Extrapolating the profile of a finite population

Soham Jana[1], Yury Polyanskiy[2], Yihong Wu[1]

[1]Yale University
[2]Massachusetts Institute of Technology

June 22, 2020

Bernoulli sampling model [Bunge and Fitzpatrick, 1993]

- Today I will be presenting my joint work with Yury Polyanskiy and Yihong Wu on profile estimation for a finite population under certain small sample regime.
- We consider an urn model with k-balls with each belonging to one of k possible types. Note that some of the colors might be empty.
- For this population we want to estimate the histogram of colors denoted by π, also known as the profile of the population [Orlitsky et al., 2005]. The j-th entry of π gives us the proportion of color groups with j-balls.
- The sampling scheme we consider is also known as the Bernoulli sampling model [Bunge and Fitzpatrick, 1993] where we observe each of the balls with probability p which might be close to 0.
- Note that even though the full distribution of colors (call it μ) contains more information, it is impossible to consistently estimate it for any vanishing p. So instead we estimate π.

Bernoulli sampling model [Bunge and Fitzpatrick, 1993]

- Today I will be presenting my joint work with Yury Polyanskiy and Yihong Wu on profile estimation for a finite population under certain small sample regime.
- We consider an urn model with k-balls with each belonging to one of k possible types. Note that some of the colors might be empty.
- For this population we want to estimate the histogram of colors denoted by π, also known as the profile of the population [Orlitsky et al., 2005]. The j-th entry of π gives us the proportion of color groups with j-balls.
- The sampling scheme we consider is also known as the Bernoulli sampling model [Bunge and Fitzpatrick, 1993] where we observe each of the balls with probability p which might be close to 0.
- Note that even though the full distribution of colors (call it μ) contains more information, it is impossible to consistently estimate it for any vanishing p. So instead we estimate π.

Bernoulli sampling model [Bunge and Fitzpatrick, 1993]

- Today I will be presenting my joint work with Yury Polyanskiy and Yihong Wu on profile estimation for a finite population under certain small sample regime.
- We consider an urn model with k-balls with each belonging to one of k possible types. Note that some of the colors might be empty.
- For this population we want to estimate the histogram of colors denoted by π, also known as the profile of the population [Orlitsky et al., 2005]. The j-th entry of π gives us the proportion of color groups with j-balls.
- The sampling scheme we consider is also known as the Bernoulli sampling model [Bunge and Fitzpatrick, 1993] where we observe each of the balls with probability p which might be close to 0.
- Note that even though the full distribution of colors (call it μ) contains more information, it is impossible to consistently estimate it for any vanishing p. So instead we estimate π.

Bernoulli sampling model [Bunge and Fitzpatrick, 1993]

- Today I will be presenting my joint work with Yury Polyanskiy and Yihong Wu on profile estimation for a finite population under certain small sample regime.
- We consider an urn model with k-balls with each belonging to one of k possible types. Note that some of the colors might be empty.
- For this population we want to estimate the histogram of colors denoted by π, also known as the profile of the population [Orlitsky et al., 2005]. The j-th entry of π gives us the proportion of color groups with j-balls.
- The sampling scheme we consider is also known as the Bernoulli sampling model [Bunge and Fitzpatrick, 1993] where we observe each of the balls with probability p which might be close to 0.
- Note that even though the full distribution of colors (call it μ) contains more information, it is impossible to consistently estimate it for any vanishing p. So instead we estimate π.

Bernoulli sampling model [Bunge and Fitzpatrick, 1993]

- Today I will be presenting my joint work with Yury Polyanskiy and Yihong Wu on profile estimation for a finite population under certain small sample regime.
- We consider an urn model with k-balls with each belonging to one of k possible types. Note that some of the colors might be empty.
- For this population we want to estimate the histogram of colors denoted by π, also known as the profile of the population [Orlitsky et al., 2005]. The j-th entry of π gives us the proportion of color groups with j-balls.
- The sampling scheme we consider is also known as the Bernoulli sampling model [Bunge and Fitzpatrick, 1993] where we observe each of the balls with probability p which might be close to 0.
- Note that even though the full distribution of colors (call it μ) contains more information, it is impossible to consistently estimate it for any vanishing p. So instead we estimate π.

We present an example first.

- Suppose we have an urn of size 17 with 5 blue balls, 4 gray balls, 2 orange, 3 red and 3 green balls.
- Then the empirical distribution is given by 5/17,4/17,2/17,3/17,3/17.
- The profile of the colors can be seen as the empirical distribution of the color deleted version of the urn.
- For the color-deleted urn we can only say that there are two colors with 3 balls each and 3 colors with 5,4, and 2 balls without the information of which color had how many balls.
- So the profile is given by $\pi_2 = \pi_4 = \pi_5 = 1/17$ and $\pi_3 = 2/17$. Note that π_0 is given by 1-distinct colors/size of the urn.

- Suppose we have an urn of size 17 with 5 blue balls, 4 gray balls, 2 orange, 3 red and 3 green balls.
- Then the empirical distribution is given by 5/17,4/17,2/17,3/17,3/17.
- The profile of the colors can be seen as the empirical distribution of the color deleted version of the urn.
- For the color-deleted urn we can only say that there are two colors with 3 balls each and 3 colors with 5,4, and 2 balls without the information of which color had how many balls.
- So the profile is given by $\pi_2 = \pi_4 = \pi_5 = 1/17$ and $\pi_3 = 2/17$. Note that π_0 is given by 1-distinct colors/size of the urn.

