Perplexity on Reduced Corpora

— Analysis of Cutoff by Power Law

Hayato Kobayashi

Yahoo Japan Corporation

Cutoff

- Removing low-frequency words from a corpus
- Common practice to save computational costs in learning

- Language modeling
  - Needed even in a distributed environment, since the feature space of k-grams is quite large [Brants+ 2007]
- Topic modeling
  - Enough for roughly analyzing topics, since low-frequency words have a small impact on the statistics [Steyvers&Griffiths 2007]

Question

- How many low-frequency words can we remove while maintaining sufficient performance?
  - More generally, how much can we reduce a corpus/model using a certain strategy?

- Many experimental studies addressing the question
  - [Stoleke 1998], [Buchsbaum+ 1998], [Goodman&Gao 2000], [Gao&Zhang 2002], [Ha+ 2006], [Hirsimaki 2007], [Church+ 2007]
  - Discussing trade-off relationships between the size of reduced corpus/model and its performance
- No theoretical study!

This work

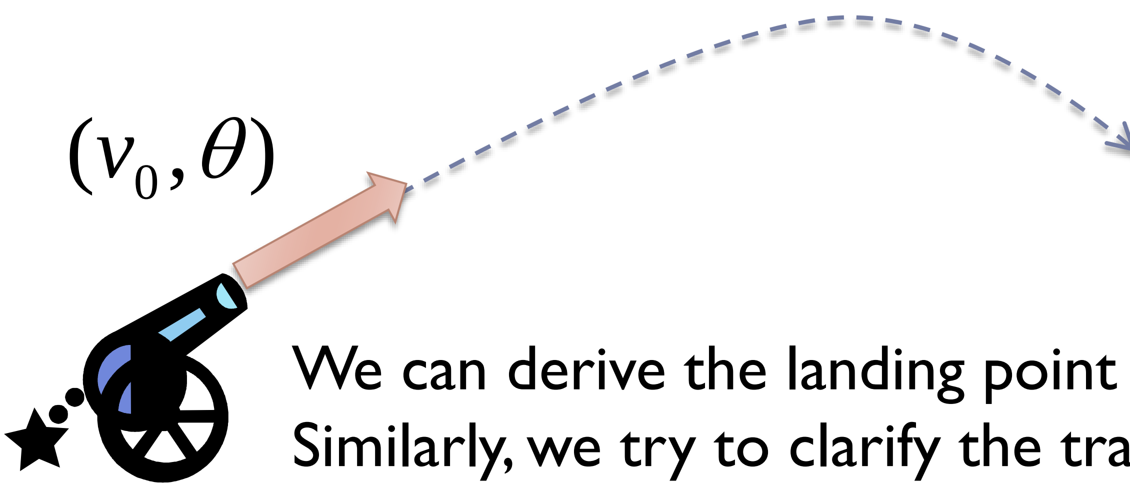
- First address the question from a theoretical standpoint

- Derive the trade-off formulae of the cutoff strategy for k-gram models and topic models
  - Perplexity vs. reduced vocabulary size
- Verify the correctness of our theory on synthetic corpora and examine the gap between theory and practice on several real corpora

Approach

- Assume a corpus follows Zipf's law (power law)
  - Empirical rule representing a long-tail property in a corpus
- Essentially the same approach as in physics
  - Constructing a theory while believing experimentally observed results (e.g., gravity acceleration g)

$$(v_0, \theta)$$



$$\frac{v_0^{\ 2}\sin(2\theta)}{g}$$

We can derive the landing point of a ball by believing g. Similarly, we try to clarify the trade-off relationships by believing Zipf's law.

Outline

- **Preliminaries**
  - Zipf's law
  - Perplexity (PP)
  - Cutoff and restoring
- PP of unigram models
- PP of k-gram models
- PP of topic models
- Conclusion

Zipf's law

- Empirical rule discovered on real corpora [Zipf, 1935]
  - Word frequency f(w) is inversely proportional to its frequency ranking r(w)

