Harvard University

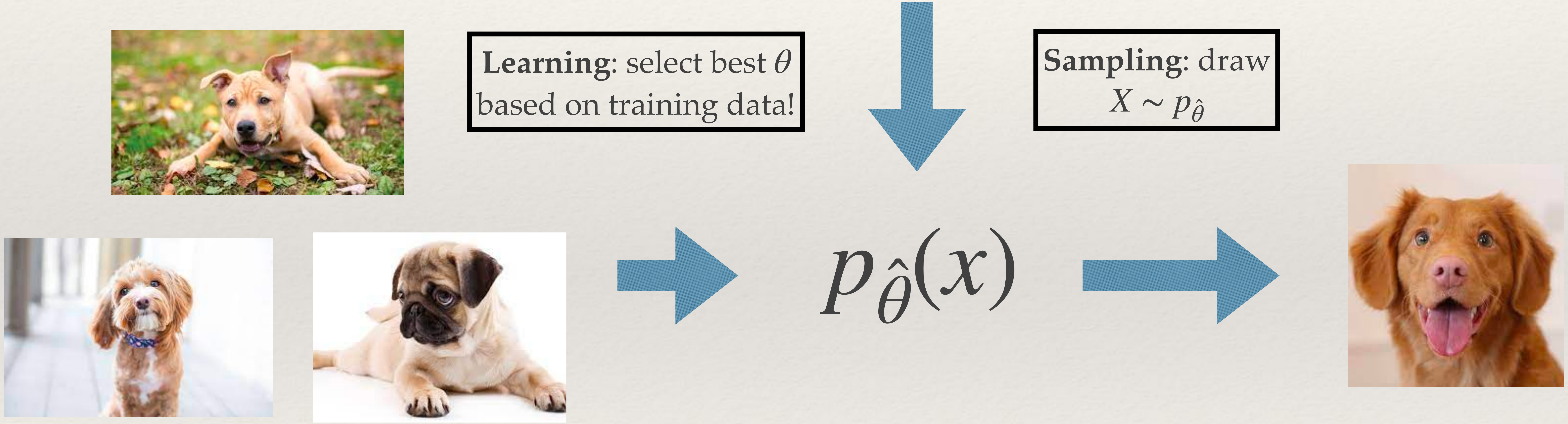
Score Matching, Efficiency, and Isoperimetry

Frederic Koehler (Stanford -> UChicago)

based on joint works with
Alexander Heckett (CMU)
Andrej Risteski (CMU)
Thuy-Duong (June) Vuong (Stanford)

Learning Distributions from Samples

$\{p_\theta : \theta \in \Theta\}$ Class of models (probability distributions) with parameter θ



Learning: select best θ based on training data!

Sampling: draw $X \sim p_{\hat{\theta}}$

$$p_{\hat{\theta}}(x)$$

Training Data $x_1, x_2, \ldots, x_n$
(iid from ground truth p^*)

Model w/ parameter $\hat{\theta}$
selected using data

Output of Model
(hopefully matches true distribution!)

Energy-Based Models

In this talk, we will mostly be interested in learning “energy-based models”

$$p_\theta(x) = \frac{1}{Z_\theta} \exp\left(f_\theta(x)\right) \qquad Z_\theta = \int \exp(f_\theta(x))dx$$

Why? Hard to find models with closed-form likelihood, these are almost as good.

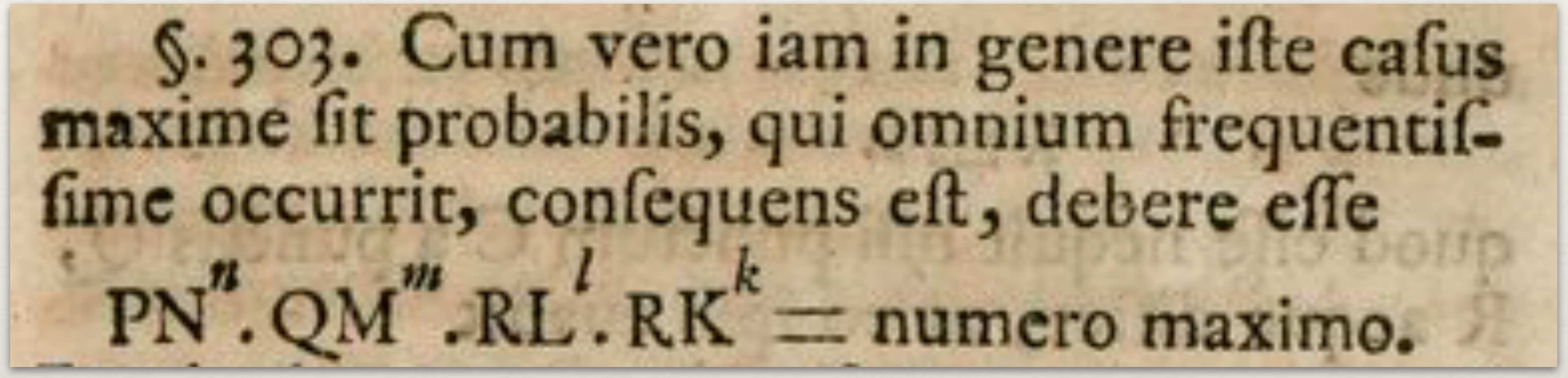
Recent survey: *How to Train Your Energy-Based Models* [Song-Kingma ’21]

How to learn the best model?

Maximum Likelihood Estimation (MLE)

“pick the model which maximizes the probability of the data”

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log p_\theta(x_i)$$



§. 303. Cum vero iam in genere iſte caſus maxime ſit probabilis, qui omnium frequentiſſime occurrit, conſequens eſt, debere eſſe

$PN^n . QM^m . RL^l . RK^k =$ numero maximo.

Lambert 1760

The good: asymptotically optimal (in sample-efficiency) as $n \to \infty$!

The bad: tricky to compute — involves sampling / partition function Z_θ.

An Alternative: Score Matching

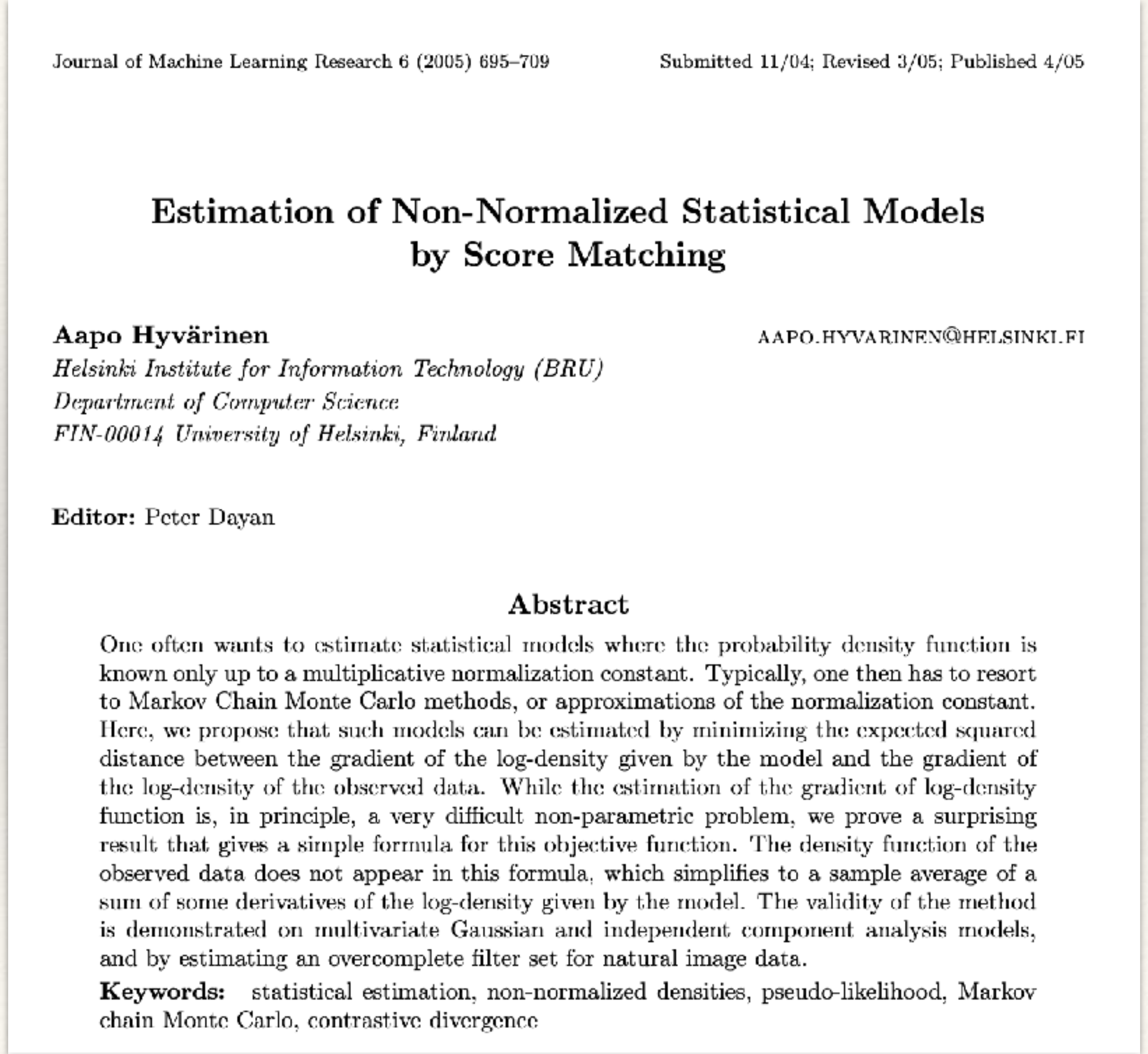
Try to fit **gradients** of a distribution, i.e. minimize

$$\sum_{i=1}^{n} \|\nabla_x \log p_\theta(x_i) - \nabla_x \log p^*(x_i)\|^2$$

Only makes sense for smooth densities.

