Bregman deviations of generic exponential families (COLT 2024)

Patrick Saux[1]

(Joint work with Sayak Ray Chowdhury[2], Odalric-Ambrym Maillard[1] and Aditya Gopalan[3])

[1] Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000, Lille, France
[2] Microsoft Research, India
[3] Indian Institute of Science, Bangalore, India

October 24, 2023

Table of Contents

Confidence sets

Problem .

- $Y_1, \dots, Y_t$ *i.i.d. samples from a distribution* ν.
- *What is* $\mu = \mathbb{E}_{Y\sim\nu}[Y]$*?*

We want not only an estimator $\widehat{\mu}_t = \frac{1}{t}\sum_{s=1}^{t} Y_s$ but also a **confidence set**:

$$\mathbb{P}\left(\mu \in \widehat{\Theta}_t^{\delta}\right) \geqslant 1-\delta\,.$$

Asymptotically (CLT) for $\delta = 5\%$:

$$\lim_{t\to+\infty} \mathbb{P}\left(\mu \in \left[\widehat{\mu}_t \pm 1.96\frac{\sigma}{\sqrt{t}}\right]\right) \approx 1-\delta\,.$$

✖ Does not tell us anything about the small sample size regime...

Nonasymptotic confidence sets

We need some assumptions...

Sub-ψ distributions:

$$\forall \lambda \in \mathcal{I} \subseteq \mathbb{R}_+, \ \log \mathbb{E}_{Y \sim \nu} \left[e^{\lambda (Y - \mu)} \right] \leqslant \psi(\lambda) \, .$$

Theorem (Chernoff bound).

$$\mathbb{P} \left(\widehat{\mu}_t - \mu \geqslant \psi_\star^{-1} \left(\frac{1}{t} \log \frac{1}{\delta} \right) \right) \leqslant \delta \, ,$$

where $\psi_\star(u) = \sup_{\lambda \in \mathcal{I}} \lambda u - \psi(\lambda)$ *is the Fenchel-Legendre conjugate of* ψ.

Explicit ψ if ν is Gaussian or bounded (not much else is known).

Nonasymptotic confidence sets

- **Fully parametric:**
  - If you know the quantiles, use them!
- **Bounded:**
  - With control of moments:
    Bennett, Berstein [Boucheron et al., 2013],
    Bentkus [Bentkus, 2004].
  - With empirical estimators of moments:
    Bernstein [Maurer and Pontil, 2009],
    Bentkus [Kuchibhotla and Zheng, 2021].
  - Sole boundedness [Phan et al., 2021],
    [Waudby-Smith and Ramdas, 2023].
- **Self-normalised sums:**
  - [Bercu et al., 2015, Bercu and Touati, 2019].

Is this all for mean estimation?

Anytime-valid statistics

Imagine you are monitoring a prospective clinical trial.

Surgery department

General practitioner

Electronic consultation

Anytime-valid statistics

Imagine you are monitoring a prospective clinical trial up to time T.

- At time t, you test an hypothesis $\theta = \theta_0$ and compute a p-value:

$$P_t = \inf \left\{ \delta \in (0,1),\ \theta_0 \notin \widehat{\Theta}_t^{\delta} \right\} .$$

- Reject if $P_t < 0.05$ (i.e. $\theta \neq \theta_0$ is deemed significant).

Anytime-valid statistics

Imagine you are monitoring a prospective clinical trial up to time T.

- At time t, you test an hypothesis $\theta = \theta_0$ and compute a p-value:

$$P_t = \inf \left\{ \delta \in (0,1),\ \theta_0 \notin \widehat{\Theta}_t^{\delta} \right\} .$$

- Reject if $P_t < 0.05$ (i.e. $\theta \neq \theta_0$ is deemed significant).

? Should you wait until T if $P_t < 0.05$? (**Early stopping.**)

Anytime-valid statistics

Imagine you are monitoring a prospective clinical trial up to time T.

- At time t, you test an hypothesis $\theta = \theta_0$ and compute a p-value:

$$P_t = \inf \left\{ \delta \in (0,1),\ \theta_0 \notin \widehat{\Theta}_t^{\delta} \right\} .$$

- Reject if $P_t < 0.05$ (i.e. $\theta \neq \theta_0$ is deemed significant).

? Should you wait until T if $P_t < 0.05$? (**Early stopping.**)

? $P_T = 0.06$, should you enrol more patients? (**Optional continuation.**)

Anytime-valid statistics

Imagine you are monitoring a prospective clinical trial up to time T.

