TURF: A Two-factor, Universal, Robust, Fast Distribution Learning Algorithm

Yi Hao, Ayush Jain, Alon Orlitsky, Vaishakh Ravindrakumar

ICML, 2022

Overview

Overview

- Learning distributions from samples - fundamental statistical problem

Overview

- Learning distributions from samples - fundamental statistical problem
- Applications - epidemiology, language modeling, GMMs in clustering

Slides at: www.vaishakhr.com/turf.pdf

Overview

- Learning distributions from samples - fundamental statistical problem
- Applications - epidemiology, language modeling, GMMs in clustering
- Non-parametric estimation of any arbitrary univariate distribution, f: discrete + continuous density

Slides at: www.vaishakhr.com/turf.pdf

Overview

- Learning distributions from samples - fundamental statistical problem
- Applications - epidemiology, language modeling, GMMs in clustering
- Non-parametric estimation of any arbitrary univariate distribution, f : discrete + continuous density
- Estimator f^{est} approximates f in ℓ_1 distance - $||f^{est} - f||_1$

Slides at: www.vaishakhr.com/turf.pdf

Overview

- Learning distributions from samples - fundamental statistical problem
- Applications - epidemiology, language modeling, GMMs in clustering
- Non-parametric estimation of any arbitrary univariate distribution, f : discrete + continuous density
- Estimator f^{est} approximates f in ℓ_1 distance - $||f^{est} - f||_1$
- Piecewise polynomial approximation to construct f^{est}

Slides at: www.vaishakhr.com/turf.pdf

A primer on ℓ_1 distance

Slides at: www.vaishakhr.com/turf.pdf

A primer on ℓ_1 distance

- Arbitrary distributions can't be learnt in ℓ_1 distance with finite samples

Slides at: www.vaishakhr.com/turf.pdf

A primer on ℓ_1 distance

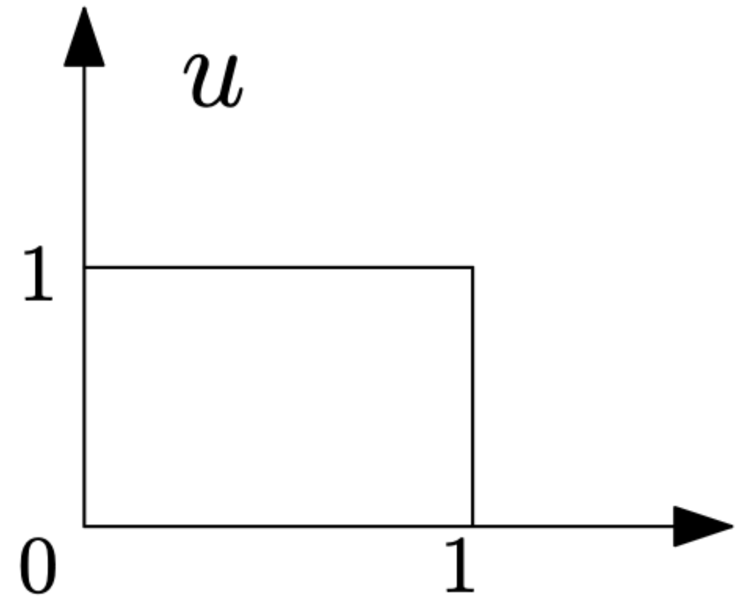
- Arbitrary distributions can't be learnt in ℓ_1 distance with finite samples
- u is the uniform distribution on $[0,1]$