Max. frequency

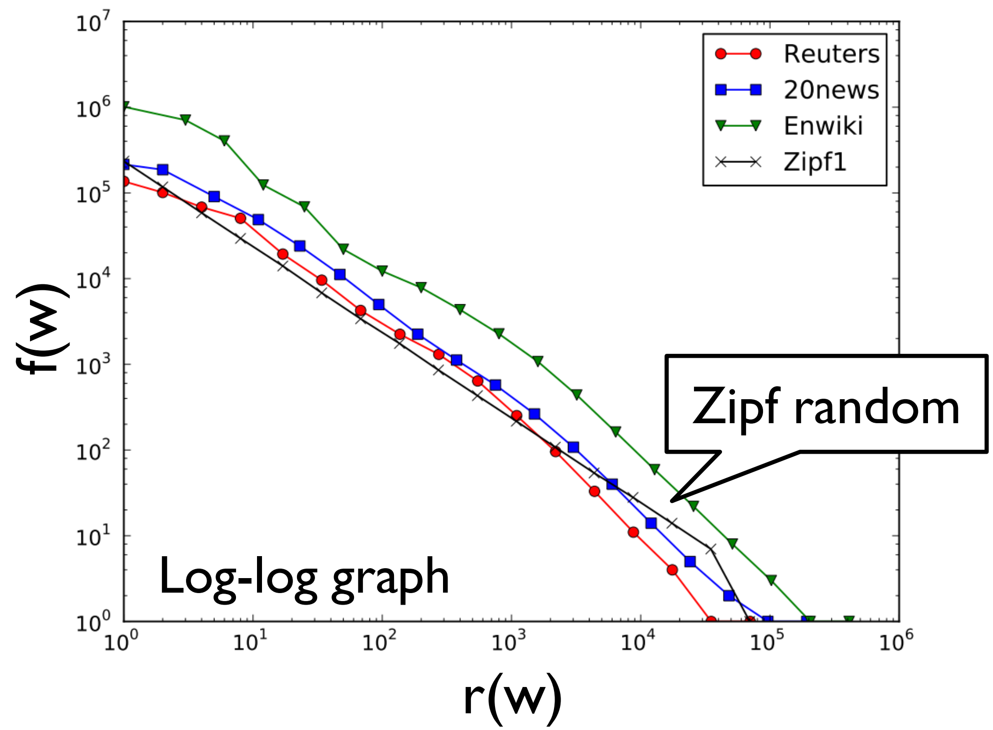
$$f(w) = \frac{C}{r(w)}$$

Frequency

Frequency ranking

(Linear on a log-log graph)

Real corpora roughly follow Zipf's law



Perplexity (PP)

- Widely used evaluation measure of statistical models
  - Geometric mean of the inverse of the per-word likelihood on the held-out test corpus

$$PP := \left(\prod_{w \in \mathbf{w}_{\mathcal{T}}} \frac{1}{p(w)} \right)^{\frac{1}{N_{\mathcal{T}}}}$$

