Statistical Evaluation of Generative Models with MAUVE Scores

Krishna Pillutla*, Lang Liu*, John Thickstun, Sean Welleck, Swabha Swayamdipta,
Rowan Zellers, Sewoong Oh, Yejin Choi, Zaid Harchaoui

Google Research · UNIVERSITY of WASHINGTON · AI2 · USC

NEURAL INFORMATION PROCESSING SYSTEMS · JMLR

Evaluating Generative Models

Divergence(Model distribution ‖ Target distr.)

- Divergence frontiers [Djolonga et al. AISTATS '20]
- **MAUVE**: evaluate open text generation [**P.** et al. NeurIPS '21]

This work: estimate metrics from samples

- Statistical bounds ↔ Empirical performance

Motivation

>> prompt: What is mauve?

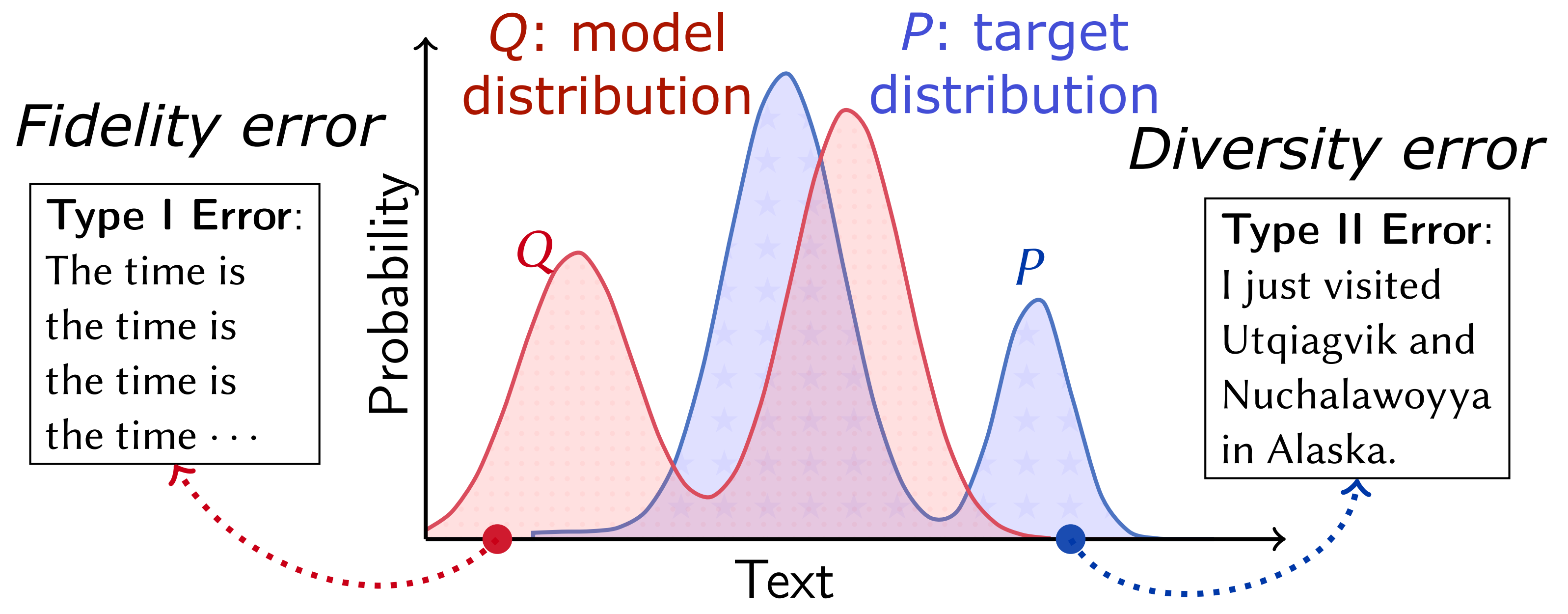
Bard: Mauve is a pale purple color named after the mallow flower (French: mauve). It is a combination of red and blue with a hint of gray.

or

How good are these generative models?

