Bregman Deviations of Generic Exponential Families

Patrick Saux[1], Sayak Ray Chowdhury[2], Odalric-Ambrym Maillard [1], Aditya Gopalan[3]

[1]Inria, Univ. Lille, CNRS, UMR 918f9 - CRIStAL, [2] Boston University, [3]Indian Institute of Science

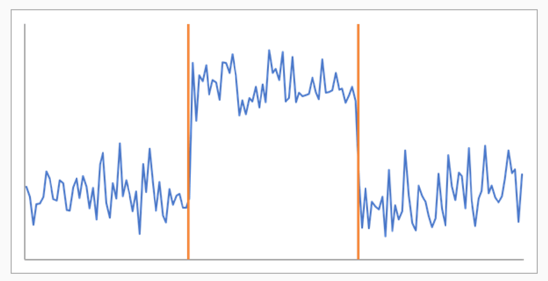
Motivations

- Observe i.i.d sequence $(X_t)_{t\in\mathbb{N}}$ drawn from parametric distributions p_θ.
- Estimate θ and confidence set $\widehat{\Theta}^\delta_t = \widehat{\Theta}^\delta_t(X_1,\dots,X_t)$.
- Time-uniformity:

$$\mathbb{P}\left(\forall t\in\mathbb{N},\ \theta\in\widehat{\Theta}^\delta_t\right)\geqslant 1-\delta \quad \text{or} \quad \mathbb{P}\left(\theta\in\widehat{\Theta}^\delta_\tau\right)\geqslant 1-\delta \text{ for } \tau \text{ stopping time}.$$

- Typical applications:

Stochastic bandits



Change-point detection

Setting: exponential families

Parametric family

$$p_\theta(x) = h(x)\exp\left(\langle\theta, F(x)\rangle - \mathcal{L}(\theta)\right)$$

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

Bregman divergence

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

Standard parameter estimate

$$\widehat{\theta}_t = \nabla\mathcal{L}^{-1}\left(\widehat{\mu}_t\right), \text{ where } \widehat{\mu}_t = \frac{1}{t}\sum_{s=1}^t F(X_s).$$

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

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

Lemma 1 (log-Laplace control) For $\theta\in\Theta$ and λ s.t $\theta+\lambda\in\Theta$,

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

Lemma 2 (Bregman duality) For any $\alpha\in[0,1]$,

$$\mathcal{B}^\star_{\mathcal{L},\theta'}\left(\alpha(\nabla\mathcal{L}(\theta)-\nabla\mathcal{L}(\theta'))\right) = \mathcal{B}_{\mathcal{L}}\left(\theta',\theta_\alpha\right),$$

where

$$\begin{aligned}\theta_\alpha &= \nabla\mathcal{L}^{-1}\left(\alpha\nabla\mathcal{L}(\theta)+(1-\alpha)\nabla L(\theta')\right),\\ \mathcal{B}^\star_{\mathcal{L},\theta'}(x) &= \sup_\lambda\langle\lambda,x\rangle - \mathcal{B}_{\mathcal{L}}\left(\theta'+\lambda,\theta'\right).\end{aligned}$$

Time-uniform Bregman deviation

Main theorem

Regularized parameter estimate

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

Bregman information gain

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

Time-uniform deviation

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

Remarks

- Valid for generic families, not just one-dimensional.
- $\gamma_{t,c}(\theta)=\frac{\dim\Theta}{2}\log\left(1+\frac{t}{c}\right)+\mathcal{O}(1) \implies$ asymptotic confidence radius is $\propto\sqrt{\frac{\log t}{t}}$.
- Implicit confidence set in θ, but easy to compute numerically.
- Explicit instantiation to many classical families:
  $\hookrightarrow$ Gaussian, Bernoulli, Exponential, Gamma, Weibull, Pareto, Poisson, χ^2.

GLR test in exponential families

Change of measure detection: distribution of X_u is $p_{\theta(u)}$

$\mathcal{H}_0$ (null): $\exists\theta_0\in\Theta,\ \forall u\in\mathbb{N},\ \theta(u)=\theta_0$ **(no change)**,

$\mathcal{H}_1$ (alt.): $\exists s\in\mathbb{N},\ \theta_1,\theta_2\in\Theta,\ \forall u\in\mathbb{N},\ \theta(u)=\theta_1\mathbb{1}_{u\leqslant s}+\theta_2\mathbb{1}_{u>s}$ **(change)**.