- Suppose we have an urn of size 17 with 5 blue balls, 4 gray balls, 2 orange, 3 red and 3 green balls.
- Then the empirical distribution is given by 5/17,4/17,2/17,3/17,3/17.
- The profile of the colors can be seen as the empirical distribution of the color deleted version of the urn.
- For the color-deleted urn we can only say that there are two colors with 3 balls each and 3 colors with 5,4, and 2 balls without the information of which color had how many balls.
- So the profile is given by $\pi_2 = \pi_4 = \pi_5 = 1/17$ and $\pi_3 = 2/17$. Note that π_0 is given by 1-distinct colors/size of the urn.

- Suppose we have an urn of size 17 with 5 blue balls, 4 gray balls, 2 orange, 3 red and 3 green balls.
- Then the empirical distribution is given by 5/17,4/17,2/17,3/17,3/17.
- The profile of the colors can be seen as the empirical distribution of the color deleted version of the urn.
- For the color-deleted urn we can only say that there are two colors with 3 balls each and 3 colors with 5,4, and 2 balls without the information of which color had how many balls.
- So the profile is given by $\pi_2 = \pi_4 = \pi_5 = 1/17$ and $\pi_3 = 2/17$. Note that π_0 is given by 1-distinct colors/size of the urn.

- Suppose we have an urn of size 17 with 5 blue balls, 4 gray balls, 2 orange, 3 red and 3 green balls.
- Then the empirical distribution is given by 5/17,4/17,2/17,3/17,3/17.
- The profile of the colors can be seen as the empirical distribution of the color deleted version of the urn.
- For the color-deleted urn we can only say that there are two colors with 3 balls each and 3 colors with 5,4, and 2 balls without the information of which color had how many balls.
- So the profile is given by $\pi_2 = \pi_4 = \pi_5 = 1/17$ and $\pi_3 = 2/17$. Note that π_0 is given by 1-distinct colors/size of the urn.

- Various important label invariant properties, such as number of distinct types, entropy and learnble through π, so estimation of π can be considered as an important problem.
- Also if we consider the small sample regime, even though the empirical distribution μ can not be consistently estimated, as mentioned before, note that the estimation of π is still possible even in certain small sample regime which gives us useful implication for estimating relevant label-invariant properties.
- Also note that estimation of π related to the program of empirical Bayes posed by [Robbins, 1951, Robbins, 1956] in the fifty's.
- Suppose we want to estimate a linear functional of the parameters and we already know of an good estimator given the value of π. In small sample regime when π is unknown we can try to plug in the estimate of π in the expression and hope for good results.

- Various important label invariant properties, such as number of distinct types, entropy and learnble through π, so estimation of π can be considered as an important problem.
- Also if we consider the small sample regime, even though the empirical distribution μ can not be consistently estimated, as mentioned before, note that the estimation of π is still possible even in certain small sample regime which gives us useful implication for estimating relevant label-invariant properties.
- Also note that estimation of π related to the program of empirical Bayes posed by [Robbins, 1951, Robbins, 1956] in the fifty's.
- Suppose we want to estimate a linear functional of the parameters and we already know of an good estimator given the value of π. In small sample regime when π is unknown we can try to plug in the estimate of π in the expression and hope for good results.

- Various important label invariant properties, such as number of distinct types, entropy and learnble through π, so estimation of π can be considered as an important problem.
- Also if we consider the small sample regime, even though the empirical distribution μ can not be consistently estimated, as mentioned before, note that the estimation of π is still possible even in certain small sample regime which gives us useful implication for estimating relevant label-invariant properties.
- Also note that estimation of π related to the program of empirical Bayes posed by [Robbins, 1951, Robbins, 1956] in the fifty's.
- Suppose we want to estimate a linear functional of the parameters and we already know of an good estimator given the value of π. In small sample regime when π is unknown we can try to plug in the estimate of π in the expression and hope for good results.

- Various important label invariant properties, such as number of distinct types, entropy and learnble through π, so estimation of π can be considered as an important problem.
- Also if we consider the small sample regime, even though the empirical distribution μ can not be consistently estimated, as mentioned before, note that the estimation of π is still possible even in certain small sample regime which gives us useful implication for estimating relevant label-invariant properties.
- Also note that estimation of π related to the program of empirical Bayes posed by [Robbins, 1951, Robbins, 1956] in the fifty's.
- Suppose we want to estimate a linear functional of the parameters and we already know of an good estimator given the value of π. In small sample regime when π is unknown we can try to plug in the estimate of π in the expression and hope for good results.

Relation to previous work

- The problem of estimating distinct elements (i.e. the problem of estimating π_0) has long line of work tracking back to [Bunge and Fitzpatrick, 1993, Charikar et al., 2000, Raskhodnikova et al., 2009, Valiant, 2011, Wu and Yang, 2018, ...].
- [Wu and Yang, 2018] has shown that for small sample regime $p = \omega\left(\frac{1}{\log k}\right)$ the rate of estimation is polynomially small in k.
- Estimation of π_0 is not possible for $p = \mathcal{O}\left(\frac{1}{\log k}\right)$.
- Our result refines this and shows that the polynomial rate is achievable for other atoms π_m for $m = o(\log k)$.
- Although when m is constnt multiple of $\log k$ estimation of π_m is hard and the rate is as big as $\Omega_p\left(\frac{1}{(\log k)^2}\right)$.