Challenge: multiple correct responses

f–Divergence Frontiers



Fidelity error

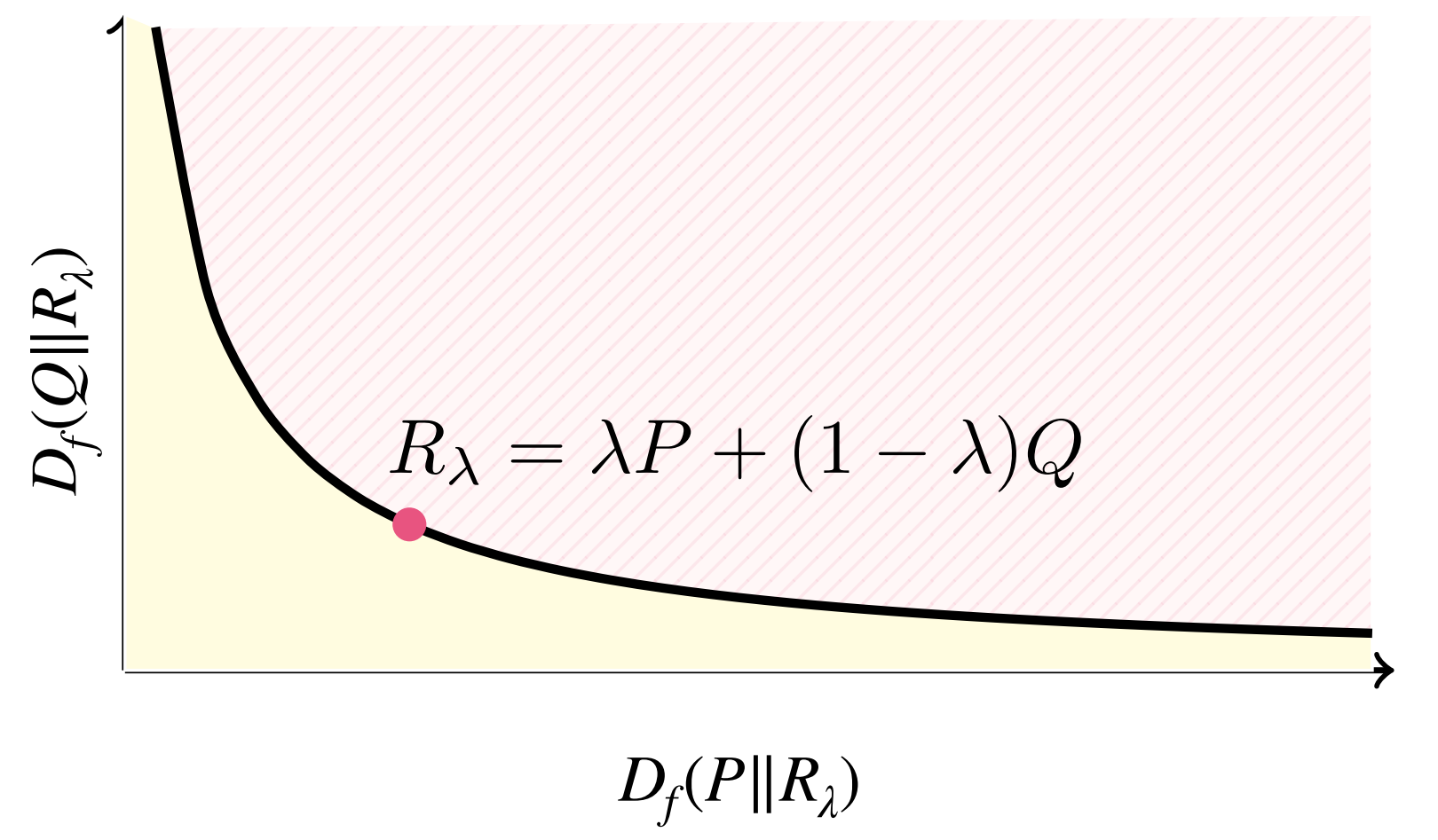
Type I Error: The time is the time is the time is the time · · ·

Diversity error

Type II Error: I just visited Utqiagvik and Nuchalawoyya in Alaska.

Type I Error $= D_f(Q|P)$ Type II Error $= D_f(P|Q)$

Softly measure both errors with f-divergences



Generalization of Renyi frontiers [Djolonga et al. AISTATS '20] and KL frontiers [**P.** et al. NeurIPS '21]