No access to $\nabla_x \log p^*(x)$, so use integration by parts trick!

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \Delta \log p_\theta(x_i) + \frac{1}{2}\|\nabla \log p_\theta(x_i)\|^2$$



Journal of Machine Learning Research 6 (2005) 695–709 Submitted 11/04; Revised 3/05; Published 4/05

Estimation of Non-Normalized Statistical Models by Score Matching

Aapo Hyvärinen AAPO.HYVARINEN@HELSINKI.FI
Helsinki Institute for Information Technology (BRU)
Department of Computer Science
FIN-00014 University of Helsinki, Finland

Editor: Peter Dayan

Abstract

One often wants to estimate statistical models where the probability density function is known only up to a multiplicative normalization constant. Typically, one then has to resort to Markov Chain Monte Carlo methods, or approximations of the normalization constant. Here, we propose that such models can be estimated by minimizing the expected squared distance between the gradient of the log-density given by the model and the gradient of the log-density of the observed data. While the estimation of the gradient of log-density function is, in principle, a very difficult non-parametric problem, we prove a surprising result that gives a simple formula for this objective function. The density function of the observed data does not appear in this formula, which simplifies to a sample average of a sum of some derivatives of the log-density given by the model. The validity of the method is demonstrated on multivariate Gaussian and independent component analysis models, and by estimating an overcomplete filter set for natural image data.

Keywords: statistical estimation, non-normalized densities, pseudo-likelihood, Markov chain Monte Carlo, contrastive divergence

Aside: integration by parts trick

$$\mathbb{E}\|\nabla \log p^* - \nabla \log p\|^2 = C_{p^*} - 2\mathbb{E}\langle \nabla \log p^*, \nabla \log p \rangle + \mathbb{E}\|\nabla \log p\|^2$$

$$\mathbb{E}\langle \nabla \log p^*, \nabla \log p \rangle = \int p^*(x)\langle \nabla \log p^*(x), \nabla \log p(x) \rangle dx$$

$$= \int \langle \nabla p^*(x), \nabla \log p(x) \rangle dx$$

$$= -\int p^*(x) \Delta \log p(x) dx$$

“Green’s first identity”
(Divergence Theorem)

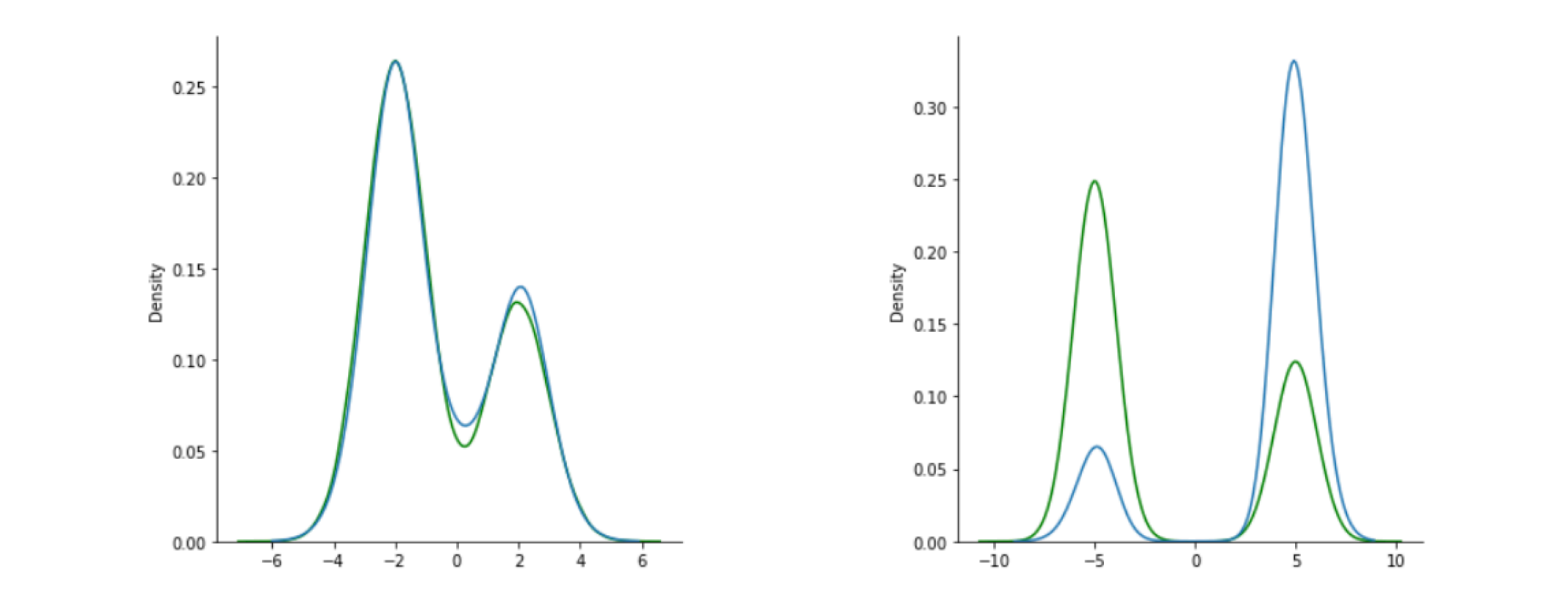
assumes decay at infinity

Key Question

what is the *statistical cost* of using score matching instead of maximum likelihood estimation ?

i.e. how many more samples will we need to achieve the same accuracy?

Score matching can struggle to learn the distribution



small separation

large separation

MLE (green, indistinguishable from truth) vs. score matching (blue)

see also [Wenliang et al ’18, Song and Ermon ’19,…]

Asymptotic Theory: Exponential Families

Exponential Families

Want to fit an **Exponential family**: $p_\theta(x) \propto \exp\left(\langle \theta, F(x) \rangle\right)$

- Ex: Gaussian ⇔ $F(x) = (x, xx^T)$. F is called the "sufficient statistic"
- I.e. an *energy-based model* with energy *linear* in "feature map" *F(x)*.

Computing MLE is possible if we can **efficiently sample** from p_θ.

But sampling can be computationally hard.

Score matching: **closed form** [Hyvarinen '07] !!!

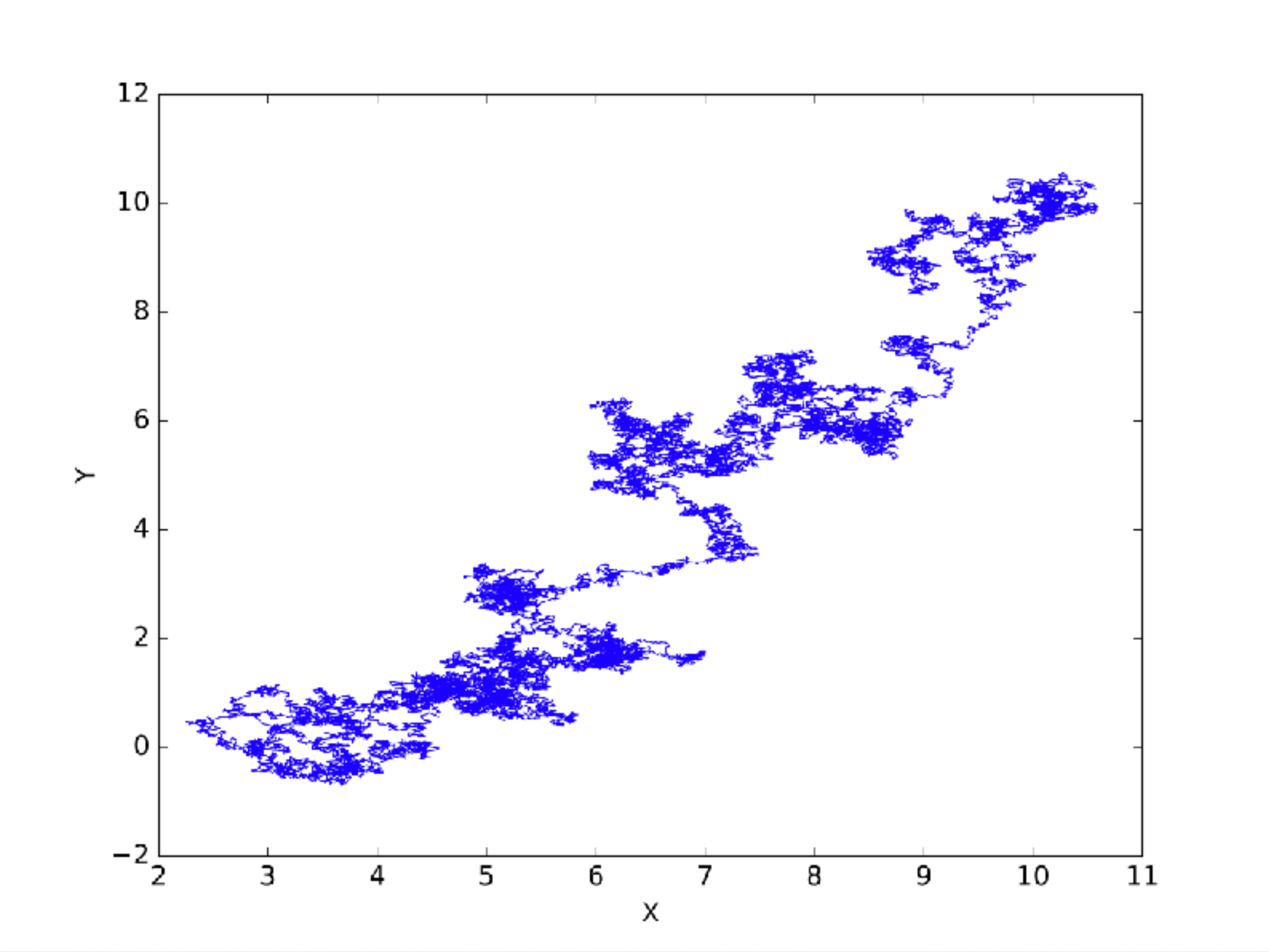
$$\hat{\theta}_{SM} = -\left(\hat{\mathbb{E}}[(JF)_X (JF)_X^T]\right)^{-1} \hat{\mathbb{E}} \Delta F, \quad \text{where} \quad \hat{E}[f] = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

Q: is there a catch? do we lose statistically for using score matching instead of MLE?