- The problem of estimating distinct elements (i.e. the problem of estimating π_0) has long line of work tracking back to [Bunge and Fitzpatrick, 1993, Charikar et al., 2000, Raskhodnikova et al., 2009, Valiant, 2011, Wu and Yang, 2018, ...].
- [Wu and Yang, 2018] has shown that for small sample regime $p = \omega\left(\frac{1}{\log k}\right)$ the rate of estimation is polynomially small in k.
- Estimation of π_0 is not possible for $p = \mathcal{O}\left(\frac{1}{\log k}\right)$.
- Our result refines this and shows that the polynomial rate is achievable for other atoms π_m for $m = o(\log k)$.
- Although when m is constnt multiple of $\log k$ estimation of π_m is hard and the rate is as big as $\Omega_p\left(\frac{1}{(\log k)^2}\right)$.

- The problem of estimating distinct elements (i.e. the problem of estimating π_0) has long line of work tracking back to [Bunge and Fitzpatrick, 1993, Charikar et al., 2000, Raskhodnikova et al., 2009, Valiant, 2011, Wu and Yang, 2018, ...].
- [Wu and Yang, 2018] has shown that for small sample regime $p = \omega\left(\frac{1}{\log k}\right)$ the rate of estimation is polynomially small in k.
- Estimation of π_0 is not possible for $p = \mathcal{O}\left(\frac{1}{\log k}\right)$.
- Our result refines this and shows that the polynomial rate is achievable for other atoms π_m for $m = o(\log k)$.
- Although when m is constnt multiple of $\log k$ estimation of π_m is hard and the rate is as big as $\Omega_p\left(\frac{1}{(\log k)^2}\right)$.

- The problem of estimating distinct elements (i.e. the problem of estimating π_0) has long line of work tracking back to [Bunge and Fitzpatrick, 1993, Charikar et al., 2000, Raskhodnikova et al., 2009, Valiant, 2011, Wu and Yang, 2018, ...].
- [Wu and Yang, 2018] has shown that for small sample regime $p = \omega\left(\frac{1}{\log k}\right)$ the rate of estimation is polynomially small in k.
- Estimation of π_0 is not possible for $p = \mathcal{O}\left(\frac{1}{\log k}\right)$.
- Our result refines this and shows that the polynomial rate is achievable for other atoms π_m for $m = o(\log k)$.
- Although when m is constnt multiple of $\log k$ estimation of π_m is hard and the rate is as big as $\Omega_p\left(\frac{1}{(\log k)^2}\right)$.

- The problem of estimating distinct elements (i.e. the problem of estimating π_0) has long line of work tracking back to [Bunge and Fitzpatrick, 1993, Charikar et al., 2000, Raskhodnikova et al., 2009, Valiant, 2011, Wu and Yang, 2018, ...].
- [Wu and Yang, 2018] has shown that for small sample regime $p = \omega\left(\frac{1}{\log k}\right)$ the rate of estimation is polynomially small in k.
- Estimation of π_0 is not possible for $p = \mathcal{O}\left(\frac{1}{\log k}\right)$.
- Our result refines this and shows that the polynomial rate is achievable for other atoms π_m for $m = o(\log k)$.
- Although when m is constnt multiple of $\log k$ estimation of π_m is hard and the rate is as big as $\Omega_p\left(\frac{1}{(\log k)^2}\right)$.

- For our paper we focus mostly on estimating π in ℓ_1 norm.
- There exist a long line of work about estimating sorted version of μ (call it μ arrow).
- We note that Valiant and Valiant showed that for general population (where the population size might be infinite) we can actually estimate the sorted version μ arrow with risk of order $\frac{1}{\sqrt{\log k}}$ when the sample size is of the order k.
- This means from the relation between π and the sorted version of μ the best possible rate of estimating π we could extract is $O\left(\frac{1}{\sqrt{\log k}}\right)$.
- We show that when the population is finite, we can do much better and the existing rate that could be extracted is loose by a square root factor.

- For our paper we focus mostly on estimating π in ℓ_1 norm.
- There exist a long line of work about estimating sorted version of μ (call it μ arrow).
- We note that Valiant and Valiant showed that for general population (where the population size might be infinite) we can actually estimate the sorted version μ arrow with risk of order $\frac{1}{\sqrt{\log k}}$ when the sample size is of the order k.
- This means from the relation between π and the sorted version of μ the best possible rate of estimating π we could extract is $O\left(\frac{1}{\sqrt{\log k}}\right)$.
- We show that when the population is finite, we can do much better and the existing rate that could be extracted is loose by a square root factor.

- For our paper we focus mostly on estimating π in ℓ_1 norm.
- There exist a long line of work about estimating sorted version of μ (call it μ arrow).
- We note that Valiant and Valiant showed that for general population (where the population size might be infinite) we can actually estimate the sorted version μ arrow with risk of order $\frac{1}{\sqrt{\log k}}$ when the sample size is of the order k.
- This means from the relation between π and the sorted version of μ the best possible rate of estimating π we could extract is $O\left(\frac{1}{\sqrt{\log k}}\right)$.
- We show that when the population is finite, we can do much better and the existing rate that could be extracted is loose by a square root factor.

