\lim_{\alpha \to 1} D_\alpha[p_{s_1} : p_{s_2}] = D_{\mathrm{KL}}[p_{s_1} : p_{s_2}]$$
- Limit case α→-1: $$\lim_{\alpha \to -1} D_\alpha[p_{s_1} : p_{s_2}] = D_{\mathrm{KL}}[p_{s_2} : p_{s_1}] = D^*_{\mathrm{KL}}[p_{s_1} : p_{s_2}]$$

- Thus by choosing values for α close to ±1, we can approximate the forward or reverse Kullback-Leibler divergence:

$$D_{\mathrm{KL}}[p_{s_1} : p_{s_2}] \simeq D_{1-\epsilon}[p_{s_1} : p_{s_2}] = \frac{1}{\epsilon(1-\epsilon)} \left(1 - \frac{\zeta((1-\epsilon)s_1 + \epsilon s_2)}{\zeta(s_1)^{1-\epsilon}\zeta(s_2)^{\epsilon}}\right)$$

Efficient method in practice to approximate the KLD between zeta distributions

Kullback-Leibler divergence between two zeta distributions

- Since zeta distributions are exponential families, it amounts to a **reverse Bregman divergence**: $D_{\mathrm{KL}}[p_{s_1} : p_{s_2}] = B_F^*(\theta_1 : \theta_2) := B_F(\theta_2 : \theta_1)$
- or equivalently a **Fenchel-Young divergence** using **dual parameterizations**:

$$D_{\mathrm{KL}}[p_{s_1} : p_{s_2}] = B_F(\theta_2 : \theta_1) = F(\theta(s_2)) + F^*(\eta(s_1)) - \theta(s_2)\eta(s_1) := Y_{F,F^*}(\theta(s_2) : \eta(s_1))$$

- using the dual parameterization: $\eta(s) = F'(\theta(s)) = E_{p_s}[t(x)] = -E_{p_s}[\log x]$
- and **negentropy convex conjugate**: $F^*(\eta(s)) = -H[p_s] = \sum_{i=1}^{\infty} p_s(i) \log p_s(i)$
- Thus the KLD between two zeta distributions is

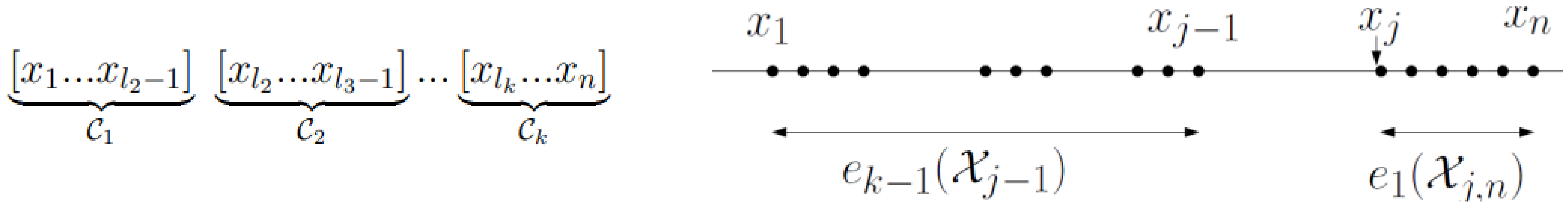
$$\begin{aligned} D_{\mathrm{KL}}[p_{s_1} : p_{s_2}] &= \log(\zeta(s_2)) - H[p_{s_1}] + s_2 E_{p_{s_1}}[\log x], \\ &= \log(\zeta(s_2)) - \sum_{i=1}^{\infty} \frac{1}{i^{s_1}\zeta(s_1)} \log(i^{s_1}\zeta(s_1)) - s_2 \frac{\zeta'(s_1)}{\zeta(s_1)}. \end{aligned}$$

$$\boxed{D_{\mathrm{KL}}[p_{s_1} : p_{s_2}] = \log(\zeta(s_2)) - \sum_{i=1}^{\infty} \frac{1}{i^s \zeta(s)} \log(i^s \zeta(s)) + s_2 \sum_{i=1}^{\infty} \frac{\Lambda(i)}{i^{s_1}}}$$

$$\eta(\theta) = \frac{\zeta'(\theta)}{\zeta(\theta)} = -\sum_{i=1}^{\infty} \frac{\Lambda(i)}{i^{\theta}},$$

Clustering finite sets of Zeta distributions

- **Agglomerative hierarchical clustering** (e.g.,single linkage, average linkage etc, get dendograms) or **partition-based k-means/k-center clustering**
- Since zeta distributions are 1D EFs, Kullback-Leibler divergence k-means clustering amounts to **1D Bregman k-means**: Optimal solution using dynamic programming. **Interval clustering** because Bregman Voronoi diagrams have connected cells.