- At time t, you test an hypothesis $\theta = \theta_0$ and compute a p-value:

$$P_t = \inf \left\{ \delta \in (0,1),\ \theta_0 \notin \widehat{\Theta}_t^{\delta} \right\} .$$

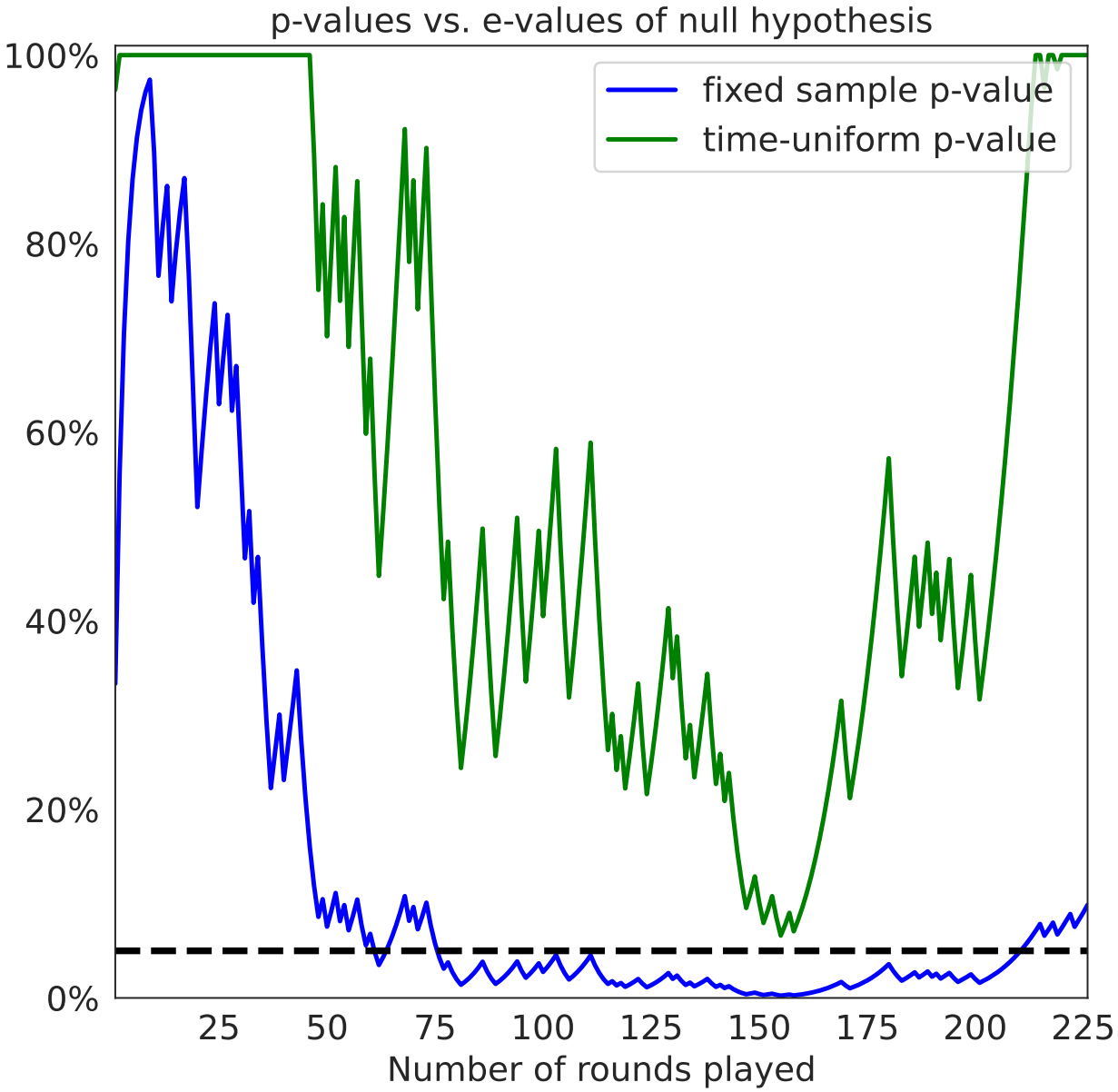
- Reject if $P_t < 0.05$ (i.e. $\theta \neq \theta_0$ is deemed significant).

? Should you wait until T if $P_t < 0.05$? (**Early stopping.**)

? $P_T = 0.06$, should you enrol more patients? (**Optional continuation.**)

? What if the trial is not randomized and you allocate patients at time t based on results at times $1, \ldots, t-1$? (**Sequential sampling**.)

Anytime-valid statistics



Anytime-valid confidence sequence

Fixed sample confidence set:

$$\forall t \in \mathbb{N},\ \mathbb{P}\left(\theta \in \widehat{\Theta}_t^{\delta}\right) \geqslant 1-\delta\,.$$

Anytime confidence sequence (CS):

$$\mathbb{P}\left(\forall t \in \mathbb{N},\ \theta \in \widehat{\Theta}_t^{\delta}\right) \geqslant 1-\delta$$

or

$$\forall \tau \textit{ stopping time},\ \mathbb{P}\left(\theta \in \widehat{\Theta}_{\tau}^{\delta}\right) \geqslant 1-\delta\,.$$

Table of Contents

Martingale and stopping time

- Stopping times are hard to deal with...
- ... but go well with martingales!

