Self-regularizing Property of Nonparametric Maximum Likelihood Estimator in Mixture Models

Yury Polyanskiy

Department of EECS
MIT

Joint work with
Yihong Wu (Yale)

ICE Speaker Series, ICE, TU-Munich, 9 Dec 2020

Mixture models

- Height distribution of a population

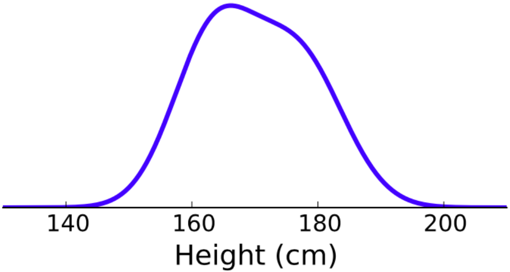


Image credit: [Hardt-Price '15]

Mixture models

- Height distribution of a population

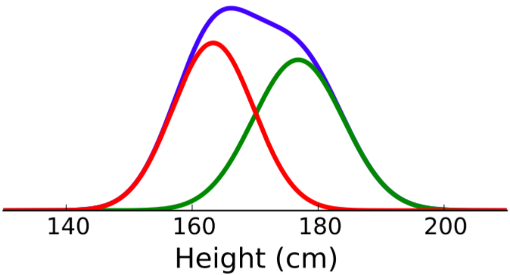


Image credit: [Hardt-Price '15]

- Model each of male and female subpopulation by a Gaussian distribution

Mixture models

- Height distribution of a population

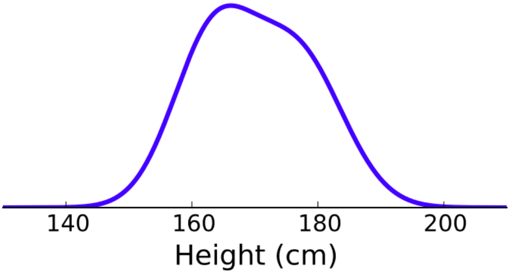


Image credit: [Hardt-Price '15]

- Model each of male and female subpopulation by a Gaussian distribution

Question

How to learn the average heights of male and female from unlabeled data?

Mixture models & empirical Bayes

HOCKEY REFERENCE

Enter Person, Team, Section, etc

Players | Teams | Seasons | Leaders | NHL Scores | Playoffs | Stathead | Newsletter | Full Site Menu Below

Nikita Kucherov

Position: RW • **Shoots**: Left
5-11, 178lb (180cm, 80kg)
Team: Tampa Bay Lightning
Born: June 17, 1993 (Age: 27-167d) in Maykop, Russian Federation
Draft: Tampa Bay, 2nd round (58th overall), 2011 NHL Entry
Amateur Teams: Quebec Remparts, Rouyn-Noranda Huskies
Pronunciation: \nih-KEE-tuh KOO-chuhr-awv\

2020 Cup Winner | 4x All Star | 2018-19 Hart | 2018-19 Pearson | 2018-19 Ross

56 86

SUMMARY	GP	G	A	PTS	+/-	PS	PIM	SH	GWG	TOI	CF%	oZS%
2019-20	68	33	52	85	26	10.5	38	210	6	18:52	56.1	59.8
Career	515	221	326	547	128	69.6	261	1483	37	17:58	54.5	59.5

Nikita Kucherov Overview | Game Logs ▾ | Splits ▾ | Scoring Logs ▾ | Game Finder | Advanced | 2019-20 Lightning ▾ | News

On this page:
NHL Standard | Player News | NHL Possession Metrics (Even Strength) | NHL Miscellaneous | NHL Playoffs | Playoff Overtime
Penalty Shots (4) | Other Standard | Similarity Scores | Leaderboards, Awards, & Honors | Other Links | Full Site Menu

NHL Standard

					Scoring					Goals				Assists			Shots			Ice Time									
Season	Age	Tm	Lg	GP	G	A	PTS	+/-	PIM	EV	PP	SH	GW	EV	PP	SH	S	S%	TSA	TOI	ATOI	FOW	FOL	FO%	BLK	HIT	TK	GV	Awards
2013-14	20	TBL	NHL	52	9	9	18	3	14	6	3	0	3	8	1	0	102	8.8	189	682	13:07	1	0	100.0	19	15	17	17	
2014-15	21	TBL	NHL	82	29	36	65	**38**	37	27	2	0	2	23	13	0	191	15.2	327	1226	14:57	0	2	0.0	28	65	22	48	AS-6,Selke-30
2015-16	22	TBL	NHL	77	30	36	66	9	30	21	9	0	4	20	16	0	209	14.4	402	1402	18:13	0	1	0.0	28	61	34	58	AS-8,Byng-53
2016-17	23	TBL	NHL	74	40	45	85	13	38	23	**17**	0	7	30	15	0	246	16.3	464	1438	19:26	0	0		20	30	54	64	**AS-2**,Hart-8,Selke-37
2017-18	24	TBL	NHL	80	39	61	100	15	42	31	8	0	7	33	28	0	279	14.0	547	1586	19:49	3	2	60.0	15	31	66	79	**AS-1**,Byng-26,Hart-6
2018-19	25	TBL	NHL	82	41	**87**	**128**	24	62	26	15	0	8	54	33	0	246	16.7	490	1637	19:58	0	3	0.0	31	44	58	89	**AS-1**,**Hart-1**,**Pearson-1**,**Ross-1**
2019-20	26	TBL	NHL	68	33	52	85	26	38	29	4	0	6	31	21	0	210	15.7	386	1283	18:52	0	1	0.0	19	37	30	59	**AS-2**,Hart-13
Career		**7 yrs**	**NHL**	**515**	**221**	**326**	**547**	**128**	**261**	**163**	**58**	**0**	**37**	**199**	**127**	**0**	**1483**	**14.9**	**2805**	**9254**	**17:58**	**4**	**9**	**30.8**	**160**	**283**	**281**	**414**	

- Nikita Kucherov scored 33 goals in 2019-2020.
- How many will he score in 2020-2021?

Setup

- $\{p_\theta : \theta \in \Theta\}$: parametric family of densities
- π: mixing distribution (prior) on Θ
- mixture density:

$$p_\pi(x) \triangleq \int_\Theta p_\theta(x)\pi(d\theta)$$

Setup

- $\{p_\theta : \theta \in \Theta\}$: parametric family of densities
- π: mixing distribution (prior) on Θ
- mixture density:

$$p_\pi(x) \triangleq \int_\Theta p_\theta(x)\pi(d\theta)$$

- Goal: given sample $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} p_\pi$, learn the mixture model (e.g. estimating π or p_π)

Running example: Gaussian location mixture

- $p_\theta(x) = \varphi(x - \theta)$: density $N(\theta, 1)$, where $\varphi(x) \triangleq \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$
- mixture density = Gaussian convolution

$$p_\pi(x) = (\varphi * \pi)(x)$$

- Special case: k-component Gaussian mixture (k-GM)

$$p_\pi(x) = \sum_{i=1}^{k} w_i \varphi(x - \theta_i), \quad \pi = \sum_{i=1}^{k} w_i \delta_{\theta_i}.$$

- Major difficulty: nonconvexity of mixture likelihood (in location parameters θ_i's)
  - Expectation-Maximization: Heuristic and suffer from spurious local maxima [Jin-Zhang-Balakrishnan-Wainwright-Jordan '16]

Nonparametric approach

Three major methodologies:

1. Method of moments: [Pearson 1895]

   learn π through estimating its moments

2. Minimum-distance estimator: [Wolfowitz '57]

$$\hat{\pi} = \arg\min_{\pi} \mathsf{dist}(p_\pi, \mathsf{empirical})$$

3. Nonparametric Maximum Likelihood: [Kiefer-Wolfowitz '56]

Nonparametric approach

Three major methodologies:

1. Method of moments: [Pearson 1895]

   learn π through estimating its moments

   Tuning param: Number of moments

2. Minimum-distance estimator: [Wolfowitz '57]

   $$\hat{\pi} = \arg\min_{\pi} \mathsf{dist}(p_\pi, \mathsf{empirical})$$

   Tuning param: choice of distance

3. Nonparametric Maximum Likelihood: [Kiefer-Wolfowitz '56]
   Tuning param: NONE!

Nonparametric Maximum Likelihood Estimator (NPMLE)

Optimizing the likelihood over the space $\mathcal{M}(\Theta)$ of all priors:

$$\hat{\pi}_{\mathrm{NPMLE}} \in \arg\max_{\pi \in \mathcal{M}(\Theta)} \frac{1}{n} \sum_{i=1}^{n} \log p_\pi(x_i)$$

- k-component mixture problem is finite dim., but non-convex
- NPMLE: ∞-dimensional but convex *(overparametrization)*

Nonparametric Maximum Likelihood Estimator (NPMLE)

Optimizing the likelihood over the space $\mathcal{M}(\Theta)$ of all priors:

$$\hat{\pi}_{\text{NPMLE}} \in \arg\max_{\pi \in \mathcal{M}(\Theta)} \frac{1}{n} \sum_{i=1}^{n} \log p_\pi(x_i)$$