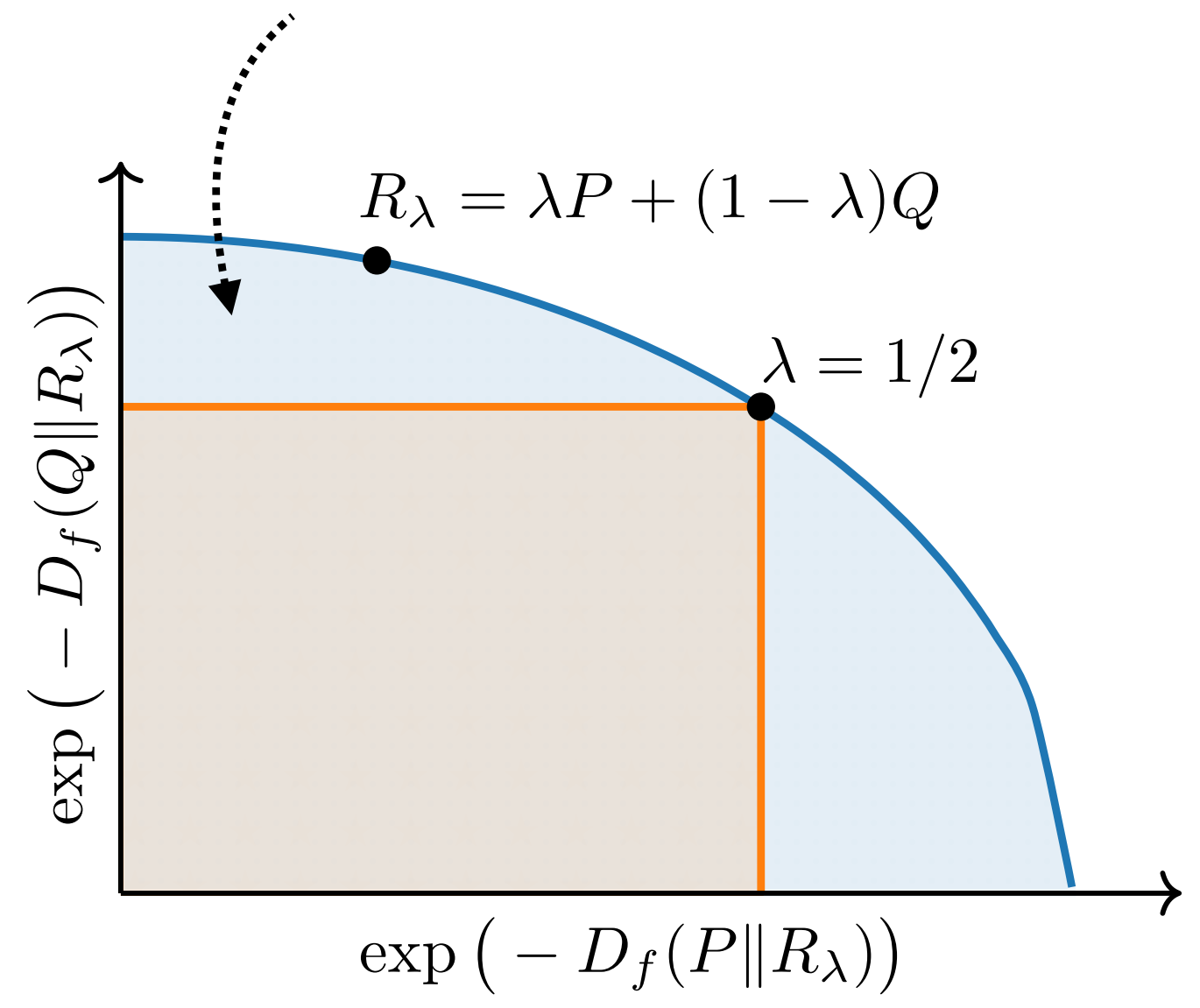
Scalar summaries of the frontiers

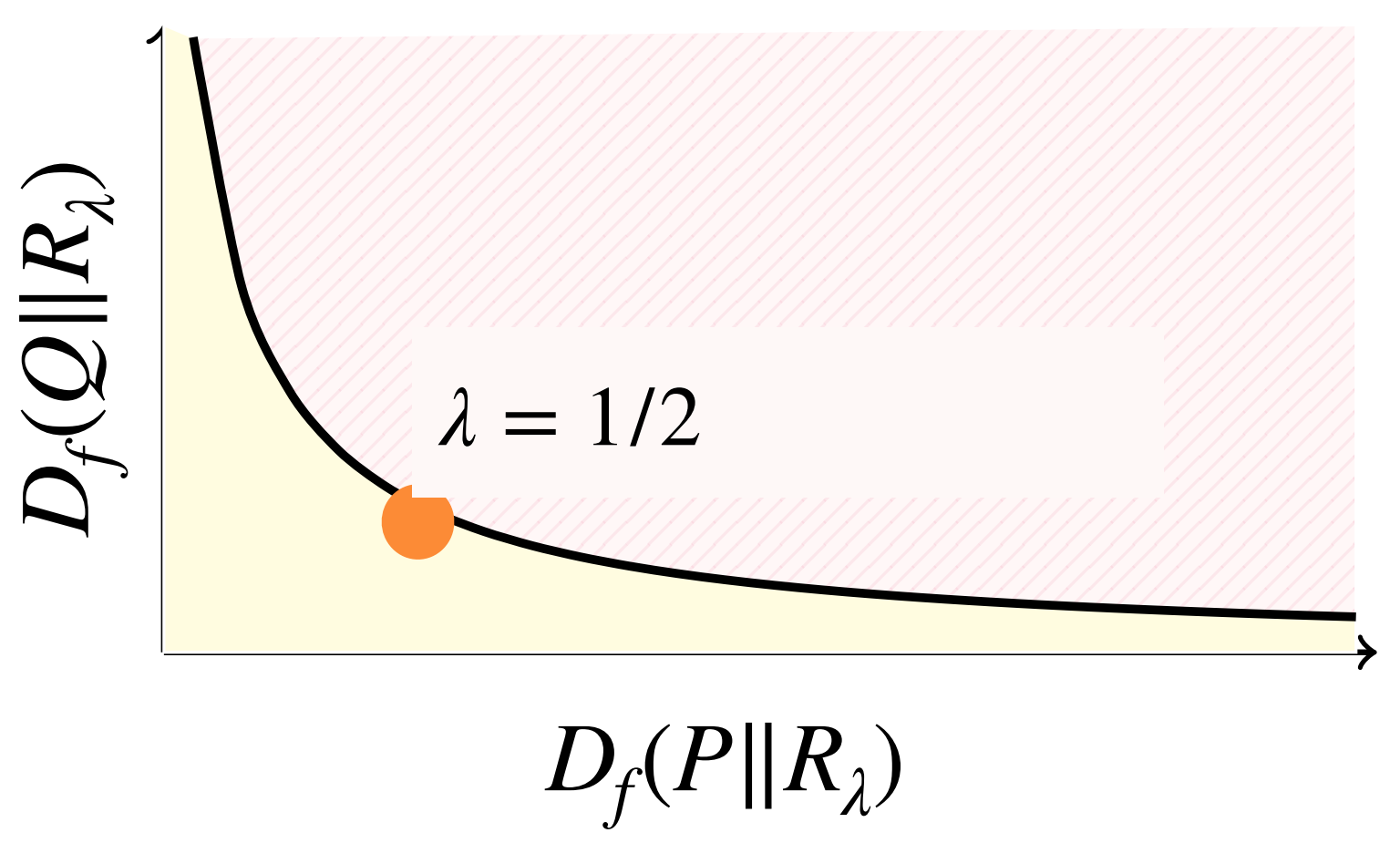
Area summary [**P.** et al. NeurIPS '21]

Mid-point summary KL → Jensen-Shannon

Integral summary [Liu, **P.** et al. NeurIPS '21]

$$\int_0^1 \Big(\lambda D(P\|R_\lambda) + (1-\lambda) D(Q\|R_\lambda)\Big) d\lambda$$

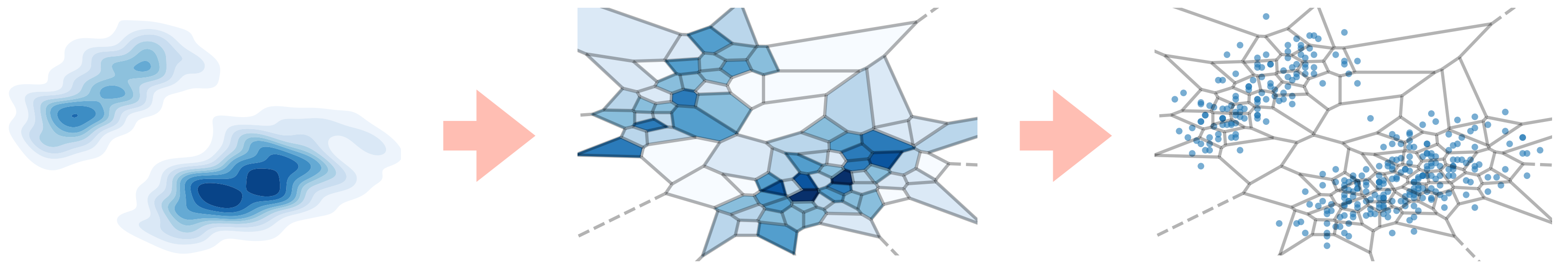




Estimation with Vector Quantization

Standard estimation procedure [Sajjadi et al. 2018, **P.** et al. 2021]

Distribution P → Quantized distribution $P_{\mathcal{S}_k}$ → Empirical distribution $\hat{P}_{\mathcal{S}_k,n}$



1. How to select the quantization size k?
2. Can we do better than the empirical estimator?
3. What is the sample complexity?