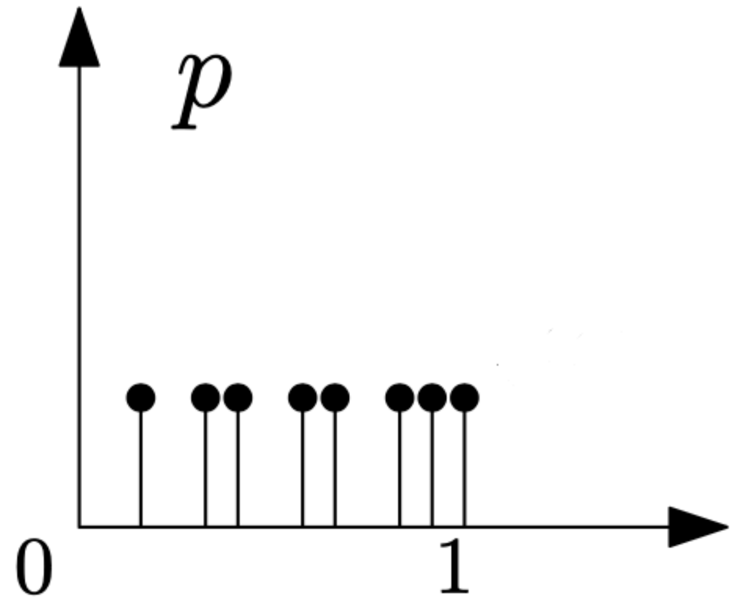
A primer on ℓ_1 distance

- Arbitrary distributions can't be learnt in ℓ_1 distance with finite samples
- u is the uniform distribution on $[0,1]$
- p is constructed by drawing $k > 0$ samples from u and assigning a mass $1/k$ at each sample location

A primer on ℓ_1 distance

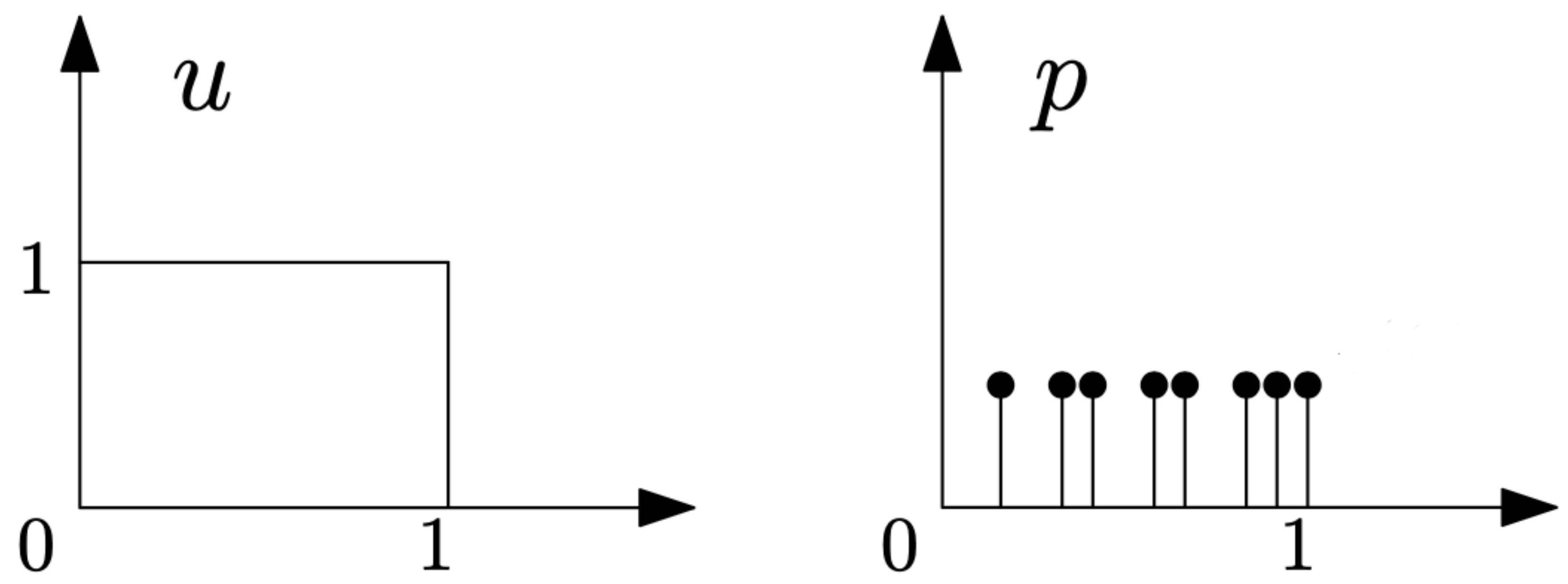
- Arbitrary distributions can't be learnt in ℓ_1 distance with finite samples
- u is the uniform distribution on $[0,1]$
- p is constructed by drawing $k > 0$ samples from u and assigning a mass $1/k$ at each sample location
-