Theorem (From NSM to anytime CS, Ville's inequality). *For $t \in \mathbb{N}$, assume we know an invertible, nondecreasing $F_t : \mathbb{R} \to \mathbb{R}_+$ s.t.*

(i) $M = \left(F_t \left(\sum_{s=1}^{t} Y_s - \mu \right) \right)_{t \in \mathbb{N}}$ *is a* $\geqslant 0$ ***supermartingale*** *(NSM)*
($\approx$ nonincreasing stochastic process);

(ii) $\mathbb{E}[M_0] \leqslant 1$.

Then for any stopping time τ, we have

$$\mathbb{P}\left(\tau \left(\widehat{\mu}_\tau - \mu \right) \geqslant F_\tau^{-1}\left(\frac{1}{\delta} \right) \right) \leqslant \delta \,.$$

To find a good NSM

The sub-ψ assumption is really a supermartingale condition:

$$\forall \lambda \in \mathcal{I},\ M^\lambda = \left(e^{\lambda t(\widehat{\mu}_t - \psi(\lambda))}\right)_{t\in\mathbb{N}} \text{ is a NSM.}$$

To find a good NSM

The sub-ψ assumption is really a supermartingale condition:

$$\forall \lambda \in \mathcal{I},\ M^\lambda = \left(e^{\lambda t(\widehat{\mu}_t - \psi(\lambda))}\right)_{t\in\mathbb{N}} \text{ is a NSM.}$$

For any probability density $q(\lambda)$ over $\mathcal{I}$,

$$M = \left(\int_{\mathcal{I}} M_t^\lambda q(\lambda) d\lambda\right)_{t\in\mathbb{N}} \text{ is also a NSM (independent of } \lambda).$$

Theorem (Sub-Gaussian mixture bound). *If $\psi(\lambda) = \sigma^2\lambda^2/2$ then*

$$\left(\widehat{\Theta}_{t,c}^{\delta}\right)_{t\in\mathbb{N}} = \left(\left[\widehat{\mu}_t \pm \sigma\sqrt{\frac{2}{t}\left(1+\frac{c}{t}\right)\log\left(\frac{2\sqrt{t/c+1}}{\delta}\right)}\right]\right)_{t\in\mathbb{N}}$$

is an anytime CS for any $c > 0$.

Remark: this corresponds to the mixing distribution $\mathcal{N}(0, 1/c)$.

Confidence width

$$\widehat{\mu}_t \pm \sigma \sqrt{\frac{2}{t}\left(1+\frac{c}{t}\right)\log\left(\frac{2\sqrt{t/c+1}}{\delta}\right)}.$$

$$\left|\widehat{\Theta}^{\delta}_{t,c}\right| = \mathcal{O}\left(\sqrt{\frac{\log t}{t}}\right) \quad \text{when } t \to +\infty.$$

- Wider than the fixed sample rate $\mathcal{O}(1/\sqrt{t})$.
- Optimal width $\mathcal{O}(\sqrt{\log\log(t)/t})$ (but rarely useful in practice).

Table of Contents

Exponential families

Parametric family indexed by $\theta \in \Theta \subseteq \mathbb{R}^d$ (open set) with densities

$$p_\theta(y) \propto e^{\langle \theta, F(y) \rangle - \mathcal{L}(\theta)} .$$

- F: feature function (of $y \in \mathbb{R}^d$),
- $\mathcal{L}$: log-partition function (of $\theta \in \Theta$), convex.
  - Assume $\det \nabla^2 \mathcal{L}(\theta) > 0$ for all $\theta \in \Theta$.

Bregman divergence:

$$\begin{aligned} \mathcal{B}_{\mathcal{L}}(\theta', \theta) &= \mathcal{L}(\theta') - \mathcal{L}(\theta) - \langle \theta' - \theta, \nabla \mathcal{L}(\theta) \rangle \\ &= \mathrm{KL}(p_\theta \parallel p_{\theta'}) . \end{aligned}$$

Examples

Gaussian $\mathcal{N}\left(\mu, \sigma^2\right)$ with known variance σ^2:

$$\theta = \mu, \Theta = \mathbb{R}\,,$$
$$\mathcal{B}_{\mathcal{L}}(\theta', \theta) = \frac{(\theta' - \theta)^2}{2\sigma^2}\,.$$

Gaussian $\mathcal{N}\left(\mu, \sigma^2\right)$:

$$\theta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^\top, \Theta = \mathbb{R} \times \mathbb{R}_-^*\,,$$
$$\mathcal{B}_{\mathcal{L}}(\theta', \theta) = \frac{1}{2}\log\frac{\theta_2}{\theta_2'} + \frac{\theta_2'}{2\theta_2} - \theta_2'\left(\frac{\theta_1'}{2\theta_2'} - \frac{\theta_1}{2\theta_2}\right)^2 - \frac{1}{2}\,.$$