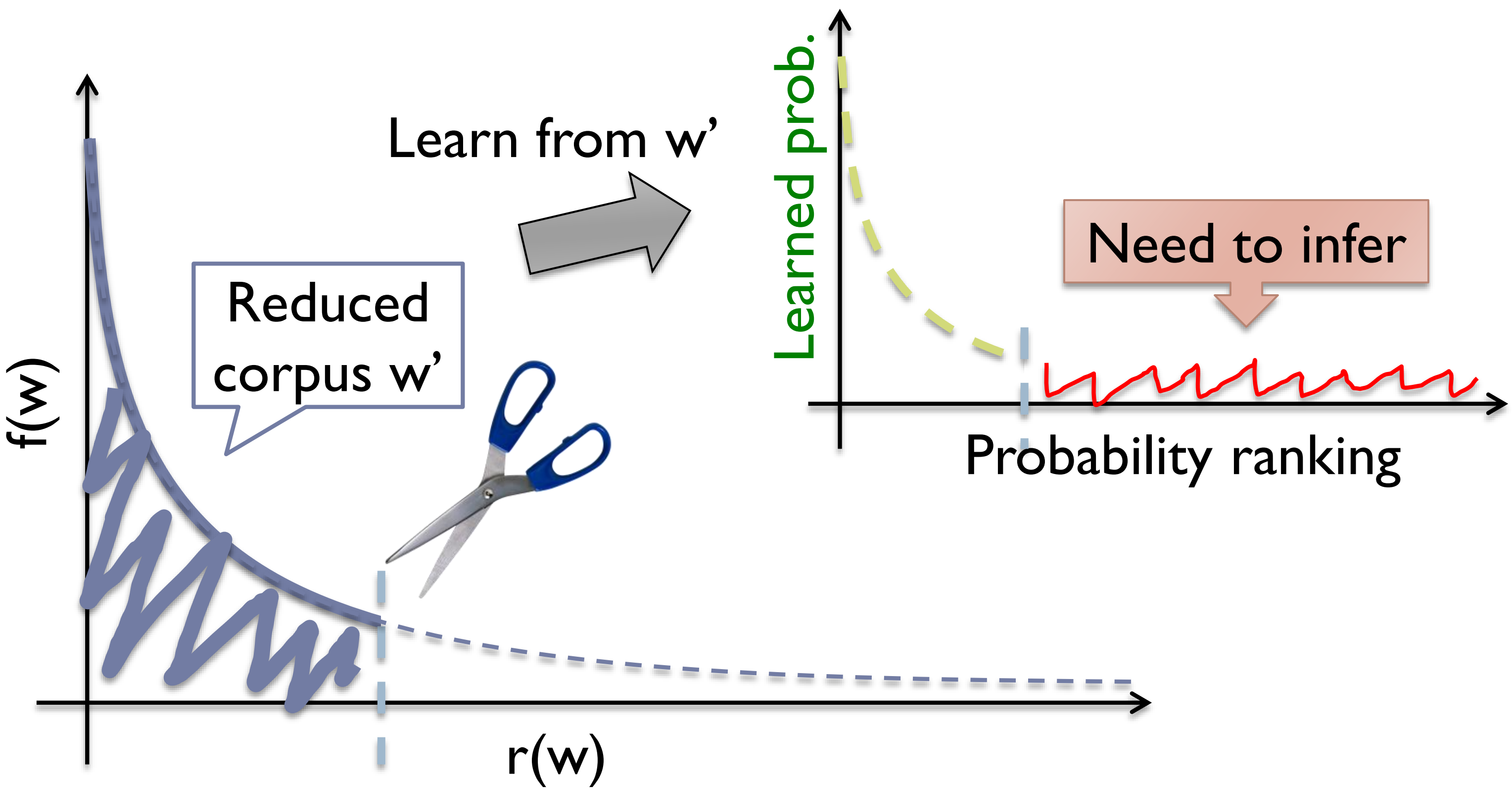
Corpus size

Test corpus

  - PP means how many possibilities one has for estimating the next word
    - Lower perplexity means better generalization performance

Cutoff

- Removing low frequency words
  - f(remaining word) ≥ f(removed word) holds



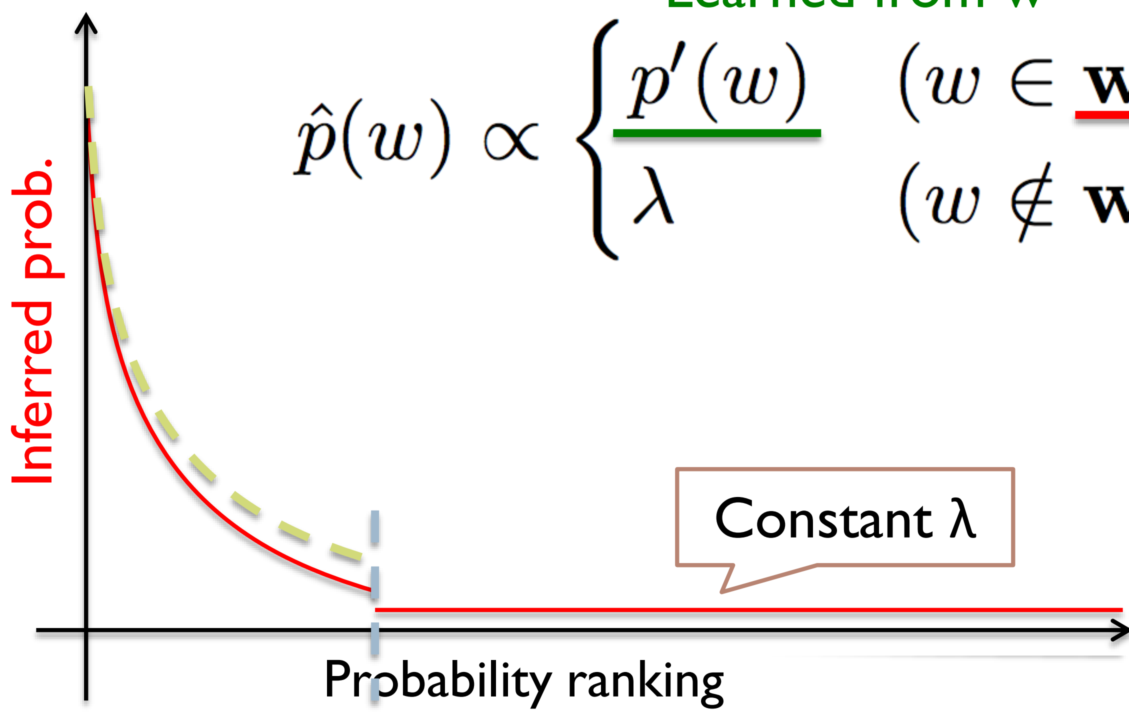
Constant restoring

- Infer the prob. of the removed words as a constant
  - Approximate the result learned from the original corpus

Learned from w'

$$\hat{p}(w) \propto \begin{cases} p'(w) & (w \in \mathbf{w}') \\ \lambda & (w \notin \mathbf{w}') \end{cases}$$

Reduced corpus



Outline

- Preliminaries
  - Zipf's law
  - Perplexity (PP)
  - Cutoff and restoring
- **PP of unigram models**
- PP of k-gram models
- PP of topic models
- Conclusion

Perplexity of unigram models

- Predictive distribution of unigram models

$$p'(w') = \frac{f(w')}{\underline{N'}}$$

Reduced corpus size

- Optimal restoring constant
  - Obtained by minimizing PP w.r.t. a constant λ, after substituting the restored probability $\hat{p}(w)$ into PP

Corpus size

$$\lambda^* = \frac{\underline{N} - N'}{(\underline{W} - \underline{W'})N'}$$

Vocab. size   Reduced vocab. size

Theorem (PP of unigram models)

- For any reduced vocabulary size W', the perplexity PP_1 of the optimal restored distribution of a unigram model is calculated as

$$\hat{PP}_1(W') = H(W) \exp\left(\frac{B(W')}{H(W)}\right) \left(\frac{W - W'}{H(W) - H(W')}\right)^{1-\frac{H(W')}{H(W)}}$$

$H(X) := \sum_{x=1}^{X} \frac{1}{x}$ Harmonic series

$B(X) := \sum_{x=1}^{X} \frac{\ln(x)}{x}$ Bertrand series (special form)

