Bregman Deviations of Generic Exponential Families

Presenter:

Akash Mondal (Sr. No. 25694)
Kintan Saha (Sr. No. 23881)

Indian Institute of Science (IISc)

1st December 2025

Outline

Motivation

- Many statistical and ML models are built from exponential families.
- Sequential problems (online learning, bandits, change-point detection, sequential testing) require **guarantees that hold uniformly over time**.
- Classical fixed-time concentration (Chernoff, Hoeffding, Bernstein) give high-probability bounds only for a fixed sample size, e.g.,

$$\Pr\left(|\hat{\theta}_n - \theta| \geq \varepsilon \right) \leq \delta.$$

Motivation

- Goal: derive **sharp, dimension-aware, time-uniform bounds** for deviations measured by Bregman divergence.
- Why Bregman? Natural geometry for exponential families; closely tied to KL divergence.
- Applications: better confidence sequences, GLR stopping rules, bandit algorithms, change-point detection.

What is a time-uniform concentration bound?

- A **time-uniform** (or uniform-in-time) bound controls a sequence of events over all n:

$$\mathbb{P}(\exists n \in \mathbb{N} : \ \mathcal{E}_n) \leq \delta.$$

- For deviations, typical goal:

$$\boxed{\mathbb{P}\Big(\exists n : \ B_L(\hat{\theta}_n, \theta) \geq \varepsilon_n\Big) \leq \delta}$$

- Useful in sequential decision-making (stopping, anytime-valid inference).

Why fixed-time bounds are insufficient

- Hoeffding/Bernstein/Chernoff provide for fixed n:

$$\mathbb{P}\big(B_L(\hat{\theta}_n, \theta) \geq \epsilon\big) \leq \alpha(n, \epsilon).$$

- Direct union bound over all n may diverge:

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} \Big\{B_L(\hat{\theta}_n, \theta) \geq \epsilon\Big\}\right) \leq \sum_{n=1}^{\infty} \alpha(n, \epsilon) = \infty.$$

- Law of the Iterated Logarithm (LIL) shows sample paths can deviate at $\sqrt{\log \log n}$-scale: fixed ε cannot hold uniformly.

Law of the Iterated Logarithm (Khinchin, 1924)

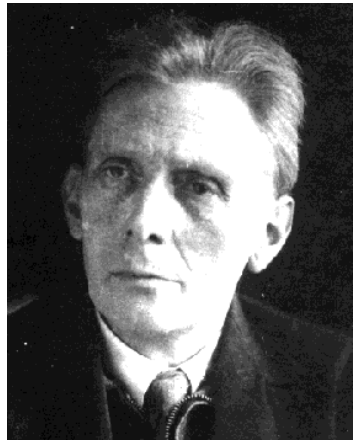
- **Due to A. Ya. Khinchin (1924):** Let $\{Y_n\}$ be i.i.d. with zero mean and unit variance, and $S_n = Y_1 + \cdots + Y_n$. Then

$$\limsup_{n\to\infty} \frac{|S_n|}{\sqrt{2n \log\log n}} = 1 \quad \text{a.s.}$$

- The classical lower bound says:

$$|\hat{\mu}_n - \mu| \approx \sqrt{\frac{2\log\log n}{n}} \quad \text{infinitely often.}$$

- Therefore no uniform bound with fixed ε is possible; ε_n must shrink and include a $\log\log n$ term.
- This motivates anytime-valid (time-uniform) bounds.



Aleksandr Yakovlevich Khinchin

(1894–1959)

Consequence: need for time-varying bounds

- Fixed–ε bounds cannot be uniformly valid due to LIL-type lower bounds.
- In exponential families, a time-uniform deviation inequality takes the form (Theorem 3):

$$(n+c)\, B_L(\theta, \theta_{n,c}(\theta)) \;\leq\; \gamma_{n,c}(\theta) + \log\left(\frac{1}{\delta}\right) \qquad \text{for all } n \geq 1.$$

- The term $\gamma_{n,c}$ is the **Bregman information gain**, which quantifies the "cost" of controlling deviations uniformly over time.
- For regular exponential families,

$$\gamma_{n,c}(\theta) \;\asymp\; \frac{d}{2} \log\left(1 + \frac{n}{c}\right)$$

giving the optimal dimension-dependent rate.

What has been solved before

- Single-parameter (1D) exponential families: sharp time-uniform bounds known (using mixture-martingale / Doob / self-normalized methods).
- Examples: Bernoulli, Gaussian mean, Poisson rate (1D).
- Gap: Arbitrary-dimensional natural parameter $\theta \in \mathbb{R}^d$ — tight, dimension-aware time-uniform concentration was open.

