Estimation, Prediction, and Classification over Large Alphabets

A. Orlitsky, UCSD

with

J. Acharya H. Das S. Pan

P. Santhanam K. Viswanathan J. Zhang

2 Probability Estimation Problems

Probability of each element

P(Slovenia)=.3, P(U.S.)=.7

P(Black)=.3, P(Blue)=.2, P(Brown)=.4 P(Green)=.1

Probability mulstiset

{.7, .3}

{.4, .3, .2, .1}

Estimate using sample

Patterns

Patterns

Replace each symbol by its order of appearance

Sequence: US, Slovenia, US, US, Slovenia

Pattern: 1, 2, 1, 1, 2

Sequence: Brown, Blue, Brown, Green, Brown, black, Blue

Pattern: 1, 2, 1, 3, 1, 4, 2

b,g,b → 1, 2, 1

 → 1, 2, 3, 2

Capture structure and frequencies, ignore symbols

Two Estimation Problems

Probability of each element

2

P(Slovenia)=.3, P(U.S.)=.7

P(Black)=.3, P(Blue)=.2, P(Brown)=.4 P(Green)=.1

Probability mulstiset

{.7, .3}

{.4, .3, .2, .1}

1

Estimate using sample

Patterns

Applications

Probability multiset {.5, .3, .1, .05, .05}

Population: species, gene mutations, sensors

Coverage: experiments, virus strains

Robustness: capacity, memory, buffer

Security: unlikely events

Information: entropy

Statistics

Who cares?

Everyone

Martianese

Martian UFO has landed

First message

@#@#@@#...@#@ (60 @'s, 40 #'s)

Statisticians analyze Martianese

Assuming independence they determine

2 symbols

one has probability .6, other .4 } {.6, .4}

Methodology

Empirical frequency or Maximum likelihood

Maximum likelihood

Distribution maximizing $P(\bar{x})$

$\bar{x}$ = @#@#@@#...@#@ (60 @, 40 #)

$p(@) = p, \quad p(\#) = 1-p$

$P(\bar{x}) = p^{60}(1-p)^{40}$ (independence)

$P'(\bar{x}) = 60\ p^{59}(1-p)^{40} - 40\ p^{60}(1-p)^{39}$

$= p^{59}(1-p)^{39}\ [\ 60(1-p) - 40p\ ] = 0$

Maximized by $p(@) = p = .6 \quad p(\#) = .4$

→ {.6, .4}

Agrees with empirical frequency, always

20 @, 50 #, 30 & → {.2, .5, .3} = {.5, .3, .2}

Works well for small alphabets

poorly large

Marchinese

這講座非常悶。。。不過請報以大笑

100 symbols, all distinct

Sequence maximum likelihood

Alphabet size ?100

Each symbol probability ?1/100

Poor model, better: ?large, possibly ∞ alphabet

Observation: Values don't matter, consider pattern

Pattern maximum likelihood

Pattern: ? 2 3 4 . . . 98 99 100

All different

Maximized by: ?

Different! Better for very large α/β

Why?
Smaller α/β?

Sequences v. patterns

Sequence independently, identically distributed (iid)

Induces probability on patterns

Ψ - pattern, e.g. 121

$P(\Psi)$ = P(observing Ψ) = P(all outcomes with pattern Ψ)

Example: Alphabet {a, b, c}

$p(a) = p(b) = p(c) = 1/3$

1 sample: $P(1) = 1$

2 samples: $P(11)$ = P(aa, bb, cc) = 3 * 1/9 = 1/3

= P(second = first)

$P(12)$ = P(ab, ac, ba, bc, ca, cb) = 6 * 1/9 = 2/3

= P(second ≠ first)

Pattern not distributed iid → Different math

Pattern maximum likelihood

Given a pattern, e.g. 1213

Find highest probability $\hat{P}$, and achieving distribution

That's it!

Length 1: $\hat{P}(1) = 1$, any distribution

2: $\hat{P}(11) = 1$, constant distributions

$\hat{P}(12) = 1$, large distributions

3: $\hat{P}(111) = 1$, $\hat{P}(123) = 1$

$\hat{P}(112) = ? = \hat{P}(121) = \hat{P}(122)$

Flip 3 coins: h,h,t

Sequence ML → {2/3, 1/3}

Pattern ML → ?

$\hat{P}(112) = 1/4$

$\geq$ Coin, P(h) = P(t) = .5

P(112) = P(hht V tth) = P(hht) + P(tth)

