# Lower Bounds for Sampling

Peter Bartlett
CS and Statistics
UC Berkeley

EPFL Open Problem Session. July 2020

# How hard is sampling?

**Problem:**

Given oracle access to a potential $f : \mathbb{R}^d \to \mathbb{R}$
(e.g., $x \mapsto f(x), \nabla f(x)$)
generate samples from $p^*(x) \propto \exp(-f(x))$.

# Positive results

**(Dalalyan, 2014)**

For smooth, strongly convex $f$, after $n = \Omega(d/\epsilon^2)$ gradient queries, overdamped Langevin MCMC has $\|p_n - p^*\|_{TV} \leq \epsilon$.

There are results of this flavor for stochastic gradient Langevin algorithms, underdamped Langevin algorithms, Metropolis-adjusted, nonconvex $f$, etc.

Lower bounds?

# Lower bound with a noisy gradient oracle

arXiv:2002.00291

## Problem:

Generate samples from $\mathbb{R}^d$ with density

$$p^*(x) \propto \exp(-f(x)),$$

with $f$ *smooth*, *strongly convex*.

Niladri Chatterji

Phil Long

## Information protocol

- Algorithm $\mathcal{A}$ is given access to a stochastic gradient oracle $Q$
- When the oracle is queried at a point $y$ it returns

$$z = \nabla f(y) + \xi,$$

  where $\xi$ is *unbiased* noise, *independent* of the query point $y$, with $\|\xi\| \leq d\sigma^2$
- The algorithm $\mathcal{A}$ is allowed to make $n$ *adaptive* queries to the oracle

# An information-theoretic lower bound

**Theorem**

*For all $d$, $\sigma^2$, $n \geq \sigma^2 d/4$ and for all $\alpha \leq \sigma^2 d/(256n)$,*

$$\inf_{\mathcal{A}} \sup_{Q} \sup_{p^*} \|\mathrm{Alg}[n; Q] - p^*\|_{\mathrm{TV}} = \Omega\Big(\sigma\sqrt{\frac{d}{n}}\Big),$$

*where the $p^*$ supremum is over $\alpha$-log smooth, $\alpha/2$-strongly log-concave distributions over $\mathbb{R}^d$.*

Hence, $\alpha$ is constant and $n = O(\sigma^2 d) \implies$

the worst-case total variation distance is larger than a constant.

For $\alpha, \sigma$ constant, matches upper bounds for stochastic gradient Langevin (Durmus, Majewski and Miasojedow, 2019).

# Proof idea

- Restrict to a finite parametric class (Gaussian) and a stochastic oracle that adds Gaussian noise.
- Like a classical comparison of statistical experiments:
  Relate the minimax TV distance to a difference of risk of two estimators, one that sees the algorithm's samples and one that sees the true distribution.
- Use Le Cam's method: relate estimation to testing.

# Open questions

- What if the noise has added structure?
  For example, what if the potential function is sum-decomposable and the oracle returns a gradient over a mini-batch of functions?

- Lower bounds for sampling with oracle access to the *exact* gradients?

Some lower bounds for related problems:


- Luis Rademacher and Santosh Vempala. Dispersion of mass and the complexity of randomized geometric algorithms. 2008.
- Rong Ge, Holden Lee, and Jianfeng Lu. Estimating normalizing constants for log-concave distributions: Algorithms and lower bounds. 2019.