Consider probabilities $\alpha_1, \ldots, \alpha_d$. Note that we'll often think of μ the discrete distribution that places weight one on each α_i. (This is kind of a different normalization from how it's defined in HJW.)

Our goal will be to use weak Schur sampling to estimate

$$\int x^k \mu(dx) = \sum_{i=1}^{d} \alpha_i^k,$$

(where bounds on error come from [AISW19]). Subsequently, we use these estimates to perform the "local moment matching" of [HJW18] to hopefully get a better sample complexity.

1 Classical component

Trying to interpret the analysis in [HJW18].

Theorem 1. *Let $\alpha_1, \ldots, \alpha_d$ be a probability distribution and suppose $B \geq \alpha_i$ for all $i \in [d]$. Fix some $K \in \mathbb{N}$. Suppose that, with probability $\geq 1 - \delta$, we have an estimate $\hat{m}_k$ for all $k \in [K]$ such that*

$$\left| \hat{m}_k - \int x^k \mu(dx) \right| < V_k.$$

Then we can find an estimate $\hat{\alpha}$ such that

$$\mathrm{E}[\|\alpha^< - \hat{\alpha}^<\|_1] \lesssim \frac{1}{K}\sqrt{Bd(1+V_1)} + 2^{9K/2} B \sum_{k=1}^{K} B^{-k} V_k + \delta.$$

[HJW18] takes $K = c_2 \ln(n)$, $B = c_1 \ln(n)/n$, and $V_k = \sqrt{d \ln(n)}(\frac{c_3}{c_1} B)^k = \sqrt{d \ln(n)}(\frac{c_3 \ln n}{n})^k$ ($c_3 > c_1$), achieving:

$$\begin{aligned}
&\mathrm{E}[\|\alpha^< - \hat{\alpha}^<\|_1] \\
&\lesssim \frac{1}{K}\sqrt{Bd(1+V_1)} + \delta + 2^{9K/2} B \sum B^{-k} V_k \\
&= \frac{1}{c_2 \ln(n)} \sqrt{\frac{c_1 \ln(n) d}{n}\left(1 + \frac{\sqrt{d}\ln^{1.5}(n) c_3}{n}\right)} + \delta + c_1 n^{\frac{9}{2}c_2 - 1} \ln(n) \sum_{k=1}^{c_2 \ln(n)} \sqrt{d \ln(n)} \left(\frac{c_3}{c_1}\right)^k \\
&\lesssim \sqrt{\frac{d}{n \ln(n)}\left(1 + \frac{\sqrt{d}\ln^{1.5}(n) c_3}{n}\right)} + \delta + n^{\frac{9}{2}c_2 - 1} \ln(n) \sqrt{d \ln(n)}\, n^{c_2 \ln(c_3/c_1)} \\
&\lesssim \sqrt{\frac{d}{n \ln(n)}} + \delta + n^{c_4 - 1}\sqrt{d}
\end{aligned}$$

Where we simply choose parameters to be sufficiently small, and c_4 can be as small as needed. Note that V_k can be quite large: the bound of $\sqrt{d \ln(n)} \cdot (O(B))^k$

can be replaced by any $d^{1-\varepsilon}(O(B))^k$ to get the $\frac{d}{\ln(d)}$ dependence we desire. (I think the whole argument still goes through even if variance is larger, in fact. It's not an artifact of having this B assumption.)

We use two pretty standard facts about polynomials: The first is Jackson's inequality, which gives an upper bound on the quality of approximation to a Lipschitz function:

Lemma 2 ([HJW18, Lemma 22])**.** *For $f : [a,b] \to \mathbb{R}$ a 1-Lipschitz function, the best polynomial approximation P of degree K satisfies*

$$|f(x) - P(x)| \leq \frac{C\sqrt{(b-a)(x-a)}}{K} \qquad \forall x \in [a,b].$$

for a universal constant C.

The second is the bound on coefficients of a bounded polynomial.

Lemma 3 ([HJW18, Lemma 27])**.** *Let $p(x) = \sum_{\nu=1}^{n} a_\nu x^\nu$ be a polynomial of degree at most n such that $|p(x)| \leq A$ for $x \in [a,b]$. Then if $a+b \neq 0$,*

$$|a_\nu| \leq 2^{7n/2} A \left|\frac{a+b}{2}\right|^{-\nu} \Big(\left|\frac{b+a}{b-a}\right|^n + 1\Big).$$

If $a+b=0$ then $|a_\nu| \leq Ab^{-\nu}(\sqrt{2}+1)^n$. (Ewin: I don't think this case is relevant for us.)