Approximation of PP of unigrams

- H(X) and B(X) can be approximated by definite integrals

$$H(X) \approx \ln X + \underline{\gamma}$$ Euler-Mascheroni const.

$$B(X) \approx \frac{1}{2} \ln^2 X$$

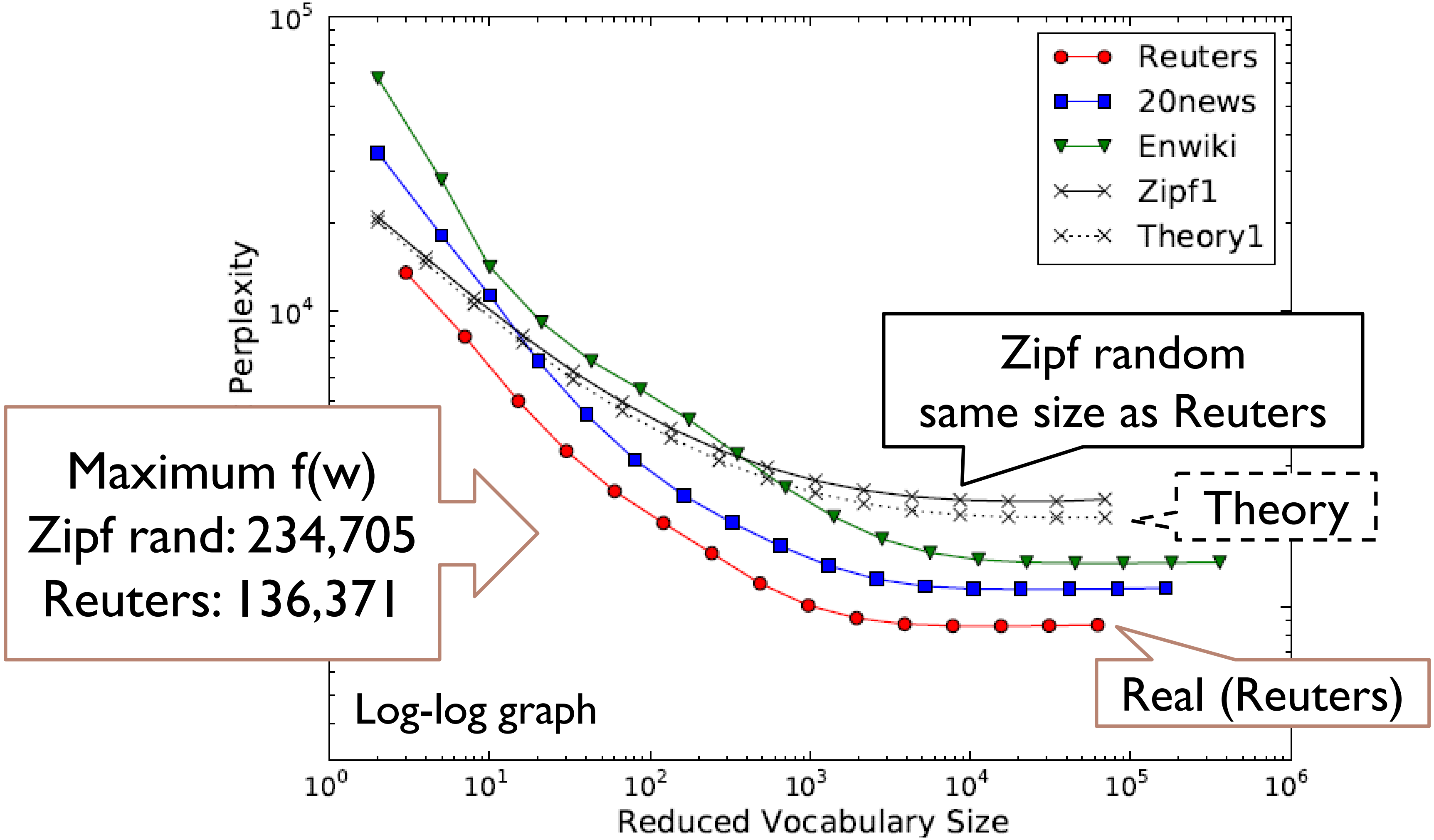
- Approximate formula $\tilde{PP}_1(W')$ is obtained as

Reduced vocab. size

$$\tilde{PP}_1(W') = \sqrt{W} \ln W \exp \frac{(\ln \underline{W'} - \ln W)^2}{2 \ln W}$$

- $\tilde{PP}_1(W')$ is quasi polynomial (quadratic)
  - Behaves as a quadratic function on a log-log graph

PP of unigrams vs. reduced vocab. size



Our theory is suited for inferring the growth rate of perplexity rather than the perplexity value itself

Outline

- Preliminaries
  - Zipf's law
  - Perplexity (PP)
  - Cutoff and restoring
- PP of unigram models
- **PP of k-gram models**
- PP of topic models
- Conclusion