Scan statistic

$$\widehat{\theta}_{a:b} = \nabla\mathcal{L}^{-1}\left(\frac{1}{b-a+1}\sum_{s=a}^b F(X_u)\right).$$

Generalized Likelihood Ratio

$$\begin{aligned}G_{1:s:t} &= \inf_{\theta_0}\sup_{\theta_1,\theta_2}\log\left(\frac{\prod_{u=1}^s p_{\theta_1}(X_u)\prod_{u=s+1}^t p_{\theta_2}(X_u)}{\prod_{u=1}^t p_{\theta_0}(X_u)}\right)\\ &= \inf_{\theta_0}\ s\mathcal{B}_{\mathcal{L}}(\theta_0,\widehat{\theta}_{1:s}) + (t-s)\mathcal{B}_{\mathcal{L}}(\theta_0,\widehat{\theta}_{s+1:t}).\end{aligned}$$

Doubly time-uniform deviation $g(t)=(t+1)\log^2(t+1)/\log(2)$,

$$\mathbb{P}_\theta\left(\exists t\in\mathbb{N},\ \exists s<t:\ (t-s+c)\mathcal{B}_{\mathcal{L}}(\theta,\widehat{\theta}_{s+1:t,c}(\theta))\geqslant\log\left(\frac{g(t)}{\delta}\right)+\gamma_{s+1:t,c}(\theta)\right)\leqslant\delta.$$

Regularized GLR test

$$\begin{aligned}\tau_{c,\delta} = \min\Big\{t\in\mathbb{N}:\ \exists s<t,\ \theta\in\Theta:\ &(s+c)\mathcal{B}_{\mathcal{L}}\left(\theta,\widehat{\theta}_{1:s,c}(\theta)\right)\geqslant\log\left(\frac{2}{\delta}\right)+\gamma_{1:s,c}(\theta)\\ \text{and } &(t-s+c)\mathcal{B}_{\mathcal{L}}\left(\theta,\widehat{\theta}_{s+1:t,c}(\theta)\right)\geqslant\log\left(\frac{2g(t)}{\delta}\right)+\gamma_{s+1:t,c}(\theta)\Big\}\end{aligned}$$

has a false alarm probability $\leqslant\delta$.

Sketch of proof

Martingale construction

Lemma 1 $\implies$ for all suitable λ and an arbitrary $c>0$,

$$M^\lambda_t = \exp\left(\left\langle\lambda,\sum_{s=1}^t F(X_s)-\mathbb{E}_\theta[F(X)]\right\rangle - t\mathcal{B}_{\mathcal{L}}(\theta+\lambda,\theta)\right) \text{ defines a } \geqslant 0 \text{ martingale}.$$

Martingale mixture For $c>0$,

$$\begin{aligned}q_\theta(\lambda|c) &\propto \exp\left(\langle\theta+\lambda, c\nabla\mathcal{L}(\theta)\rangle - c\mathcal{L}(\theta)\right),\\ M_t &= \int M^\lambda_t q_\theta(\lambda|c)\,d\lambda.\end{aligned}$$

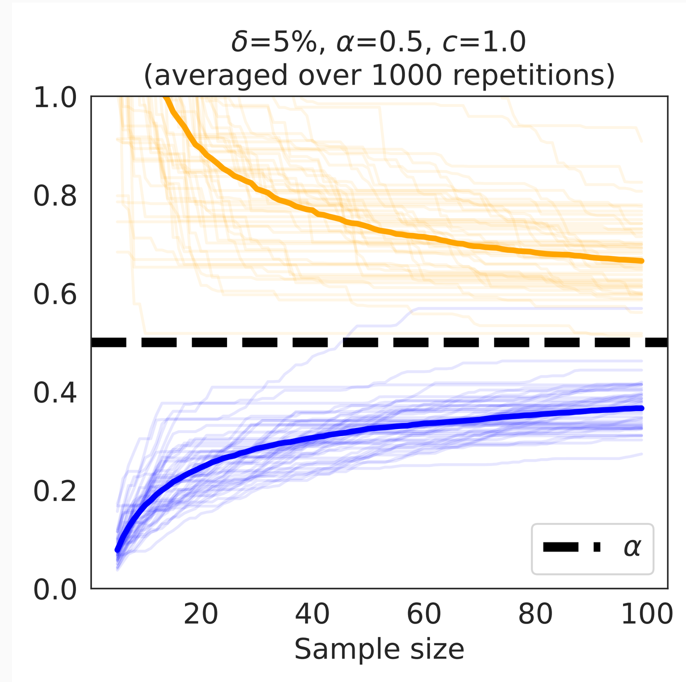
Rewriting

Lemma 2 $\implies M_t = \exp\left((t+c)\mathcal{B}_{\mathcal{L}}\left(\theta,\widehat{\theta}_{t,c}(\theta)\right)-\gamma_{t,c}(\theta)\right)$.