- k-component mixture problem is finite dim., but non-convex
- NPMLE: ∞-dimensional but convex *(overparametrization)*
- NPMLE is a form of minimum-distance estimator:

$$\hat{\pi}_{\text{NPMLE}} = \arg\min_{\pi} D(\hat{P}_n \| P_\pi) \qquad D(P||Q) = \int dP \log \frac{dP}{dQ}$$

($\hat{P}_n$ is empirical distribution of samples)

Nonparametric Maximum Likelihood Estimator (NPMLE)

Optimizing the likelihood over the space $\mathcal{M}(\Theta)$ of all priors:

$$\hat{\pi}_{\mathrm{NPMLE}} \in \arg \max_{\pi \in \mathcal{M}(\Theta)} \frac{1}{n} \sum_{i=1}^{n} \log p_\pi(x_i)$$

- k-component mixture problem is finite dim., but non-convex
- NPMLE: ∞-dimensional but convex *(overparametrization)*
- NPMLE is a form of minimum-distance estimator:

$$\hat{\pi}_{\mathrm{NPMLE}} = \arg\min_{\pi} D(\hat{P}_n \| P_\pi) \qquad D(P\|Q) = \int dP \log \frac{dP}{dQ}$$

  ($\hat{P}_n$ is empirical distribution of samples)
- NPMLE is related to rate-distortion theory (with source $\sim \hat{P}_n$):

$$\min_{\pi} D(\hat{P}_n \| P_\pi) = \min_{P_{\theta,X}: P_X = \hat{P}_n} I(\theta; X) + \frac{1}{2\sigma^2} \mathbb{E}[\|\theta - X\|^2]$$

- ... and also to entropic optimal transport [Weed-Rigollete '18]

$$\min_{\pi} D(\hat{P}_n \| P_\pi) = \min_{\pi} W_2^{(\sigma)}(\pi, \hat{P}_n)$$

Nonparametric Maximum Likelihood Estimator (NPMLE)

Optimizing the likelihood over the space $\mathcal{M}(\Theta)$ of all priors:

$$\hat{\pi}_{\mathrm{NPMLE}} \in \arg\max_{\pi\in\mathcal{M}(\Theta)} \frac{1}{n}\sum_{i=1}^{n} \log p_\pi(x_i)$$

- *Information-theoretic literature*:
  - Iterative algo [Richardson '70] (for astronomy imaging)
  - Proof of convergence and connections to Blahut-Arimoto algo [Csiszar-Tusnady '82]

Nonparametric Maximum Likelihood Estimator (NPMLE)

Optimizing the likelihood over the space $\mathcal{M}(\Theta)$ of all priors:

$$\hat{\pi}_{\mathrm{NPMLE}} \in \arg\max_{\pi \in \mathcal{M}(\Theta)} \frac{1}{n} \sum_{i=1}^{n} \log p_\pi(x_i)$$

- *Information-theoretic literature*:
  - Iterative algo [Richardson '70] (for astronomy imaging)
  - Proof of convergence and connections to Blahut-Arimoto algo [Csiszar-Tusnady '82]
- *Stats literature*: for mixture model in one dimension
  - Basic properties (existence, uniqueness, discreteness) are well understood [Simar '76, Jewell '82, Lindsay '83,'95]
  - Other kinds of iterative algorithms:
    - vertex exchange method [Böhning '81, Lindsay '83]
    - Grid-based: discretization [Koenker-Mizera '14]
  - $\sim 10^2$ papers on density estimation via NPMLE, NPMLE for empirical Bayes, shape-constrained NPMLE (Grenander)...

Pros and Cons of NPMLE

Advantages:

- Flexibility: no tuning parameters, no penalty, assumes no upper bound on the mixture order
- Computation: does not suffer from non-convexity
- Accuracy: near-parametric rate in density estimation [Zhang '09, Saha-Guntuboyina '20]

Pros and Cons of NPMLE

Advantages:

- Flexibility: no tuning parameters, no penalty, assumes no upper bound on the mixture order
- Computation: does not suffer from non-convexity
- Accuracy: near-parametric rate in density estimation [Zhang '09, Saha-Guntuboyina '20]
- Widely used in empirical Bayes and superior in both theoretical and practical performance

Pros and Cons of NPMLE

Advantages:

- Flexibility: no tuning parameters, no penalty, assumes no upper bound on the mixture order
- Computation: does not suffer from non-convexity
- Accuracy: near-parametric rate in density estimation [Zhang '09, Saha-Guntuboyina '20]
- Widely used in empirical Bayes and superior in both theoretical and practical performance

Potential issues:

- An extreme form of overparameterization
- Runs the risk of overfitting

Pros and Cons of NPMLE

Advantages:

- Flexibility: no tuning parameters, no penalty, assumes no upper bound on the mixture order
- Computation: does not suffer from non-convexity
- Accuracy: near-parametric rate in density estimation [Zhang '09, Saha-Guntuboyina '20]
- Widely used in empirical Bayes and superior in both theoretical and practical performance

Potential issues:

- An extreme form of overparameterization
- Runs the risk of overfitting

Major question

- Does NPMLE "overfit" if the data are drawn from a finite mixture?
- If data are generated from a k-GM, say, $k = 2$, what is the typical model size fitted by NPMLE?

Pros and Cons of NPMLE

Advantages:

- Flexibility: no tuning parameters, no penalty, assumes no upper bound on the mixture order
- Computation: does not suffer from non-convexity
- Accuracy: near-parametric rate in density estimation [Zhang '09, Saha-Guntuboyina '20]
- Widely used in empirical Bayes and superior in both theoretical and practical performance

Potential issues:

- An extreme form of overparameterization
- Runs the risk of overfitting

Major question

- Does NPMLE "overfit" if the data are drawn from a finite mixture?
- If data are generated from a k-GM, say, $k = 2$, what is the typical model size fitted by NPMLE?

These questions are not answered by classical theory.

Structural property of NPMLE

Optimality condition

- Objective function: $\ell(\pi) = \frac{1}{n}\sum_{i=1}^{n} \log p_\pi(x_i)$, maximized by $\hat{\pi} = \hat{\pi}_{\text{NPMLE}}$.

Optimality condition

- Objective function: $\ell(\pi) = \frac{1}{n}\sum_{i=1}^n \log p_\pi(x_i)$, maximized by $\hat{\pi} = \hat{\pi}_{\mathrm{NPMLE}}$.
- For any $\epsilon \in [0,1]$ and any $\theta \in \mathbb{R}$,

$$\ell(\hat{\pi}) \geq \ell((1-\epsilon)\hat{\pi} + \epsilon\delta_\theta) \implies \underbrace{\frac{d}{d\epsilon}\ell((1-\epsilon)\hat{\pi} + \epsilon\delta_\theta)\Big|_{\epsilon=0}}_{\frac{1}{n}\sum_{i=1}^n \frac{p_\theta(x_i)}{p_{\hat{\pi}}(x_i)} - 1} \leq 0$$

Optimality condition

- Objective function: $\ell(\pi) = \frac{1}{n}\sum_{i=1}^n \log p_\pi(x_i)$, maximized by $\hat{\pi} = \hat{\pi}_{\mathrm{NPMLE}}$.
- For any $\epsilon \in [0,1]$ and any $\theta \in \mathbb{R}$,

$$\ell(\hat{\pi}) \geq \ell((1-\epsilon)\hat{\pi} + \epsilon\delta_\theta) \implies \underbrace{\frac{d}{d\epsilon}\ell((1-\epsilon)\hat{\pi} + \epsilon\delta_\theta)\Big|_{\epsilon=0}}_{\frac{1}{n}\sum_{i=1}^n \frac{p_\theta(x_i)}{p_{\hat{\pi}}(x_i)} - 1} \leq 0$$

First-order optimality condition

$$\hat{\pi} \text{ is optimal} \iff D_{\hat{\pi}}(\theta) \triangleq \frac{1}{n}\sum_{i=1}^n \frac{p_\theta(x_i)}{p_{\hat{\pi}}(x_i)} \leq 1, \quad \forall \theta \in \mathbb{R}$$

Optimality condition

- Objective function: $\ell(\pi) = \frac{1}{n}\sum_{i=1}^n \log p_\pi(x_i)$, maximized by $\hat\pi = \hat\pi_{\mathrm{NPMLE}}$.
- For any $\epsilon \in [0,1]$ and any $\theta \in \mathbb{R}$,

$$\ell(\hat\pi) \ge \ell((1-\epsilon)\hat\pi + \epsilon\delta_\theta) \implies \underbrace{\frac{d}{d\epsilon}\ell((1-\epsilon)\hat\pi + \epsilon\delta_\theta)\Big|_{\epsilon=0}}_{\frac{1}{n}\sum_{i=1}^n \frac{p_\theta(x_i)}{p_{\hat\pi}(x_i)} - 1} \le 0$$

First-order optimality condition

$$\hat\pi \text{ is optimal} \iff D_{\hat\pi}(\theta) \triangleq \frac{1}{n}\sum_{i=1}^n \frac{p_\theta(x_i)}{p_{\hat\pi}(x_i)} \le 1, \quad \forall\theta \in \mathbb{R}$$

Consequence:

- Averaging the LHS over $\hat\pi \implies \int \hat\pi(d\theta) D_{\hat\pi}(\theta) = 1$

Optimality condition

- Objective function: $\ell(\pi) = \frac{1}{n}\sum_{i=1}^n \log p_\pi(x_i)$, maximized by $\hat\pi = \hat\pi_{\mathrm{NPMLE}}$.
- For any $\epsilon \in [0,1]$ and any $\theta \in \mathbb{R}$,

$$\ell(\hat\pi) \geq \ell((1-\epsilon)\hat\pi + \epsilon\delta_\theta) \implies \underbrace{\frac{d}{d\epsilon}\ell((1-\epsilon)\hat\pi + \epsilon\delta_\theta)\Big|_{\epsilon=0}}_{\frac{1}{n}\sum_{i=1}^n \frac{p_\theta(x_i)}{p_{\hat\pi}(x_i)} - 1} \leq 0$$

First-order optimality condition

$$\hat\pi \text{ is optimal} \iff D_{\hat\pi}(\theta) \triangleq \frac{1}{n}\sum_{i=1}^n \frac{p_\theta(x_i)}{p_{\hat\pi}(x_i)} \leq 1, \quad \forall \theta \in \mathbb{R}$$

Consequence:

- Averaging the LHS over $\hat\pi \implies \int \hat\pi(d\theta) D_{\hat\pi}(\theta) = 1$
- Thus

$$\mathrm{supp}(\hat\pi_{\mathrm{NPMLE}}) \subset \{\text{Global maximizers of } D_{\hat\pi}\} \subset \{\text{Critical points of } D_{\hat\pi}\}$$

Gaussian mixture

- Note that

$$D_{\hat{\pi}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{p_{\hat{\pi}}(x_i)}\phi(x_i - \theta) \propto \sum_{i=1}^{n} w_i \phi(x_i - \theta)$$

which is an n-GM density, with centers at the datapoints.

Gaussian mixture

- Note that

$$D_{\hat{\pi}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{p_{\hat{\pi}}(x_i)} \phi(x_i - \theta) \propto \sum_{i=1}^{n} w_i \phi(x_i - \theta)$$

which is an n-GM density, with centers at the datapoints.

- Fact: n-GM density in 1D has at most n modes [Polya-Szegö '25, Hummel-Gidas '84]

Theorem (Lindsay '83)

$\hat{\pi}_{\mathrm{NPMLE}}$ *exists and is unique and discrete with at most* n *atoms*

Gaussian mixture

- Note that

$$D_{\hat{\pi}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{p_{\hat{\pi}}(x_i)}\phi(x_i - \theta) \propto \sum_{i=1}^{n} w_i \phi(x_i - \theta)$$

which is an n-GM density, with centers at the datapoints.

- Fact: n-GM density in 1D has at most n modes [Polya-Szegö '25, Hummel-Gidas '84]

Theorem (Lindsay '83)

$\hat{\pi}_{\mathrm{NPMLE}}$ *exists and is unique and discrete with at most* n *atoms*

- This deterministic result is tight in the worst case
- In practice, model fitted by NPMLE is much simpler

Gaussian mixture

- Note that

$$D_{\hat{\pi}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{p_{\hat{\pi}}(x_i)}\phi(x_i - \theta) \propto \sum_{i=1}^{n} w_i \phi(x_i - \theta)$$

  which is an n-GM density, with centers at the datapoints.
- Fact: n-GM density in 1D has at most n modes [Polya-Szegö '25, Hummel-Gidas '84]

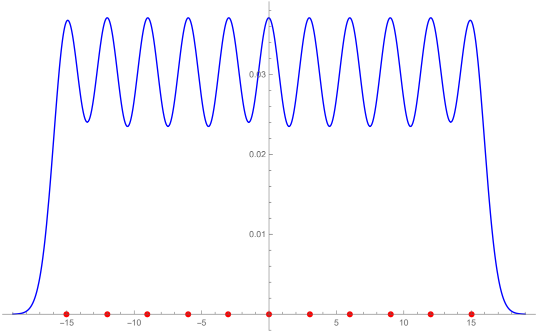
Theorem (Lindsay '83)

$\hat{\pi}_{\text{NPMLE}}$ *exists and is unique and discrete with at most n atoms*

- This deterministic result is tight in the worst case
- In practice, model fitted by NPMLE is much simpler
- Question: can we improve it when $x_1, \ldots, x_n$ are random?

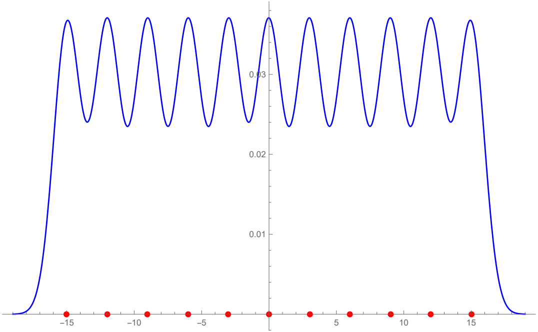
Example 1: datapoints well-spread out

```
sample=[-15. -12.  -9.  -6.  -3.   0.   3.   6.   9.  12.  15.]

NPMLE output:

weights= [0.09100201 0.09084195 0.09092767 0.09092682 0.09092779 0.09083749
0.09083684 0.09092779 0.09092766 0.09084195 0.09100201]
centers= [-14.96996997 -11.996997    -8.99399399  -5.99099099  -2.98798799
0.01501502   2.98798799   5.99099099   8.99399399  11.996997 14.96996997]
```



Example 1: datapoints well-spread out

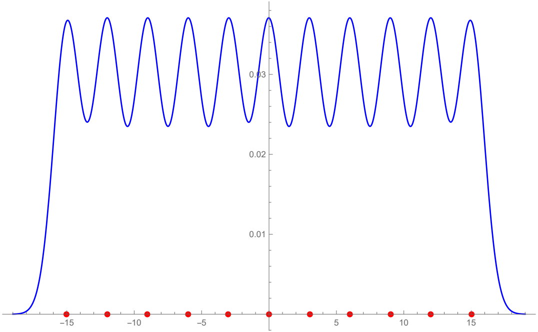
```
sample=[-15. -12.  -9.  -6.  -3.   0.   3.   6.   9.  12.  15.]

NPMLE output:

weights= [0.09100201 0.09084195 0.09092767 0.09092682 0.09092779 0.09083749
0.09083684 0.09092779 0.09092766 0.09084195 0.09100201]
centers= [-14.96996997 -11.996997    -8.99399399  -5.99099099  -2.98798799
0.01501502   2.98798799   5.99099099   8.99399399  11.996997 14.96996997]
```



- Bad news: NPMLE fits an n-component Gaussian mixture

Example 1: datapoints well-spread out

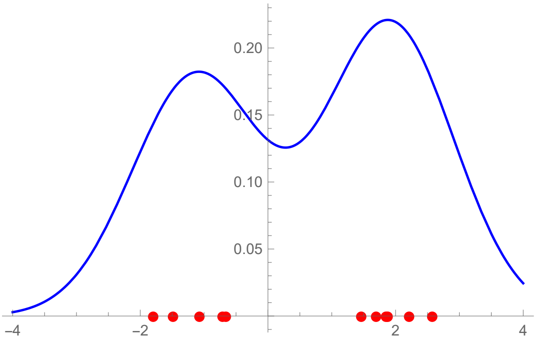
```
sample=[-15. -12.  -9.  -6.  -3.   0.   3.   6.   9.  12.  15.]

NPMLE output:

weights= [0.09100201 0.09084195 0.09092767 0.09092682 0.09092779 0.09083749
0.09083684 0.09092779 0.09092766 0.09084195 0.09100201]
centers= [-14.96996997 -11.996997    -8.99399399  -5.99099099  -2.98798799
0.01501502   2.98798799   5.99099099   8.99399399  11.996997 14.96996997]
```



- Bad news: NPMLE fits an n-component Gaussian mixture
- Good news: this sample is atypical!

