Sharper Bounds for Chebyshev Moment Matching with Applications

(Apoorv Vikram Singh)
(NYU)

Joint work with: Cameron Musco, Christopher Musco, and Lucas Rosenblatt
(UMass Amherst) (NYU) (NYU)

Method of Moments

- Probability distribution p supported on $[-1,1]$
- Given noisy moments $\tilde{m}_1, \tilde{m}_2, \ldots$ estimate p

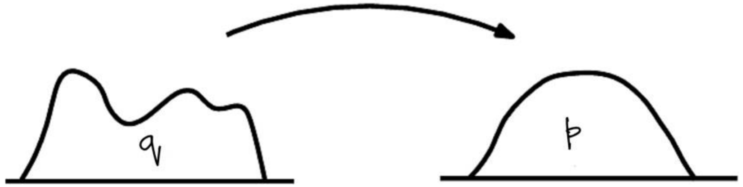
$$\tilde{m}_j = \int x^j p(x)dx + noise$$

Wasserstein 1 Distance

- Given ε, and noisy moments $\tilde{m}_1, \tilde{m}_2, \ldots$, output q such that $W_1(p,q) \leq \varepsilon$.

Wasserstein Distance

Wasserstein-1 distance: Minimum over all schemes of "moving" one distribution to another, where the cost of moving one unit of mass from x_1 in p to x_2 in q is $|x_1 - x_2|$



Dual of Wasserstein Distance

- Wasserstein-1 distance: Minimum over all schemes of “moving” one distribution to another, where the cost of moving one unit of mass from x_1 in p to x_2 in q is $|x_1 - x_2|$.

- Dual:

$$W_1(p, q) = \sup_{f:1-\text{Lipschitz}} \int f(x)\big(p(x) - q(x)\big)dx$$

Aim:

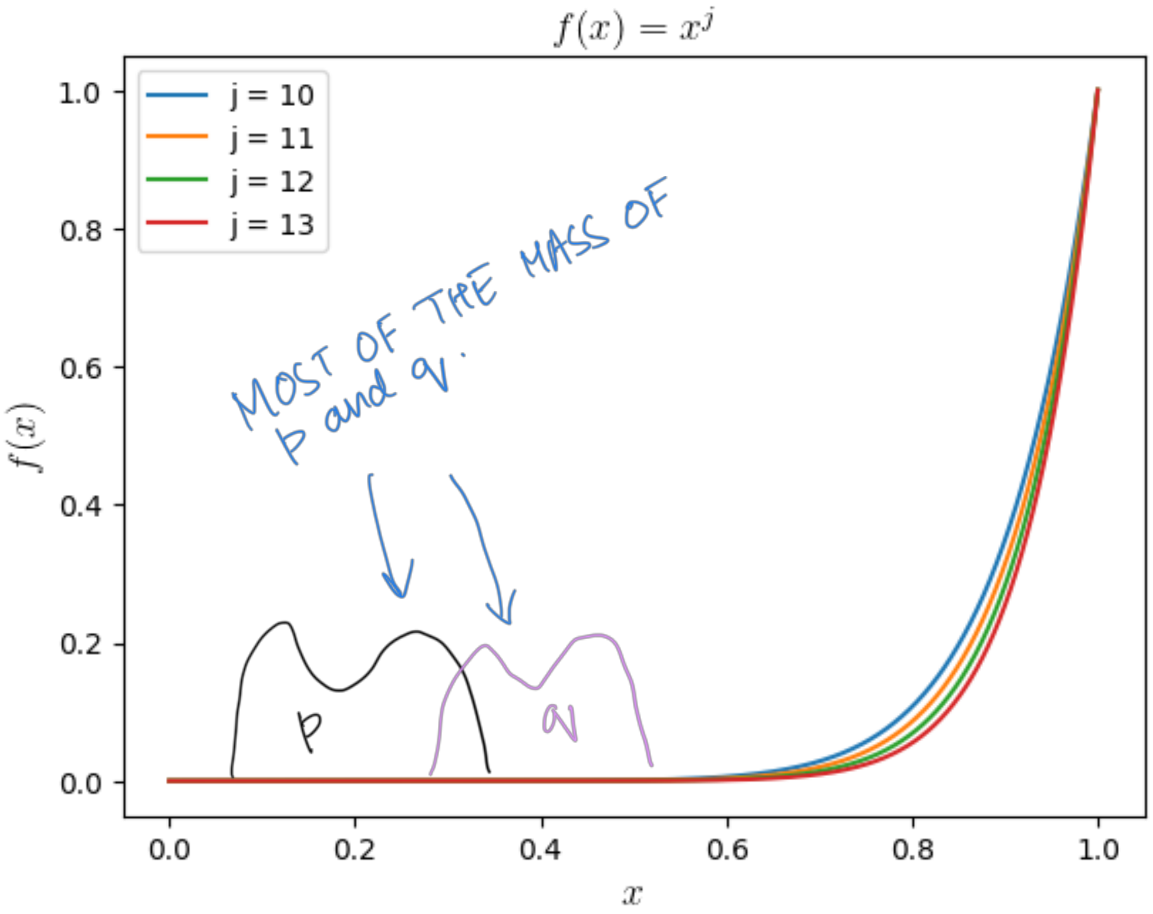
- Given noisy moment estimates of p, output distribution q such that