Slides at: www.vaishakhr.com/turf.pdf

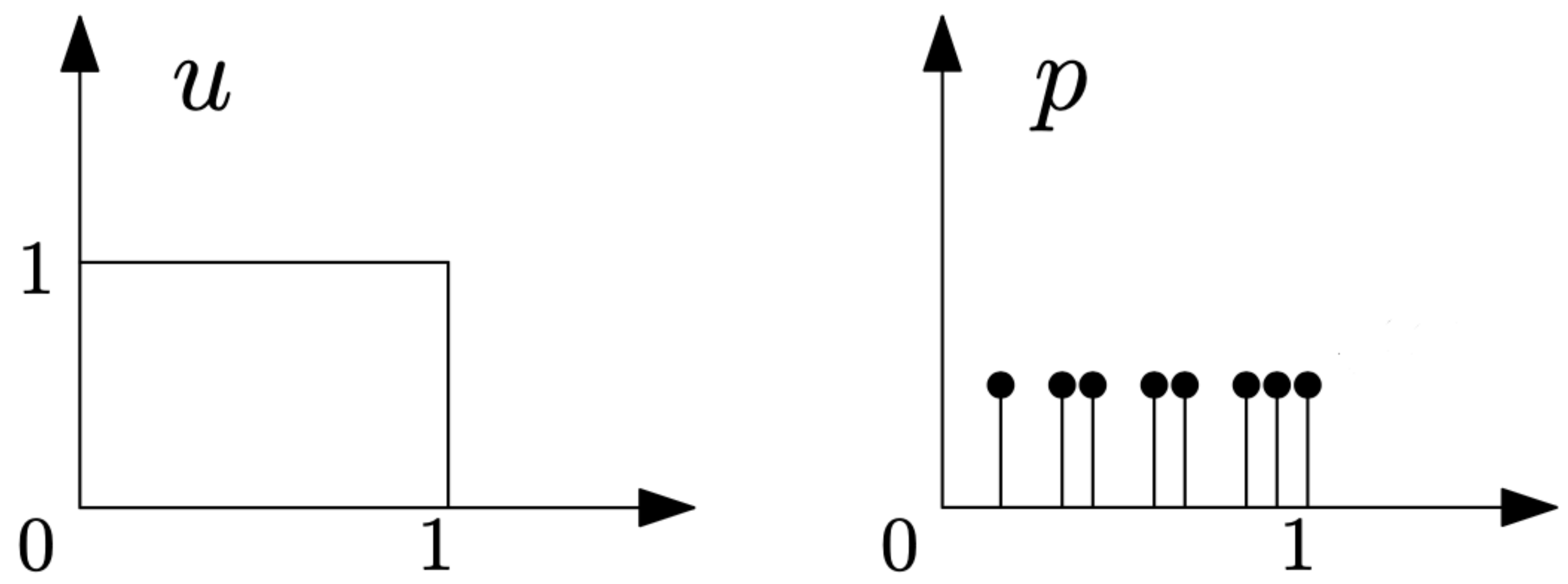
A primer on ℓ_1 distance

- Arbitrary distributions can't be learnt in ℓ_1 distance with finite samples
- u is the uniform distribution on $[0,1]$
- p is constructed by drawing $k > 0$ samples from u and assigning a mass $1/k$ at each sample location
- 

- u and p cannot be distinguished using any $n > 0$ many samples if $k \gg n$

Slides at: www.vaishakhr.com/turf.pdf

A primer on ℓ_1 distance

- Arbitrary distributions can't be learnt in ℓ_1 distance with finite samples
- u is the uniform distribution on $[0,1]$
- p is constructed by drawing $k > 0$ samples from u and assigning a mass $1/k$ at each sample location
- 