A Connection of Learning and Sampling

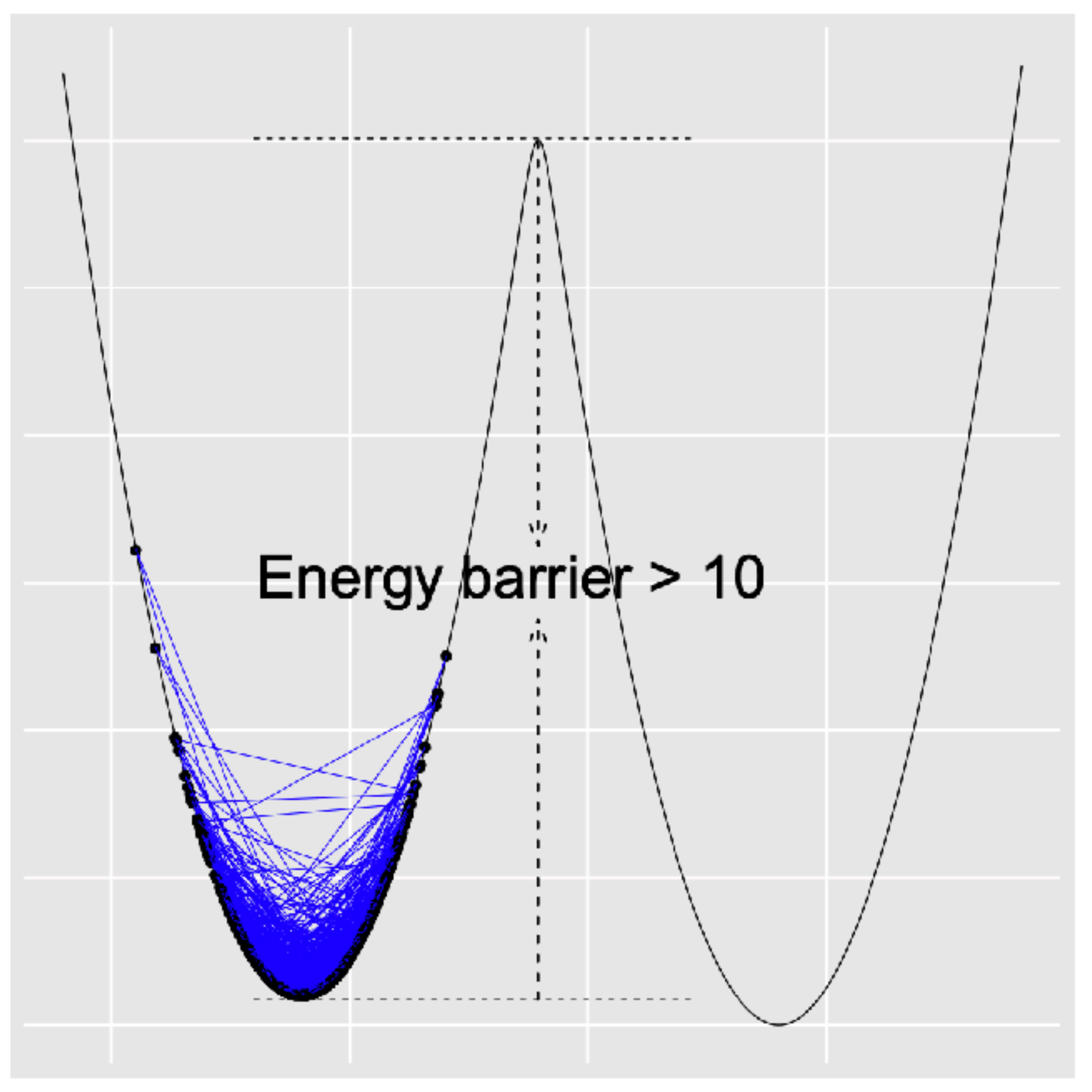
- We find **statistical performance** of score matching is deeply connected to **algorithmic performance** of the Langevin dynamics, a **canonical sampling algorithm.**
- **Langevin diffusion = "Gradient ascent plus noise"**

$$dX_t = \nabla \log p(X_t)dt + \sqrt{2}dB_t$$



Langevin dynamics for a Gaussian, started at (10,10) ("rapid mixing")

Torpid mixing of Langevin



https://waynedw.github.io/posts/CSGLD/

Quick Summary

Informal Theorem: Score matching is almost as sample-efficient as MLE if *Langevin mixes rapidly*!

roughly, $\|\hat{\theta}_{SM} - \theta\|^2 = O\left(\|\hat{\theta}_{MLE} - \theta\|^2\right)$ for sufficiently large # of samples n

Informal Converse: If Langevin mixes slowly, then **score matching pays a corresponding sample efficiency penalty** in any *sufficiently rich* exponential family.

What is relative (sample) efficiency?

- Let's be precise about *sample complexity.*
- *Notation:* $\Sigma_X = \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T$ is the covariance matrix of random vector X.
- ***Asymptotic normality of MLE:*** $\sqrt{n}(\hat{\theta}_{MLE} - \theta) \to N(0,\Gamma_{MLE})$
  - Here $\Gamma_{MLE} = \Sigma_F^{-1}$ is the "inverse Fisher information" and it is *optimal* in some strong (technically involved) sense.
- Score matching is also asymptotically normal: $\sqrt{n}(\hat{\theta}_{SM} - \theta) \to N(0,\Gamma_{SM})$

$$\Gamma_{SM} = \mathbb{E}[(JF)_X(JF)_X^T]^{-1}\,\Sigma_{(JF)_X(JF)_X^T\theta+\Delta F}\,\mathbb{E}[(JF)_X(JF)_X^T]^{-1}\,.$$

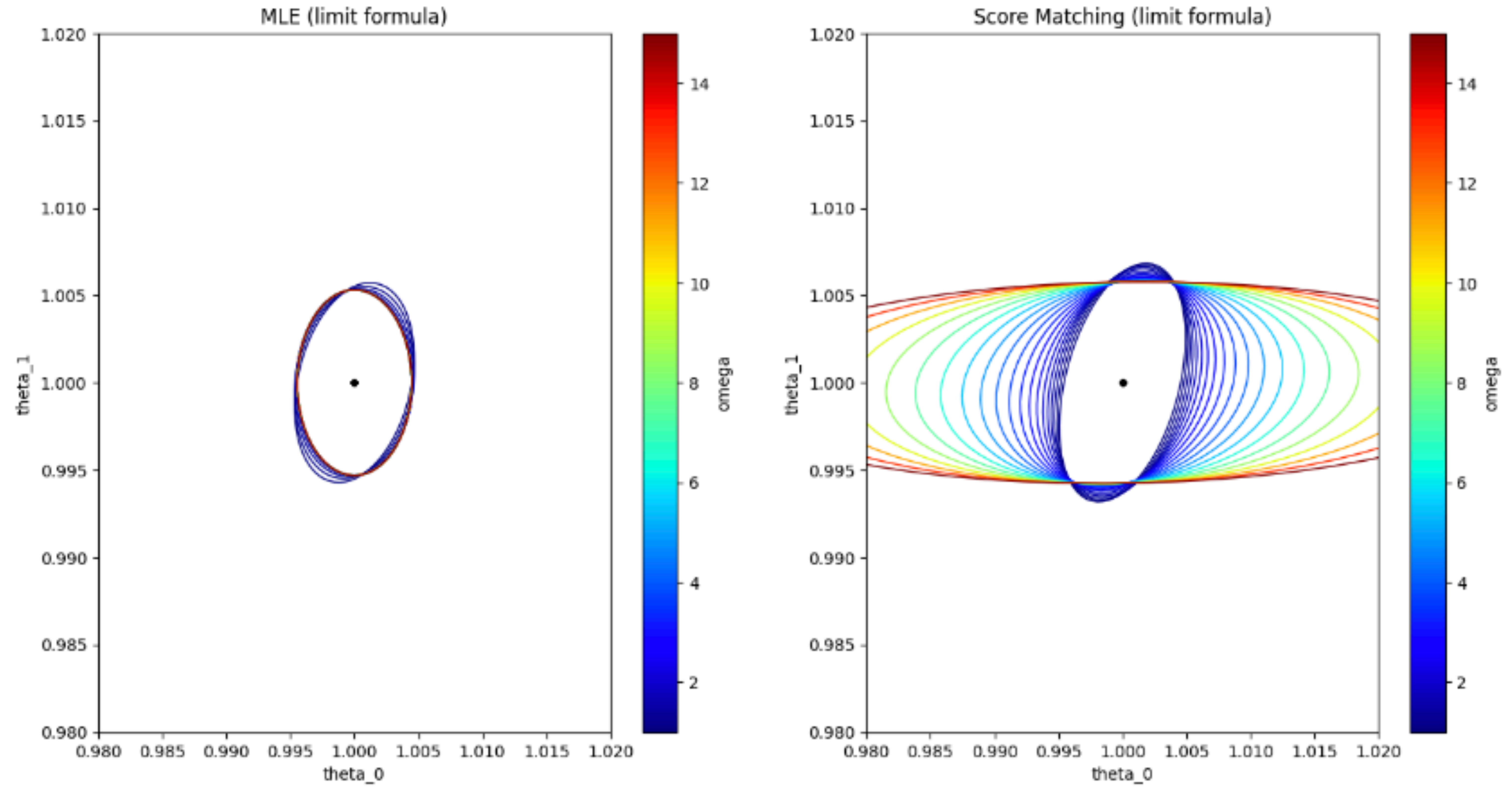
- "All we need to do": see how much bigger Γ_{SM} is compared to Γ_{MLE}