Perplexity of k-gram models

- Simple model where k-grams are calculated from a random word sequence based on Zipf's law
- The model is "stupid"
  - Bigram "is is" is quite frequent

$$p(\text{"is is"}) = p(\text{"}is\text{"})\,p(\text{"}is\text{"})$$

  - Two bigrams "is a" and "a is" have the same frequency

$$p(\text{"is a"}) = p(\text{"}is\text{"})\,p(\text{"}a\text{"}) = p(\text{"a is"})$$

- Later experiment will uncover the fact that the model can roughly capture the behavior of real corpora

Frequency of a k-gram

- Frequency f_k of a k-gram w_k is defined by

$$f_k(w_k) = \frac{C_k}{g_k(r_k(w_k))}$$

Decay function

- Decay function g_2 of bigrams is as follows

$$\begin{aligned}(g_2(i))_i &:= (g_2(1),\ g_2(2),\ g_2(3), \cdots) \\ &= (1\cdot 1,\ 1\cdot 2,\ 2\cdot 1,\ 1\cdot 3,\ 3\cdot 1, \cdots) \\ &= (1,\ 2,\ 2,\ 3,\ 3,\ 4,\ 4,\ 4,\ 5,\ 5,\ 6, \cdots)\end{aligned}$$

- Decay function g_k of k-grams is defined through its inverse: $g_k^{-1}(\ell) := \sum_{n=1}^{\ell} d_k(n)$

$$d_k(n) := \sum_{i_1 \cdot i_2 \cdots i_k = n} 1$$

Piltz divisor function that represents # of divisors of n

Exponent of k-gram distributions

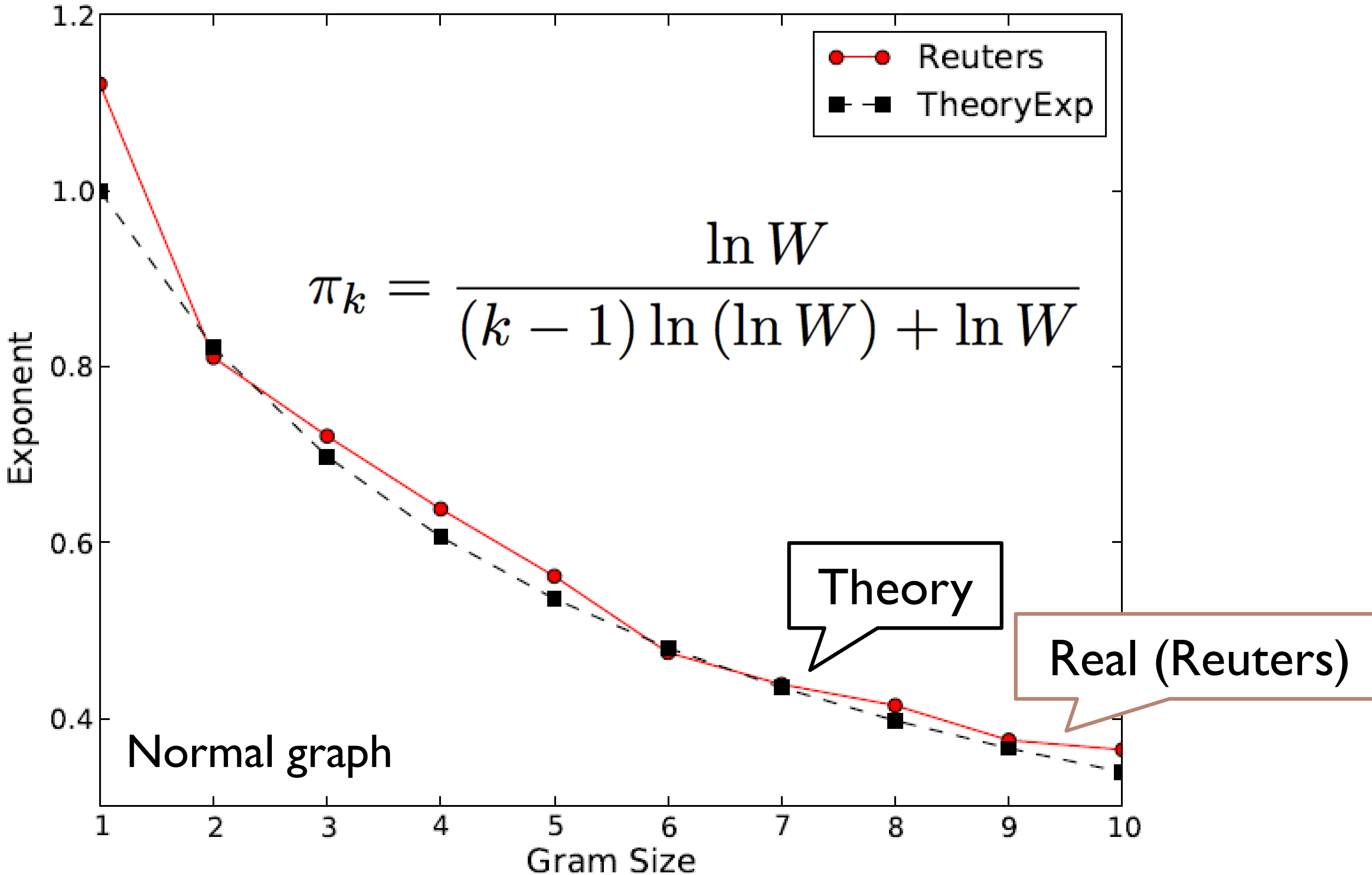
- Assume k-gram frequencies follow a power law
  - [Ha+ 2006] found k-gram frequencies roughly follow a power law, whose exponent π_k is smaller than 1 (k>1)