Overview of Techniques

- **Mixture martingales:** Construct likelihood-ratio martingales indexed by a parameter and integrate over a prior to obtain a single nonnegative supermartingale. This yields time-uniform bounds directly via Ville's inequality.
- **Stitching (background technique):** A general method to combine many fixed-time bounds into a single uniform-in-time bound using a geometric time grid and a union-bound over allocated failure probabilities.
- **Self-normalization (background technique):** Uses empirical or data-dependent variance processes to normalize deviations, producing scale-invariant or variance-adaptive bounds. Widely used in classical martingale theory and bandits.

Mixture Martingales (intuition)

- In exponential families, the process

$$M_n(\lambda) = \exp\big(n\langle \lambda, \bar{F}_n - \mu\rangle - nB_{L,\theta}(\lambda)\big)$$

  is a nonnegative martingale for each fixed λ.
- By integrating over a prior density $q(\lambda)$, we form a **mixture supermartingale**:

$$M_n = \int M_n(\lambda)\, q(\lambda)\, d\lambda.$$

- Applying **Ville's inequality(Genralization of Markov's Inequality)**:

$$\Pr\big(\exists n:\ M_n \geq \tfrac{1}{\delta}\big) \leq \delta,$$

  yields a time-uniform deviation bound automatically.
- This method exploits the geometry of exponential families through the log-partition function, giving sharp and dimension-aware bounds.

Stitching and Related Approaches

- **Stitching:** Choose a geometric time grid $n_k = \lfloor (1+\eta)^k \rfloor$. Apply fixed-time bounds at these points with allocated failures δ_k satisfying $\sum_{k=1}^{\infty} \delta_k = \delta$.
- Use the union bound:

$$\Pr(\exists k : E_{n_k}) \leq \sum_{k=1}^{\infty} \delta_k = \delta,$$

  to guarantee validity at all grid times.
- Extend the guarantee to all intermediate times using monotonicity or interpolation arguments.
- Stitching is flexible and model-agnostic, but typically produces looser bounds than mixture martingales and does not fully exploit exponential-family structure.

Exponential Family Setup

- Exponential family density (canonical form):

$$p_\theta(x) = h(x)\exp(\langle\theta, F(x)\rangle - L(\theta)), \qquad \theta \in \mathbb{R}^d.$$

- $F : X \to \mathbb{R}^d$: feature function; $h : X \to \mathbb{R}_+$: base function.
- Log-partition function:

$$L(\theta) = \log \int_X h(x)\exp(\langle\theta, F(x)\rangle)\, dx.$$

- **Examples in the exponential family:**
  - Gaussian (known variance), Bernoulli, Binomial
  - Poisson, Exponential, Gamma, Beta, Dirichlet
- **Not exponential families (do not admit a finite-dimensional sufficient statistic):**
  - Cauchy distribution
  - Uniform distribution with unknown interval

Regularity and Parameters

- Domain:
$$\mathcal{D} = \{\theta \in \mathbb{R}^d : L(\theta) < \infty\}.$$
- Regular set:
$$\mathcal{I} = \{\theta \in \mathcal{D} : \det \nabla^2 L(\theta) > 0\}.$$
- Expectation (dual) parameter:
$$\mu(\theta) = \nabla L(\theta) = \mathbb{E}_\theta[F(X)].$$
- One-to-one mapping: $\theta \leftrightarrow \mu$ on $\mathcal{I}$.

Bregman Divergence and KL

- Bregman divergence for convex L:

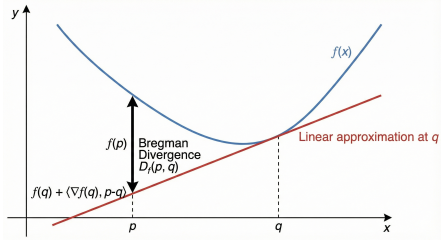
$$B_L(\theta', \theta) = L(\theta') - L(\theta) - \langle \theta' - \theta, \nabla L(\theta) \rangle.$$

- For exponential families:

$$\mathrm{KL}\big(p_\theta \| p_{\theta'}\big) = B_L(\theta', \theta).$$

- Localized form around θ:

$$B_{L,\theta}(\lambda) = L(\theta + \lambda) - L(\theta) - \langle \lambda, \nabla L(\theta) \rangle.$$