Aside: asymptotic normality

$$\hat{\theta}_{SM} = -\left(\hat{\mathbb{E}}[(JF)_X(JF)_X^T]\right)^{-1}\hat{\mathbb{E}}\Delta F, \quad \text{where} \quad \hat{E}[f] = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

- “Delta method” (see [Forbes-Lauritzen ’14]):
  - $\hat{\mathbb{E}}[(JF)_X(JF)_X^T] = \mathbb{E}[(JF)_X(JF)_X^T] + \Delta_1/\sqrt{n}$
  - $\hat{\mathbb{E}}\Delta F = \mathbb{E}\Delta F + \Delta_2/\sqrt{n}$
  - (Δ_1, Δ_2) jointly Gaussian by Central Limit Theorem.
  - Calculate it out…
- Determining Γ_{SM} is “standard”, but **analyzing it is the real challenge !**

Visualization of limiting covariances



(Restricted) Poincare constant

Definition: the *Poincare constant* of p_θ is the minimal $C_P > 0$ so that for every $\theta \in \Theta$ and function f,

$$\mathrm{Var}_{p_\theta}(f(x)) \leq C_P \, \mathbb{E}_{x \sim p_\theta} \|\nabla f(x)\|^2$$

Restricted Poincare constant: smallest C_P for all $f(x) = \langle \tau, F(x) \rangle$ where F(x) is the sufficient statistic. Can be smaller!

1. Poincare constant equivalent to *relaxation time* of Langevin dynamics.

2. We control performance of score matching by *restricted Poincare constant*.

Upper bound

- **Theorem**: if C_P is the restricted Poincare constant, then

$$\|\Gamma_{SM}\|_{OP} \leq 2C_P^2 \|\Gamma_{MLE}\|_{OP}^2 \left(\|\theta\|^2 \mathbb{E}\|(JF)_X\|_{OP}^4 + \mathbb{E}\|\Delta F\|_2^2 \right).$$

- So… our loss compared to MLE is controlled by:
  - Restricted Poincare constant C_P
  - Smoothness $\left(\|\theta\|^2 \mathbb{E}\|(JF)_X\|_{OP}^4 + \mathbb{E}\|\Delta F\|_2^2 \right)$
  - A quadratic factor.
- Dependence on restricted Poincare+"smoothness" is required (lower bound)

Proof idea

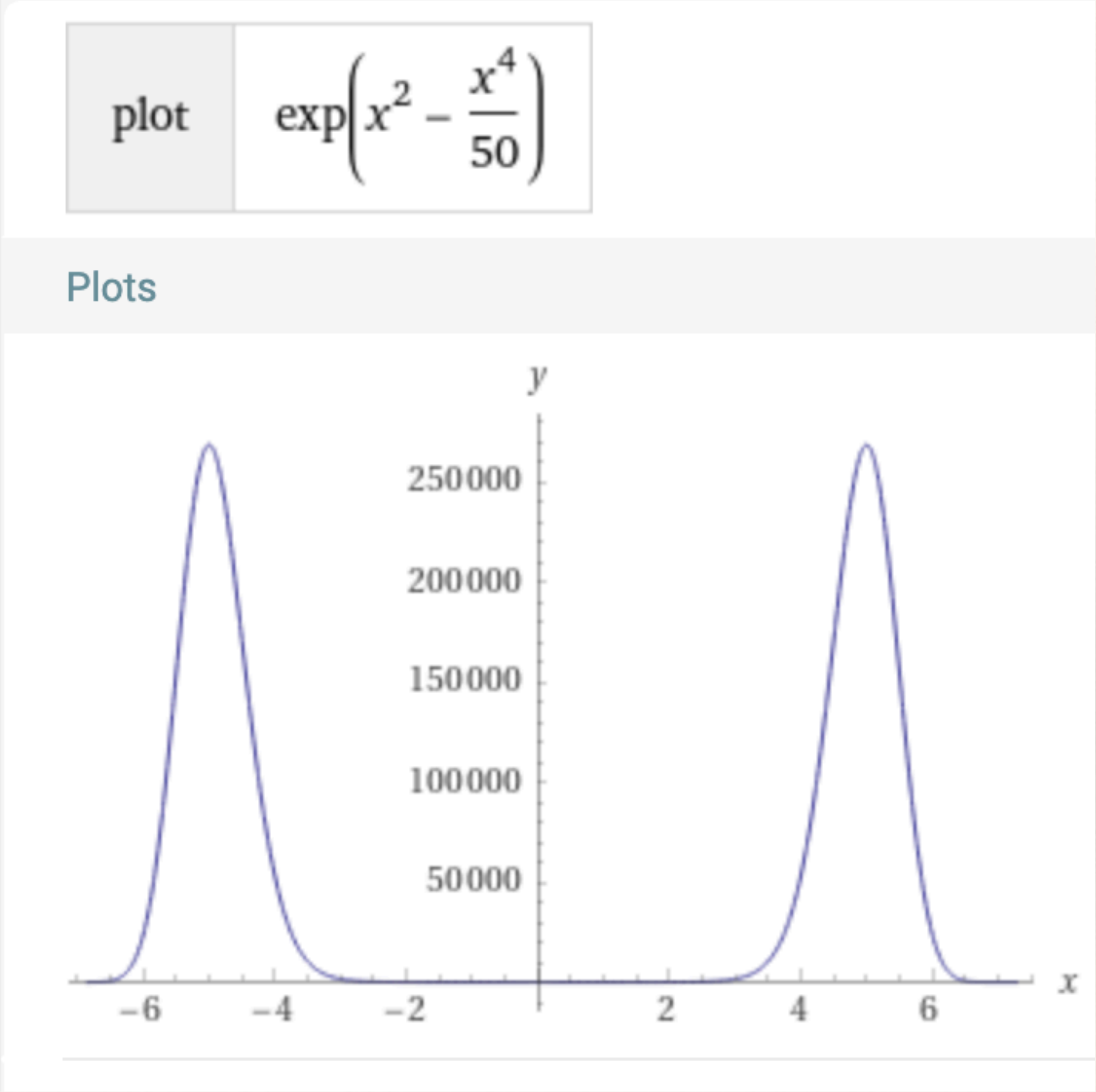
- Recall… $\Gamma_{SM} = \mathbb{E}[(JF)_X(JF)_X^T]^{-1}\,\Sigma_{(JF)_X(JF)_X^T\theta+\Delta F}\,\mathbb{E}[(JF)_X(JF)_X^T]^{-1}$.
- Proof straightforward given **Key Lemma.**
- **Key Lemma:** $\mathbb{E}[(JF)_X(JF)_X^T]^{-1} \leq C_P\Sigma_F^{-1}$
  - Proof: for any vector w,

$$\begin{aligned}\langle w, \mathbb{E}[(JF)_X(JF)_X^T]w\rangle &= \mathbb{E}\|\nabla_x\langle w, F(x)\rangle\,|_X\|_2^2 \\ &\geq \frac{1}{C_P}Var(\langle w, F(x)\rangle) = \frac{1}{C_P}\langle w, \Sigma_F w\rangle\end{aligned}$$

A bit about the lower bound

- Suppose that $p_\theta \propto \exp\left(\langle \theta, F_1(x)\rangle\right)$ has a large (unrestricted) Poincare constant C_P and F_1. *This does not imply a large restricted Poincare constant, nor the failure of score matching.* (Examples on next slides)
- However, we can view p_θ as an element of a a *larger exponential family* by $p_\theta(x) \propto \exp(\langle \theta, F_1(x)\rangle + \theta_2 F_2(x))$ where $\theta_2 = 0$ and F_2 is carefully chosen.
- In this *larger family*, restricted Poincare is big and score matching performs poorly —- there exists w so that

$$\frac{\langle w, \Gamma_{SM} w\rangle}{\langle w, \Gamma_{MLE} w\rangle} = \frac{\Omega_d\left(C_P\right)}{\text{smoothness}}$$



plot $\exp\left(x^2 - \frac{x^4}{50}\right)$

Plots

MLE
Score Matching
log(accuracy)
−5.0
−5.1
−5.2
−5.3
−5.4
−5.5
−5.6
−5.7
offset
1
2
3
4
5
6
7

Figure 3: Here we see the result of running an identical experiment to Figure 1, only we remove the second sufficient statistic, so our distribution is now $p_\theta(x) \propto e^{\theta_0(x^2 - x^4/(2a^2))}$ where $\theta_0 = 1$ and we again vary the offset a between 1 and 7. With only the single sufficient statistic, score matching performs comparably to MLE.

