Bregman Deviations of Generic Exponential Families (and some extras)

Patrick Saux[1]

[1] Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000, Lille, France

June 2, 2022

Who?

Odalric-Ambrym Maillard

Sayak Ray Chowdhury

Aditya Gopalan

Table of Contents

Appetizer: local Dvoretzky-Kiefer-Wolfowitz confidence

DKW (Massart, 1990):

$$\mathbb{P}\left(\sup_{x\in\mathcal{X}} \widehat{F}_n(x) - F(x) > \epsilon\right) \leq e^{-2n\epsilon^2}\,.$$

Appetizer: local Dvoretzky-Kiefer-Wolfowitz confidence

DKW (Massart, 1990):

$$\mathbb{P}\left(\sup_{x\in[0,1]} \widehat{U}_n(x) - U(x) > \epsilon\right) \leq e^{-2n\epsilon^2} .$$

Appetizer: local Dvoretzky-Kiefer-Wolfowitz confidence

Local DKW (?):

$$\mathbb{P}\left(\sup_{x \in [\alpha, \beta]} \widehat{U}_n(x) - U(x) > \epsilon \right) \leq \textbf{?}\,.$$

Appetizer: local Dvoretzky-Kiefer-Wolfowitz confidence

Local DKW (Maillard, 2022):

$$\mathbb{P}\left(\sup_{x\in[\alpha,\beta]} \widehat{U}_n(x) - U(x) > \epsilon\right) = \sum_{\ell=0}^{\overline{n}_{\alpha,\epsilon}-1} \binom{n}{\ell} \beta_{\ell+1,\epsilon}^{n-\ell} \ell! I_\ell(1;\beta_{1,\epsilon},\dots\beta_{\ell,\epsilon}),$$

where

$$I_k(x;a_1,\dots,a_k) = \int_{a_1}^{x}\int_{a_2}^{t_1}\cdots\int_{a_k}^{t_{k-1}} dt_1\dots dt_k, \text{ for } x \geq a_1 \geq \cdots \geq a_k \in \mathbb{R},$$

$$\beta_{k,\epsilon} = \min(\beta,(n-k+1)/n-\epsilon),$$

$$\overline{n}_{\alpha,\epsilon} = \lceil n(1-\alpha-\epsilon)\rceil .$$

Appetizer: local Dvoretzky-Kiefer-Wolfowitz confidence

Local DKW (Maillard, 2022):

$$\mathbb{P}\left(\sup_{x\in[\alpha,\beta]} \widehat{U}_n(x) - U(x) > \epsilon\right) = \sum_{\ell=0}^{\overline{n}_{\alpha,\epsilon}-1} \binom{n}{\ell} \beta_{\ell+1,\epsilon}^{n-\ell} \ell! I_\ell(1; \beta_{1,\epsilon}, \dots \beta_{\ell,\epsilon}),$$

where

$$I_k(x; a_1, \dots, a_k) = \int_{a_1}^{x} \int_{a_2}^{t_1} \cdots \int_{a_k}^{t_{k-1}} dt_1 \dots dt_k, \text{ for } x \geq a_1 \geq \dots \geq a_k \in \mathbb{R},$$

$$\beta_{k,\epsilon} = \min(\beta, (n-k+1)/n - \epsilon),$$

$$\overline{n}_{\alpha,\epsilon} = \lceil n(1-\alpha-\epsilon) \rceil .$$

$\hookrightarrow$ time-uniform (peeling), application to cVaR, spectral risk measures...

Table of Contents

Exponential families

Parametric family indexed by $\theta \in \Theta$ (open set) of distributions ν_θ over $\mathbb{R}^d$ given by

$$\frac{d\nu_\theta}{d\nu_{\theta_\circ}}(x) = h(x)e^{\langle \theta, F(x) \rangle - \mathcal{L}(\theta)} .$$

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

$\hookrightarrow$ **Goal:** time-uniform confidence around θ.

