Mean Estimation from One-bit Measurements

Alon Kipnis (Stanford)
John Duchi (Stanford)

Allerton
October 2017

Table of Contents

Motivation

Point estimation under communication constraints:

$$p_{\theta}(x) \rightarrow X_1, \ldots, X_n \longrightarrow \boxed{\begin{array}{l}\text{limited bits}\\ \text{per sample}\end{array}} \longrightarrow \widehat{\theta}$$

Motivation

Point estimation under communication constraints:

$$p_{\theta}(x) \rightarrow X_1, \ldots, X_n \longrightarrow \boxed{\begin{array}{l}\text{limited bits}\\ \text{per sample}\end{array}} \longrightarrow \widehat{\theta}$$

Estimation error is due to:

(i) limited data

(ii) limited bits

Motivation

Point estimation under communication constraints:

$$p_{\theta}(x) \rightarrow X_1, \ldots, X_n \longrightarrow \boxed{\begin{array}{l}\text{limited bits}\\ \text{per sample}\end{array}} \longrightarrow \widehat{\theta}$$

Estimation error is due to:

(i) limited data

(ii) limited bits

Relevant scenarios:

Motivation

Point estimation under communication constraints:

$$p_{\theta}(x) \to X_1, \ldots, X_n \longrightarrow \boxed{\begin{array}{l}\text{limited bits}\\ \text{per sample}\end{array}} \longrightarrow \widehat{\theta}$$

Estimation error is due to:

(i) limited data

(ii) limited bits

Relevant scenarios:

- big data

Motivation

Point estimation under communication constraints:

$$p_{\theta}(x) \rightarrow X_1, \ldots, X_n \longrightarrow \boxed{\text{limited bits per sample}} \longrightarrow \widehat{\theta}$$

Estimation error is due to:

(i) limited data

(ii) limited bits

Relevant scenarios:

- big data
- low-power sensors

Motivation

Point estimation under communication constraints:

$$p_{\theta}(x) \rightarrow X_1, \ldots, X_n \longrightarrow \boxed{\text{limited bits per sample}} \longrightarrow \widehat{\theta}$$

Estimation error is due to:

(i) limited data

(ii) limited bits

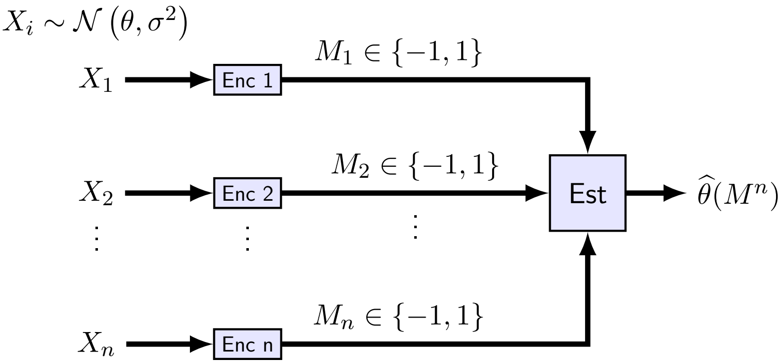
Relevant scenarios:

- big data
- low-power sensors
- distributed computing / optimization

This talk:

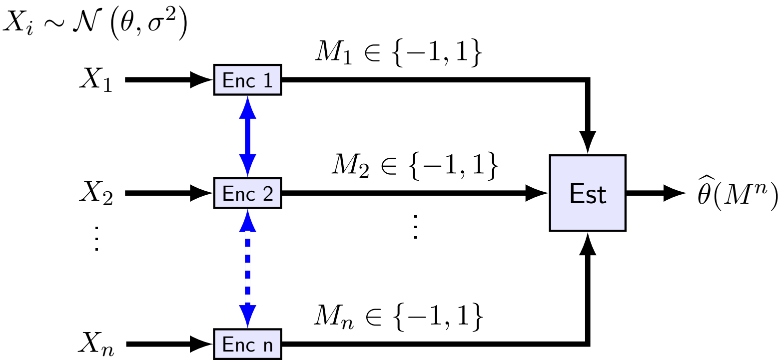
Estimating the mean θ of a normal distribution $\mathcal{N}(\theta, \sigma^2)$ from one-bit per sample (σ is known)