- For our paper we focus mostly on estimating π in ℓ_1 norm.
- There exist a long line of work about estimating sorted version of μ (call it μ arrow).
- We note that Valiant and Valiant showed that for general population (where the population size might be infinite) we can actually estimate the sorted version μ arrow with risk of order $\frac{1}{\sqrt{\log k}}$ when the sample size is of the order k.
- This means from the relation between π and the sorted version of μ the best possible rate of estimating π we could extract is $O\left(\frac{1}{\sqrt{\log k}}\right)$.
- We show that when the population is finite, we can do much better and the existing rate that could be extracted is loose by a square root factor.

- For our paper we focus mostly on estimating π in ℓ_1 norm.
- There exist a long line of work about estimating sorted version of μ (call it μ arrow).
- We note that Valiant and Valiant showed that for general population (where the population size might be infinite) we can actually estimate the sorted version μ arrow with risk of order $\frac{1}{\sqrt{\log k}}$ when the sample size is of the order k.
- This means from the relation between π and the sorted version of μ the best possible rate of estimating π we could extract is $O\left(\frac{1}{\sqrt{\log k}}\right)$.
- We show that when the population is finite, we can do much better and the existing rate that could be extracted is loose by a square root factor.

Our main result is the following:

- Given any p the upper bound for minimax risk is constant multiple of $\frac{1}{p \log k}$ with the upper bound being achieved by minimum distance estimator that comes from solving a linear program and can be calculated in polynomial time.
- The lower bound is also constant mulple of $\frac{1}{p \log k}$ until p becomes smaller than $\frac{1}{\sqrt{\log k}}$.
- This shows that in constant p regime we get the exact minimax rate of $\frac{1}{\log k}$ that is an improvement over existing rate.
- This also shows that in vanishing p regime of $\omega\left(\frac{1}{\log k}\right)$ consistent estimation of π is possible.

Our main result is the following:

- Given any p the upper bound for minimax risk is constant multiple of $\frac{1}{p \log k}$ with the upper bound being achieved by minimum distance estimator that comes from solving a linear program and can be calculated in polynomial time.
- The lower bound is also constant mulple of $\frac{1}{p \log k}$ until p becomes smaller than $\frac{1}{\sqrt{\log k}}$.
- This shows that in constant p regime we get the exact minimax rate of $\frac{1}{\log k}$ that is an improvement over existing rate.
- This also shows that in vanishing p regime of $\omega\left(\frac{1}{\log k}\right)$ consistent estimation of π is possible.

Our main result is the following:

- Given any p the upper bound for minimax risk is constant multiple of $\frac{1}{p\log k}$ with the upper bound being achieved by minimum distance estimator that comes from solving a linear program and can be calculated in polynomial time.
- The lower bound is also constant mulple of $\frac{1}{p\log k}$ until p becomes smaller than $\frac{1}{\sqrt{\log k}}$.
- This shows that in constant p regime we get the exact minimax rate of $\frac{1}{\log k}$ that is an improvement over existing rate.
- This also shows that in vanishing p regime of $\omega\left(\frac{1}{\log k}\right)$ consistent estimation of π is possible.

Our main result is the following:

- Given any p the upper bound for minimax risk is constant multiple of $\frac{1}{p \log k}$ with the upper bound being achieved by minimum distance estimator that comes from solving a linear program and can be calculated in polynomial time.
- The lower bound is also constant mulple of $\frac{1}{p \log k}$ until p becomes smaller than $\frac{1}{\sqrt{\log k}}$.
- This shows that in constant p regime we get the exact minimax rate of $\frac{1}{\log k}$ that is an improvement over existing rate.
- This also shows that in vanishing p regime of $\omega\left(\frac{1}{\log k}\right)$ consistent estimation of π is possible.

The connection of minimum distance estimation with our minimax problem is the following.

- Consider a more general version of the set-up with parameter space Θ and family of distribution $\{P_\theta, \theta \in \Theta\}$ with semi-norm ρ.
- We want to analyze the risk $R(k)$ for some general semi-norm d. under some weight constraints and based on the data X_j distributed as P_{θ_j}.
- Then the empirical distribution of the sample $\hat{\nu}$ has expectation equal to πP.
- This motivates the following estimator of minimum distance type.

The connection of minimum distance estimation with our minimax problem is the following.

- Consider a more general version of the set-up with parameter space Θ and family of distribution $\{P_\theta, \theta \in \Theta\}$ with semi-norm ρ.
- We want to analyze the risk $R(k)$ for some general semi-norm d. under some weight constraints and based on the data X_j distributed as P_{θ_j}.
- Then the empirical distribution of the sample $\hat{\nu}$ has expectation equal to πP.
- This motivates the following estimator of minimum distance type.

The connection of minimum distance estimation with our minimax problem is the following.

- Consider a more general version of the set-up with parameter space Θ and family of distribution $\{P_\theta, \theta \in \Theta\}$ with semi-norm ρ.
- We want to analyze the risk $R(k)$ for some general semi-norm d. under some weight constraints and based on the data X_j distributed as P_{θ_j}.
- Then the empirical distribution of the sample $\hat{\nu}$ has expectation equal to πP.
- This motivates the following estimator of minimum distance type.

The connection of minimum distance estimation with our minimax problem is the following.