- u and p cannot be distinguished using any $n > 0$ many samples if $k \gg n$
- Since $\|u - p\|_1 = 2$ any f^{est} suffers either $\|f^{est} - u\|_1 \geq 1$ or $\|f^{est} - p\|_1 \geq 1$

Slides at: www.vaishakhr.com/turf.pdf

Piecewise polynomial approximation

Piecewise polynomial approximation

- While arbitrary f can't be learnt in ℓ_1 distance structured distributions can be learnt

Piecewise polynomial approximation

- While arbitrary f can't be learnt in ℓ_1 distance structured distributions can be learnt
- Idea is to approximate f with the class of t-piece degree-d polynomials - $\mathscr{P}_{t,d}$

Piecewise polynomial approximation

- While arbitrary f can't be learnt in ℓ_1 distance structured distributions can be learnt
- Idea is to approximate f with the class of t-piece degree-d polynomials - $\mathscr{P}_{t,d}$
- $\mathscr{P}_{t,d}$: t disjoint intervals. In each interval it is a degree-d polynomial. E.g. $\mathscr{P}_{t,0}$ is the t-piece histogram class

Slides at: www.vaishakhr.com/turf.pdf

Piecewise polynomial approximation

- While arbitrary f can't be learnt in ℓ_1 distance structured distributions can be learnt
- Idea is to approximate f with the class of t-piece degree-d polynomials - $\mathscr{P}_{t,d}$
- $\mathscr{P}_{t,d}$: t disjoint intervals. In each interval it is a degree-d polynomial. E.g. $\mathscr{P}_{t,0}$ is the t-piece histogram class
- If $f \in \mathscr{P}_{t,d}$, with n samples, we'd like to learn to the min-max rate of the class - $\mathscr{R}_n(\mathscr{P}_{t,d}) = \mathcal{O}\left(\sqrt{t(d+1)/n}\right)$

Slides at: www.vaishakhr.com/turf.pdf

Piecewise polynomial approximation

- While arbitrary f can't be learnt in ℓ_1 distance structured distributions can be learnt
- Idea is to approximate f with the class of t-piece degree-d polynomials - $\mathcal{P}_{t,d}$
- $\mathcal{P}_{t,d}$: t disjoint intervals. In each interval it is a degree-d polynomial. E.g. $\mathcal{P}_{t,0}$ is the t-piece histogram class
- If $f \in \mathcal{P}_{t,d}$, with n samples, we'd like to learn to the min-max rate of the class - $\mathcal{R}_n(\mathcal{P}_{t,d}) = \mathcal{O}\left(\sqrt{t(d+1)/n}\right)$
- For general $f \notin \mathcal{P}_{t,d}$ learn to closest approximation distance $\|f - \mathcal{P}_{t,d}\|_1$ plus $\mathcal{O}\left(\sqrt{t(d+1)/n}\right)$

Slides at: www.vaishakhr.com/turf.pdf

Piecewise polynomial approximation

- While arbitrary f can't be learnt in ℓ_1 distance structured distributions can be learnt
- Idea is to approximate f with the class of t-piece degree-d polynomials - $\mathscr{P}_{t,d}$
- $\mathscr{P}_{t,d}$: t disjoint intervals. In each interval it is a degree-d polynomial. E.g. $\mathscr{P}_{t,0}$ is the t-piece histogram class
- If $f \in \mathscr{P}_{t,d}$, with n samples, we'd like to learn to the min-max rate of the class - $\mathscr{R}_n(\mathscr{P}_{t,d}) = \mathcal{O}\left(\sqrt{t(d+1)/n}\right)$
- For general $f \notin \mathscr{P}_{t,d}$ learn to closest approximation distance $\|f - \mathscr{P}_{t,d}\|_1$ plus $\mathcal{O}\left(\sqrt{t(d+1)/n}\right)$
- We'd like an estimator f^{est} such that $\mathbb{E}\|f^{est} - f\|_1 \leq c \cdot \|f - \mathscr{P}_{t,d}\|_1 + \mathcal{O}(\mathscr{R}_n(\mathscr{P}_{t,d}))$