Score matching does work in *some* multimodal examples. Explained by “restricted Poincare constant”.

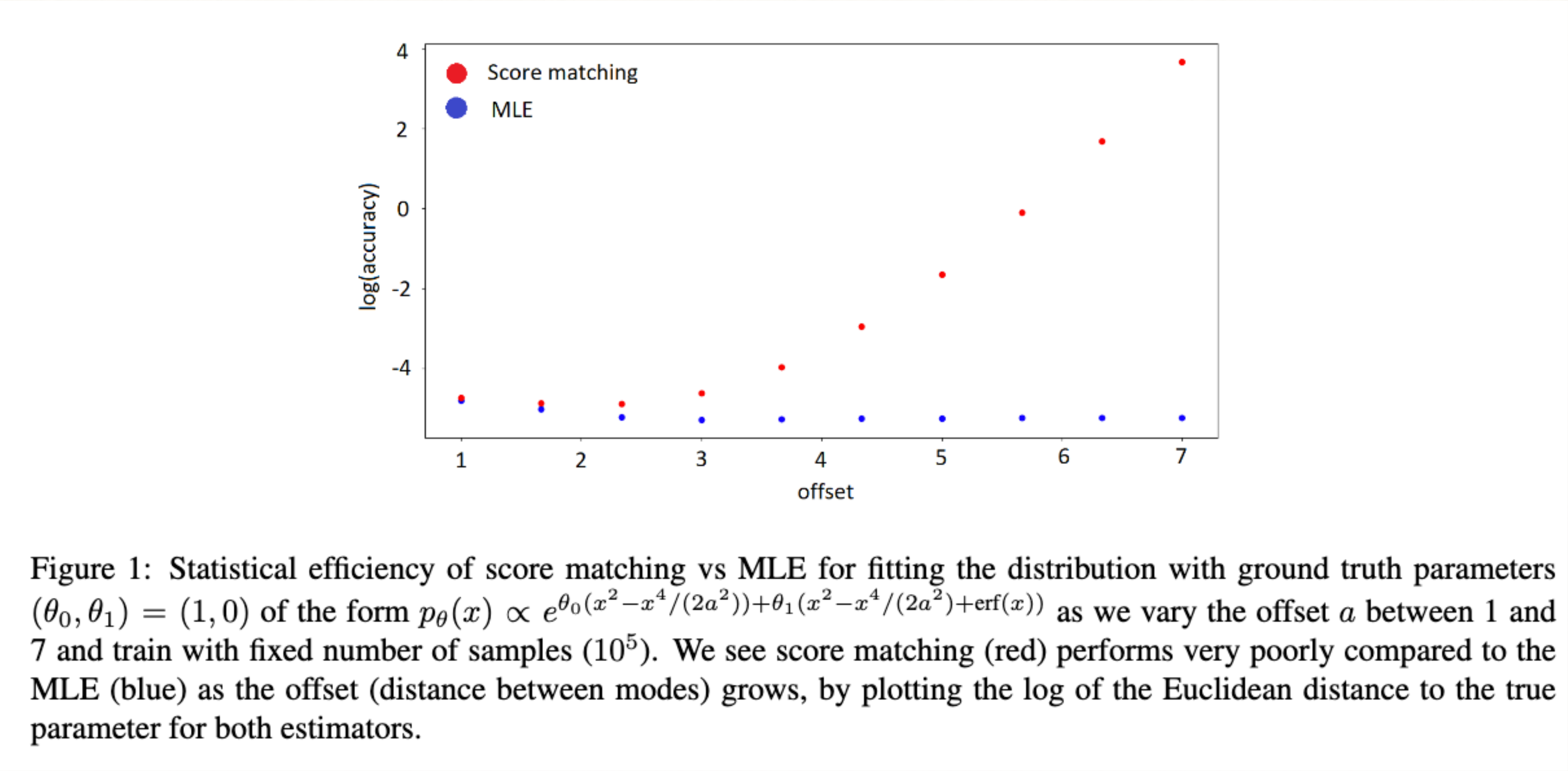
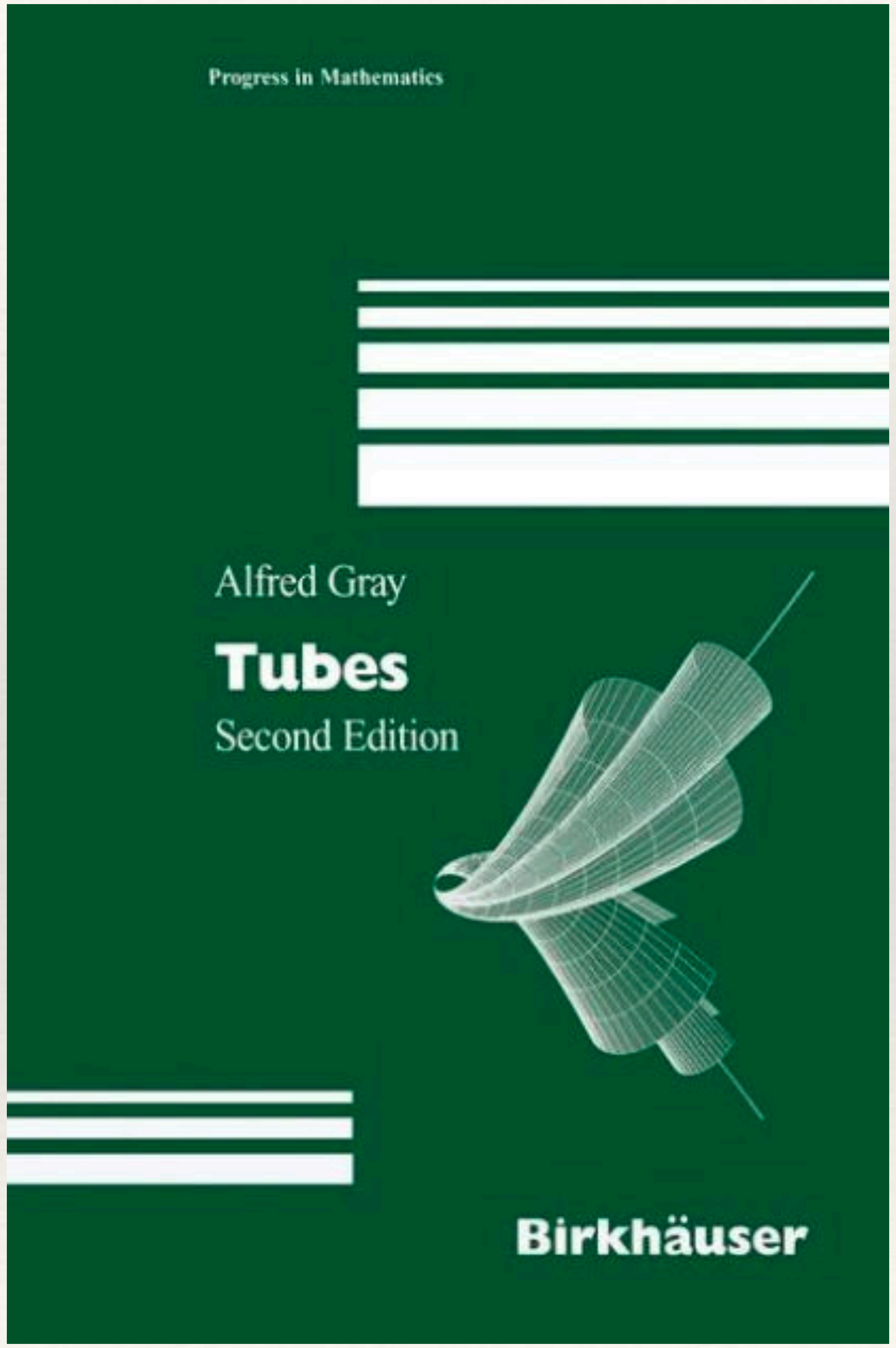


Figure 1: Statistical efficiency of score matching vs MLE for fitting the distribution with ground truth parameters $(\theta_0, \theta_1) = (1, 0)$ of the form $p_\theta(x) \propto e^{\theta_0(x^2-x^4/(2a^2))+\theta_1(x^2-x^4/(2a^2)+\mathrm{erf}(x))}$ as we vary the offset a between 1 and 7 and train with fixed number of samples (10^5). We see score matching (red) performs very poorly compared to the MLE (blue) as the offset (distance between modes) grows, by plotting the log of the Euclidean distance to the true parameter for both estimators.