Three Encoding Scenarios



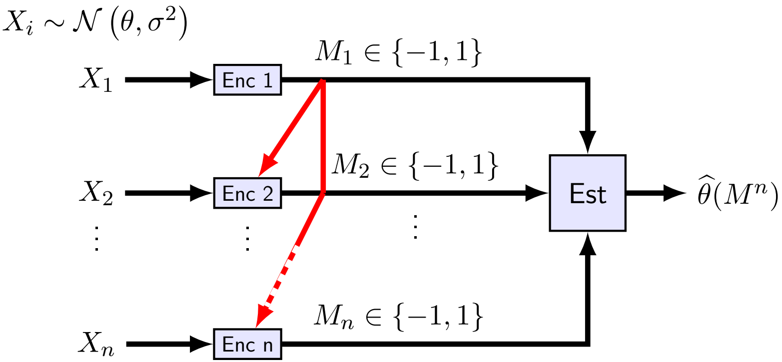
- Distributed: $M_i = f_i(X_i)$

Three Encoding Scenarios



- Distributed: $M_i = f_i(X_i)$
- Centralized: $M^n = (M_1, \ldots, M_n) = f(X_1, \ldots, X_n)$

Three Encoding Scenarios



- Distributed: $M_i = f_i(X_i)$
- Centralized: $M^n = (M_1, \ldots, M_n) = f(X_1, \ldots, X_n)$
- Adaptive / Sequential: $M_i = f_i(X_i, M^{i-1})$

Related Work

Related Work

- Estimation via compressed information [Han '87], [Zhang & Berger '88] (centralized)

Related Work

- Estimation via compressed information [Han '87], [Zhang & Berger '88] (centralized)
- Estimation from multiple machines subject to a bit constraint [Zhang, Duchi, Jordan, Wainwright '13] (distributed / adaptive)

Related Work

- Estimation via compressed information [Han '87], [Zhang & Berger '88] (centralized)
- Estimation from multiple machines subject to a bit constraint [Zhang, Duchi, Jordan, Wainwright '13] (distributed / adaptive)
- Remote multiterminal source coding (CEO) [Berger, Zhang, Wiswanathan '96], [Oohama '97] (distributed)

Consistency

Q: in what setting consistent estimation is possible?

Consistency

Q: in what setting consistent estimation is possible?

A: all !

Consistency

Q: in what setting consistent estimation is possible?

A: all ! **Proof:**

$$M_i = \mathbf{1}(X_i > 0), \quad i = 1, \ldots, n$$

(distributed setting)

Consistency

Q: in what setting consistent estimation is possible?

A: all ! **Proof:**

$$M_i = \mathbf{1}(X_i > 0), \quad i = 1, \ldots, n$$

(distributed setting)

$$\frac{1}{n} \sum_{i=1}^{n} M_i \to \mathbb{P}\left(X > 0\right) = \Phi\left(\theta / \sigma\right)$$

Consistency

Q: in what setting consistent estimation is possible?

A: all ! **Proof:**

$$M_i = \mathbf{1}(X_i > 0), \quad i = 1, \ldots, n$$

(distributed setting)

$$\frac{1}{n} \sum_{i=1}^{n} M_i \to \mathbb{P}\left(X > 0\right) = \Phi\left(\theta / \sigma\right)$$

$$\sigma \Phi^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} M_i \right) \to \theta$$

Efficiency

Definition: *asymptotic relative efficiency (ARE)*:

$$\mathrm{ARE}(\widehat{\theta}_n) \triangleq \lim_{n\to\infty} \mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right] / \left(\sigma^2/n\right)$$

(relative to sample mean $\bar{\theta}$)

Efficiency

Definition: *asymptotic relative efficiency (ARE)*:

$$\mathrm{ARE}(\widehat{\theta}_n) \triangleq \lim_{n\to\infty} \mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right] / \left(\sigma^2/n\right)$$

(relative to sample mean $\bar{\theta}$)

Proposition

ARE under centralized encoding is 1

Efficiency

Definition: *asymptotic relative efficiency (ARE)*:

$$\mathrm{ARE}(\widehat{\theta}_n) \triangleq \lim_{n\to\infty} \mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right] / \left(\sigma^2/n\right)$$