Regularized Estimators

- Let $X_1, \ldots, X_n \sim p_\theta$ be i.i.d. samples, where $\theta \in \mathcal{I}$, and let θ_0 be a fixed reference parameter.
- For any constant $c > 0$, define the regularized (or smoothed) estimator

$$\theta_{n,c}(\theta_0) = (\nabla L)^{-1}\left(\frac{\sum_{t=1}^{n} F(X_t) + c\,\nabla L(\theta_0)}{n + c}\right).$$

- This estimator $\theta_{n,c}$ is in fact the **MAP estimator** under a conjugate prior centered at θ_0, with strength c.
- The associated "information gain" compares how sharply the posterior concentrates around $\theta_{n,c}$ relative to the reference θ_0. This quantity will later determine the width of time-uniform deviation bounds.

Why Bregman divergence

- Bregman divergence is the natural geometry induced by the log-partition function L.
- It coincides with KL divergence within exponential families:

$$\mathrm{KL}(p_\theta \,\|\, p_{\theta'}) \;=\; B_L(\theta', \theta).$$

- **Order matters:** $B_L(\theta', \theta)$ corresponds to $\mathrm{KL}(p_\theta \| p_{\theta'})$; swapping arguments changes the value.
- Working in the Bregman geometry translates KL/log-likelihood ratios directly into convex-analytic quantities, enabling mixture martingale analysis.

Why not work directly with KL?

- KL between composite (empirical) and true distributions behaves like Bregman but is awkward in high dimensions.
- Bregman on parameter space aligns with parameter estimates and with convex-analytic tools (Legendre duality).
- KL of distributions reduces to Bregman on parameters for exponential families — making parameter-space bounds both interpretable and tractable.

Bregman information gain (definition)

- Regularized estimator:

$$\theta_{n,c}(\theta_0) = (\nabla L)^{-1}\left(\frac{\sum_{t=1}^{n} F(X_t) + c\,\nabla L(\theta_0)}{n+c}\right).$$

- **Bregman information gain**

$$\gamma_{n,c}(\theta_0) = \log\left(\frac{\int \exp(-c\,B_L(\theta', \theta_0))\; d\theta'}{\int \exp(-(n+c)\,B_L(\theta', \theta_{n,c}(\theta_0)))\; d\theta'}\right).$$

- Intuition: compares posterior normalizers; quantifies how much information the data provided relative to prior regularization.
- measures a form of information gain about the true parameter after n samples expressed in the Bregman divergence.

Main Theorem (informal)

- **Informal:** With probability at least $1-\delta$, for all $n \geq 1$,

$$(n+c)\, B_L(\theta_0, \theta_{n,c}(\theta_0)) \leq \log\left(\tfrac{1}{\delta}\right) + \gamma_{n,c}(\theta_0).$$

- Holds for any regular exponential family and any reference $\theta_0 \in \mathcal{I}$.
- Interpretable: Bregman divergence times sample size is bounded by prior-data information $+ \log(1/\delta)$.
- $$\boxed{\Theta_{n,c}(\delta) = \left\{\theta_0 \in \Theta : (n+c)\, B_L\big(\theta_0, \theta_{n,c}(\theta_0)\big) \leq \log\left(\tfrac{1}{\delta}\right) + \gamma_{n,c}(\theta_0)\right\}.}$$
- The confidence sequence satisfies the **anytime-valid bound**:

$$\mathbb{P}_\theta(\theta \notin \Theta_{n,c}(\delta) \text{ for some } n \geq 1) \leq \delta.$$

- Equivalently,

$$\mathbb{P}_\theta(\theta \in \Theta_{n,c}(\delta) \;\; \forall n \geq 1) \geq 1-\delta.$$

Main Theorem (consequences)

- Rearranged gives confidence sequences:

$$B_L(\theta_0, \theta_{n,c}(\theta_0)) \leq \frac{\log(1/\delta) + \gamma_{n,c}(\theta_0)}{n + c}.$$

- Typical growth of $\gamma_{n,c}$ is $\frac{d}{2}\log(1 + n/c) + O(1)$ for regular models.

Proof Sketch - Lemma 1: Tail & Duality Properties

- **MGF identity for exponential families:** For $\theta \in \Theta_I$ and any λ such that $\theta + \lambda \in \Theta_D$,

$$\log \mathbb{E}_\theta \left[\exp\left(\langle \lambda, F(X) - \nabla L(\theta) \rangle \right) \right] = B_{L,\theta}(\lambda),$$

where $B_{L,\theta}(\lambda) = L(\theta + \lambda) - L(\theta) - \langle \lambda, \nabla L(\theta) \rangle = B_L(\theta + \lambda, \theta)$.