- Consider a more general version of the set-up with parameter space Θ and family of distribution $\{P_\theta, \theta \in \Theta\}$ with semi-norm ρ.
- We want to analyze the risk $R(k)$ for some general semi-norm d. under some weight constraints and based on the data X_j distributed as P_{θ_j}.
- Then the empirical distribution of the sample $\hat{\nu}$ has expectation equal to πP.
- This motivates the following estimator of minimum distance type.

- Next we note that if we have some additional concentration bound on ν around πP then it translate into upeer bound on the risk based on the follwing linear program $\delta(t)$.
- Note that when ρ is taken to be the total variation distance then we also get lower bound based on the same linear program $\delta(t)$.
- When these bounds match we get the optimal rate for the risk.

- Next we note that if we have some additional concentration bound on ν around πP then it translate into upeer bound on the risk based on the follwing linear program $\delta(t)$.
- Note that when ρ is taken to be the total variation distance then we also get lower bound based on the same linear program $\delta(t)$.
- When these bounds match we get the optimal rate for the risk.

- Next we note that if we have some additional concentration bound on ν around πP then it translate into upeer bound on the risk based on the follwing linear program $\delta(t)$.
- Note that when ρ is taken to be the total variation distance then we also get lower bound based on the same linear program $\delta(t)$.
- When these bounds match we get the optimal rate for the risk.

- For the bernoulli sampling problem the family of distribution is given by the Binomial Markov kernel P.
- Also the corresponding linear function is given by δ_{TV} and we get the following relation.
- Note that the value of $\delta_{\mathrm{TV}}(t)$ is of logarithmic order which makes the upper and lower bounds similar for constant p regime.

- For the bernoulli sampling problem the family of distribution is given by the Binomial Markov kernel P.
- Also the corresponding linear function is given by δ_{TV} and we get the following relation.
- Note that the value of $\delta_{\mathrm{TV}}(t)$ is of logarithmic order which makes the upper and lower bounds similar for constant p regime.

- For the bernoulli sampling problem the family of distribution is given by the Binomial Markov kernel P.
- Also the corresponding linear function is given by δ_{TV} and we get the following relation.
- Note that the value of $\delta_{\mathrm{TV}}(t)$ is of logarithmic order which makes the upper and lower bounds similar for constant p regime.

- We note that bounding $\delta_{\mathrm{TV}}(t)$ is difficult as it involves optimizing over set of probability mass functions.
- Instead of bounding $\delta_{\mathrm{TV}}(t)$ we consider another linear program $\delta_*(t)$ that is defined via generating functions.
- Given any function g we first define its A norm, which is given by the sum of absolute values of the functions power series coefficients.
- Define $\delta_*(t)$ as the suppremum of the A norm of all analytic functions f whose derivative's A norm is bounded by 1 and the A norm of p-transform f_p, which is given by f of $\bar{p} + pz$, bounded by t.
- Then we can show that the new linear program differs from $\delta_{\mathrm{TV}}(t)$ by polynomially small quantity in t. So it is sufficient to work with $\delta_*(t)$ instead.
- Next we bound $\delta_*(t)$ using complex analytic techniques.

- We note that bounding $\delta_{\mathrm{TV}}(t)$ is difficult as it involves optimizing over set of probability mass functions.
- Instead of bounding $\delta_{\mathrm{TV}}(t)$ we consider another linear program $\delta_*(t)$ that is defined via generating functions.
- Given any function g we first define its A norm, which is given by the sum of absolute values of the functions power series coefficients.
- Define $\delta_*(t)$ as the suppremum of the A norm of all analytic functions f whose derivative's A norm is bounded by 1 and the A norm of p-transform f_p, which is given by f of $\bar{p} + pz$, bounded by t.
- Then we can show that the new linear program differs from $\delta_{\mathrm{TV}}(t)$ by polynomially small quantity in t. So it is sufficient to work with $\delta_*(t)$ instead.
- Next we bound $\delta_*(t)$ using complex analytic techniques.

- We note that bounding $\delta_{\mathrm{TV}}(t)$ is difficult as it involves optimizing over set of probability mass functions.
- Instead of bounding $\delta_{\mathrm{TV}}(t)$ we consider another linear program $\delta_*(t)$ that is defined via generating functions.
- Given any function g we first define its A norm, which is given by the sum of absolute values of the functions power series coefficients.
- Define $\delta_*(t)$ as the suppremum of the A norm of all analytic functions f whose derivative's A norm is bounded by 1 and the A norm of p-transform f_p, which is given by f of $\bar{p} + pz$, bounded by t.
- Then we can show that the new linear program differs from $\delta_{\mathrm{TV}}(t)$ by polynomially small quantity in t. So it is sufficient to work with $\delta_*(t)$ instead.
- Next we bound $\delta_*(t)$ using complex analytic techniques.

- We note that bounding $\delta_{\mathrm{TV}}(t)$ is difficult as it involves optimizing over set of probability mass functions.
- Instead of bounding $\delta_{\mathrm{TV}}(t)$ we consider another linear program $\delta_*(t)$ that is defined via generating functions.
- Given any function g we first define its A norm, which is given by the sum of absolute values of the functions power series coefficients.
- Define $\delta_*(t)$ as the suppremum of the A norm of all analytic functions f whose derivative's A norm is bounded by 1 and the A norm of p-transform f_p, which is given by f of $\bar{p} + pz$, bounded by t.
- Then we can show that the new linear program differs from $\delta_{\mathrm{TV}}(t)$ by polynomially small quantity in t. So it is sufficient to work with $\delta_*(t)$ instead.
- Next we bound $\delta_*(t)$ using complex analytic techniques.