(relative to sample mean $\bar{\theta}$)

Proposition

ARE under centralized encoding is 1

Proof:

$$\mathbb{E}\left(\theta - \widehat{\theta}\right)^2 = \overbrace{\mathbb{E}\left(\theta - \bar{\theta}\right)^2}^{\sigma^2/n} + \mathbb{E}\left(\bar{\theta} - \widehat{\theta}\right)^2$$

Efficiency

Definition: *asymptotic relative efficiency (ARE)*:

$$\mathrm{ARE}(\widehat{\theta}_n) \triangleq \lim_{n\to\infty} \mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right] / \left(\sigma^2/n\right)$$

(relative to sample mean $\bar{\theta}$)

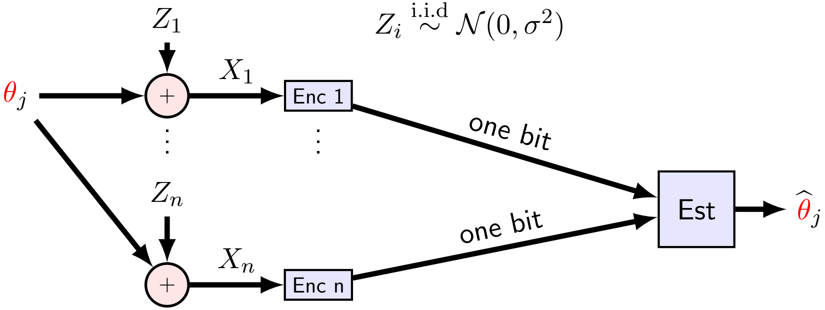
Proposition

ARE under centralized encoding is 1

Proof:

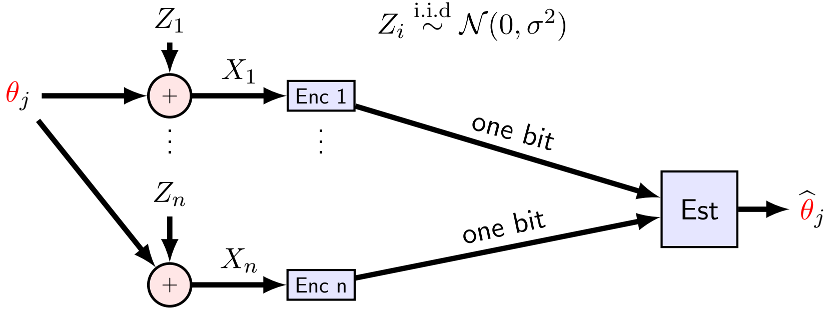
$$\mathbb{E}\left(\theta - \widehat{\theta}\right)^2 = \overbrace{\mathbb{E}\left(\theta - \bar{\theta}\right)^2}^{\sigma^2/n} + \underbrace{\mathbb{E}\left(\bar{\theta} - \widehat{\theta}\right)^2}_{O(2^{-2n})}$$

Relation to CEO



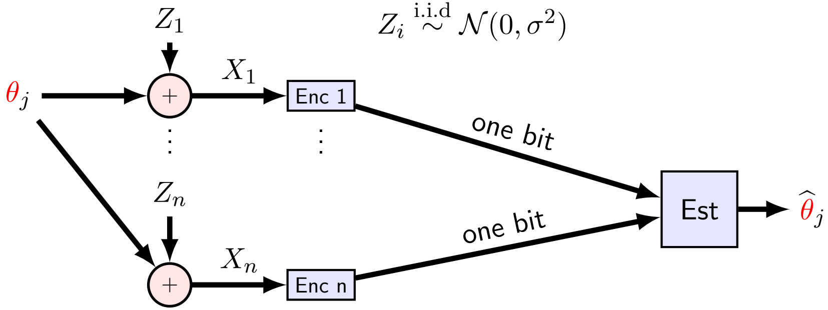
- $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(\theta, \sigma^2)$