$$W_1(p,q) \le \varepsilon$$

Question: How much noise can we tolerate?

[KV’17, JMSS’23]: Need to estimate $m_1, \ldots, m_{1/\varepsilon}$ moments to accuracy $\pm \exp(-1/\varepsilon)$

[KV'17, JMSS'23]: Need to estimate $m_1, \ldots, m_{1/\varepsilon}$ moments to accuracy $\pm \exp(-1/\varepsilon)$ to output q such that $W_1(p, q) \leq \varepsilon$



Imagine integrating

$$m_j = \int x^j p(x) dx$$

How many samples do we need?

Getting the Moment Estimates

[KV'17, JMSS'23]: Need to estimate $m_1, \dots, m_{1/\varepsilon}$ moments to accuracy $\pm \exp(-1/\varepsilon)$ to output q such that $W_1(p,q) \le \varepsilon$

- For "vanilla" moments $m_j = \int x^j p(x) dx$, since p is supported on $[-1,1]$, we have that $|m_j| \le 1$.
- Let $X_1, \dots, X_n \sim_{\text{iid}} p$, then, let $\widetilde{m_j} = \frac{1}{n} \sum_{i \in [n]} X_i^j$.
- By Hoeffding's, we get that

$$\mathrm{P}\left(\left|\widetilde{m_j} - m_j\right| \ge t\right) \le 2\exp(-nt^2)$$

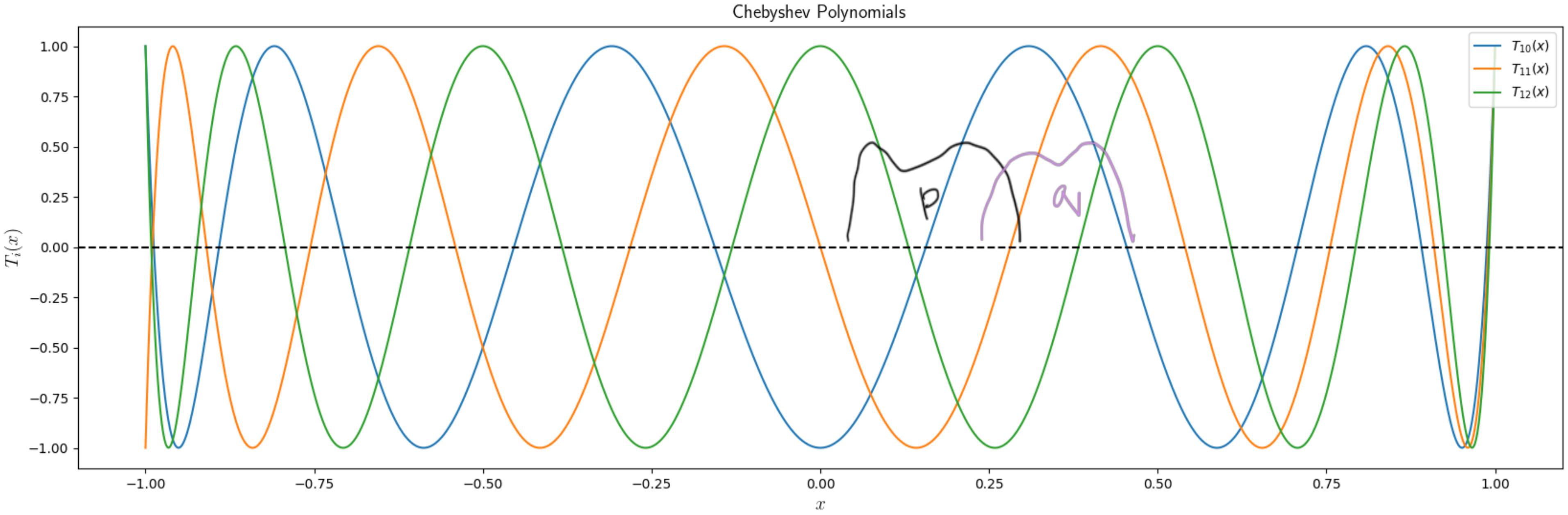
- Need $n = \exp\left(O\left(\frac{1}{\varepsilon}\right)\right)$