Bernoulli $\mathcal{B}(p)$:

$$\theta = p, \Theta = (0,1)\,,$$
$$\mathcal{B}_{\mathcal{L}}(\theta', \theta) = \theta\log\frac{\theta}{\theta'} + (1-\theta)\log\frac{1-\theta}{1-\theta'}\,.$$

Bregman martingale

Lemma . *For $\theta \in \Theta$ and λ s.t. $\theta + \lambda \in \Theta$,*

$$\log \mathbb{E}_\theta \left[e^{\langle \lambda, F(Y) - \mathbb{E}_\theta[F(Y)] \rangle} \right] = \mathcal{B}_\mathcal{L}(\theta + \lambda, \theta) \, .$$

- The Bregman divergence $\mathcal{B}_\mathcal{L}$ plays the role of ψ.
- If $\widehat{\mu}_t = \frac{1}{t} \sum_{s=1}^{t} F(Y_s)$ and $\mu = \mathbb{E}_\theta [F(Y)]$ then

$$\forall \lambda, \; M^\lambda = \left(e^{\langle \lambda, t(\widehat{\mu}_t - \mu) \rangle - t\mathcal{B}_\mathcal{L}(\theta+\lambda, \theta)} \right)_{t \in \mathbb{N}} \text{ is a NSM.}$$

- **Mixture:** for any $c > 0$ and $q_\theta(\lambda | c) \propto e^{\langle \theta + \lambda, c \nabla \mathcal{L}(\theta) \rangle - c\mathcal{L}(\theta)}$,

$$M_t = \int M_t^\lambda q_\theta(\lambda | c) d\lambda \quad \text{is a NSM.}$$

Bregman confidence sequence

Theorem (Bregman, mixture bound for exponential families).

$$\left(\widehat{\Theta}_t^{\delta}\right)_{t\in\mathbb{N}} = \left(\left\{\theta\in\Theta,\ (t+c)\mathcal{B}_{\mathcal{L}}\left(\theta,\widehat{\theta}_{t,c}(\theta)\right)\leqslant \log\frac{1}{\delta}+\gamma_{t,c}(\theta)\right\}\right)_{t\in\mathbb{N}}$$

is an anytime CS.

Regularised parameter estimate ($c=0 \iff$ MLE):

$$\widehat{\theta}_{t,c}(\theta) = \nabla\mathcal{L}^{-1}\left(\frac{t}{t+c}\widehat{\mu}_t + \frac{c}{t+c}\mathcal{L}(\theta)\right).$$

Bregman information gain:

$$\gamma_{t,c}(\theta) = \log\frac{\int_\Theta e^{-c\mathcal{B}_{\mathcal{L}}(\theta',\theta)}d\theta'}{\int_\Theta e^{-(t+c)\mathcal{B}_{\mathcal{L}}(\theta',\widehat{\theta}_{t,c}(\theta))}d\theta'}.$$

Bregman confidence sequence

$$(t+c)\mathcal{B}_{\mathcal{L}}\left(\theta, \widehat{\theta}_{t,c}(\theta)\right) \leqslant \log \frac{1}{\delta} + \gamma_{t,c}(\theta)\,.$$

- Laplace's method for approximating integrals: when $t \to +\infty$,

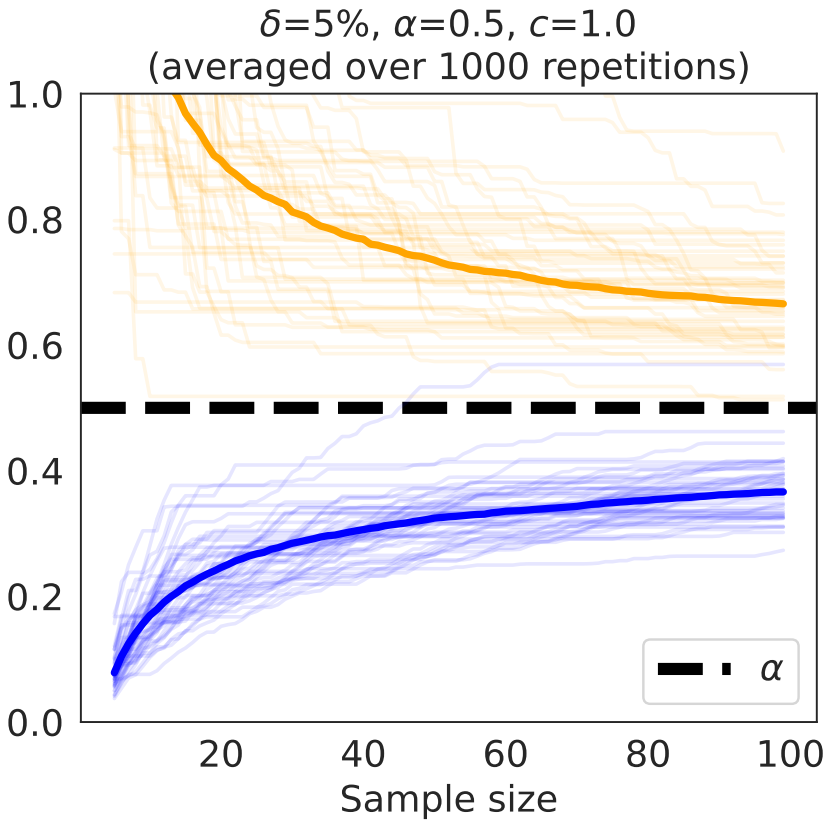
$$\mathcal{B}_{\mathcal{L}}(\theta, \widehat{\theta}_{t,c}) \underset{\sim}{\propto} \|\theta - \widehat{\theta}_t\|^2\,,$$