Relation to CEO



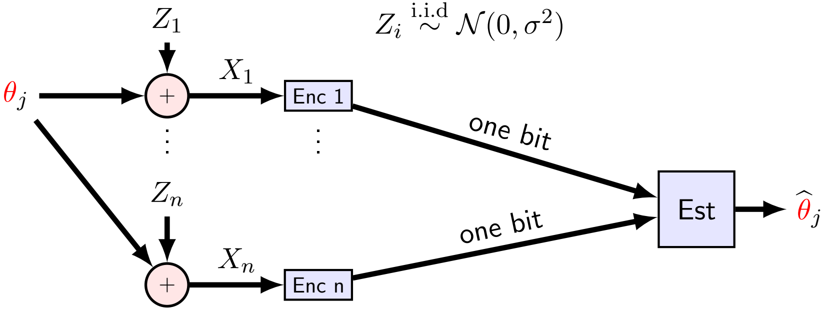
- $X_1,\ldots,X_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(\theta,\sigma^2)$
- Replicate k times:
  - $\theta_1,\ldots,\theta_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma_0^2)$
  - $\theta_j \to X_{j,1},\ldots,X_{j,n}$

Relation to CEO



- $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(\theta, \sigma^2)$
- Replicate k times:
  - $\theta_1, \ldots, \theta_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_0^2)$
  - $\theta_j \to X_{j,1}, \ldots, X_{j,n}$
- Encoder i block encodes $X_{1,i}, \ldots, X_{k,i}$ using k bits

Relation to CEO



- $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(\theta, \sigma^2)$
- Replicate k times:
  - $\theta_1, \ldots, \theta_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_0^2)$
  - $\theta_j \to X_{j,1}, \ldots, X_{j,n}$
- Encoder i block encodes $X_{1,i}, \ldots, X_{k,i}$ using k bits
- $D_{CEO} = \inf_k \frac{1}{k} \sum_{j=1}^k \mathbb{E} \left(\theta_j - \widehat{\theta}_j \right)^2$

Quadratic Gaussian CEO under optimal rate allocation [Chen, Zhang, Berger, Wicker '04] :

$$D_{CEO} \geq \frac{4}{3}\frac{\sigma^2}{n} + o(1/n)$$

Quadratic Gaussian CEO under optimal rate allocation [Chen, Zhang, Berger, Wicker '04] :

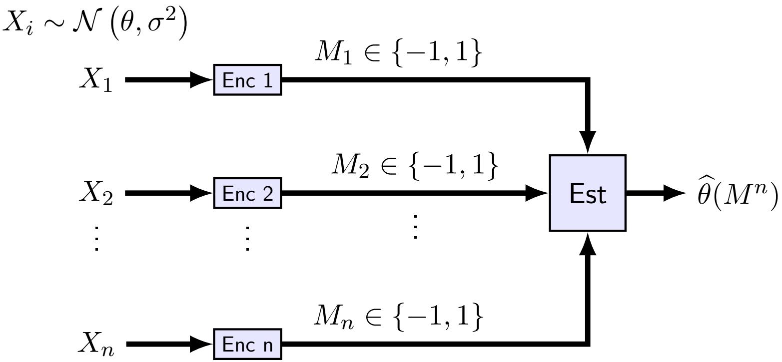
$$D_{CEO} \geq \frac{4}{3}\frac{\sigma^2}{n} + o(1/n)$$

Conclusion

Distributed encoding will hurt you (even if you can repeat experiment and encode over blocks)