= 1/8 + 1/8 = 1/4

$\leq$ P(112) = $\Sigma_x\ p^2(x) \cdot (1-p(x))$

= $\Sigma_x\ p(x) \cdot [\ p(x)\ (1-p(x))\ ]$

$\leq \Sigma_x\ p(x) \cdot [\ 1/4\ ]$

= 1/4

h,h,t: Sequence ML: {2/3, 1/3}, Pattern ML: {.5, .5}

Which more logical? Not p(h) & p(t), just multiset

3 flips, can't get 1.5 each → 2&1 closest to uniform

{2/3, 1/3} & others → 3 same more likely → 2&1 less likely

Theoretical Results

Skip

Experiments

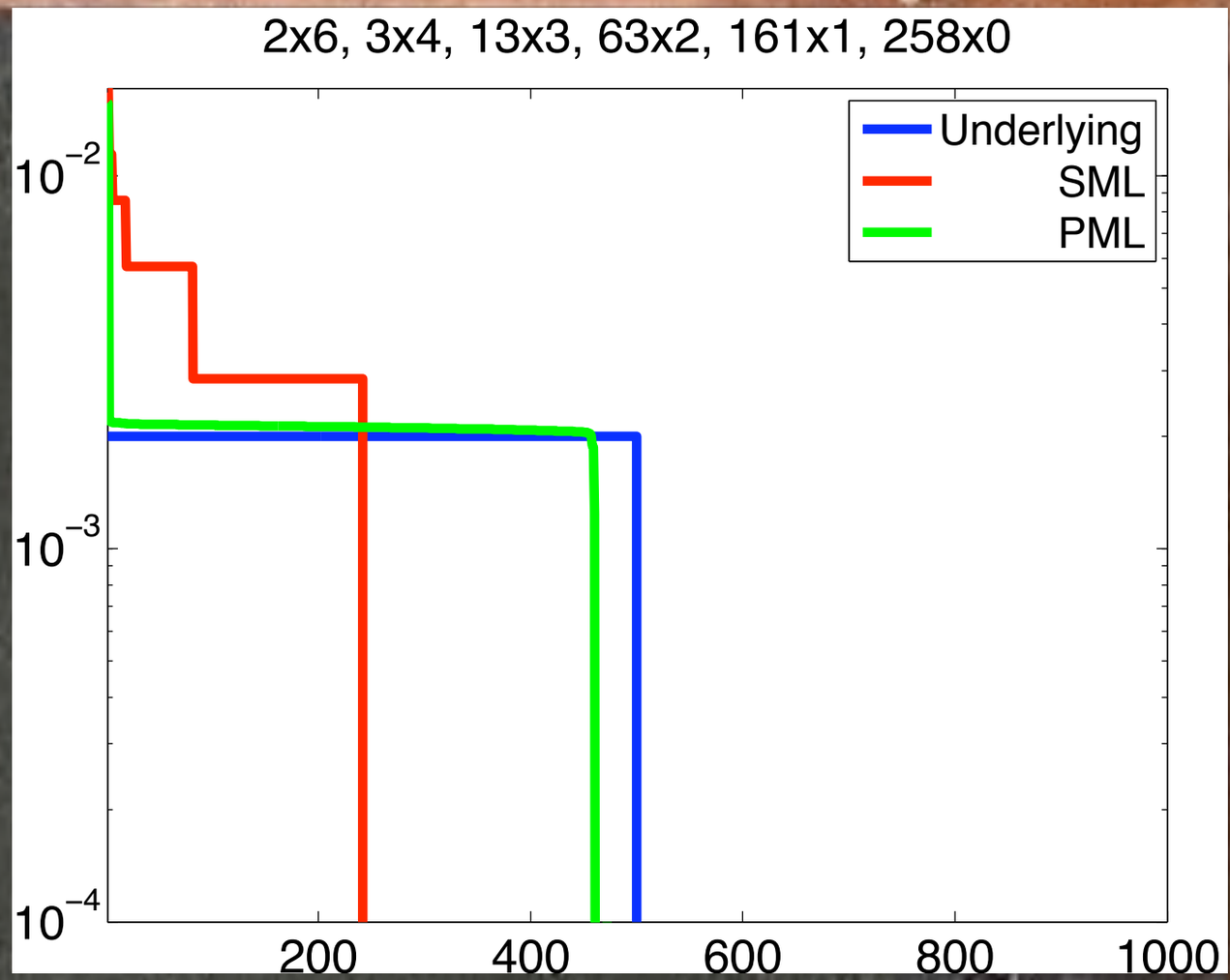
Uniform

500 symbols

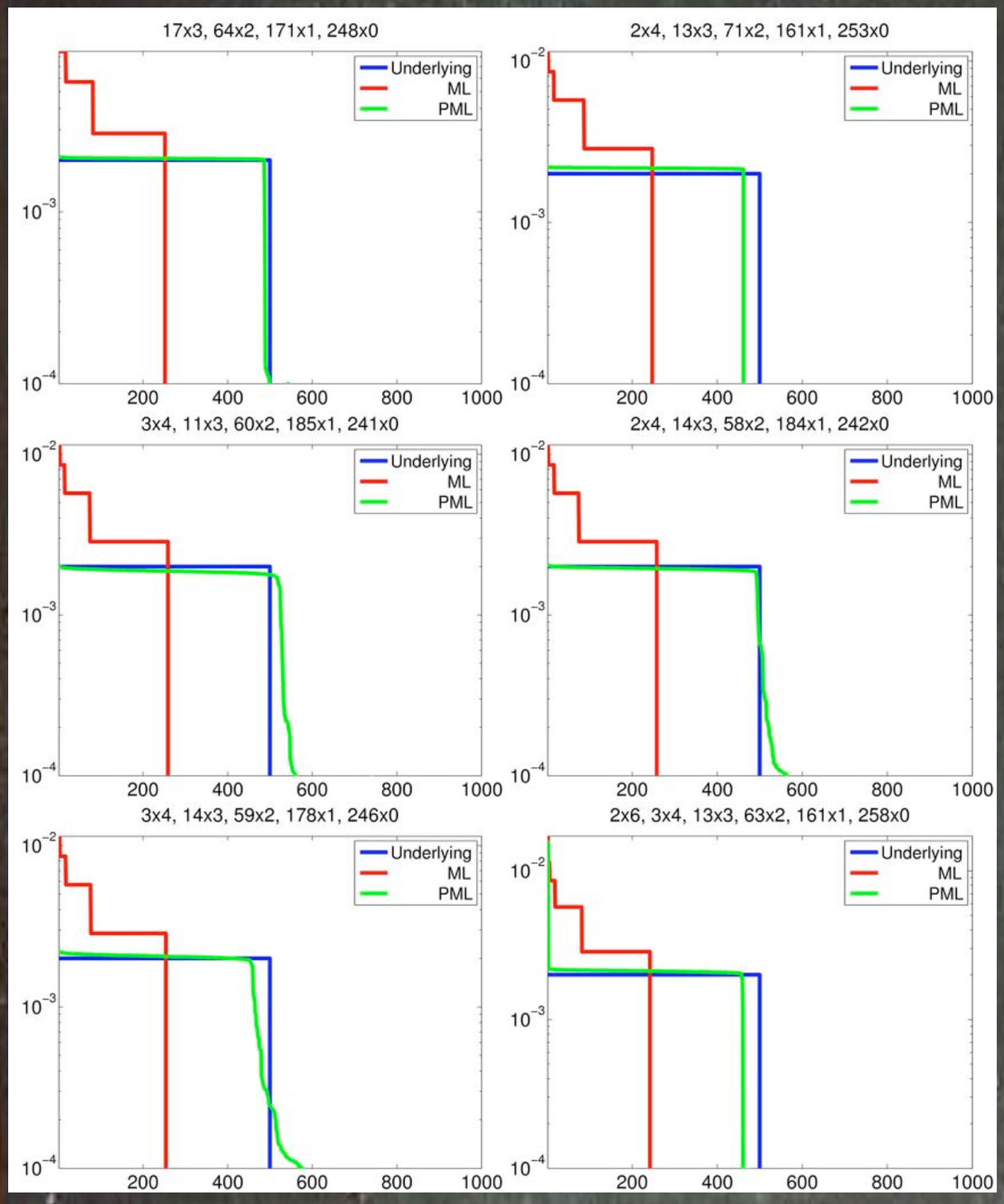
350 samples

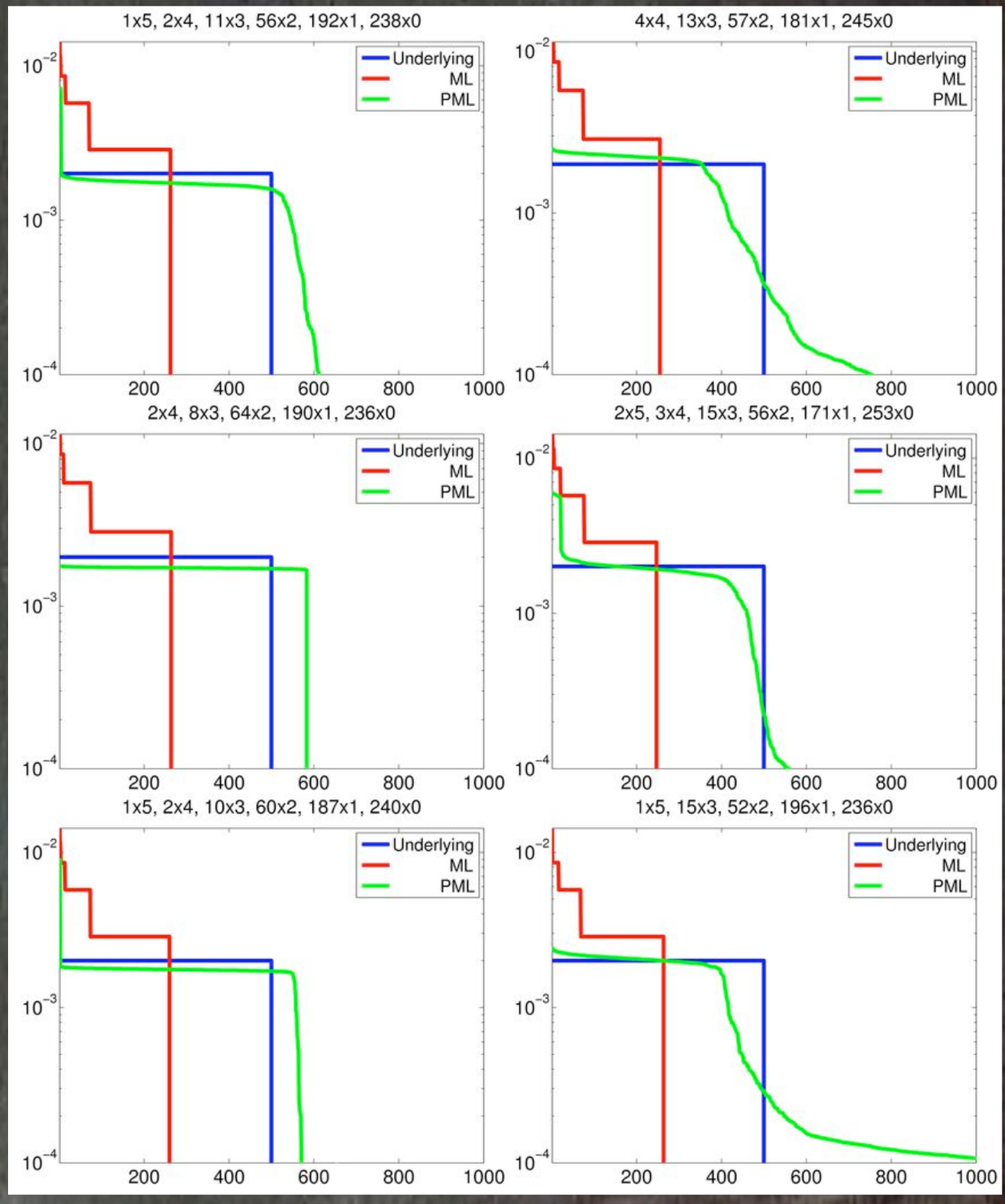
2x6, 3x4, 13x3, 63x2, 161x1

242 appeared, 258 did not



U[500], 350x, 12 experiments