Example 2: datapoints clustered

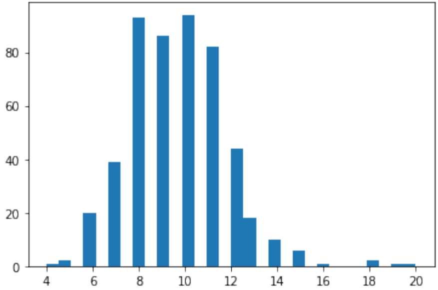
```
sample= [ 1.86797447  1.4552763  -1.80237513 -0.7244036   2.22400636  1.85900276
  2.57612104  1.69214083 -0.64707404 -1.48164282 -1.07169643]

NPMLE output:

weights=  [0.45098479 0.5490152 ]
centers=  [-1.12302888  1.90554054]
```



NPMLE fits a 2-component Gaussian mixture

Further experiment

- True distribution: $N(0,1)$ (single component)
- Sample size $n = 10000$



Histogram of $|\mathrm{supp}(\hat{\pi}_{\mathrm{NPMLE}})|$ in 500 trials

Main result

Theorem (P.-Wu '20)

- *Exists absolute constant* C_0 *s.t. for Gaussian location mixtures,*

$$|\operatorname{supp}(\hat{\pi}_{\mathrm{NPMLE}})| \le C_0(x_{\max}-x_{\min})^2, \quad x_{\min} = \min_{i\in[n]} x_i, x_{\max} = \max_{i\in[n]} x_i$$

Main result

Theorem (P.-Wu '20)

- *Exists absolute constant C_0 s.t. for Gaussian location mixtures,*

$$|\operatorname{supp}(\hat{\pi}_{\mathrm{NPMLE}})| \le C_0(x_{\max} - x_{\min})^2, \quad x_{\min} = \min_{i\in[n]} x_i, x_{\max} = \max_{i\in[n]} x_i$$

- *Thus, if $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \pi * N(0,1)$ for some 1-subgaussian mixing distribution π, then w.h.p.*

$$|\operatorname{supp}(\hat{\pi}_{\mathrm{NPMLE}})| \le O(\log n)$$

Main result

Theorem (P.-Wu '20)

- *Exists absolute constant C_0 s.t. for Gaussian location mixtures,*

$$|\operatorname{supp}(\hat{\pi}_{\mathrm{NPMLE}})| \leq C_0(x_{\max} - x_{\min})^2, \quad x_{\min} = \min_{i \in [n]} x_i, x_{\max} = \max_{i \in [n]} x_i$$

- *Thus, if $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \pi * N(0,1)$ for some 1-subgaussian mixing distribution π, then w.h.p.*

$$|\operatorname{supp}(\hat{\pi}_{\mathrm{NPMLE}})| \leq O(\log n)$$

- Significantly improves the worst-case bound (n)
- If data are drawn from a finite k-GM, NPMLE typically fits an $O(\log n)$-GM
- Universality of $\log n$: analogous result holds for exponential families with tail $\exp(-|x|^C)$ for $C > 1$

Optimality of $\log n$

- Is our estimate $|\operatorname{supp} \hat{\pi}_{\mathrm{NPMLE}}| \lesssim \log n$ tight? YES!
  - Inapproximability result [Wu-Verdú '10]:

$$\inf_{\pi:k\text{-atomic}} H(p_\pi, N(0,2)) \geq e^{-O(k)}$$

  - Thus, if $X_i \overset{iid}{\sim} N(0,2)$ then **for any mixture density estimator**:

$$H(P_{\hat{\pi}}, N(0,2)) = \mathrm{poly}(n) \quad \Longrightarrow \quad |\operatorname{supp} \hat{\pi}| = \Omega(\log n)$$

  - For any subgaussian π: $H(p_{\hat{\pi}_{\mathrm{NPMLE}}}, p_\pi) = O_P(\frac{\log n}{\sqrt{n}})$ [Zhang '09]

Optimality of $\log n$

- Is our estimate $|\operatorname{supp}\hat\pi_{\mathrm{NPMLE}}| \lesssim \log n$ tight? YES!
  - Inapproximability result [Wu-Verdú '10]:

$$\inf_{\pi:k\text{-atomic}} H(p_\pi, N(0,2)) \ge e^{-O(k)}$$

  - Thus, if $X_i \overset{iid}{\sim} N(0,2)$ then **for any mixture density estimator**:

$$H(P_{\hat\pi}, N(0,2)) = \mathrm{poly}(n) \qquad \Longrightarrow \qquad |\operatorname{supp}\hat\pi| = \Omega(\log n)$$

  - For any subgaussian π: $H(p_{\hat\pi_{\mathrm{NPMLE}}}, p_\pi) = O_P(\frac{\log n}{\sqrt{n}})$ [Zhang '09]
- Why $\ge \log n$ mixture components is useless?
  - Approximability result (m.o.m.):

$$\forall \pi \in \mathrm{SubGauss} \exists \pi' - k\text{-atomic}: \qquad H(P_\pi, P_{\pi'}) = o(1/\sqrt{n})$$

    and $\boxed{k = O(\log n)}$.
  - IOW, n-sample $X_i \overset{iid}{\sim} P_\pi$ is statistically indistinguihsable from $X_i' \overset{iid}{\sim} P_{\pi'}$

Self-regularization property of the NPMLE

Recap:

- We have a sequence of models

$$\mathcal{M}_1 \subset \mathcal{M}_2 \subset \cdots \mathcal{M}$$

$$\begin{aligned}
\mathcal{M} &= \{P_\pi = \pi * N(0,1) : \pi \text{ is 1-subgaussian } \} \\
\mathcal{M}_k &= \{P_\pi = \pi * N(0,1) : \pi \text{ is 1-subgaussian and } k\text{-atomic}\}
\end{aligned}$$

- We know that *statistical degree* is $\Theta(\log n)$.
  I.e. for any $f \in \mathcal{M}$ there exists $f_k \in \mathcal{M}_k$ with $k \asymp \log n$ such that

$$\mathrm{TV}(f^{\otimes n}, f_k^{\otimes n}) = o(1)\,.$$

- Surprise: NPMLE automatically selects density estimate $\hat{f} \in \mathcal{M}_k$ with $k \asymp \log n$!

Self-regularization property of the NPMLE

Recap:

- We have a sequence of models

$$\mathcal{M}_1 \subset \mathcal{M}_2 \subset \cdots \mathcal{M}$$

$$\begin{aligned}\mathcal{M} &= \{P_\pi = \pi * N(0,1) : \pi \text{ is 1-subgaussian }\}\\ \mathcal{M}_k &= \{P_\pi = \pi * N(0,1) : \pi \text{ is 1-subgaussian and } k\text{-atomic}\}\end{aligned}$$

- We know that *statistical degree* is $\Theta(\log n)$.
  I.e. for any $f \in \mathcal{M}$ there exists $f_k \in \mathcal{M}_k$ with $k \asymp \log n$ such that

$$\mathrm{TV}(f^{\otimes n}, f_k^{\otimes n}) = o(1)\,.$$

- Surprise: NPMLE automatically selects density estimate $\hat{f} \in \mathcal{M}_k$ with $k \asymp \log n$!

Self-regularization of NPMLE

Absent any explicit form of model selection, NPMLE automatically chooses the model of order-optimal complexity.

Model selection and penalized MLE

- The likelihood of the best k-GM fit (non-convex):

$$L_{\mathrm{opt}}(k) \triangleq \max_{\pi:k\text{-atomic}} \frac{1}{n}\sum_{i=1}^{n} \log p_\pi(x_i).$$

- Penalized MLE: for some pre-defined maximal model size K,

$$\max_{k=1,\ldots,K} \{L_{\mathrm{opt}}(k) - \mathrm{pen}(k)\}$$

Model selection and penalized MLE

- The likelihood of the best k-GM fit (non-convex):

$$L_{\text{opt}}(k) \triangleq \max_{\pi:k\text{-atomic}} \frac{1}{n}\sum_{i=1}^{n} \log p_\pi(x_i).$$

- Penalized MLE: for some pre-defined maximal model size K,

$$\max_{k=1,\dots,K} \{L_{\text{opt}}(k) - \text{pen}(k)\}$$

- New result shows: w.h.p. $k \mapsto L_{\text{opt}}(k)$ flattens after $k \geq C \log n$.

Model selection and penalized MLE

- The likelihood of the best k-GM fit (non-convex):

$$L_{\text{opt}}(k) \triangleq \max_{\pi: k\text{-atomic}} \frac{1}{n} \sum_{i=1}^{n} \log p_\pi(x_i).$$

- Penalized MLE: for some pre-defined maximal model size K,

$$\max_{k=1,\ldots,K} \{L_{\text{opt}}(k) - \text{pen}(k)\}$$

- New result shows: w.h.p. $k \mapsto L_{\text{opt}}(k)$ flattens after $k \geq C \log n$.
- To achieve model selection consistency, penality is probably needed e.g. BIC $\text{pen}(k) = \frac{k}{2} \log n$ [Leroux '92, Keribin '00]
- NPMLE exhibits some mild overfitting, a modest (and fair) price for being completely automatic and computationally attractive.

Analogy with shape-constrained NPMLE

$x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} f$, a monotone density on $[0, 1]$ [Grenander '56]

- $\hat{f}_{\mathrm{NPMLE}}$ (Grenander estimator) is piecewise constant with k_n pieces.
- Deterministically $k_n \leq n$

Analogy with shape-constrained NPMLE

$x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} f$, a monotone density on $[0,1]$ [Grenander '56]

- $\hat{f}_{\mathrm{NPMLE}}$ (Grenander estimator) is piecewise constant with k_n pieces.
- Deterministically $k_n \le n$
- Typically
  - Under conditions on f: $k_n = O_P(n^{1/3})$ [Groeneboom '11]
  - For uniform f: $k_n \approx N(\log n, \log n)$ [Groeneboom-Lopuhaa '93]

Analogy with shape-constrained NPMLE

$x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} f$, a monotone density on $[0,1]$ [Grenander '56]

- $\hat{f}_{\mathrm{NPMLE}}$ (Grenander estimator) is piecewise constant with k_n pieces.
- Deterministically $k_n \le n$
- Typically
  - Under conditions on f: $k_n = O_P(n^{1/3})$ [Groeneboom '11]
  - For uniform f: $k_n \approx N(\log n, \log n)$ [Groeneboom-Lopuhaa '93]
- Thanks to an explicit characterization of $\hat{f}_{\mathrm{NPMLE}}$ in terms of empirical processes (no such result for mixture models)

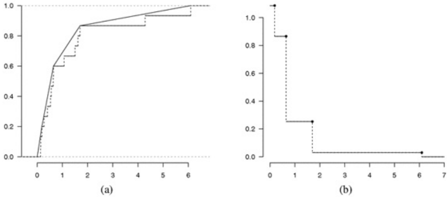


Figure 2.4 (a) Empirical distribution function of a sample of size $n = 15$ and its concave majorant. (b) The resulting Grenander estimate.