$$\gamma_{t,c}(\theta) = \frac{\dim \Theta}{2} \log\left(1 + \frac{t}{c}\right) + \mathcal{O}(1)\,,$$

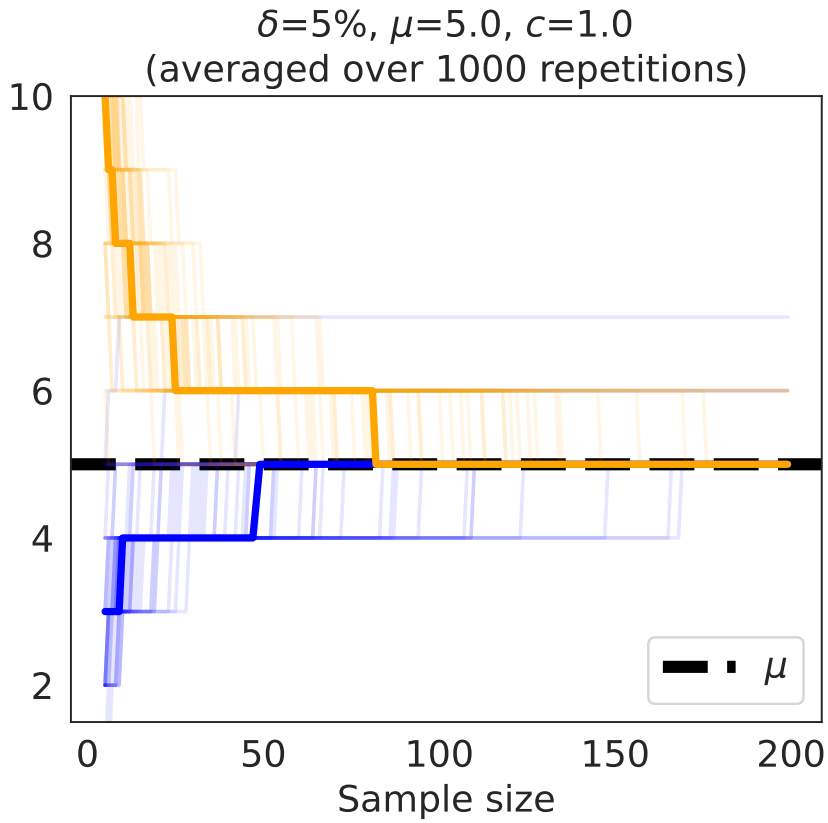
$$\left|\widehat{\Theta}_t^{\delta}\right| = \mathcal{O}\left(\sqrt{\frac{\log t}{t}}\right)\,.$$

- Implicit confidence set in θ...
  - ...but convex: easy numerical solution.