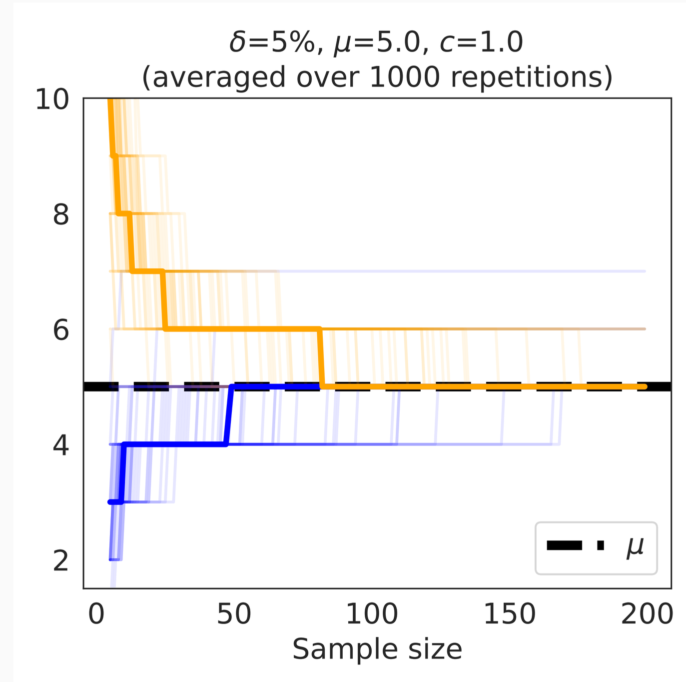
Conclusion

Ville's inequality (supermartingale + Doob's optional stopping). □

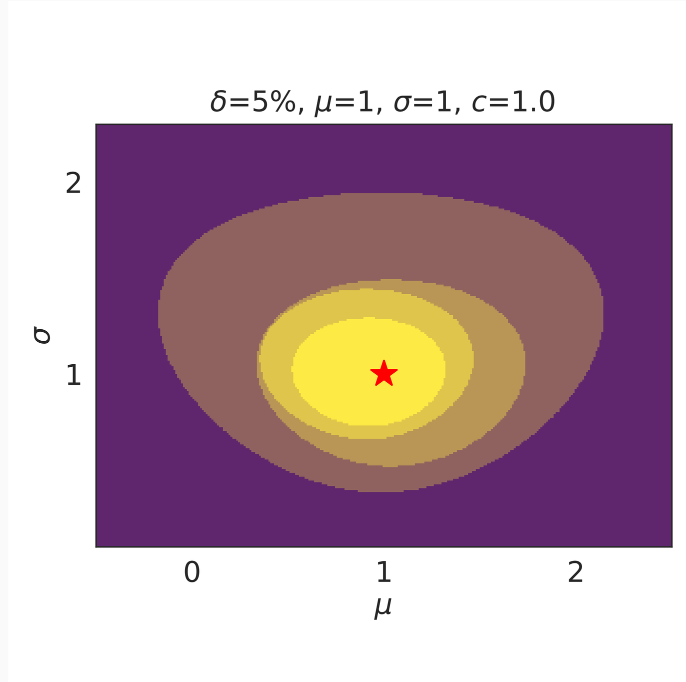
Numerical experiments



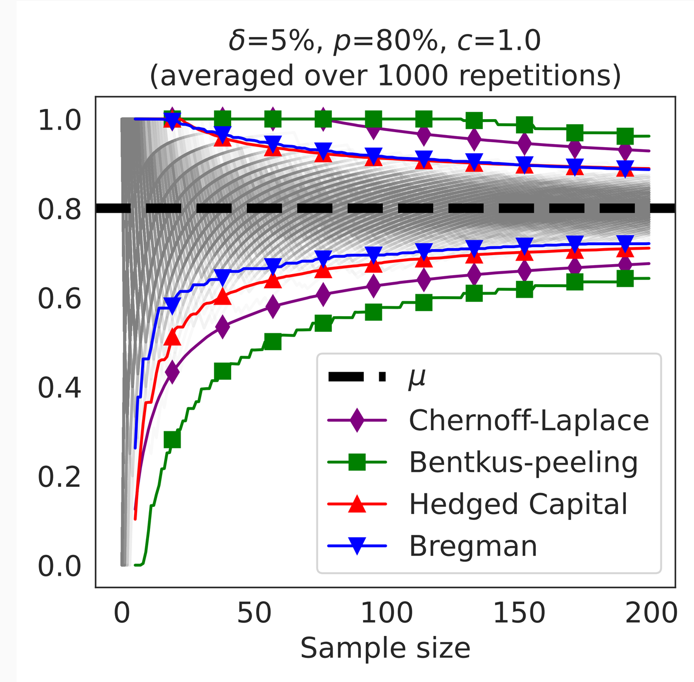
Pareto



Chi-square



Gaussian mean-variance $t\in\{10,25,50,100\}$



Comparison of median confidence envelopes around the mean for Bernoulli $\mathcal{B}(0.8)$

Aknowledgments: The authors acknowledge the funding of the French National Research Agency, the French Ministry of Higher Education and Research, Inria, the MEL and the I-Site ULNE regarding project R-PILOTE-19-004-APPRENF and Bandits For Health (B4H).