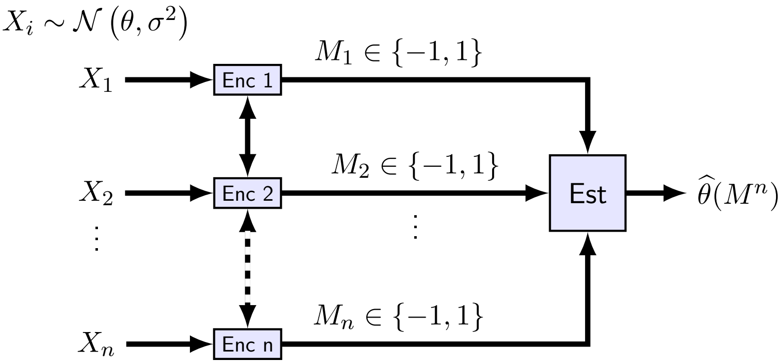
Table of Contents

Three Encoding Scenarios



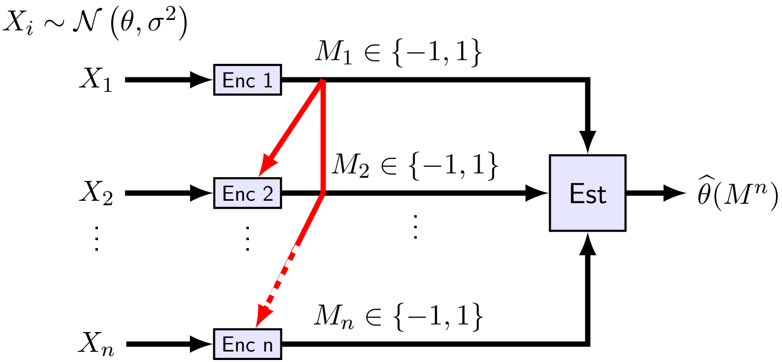
- Distributed: $M_i = f_i(X_i)$ ARE $\geq 4/3$

Three Encoding Scenarios



- Distributed: $M_i = f_i(X_i)$ ARE $\geq 4/3$
- Centralized: $M^n = f(X_1, \ldots, X_n)$ ARE $= 1$

Three Encoding Scenarios



- Distributed: $M_i = f_i(X_i)$ ARE $\geq 4/3$
- Centralized: $M^n = f(X_1, \ldots, X_n)$ ARE $= 1$
- Next: adaptive: $M_i = f_i(X_i, M^{i-1})$ ARE$= \pi/2$

Main Results (adaptive encoding)

Theorem (achievability)

There exists an estimator with ARE $\pi/2$

Theorem (converse)

No estimator have ARE lower than $\pi/2$

Achievability

existence of an estimator with ARE $= \pi/2$

(i) For $X \sim \mathcal{N}(\theta, \sigma^2)$, med$(X) = \theta$

Achievability

existence of an estimator with ARE $= \pi/2$

(i) For $X \sim \mathcal{N}(\theta, \sigma^2)$, med$(X) = \theta$

(ii) med$(X) = \operatorname*{argmin}_m \mathbb{E}\,|X - m|$

Achievability

existence of an estimator with ARE $= \pi/2$

(i) For $X \sim \mathcal{N}(\theta, \sigma^2)$, med$(X) = \theta$

(ii) med$(X) = \operatorname*{argmin}_m \mathbb{E}\,|X - m|$

(iii) Stochastic gradient descent on $\mathbb{E}\,|X - \theta|$:

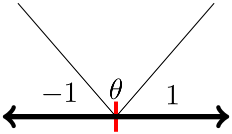
Achievability

existence of an estimator with ARE $= \pi/2$

(i) For $X \sim \mathcal{N}(\theta, \sigma^2)$, $\mathrm{med}(X) = \theta$

(ii) $\mathrm{med}(X) = \operatorname*{argmin}_m \mathbb{E}\,|X - m|$

(iii) Stochastic gradient descent on $\mathbb{E}\,|X - \theta|$:



$$\theta_n = \theta_{n-1} + \gamma_n \mathrm{sign}(X_n - \theta_n)$$

$$\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \theta_i$$

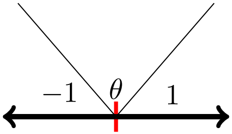
Achievability

existence of an estimator with ARE $= \pi/2$

(i) For $X \sim \mathcal{N}(\theta, \sigma^2)$, $\mathsf{med}(X) = \theta$

(ii) $\mathsf{med}(X) = \operatorname*{argmin}_m \mathbb{E}\,|X - m|$

(iii) Stochastic gradient descent on $\mathbb{E}\,|X - \theta|$:



$$\theta_n = \theta_{n-1} + \gamma_n \mathsf{sign}(X_n - \theta_n)$$

$$\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \theta_i$$

From [Polyak & Juditsky '92] (under conditions on (γ_n)):

$$\sqrt{n}(\theta - \widehat{\theta}_n) \to \mathcal{N}\left(0, \sigma^2 \pi/2\right)$$