- **Bregman duality relations):** If ∇L is one-to-one, then for all $\theta, \theta' \in \Theta_D$ where G^* is the Legendre-Frenchel dual of G,

$$B_L(\theta', \theta) = B^\star_{L,\theta'}(\nabla L(\theta) - \nabla L(\theta')) = B_{L^\star}(\nabla L(\theta), \nabla L(\theta')).$$

- **Interpolation identity under expectation parametrization:** For any $\alpha \in [0, 1]$,

$$B^\star_{L,\theta'}(\alpha(\nabla L(\theta) - \nabla L(\theta'))) = B_L(\theta', \theta_\alpha),$$

with

$$\theta_\alpha = \nabla L^{-1}\big(\alpha \nabla L(\theta) + (1 - \alpha) \nabla L(\theta')\big).$$

Proof Sketch - Step 2: Martingale construction

- Given i.i.d. samples $X_1, \ldots, X_n \sim p_\theta$, define:

$$\bar{F}_n = \frac{1}{n} \sum_{t=1}^{n} F(X_t), \quad \mu = \nabla L(\theta) = \mathbb{E}_\theta[F(X)].$$

- For any $\lambda \in \Lambda_\theta := \{\lambda : \theta + \lambda \in \Theta\}$, construct:

$$M_n^\lambda = \exp\big(n\langle \lambda, \bar{F}_n - \mu \rangle - nB_{L,\theta}(\lambda)\big) .$$

- Then:

$$\mathbb{E}_\theta[M_n^\lambda \mid \mathcal{F}_{n-1}] = M_{n-1}^\lambda, \quad \Rightarrow M_n^\lambda \text{ is a martingale}, \quad \mathbb{E}_\theta[M_n^\lambda] = 1.$$

Proof Sketch - Step 3: Mixture Martingale

- **Define the exponential-family induced conjugate prior:**

$$q(\theta \mid \alpha, \beta) = H(\alpha, \beta) \exp(\langle \theta, \alpha \rangle - \beta L(\theta)),$$

where $H(\alpha, \beta)$ is the normalizing constant ensuring q integrates to 1.

- **Construct the mixture martingale:**

$$M_n = \int_{\Lambda_\theta} M_n^\lambda \, q(\theta + \lambda \mid \alpha, \beta) \, d\lambda,$$

where $M_n^\lambda = \exp\big(n\langle \lambda, \bar{F}_n - \mu \rangle - nB_{L,\theta}(\lambda)\big)$ is the pointwise martingale from Step-2.

- Taking $\alpha = c\nabla L(\theta), \quad \beta = c$

$$M_n = G(\theta, c) \int_{\Lambda_\theta} \exp\Big(n\langle \lambda, \bar{F}_n - \mu \rangle - (n + c) B_{L,\theta}(\lambda)\Big) \, d\lambda,$$

where

$$G(\theta, c) = \left[\int_\Theta \exp(-cB_L(\theta', \theta)) d\theta' \right]^{-1}.$$

Proof Sketch - Step 4: Ville's inequality

- $M_n \geq 0$ is a **non-negative supermartingale** under $\mathbb{P}_\theta$ and $\mathbb{E}_\theta[M_n] \leq 1$.
- Since $M_n \geq 0$ is a supermartingale:

$$\mathbb{P}_\theta\left(\exists n \geq 1 : M_n \geq \frac{1}{\delta}\right) \leq \delta \quad \text{(Ville's Inequality)}.$$

Thus for any stopping time N:

$$\mathbb{P}_\theta(M_N \geq 1/\delta) \leq \delta.$$

- Using the duality properties

$$M_n = \exp\big((n+c)B^\star_{L,\theta}(x)\big)\,\frac{G(\theta, c)}{G(\theta_{n,c}(\theta),\, n+c)}.$$

Proof Sketch - Step 5: Ville's inequality on M_n

- Using the explicit expression

$$M_n = \exp\big((n+c)B^{\star}_{L,\theta}(x)\big)\,\frac{G(\theta,c)}{G(\theta_{n,c}(\theta),\,n+c)}, \quad x = \frac{n}{n+c}(\bar{F}_n - \mu),$$

  we see that the event

$$\left\{(n+c)B^{\star}_{L,\theta}(x) \geq \log\frac{G(\theta_{n,c}(\theta),\,n+c)}{\delta\,G(\theta,c)}\right\} \subseteq \{M_n \geq 1/\delta\}.$$