- We note that bounding $\delta_{\mathrm{TV}}(t)$ is difficult as it involves optimizing over set of probability mass functions.
- Instead of bounding $\delta_{\mathrm{TV}}(t)$ we consider another linear program $\delta_*(t)$ that is defined via generating functions.
- Given any function g we first define its A norm, which is given by the sum of absolute values of the functions power series coefficients.
- Define $\delta_*(t)$ as the suppremum of the A norm of all analytic functions f whose derivative's A norm is bounded by 1 and the A norm of p-transform f_p, which is given by f of $\bar{p} + pz$, bounded by t.
- Then we can show that the new linear program differs from $\delta_{\mathrm{TV}}(t)$ by polynomially small quantity in t. So it is sufficient to work with $\delta_*(t)$ instead.
- Next we bound $\delta_*(t)$ using complex analytic techniques.

- We note that bounding $\delta_{\mathrm{TV}}(t)$ is difficult as it involves optimizing over set of probability mass functions.
- Instead of bounding $\delta_{\mathrm{TV}}(t)$ we consider another linear program $\delta_*(t)$ that is defined via generating functions.
- Given any function g we first define its A norm, which is given by the sum of absolute values of the functions power series coefficients.
- Define $\delta_*(t)$ as the suppremum of the A norm of all analytic functions f whose derivative's A norm is bounded by 1 and the A norm of p-transform f_p, which is given by f of $\bar{p} + pz$, bounded by t.
- Then we can show that the new linear program differs from $\delta_{\mathrm{TV}}(t)$ by polynomially small quantity in t. So it is sufficient to work with $\delta_*(t)$ instead.
- Next we bound $\delta_*(t)$ using complex analytic techniques.

- For upper bound on $\delta_*(t)$ we first write down the objective function in terms of its power series coefficients.
- For the tail part of the sum of the coefficients after term $\log(1/t)$ we use constraints on derivative to get $O_p\left(\frac{1}{\log k}\right)$ bound.
- For the first part of the sum we do a term by term analysis. For each term in the summand we write down LP $\delta_m(t)$ that maximizes the value of the coefficient over the same set of constraints as in $\delta_*(t)$. The sum of these LPs bound the original LP from above.
- Then it remains to show that each of these new LP's are negligible compared to order of $\frac{1}{\log(1/t)}$.
- Combining these we finally get the desired logarithmic bound. We use Hadamard's three line theorem for bounding $\delta_m(t)$.

- For upper bound on $\delta_*(t)$ we first write down the objective function in terms of its power series coefficients.
- For the tail part of the sum of the coefficients after term $\log(1/t)$ we use constraints on derivative to get $O_p\left(\frac{1}{\log k}\right)$ bound.
- For the first part of the sum we do a term by term analysis. For each term in the summand we write down LP $\delta_m(t)$ that maximizes the value of the coefficient over the same set of constraints as in $\delta_*(t)$. The sum of these LPs bound the original LP from above.
- Then it remains to show that each of these new LP's are negligible compared to order of $\frac{1}{\log(1/t)}$.
- Combining these we finally get the desired logarithmic bound. We use Hadamard's three line theorem for bounding $\delta_m(t)$.

- For upper bound on $\delta_*(t)$ we first write down the objective function in terms of its power series coefficients.
- For the tail part of the sum of the coefficients after term $\log(1/t)$ we use constraints on derivative to get $O_p\left(\frac{1}{\log k}\right)$ bound.
- For the first part of the sum we do a term by term analysis. For each term in the summand we write down LP $\delta_m(t)$ that maximizes the value of the coefficient over the same set of constraints as in $\delta_*(t)$. The sum of these LPs bound the original LP from above.
- Then it remains to show that each of these new LP's are negligible compared to order of $\frac{1}{\log(1/t)}$.
- Combining these we finally get the desired logarithmic bound. We use Hadamard's three line theorem for bounding $\delta_m(t)$.

- For upper bound on $\delta_*(t)$ we first write down the objective function in terms of its power series coefficients.
- For the tail part of the sum of the coefficients after term $\log(1/t)$ we use constraints on derivative to get $O_p\left(\frac{1}{\log k}\right)$ bound.
- For the first part of the sum we do a term by term analysis. For each term in the summand we write down LP $\delta_m(t)$ that maximizes the value of the coefficient over the same set of constraints as in $\delta_*(t)$. The sum of these LPs bound the original LP from above.
- Then it remains to show that each of these new LP's are negligible compared to order of $\frac{1}{\log(1/t)}$.
- Combining these we finally get the desired logarithmic bound. We use Hadamard's three line theorem for bounding $\delta_m(t)$.

- For upper bound on $\delta_*(t)$ we first write down the objective function in terms of its power series coefficients.
- For the tail part of the sum of the coefficients after term $\log(1/t)$ we use constraints on derivative to get $O_p\left(\frac{1}{\log k}\right)$ bound.
- For the first part of the sum we do a term by term analysis. For each term in the summand we write down LP $\delta_m(t)$ that maximizes the value of the coefficient over the same set of constraints as in $\delta_*(t)$. The sum of these LPs bound the original LP from above.
- Then it remains to show that each of these new LP's are negligible compared to order of $\frac{1}{\log(1/t)}$.
- Combining these we finally get the desired logarithmic bound. We use Hadamard's three line theorem for bounding $\delta_m(t)$.