Converse

ARE $\geq \pi/2$

The van-Trees inequality (e.g. [Tsybakov '08]) implies

$$\mathbb{E}\left(\theta - \widehat{\theta}\right)^2 \geq \frac{1}{I_\theta(M^n) + c} = \frac{1}{\sum_{i=1}^n I_\theta(M_i|M^{i-1}) + c}$$

Converse

ARE $\geq \pi/2$

The van-Trees inequality (e.g. [Tsybakov '08]) implies

$$\mathbb{E}\left(\theta - \widehat{\theta}\right)^2 \geq \frac{1}{I_\theta(M^n) + c} = \frac{1}{\sum_{i=1}^n I_\theta(M_i|M^{i-1}) + c}$$

Lemma (K. & Duchi '17)

$$I_\theta(M_i|M^{i-1}) \leq \frac{2}{\pi\sigma^2}$$

Converse

ARE $\geq \pi/2$

The van-Trees inequality (e.g. [Tsybakov '08]) implies

$$\mathbb{E}\left(\theta - \widehat{\theta}\right)^2 \geq \frac{1}{I_\theta(M^n) + c} = \frac{1}{\sum_{i=1}^n I_\theta(M_i|M^{i-1}) + c}$$

Lemma (K. & Duchi '17)

$$I_\theta(M_i|M^{i-1}) \leq \frac{2}{\pi\sigma^2}$$

Proof:
Stein's identity implies that detection region maximizing the information is a threshold: $M_i^{-1}(1) = (\theta, \infty)$

Table of Contents

Distributed Encoding

Threshold Detection

We consider only messages of the form

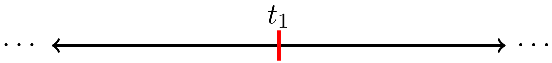
$$M_i = \mathsf{sign}(X_i - t_i), \quad i = 1, \ldots, n$$

Distributed Encoding

Threshold Detection

We consider only messages of the form

$$M_i = \mathsf{sign}(X_i - t_i), \quad i = 1, \ldots, n$$

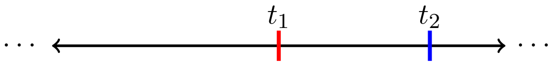


Distributed Encoding

Threshold Detection

We consider only messages of the form

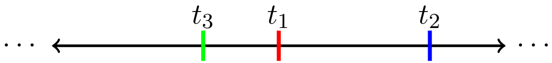
$$M_i = \mathsf{sign}(X_i - t_i), \quad i = 1, \ldots, n$$



Distributed Encoding
Threshold Detection

We consider only messages of the form

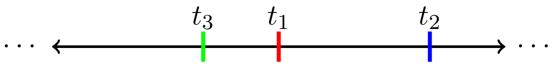
$$M_i = \mathsf{sign}(X_i - t_i), \quad i = 1, \ldots, n$$



Distributed Encoding
Threshold Detection

We consider only messages of the form

$$M_i = \mathsf{sign}(X_i - t_i), \quad i = 1, \ldots, n$$



Assume:

$$\lambda_n([a,b]) = \frac{1}{n}\mathsf{card}\left([a,b] \cap \{t_i\}\right)$$

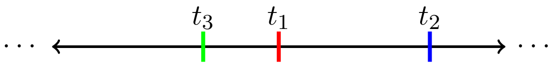
converges weakly to a probability distribution λ

Distributed Encoding

Threshold Detection

We consider only messages of the form

$$M_i = \mathsf{sign}(X_i - t_i), \quad i = 1, \ldots, n$$



Assume:

$$\lambda_n([a,b]) = \frac{1}{n}\mathsf{card}\left([a,b] \cap \{t_i\}\right)$$

converges weakly to a probability distribution λ

Example: $t_1, \ldots, t_n$ drawn i.i.d. from a distribution λ on $\mathbb{R}$

Main Results (distributed encoding)

Main Results (distributed encoding)

Theorem

(i) *The Maximum likelihood estimator* $\widehat{\theta}_{ML}$ *satisfies*

$$\sqrt{n}(\theta - \widehat{\theta}_{ML}) \to \mathcal{N}\left(0, \sigma^2 / K_\lambda(\theta)\right)$$

where:

$$K_\lambda(\theta) = \int_{\mathbb{R}} \eta\left(\frac{t-\theta}{\sigma}\right) \lambda(dt)$$

$$\eta(x) = \frac{\phi^2(x)}{\Phi(x)\Phi(-x)}$$

Main Results (distributed encoding)