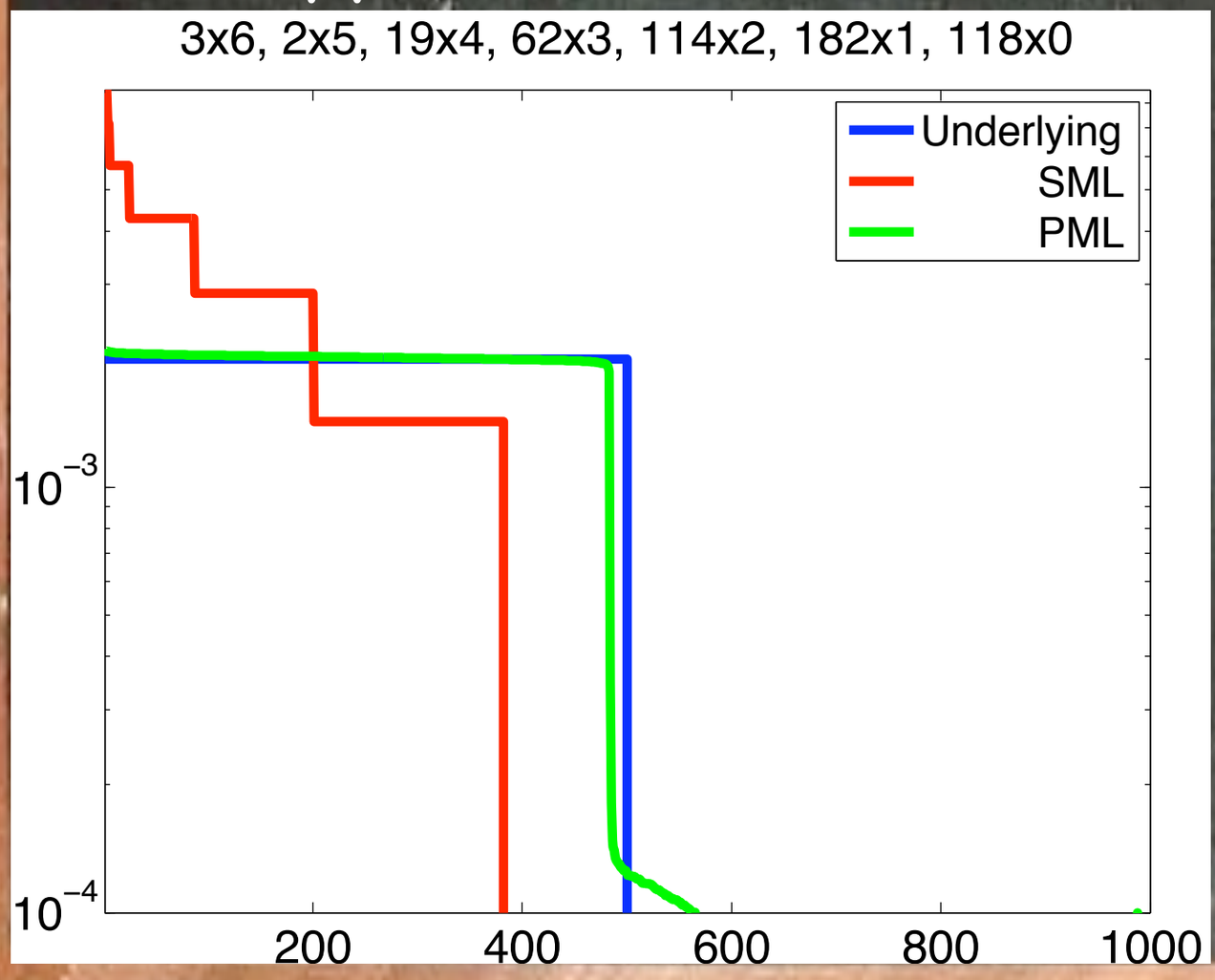
Uniform

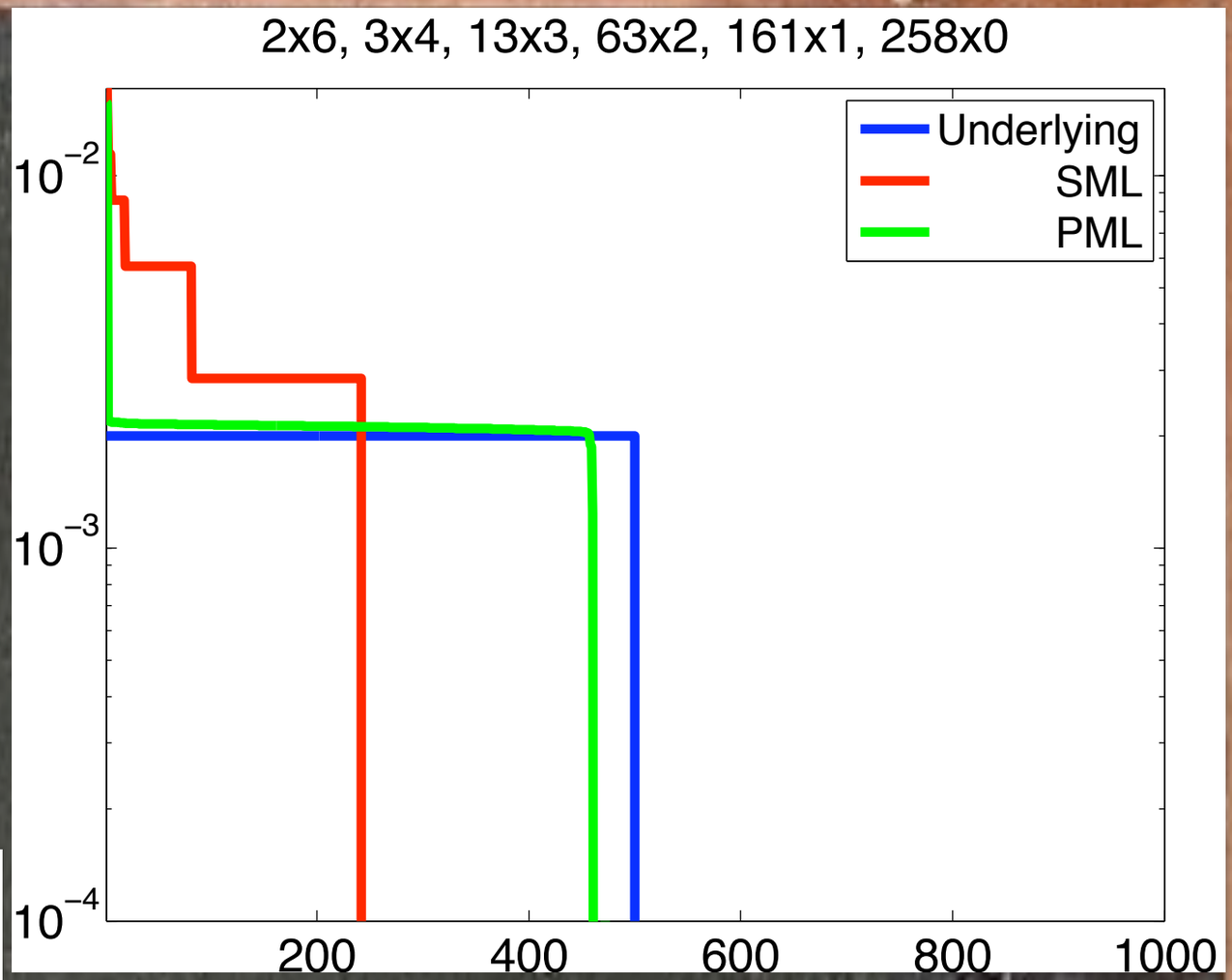
500 symbols

350 samples

2x6, 3x4, 13x3, 63x2, 161x1

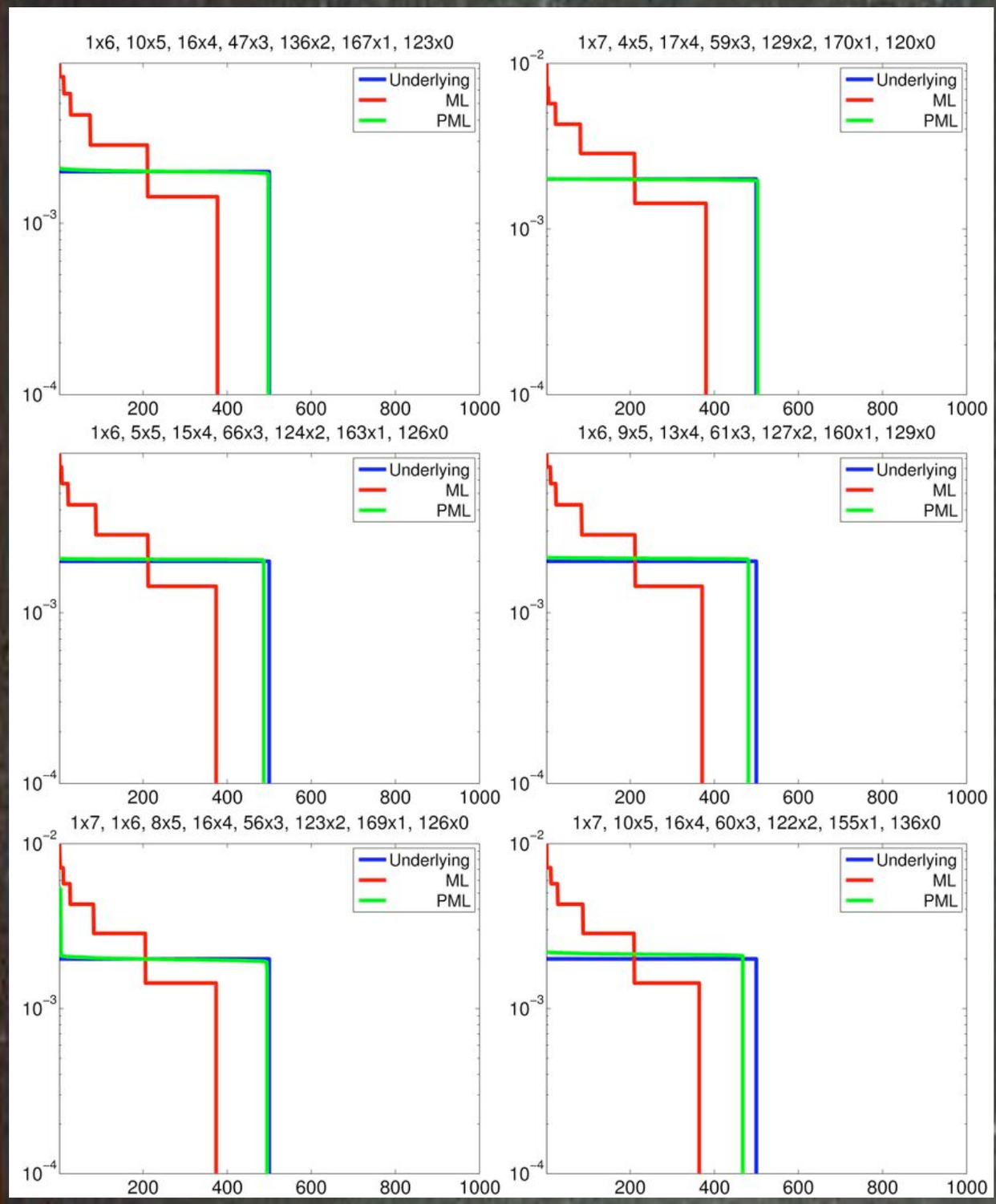
248 appeared, 258 did not

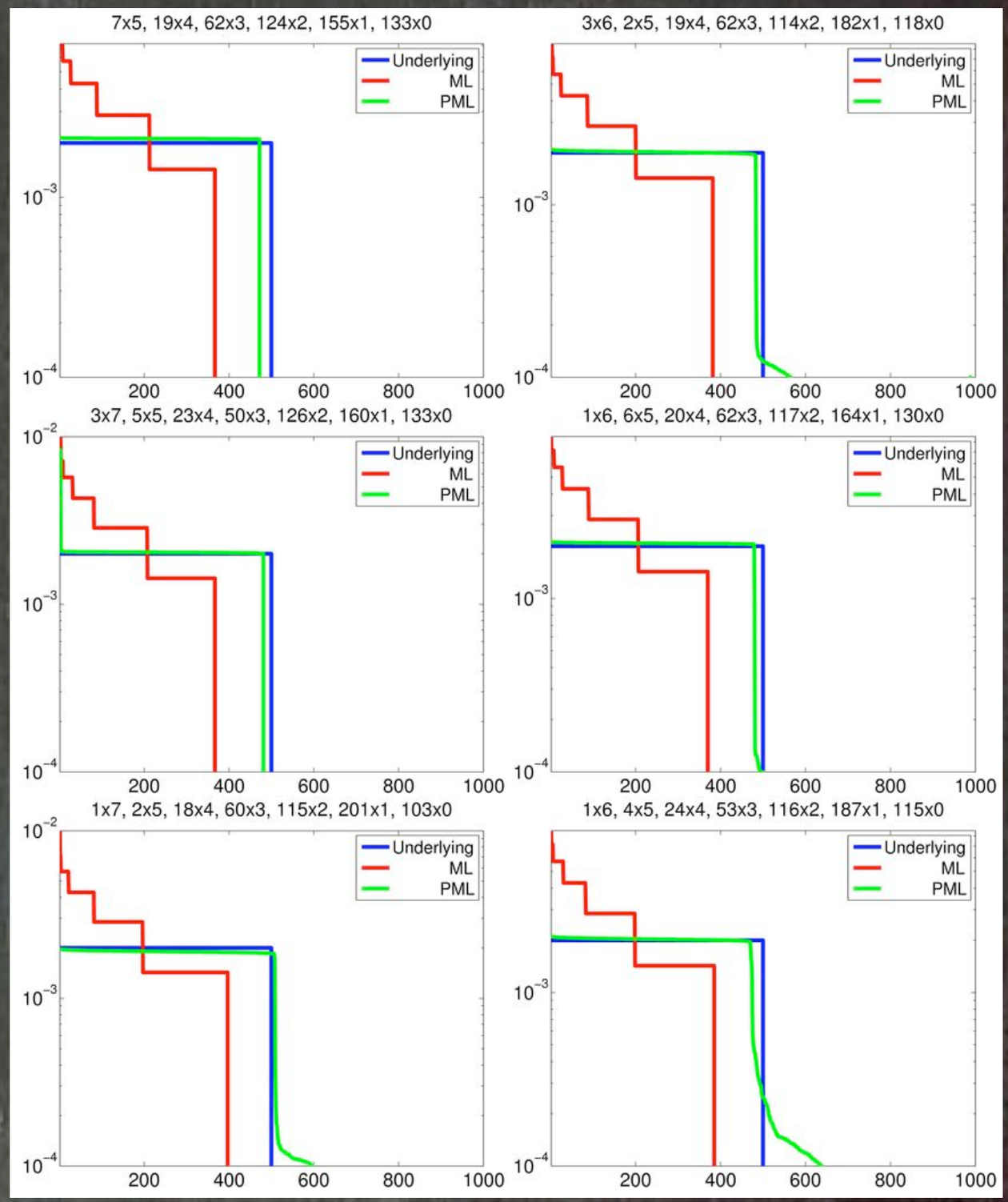




700 samples

U[500], 700x, 12 experiments





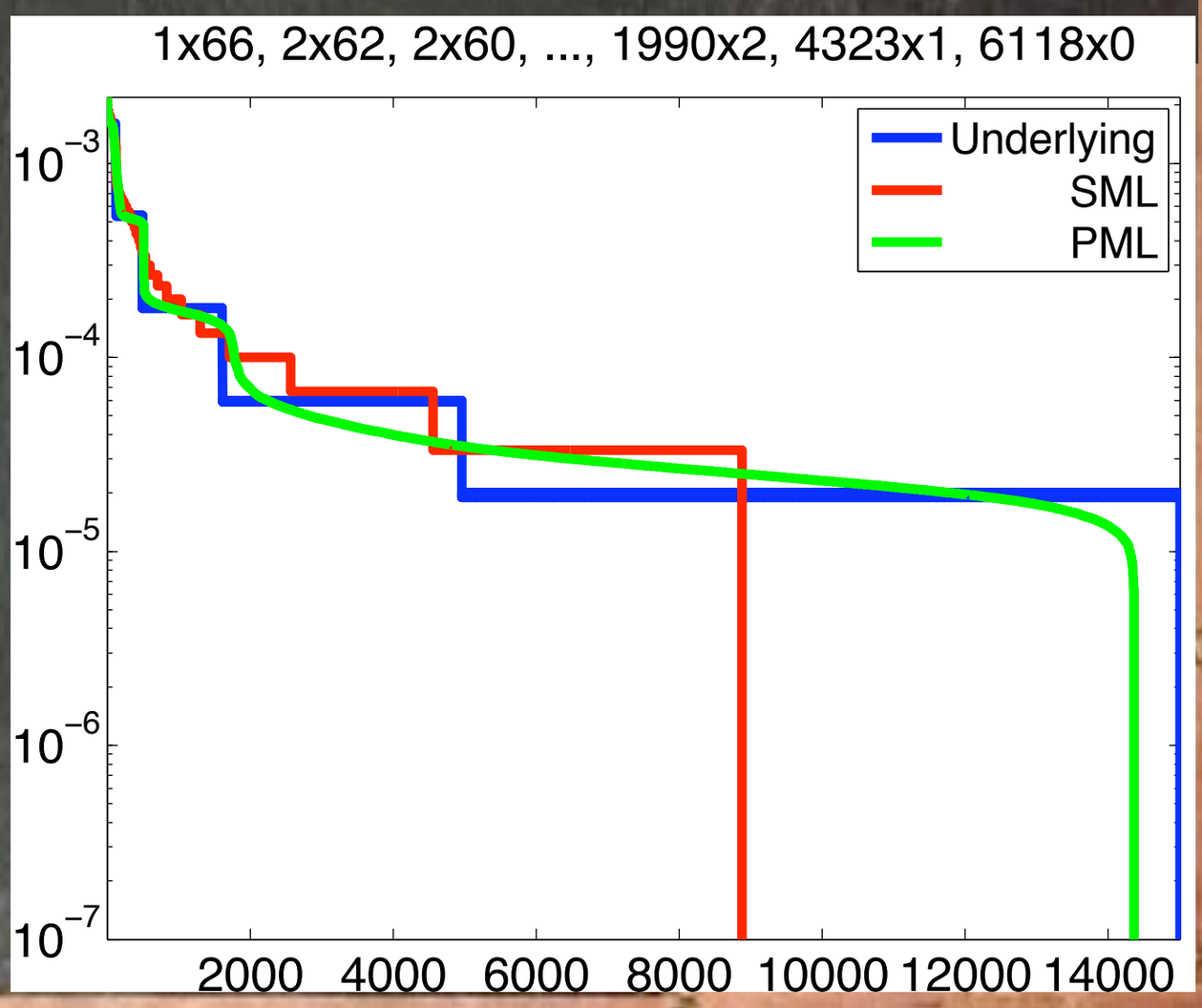
Staircase