- Multivariate product Zeta distributions can be clustered using Lloyd's k-means heuristics but then best k-means is then **NP-hard**

"Optimal interval clustering: Application to Bregman clustering and statistical mixture learning." IEEE Signal Processing Letters 21.10 (2014): 1289-1292.

Clustering sets of Zipf's distributions

- Problem: Zipf's distributions do not have the same support
- Consider **k-means prototypes** (cluster centers) as zeta distributions.
- Thus we need a distance between a Zipf's distribution (=right truncated zeta distribution) and a zeta distribution. This is a **duo Bregman divergence** (with two generators in dominance relation):

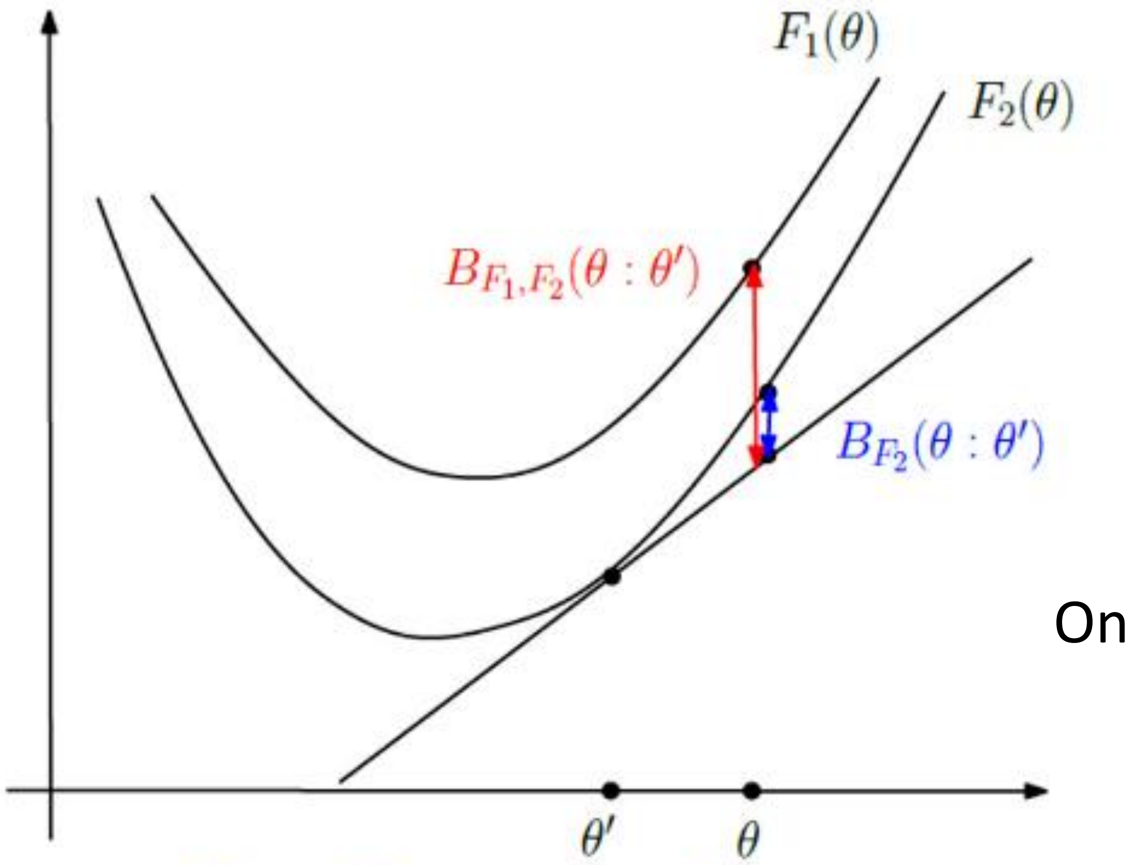
$$D_{\mathrm{KL}}[p_{\theta,N} : q_s] = B_{F_2,F_1}(s:\theta) = F_2(s) - F_1(\theta) - (s-\theta)F_1'(\theta)$$