Numerical experiments



(a) Pareto



(b) Chi-square

Numerical experiments

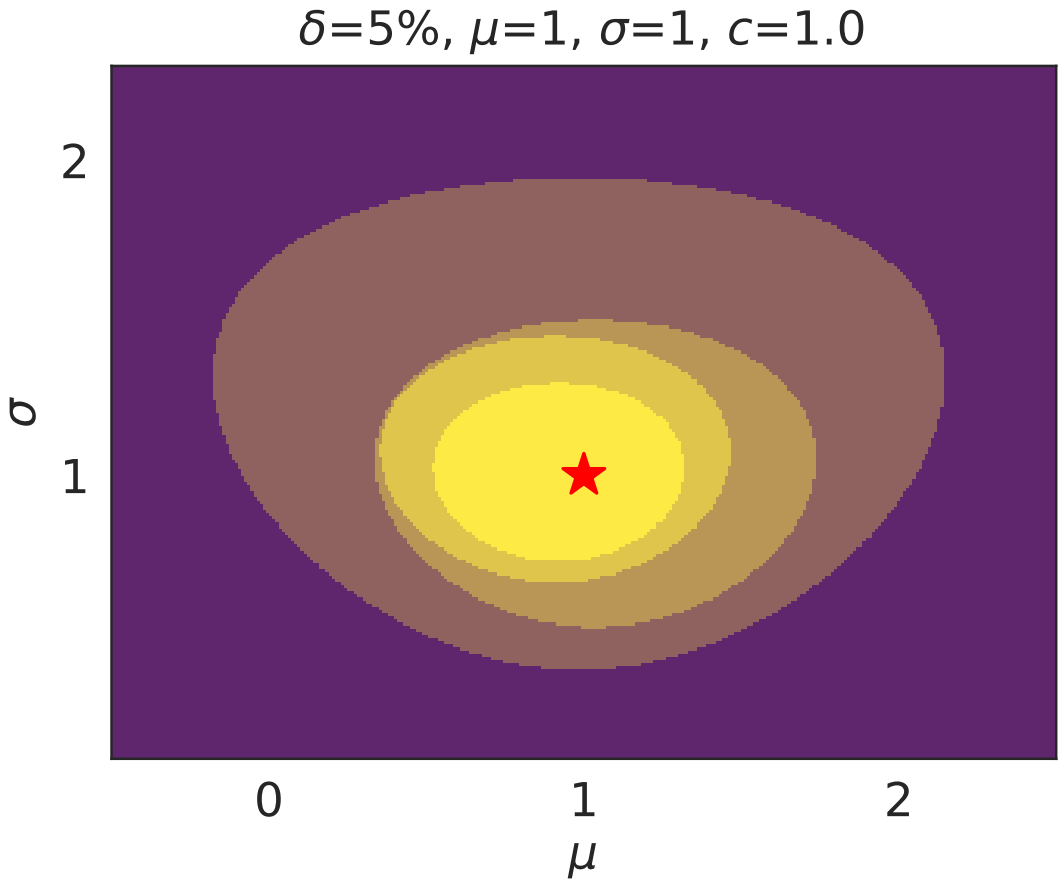
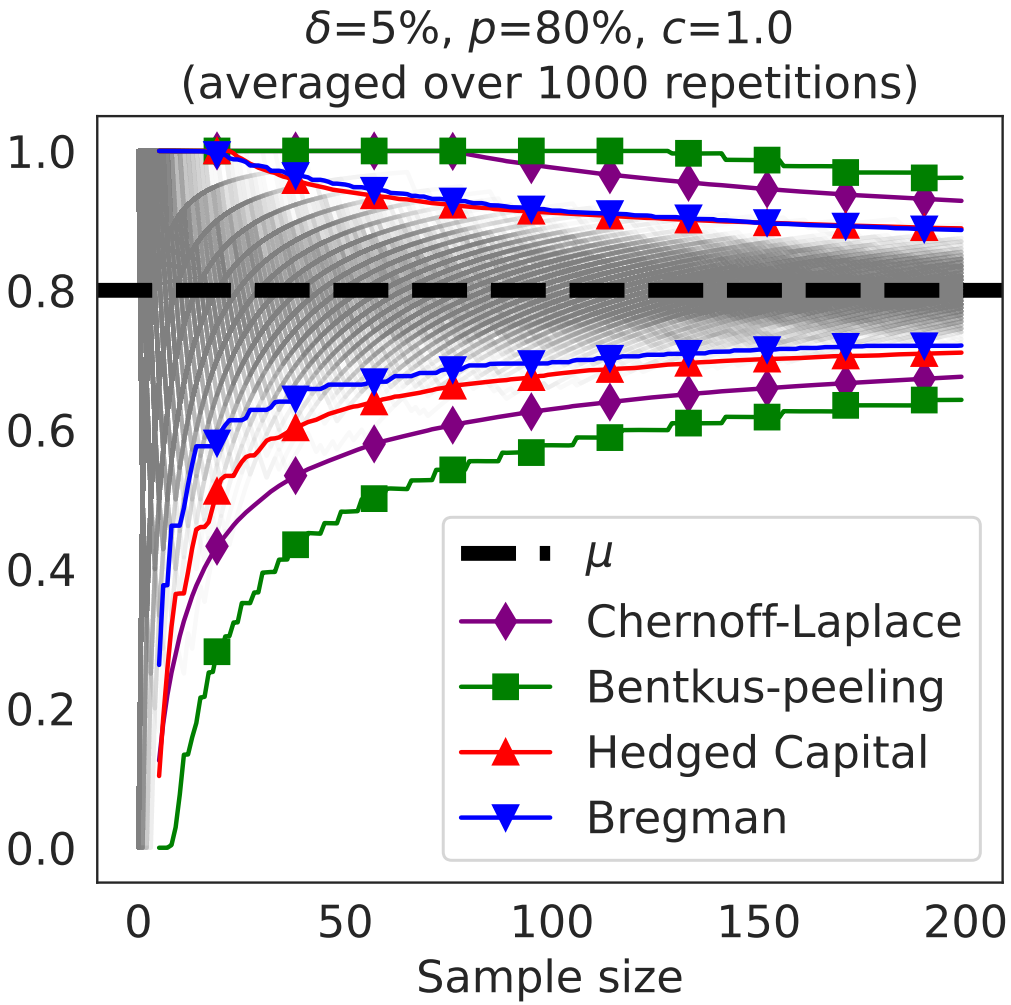


Figure: Gaussian (mean and variance) for $t \in \{10, 25, 50, 100\}$ observations

Numerical experiments



Comparison of median confidence envelopes around the mean for $\mathcal{B}(0.8)$. Grey lines are trajectories of empirical means $\widehat{\mu}_t$.