Slides at: www.vaishakhr.com/turf.pdf

Piecewise polynomial approximation

- While arbitrary f can't be learnt in ℓ_1 distance structured distributions can be learnt
- Idea is to approximate f with the class of t-piece degree-d polynomials - $\mathscr{P}_{t,d}$
- $\mathscr{P}_{t,d}$: t disjoint intervals. In each interval it is a degree-d polynomial. E.g. $\mathscr{P}_{t,0}$ is the t-piece histogram class
- If $f \in \mathscr{P}_{t,d}$, with n samples, we'd like to learn to the min-max rate of the class - $\mathscr{R}_n(\mathscr{P}_{t,d}) = \mathscr{O}\left(\sqrt{t(d+1)/n}\right)$
- For general $f \notin \mathscr{P}_{t,d}$ learn to closest approximation distance $\|f - \mathscr{P}_{t,d}\|_1$ plus $\mathscr{O}\left(\sqrt{t(d+1)/n}\right)$
- We'd like an estimator f^{est} such that $\mathbb{E}\|f^{est} - f\|_1 \leq c \cdot \|f - \mathscr{P}_{t,d}\|_1 + \mathscr{O}(\mathscr{R}_n(\mathscr{P}_{t,d}))$
- We call it a c-factor approximation for $\mathscr{P}_{t,d}$

Slides at: www.vaishakhr.com/turf.pdf

Applications of c-factor estimator

Applications of c-factor estimator

- Why 'piecewise' polynomials? Versatile approximation objects.

$\mathscr{P}_{t,d}$: t - # pieces, d - degree

Applications of c-factor estimator

- Why 'piecewise' polynomials? Versatile approximation objects.

- Suppose f^{est} is a c-factor estimator for $\mathscr{P}_{t,d}$. That is,
  $\mathbb{E}\|f^{est} - f\|_1 \leq c \cdot \|f - \mathscr{P}_{t,d}\|_1 + \mathcal{O}(\mathscr{R}_n(\mathscr{P}_{t,d}))$

Slides at: www.vaishakhr.com/turf.pdf

$\mathscr{P}_{t,d}$: t - # pieces, d - degree

Applications of c-factor estimator

- Why 'piecewise' polynomials? Versatile approximation objects.
- Suppose f^{est} is a c-factor estimator for $\mathscr{P}_{t,d}$. That is,
  $\mathbb{E}\|f^{est} - f\|_1 \leq c \cdot \|f - \mathscr{P}_{t,d}\|_1 + \mathcal{O}(\mathscr{R}_n(\mathscr{P}_{t,d}))$
- Depending on prior knowledge of f, use approximation theory to bound $\|f - \mathscr{P}_{t,d}\|_1$

$\mathscr{P}_{t,d}$: t - # pieces, d - degree

Applications of c-factor estimator

- Why 'piecewise' polynomials? Versatile approximation objects.
- Suppose f^{est} is a c-factor estimator for $\mathscr{P}_{t,d}$. That is,
  $\mathbb{E}\|f^{est} - f\|_1 \leq c \cdot \|f - \mathscr{P}_{t,d}\|_1 + \mathcal{O}(\mathscr{R}_n(\mathscr{P}_{t,d}))$
- Depending on prior knowledge of f, use approximation theory to bound $\|f - \mathscr{P}_{t,d}\|_1$
- If f is unimodal, choose $t = \mathcal{O}(n^{1/3}),\ d = 0$ to achieve uni-modal's min-max rate $\mathcal{O}(1/n^{1/3})$

Slides at: www.vaishakhr.com/turf.pdf

$\mathscr{P}_{t,d}$: t - # pieces, d - degree

Applications of c-factor estimator

- Why 'piecewise' polynomials? Versatile approximation objects.