Theorem

(i) *The Maximum likelihood estimator* $\widehat{\theta}_{ML}$ *satisfies*

$$\sqrt{n}(\theta - \widehat{\theta}_{ML}) \to \mathcal{N}\left(0, \sigma^2 / K_\lambda(\theta)\right)$$

where:

$$K_\lambda(\theta) = \int_{\mathbb{R}} \eta\left(\frac{t-\theta}{\sigma}\right)\lambda(dt)$$
$$\eta(x) = \frac{\phi^2(x)}{\Phi(x)\Phi(-x)}$$

(ii) *For any estimator* $\widehat{\theta}(M_1, \ldots, M_n)$*:*

$$\liminf_{c\to\infty} \liminf_{n\to\infty} \sup_{\tau\,:\,|\tau-\theta|\leq\frac{c}{\sqrt{n}}} n\mathbb{E}\left(\widehat{\theta} - \tau\right)^2 \geq \sigma^2 / K_\lambda(\theta),$$

Interpretations

Interpretations

- MLE is local asymptotically minimax

Interpretations

- MLE is local asymptotically minimax
- ARE of MLE is $1/K_\lambda(\theta)$ – only depends on asymptotic threshold density λ

Interpretations

- MLE is local asymptotically minimax
- ARE of MLE is $1/K_\lambda(\theta)$ – only depends on asymptotic threshold density λ
- $$\mathrm{ARE} = \frac{1}{K_\lambda(\theta)} = \frac{\sigma^2}{\int \eta\left(\frac{t-\theta}{\sigma}\right)\lambda(dt)} > \pi/2$$

  (equality iff $\lambda = \delta_\theta$)

Table of Contents

Summary

Summary

- Centralized encoding: ARE$= 1$

Summary

- Centralized encoding: ARE$= 1$
- Adaptive:
  - ARE is $\pi/2$

Summary

- Centralized encoding: ARE$= 1$
- Adaptive:
  - ARE is $\pi/2$
  - ~ 1.57 more samples are required due to 1-bit constraints

Summary

- Centralized encoding: ARE$= 1$
- Adaptive:
  - ARE is $\pi/2$
  - ~ 1.57 more samples are required due to 1-bit constraints
- Distributed threshold detection:
  - ARE of MLE characterized by density of threshold values

Summary

- Centralized encoding: ARE$= 1$
- Adaptive:
  - ARE is $\pi/2$
  - ~ 1.57 more samples are required due to 1-bit constraints
- Distributed threshold detection:
  - ARE of MLE characterized by density of threshold values
  - MLE is local asymptotically optimal

Summary

- Centralized encoding: ARE$= 1$
- Adaptive:
  - ARE is $\pi/2$
  - ~ 1.57 more samples are required due to 1-bit constraints
- Distributed threshold detection:
  - ARE of MLE characterized by density of threshold values
  - MLE is local asymptotically optimal

Summary

- Centralized encoding: ARE$= 1$
- Adaptive:
  - ARE is $\pi/2$
  - ~ 1.57 more samples are required due to 1-bit constraints

- Distributed threshold detection:
  - ARE of MLE characterized by density of threshold values
  - MLE is local asymptotically optimal

Open question

Is there a distributed encoding scheme with ARE that is both finite and independent of radius of Θ ?