15K elements, 5 steps, ~3x

30K samples

Observe 8,882 elts

6,118 missing



Zipf

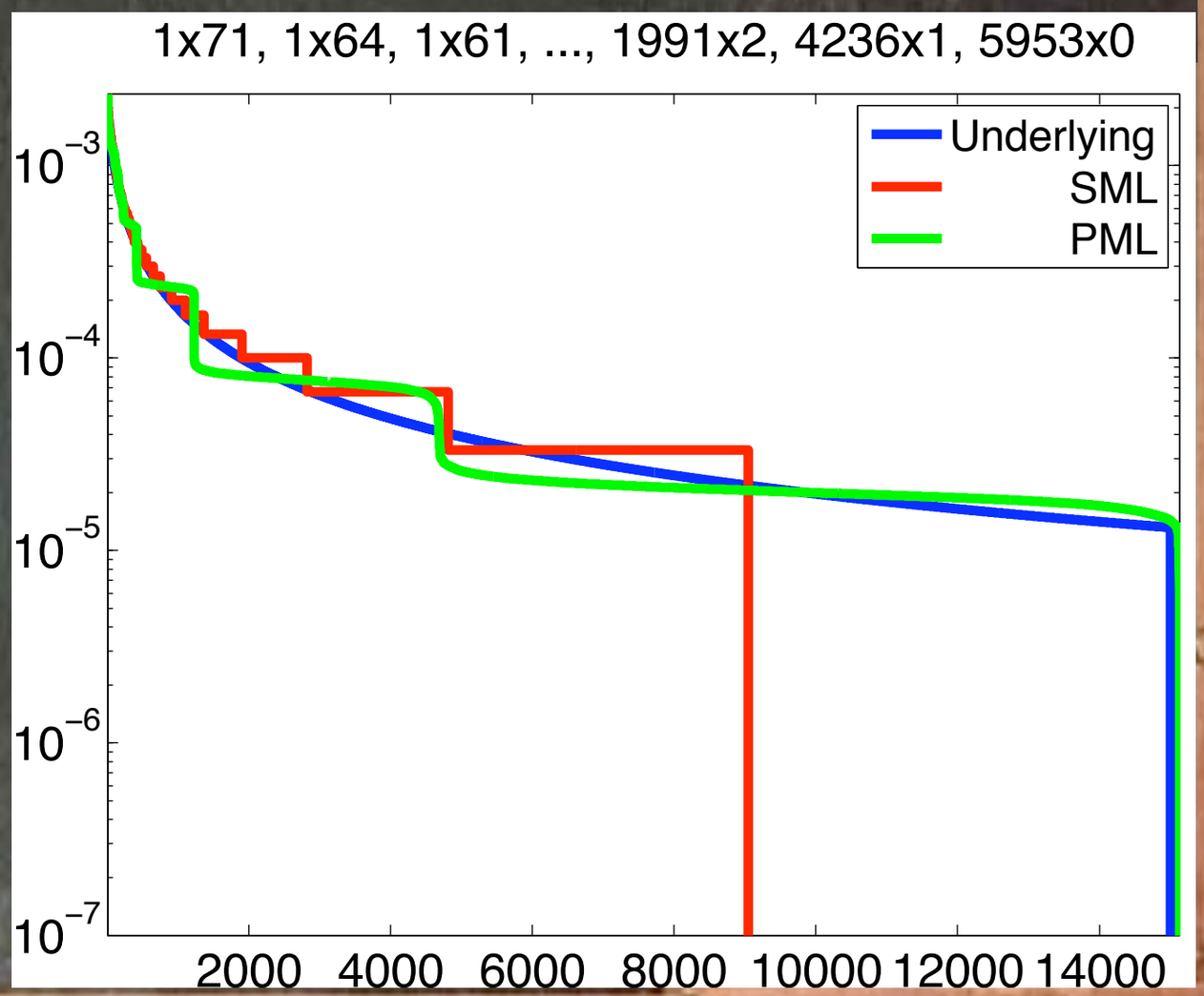
Underlies many natural phenomena

$p_i = C/i$, i=100...15,000

30,000 samples

Observe 9,047 elts

5,953 missing



1990 Census - Last names

SMITH	1.006	1.006	1
JOHNSON	0.810	1.816	2
WILLIAMS	0.699	2.515	3
JONES	0.621	3.136	4
BROWN	0.621	3.757	5
DAVIS	0.480	4.237	6
MILLER	0.424	4.660	7
WILSON	0.339	5.000	8
MOORE	0.312	5.312	9
TAYLOR	0.311	5.623	10
⋮			
AMEND	0.001	77.478	18835
ALPHIN	0.001	77.478	18836
ALLBRIGHT	0.001	77.479	18837
AIKIN	0.001	77.479	18838
ACRES	0.001	77.480	18839
ZUPAN	0.000	77.480	18840
ZUCHOWSKI	0.000	77.481	18841
ZEOLLA	0.000	77.481	18842