Return Mauve$(\hat{P}_{\mathcal{S}_k,n}, \hat{Q}_{\mathcal{S}_k,n})$

Estimation error bounds

Statistical error: For discrete P, Q with support size k

$$\mathbb{E}\,|D_f(\hat{P}_n\|\hat{Q}_n) - D_f(P\|Q)| \lesssim \sqrt{\frac{k}{n}}$$

Total error: For any P, Q and k, there exists a partitioning $\mathcal{S}_k$ such that

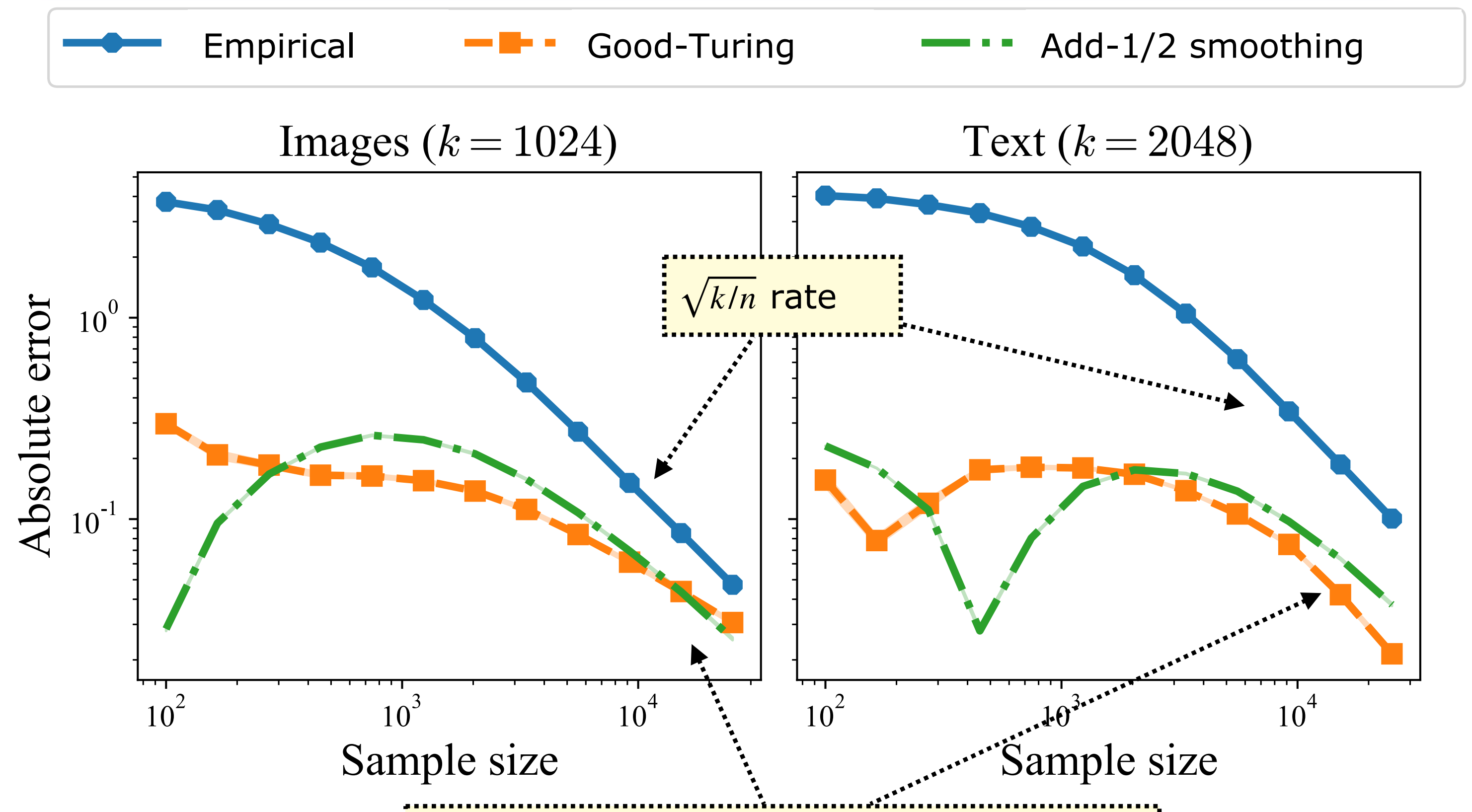
$$\mathbb{E}\left|D_f(\hat{P}_{\mathcal{S}_k,n}\|\hat{Q}_{\mathcal{S}_k,n}) - D_f(P\|Q)\right| \lesssim \sqrt{\frac{k}{n}} + \frac{1}{k}$$