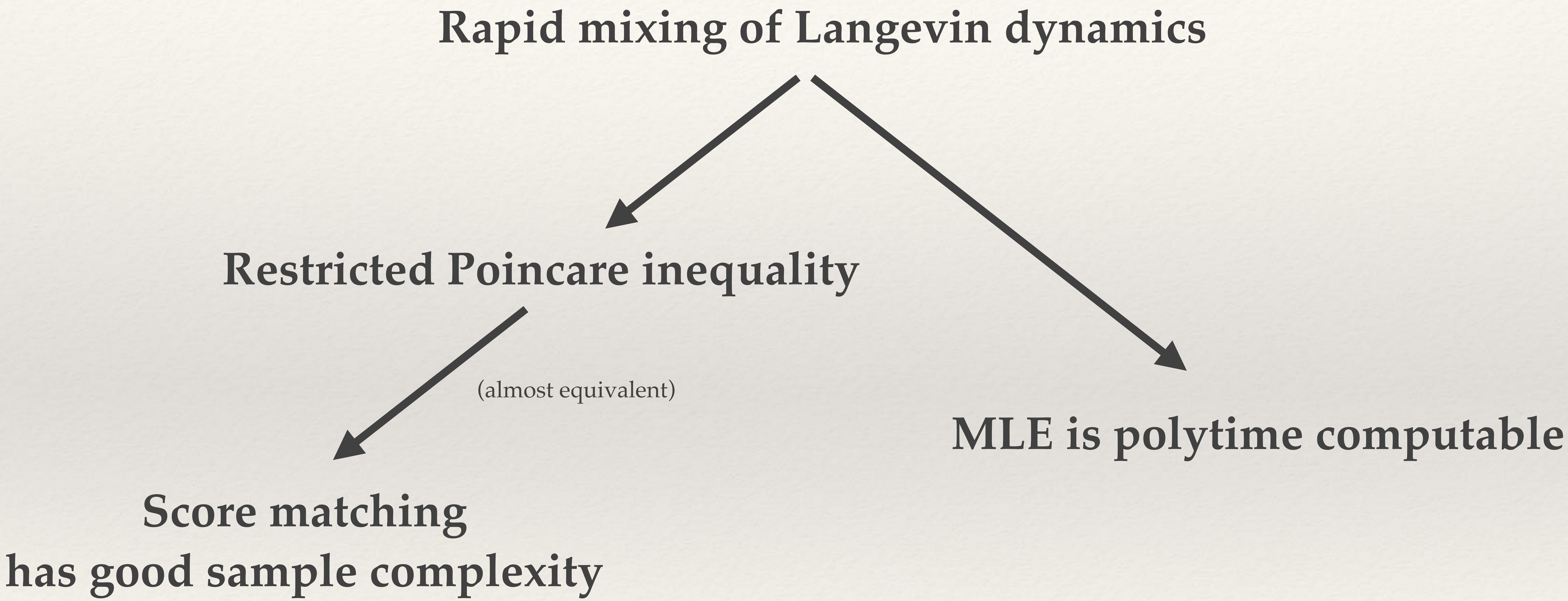
Adding a sufficient statistic with coefficient $\theta_1 = 0$ causes failure of score matching.

Proof idea

- Large Poincare constant implies a sparse cut exists
  - “Hard direction” of Cheeger's inequality.
- Add a new sufficient statistic corresponding to *smoothed indicator* of the sparse cut.
- Prove estimated coefficient for the new statistic has large variance.
  - uses formula for Γ_{MLE} and analysis on *tubular neighborhoods* of cut surface.



Summary (asymptotics)



Rapid mixing of Langevin dynamics

Restricted Poincare inequality

(almost equivalent)

Score matching
has good sample complexity

MLE is polytime computable

In general, MLE is hard to compute and score matching is statistically bad.
Rapid mixing of Langevin fixes both problems.

Nonasymptotic theory
(and some new connections!)

From now on, we are no longer specializing to exponential families.

Nonasymptotic theory

Log-sobolev constant C_q defined s.t. $\mathrm{KL}(p, q) \leq C_q \mathbb{E}\|\nabla \log p - \nabla \log q\|^2$

- Strengthening of Poincare. Mixing in KL Divergence.
- Relates KL divergence (what we care about) to *"population risk"* of score matching (i.e. "test error").

Score matching estimator is an *"empirical risk minimizer"* so *statistical learning theory* tells us how to control population risk.

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \Delta \log p_\theta(x_i) + \frac{1}{2}\|\nabla \log p_\theta(x_i)\|^2$$

Nonasymptotic theory (summary)

Suppose $p^* = p_{\theta^*}$ “well-specified” (for simplicity).

$$KL(p, p_{\hat{\theta}}) \leq C_{p_{\hat{\theta}}} \mathbb{E} \|\nabla \log p^* - \nabla \log p_{\hat{\theta}}\|^2 \leq 2 C_{p_{\hat{\theta}}} \mathcal{R}_n \text{ where}$$

log-Sobolev

“symmetrization”

$$\mathcal{R}_n := \mathbb{E}_{X_1, \ldots, X_n, \epsilon_1, \ldots, \epsilon_n} \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \left[\Delta \log p_\theta(X_i) + \frac{1}{2} \|\nabla \log p_\theta(X_i)\|^2 \right]$$

$$\epsilon_1, \ldots, \epsilon_n \sim Uni\{\pm 1\}$$

Rademacher complexity

Example

Example 1. *Suppose we are fitting an isotropic Gaussian in* d *dimensions with unknown mean* μ^* *satisfying* $\|\mu^*\| \leq R$. *The class of distributions* $\mathcal{P}$ *is* q_μ *with* $\|\mu\| \leq R$ *of the form* $q_\mu(x) \propto \exp\left(-\|x-\mu\|^2/2\right)$ *so the expected Rademacher complexity can be upper bounded as so:*

$$\mathcal{R}_n = \mathbb{E}\sup_\mu \frac{1}{n}\sum_{i=1}^n \epsilon_i \left[-d/2 + \frac{1}{2}\|X_i - \mu\|^2\right]$$

$$= \mathbb{E}\sup_\mu \left\langle \frac{1}{n}\sum_{i=1}^n \epsilon_i X_i, \mu \right\rangle = R\,\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i\right\| \leq R\sqrt{\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i\right\|^2} = R\sqrt{\frac{R^2+d}{n}}$$

where the inequality is Jensen's inequality and in the last step we expanded the square and used that $\mathbb{E}\epsilon_i\epsilon_j = 1(i=j)$ *and* $\mathbb{E}\|X_i\|^2 \leq R^2 + d$. *Recall that the standard Gaussian distribution is* 1*-strongly log concave so* $C_{LS} \leq 1/2$. *Hence we have the concrete bound* $\mathbb{E}\,\mathbf{KL}(p,\hat{p}) \leq R\sqrt{\frac{R^2+d}{n}}$.

More on Log-Sobolev

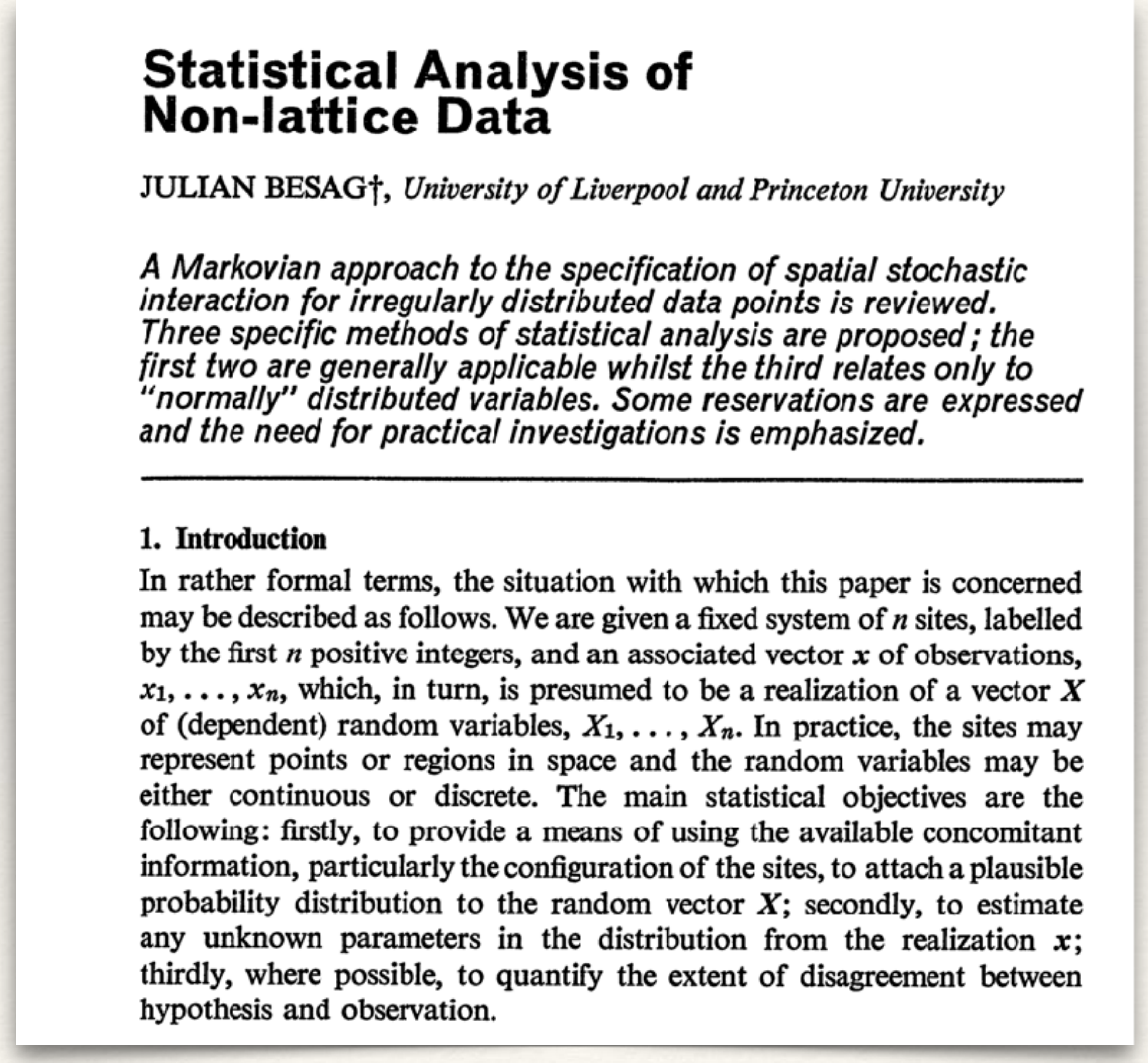
- In general, log-Sobolev is defined for an *arbitrary Markov semigroup*.
  - So far we only talked about *log-Sobolev for Langevin semigroup*.
- The "other" famous Markov chain is *Glauber dynamics* (Gibbs sampler)
  - At every step, pick a random coordinate i of X, resample $X_i \mid X_{\sim i}$
  - Makes sense for both continuous and discrete distribution.
- We can now answer: score matching is to Langevin as **???????** is to Glauber