Image credit: [Groeneboom-Jongbloed '14]

Statistical consequence: density estimation

Theorem (Zhang '09)

Let $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} p_\pi \triangleq \pi * \varphi$. *For any 1-subgaussian* π*,*

$$\mathbb{E}_\pi[H^2(p_{\hat{\pi}_{\mathrm{NPMLE}}}, p_\pi)] \lesssim \frac{\log^2 n}{n},$$

Statistical consequence: density estimation

Theorem (Zhang '09)

Let $x_1, \ldots, x_n \overset{i.i.d.}{\sim} p_\pi \triangleq \pi * \varphi$. *For any 1-subgaussian* π,

$$\mathbb{E}_\pi[H^2(p_{\hat{\pi}_{\mathrm{NPMLE}}}, p_\pi)] \lesssim \frac{\log^2 n}{n},$$

- Std. analysis of NPMLE is via empirical process theory:

$$\sup_{f \in \mathcal{F}} |\hat{\mathbb{E}}_n[f] - \mathbb{E}[f]| \lesssim \epsilon + \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, \epsilon)}{n}}$$

$+$ metric entropy bounds

Statistical consequence: density estimation

Theorem (Zhang '09)

Let $x_1, \ldots, x_n \overset{i.i.d.}{\sim} p_\pi \triangleq \pi * \varphi$. *For any 1-subgaussian* π,

$$\mathbb{E}_\pi[H^2(p_{\hat{\pi}_{\mathrm{NPMLE}}}, p_\pi)] \lesssim \frac{\log^2 n}{n},$$

- Std. analysis of NPMLE is via empirical process theory:

$$\sup_{f \in \mathcal{F}} |\hat{\mathbb{E}}_n[f] - \mathbb{E}[f]| \lesssim \epsilon + \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, \epsilon)}{n}}$$

  $+$ metric entropy bounds

- Our guess: $\hat{\pi}_{\mathrm{NPMLE}}$ has $C \log n$ atoms so we expect

$$H^2(P_{\hat{\pi}_{\mathrm{NPMLE}}}, P_\pi) \lesssim \frac{\log n}{n}$$

  Unfortunately, rigorous proof picks up another $\log n$.
  (Best lower bound on H^2 density estimation $\Omega\left(\frac{\log n}{n}\right)$ [Kim '14]).

Statistical consequence: density estimation

Theorem (Zhang '09)

Let $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} p_\pi \triangleq \pi * \varphi$. For any 1-subgaussian π,

$$\mathbb{E}_\pi[H^2(p_{\hat{\pi}_{\mathrm{NPMLE}}}, p_\pi)] \lesssim \frac{\log^2 n}{n},$$

Proof based on self-regularization: Let $k = C \log n$.

- On the event that $\hat{\pi}_{\mathrm{NPMLE}}$ is k-atomic, $\hat{\pi}_{\mathrm{NPMLE}}$ coincides with parametric MLE over k-GMs
- Find k-GM $p_{\pi'}$ such that $\mathrm{TV}(p_{\pi'}, p_\pi) \le n^{-10}$, so we can couple $(x_1, \ldots, x_n)$ to $(x'_1, \ldots, x'_n) \overset{\text{i.i.d.}}{\sim} p_{\pi'}$ with probability $1 - n^{-9}$
- This reduces the problem to k-GM and allows invoking existing guarantee for parametric MLE: $O(\frac{k}{n} \log \frac{n}{k})$ [Maugis-Michel '11]

Self-regularization: cartoon example

- Model sequence on $\mathbb{Z}_+$:

$$\mathcal{M}_1 \subset \mathcal{M}_2 \subset \cdots \subset \mathcal{M}$$

$$\begin{aligned} \mathcal{M} &= \{\pi : \pi[\{m\}] \le 2^{-m}, m \in \mathbb{Z}_+\} \\ \mathcal{M}_k &= \{\pi \in \mathcal{M} : \pi[\{m\}] = 0, m > k\} \end{aligned}$$

- By truncation we see:

$$\forall \pi \in \mathcal{M} \quad \exists \pi' \in \mathcal{M}_k : \mathrm{TV}(\pi, \pi') = o(1/n)$$

with $k \asymp \log n$ – *statistical degree*.

Self-regularization: cartoon example

- Model sequence on $\mathbb{Z}_+$:

$$\mathcal{M}_1 \subset \mathcal{M}_2 \subset \cdots \subset \mathcal{M}$$

$$\begin{aligned} \mathcal{M} &= \{\pi : \pi[\{m\}] \le 2^{-m}, m \in \mathbb{Z}_+\} \\ \mathcal{M}_k &= \{\pi \in \mathcal{M} : \pi[\{m\}] = 0, m > k\} \end{aligned}$$

- By truncation we see:

$$\forall \pi \in \mathcal{M} \quad \exists \pi' \in \mathcal{M}_k : \mathrm{TV}(\pi, \pi') = o(1/n)$$

with $\boxed{k \asymp \log n}$ – *statistical degree*.

- OTOH, given $X_i \stackrel{iid}{\sim} \pi$ we have

$$\hat{\pi}_{\mathrm{NPMLE}}[\{m\}] = \frac{1}{n} \sum_{t=1}^{n} 1\{X_i = m\},$$

and clearly

$$\mathbb{P}[\hat{\pi}_{\mathrm{NPMLE}} \in \mathcal{M}_k] = 1 - o(1)$$

whenever $\boxed{k \gtrsim \log n}$.

Self-regularization: cartoon example

- Model sequence on $\mathbb{Z}_+$:

$$\mathcal{M}_1 \subset \mathcal{M}_2 \subset \cdots \subset \mathcal{M}$$

$$\begin{aligned}\mathcal{M} &= \{\pi : \pi[\{m\}] \le 2^{-m}, m \in \mathbb{Z}_+\} \\ \mathcal{M}_k &= \{\pi \in \mathcal{M} : \pi[\{m\}] = 0, m > k\}\end{aligned}$$

- By truncation we see:

$$\forall \pi \in \mathcal{M} \quad \exists \pi' \in \mathcal{M}_k : \mathrm{TV}(\pi, \pi') = o(1/n)$$

with

- OTOH

Observation: More samples "unlock" new dimensions in $\mathcal{M}$ and NPMLE adapts to it.

$$\hat{\pi}_{\mathrm{NPMLE}}[\{m\}] = \frac{1}{n} \sum_{t=1} 1\{X_i = m\},$$

and clearly

$$\mathbb{P}[\hat{\pi}_{\mathrm{NPMLE}} \in \mathcal{M}_k] = 1 - o(1)$$

whenever $\boxed{k \gtrsim \log n}$.

Proof of the main result

Warmup: Poisson model

Self-regularization in Poisson mixture is easy to prove:

- $p_\theta(x) = \frac{\theta^x}{x!} e^{-\theta}$ and $x \in \mathbb{Z}_+$.

Warmup: Poisson model

Self-regularization in Poisson mixture is easy to prove:

- $p_\theta(x) = \frac{\theta^x}{x!} e^{-\theta}$ and $x \in \mathbb{Z}_+$.
- Gradient

$$D_{\hat{\pi}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{p_\theta(x_i)}{p_{\hat{\pi}}(x_i)} = e^{-\theta} \underbrace{\sum_{i=1}^{n} w_i \theta^{x_i}}_{\text{deg-}x_{\max} \text{ polynomial in } \theta}$$

So

$$D'_{\hat{\pi}}(\theta) = e^{-\theta} \times \mathrm{poly}(\theta)$$

has at most $\deg \mathrm{poly} \le x_{\max}$ roots!

Warmup: Poisson model

Self-regularization in Poisson mixture is easy to prove:

- $p_\theta(x) = \frac{\theta^x}{x!}e^{-\theta}$ and $x \in \mathbb{Z}_+$.
- Gradient

$$D_{\hat{\pi}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{p_\theta(x_i)}{p_{\hat{\pi}}(x_i)} = e^{-\theta} \underbrace{\sum_{i=1}^{n} w_i \theta^{x_i}}_{\text{deg-}x_{\max}\text{ polynomial in }\theta}$$

  So

$$D'_{\hat{\pi}}(\theta) = e^{-\theta} \times \mathrm{poly}(\theta)$$

  has at most $\deg \mathrm{poly} \le x_{\max}$ roots!
- For nice (e.g. subexponential) mixing distribution π, $x_{\max} = O_P(\log n)$

Warmup: Poisson model

Self-regularization in Poisson mixture is easy to prove:

- $p_\theta(x) = \frac{\theta^x}{x!} e^{-\theta}$ and $x \in \mathbb{Z}_+$.
- Gradient

$$D_{\hat\pi}(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{p_\theta(x_i)}{p_{\hat\pi}(x_i)} = e^{-\theta} \underbrace{\sum_{i=1}^n w_i \theta^{x_i}}_{\text{deg-}x_{\max}\text{ polynomial in }\theta}$$

  So

$$D'_{\hat\pi}(\theta) = e^{-\theta} \times \mathrm{poly}(\theta)$$

  has at most $\deg \mathrm{poly} \le x_{\max}$ roots!