- Hence, for any fixed n,

$$\mathbb{P}_\theta\left(B^{\star}_{L,\theta}\Big(\tfrac{n}{n+c}(\bar{F}_n - \mu)\Big) \geq \frac{1}{n+c}\log\frac{G(\theta_{n,c}(\theta),\,n+c)}{\delta\,G(\theta,c)}\right) \leq \delta.$$

Proof Sketch - Step 6: From $B^\star$ to B_L and $\gamma_{n,c}$

- By Bregman duality (Lemma 1),

$$B^\star_{L,\theta}\Big(\tfrac{n}{n+c}(\bar{F}_n - \mu)\Big) = B_L\big(\theta, \theta_{n,c}(\theta)\big).$$

- Define the Bregman information gain:

$$\gamma_{n,c}(\theta) = \log \frac{\displaystyle\int_\Theta e^{-cB_L(\theta',\theta)}\, d\theta'}{\displaystyle\int_\Theta e^{-(n+c)B_L(\theta',\theta_{n,c}(\theta))}\, d\theta'} = \log \frac{G(\theta_{n,c}(\theta),\, n+c)}{G(\theta, c)}.$$

- Plugging this and duality into the previous inequality yields

$$(n+c)\, B_L\big(\theta, \theta_{n,c}(\theta)\big) \le \log \frac{1}{\delta} + \gamma_{n,c}(\theta),$$

which is the desired deviation inequality for a fixed n.

Remarks - Comparison with Kaufmann & Koolen

- Kaufmann and Koolen in *Mixture martingales revisited with applications to sequential tests and confidence intervals*, JMLR 2021 had tried to use the mixture martingale framework to derive concentration results for exponential families.
- Their mixture construction uses **discrete mixtures + stitching**, limited to 1-D families (Gaussian with known variance, Gamma with fixed shape).
- Their discrete prior leads to technical constants not intrinsic to the exponential-family geometry.
- The paper's prior is exponential-family induced and produces the natural normalizer ratio

$$\log \frac{G(\theta_{n,c}(\theta),\, n+c)}{G(\theta, c)} = \gamma_{n,c}(\theta),$$

yielding a shorter and more elegant proof in Bregman geometry.

Remarks - Asymptotics via Laplace

- Since $\nabla^2 L(\theta) \succ 0$ for $\theta \in \Theta_I$, locally

$$B_L(\theta + \lambda, \theta) = \tfrac{1}{2}\lambda^\top \nabla^2 L(\theta)\lambda + o(\|\lambda\|^2).$$

- Applying Laplace's method to the denominator integral

$$\int_\Theta \exp(-(n+c)B_L(\theta', \theta_{n,c}(\theta)))d\theta',$$

gives the asymptotic

$$\gamma_{n,c}(\theta) = \frac{d}{2}\log(1 + n/c) + O(1).$$

- The scaling is worse than that provided by LIL and is a well known drawback of the method of mixtures. This is compensated for by better non asymptotic performance
- A size-dependent $c_n \propto n$ would break the supermartingale property and is ruled out by the LIL; therefore, c must remain fixed.

Remarks - On the role of a Legendre regularizer

- Switching from a constant c to a **Legendre function** $L_0(\theta)$ enables **global regularization** that does *not* rely on the true data-generating parameter.
- This yields **more explicit and flexible confidence sets**, since the estimator and information gain no longer depend on the unknown ground-truth parameter.
- The Legendre prior allows handling **non-independent or non-identically distributed data**, extending guarantees beyond the classical i.i.d. setting. (Useful in contextual and adaptive online learning.)
- However, unlike local c-based regularization, the Legendre approach requires **integrability of the induced prior**, which may fail for some families, making closed-form computation rarer and often requiring numerical inversion.
- A natural (but not mandatory) choice is to use the **log-partition function itself as** L_0 to preserve geometry, but the framework remains valid for any smooth Legendre function satisfying the integrability condition.
- The authors also argue such a global regularization is actually less convenient to use except for univariate and multivariate Gaussian distributions which have invariant geometry thus removing the need for a regularizer.

Specialization to Classical Families

- The authors derive the explicit Bregman information gain and confidence sets for various classical families like Gaussian (unknown mean and variance), Bernoulli, Exponential, Pareto, Chi-Square etc
- The derived results allow for the construction of high probability confidence sets for Chi-Square distributions which hasn't been studied extensively in prior work.