Examples

Table 2: Summary of Bregman confidence sets given by Theorem 3.2 for representative families
Throughout, the following notations are used:

$S_n = \sum_{t=1}^n X_t$, $\widehat{\mu}_n = \frac{S_n}{n}$, $Z_n(\mu,\sigma) = \frac{1}{\sigma^2}\sum_{t=1}^n (X_t-\mu)^2$,
$S_n^{(k)} = \sum_{t=1}^n X_t^k$, $L_n = \sum_{t=1}^n \log X_t$, $K_n = \sum_{t=1}^n \log X_t/2$,
$I(a,b) = \int_{-\infty}^{+\infty} e^{-ae^\theta + b\theta} d\theta$,
$J(a,b) = \int_0^\infty \exp\left(-a\log\Gamma\left(\frac{k}{2}\right) + b\frac{k}{2}\right) dk$ (if $k \in \mathbb{R}_+$),
$J(a,b) = \sum_{k'=1}^\infty \exp\left(-a\log\Gamma\left(\frac{k'}{2}\right) + b\frac{k'}{2}\right)$ (if $k \in \mathbb{N}$).

Name	Parameters	Formula
Gaussian	$\mu \in \mathbb{R}$	$\frac{1}{n+c}\frac{(S_n-n\mu)^2}{2\sigma^2} \leqslant \log\frac{1}{\delta} + \frac{1}{2}\log\frac{n+c}{c}$
Gaussian	$\sigma \in \mathbb{R}_+$	$\frac{Q_n(\mu)}{2\sigma^2} - \left(\frac{n+c}{2}+1\right)\log\left(\frac{Q_n(\mu)}{2\sigma^2}+\frac{c}{2}\right)$ $\leqslant \log\frac{1}{\delta} + \log\frac{\Gamma\left(\frac{c}{2}\right)}{\Gamma\left(\frac{n+c}{2}\right)} - \log\left(\frac{n+c}{2}\right) - \frac{c}{2}\log\frac{c}{2}$
Gaussian	$\mu \in \mathbb{R}$ $\sigma \in \mathbb{R}_+$	$\frac{1}{2}Z_n(\mu,\sigma) - \frac{n+c+3}{2}\log\left(\frac{n}{n+c}Z_n(\widehat{\mu}_n,\sigma) + \frac{c}{n+c}Z_n(\mu,\sigma) + c\right)$ $\leqslant \log\frac{1}{\delta} - \frac{n}{2}\log 2 - \left(\frac{c}{2}+2\right)\log c + \frac{1}{2}\log(n+c)$ $+ \log\Gamma\left(\frac{c+3}{2}\right) - \log\Gamma\left(\frac{n+c+3}{2}\right)$
Bernoulli	$\mu \in [0,1]$	$S_n\log\frac{1}{\mu} + (n-S_n)\log\frac{1}{1-\mu} + \log\frac{\Gamma(S_n+c\mu)\Gamma(n-S_n+c(1-\mu))}{\Gamma(c\mu)\Gamma(c(1-\mu))}$ $\leqslant \log\frac{1}{\delta} + \log\frac{\Gamma(n+c)}{\Gamma(c)}$
Exponential	$\mu \in \mathbb{R}_+$	$\frac{S_n}{\mu} - (n+c+1)\log\left(\frac{S_n}{\mu}+c\right)$ $\leqslant \log\frac{1}{\delta} + \log\frac{\Gamma(c)}{\Gamma(n+c)} - \log(n+c) - c\log c$
Gamma	$\lambda \in \mathbb{R}_+$	$\frac{S_n}{\lambda} - ((n+c)k+1)\log\left(\frac{S_n}{\lambda}+ck\right)$ $\leqslant \log\frac{1}{\delta} + \log\frac{\Gamma(ck)}{\Gamma((n+c)k)} - \log((n+c)k) - ck\log ck$
Weibull	$\lambda \in \mathbb{R}_+$	$\frac{S_n^{(k)}}{\lambda^k} - (n+c+1)\log\left(\frac{S_n^{(k)}}{\lambda^k}+c\right)$ $\leqslant \log\frac{1}{\delta} + \log\frac{\Gamma(c)}{\Gamma(n+c)} - \log(n+c) - c\log c$
Pareto	$\alpha \in \mathbb{R}$	$\alpha L_n - (n+c+1)\log(\alpha L_n + c)$ $\leqslant \log\frac{1}{\delta} + \log\frac{\Gamma(c)}{\Gamma(n+c)} - \log(n+c) - c\log c$
Poisson	$\lambda \in \mathbb{R}_+$	$n\lambda - S_n\log\lambda$ $\leqslant \log\frac{1}{\delta} + \log I(c, c\lambda) - \log I(n+c, S_n+c\lambda)$
Chi-square	$k \in \mathbb{N}$ or $k \in \mathbb{R}_+$	$n\log\Gamma\left(\frac{k}{2}\right) - \frac{k}{2}K_n - \log J\left(c, c\psi_0\left(\frac{k}{2}\right)\right)$ $+ \log J\left(n+c, K_n + c\psi_0\left(\frac{k}{2}\right)\right) \leqslant \log\frac{1}{\delta}$