- For nice (e.g. subexponential) mixing distribution π, $x_{\max} = O_P(\log n)$
- This does <u>not</u> work for Gaussian: $D(\theta)$ not a poly!

Recall from optimality condition

- $|\mathrm{supp}(\hat{\pi}_{\mathrm{NPMLE}})| \le \#$ of critical points of

$$D_{\hat{\pi}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{p_{\hat{\pi}}(x_i)}\phi(x_i - \theta) \propto \sum_{i=1}^{n} w_i \phi(x_i - \theta) = (\pi * \varphi)(\theta)$$

where π is supported on $\{x_1, \ldots, x_n\} \subset [x_{\min}, x_{\max}]$

Recall from optimality condition

- $|\operatorname{supp}(\hat{\pi}_{\mathrm{NPMLE}})| \le \#$ of critical points of

$$D_{\hat{\pi}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_{\hat{\pi}}(x_i)}\phi(x_i-\theta) \propto \sum_{i=1}^{n} w_i\phi(x_i-\theta) = (\pi * \varphi)(\theta)$$

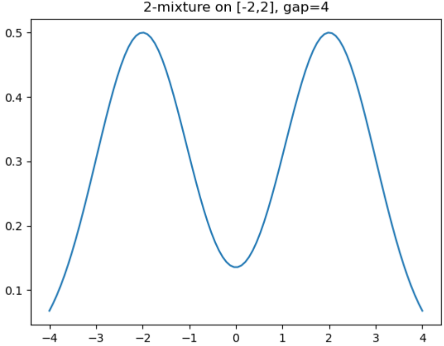
  where π is supported on $\{x_1,\ldots,x_n\} \subset [x_{\min}, x_{\max}]$
- Reduces to counting critical points of Gaussian convolved with compactly supported measure

Number of modes of Gaussian mixtures

- **Key analytic puzzle:** Given $a > 0$ and a measure π on $[-a, a]$ how many modes can $P_\pi = \pi * \phi$ have?

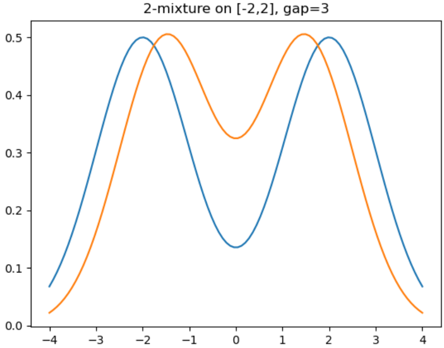
Number of modes of Gaussian mixtures

- **Key analytic puzzle:** Given $a > 0$ and a measure π on $[-a, a]$ how many modes can $P_\pi = \pi * \phi$ have?
- Let us try to understand it:



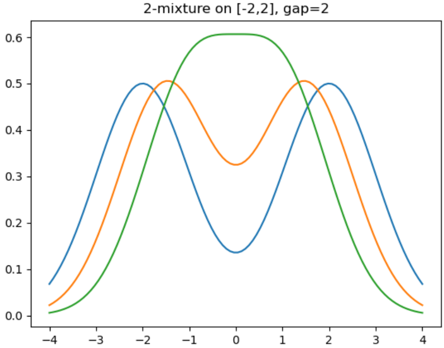
Number of modes of Gaussian mixtures

- **Key analytic puzzle:** Given $a > 0$ and a measure π on $[-a, a]$ how many modes can $P_\pi = \pi * \phi$ have?
- Let us try to understand it:



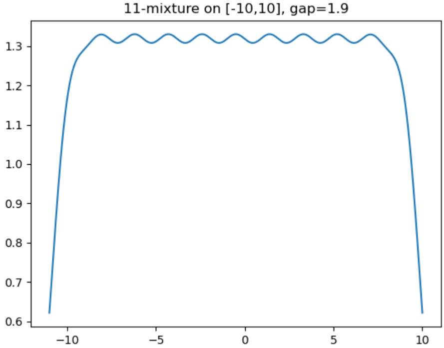
Number of modes of Gaussian mixtures

- **Key analytic puzzle:** Given $a > 0$ and a measure π on $[-a, a]$ how many modes can $P_\pi = \pi * \phi$ have?
- Let us try to understand it:



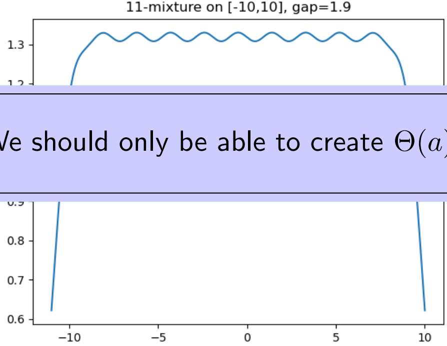
Number of modes of Gaussian mixtures

- **Key analytic puzzle:** Given $a > 0$ and a measure π on $[-a, a]$ how many modes can $P_\pi = \pi * \phi$ have?
- Let us try to understand it:



Number of modes of Gaussian mixtures

- **Key analytic puzzle:** Given $a > 0$ and a measure π on $[-a, a]$ how many modes can $P_\pi = \pi * \phi$ have?
- Let us try to understand it:



We should only be able to create $\Theta(a)$ modes?

Maxima of Gaussian mixtures

Theorem

Let π be supported on $[-a, a]$. Then $\pi * \varphi$ has at most $C_0 a^2$ critical points (C_0–absolute constant). Furthermore, this bound is order-tight as $a \to \infty$.

Maxima of Gaussian mixtures

Theorem

Let π be supported on $[-a, a]$. Then $\pi * \varphi$ has at most $C_0 a^2$ critical points (C_0–absolute constant). Furthermore, this bound is order-tight as $a \to \infty$.

- A concurrent (and totally independent) work in IT on amplitude-constrained channel capacity contains the same upper bound! [Dytso-Yagli-Poor-Shamai '20]
- They further conjectured that $O(a)$ should be tight...
- Our lower bound :

Maxima of Gaussian mixtures

Theorem

Let π be supported on $[-a, a]$. Then $\pi * \varphi$ has at most $C_0 a^2$ critical points (C_0–absolute constant). Furthermore, this bound is order-tight as $a \to \infty$.

- A concurrent (and totally independent) work in IT on amplitude-constrained channel capacity contains the same upper bound! [Dytso-Yagli-Poor-Shamai '20]
- They further conjectured that $O(a)$ should be tight...
- Our lower bound :

Lemma

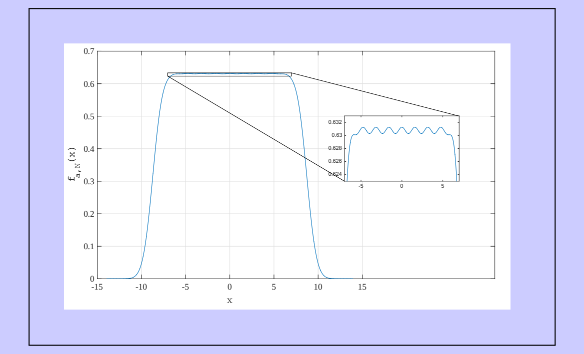
Take $\pi = (1 + \sin(\omega x))1\{|x| \le a\}$. For $\omega \asymp a$ density $\pi * \phi$ has $\Omega(a^2)$ modes.

Maxima of Gaussian mixtures

Theorem

Let π be supported on $[-a, a]$. Then $\pi * \varphi$ has at most $C_0 a^2$ critical points (C_0–absolute constant). Furthermore, this bound is order-tight as $a \to \infty$.

- A con
  amplit ame upper
  bound
- They
- Our lo



Lemma

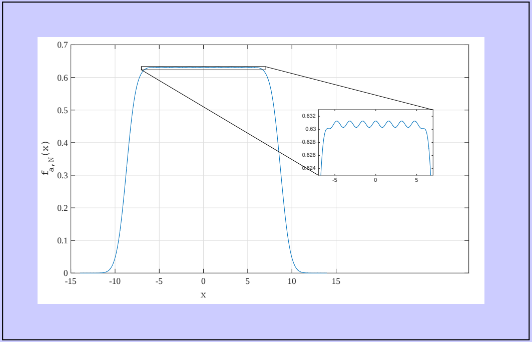
Take $\pi = ($ ϕ has $\Omega(a^2)$ modes.

Maxima of Gaussian mixtures

Theorem

Let π be supported on $[-a, a]$. Then $\pi * \varphi$ has at most $C_0 a^2$ critical points (C_0–absolute constant). Furthermore, this bound is order-tight as $a \to \infty$.