- Note that for any analytic function f we can define its suppremum norm over set C. Call it $\|f\|_{H^\infty(C)}$.
- Consider unit disk D and horodisks disks D_p given by $\bar{p} + pD$.
- For any two constants q, q_1 in unit interval we can define horodisks D_q and D_{q_1} inside the unit disk.
- Then Hadamard's three line theorem says that the suppremum of any analytic function f over the middle most horodisk can be bounded from above in terms of the suppremum of the same function over the unit disk and the suppremum over the innermost horodisk.

- Note that for any analytic function f we can define its suppremum norm over set C. Call it $\|f\|_{H^\infty(C)}$.
- Consider unit disk D and horodisks disks D_p given by $\bar{p} + pD$.
- For any two constants q, q_1 in unit interval we can define horodisks D_q and D_{q_1} inside the unit disk.
- Then Hadamard's three line theorem says that the suppremum of any analytic function f over the middle most horodisk can be bounded from above in terms of the suppremum of the same function over the unit disk and the suppremum over the innermost horodisk.

- Note that for any analytic function f we can define its suppremum norm over set C. Call it $\|f\|_{H^\infty(C)}$.
- Consider unit disk D and horodisks disks D_p given by $\bar{p} + pD$.
- For any two constants q, q_1 in unit interval we can define horodisks D_q and D_{q_1} inside the unit disk.
- Then Hadamard's three line theorem says that the suppremum of any analytic function f over the middle most horodisk can be bounded from above in terms of the suppremum of the same function over the unit disk and the suppremum over the innermost horodisk.

- Note that for any analytic function f we can define its suppremum norm over set C. Call it $\|f\|_{H^\infty(C)}$.
- Consider unit disk D and horodisks disks D_p given by $\bar{p} + pD$.
- For any two constants q, q_1 in unit interval we can define horodisks D_q and D_{q_1} inside the unit disk.
- Then Hadamard's three line theorem says that the suppremum of any analytic function f over the middle most horodisk can be bounded from above in terms of the suppremum of the same function over the unit disk and the suppremum over the innermost horodisk.

- Using ordering between A norm and the sup norm over the unit disk, for each feasible function f for $\delta_m(t)$ we can bound the suppremum of f over unit disk D and horodisk D_p by using the constraints on $\delta_m(t)$, and Cauchy integral formula .
- We can also use Cauchy's integral formula to bound the objective function of $\delta_m(t)$ in terms of suppremum over horodisk $D_{1/2}$.
- For $p < \frac{1}{2}$ we get the inclusion of horodisks as required for applying Hadamard's theorem and we bound the suppremum over the horodisk $D_{1/2}$ in terms of suppremum over D_p and suppremum over unit disk D.
- For $p > \frac{1}{2}$ we get direct bound on the objective function based on horodisk D_p.
- Finally we sum all the bounds on $\delta_m(t)$ to arrive at order $\frac{1}{\log(1/t)}$ bound, as desired.

- Using ordering between A norm and the sup norm over the unit disk, for each feasible function f for $\delta_m(t)$ we can bound the suppremum of f over unit disk D and horodisk D_p by using the constraints on $\delta_m(t)$, and Cauchy integral formula .
- We can also use Cauchy's integral formula to bound the objective function of $\delta_m(t)$ in terms of suppremum over horodisk $D_{1/2}$.
- For $p < \frac{1}{2}$ we get the inclusion of horodisks as required for applying Hadamard's theorem and we bound the suppremum over the horodisk $D_{1/2}$ in terms of suppremum over D_p and suppremum over unit disk D.
- For $p > \frac{1}{2}$ we get direct bound on the objective function based on horodisk D_p.
- Finally we sum all the bounds on $\delta_m(t)$ to arrive at order $\frac{1}{\log(1/t)}$ bound, as desired.

- Using ordering between A norm and the sup norm over the unit disk, for each feasible function f for $\delta_m(t)$ we can bound the suppremum of f over unit disk D and horodisk D_p by using the constraints on $\delta_m(t)$, and Cauchy integral formula .
- We can also use Cauchy's integral formula to bound the objective function of $\delta_m(t)$ in terms of suppremum over horodisk $D_{1/2}$.
- For $p < \frac{1}{2}$ we get the inclusion of horodisks as required for applying Hadamard's theorem and we bound the suppremum over the horodisk $D_{1/2}$ in terms of suppremum over D_p and suppremum over unit disk D.
- For $p > \frac{1}{2}$ we get direct bound on the objective function based on horodisk D_p.
- Finally we sum all the bounds on $\delta_m(t)$ to arrive at order $\frac{1}{\log(1/t)}$ bound, as desired.

- Using ordering between A norm and the sup norm over the unit disk, for each feasible function f for $\delta_m(t)$ we can bound the suppremum of f over unit disk D and horodisk D_p by using the constraints on $\delta_m(t)$, and Cauchy integral formula .
- We can also use Cauchy's integral formula to bound the objective function of $\delta_m(t)$ in terms of suppremum over horodisk $D_{1/2}$.
- For $p < \frac{1}{2}$ we get the inclusion of horodisks as required for applying Hadamard's theorem and we bound the suppremum over the horodisk $D_{1/2}$ in terms of suppremum over D_p and suppremum over unit disk D.
- For $p > \frac{1}{2}$ we get direct bound on the objective function based on horodisk D_p.
- Finally we sum all the bounds on $\delta_m(t)$ to arrive at order $\frac{1}{\log(1/t)}$ bound, as desired.

