Mean Estimation from One-bit Measurements

Alon Kipnis (Stanford)
John Duchi (Stanford)

Allerton
October 2017

Table of Contents

Motivation

Point estimation under communication constraints:

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

Estimation error is due to:

(i) limited data

(ii) limited bits

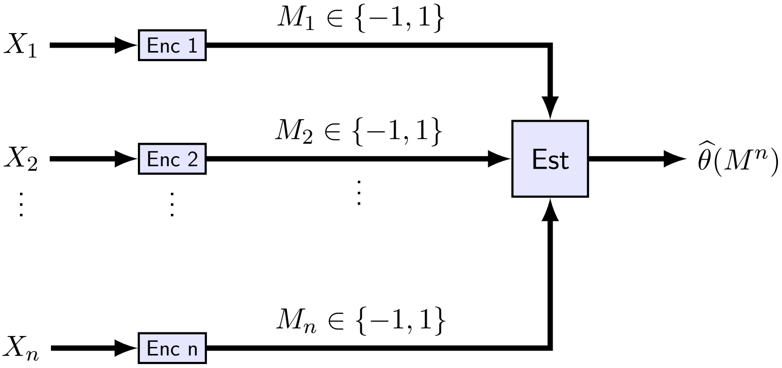
Relevant scenarios:

- big data
- low-power sensors
- distributed computing / optimization (bottleneck is due to communication between processing units)

This talk:

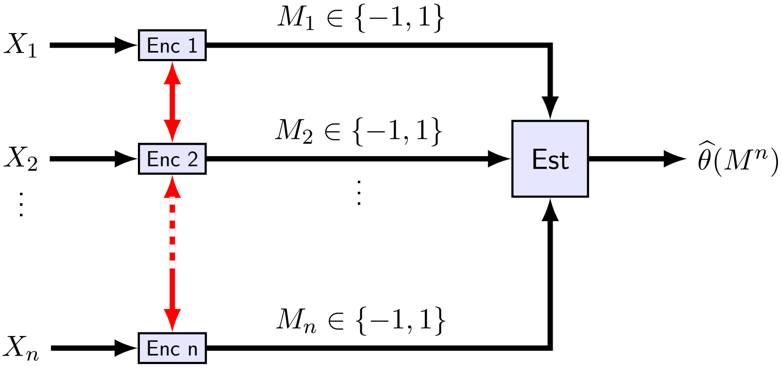
Estimating the mean θ of a normal distribution $\mathcal{N}(\theta, \sigma^2)$ from one-bit per sample (σ is known)

Three Encoding Scenarios



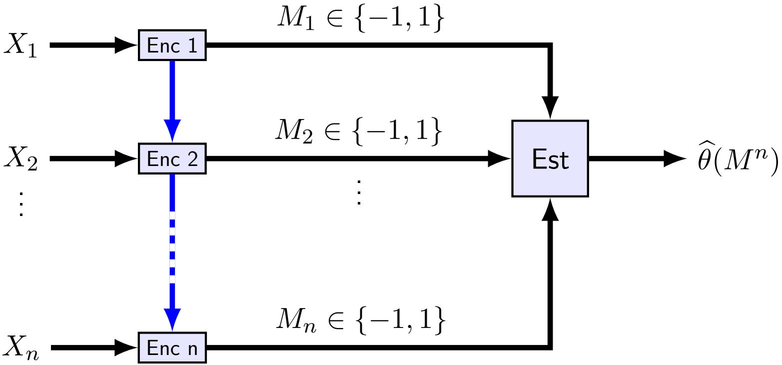
- Distributed: $M_i = f_i(X_i)$

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

- Estimation via compressed information [Han '87], [Zhang & Berger '88]
- Distributed hypothesis testing under quantization [Tsitsiklis '88]
- Estimation from multiple machines subject to a bit constraint [Zhang, Duchi, Jordan, Wainwright '13]
- Remote multiterminal source coding (CEO) [Berger, Zhang, Wiswanathan '96], [Oohama '97]

Consistency

Q: in what setting consistent estimation is possible?