- A conc … amplit … ame upper bound
- They f
- Our lo

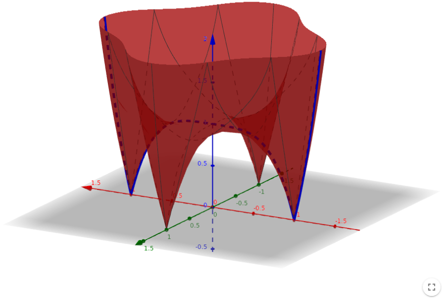


Lemma

Take $\pi = ($ … ϕ has $\Omega(a^2)$ modes.

Another construction in [Kashyap-Krishnapur '20]

Main tool: complex analysis



via http://geogebra.org

- Imagine a poly with many roots in the unit circle

$$p(z) = c\prod_{j=1}^{n}(z-a_j)$$

- Then its magnitude on a far-away circle should be very large:

$$\frac{|p(z)|}{|p(0)|} \gtrsim |z|^n, \qquad |z| \gg 1$$

- This generalizes: holomorphic functions with many zeros must grow very fast at infinity.

Main tool: complex analysis

Jensen's formula

- Let g be an analytic function. Then

$$\sum_k \log \frac{|a_k|}{R} = \frac{1}{2\pi} \int_0^{2\pi} \log \frac{|g(Re^{i\theta})|}{|g(0)|} d\theta$$

where $a_1, a_2, \ldots$ are the zeros of g inside disk of radius R

Main tool: complex analysis

Jensen's formula

- Let g be an analytic function. Then

$$\sum_k \log \frac{|a_k|}{R} = \frac{1}{2\pi} \int_0^{2\pi} \log \frac{|g(Re^{i\theta})|}{|g(0)|} d\theta$$

  where $a_1, a_2, \ldots$ are the zeros of g inside disk of radius R
- Consequence: for $r < R$,

$$\# \{\text{zeros of } g \text{ inside disk of radius } r\} \ \leq \frac{\log \frac{M}{|g(0)|}}{\log \frac{R}{r}}$$

  where $M = \sup_{|z|=R} |g(z)|$

Proof for Gaussian mixtures

Let $U \sim \pi$ and $p(x) = (\pi * \varphi)(x) = \mathbb{E}[\varphi(x - U)]$.

- **Step 1: Localize roots.** All real roots of p' are in $[-a, a]$, since

$$p'(x) = \mathbb{E}[(U - x)\varphi(x - U)], \quad |U| \le a.$$

Pick $z_0 = -2a$, $r = 3a$, $R = 4a$

Proof for Gaussian mixtures

Let $U \sim \pi$ and $p(x) = (\pi * \varphi)(x) = \mathbb{E}[\varphi(x - U)]$.

- **Step 1: Localize roots.** All real roots of p' are in $[-a, a]$, since

$$p'(x) = \mathbb{E}[(U - x)\varphi(x - U)], \quad |U| \leq a.$$

  Pick $z_0 = -2a$, $r = 3a$, $R = 4a$

- **Step 2: Lower bound $|p'(2a)|$.**

$$p'(-2a) \geq ae^{-Ca^2}$$

Proof for Gaussian mixtures

Let $U \sim \pi$ and $p(x) = (\pi * \varphi)(x) = \mathbb{E}[\varphi(x - U)]$.

- **Step 1: Localize roots.** All real roots of p' are in $[-a, a]$, since

$$p'(x) = \mathbb{E}[(U - x)\varphi(x - U)], \quad |U| \leq a.$$

  Pick $z_0 = -2a$, $r = 3a$, $R = 4a$

- **Step 2: Lower bound $|p'(2a)|$.**

$$p'(-2a) \geq ae^{-Ca^2}$$

- **Step 3: Upper bound $|p'|$ on complex circle**

$$|p'(w)| \leq e^{Ca^2}, \quad \forall w \in R \triangleq \{z : |z + 2a| \leq 4a\}$$

Proof for Gaussian mixtures

Let $U \sim \pi$ and $p(x) = (\pi * \varphi)(x) = \mathbb{E}[\varphi(x-U)]$.

- **Step 1: Localize roots.** All real roots of p' are in $[-a, a]$, since

$$p'(x) = \mathbb{E}[(U-x)\varphi(x-U)], \quad |U| \le a.$$

  Pick $z_0 = -2a$, $r = 3a$, $R = 4a$

- **Step 2: Lower bound $|p'(2a)|$.**

$$p'(-2a) \ge ae^{-Ca^2}$$

- **Step 3: Upper bound $|p'|$ on complex circle**

$$|p'(w)| \le e^{Ca^2}, \quad \forall w \in R \triangleq \{z : |z+2a| \le 4a\}$$

- Jensen's formula: p' has at most $C_0 a^2$ in R (which contains all its real roots).

Proof for Gaussian mixtures

Let $U \sim \pi$ and $p(x) = (\pi * \varphi)(x) = \mathbb{E}[\varphi(x - U)]$.

- **Step 1: Localize roots.** All real roots of p' are in $[-a, a]$, since

$$p'(x) = \mathbb{E}[(U - x)\varphi(x - U)], \quad |U| \le a.$$

  Pick $z_0 = -2a$, $r = 3a$, $R = 4a$
- **Step 2: Lower bound $|p'(2a)|$.**

$$p'(-2a) \ge a e^{-Ca^2}$$

- **Step 3: Upper bound $|p'|$ on complex circle**

$$|p'(w)| \le e^{Ca^2}, \quad \forall w \in R \triangleq \{z : |z + 2a| \le 4a\}$$

- Jensen's formula: p' has at most $C_0 a^2$ in R (which contains all its real roots).
- Proof for exponential family is more complicated.

Discussion and Open problems

Limitation of current proof technique

Mixture of exponentials [Jewell '82]

- $\mathrm{Exp}(\theta)$ with density $p_\theta(x) = \theta e^{-\theta x} \mathbf{1}\{x > 0\}$ and $\theta > 0$.

$$\hat{\pi}_{\mathrm{NPMLE}} = \arg \max_{\pi \in \mathcal{M}(\mathbb{R}_+)} \frac{1}{n} \sum_{i=1}^{n} \log p_\pi(x_i), \quad p_\pi(x) = \int \theta e^{-\theta x} \pi(d\theta).$$

Limitation of current proof technique

Mixture of exponentials [Jewell '82]

- $\mathrm{Exp}(\theta)$ with density $p_\theta(x) = \theta e^{-\theta x}\mathbf{1}\{x > 0\}$ and $\theta > 0$.

$$\hat{\pi}_{\mathrm{NPMLE}} = \arg\max_{\pi \in \mathcal{M}(\mathbb{R}_+)} \frac{1}{n}\sum_{i=1}^{n} \log p_\pi(x_i), \quad p_\pi(x) = \int \theta e^{-\theta x}\pi(d\theta).$$

- Similar analysis yields:

$$|\operatorname{supp}(\hat{\pi}_{\mathrm{NPMLE}})| \lesssim \frac{x_{\max}}{x_{\min}}.$$

- If $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(1)$, then whp $x_{\max} \asymp \log n, x_{\min} \asymp \frac{1}{n}$ and $|\operatorname{supp}(\hat{\pi}_{\mathrm{NPMLE}})| \lesssim n \log n$ is useless

Limitation of current proof technique

Mixture of exponentials [Jewell '82]

- $\mathrm{Exp}(\theta)$ with density $p_\theta(x) = \theta e^{-\theta x}\mathbf{1}\{x > 0\}$ and $\theta > 0$.

$$\hat{\pi}_{\mathrm{NPMLE}} = \arg\max_{\pi \in \mathcal{M}(\mathbb{R}_+)} \frac{1}{n}\sum_{i=1}^{n} \log p_\pi(x_i), \quad p_\pi(x) = \int \theta e^{-\theta x}\pi(d\theta).$$

- Similar analysis yields:

$$|\operatorname{supp}(\hat{\pi}_{\mathrm{NPMLE}})| \lesssim \frac{x_{\max}}{x_{\min}}.$$

- If $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(1)$, then whp $x_{\max} \asymp \log n, x_{\min} \asymp \frac{1}{n}$ and $|\operatorname{supp}(\hat{\pi}_{\mathrm{NPMLE}})| \lesssim n \log n$ is useless
- Open question: does self-regularization fail for exp mixture?

Stronger self-regularization for constrained solution

Suppose true mixing distribution π is known to be supported on $[-1, 1]$.

- Correct model complexity reduces to $O(\frac{\log n}{\log \log n})$ components.

Stronger self-regularization for constrained solution

Suppose true mixing distribution π is known to be supported on $[-1, 1]$.

- Correct model complexity reduces to $O(\frac{\log n}{\log \log n})$ components.
- Constrained NPMLE:

$$\hat{\pi}_{\mathrm{NPMLE}} = \arg \max_{\pi \in \mathcal{M}([-1,1])} \frac{1}{n} \sum_{i=1}^{n} \log(\pi * \varphi)(x_i)$$

Stronger self-regularization for constrained solution

Suppose true mixing distribution π is known to be supported on $[-1, 1]$.