Pseudolikelihood

- *Log-sobolev inequality* for Glauber dynamics on Q implies *approximate tensorization of entropy* [Marton '14, Caputo-Menz-Tetali '15]

$$\mathrm{KL}(P, Q) \leq C_q \sum_{i=1}^{n} \mathbb{E}_{X_{\sim i} \sim P} \mathrm{KL}(P(X_i \mid X_{\sim i}), Q(X_i \mid X_{\sim i}))$$

- This implies rapid mixing of Glauber.
- RHS is the (population) objective for the famous *pseudolikelihood* estimator [Besag '71] !!!
- Nonasymptotic bounds via symmetrization.
- C.f. [Hyvarinen '06, '07a, '07b, Lyu '09,…]



Statistical Analysis of Non-lattice Data

JULIAN BESAG†, *University of Liverpool and Princeton University*

A Markovian approach to the specification of spatial stochastic interaction for irregularly distributed data points is reviewed. Three specific methods of statistical analysis are proposed; the first two are generally applicable whilst the third relates only to "normally" distributed variables. Some reservations are expressed and the need for practical investigations is emphasized.

1. Introduction

In rather formal terms, the situation with which this paper is concerned may be described as follows. We are given a fixed system of n sites, labelled by the first n positive integers, and an associated vector $\boldsymbol{x}$ of observations, $x_1, \ldots, x_n$, which, in turn, is presumed to be a realization of a vector $\boldsymbol{X}$ of (dependent) random variables, $X_1, \ldots, X_n$. In practice, the sites may represent points or regions in space and the random variables may be either continuous or discrete. The main statistical objectives are the following: firstly, to provide a means of using the available concomitant information, particularly the configuration of the sites, to attach a plausible probability distribution to the random vector $\boldsymbol{X}$; secondly, to estimate any unknown parameters in the distribution from the realization $\boldsymbol{x}$; thirdly, where possible, to quantify the extent of disagreement between hypothesis and observation.

Summary & Thoughts (up to now)

- Score matching & pseudolikelihood vs MLE.
  - A kind of computational-statistical tradeoff. (See also follow up works e.g. [Pabbaraju et al ’23]…)
- Learning meets sampling.
  - Proving LSI/ATE has useful statistical implications!

A Teaser

- Message so far: *vanilla score matching doesn't work well with multimodal data and large function classes*
- There are many ways to fix it. One popular method is learning the *annealed score functions* (a type of diffusion model).
- But…could we use the vanilla score function anyway?

Inspiration from previous experiments

- Right: [Xie-Lu-Zhu-Wu ’16]
- Train an energy-based model via *contrastive divergence* (CD) *[Hinton ’06]*
- Key point: use training data to initialize the sampler!
- An old but effective method?

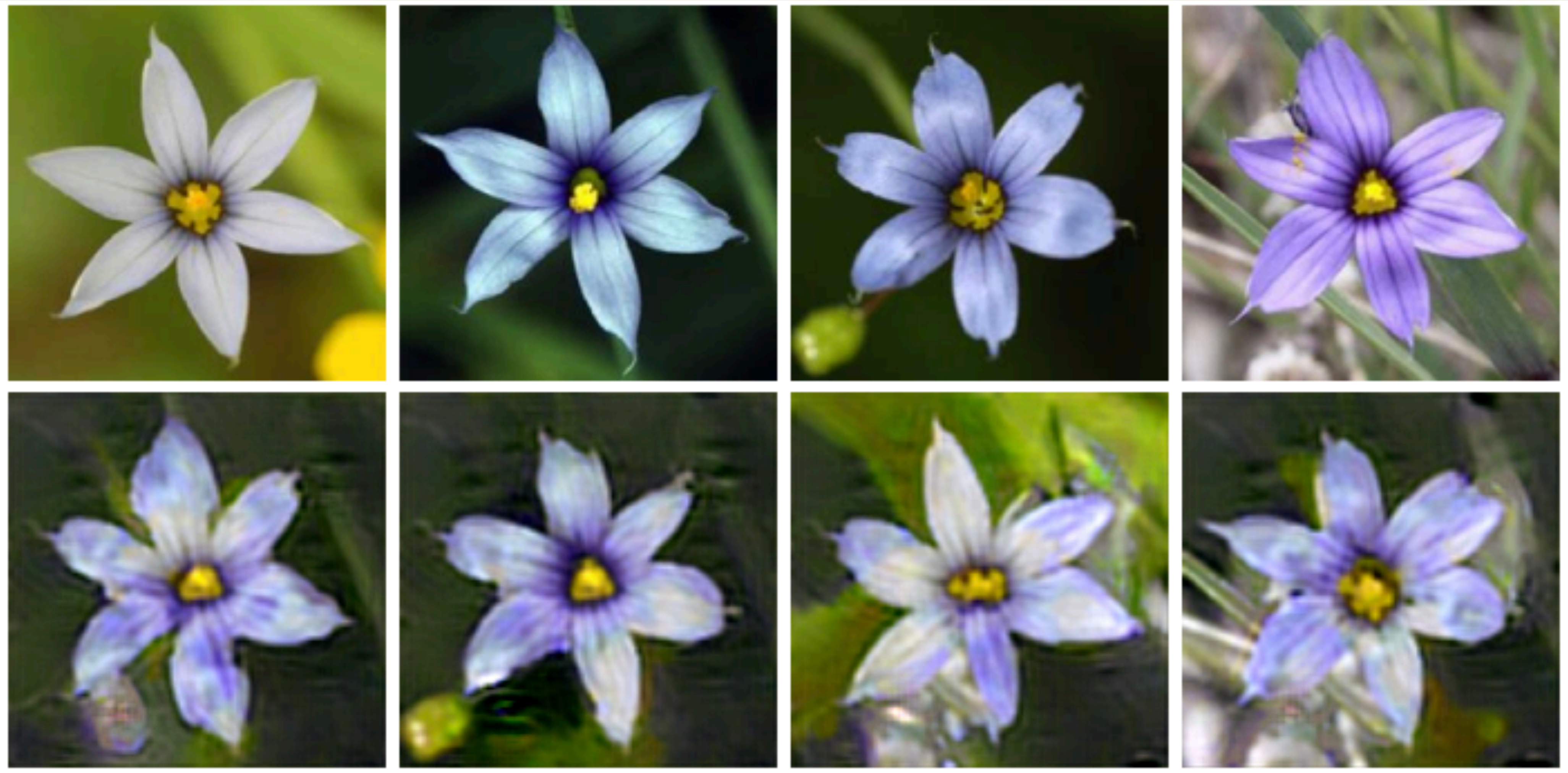


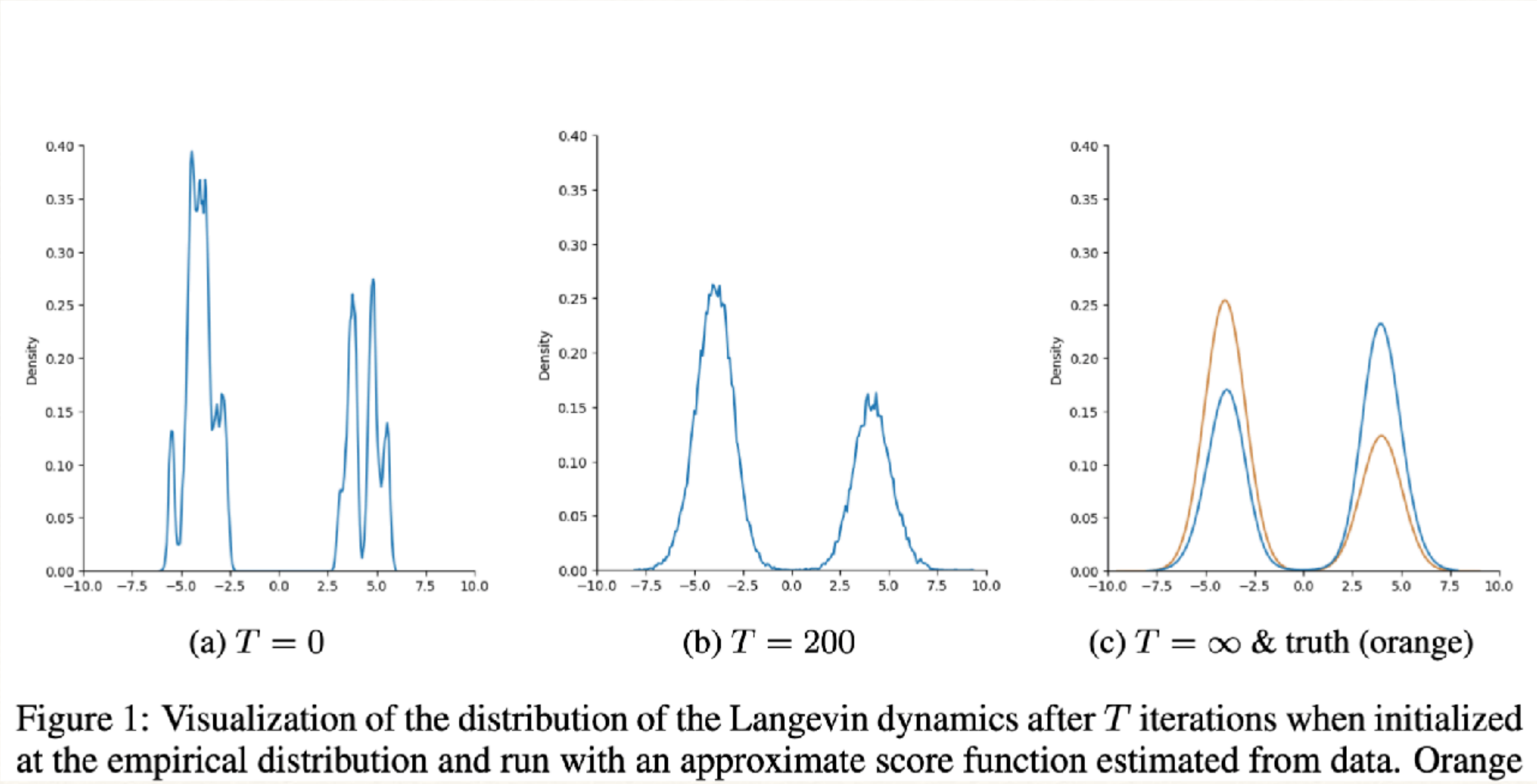
Figure 2. Generating object patterns. For each category, the first row displays 4 of the training images, and the second row displays 4 of the images generated by the learning algorithm.