$F_2(s) = \log \zeta(s)$ is larger than $F_1(\theta) = \log H_{N,\theta} = \log \sum_{i=1}^{N} \frac{1}{i^\theta}$

- Since $F_1'(\theta) = \frac{H'_{N,\theta}}{H_{N,\theta}} = -\sum_{i=1}^{N} \frac{\log i}{i^\theta H_{N,\theta}}$, we get

$$\boxed{D_{\mathrm{KL}}[p_{\theta,N} : q_s] = \log \frac{\zeta(s)}{\sum_{i=1}^{N} \frac{1}{i^\theta}} + (s-\theta) \sum_{i=1}^{N} \frac{\log i}{i^\theta H_{N,\theta}}}$$

Visualizing duo Bregman (pseudo-)divergences



Duo Bregman divergence

$$B_{F_1,F_2}(\theta:\theta') = F_1(\theta) - F_2(\theta') - (\theta-\theta')^\top \nabla F_2(\theta')$$

Generator F_1 dominates generator F_2

Only **pseudo divergence** because strictly positive

Recover Bregman divergences when $F=F_1=F_2$

"Statistical Divergences between Densities of Truncated Exponential Families with Nested Supports: Duo Bregman and Duo Jensen Divergences." Entropy 24.3 (2022)

Comparisons of discrete vs continuous power laws

	Zeta distribution	Pareto distribution
	Exponential family $\exp(\theta t(x) - F(\theta))$	
	Discrete EF	Continuous EF
PMF/PDF	$p_s(x) = \frac{1}{x^s \zeta(s)}, \quad \zeta(s) = \sum_{i=1}^{\infty} \frac{1}{i^s}$	$q_s(x) = \frac{s-1}{x^s}$
Support $\mathcal{X}$	$\mathbb{N} = \{1, 2, \ldots\}$	$(1, \infty)$
Natural parameter θ	$s \in \Theta = (1, \infty)$	$s \in \Theta = (0, \infty)$
Cumulant $F(\theta)$	$\log \zeta(\theta)$	$-\log(\theta - 1)$
Sufficient statistic $t(x)$	$-\log x$	$-\log x$
Moment parameter $\eta = -E[\log x]$	$\frac{\zeta'(\theta)}{\zeta(\theta)}$	$-\frac{1}{s-1}$
Mean	$\frac{\zeta(s-1)}{\zeta(s)}$	$\frac{s}{s-1}$
Variance	$\frac{\zeta(s)\zeta(s-2)-\zeta(s-1)^2}{\zeta(x)^2}, s > 3$	$\frac{s}{(s-1)^2(s-2)}, s > 2$
Conjugate $F^*(\eta)$	$-H[p_s] = -\sum_{i=1}^{\infty} \frac{1}{i^s \zeta(s)} \log(i^s \zeta(s))$	$\eta - 1 - \log(-\eta)$
Maximum likelihood estimator	$\hat{\eta} = \frac{\zeta'(\hat{\theta})}{\zeta(\hat{\theta})} = -\frac{1}{n} \sum_{i=1}^{n} \log x_i$	$\hat{s} = \frac{n}{\sum_{i=1}^{n} \log x_i}$
Fisher information	$\sum_{i=0}^{\infty} \Lambda(i) \log(i) i^{-s}$	$\frac{1}{(s-1)^2}$
Entropy $-F^*(\eta(s))$	$\sum_{i=1}^{\infty} \frac{1}{i^s \zeta(s)} \log(i^s \zeta(s))$	$1 + \frac{1}{s-1} - \log(s-1)$
Bhattacharyya coefficient I_α	$\frac{\zeta(\alpha s_1 + (1-\alpha) s_2)}{\zeta(s_1)^\alpha \zeta(s_2)^{1-\alpha}}$	$\frac{\alpha s_1 + (1-\alpha) s_2}{s_1^\alpha s_2^{1-\alpha}}$
Kullback-Leibler divergence D_{KL}	$\log(\zeta(s_2)) - \sum_{i=1}^{\infty} \frac{1}{i^s \zeta(s)} \log(i^s \zeta(s)) - s_2 \frac{\zeta'(s_1)}{\zeta(s_1)}$	$\log\left(\frac{s_1 - 1}{s_2 - 1}\right) + \frac{s_2 - s_1}{s_1 - 1}$

Notice: Differential entropy may be negative (large s) but never the discrete entropy!