- Suppose f^{est} is a c-factor estimator for $\mathscr{P}_{t,d}$. That is, $\mathbb{E}\|f^{est} - f\|_1 \leq c \cdot \|f - \mathscr{P}_{t,d}\|_1 + \mathcal{O}(\mathscr{R}_n(\mathscr{P}_{t,d}))$

- Depending on prior knowledge of f, use approximation theory to bound $\|f - \mathscr{P}_{t,d}\|_1$

- If f is unimodal, choose $t = \mathcal{O}(n^{1/3}),\ d = 0$ to achieve uni-modal's min-max rate $\mathcal{O}(1/n^{1/3})$

- If f log-concave, choose $t = \mathcal{O}(n^{1/5}),\ d = 1$ to achieve log-concave's min-max rate of $\mathcal{O}(1/n^{2/5})$

Slides at: www.vaishakhr.com/turf.pdf

$\mathscr{P}_{t,d}$: t - # pieces, d - degree

Applications of c-factor estimator

- Why 'piecewise' polynomials? Versatile approximation objects.
- Suppose f^{est} is a c-factor estimator for $\mathscr{P}_{t,d}$. That is,
  $\mathbb{E}\|f^{est} - f\|_1 \leq c \cdot \|f - \mathscr{P}_{t,d}\|_1 + \mathcal{O}(\mathscr{R}_n(\mathscr{P}_{t,d}))$
- Depending on prior knowledge of f, use approximation theory to bound $\|f - \mathscr{P}_{t,d}\|_1$
- If f is unimodal, choose $t = \mathcal{O}(n^{1/3}),\ d = 0$ to achieve uni-modal's min-max rate $\mathcal{O}(1/n^{1/3})$
- If f log-concave, choose $t = \mathcal{O}(n^{1/5}),\ d = 1$ to achieve log-concave's min-max rate of $\mathcal{O}(1/n^{2/5})$
- Similarly Gaussian and their mixtures

Slides at: www.vaishakhr.com/turf.pdf

$\mathscr{P}_{t,d}$: t - # pieces, d - degree

Contribution - approximation factor

$\mathscr{P}_{t,d}$: t - # pieces, d - degree

Contribution - approximation factor

- We'd like an f^{est} such that $\mathbb{E}||f^{est} - f||_1 \leq c \cdot ||f - \mathcal{P}_{t,d}||_1 + \mathcal{O}(\mathcal{R}_n(\mathcal{P}_{t,d}))$

Slides at: www.vaishakhr.com/turf.pdf

$\mathcal{P}_{t,d}$: t - # pieces, d - degree

Contribution - approximation factor

- We'd like an f^{est} such that $\mathbb{E}||f^{est} - f||_1 \leq c \cdot ||f - \mathscr{P}_{t,d}||_1 + \mathcal{O}(\mathscr{R}_n(\mathscr{P}_{t,d}))$
- For any f^{est}, $c \geq 2$ if $t \geq 2,\ d \geq 0$ [Chan et. al. 14]. In this work we show $c \geq 2$ even if $t \geq 1,\ d \geq 1$

Slides at: www.vaishakhr.com/turf.pdf

$\mathscr{P}_{t,d}$: t - # pieces, d - degree

Contribution - approximation factor

- We'd like an f^{est} such that $\mathbb{E}||f^{est} - f||_1 \leq c \cdot ||f - \mathscr{P}_{t,d}||_1 + \mathcal{O}(\mathscr{R}_n(\mathscr{P}_{t,d}))$
- For any f^{est}, $c \geq 2$ if $t \geq 2,\ d \geq 0$ [Chan et. al. 14]. In this work we show $c \geq 2$ even if $t \geq 1,\ d \geq 1$
- Previous estimators achieve $c \in [2.25, 3]$ [Yatracos 85, Acharya et. al. 15, Hao et. al. 20] w.r.t. $\mathscr{P}_{t,d}$ depending on the degree d for any # pieces $t \geq 1$