Chebyshev Polynomials (to the rescue)

- The i-th Chebyshev polynomial is denoted by T_i for $i = 0,1,2,\ldots$
- Defined recursively:

$$T_0(x) = 1, \qquad T_1(x) = x$$

$$T_i(x) = 2x\,T_{i-1}(x) - T_{i-2}(x), \text{ for } i \geq 2$$



Chebyshev Moments

- For a distribution p, its j-th Chebyshev moment is

$$t_j := \int T_j(x)p(x)dx$$

- [BKM'22]: Estimating t_j up to error $\pm\tilde{O}(\varepsilon)$ for $j = 1,2,\ldots,1/\varepsilon$ suffices to o/p a distribution q such that $W_1(p,q) \leq \varepsilon$
- By a similar analysis as before, we only need

$n = \widetilde{O}(1/\varepsilon^2)$ samples

What If We Have Access to Only Moments?

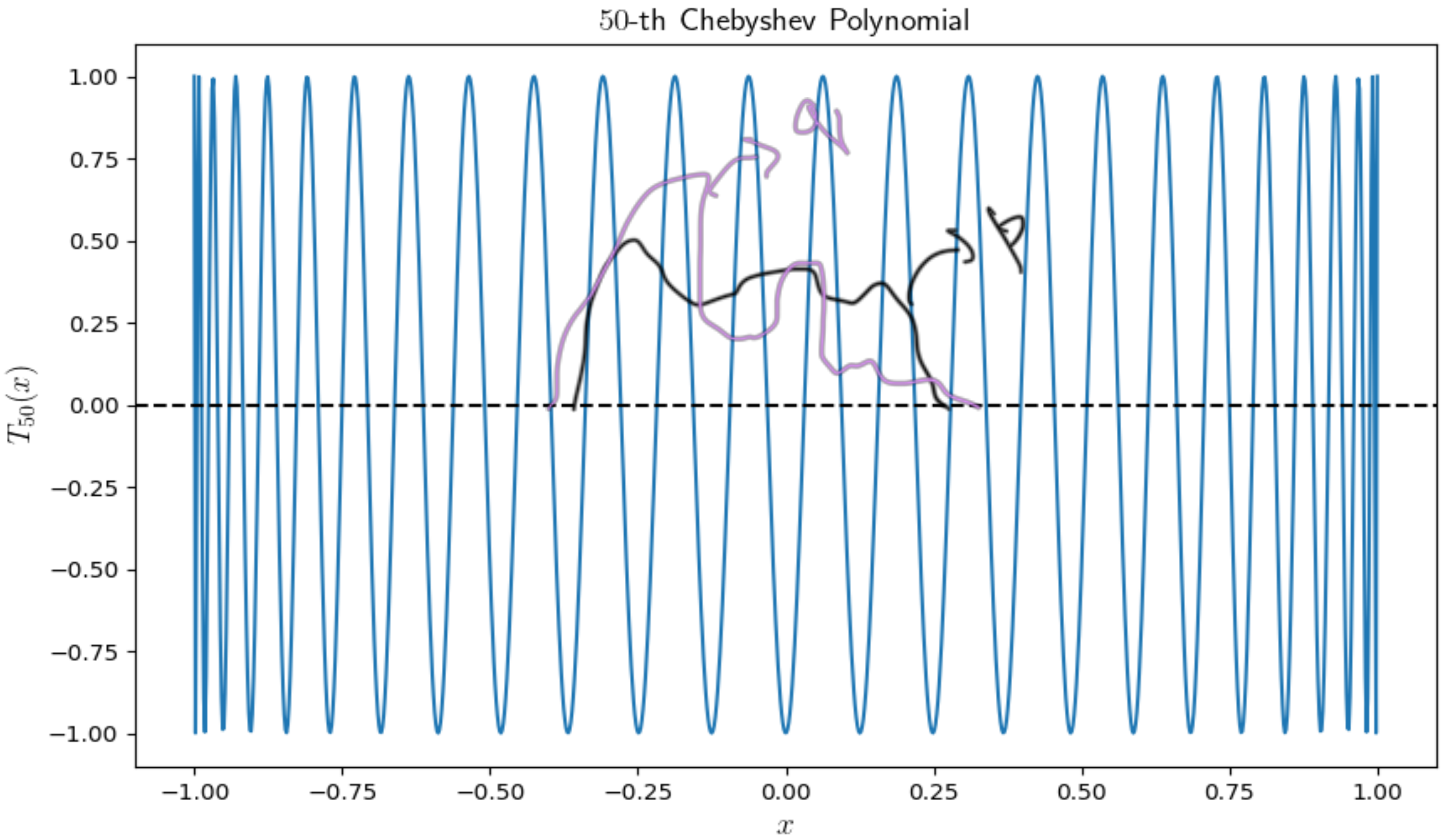
- In some applications, we do not have access to iid samples
- We have access to noisy moments of the distribution, and we want to recover the underlying distribution