Consistency

Q: in what setting consistent estimation is possible?

A: all !

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

(as in the distributed setting)

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

Efficiency

Definition: *asymptotic relative efficiency (ARE)* of an estimator:

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

(σ^2/n is the minimax risk without communication constraints)

Efficiency

Definition: *asymptotic relative efficiency (ARE)* of an estimator:

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

(σ^2/n is the minimax risk without communication constraints)

Q: in what scenarios the ARE is finite?

Efficiency

Definition: *asymptotic relative efficiency (ARE)* of an estimator:

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

(σ^2/n is the minimax risk without communication constraints)

Q: in what scenarios the ARE is finite?

This talk: all three scenarios

ARE under Centralized Encoding

Proposition

If the parameter space Θ is bounded, then the ARE under centralized encoding is 1

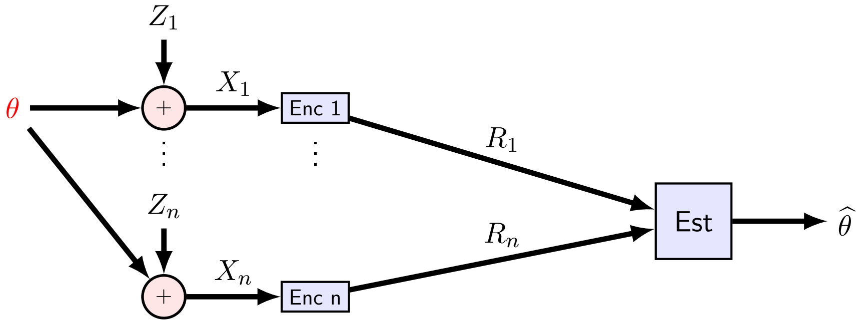
Proof:

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

- Encoder is required to describe $\bar{\theta}$ using n bits
  - divide parameter space Θ into 2^n regions of equal size
  - send region index where $\bar{\theta}$ falls
- MSE in estimating $\bar{\theta}$ decreases exponentially in n

Note: globally optimal strategy for a finite n is hard to derive since mean of $\bar{\theta}$ is unknown

Relation to CEO



Assume:

- $\theta \sim \mathcal{N}(0, \sigma_\theta^2)$,
- $Z_1, \ldots, Z_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$

Encode k instances:

- $R_1 = \ldots = R_n = 1$
- $D_{CEO} = \frac{1}{k} \sum_{j=1}^{k} \mathbb{E}\left(\theta_j - \widehat{\theta}_j\right)^2$

From [K., Rini, Goldsmith '17]:

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

ARE of $4/3$ can be attained in a fully distributed encoding (with encoding over blocks of multiple problem instances)

Conclusion

Distributed encoding is almost not a limiting factor (although inability to exploit concentration of measure in high dimension – might!)

Table of Contents

Main Results (adaptive encoding)

Theorem (converse)

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

Theorem (achievability)

Assume that Θ *is a bounded interval. There exists an estimator with ARE equals to* $\pi/2$

Theorem (one-step optimal strategy)

The next step one-bit message that minimizes the MSE is of the form $M = \mathrm{sign}(X - \tau)$ *where* τ *satisfies the fixed-point equation*

$$\tau = \frac{1}{2}\left(\frac{\int_{-\infty}^{\tau} \theta\pi(d\theta)}{\int_{-\infty}^{\tau} \pi(d\theta)} + \frac{\int_{\tau}^{\infty} \theta\pi(d\theta)}{\int_{\tau}^{\infty} \pi(d\theta)} \right)$$

Proof

converse (ARE $\geq \pi/2$)

Assume a prior $\pi(d\theta)$ on Θ with location Fisher information I_π.
The van-Trees inequality (e.g. [Tsybakov '08]) implies

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

Lemma

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

(proof by induction over a finite set of intervals approximating $M_i^{-1}(1)$ given M^{i-1})

Proof

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

[Polyak & Juditsky '92]:

$$\begin{cases} \theta_i = \theta_{i-1} + \gamma_i \varphi(X_i - \theta_i) \quad i = 1, \ldots, n \\ \widehat{\theta} = \frac{1}{n} \sum_{i=1}^n \theta_i \end{cases}$$

where:

(i) $\gamma_n \to 0^+$ "not too slow"

(ii) $\psi(x) = \mathbb{E}\varphi(x + Z)$

(iii) $\chi(x) = \mathbb{E}\varphi^2(x + Z)$

(iv) some regularity conditions on φ, χ, ψ

Then

$$\sqrt{n}(\theta - \widehat{\theta}) \to \mathcal{N}(0, V)$$

where $V = \chi(0)/\psi'^2(0)$.

Proof

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

[Polyak & Juditsky '92]:

$$\begin{cases} \theta_i = \theta_{i-1} + \gamma_i \varphi(X_i - \theta_i) \quad i = 1, \ldots, n \\ \widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \theta_i \end{cases}$$

where:

(i) $\gamma_n \to 0^+$ "not too slow"

(ii) $\psi(x) = \mathbb{E}\varphi(x + Z)$

(iii) $\chi(x) = \mathbb{E}\varphi^2(x + Z)$

(iv) some regularity conditions on φ, χ, ψ

Then

$$\sqrt{n}(\theta - \widehat{\theta}) \to \mathcal{N}(0, V)$$

where $V = \chi(0)/\psi'^2(0)$.

Proof of theorem: take $\varphi(x) = \text{sign}(x)$

One-step optimality

Theorem

Let $\pi(\theta)$ be an absolutely continuous log-concave distribution. For $X \sim \mathcal{N}(\theta, \sigma^2)$ let

$$M = \mathsf{sign}(X - \tau)$$

where τ is the unique solution to

$$\tau = \frac{1}{2}\left(\frac{\int_{-\infty}^{\tau} \theta\pi(d\theta)}{\int_{-\infty}^{\tau} \pi(d\theta)} + \frac{\int_{\tau}^{\infty} \theta\pi(d\theta)}{\int_{\tau}^{\infty} \pi(d\theta)}\right)$$

Then for any $M'(X) \in \{-1, 1\}$ and $\widehat{\theta}(M')$:

$$\mathbb{E}\left(\theta - \widehat{\theta}(M')\right)^2 \geq \mathbb{E}\left(\theta - \mathbb{E}[\theta|M]\right)^2$$

Interpertation:

The optimal one-bit message is a threshold detector. The threshold is the fixed-point that balances conditional center of masses given the message

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

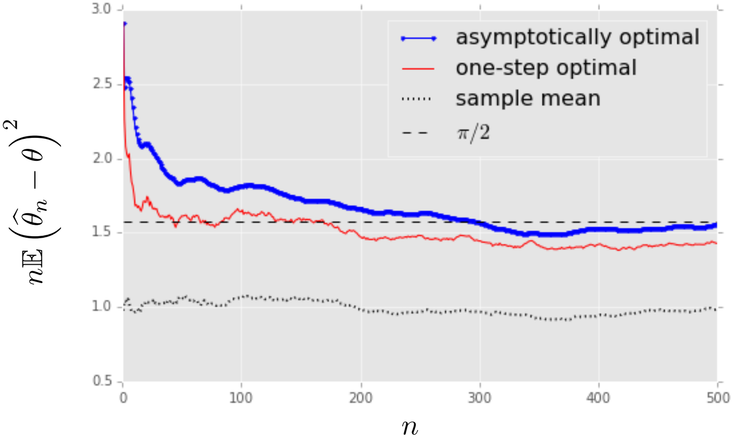
(iii) Find τ_n from

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

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

Numerical Example

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



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

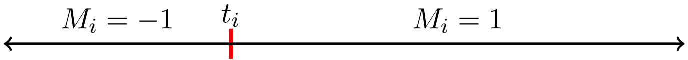
Table of Contents

Distributed Encoding

Threshold Detection

We consider only messages of the form

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



Assume:

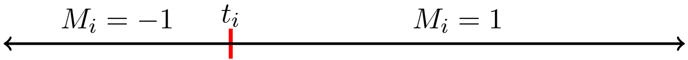
$$\lambda_n([a,b]) = \frac{1}{n} \left|[a,b] \cap \{t_i\}\right|$$

converges weakly to a probability distribution λ

Distributed Encoding
Threshold Detection

We consider only messages of the form

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



Assume:

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

converges weakly to a probability distribution λ

Example: $t_1, \dots, t_n$ are drawn independently from a probability distribution λ on $\mathbb{R}$

Main Results (distributed encoding)

Theorem

(i) *For any estimator* $\widehat{\theta}$*:*

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

where:

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

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

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

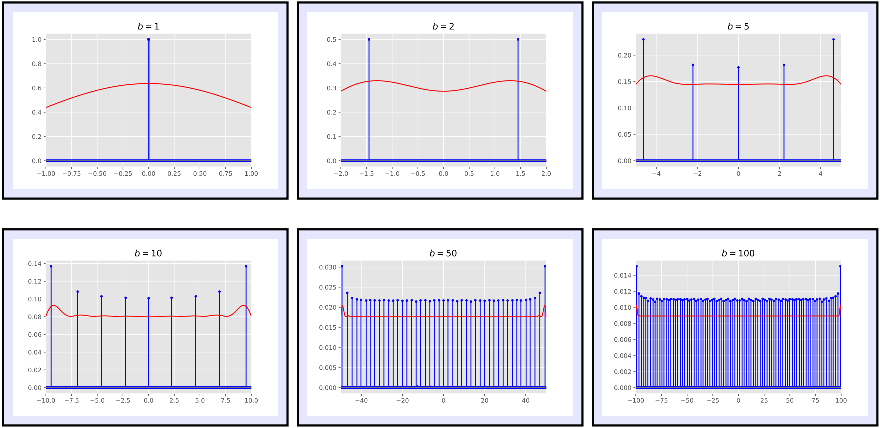
$$\sqrt{n}(\theta-\widehat{\theta}_{ML}) \to \mathcal{N}\left(0,\sigma^2/K(\theta)\right)$$

Interpretations

- ML estimator is local asymptotically minimax
- ARE of ML is $1/K(\theta)$ – depends only in the asymptotic threshold density λ
- $$1/K(\theta) = \frac{1}{\int \eta\left(\frac{t-\theta}{\sigma}\right)\lambda(dt)} \geq \frac{1}{\int \eta\left(0\right)\lambda(dt)} = \pi/2$$
  (attained by $\lambda(dt) = \delta_\theta$)
- 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

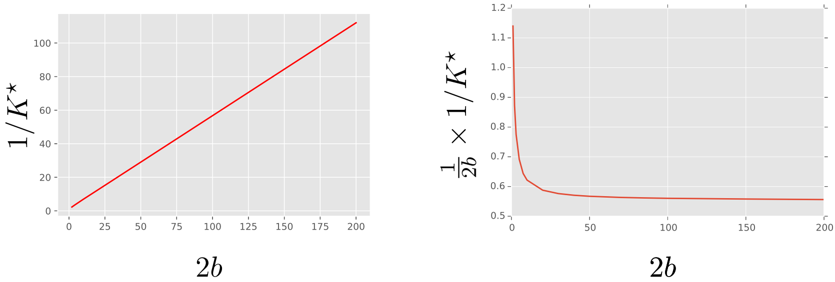
support of optimal threshold density $\lambda^\star$



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

Minimax λ

Minimax ARE vs size of parameter space



- ARE increases with size of parameter space

Table of Contents

Summary

- Asymptotic relative efficiency in adaptive setting is $\pi/2$ regardless of size of parameter space – only ~ 1.57 more samples are required due to 1-bit constraints
- One-step optimal one-bit message is a threshold detector
- ARE in distributed setting is finite
- ML estimator is local asymptotically optimal for threshold detection
- ARE of ML is characterized by asymptotic density of threshold values
- Minimax ARE of ML depends on size of parameter space

Open question

Is there a distributed encoding scheme with ARE that is both finite and independent of size of parameter space ?