Exponential families

$$\frac{d\nu_\theta}{d\nu_{\theta_\circ}}(x) = h(x)e^{\langle \theta, F(x)\rangle - \mathcal{L}(\theta)} \,.$$

MLE:

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

Bregman divergence:

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

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

$$\widehat{\mu}_t = \frac{1}{t}\sum_{s=1}^{t} F(X_s) \quad \text{and} \quad \mu = \mathbb{E}_\theta\left[F(X)\right].$$

Bregman martingale

$$\widehat{\mu}_t = \frac{1}{t}\sum_{s=1}^{t} F(X_s) \quad \text{and} \quad \mu = \mathbb{E}_\theta\left[F(X)\right] .$$

Nonnegative martingale: For any $\lambda \in \Theta$,

$$M_t^\lambda = e^{\langle \lambda, t(\widehat{\mu}_t - \mu)\rangle - t\mathcal{B}_{\mathcal{L}}(\theta+\lambda,\theta)} ,$$

Bregman martingale

$$\widehat{\mu}_t = \frac{1}{t}\sum_{s=1}^{t} F(X_s) \quad \text{and} \quad \mu = \mathbb{E}_\theta\left[F(X)\right] .$$

Nonnegative martingale: For any $\lambda \in \Theta$,

$$M_t^\lambda = e^{\langle \lambda, t(\widehat{\mu}_t - \mu)\rangle - t\mathcal{B}_\mathcal{L}(\theta+\lambda,\theta)} ,$$

Mixture: for $c > 0$,

$$q_\theta(\lambda|c) \propto e^{\langle \theta+\lambda, c\nabla\mathcal{L}(\theta)\rangle - c\mathcal{L}(\theta)} ,$$

$$M_t = \int M_t^\lambda q_\theta(\lambda|c)d\lambda .$$

Bregman martingale

Nonnegative

Mixture: for

Bregman-Laplace confidence set

Regularized parameter estimate:

$$\widehat{\theta}_{t,c}(\theta_0) = \nabla\mathcal{L}^{-1}\left(\frac{t}{t+c}\widehat{\mu}_t + \frac{c}{t+c}\mathcal{L}(\theta_0)\right) .$$

Bregman-Laplace confidence set

Regularized parameter estimate:

$$\widehat{\theta}_{t,c}(\theta_0) = \nabla\mathcal{L}^{-1}\left(\frac{t}{t+c}\widehat{\mu}_t + \frac{c}{t+c}\mathcal{L}(\theta_0)\right) .$$

Bregman information gain:

$$\gamma_{t,c}(\theta_0) = \log \frac{\int_\Theta e^{-c\mathcal{B}_\mathcal{L}(\theta',\theta_0)} d\theta'}{\int_\Theta e^{-(t+c)\mathcal{B}_\mathcal{L}(\theta',\widehat{\theta}_{t,c}(\theta_0))} d\theta'} .$$

Bregman-Laplace confidence set

Regularized parameter estimate:

$$\widehat{\theta}_{t,c}(\theta_0) = \nabla\mathcal{L}^{-1}\left(\frac{t}{t+c}\widehat{\mu}_t + \frac{c}{t+c}\mathcal{L}(\theta_0)\right) .$$

Bregman information gain:

$$\gamma_{t,c}(\theta_0) = \log \frac{\int_\Theta e^{-c\mathcal{B}_\mathcal{L}(\theta',\theta_0)}d\theta'}{\int_\Theta e^{-(t+c)\mathcal{B}_\mathcal{L}(\theta',\widehat{\theta}_{t,c}(\theta_0))}d\theta'} .$$

Theorem (Bregman-Laplace mixture bound for exponential families)

For any stopping time τ (adapted to the natural filtration...) and any $c > 0$,

$$\mathbb{P}\left((\tau+c)\mathcal{B}_\mathcal{L}\left(\theta, \widehat{\theta}_{\tau,c}(\theta)\right) \geq \log\frac{1}{\delta} + \gamma_{\tau,c}(\theta)\right) \leq \delta$$

Remarks

$$\mathbb{P}\left((\tau + c)\mathcal{B}_{\mathcal{L}}\left(\theta, \widehat{\theta}_{\tau,c}(\theta)\right) \geq \log\frac{1}{\delta} + \gamma_{\tau,c}(\theta)\right) \leq \delta$$

- Implicit confidence set...
  - ...but essentially level sets of convex functions: easy numerical solution.