$$f_k(w_k) \propto r_k(w_k)^{-\pi_k}$$

- Optimal exponent in our model based on the assumption
  - By minimizing the sum of squared errors between the inverse gradients $g_k^{-1}(r)$ and r^{1/π_k} on a log-log graph

$$\pi_k = \frac{\ln W}{(k-1)\ln(\ln W) + \ln W}$$

Exponent of k-grams vs. gram size



$$\pi_k = \frac{\ln W}{(k-1)\ln(\ln W) + \ln W}$$

Corollary (PP of k-gram models)

- For any reduced vocabulary size W', the perplexity of the optimal restored distribution of a k-gram model is calculated as

$$\hat{PP}_k(W') = H_{\pi_k}(W) \exp\left(\frac{B_{\pi_k}(W')}{H_{\pi_k}(W)}\right) \left(\frac{W - W'}{H_{\pi_k}(W) - H_{\pi_k}(W')}\right)^{1-\frac{H_{\pi_k}(W')}{H_{\pi_k}(W)}}$$

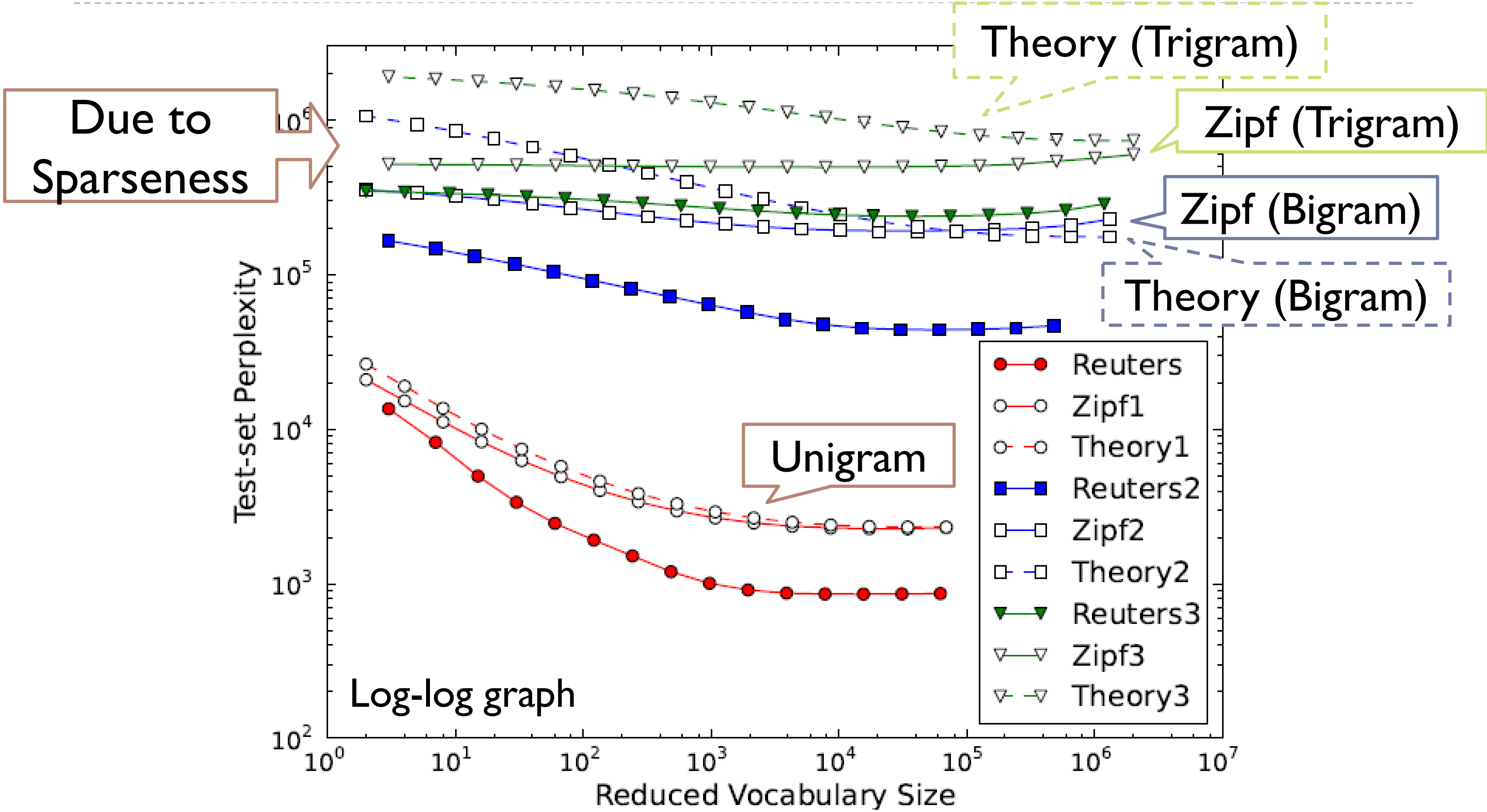
$$H_a(X) := \sum_{x=1}^{X} \frac{1}{x^a}$$