Slides at: www.vaishakhr.com/turf.pdf

$\mathscr{P}_{t,d}$: t - # pieces, d - degree

Contribution - approximation factor

- We'd like an f^{est} such that $\mathbb{E}\|f^{est} - f\|_1 \leq c \cdot \|f - \mathscr{P}_{t,d}\|_1 + \mathcal{O}(\mathscr{R}_n(\mathscr{P}_{t,d}))$
- For any f^{est}, $c \geq 2$ if $t \geq 2,\ d \geq 0$ [Chan et. al. 14]. In this work we show $c \geq 2$ even if $t \geq 1,\ d \geq 1$
- Previous estimators achieve $c \in [2.25, 3]$ [Yatracos 85, Acharya et. al. 15, Hao et. al. 20] w.r.t. $\mathscr{P}_{t,d}$ depending on the degree d for any # pieces $t \geq 1$
- In this work we achieve the optimal $c = 2$

$\mathscr{P}_{t,d}$: t - # pieces, d - degree

Contributions - crossvalidation

$\mathscr{P}_{t,d}$: t - # pieces, d - degree

Contributions - crossvalidation

- We had chosen $t = \mathcal{O}(n^{1/3})$ to achieve uni-modal's min-max rate. But if log-concave, need to choose $t = \mathcal{O}(n^{1/5})$ to achieve its min-max rate

Contributions - crossvalidation

- We had chosen $t = \mathcal{O}(n^{1/3})$ to achieve uni-modal's min-max rate. But if log-concave, need to choose $t = \mathcal{O}(n^{1/5})$ to achieve its min-max rate

- In general we do not know anything about f i.e. whether or not it is unimodal

Contributions - crossvalidation

- We had chosen $t = \mathcal{O}(n^{1/3})$ to achieve uni-modal's min-max rate. But if log-concave, need to choose $t = \mathcal{O}(n^{1/5})$ to achieve its min-max rate
- In general we do not know anything about f i.e. whether or not it is unimodal
- Choice of t varies significantly so how to select t?

Contributions - crossvalidation

- We had chosen $t = \mathcal{O}(n^{1/3})$ to achieve uni-modal's min-max rate. But if log-concave, need to choose $t = \mathcal{O}(n^{1/5})$ to achieve its min-max rate
- In general we do not know anything about f i.e. whether or not it is unimodal
- Choice of t varies significantly so how to select t?
- Given c-factor estimates f_t^{est} for $t \in \{1,2,\ldots,n\}$, we'd like to estimate t^{est} such that for some $c' \geq c$, $\mathbb{E}\|f_{t^{est}}^{est} - f\|_1 \leq \min_{0 \leq t \leq n} \left(c' \cdot \|f - \mathcal{P}_{t,d}\|_1 + \mathcal{O}\left(\mathcal{R}_n(\mathcal{P}_{t,d})\right)\right)$

Slides at: www.vaishakhr.com/turf.pdf

$\mathcal{P}_{t,d}$: t - # pieces, d - degree

Contributions - crossvalidation

- We had chosen $t = \mathcal{O}(n^{1/3})$ to achieve uni-modal's min-max rate. But if log-concave, need to choose $t = \mathcal{O}(n^{1/5})$ to achieve its min-max rate
- In general we do not know anything about f i.e. whether or not it is unimodal
- Choice of t varies significantly so how to select t?
- Given c-factor estimates f_t^{est} for $t \in \{1,2,\ldots,n\}$, we'd like to estimate t^{est} such that for some $c' \geq c$, $\mathbb{E}\|f_{t^{est}}^{est} - f\|_1 \leq \min_{0 \leq t \leq n} \left(c' \cdot \|f - \mathcal{P}_{t,d}\|_1 + \mathcal{O}\left(\mathcal{R}_n(\mathcal{P}_{t,d})\right)\right)$
- Previous works achieve $c' = 3c$. We achieve the optimal $c' = c$

Slides at: www.vaishakhr.com/turf.pdf

$\mathcal{P}_{t,d}$: t - # pieces, d - degree

Thank you!