Drawing discrete/continuous power law variates

Zeta variates:

$$p_s(x) = \frac{1}{\zeta(s)} \frac{1}{x^s}$$

Acceptance/rejection method:

- Draw $u_1 \sim \mathrm{Unif}(0,1)$ and $u_2 \sim \mathrm{Unif}(0,1)$
- Let $x = \left\lfloor u_1^{-\frac{1}{s-1}} \right\rfloor$ and $t = (1 + \frac{1}{x})^{s-1}$.
- Accept x if $x \leq \frac{t}{t-1} \frac{2^{s-1}-1}{2^{s-1}u_2}$

may require several rounds

Truncated Pareto variates:

$$q_s^{a,b}(x) = \frac{(s-1)a^{s-1}}{\left(1 - \left(\frac{a}{b}\right)^{s-1}\right) \; x^s} \qquad \mathcal{X} = (a,b)$$

Inverse transform method:

$$u \sim \mathrm{Unif}(0,1)$$

$$x = ab\left(b^{s-1} - u(b^{s-1} - a^{s-1})\right)^{-\frac{1}{s-1}}$$

Thank you!

[arXiv:2104.10548]

Related work:
Discrete vs Continuous Normal distribution

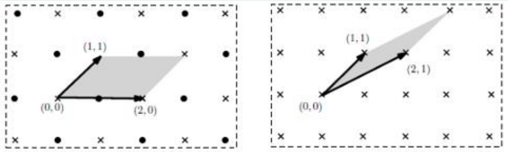
The Kullback–Leibler Divergence Between Lattice Gaussian Distributions

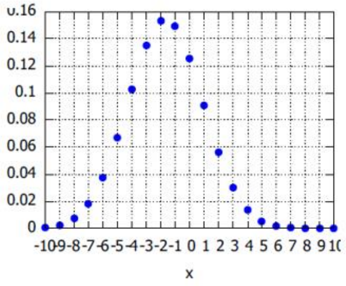
J. Indian Inst. Sci.
A Multidisciplinary Reviews Journal
ISSN: 0970-4140 Coden-JIISAD

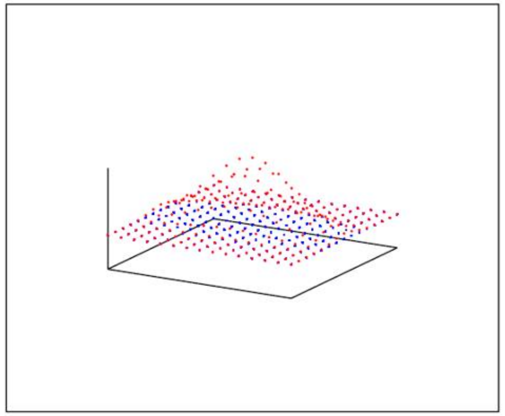
$$p_\xi(l) = \frac{1}{\theta_\Lambda(\xi)} \exp\left(2\pi\left(-\frac{1}{2} l^\top \xi_2 l + l^\top \xi_1\right)\right), \quad l \in \Lambda,$$

$$\theta_\Lambda(\xi) := \sum_{l\in\Lambda} \exp\left(2\pi\left(-\frac{1}{2} l^\top \xi_2 l + l^\top \xi_1\right)\right),$$

$$\theta_\Lambda(\xi) = \theta(-iL^\top \xi_1, iL^\top \xi_2 L),$$







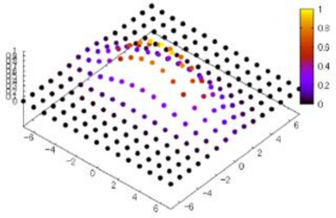


Table 1: Summary of statistical divergences with corresponding formula for lattice Gaussian distributions with partition function $\theta_\Lambda(\xi)$. Ordinary parameterization $\lambda(\xi) = (\mu = E_{p_\xi}[X], \Sigma = \mathrm{Cov}_{p_\xi}[X])$ for $X \sim N_\Lambda(\xi)$.