- Correct model complexity reduces to $O(\frac{\log n}{\log\log n})$ components.
- Constrained NPMLE:

$$\hat{\pi}_{\mathrm{NPMLE}} = \arg \max_{\pi \in \mathcal{M}([-1,1])} \frac{1}{n} \sum_{i=1}^{n} \log(\pi * \varphi)(x_i)$$

- Open question: Is $\hat{\pi}_{\mathrm{NPMLE}}$ $O(\frac{\log n}{\log\log n})$-atomic?

Stronger self-regularization for constrained solution

Suppose true mixing distribution π is known to be supported on $[-1, 1]$.

- Correct model complexity reduces to $O(\frac{\log n}{\log \log n})$ components.
- Constrained NPMLE:

$$\hat{\pi}_{\mathrm{NPMLE}} = \arg \max_{\pi \in \mathcal{M}([-1,1])} \frac{1}{n} \sum_{i=1}^{n} \log(\pi * \varphi)(x_i)$$

- Open question: Is $\hat{\pi}_{\mathrm{NPMLE}}$ $O(\frac{\log n}{\log \log n})$-atomic?
- Our method disregards special structure of weights $\frac{1}{P_\pi(X_i)}$ (and provably fails)

Multiple dimensions

- Existence and uniqueness not resolved
  - Major difficulty: $\#$ modes of n-GM in d dim $= \Omega(n^d)$ [Améndola-Engström-Haase '17]

Multiple dimensions

- Existence and uniqueness not resolved
  - Major difficulty: $\#$ modes of n-GM in d dim $= \Omega(n^d)$ [Améndola-Engström-Haase '17]
- Fundamental distinction with 1D:
  - With positive (but low) probability $\hat{\pi}_{\mathrm{NPMLE}}$ is not unique (e.g. Poisson mixture in 2D) [P.-Wu '20]

Multiple dimensions

- Existence and uniqueness not resolved
  - Major difficulty: $\#$ modes of n-GM in d dim $= \Omega(n^d)$ [Améndola-Engström-Haase '17]
- Fundamental distinction with 1D:
  - With positive (but low) probability $\hat{\pi}_{\mathrm{NPMLE}}$ is not unique (e.g. Poisson mixture in 2D) [P.-Wu '20]
- Open problems:
  - Understanding the typical structural NPMLE in d dimensions
  - Scalable algorithms

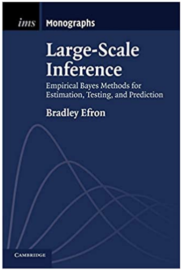
NPMLE in Empirical Bayes

Empirical Bayes formalism

Large-scale inference

$$X_i \overset{\text{ind}}{\sim} P_{\theta_i}, \quad i = 1, \ldots, n$$

Goal: Estimate $\theta_1, \ldots, \theta_n$



- EB setting: $\theta_i \overset{\text{i.i.d.}}{\sim} \pi$.
  - Metric: compete with oracle (Bayes) who knows π
- Compound setting: θ_i deterministic.
  - Metric: compete with oracle who knows empirical distribution of θ_i's
- Competetive optimality offers a meaningful framework to go beyond (pessimistic) minimax setting

Empirical Bayes estimator

Bayes estimator $\hat{\theta}_{\mathsf{Bayes}}(\cdot;\pi)$ depends on the unknown π.

Robbins' meta-principle

- Learn the prior $\hat{\pi}$ (empirical distribution) from data
- Execute Bayes strategy with learned prior $\hat{\theta}_{\mathsf{Bayes}}(\cdot;\hat{\pi})$

Empirical Bayes estimator

Bayes estimator $\hat{\theta}_{\mathsf{Bayes}}(\cdot;\pi)$ depends on the unknown π.

Robbins' meta-principle

- Learn the prior $\hat{\pi}$ (empirical distribution) from data
- Execute Bayes strategy with learned prior $\hat{\theta}_{\mathsf{Bayes}}(\cdot;\hat{\pi})$
- NPMLE is a principled way of learning prior from data.

Empirical Bayes estimator

Bayes estimator $\hat{\theta}_{\text{Bayes}}(\cdot;\pi)$ depends on the unknown π.

Robbins' meta-principle

- Learn the prior $\hat{\pi}$ (empirical distribution) from data
- Execute Bayes strategy with learned prior $\hat{\theta}_{\text{Bayes}}(\cdot;\hat{\pi})$
- NPMLE is a principled way of learning prior from data.

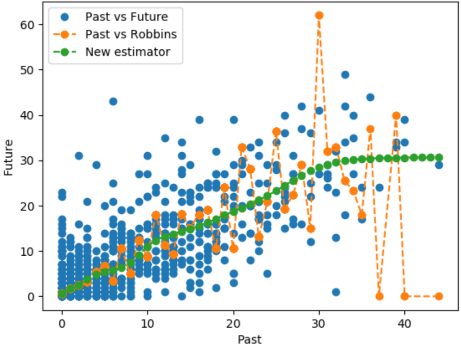
Robbins' ad hoc scheme for **Poisson** model

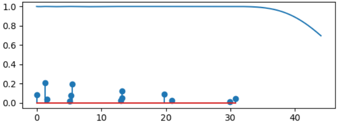
- Given $X = x$ drawn from $\mathrm{Poi}(\theta)$ and $\theta \sim \pi$,

$$\hat{\theta}_{\text{Bayes}}(x;\pi) = (x+1)\frac{p_\pi(x+1)}{p_\pi(x)} \implies \hat{\theta}_{\text{Robbins}}(x) = (x+1)\frac{p_{\text{emp}}(x+1)}{p_{\text{emp}}(x)}$$

A real-data experiment

- NHL data: goals of a player in season 2017 and 2018
- NPMLE is much more stable than Robbins





Data: `https://www.hockey-reference.com/leagues/NHL_2019_skaters-advanced.html`

Optimal regret in EB

$$\mathrm{Regret}_n = \inf_{\hat\theta} \sup_{\pi\in\Pi} \sum_{i=1}^n \{\mathbb{E}_\pi[(\hat\theta_i - \theta_i)^2] - R_{\mathsf{Bayes}}(\pi)\}$$

- Robbins showed sublinear regret $\mathrm{Regret}_n = o(n)$ is possible (aka "borrowing strength", or "learning from experience of others")
- ... but as we saw his estimator is very finicky.
- ... so many improvements over the years.

Optimal regret in EB

$$\mathrm{Regret}_n = \inf_{\hat\theta} \sup_{\pi \in \Pi} \sum_{i=1}^{n} \{\mathbb{E}_\pi[(\hat\theta_i - \theta_i)^2] - R_{\mathsf{Bayes}}(\pi)\}$$

- Robbins showed sublinear regret $\mathrm{Regret}_n = o(n)$ is possible (aka "borrowing strength", or "learning from experience of others")
- ... but as we saw his estimator is very finicky.
- ... so many improvements over the years.
- Question 1: Does NPMLE provably improve over Robbins?
- Question 2: How does Regret_n scale with n?

Optimal regret in EB

$$\text{Regret}_n = \inf_{\hat\theta} \sup_{\pi\in\Pi} \sum_{i=1}^{n} \{\mathbb{E}_\pi[(\hat\theta_i - \theta_i)^2] - R_{\text{Bayes}}(\pi)\}$$

Theorem (P.-Wu '20)

Consider compactly supported priors and $P_\theta = \mathrm{Poi}(\theta)$ *(Poisson model)*

$$\text{Regret}_n \asymp \left(\frac{\log n}{\log\log n}\right)^2$$

and achieved by Robbins' estimator.

Optimal regret in EB

$$\text{Regret}_n = \inf_{\hat\theta} \sup_{\pi \in \Pi} \sum_{i=1}^n \{\mathbb{E}_\pi[(\hat\theta_i - \theta_i)^2] - R_{\text{Bayes}}(\pi)\}$$

Theorem (P.-Wu '20)

Consider compactly supported priors and $P_\theta = \mathrm{Poi}(\theta)$ (Poisson model)

$$\text{Regret}_n \asymp \left(\frac{\log n}{\log \log n}\right)^2$$

and achieved by Robbins' estimator.

- Long-standing conjecture was to prove an $\omega(1)$ lower bound [Singh '79]
- Upper bound via Robbins is from [Brown-Greenshtein-Ritov '13]
- For $P_\theta = N(\theta, 1)$ (normal means) we show $\text{Regret}_n \gtrsim \left(\frac{\log n}{\log \log n}\right)^2$

  Best upper bound $O(\log^5 n)$ by NPMLE [Jiang-Zhang '09]

Concluding remarks

Main result:

- Self-regularizng property of NPMLE for certain mixture models: automatically tunes to the correct model size

Many open problems

- Better self-regularization with constraints
- Theory and algorithms for NPMLE in multiple/high dimensions
- Regret optimality of NPMLE in empirical Bayes

References

- Y. Polyanskiy and W. *Self-regularizing Property of Nonparametric Maximum Likelihood Estimator in Mixture Models*, arxiv:2008.08244.
- Y. Polyanskiy and W. *Sharp regret bound for empirical Bayes and compound decision problems (or: The optimality of Robbins' scheme)*, draft, 2020.