Hyper harmonic series

$$B_a(X) := \sum_{x=1}^{X} \frac{a \ln x}{x^a}$$

Bertrand series (another special form)

PP of k-grams vs. reduced vocab. size



We need to make assumptions that include backoff and smoothing for higher order k-grams

Additional properties by power-law

- Treat as a variant of the coupon collector's problem
  - How many trials are needed for collecting all coupons whose occurrence probabilities follow some stable distribution
  - There exists several works about power law distributions
- Corpus size for collecting all of the k-grams, according to [Boneh&Papanicolaou 1996]
  - When $\pi_k = 1$, $W \ln^2 W$, otherwise, $\dfrac{kW^k}{1-\pi_k}$
- Lower and upper bound of the number of k-grams from the corpus size N and vocab. size W, according to [Atsonios+ 2011]

$$(\pi_k + 1)\left(1 - e^{-\frac{(1-\pi_k)N}{W^k - 1} - \ln \frac{W^k - 1}{W^k}}\right) \leq \tilde{W}_k \leq \frac{\pi_k}{\pi_k - 1}\left(\frac{N}{H_{\pi_k}(W^k)}\right)^{\frac{1}{\pi}} - \frac{N}{(\pi_k - 1)H_{\pi_k}(W^k)} W^{1-\pi_k}$$

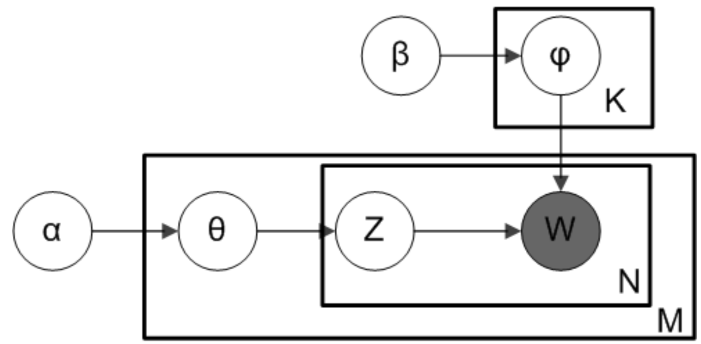
Outline

- Preliminaries
  - Zipf's law
  - Perplexity (PP)
  - Cutoff and restoring
- PP of unigram models
- PP of k-gram models
- **PP of topic models**
- Conclusion

Perplexity of topic models

- Latent Dirichlet Allocation (LDA) [Blei+ 2003]

$$
\begin{aligned}
\theta_{d_i} &\sim Dirichlet(\alpha) \\
z_i|\theta_{d_i} &\sim Multi(\theta_{d_i}) \\
\phi_{z_i} &\sim Dirichlet(\beta) \\
w_i|z_i,\phi_{z_i} &\sim Multi(\phi_{z_i}),
\end{aligned}
$$



[Griffiths&Steyvers 2004]

- Learning with Gibbs sampling
  - Obtain a "good" topic assignment z_i for each word w_i
- Posterior distributions of two hidden parameters

$$\hat{\theta}_d(z) \propto n_z^{(d)} + \alpha$$

Document-topic distribution
Mixture rate of topic z in document d

$$\hat{\varphi}_z(w) \propto n_z^{(w)} + \beta$$

Topic-word distribution
Occurrence rate of word w in topic z

Rough assumptions of ϕ and θ

- Assumption of ϕ
  - Word distribution ϕ_z of each topic z follows Zipf's law

It is natural, regarding each topic as a corpus

- Assumptions of θ (two extreme cases)
  - Case All: Each document evenly has all topics
  - Case One: Each document only has one topic (uniform dist.)

The curve of actual perplexity is expected to be between their values

- Case All: PP of a topic model ≈ PP of a unigram
  - Marginal predictive distribution is independent of d

$$\sum_{z=1}^{T} \underline{\hat{\theta}_d(z)}\hat{\phi}_z(w) \propto \sum_{z=1}^{T} \frac{n_z^{(w)}+\beta}{T} \underset{\sim}{\propto} f(w)$$