Divergence	Definition
	Closed-form formula for lattice Gaussians
Kullback–Leibler divergence	$D_{\mathrm{KL}}[p_\xi : p_{\xi'}] = \sum_{l\in\Lambda} p_\xi(l) \log \frac{p_\xi(l)}{p_{\xi'}(l)}$
	$D_{\mathrm{KL}}[p_\xi : p_{\xi'}] = \log\left(\frac{\theta_\Lambda(\xi')}{\theta_\Lambda(\xi)}\right)$
	$-2\pi\mu^\top(\xi_1' - \xi_1) + \pi\,\mathrm{tr}\left((\xi_2' - \xi_2)(\Sigma + \mu\mu^\top)\right)$
squared Hellinger divergence	$D^2_{\mathrm{Hellinger}}[p_\xi : p_{\xi'}] = \frac{1}{2}\sum_{l\in\Lambda}(\sqrt{p_\xi(l)} - \sqrt{p_{\xi'}(l)})^2$
	$D^2_{\mathrm{Hellinger}}[p_\xi : p_{\xi'}] = 1 - \frac{\theta_\Lambda\left(\frac{\xi+\xi'}{2}\right)}{\sqrt{\theta_\Lambda(\xi)\theta_\Lambda(\xi')}}$
Rényi α-divergence	$D_\alpha[p_\xi : p_{\xi'}] = \frac{1}{\alpha-1}\log\left(\sum_{l\in\Lambda} p_\xi(l)^\alpha p_{\xi'}(l)^{1-\alpha}\right)$
$(\alpha > 0, \alpha \neq 1)$	$D_\alpha[p_\xi : p_{\xi'}] = \frac{\alpha}{1-\alpha}\log\frac{\theta_\Lambda(\xi)}{\theta_\Lambda(\alpha\xi+(1-\alpha)\xi')} + \log\frac{\theta_\Lambda(\xi')}{\theta_\Lambda(\alpha\xi+(1-\alpha)\xi')}$
	$\lim_{\alpha\to 1} D_\alpha[p_\xi : p_{\xi'}] = D_{\mathrm{KL}}[p_\xi : p_{\xi'}]$
γ-divergence	$\vec{D}_\gamma[p_\xi : p_{\xi'}] = \frac{1}{\gamma(\gamma-1)}\log\left(\frac{\left(\sum_{l\in\Lambda} p_\xi^\gamma(x)\right)\left(\sum_{l\in\Lambda} p_{\xi'}^\gamma(l)\right)^{\gamma-1}}{\left(\sum_{l\in\Lambda} p_\xi(l) p_{\xi'}^{\gamma-1}(l)\right)^\gamma}\right)$
$(\gamma > 1)$	$\vec{D}_\gamma[p_\xi : p_{\xi'}] = \frac{1}{\gamma(\gamma-1)}\log\left(\frac{\theta_\Lambda(\gamma\xi)\theta_\Lambda(\gamma\xi')^{\gamma-1}}{\theta_\Lambda(\xi+(\gamma-1)\xi')^\gamma}\right)$
	$\lim_{\gamma\to 1}\vec{D}_\gamma[p_\xi : p_{\xi'}] = D_{\mathrm{KL}}[p_\xi : p_{\xi'}]$
Hölder divergence	$\vec{D}^H_{\alpha,\gamma}[r : s] := \left\lvert\log\left(\frac{\sum_{x\in\mathcal{X}} r(x)^{\gamma/\alpha} s(x)^{\gamma/\beta}}{\left(\sum_{x\in\mathcal{X}} r(x)^\gamma\right)^{1/\alpha}\left(\sum_{x\in\mathcal{X}} s(x)^\gamma\right)^{1/\beta}}\right)\right\rvert$
$(\gamma > 0, \frac{1}{\alpha} + \frac{1}{\beta} = 1)$	$\vec{D}^H_{\alpha,\gamma}[p_\xi : p_{\xi'}] = \left\lvert\log\frac{\theta_\Lambda(\gamma\xi)^{\frac{1}{\alpha}}\theta_\Lambda(\gamma\xi')^{\frac{1}{\beta}}}{\theta_\Lambda(\frac{\gamma}{\alpha}\xi + \frac{\gamma}{\beta}\xi')}\right\rvert$
Cauchy–Schwarz divergence	$\vec{D}_{\mathrm{CS}}[r : s] := -\log\frac{\sum_{x\in\mathcal{X}} r(x)s(x)}{\sqrt{\left(\sum_{x\in\mathcal{X}} r^2(x)\right)\left(\sum_{x\in\mathcal{X}} s^2(x)\right)}}$
(Hölder with $\alpha = \beta = \gamma = 2$)	$\vec{D}_{\mathrm{CS}}[p_\xi : p_{\xi'}] = \log\frac{\sqrt{\theta_\Lambda(2\xi)\theta_\Lambda(2\xi')}}{\theta_\Lambda(\xi+\xi')}$

"The Kullback–Leibler Divergence Between Lattice Gaussian Distributions." Journal of the Indian Institute of Science (2022): 1-12.