18,839 names

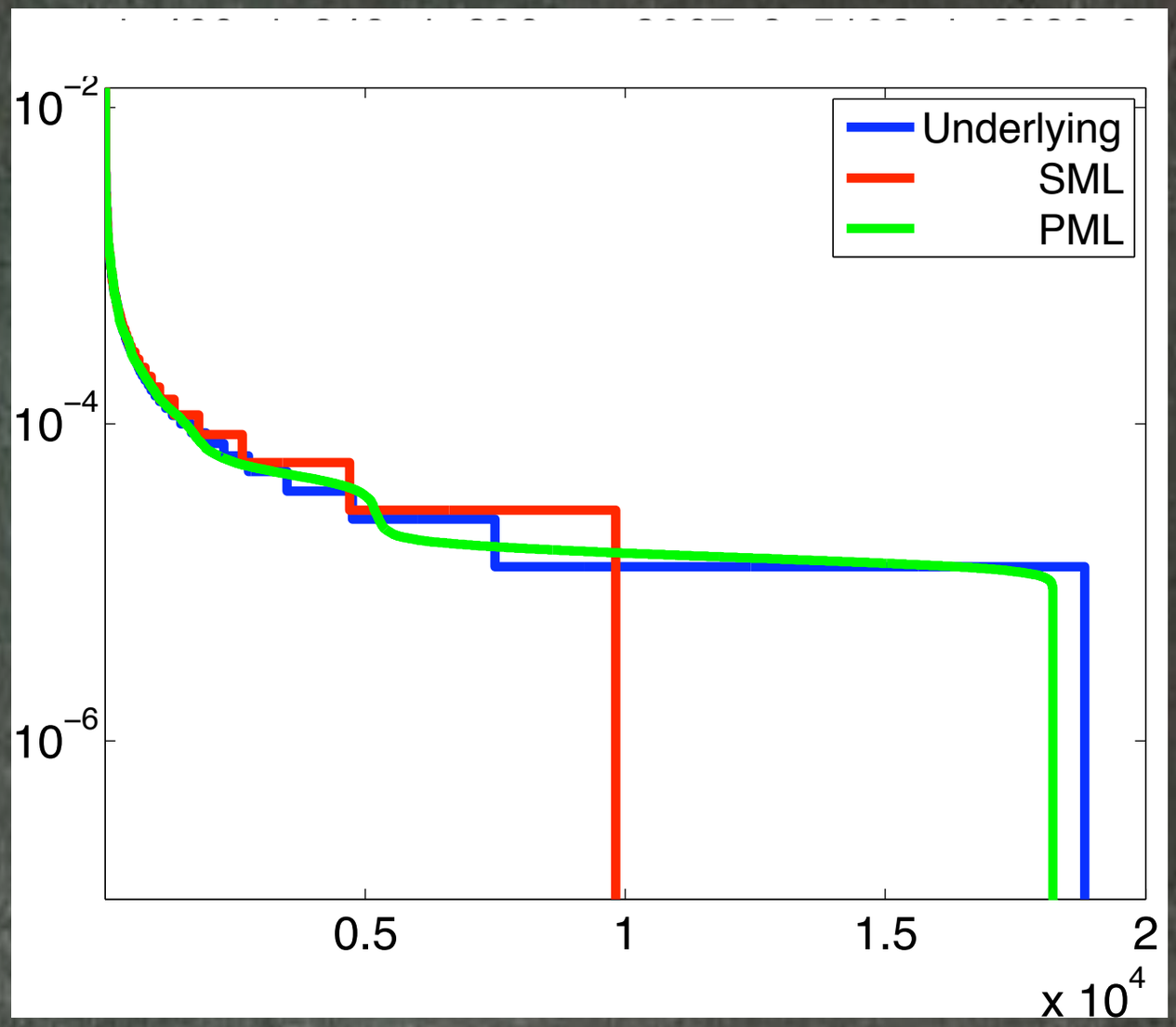
77.48% population

~230 million

1990 Census - Last names

18,839 last names based on ~230 million

35,000 samples, observed 9,813 names



Population estimation

Malayan butterflies [Fisher Corbet Williams 43]

Maximum likelihood doesn't work

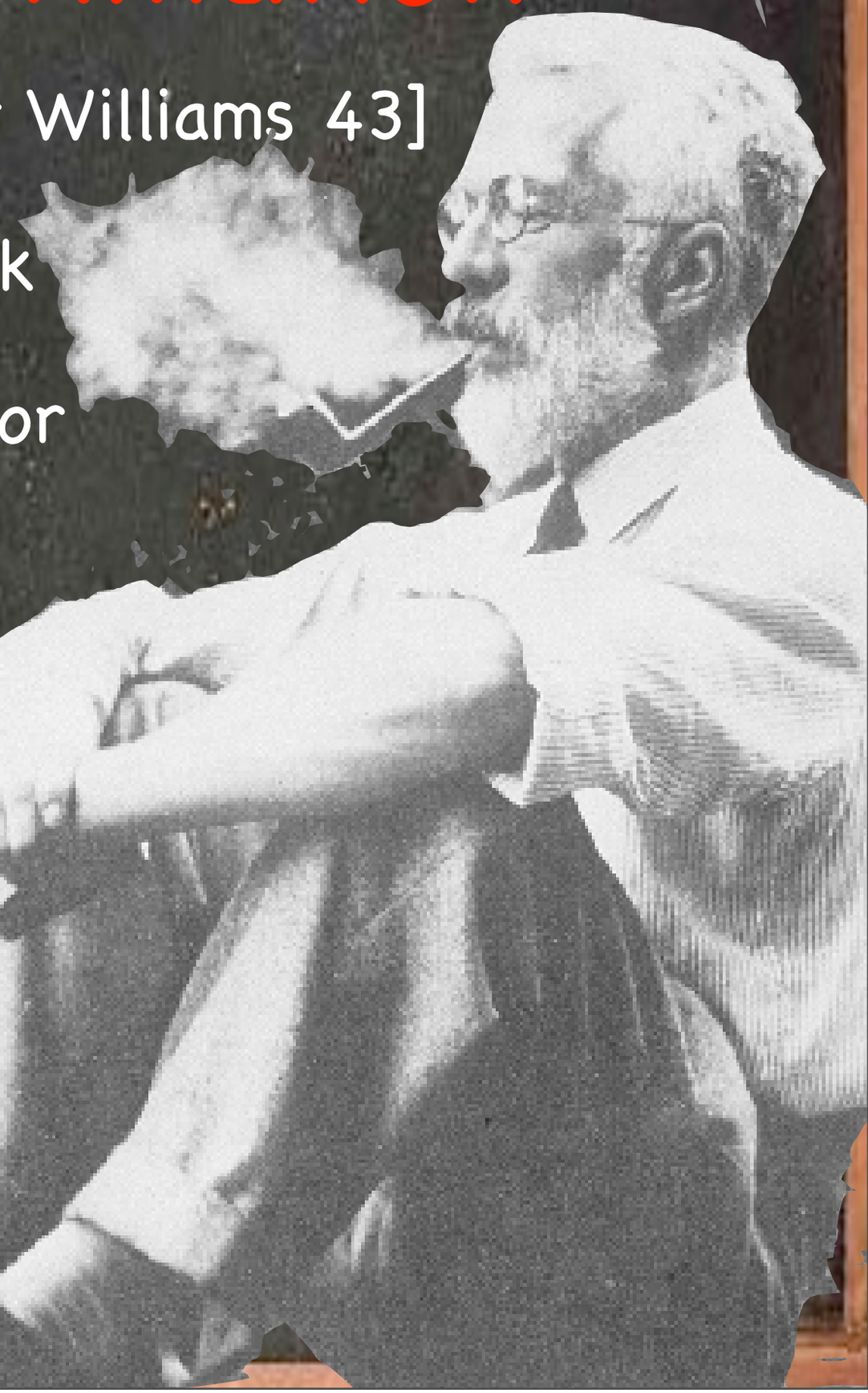
Poisson distribution, Gamma prior

New species [Good Toulmin 56]