Remarks

$$\mathbb{P}\left((\tau + c)\mathcal{B}_{\mathcal{L}}\left(\theta, \widehat{\theta}_{\tau,c}(\theta)\right) \geq \log\frac{1}{\delta} + \gamma_{\tau,c}(\theta)\right) \leq \delta$$

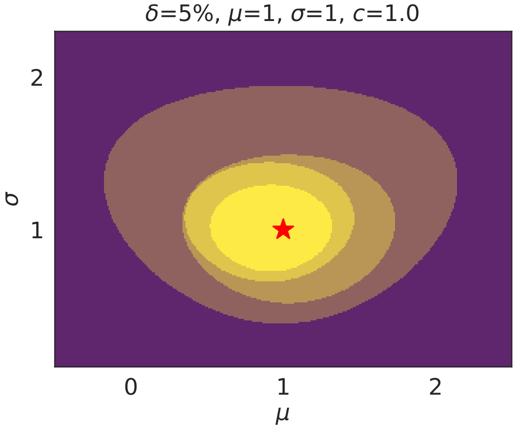
- Implicit confidence set...
  - ...but essentially level sets of convex functions: easy numerical solution.

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

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

  - Gaussian case: confidence width $\approx \mathcal{O}\left(\sqrt{\frac{\log t}{t}}\right)$.

Numerical experiments



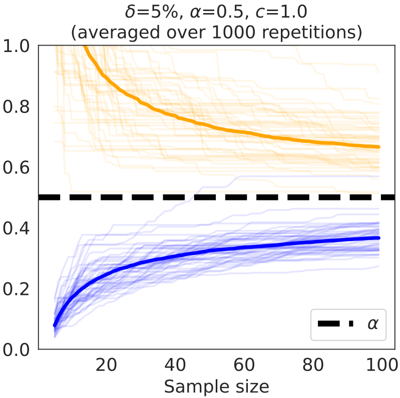
Gaussian (mean and variance) for $t \in \{10, 25, 50, 100\}$ observations

Questions?

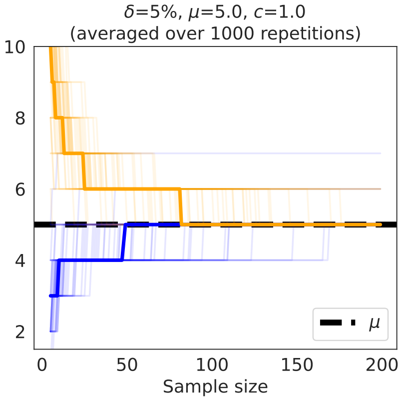
S. R. Chowdhury, P. Saux, O.-A. Maillard, and A. Gopalan. Bregman deviations of generic exponential families. arXiv preprint arXiv:2201.07306, 2022.

O.-A. Maillard. Local Dvoretzky–Kiefer–Wolfowitz confidence bands. Mathematical Methods of Statistics, 30(1):16–46, Jan 2021. ISSN 1934-8045.

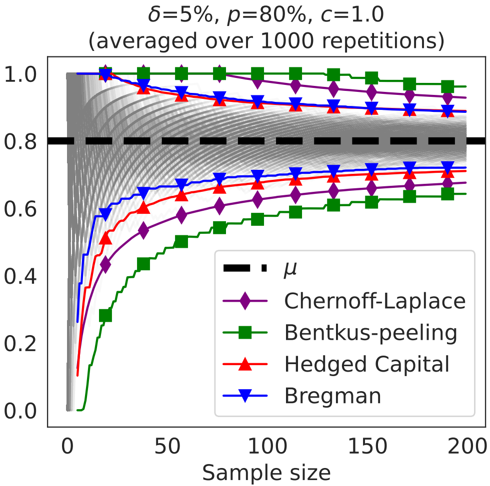
Numerical experiments



Pareto



Chi-square



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