"Vanilla" moments require very precise estimates of the moments

- [BKM'22]: Estimating first $1/\varepsilon$ Chebyshev moments up to error $\pm\,\tilde{O}(\varepsilon)$ suffices to o/p a distribution q such that $W_1(p,q) \le \varepsilon$

Our Result → Estimating t_j to accuracy $\pm O(\sqrt{j}\,\varepsilon)$, for $j = 1, \ldots, 1/\varepsilon$ suffices!

Why can we afford lower accuracy in higher moments?



For p and q far, the higher Chebyshev moments will differ more.

Formal Result

- Let p and q be two distributions supported on $[-1,1]$. Let k be an integer

Denote $\mathrm{E}_{x\sim p}\left[T_j(x)\right] := \int T_j(x)p(x)dx$

If $\sum_{j=1}^{k} \frac{1}{j^2}\left(\mathrm{E}_{x\sim p}\left[T_j(x)\right] - \mathrm{E}_{x\sim q}\left[T_j(x)\right]\right)^2 \le \Gamma^2$, then

$$W_1(p,q) \le \frac{c}{k} + \Gamma$$

- Set $k = 1/\varepsilon, \Gamma = \varepsilon$, we get $W_1(p,q) \le O(\varepsilon)$

Application 1: SDE

- Spectral Density Estimation: For a symmetric matrix $A \in \mathbb{R}^{n \times n}$, $\|A\|_2 \leq 1$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$, its spectral density is
$$p := \text{Unif}(\{\lambda_1, \ldots, \lambda_n\})$$
  - Aim: Output a distribution q such that $W_1(p, q) \leq \varepsilon$
  - Equivalently: Output a list of eigenvalues $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_n$ such that $\frac{1}{n}\sum_{i=1}^{n} |\lambda_i - \tilde{\lambda}_i| \leq \varepsilon$
  - Matrix-Vector Query Model: Given v, we get to observe Av
  - Goal: Minimize the number of matrix-vector queries to A

Spectral Density Estimation

- [BKM'22]: For matrix of size $n = \widetilde{\Omega}(1/\varepsilon^2)$, $\tilde{O}(1/\varepsilon)$ matrix-vector product with A suffices to estimate the spectral density of A.

Our Result: For matrix of any size, $\tilde{O}(1/\varepsilon)$ matrix-vector products with A suffices to estimate the spectral density of A.

Our Result: The number of queries is tight up to $\log$ factors

Application 2: Differential Privacy

- Given a data-set $x_1, \ldots, x_n$, we want to generate a differentially private synthetic data-set which is close to the original data-set
- Motivation: Perform downstream task without the need for a differentially private algorithm for each use-case
- Our Idea: Noise the Chebyshev moments of the uniform distribution over the data-set.
  - Can noise higher moments more.
  - Still recover a distribution close to the original distribution in W_1 distance.