Smoothing: For the add-b estimator $\hat{P}_{\mathcal{S}_k,n,b}$ of P

$$\mathbb{E}\left|D_f(\hat{P}_{\mathcal{S}_k,n,b}\|\hat{Q}_{\mathcal{S}_k,n,b}) - D_f(P\|Q)\right| \lesssim \frac{\sqrt{nk}+bk}{n+bk} + \frac{1}{k}$$

Quantization error

Empirical behavior of statistical error



$\sqrt{k/n}$ rate

Smoothing helps especially at small n

Other estimation methods

Non-parametric: estimate $P(x)/Q(x)$ using k-NN
Rate $= (k/n)^{1/d} + 1/k$ [Noshad et al. ISIT 2017]

Classifier: estimate $P(x)/Q(x)$ w/ logistic regression

Parametric: Approximate P, Q w/ Gaussians

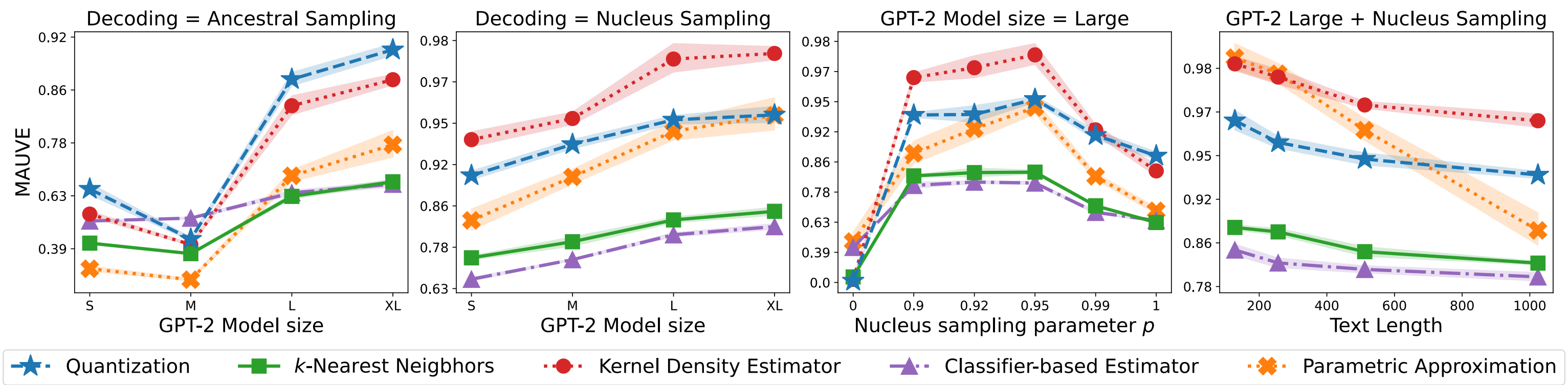
Result: All estimation methods work in practice

- Parametric is non-robust to hyperparams

Larger models are better

Quantify benefits of truncation heuristics

Shorter generations are better



References

Djolong, Lucic, Cuturi, Bachem, Bousquet, Gelly. AISTATS 2020.
Precision-Recall Curves Using Information Divergence Frontiers.

P., Swayamdipta, Zellers, Thickstun, Welleck, Choi, Harchaoui.
NeurIPS 2021 (*Outstanding Paper Award*).
MAUVE: Measuring the Gap Between Neural Text and Human Text.

Liu, **P.**, Welleck, Oh, Choi, Harchaoui. NeurIPS 2021.
Divergence Frontiers for Generative Models

P.*, Liu*, Thickstun, Welleck, Swayamdipta, Zellers, Oh, Choi, Harchaoui.
JMLR 2023.
MAUVE Scores for Generative Models: Theory and Practice.

Software

SCAN ME

www krishnap25.github.io KrishnaPillutla