Proof. Recall the derivation done in [HJW18]: consider our given probability vector $\alpha = (\alpha_1, \ldots, \alpha_d)$, along with an estimate probability vector $\beta = (\beta_1, \ldots, \beta_d)$ that is formed by considering a measure ν and discretizing it as in [HJW18, Definition 8]. Then

$$\begin{aligned}
\mathrm{E}[\|\alpha^< - \beta^<\|_1] &= \mathrm{E}[W(\mu, \mu_\beta)] && \text{by [HJW18, Lemma 7]} \\
&= W(\mu, \nu) && \text{by [HJW18, Lemma 9]} \\
&= \sup_{f:\|f\|_{\mathrm{Lip}} \leq 1} \int f(x)(\mu(dx) - \nu(dx)). && \text{by [HJW18, Lemma 10]}
\end{aligned}$$

So, the goal is to find some measure ν (not necessarily discrete) that cannot be distinguished from the target distribution μ via 1-Lipschitz functions.

Let $\hat{\mu}$ be any measure satisfying $\hat{\mu}([0,1]) = \mu([0,1]) = n$ and

$$\left|\hat{m}_k - \int x^k \hat{\mu}(dx)\right| < V_k$$

for all $k \in [K]$. From the proof assumption, such a $\hat{\mu}$ exists with probability $\geq 1-\delta$, and by triangle inequality this $\hat{\mu}$ satisfies

$$\left|\int x^k \hat{\mu}(dx) - \int x^k \mu(dx)\right| < 2V_k. \tag{1}$$

This will be our proposed estimate measure. So, consider some $f : \mathbb{R} \to \mathbb{R}$ that is 1-Lipschitz satisfying $f(0) = 0$ (without loss of generality). We will take a polynomial approximation to f. Consider a polynomial $P(x) = \sum_{k=0}^{K} a_k x^k$, which we take to be the polynomial minimizing pointwise deviation from $f(x)$, $P := \arg\min_Q \max_x |Q(x) - f(x)|$. Then

$$
\begin{aligned}
&\left| \int f(x)(\mu(dx) - \hat{\mu}(dx)) \right| \\
&\leq \left| \int (f(x) - P(x))(\mu(dx) - \hat{\mu}(dx)) \right| + \left| \int P(x)(\mu(dx) - \hat{\mu}(dx)) \right| \\
&\leq \int |f(x) - P(x)|(\mu(dx) + \hat{\mu}(dx)) + \sum_{k=1}^{K} |a_k| \cdot 2V_k \\
&\leq \int |f(x) - P(x)|(\mu(dx) + \hat{\mu}(dx)) + 2\sum_{k=1}^{K} |a_k| V_k
\end{aligned}
$$

By appying Lemma 2 with $[a, b] = [0, B]$, we can bound the first term.

$$
\begin{aligned}
&\int |f(x) - P(x)|(\mu(dx) - \hat{\mu}(dx)) \\
&\leq \frac{C\sqrt{B}}{K} \int \sqrt{x}(\mu(dx) + \hat{\mu}(dx)) \\
&\leq \frac{C\sqrt{B}}{K} \sqrt{\Big(\int \sqrt{x}^2 (\mu(dx) + \hat{\mu}(dx)) \Big) \Big(\int 1^2 (\mu(dx) + \hat{\mu}(dx)) \Big)} && \text{by Cauchy-Schwarz} \\
&= \frac{C\sqrt{2Bd}}{K} \sqrt{\int x(\mu(dx) + \hat{\mu}(dx))} \\
&= \frac{C\sqrt{2Bd}}{K} \sqrt{\int x(2\mu(dx)) + \int x(\hat{\mu}(dx) - \mu(dx))} \\
&= \frac{C\sqrt{2Bd}}{K} \sqrt{2 + 2V_1} && \text{by Eq. (1)} \\
&\lesssim \frac{1}{K} \sqrt{Bd(1 + V_1)}
\end{aligned}
$$

To upper bound the second part, we need upper bounds on the coefficients.