Mutations, Strains

Sensors

Many others



Shakespeare

899,306 words, 27,689 distinct (half appear once)

[76]:

Estimating the number of unseen species: How many words did Shakespeare know?

BY BRADLEY EFRON AND RONALD THISTED

Department of Statistics, Stanford University, California

In equal size: 11,430 new words

[85, Taylor]: "Shall I die?"

429 words, 9 new: joying, scanty, speck,....,

[87]:

Did Shakespeare write a newly-discovered poem?

BY RONALD THISTED

Department of Statistics, University of Chicago, Chicago, Illinois 60637, U.S.A.

AND BRADLEY EFRON

Department of Statistics, Stanford University, Stanford, California 94305, U.S.A.

Expect 6.93 new

Pattern maximum likelihood: 6.98 new

new symbols

Zipf distribution over 15K elements

Sample 30K times

Estimate: # new symbols in sample of size λ * 30K

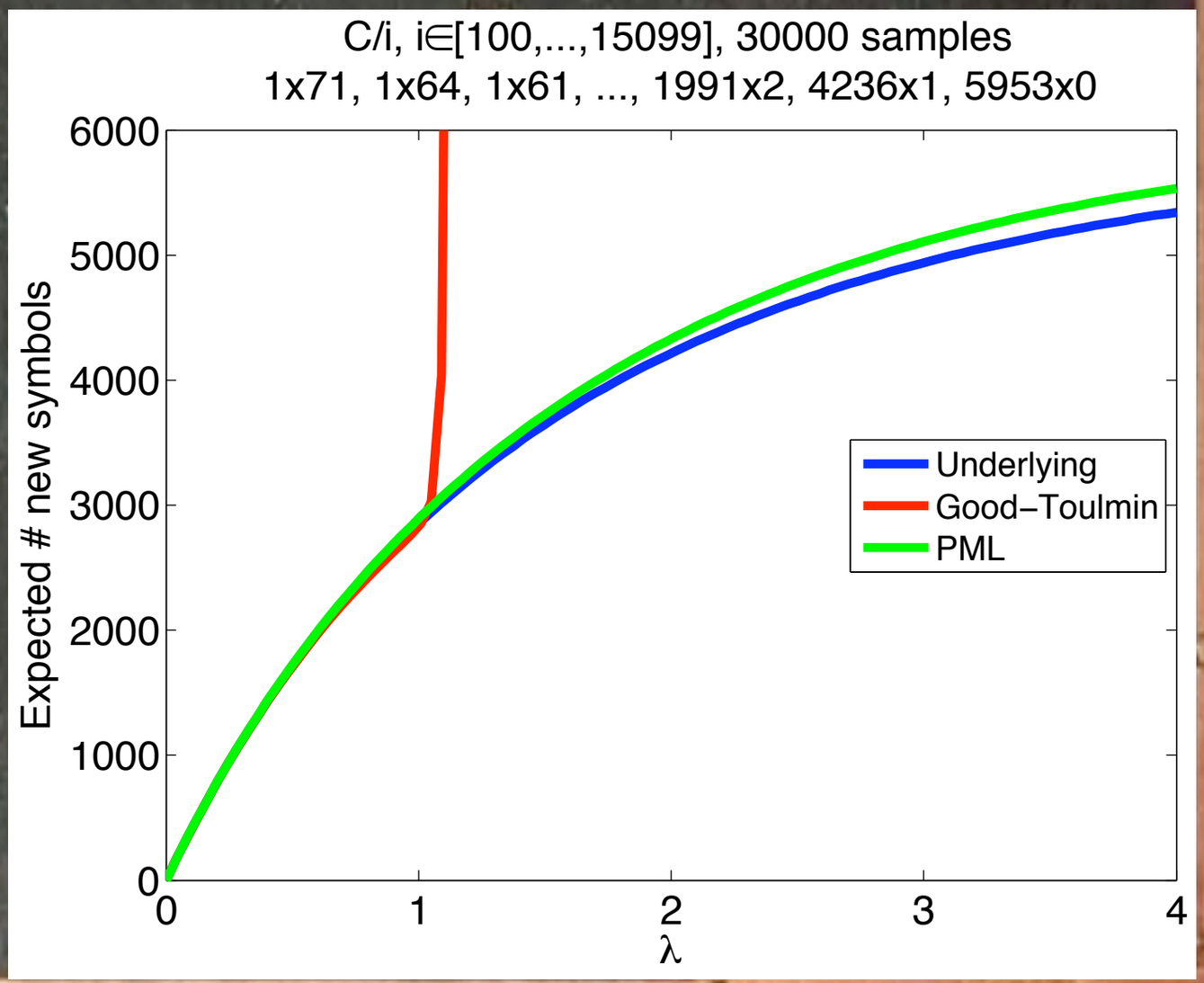
Good-Toulmin: $\lambda < 1$

$\lambda > 1$

Estimate PML & predict

Extends to $\lambda > 1$

Applies to other predictions



Probability of Specific Elements

Can Patterns Work?

US, Slovenia, US, US -> 1, 2, 1, 1

Lost connection to original?

No

1 - US, 2 - Slovenia

After 1, 2, 1, 1, estimate probabilities of 1, 2, 3

Probabilities of observed symbols & NEW

Text classification

Train: Documents pre-classified into topics

Business, technology, politics, sports,...

Test: New document

Find topic

Existing methods: Many

Nearest Neighbor, Boosting, ...

Support vector machines (SVM)

Current approach

Plain probability estimation

Comparison

Methods: Nearest neighbor, Laplace, SVM, Patterns

Data sets: Newsgroups, Reuters R52, CADE

Platform: Rainbow (CMU, UMass)

	R52	Newsgroups	CADE
Laplace	88.43	80.92	57.27
Nrst nbr	83.22	75.93	51.20
SVM	93.57	82.84	52.84
Patterns	91.98	82.92	59.24

1 2 3 4 5 6 7 8 9

Thank You!