Formal Result

- Let p and q be two distributions supported on $[-1,1]$. Let k be an integer

Denote $\mathrm{E}_{x\sim p}\left[T_j(x)\right] := \int T_j(x)p(x)dx$

If $\sum_{j=1}^{k} \frac{1}{\mathrm{j}^2}\left(\mathrm{E}_{x\sim p}\left[T_j(x)\right] - \mathrm{E}_{x\sim q}\left[T_j(x)\right]\right)^2 \leq \Gamma^2$, then

$$W_1(p,q) \leq \frac{c}{k} + \Gamma$$

- Set $k = 1/\varepsilon, \Gamma = \varepsilon$, we get $W_1(p,q) \leq O(\varepsilon)$

Proof Sketch

- Recall: $W_1(p,q) = \sup_{f:1-\text{Lipschitz}} \int f(x)\big(p(x) - q(x)\big)dx$
- Idea: Represent f, p, q in Chebyshev polynomial basis
$$f = c_0 + c_1 T_1(x) + \cdots + c_k T_k(x) + c_{k+1} + \cdots$$
- Jackson's Theorem: $f_k \coloneqq c_0 + c_1 T_1(x) + \cdots + c_k T_k(x)$ is a good uniform approximation, $\|f - f_k\|_\infty \leq O\left(\frac{1}{k}\right)$.
- $\int f(x)\big(p(x) - q(x)\big)dx = \int f_k(x)\big(p(x) - q(x)\big)dx + \int (f - f_k)\big(p(x) - q(x)\big)dx$

c/k

Focus: $\int f_k(x)\big(p(x) - q(x)\big)dx$

If $\sum_{j=1}^{k} \frac{1}{j^2}\left(\mathrm{E}_{x\sim p}[T_j(x)] - \mathrm{E}_{x\sim q}[T_j(x)]\right)^2 \leq \Gamma^2$, then $W_1(p,q) \leq \frac{c}{k} + \Gamma$

Recall: $f_k := c_0 + c_1 T_1(x) + \cdots + c_k T_k(x)$

- After some calculations,

$$\int f_k(x)\big(p(x) - q(x)\big)dx = \sum_{j=1}^{k} \int c_j \big(p(x) - q(x)\big)T_j(x)dx$$

$\to \left(\mathrm{E}_{x\sim p}[T_j(x)] - \mathrm{E}_{x\sim q}[T_j(x)]\right)$

- Cauchy Schwarz:

$$\leq \left(\sum_{j=1}^{k} (j^2 c_j^2)\right)^{\frac{1}{2}} \cdot \left(\sum_{j=1}^{k} \frac{1}{j^2}\left(\mathrm{E}_{x\sim p}[T_j(x)] - \mathrm{E}_{x\sim q}[T_j(x)]\right)^2\right)^{1/2}$$

$\to \leq \Gamma$

$\int f(x)\big(p(x)-q(x)\big)dx = \int f_k(x)\big(p(x)-q(x)\big)dx + \int (f-f_k)\,\big(p(x)-q(x)\big)dx$

Story Till Now

- Status: $\int f_k(x)\big(p(x)-q(x)\big)dx \le \left(\sum_{j=1}^{k}(j^2 c_j^2)\right)^{\frac{1}{2}}.\,\Gamma$

Our Result:

Let $f = c_0 + c_1T_1(x) + c_2T_2(x) + \ldots$ be a 1-Lipschitz function. Then,

$$\sum_{j=0}^{\infty}(j^2c_j^2) \le \frac{\pi}{2}$$

- We get $\int f_k(x)\big(p(x)-q(x)\big)dx \le \sqrt{\pi/2}\,.\,\Gamma$
- $\int f(x)\big(p(x)-q(x)\big)dx = \sqrt{\frac{\pi}{2}}\,\Gamma + \frac{c}{k}.$

Open Problem

Let $f = c_0 + c_1 T_1(x) + c_2 T_2(x) + \ldots$ be a 1-Lipschitz function. Then,

$$\sum_{j=0}^{\infty} (j^2 c_j^2) \leq \frac{\pi}{2}$$

- This result does not characterize 1-Lipschitz functions.
- Is there a characterization of 1-Lipschitz functions that maximizes W_1 distance?