=1/T

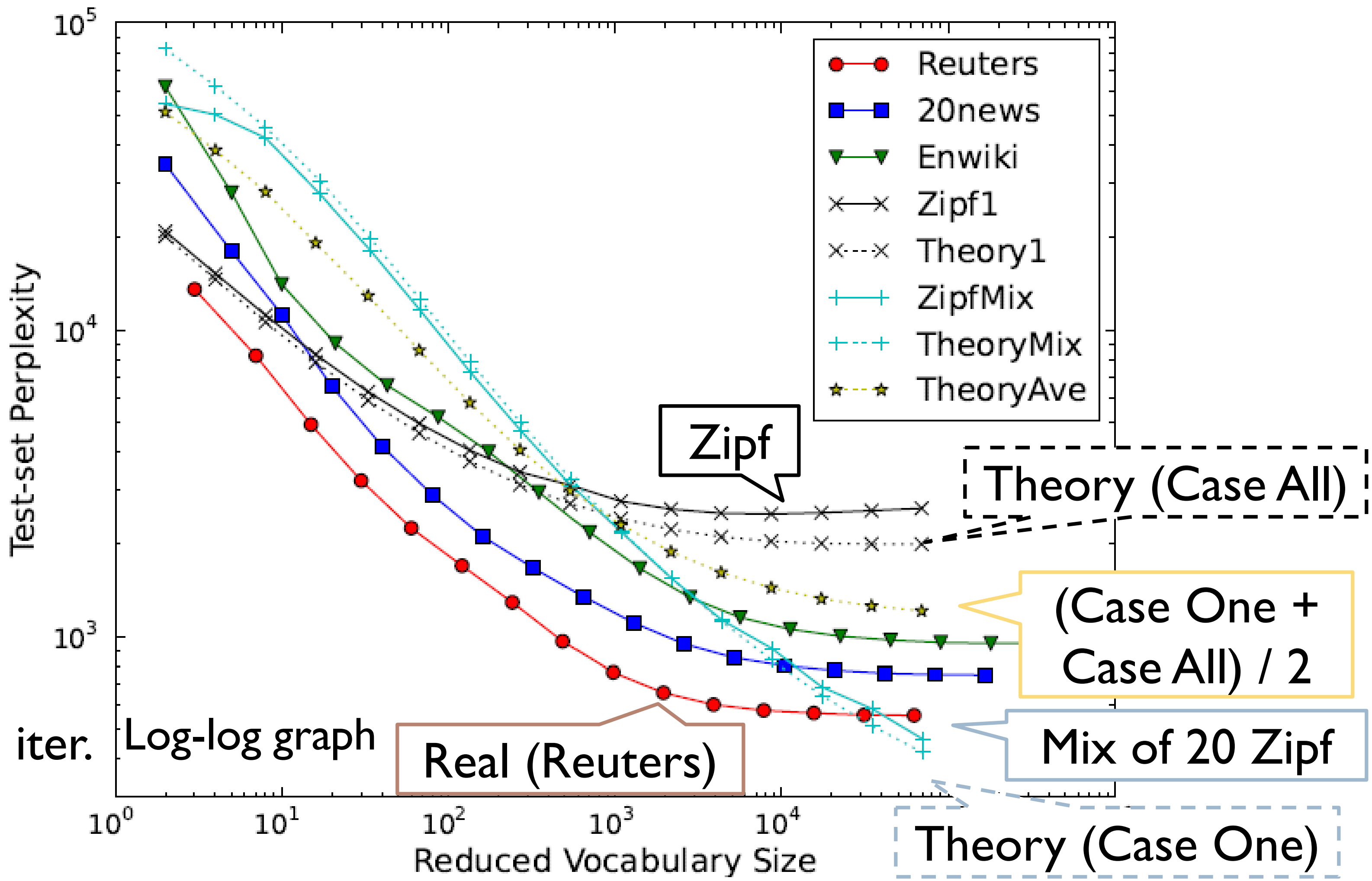
Theorem(PP of LDA models: Case One)

- For any reduced vocabulary size W', the perplexity of the optimal restored distribution of a topic model in the Case One is calculated as

$$\hat{PP}_{Mix}(W') = H(W/T)\exp\left(\frac{B(W'/T)}{H(W/T)}\right)\left(\frac{W - W'}{H(W/T) - H(W'/T)}\right)^{1-\frac{H(W'/T)}{H(W/T)}}$$

T : # of topics in LDA

PP of LDA models vs. reduced vocab. size



T=20
CGS w/ 100 iter.
α=β=0.1

Time, memory, and PP of LDA learning

- Results of Reuters corpus

corpus	time	memory	perplexity
original	4m3.80s	71,548KB	500
(1/10)	3m55.70s	46,648KB	550
(1/20)	3m42.63s	34,024KB	611

- Memory usage of the (1/10)-corpus is only 60% of that of the original corpus
  - Helps in-memory computing for a larger corpus, although the computational time decreased a little

Outline

- Preliminaries
  - Zipf's law
  - Perplexity (PP)
  - Cutoff and restoring
- PP of unigram models
- PP of k-gram models
- PP of topic models
- **Conclusion**

Conclusion

- Trade-off formulae of the cutoff strategy for k-gram models and topic models based on Zipf'law
  - Perplexity vs. reduced vocabulary size
- Experiments on real corpora showed that the estimation of the perplexity growth rate is reasonable
- We can get the best cutoff parameter by maximizing the reduction rate ensuring an acceptable (relative) perplexity
- Possibility that we can theoretically derive empirical parameters, or "rules of thumb", for different NLP problems

Can we derive other "rules of thumb" based on Zipf's law?

Thank you