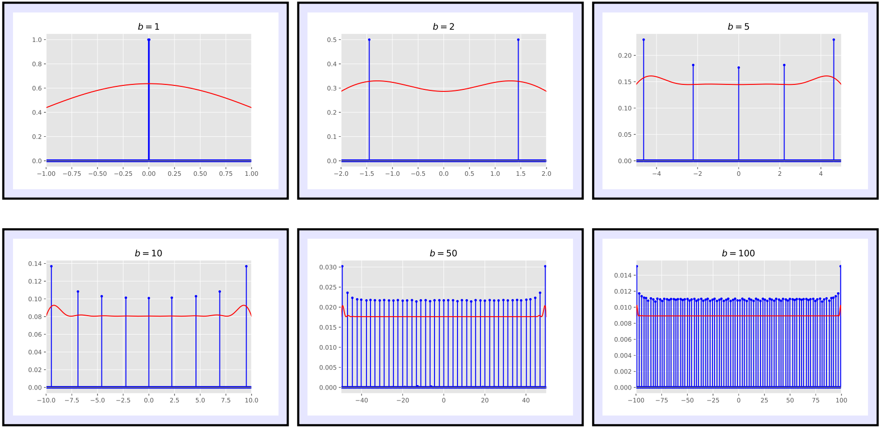
Minimax threshold density

Minimax λ for $\theta \in (-b\sigma, b\sigma)$:

$$\begin{aligned} \text{maximize} \quad & \inf_{\tau \in (-b,b)} \int \eta(t-\tau)\lambda(dt) \\ \text{subject to} \quad & \lambda(dt) \geq 0, \quad \int \lambda(dt) \leq 1. \end{aligned}$$

Minimax λ

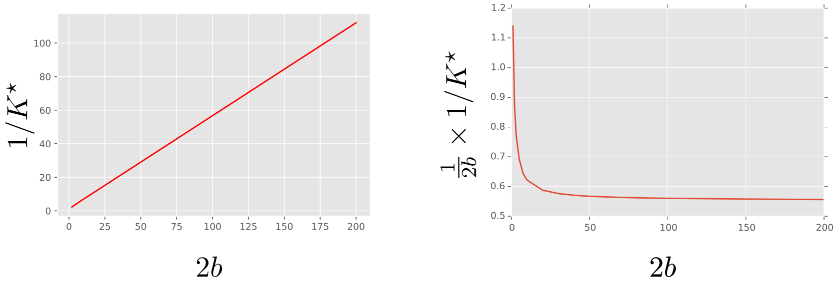
support of optimal threshold density $\lambda^\star$



$$K^\star = \inf_\theta K^\star(\theta) = \inf_\theta \int \eta(t-\theta)\lambda^\star(dt)$$

Minimax λ

Minimax ARE vs size of parameter space



- ARE increases with size of parameter space

One-step Optimal Scheme

Initialization: $P_0(t) = \pi(\theta)$
Repeat for $n \geq 1$:

(i) $P_n(t) = \mathbb{P}(\theta = t | M^n) = \alpha_n P_{n-1}(t) \Phi\left(M_n \frac{t - \tau_{n-1}}{\sigma}\right)$

(ii) $\widehat{\theta} = \mathbb{E}[\theta | M^n] = \int t P_n(t) dt$

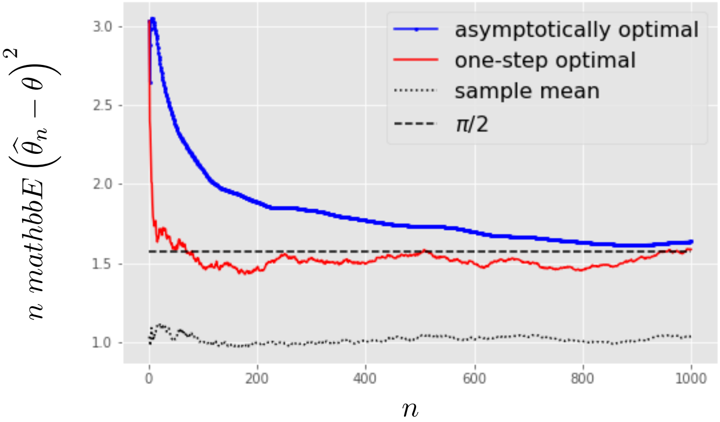
(iii) Find τ_n from

$$\tau_n = \frac{1}{2}\left(\frac{\int_{-\infty}^{\tau} t P_n(t) dt}{\int_{-\infty}^{\tau} P_n(t) dt} + \frac{\int_{\tau}^{\infty} t P_n(t) dt}{\int_{\tau}^{\infty} P_n(t) dt}\right)$$

(iv) $M_{n+1} = \mathrm{sign}(X_{n+1} - \tau_n)$

Numerical Example

Normalized empirical risk versus number of samples n
(1000 Monte Carlo experiments)



$\theta \sim \text{unif}(-3, 3)$