- Using ordering between A norm and the sup norm over the unit disk, for each feasible function f for $\delta_m(t)$ we can bound the suppremum of f over unit disk D and horodisk D_p by using the constraints on $\delta_m(t)$, and Cauchy integral formula .
- We can also use Cauchy's integral formula to bound the objective function of $\delta_m(t)$ in terms of suppremum over horodisk $D_{1/2}$.
- For $p < \frac{1}{2}$ we get the inclusion of horodisks as required for applying Hadamard's theorem and we bound the suppremum over the horodisk $D_{1/2}$ in terms of suppremum over D_p and suppremum over unit disk D.
- For $p > \frac{1}{2}$ we get direct bound on the objective function based on horodisk D_p.
- Finally we sum all the bounds on $\delta_m(t)$ to arrive at order $\frac{1}{\log(1/t)}$ bound, as desired.

- Finally we end the presentation with sketch of proof of the lower bound on $\delta_*(t)$.
- Using ordering between A norm and the sup norm over the unit disk we relax constraints and objective function of $\delta_*(t)$ to come up with a different linear program δ_{H^∞} that is easier to solve.
- It turns out that the value of the new linear program is also of order $\frac{1}{\log(1/t)}$. We find out the function that achieves it.
- Then we modify the solution via linear transform to get feasible solution to $\delta_*(t)$.
- We relate the coefficients of the modified function to the Laguerre polynomials and by using the properties of Laguerre polynomials we get the result.

- Finally we end the presentation with sketch of proof of the lower bound on $\delta_*(t)$.
- Using ordering between A norm and the sup norm over the unit disk we relax constraints and objective function of $\delta_*(t)$ to come up with a different linear program δ_{H^∞} that is easier to solve.
- It turns out that the value of the new linear program is also of order $\frac{1}{\log(1/t)}$. We find out the function that achieves it.
- Then we modify the solution via linear transform to get feasible solution to $\delta_*(t)$.
- We relate the coefficients of the modified function to the Laguerre polynomials and by using the properties of Laguerre polynomials we get the result.

- Finally we end the presentation with sketch of proof of the lower bound on $\delta_*(t)$.
- Using ordering between A norm and the sup norm over the unit disk we relax constraints and objective function of $\delta_*(t)$ to come up with a different linear program δ_{H^∞} that is easier to solve.
- It turns out that the value of the new linear program is also of order $\frac{1}{\log(1/t)}$. We find out the function that achieves it.
- Then we modify the solution via linear transform to get feasible solution to $\delta_*(t)$.
- We relate the coefficients of the modified function to the Laguerre polynomials and by using the properties of Laguerre polynomials we get the result.

- Finally we end the presentation with sketch of proof of the lower bound on $\delta_*(t)$.
- Using ordering between A norm and the sup norm over the unit disk we relax constraints and objective function of $\delta_*(t)$ to come up with a different linear program δ_{H^∞} that is easier to solve.
- It turns out that the value of the new linear program is also of order $\frac{1}{\log(1/t)}$. We find out the function that achieves it.
- Then we modify the solution via linear transform to get feasible solution to $\delta_*(t)$.
- We relate the coefficients of the modified function to the Laguerre polynomials and by using the properties of Laguerre polynomials we get the result.

- Finally we end the presentation with sketch of proof of the lower bound on $\delta_*(t)$.
- Using ordering between A norm and the sup norm over the unit disk we relax constraints and objective function of $\delta_*(t)$ to come up with a different linear program δ_{H^∞} that is easier to solve.
- It turns out that the value of the new linear program is also of order $\frac{1}{\log(1/t)}$. We find out the function that achieves it.
- Then we modify the solution via linear transform to get feasible solution to $\delta_*(t)$.
- We relate the coefficients of the modified function to the Laguerre polynomials and by using the properties of Laguerre polynomials we get the result.

Bunge, J. and Fitzpatrick, M. (1993).
Estimating the number of species: a review.
Journal of the American Statistical Association, 88(421):364–373.

Charikar, M., Chaudhuri, S., Motwani, R., and Narasayya, V. (2000).
Towards estimation error guarantees for distinct values.
In *Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pages 268–279. ACM.

Orlitsky, A., Santhanam, N., Viswanathan, K., and Zhang, J. (2005).
Convergence of profile based estimators.
In *Proc. 2005 IEEE Int. Symp. Inf. Theory (ISIT)*, pages 1843–1847. IEEE.

Raskhodnikova, S., Ron, D., Shpilka, A., and Smith, A. (2009).
Strong lower bounds for approximating distribution support size and the distinct elements problem.
SIAM Journal on Computing, 39(3):813–842.

Robbins, H. (1951).
Asymptotically subminimax solutions of compound statistical decision problems.
In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. The Regents of the University of California.

Robbins, H. (1956).
An empirical bayes approach to statistics.
In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Valiant, P. (2011).
Testing symmetric properties of distributions.
SIAM Journal on Computing, 40(6):1927–1968.

Wu, Y. and Yang, P. (2018).
Sample complexity of the distinct element problem.
Mathematical Statistics and Learning, 1(1):37–72.