$$|P(x)| \leq |P(x) - f(x)| + |f(x)| \leq \frac{CB}{K} + B$$

so, using Lemma 3, for all $k \in [K]$

$$
\begin{aligned}
|a_k| &\leq 2^{7K/2+1} B \Big(1 + \frac{C}{K}\Big) \Big(\frac{B}{2}\Big)^{-k} \\
&\leq 2^{9K/2+1} \Big(1 + \frac{C}{K}\Big) B^{1-k}
\end{aligned}
$$

and

$$
\begin{aligned}
2\sum_{k=1}^{K}|a_k|V_k &\leq 2\sum_{k=1}^{K}2^{9K/2+1}\Big(1+\frac{C}{K}\Big)B^{1-k}V_k \\
&\leq (1+C)2^{9K/2+2}\sum B^{1-k}V_k \lesssim 2^{9K/2}\sum B^{1-k}V_k
\end{aligned}
$$

Putting everything together, we have

$$
\begin{aligned}
&\mathrm{E}\,\|\alpha^{<}-\beta^{<}\|_1 \\
&= \mathrm{E}\sup_{f:\|f\|_{\mathrm{Lip}}\leq 1}\int_{\mathbb{R}} f(x)(\mu(dx)-\hat{\mu}(dx)) \\
&\lesssim \Big(\frac{1}{K}\sqrt{Bd(1+V_1)}+2^{9K/2}\sum B^{1-k}V_k\Big)+(\max\|P^{<}-\hat{P}^{<}\|_1)(\Pr[\text{alg fails}]) \\
&\lesssim \frac{1}{K}\sqrt{Bd(1+V_1)}+2^{9K/2}\sum B^{1-k}V_k+\delta
\end{aligned}
$$

□

2 Quantum component

Lemma 4 ([AISW19, Lemma 9]). *There is a constant C_k depending only on k such that*

$$
\mathrm{E}[\frac{1}{n^{\underline{k}}}p^{\#}_{(k)}(\boldsymbol{\lambda})] = \int x^k\mu(dx) \tag{2}
$$

$$
\mathrm{Var}[\frac{1}{n^{\underline{k}}}p^{\#}_{(k)}(\boldsymbol{\lambda})] = C_k\Big(n^{-k}+n^{-1}\int x^{2k-1}\mu(dx)\Big) \tag{3}
$$

Here, $p^{\#}_{(k)}(\boldsymbol{\lambda})$ is the estimator arising from what I think is the naive thing, which is writing the power sum (which they call $M_k(\alpha)$) in the Schur polynomial basis $\{s_\lambda(\alpha)\}$ and using the coefficients to renormalize the probabilities such that the estimator is unbiased. $n^{\underline{k}}$ is the falling power $(n)(n-1)\cdots(n-k+1)$.

3 Combining the two

For simplicity, we'll take the case in [HJW18] where $B = O(\ln^2 d/d)$.

$$
\begin{aligned}
\mathrm{E}[\|\alpha^{<}-\hat{\alpha}^{<}\|_1] &\lesssim \frac{1}{K}\sqrt{Bd(1+V_1)}+\delta+2^{9K/2}B\sum B^{-k}V_k \\
&\lesssim \frac{1}{K}\ln(d)\sqrt{1+V_1}+\delta+\frac{2^{9K/2}}{d}(\operatorname{poly}\log(d))\sum V_k d^k
\end{aligned}
$$

We'll do a very rough sanity check by judging V_k to be like square root of variance and taking $C_k = 1$ (though this will make a difference in the log factors), making this

$$
\lesssim \frac{\ln d}{K}+\delta+\frac{2^{9K/2}}{d}\operatorname{poly}\log d\sum d^k\sqrt{n^{-k}+n^{-1}d^{3-2k}}
$$

Now, we take $K = O(\ln d)$ for sufficiently small constant.

$$\lesssim 1 + \delta + d^{\varepsilon - 1} \operatorname{poly}\log d \sum d^k \sqrt{n^{-k} + n^{-1} d^{3-2k}}$$
$$\lesssim 1 + \delta + d^{\hat{\varepsilon} - 1} \sum d^k \sqrt{n^{-k} + n^{-1} d^{3-2k}}$$

Where we bound V_k^2 as follows:

$$n^{-k} + n^{-1} \int x^{2k-1} \mu(dx) \leq n^{-k} + n^{-1} d B^{2k-2} \lesssim n^{-k} + n^{-1} d^{3-2k}$$

The final quantity should be $O(1)$ when we take $n = O(d^2 / \operatorname{poly}\log(d))$, which is the desired dependence (presumably one can add dependence on ε later).

References

[AISW19] Jayadev Acharya, Ibrahim Issa, Nirmal V. Shende, and Aaron B. Wagner. Measuring quantum entropy. In *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, July 2019.

[HJW18] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under wasserstein distance. volume 75 of *Proceedings of Machine Learning Research*, pages 3189–3221. PMLR, 06–09 Jul 2018.