Aside: what is contrastive divergence?

- Suppose $p_\theta(x) = \exp(f_\theta(x) - \log Z_\theta)$ where $Z_\theta = \int \exp(f_\theta(x))$
- MLE: maximize $\mathbb{E} \log p_\theta(x) = \mathbb{E} f_\theta(x) - \log Z_\theta$
- Gradient: $\mathbb{E}\, \nabla f_\theta - \mathbb{E}_\theta \nabla f_\theta$
- Run GD + approximate second expectation using data-based MCMC

Sampling multimodal distributions with the vanilla score

THM [K + Vuong ‘23+]: for mixtures of log-concave, Langevin with data-based initialization + early stopping succeeds with any vanilla score matched model.



(a) $T = 0$

(b) $T = 200$

(c) $T = \infty$ & truth (orange)

Figure 1: Visualization of the distribution of the Langevin dynamics after T iterations when initialized at the empirical distribution and run with an approximate score function estimated from data. Orange

Insights into contrastive divergence?

- Hyvarinen: score matching can be recovered as a limit of contrastive divergence training as step size goes to zero.

- It seems CD implicitly fits the vanilla score function, so our theory applies.

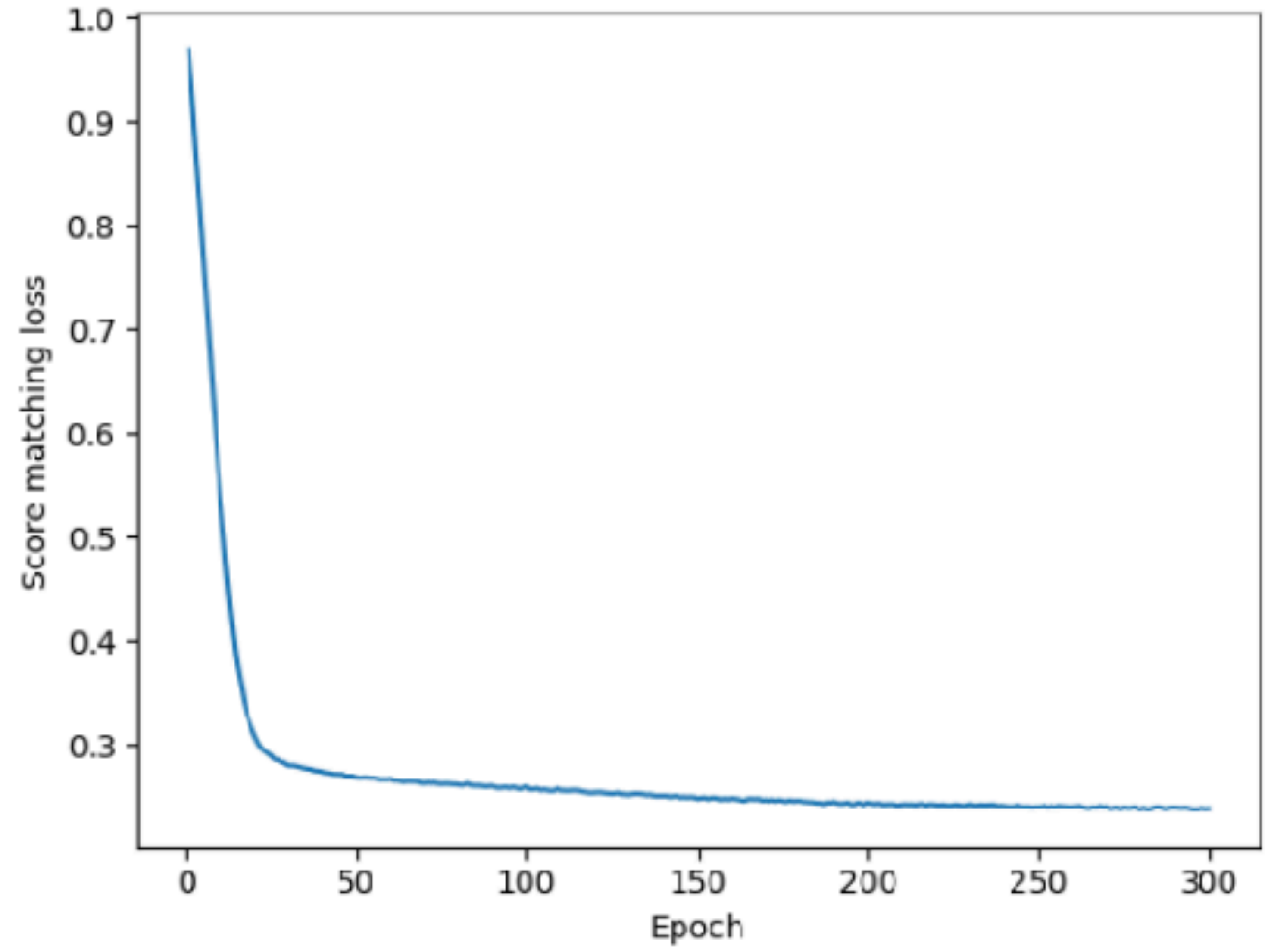
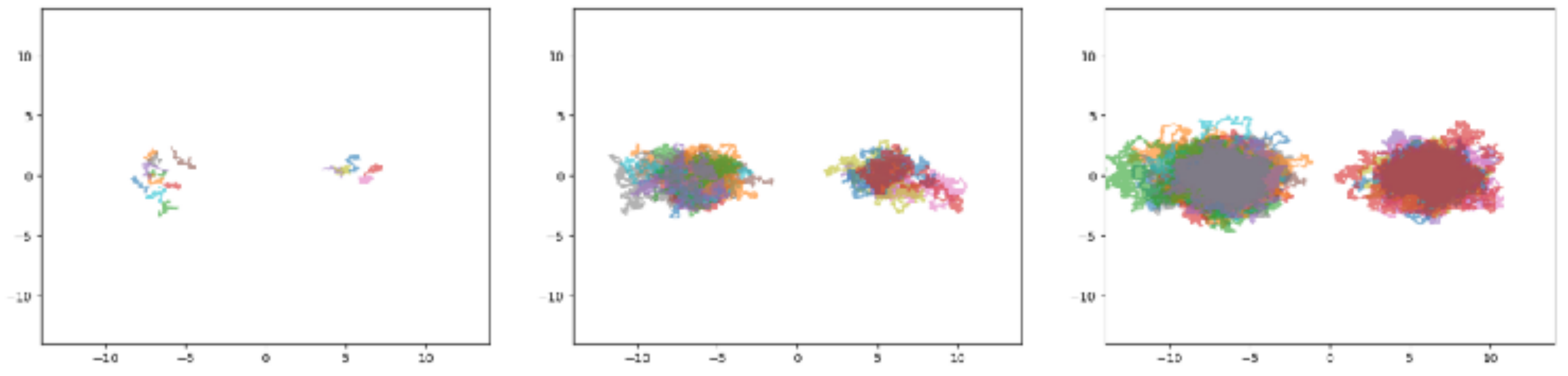


Figure 5: Score matching training loss (precisely the same loss used to train the models in Figures 2 and 3) curve for the CD-trained model in Figure 2. Although the score matching loss is not being explicitly optimized, we see it goes down monotonically over the epochs of CD training nonetheless.



(a) $T = 300$ (b) $T = 12000$ (c) $T = 120000$

Further Thoughts

- Please read our papers for more!
- What applications is vanilla score matching/CD the *right* approach for?
  - Seems unlikely *one* approach to generative modeling is the best for *all* types of data…
  - Clear benefits of pursuing a diversity of approaches. Otherwise, we would still be using GANs for everything.