Conclusion

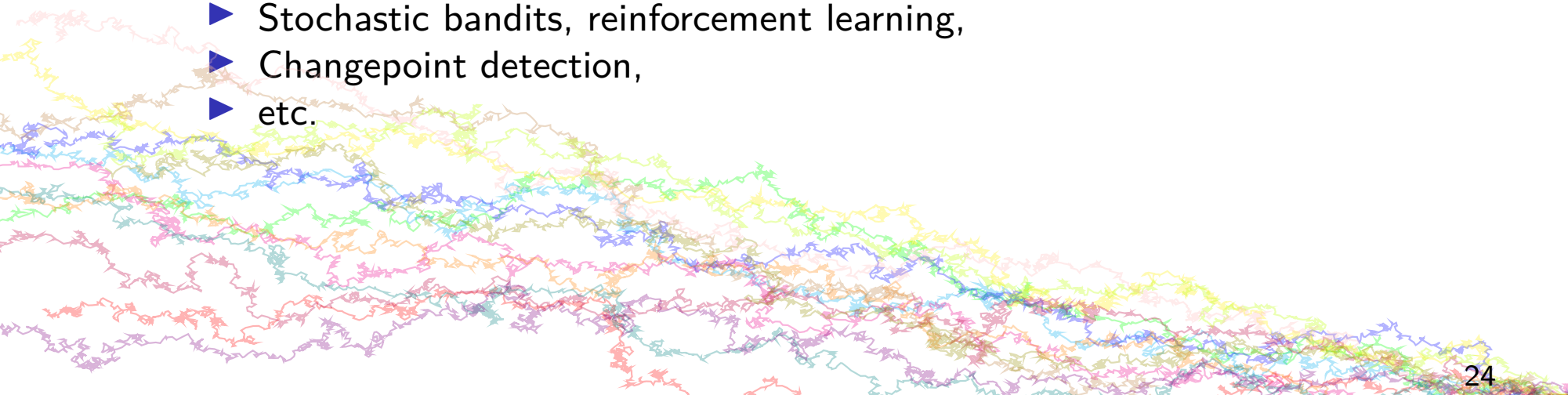
⚠ Beware of the sampling mechanism/early stopping/optimal continuation!

✔ Anytime valid CS:

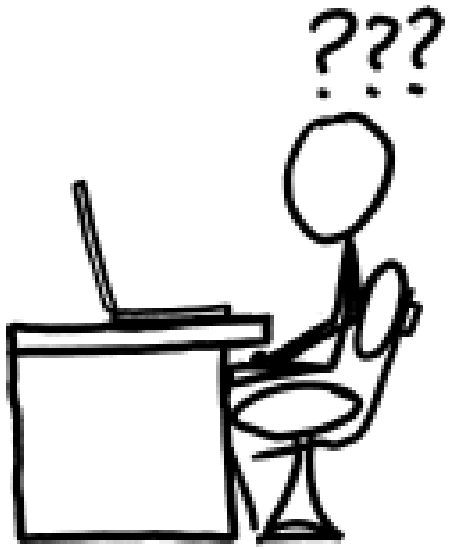
- Bounded [Waudby-Smith and Ramdas, 2023],
- Bregman [Chowdhury et al., 2023],
- Sub-Gaussian (cf. my thesis).

🎓 Many applications:

- Safe statistical inference,
- Stochastic bandits, reinforcement learning,
- Changepoint detection,
- etc.



Questions?



References I

V. Bentkus. On hoeffding's inequalities. The Annals of Probability, 32(2):1650–1673, 2004.

B. Bercu and T. Touati. New insights on concentration inequalities for self-normalized martingales. Electronic Communications in Probability, 24:1–12, 2019.

B. Bercu, B. Delyon, and E. Rio. Concentration inequalities for sums and martingales. Springer, 2015.

S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities: A nonasymptotic theory of independence. Oxford university press, 2013.

S. R. Chowdhury, P. Saux, O. Maillard, and A. Gopalan. Bregman deviations of generic exponential families. In The Thirty Sixth Annual Conference on Learning Theory, pages 394–449. PMLR, 2023.

A. K. Kuchibhotla and Q. Zheng. Near-optimal confidence sequences for bounded random variables. In International Conference on Machine Learning, pages 5827–5837. PMLR, 2021.

A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. 2009.

M. Phan, P. Thomas, and E. Learned-Miller. Towards practical mean bounds for small samples. In International Conference on Machine Learning, pages 8567–8576. PMLR, 2021.

I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2023.