Bregman information gain - Gaussian and Bernoulli

Gaussian (unknown mean and variance). Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Define

$$S_n = \sum_{t=1}^{n} X_t, \quad \hat{\mu}_n = \frac{S_n}{n}, \quad Z_n(\mu, \sigma) = \frac{1}{\sigma^2} \sum_{t=1}^{n} (X_t - \mu)^2.$$

For reference $\mu_0 \in \mathbb{R}$, $\sigma_0 \in \mathbb{R}_+$, $c > 0$, the Bregman information gain is

$$\gamma_{n,c}^{\mathcal{N}}(\mu_0, \sigma_0) = \frac{3}{2} \log\left(\frac{n}{n+c} Z_n(\hat{\mu}_n, \sigma_0) + \frac{c}{n+c} Z_n(\mu_0, \sigma_0) + c \right) + f_{n,c},$$

Bernoulli. Let $X \sim \mathrm{Bernoulli}(\mu)$. For reference $\mu_0 \in \mathbb{R}$, $c > 0$, define

$$\mu_{n,c}(\mu_0) = \frac{S_n + c\mu_0}{n+c}.$$

Then

$$\gamma_{n,c}^{\mathrm{Bernoulli}}(\mu_0) = c\, H(\mu_0) - (n+c)\, H(\mu_{n,c}(\mu_0)) + \log \frac{\mathrm{B}(c\mu_0,\, c(1-\mu_0))}{\mathrm{B}((n+c)\mu_{n,c}(\mu_0),\, (n+c)(1-\mu_{n,c}(\mu_0)))},$$

where $\mathrm{B}(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\, du$ and $H(\cdot)$ is Bernoulli entropy.

Bregman information gain - Exponential, Pareto, Chi-square

Exponential. Let $X \sim \mathrm{Exp}(1/\mu)$ with mean $\mu > 0$. For reference $\mu_0 > 0$, $c > 0$,

$$\gamma_{n,c}^{\mathrm{Exp}}(\mu_0) = \log\Big(\frac{S_n}{\mu_0} + c\Big) + \log\frac{\Gamma(c)}{\Gamma(n+c)} + (n+c-1)\log(n+c) - c\log c - n.$$

Pareto (fixed scale). Let $X \sim \mathrm{Pareto}(\alpha)$ with unknown shape $\alpha > 0$, and $L_n = \sum_{t=1}^{n} \log X_t$. For reference $\alpha_0 > 0$, $c > 0$,

$$\gamma_{n,c}^{\mathrm{Pareto}}(\alpha_0) = \log(\alpha_0 L_n + c) + \log\frac{\Gamma(c)}{\Gamma(n+c)} + (n+c-1)\log(n+c) - c\log c - n.$$

Chi-square. Let $X \sim \chi^2(k)$ with unknown degrees of freedom $k \in \mathbb{N}$, and $K_n = \sum_{t=1}^{n} \log X_t$. For reference $k_0 \in \mathbb{N}$, $c > 0$, let $k_{n,c}(k_0)$ solve

$$\psi_0\Big(\frac{k_{n,c}(k_0)}{2}\Big) = \frac{K_n + c\,\psi_0(k_0/2)}{n+c},$$

where ψ_0 is the digamma function. Then

$$\gamma_{n,c}^{\chi^2}(k_0) = \frac{k_{n,c}(k_0)}{2}\Big(K_n + c\,\psi_0(k_0/2)\Big) - (n+c)\log\Gamma\Big(\frac{k_{n,c}(k_0)}{2}\Big) + c\log\Gamma\Big(\frac{k_0}{2}\Big) - c\frac{k_0}{2}\psi_0\Big(\frac{k_0}{2}\Big) + \log\frac{J(c,\, c\psi_0(k_0/2))}{J(n+c,\, (K_n + c\psi_0(k_0/2)))},$$

with $J(a,b) = \sum_{k'=1}^{\infty} \exp\Big(-a\log\Gamma\Big(\frac{k'}{2}\Big) + b\,\frac{k'}{2}\Big)$.

Confidence Sequences

- Construct confidence regions on θ for all n:

$$\mathcal{C}_n = \{\theta' : B_L(\theta', \theta_{n,c}(\theta')) \leq \frac{\log(1/\delta)+\gamma_{n,c}(\theta')}{n+c}\}.$$

- These are valid simultaneously over time with coverage $1 - \delta$.

Application: Confidence Sequences

Distribution	Parameters	Confidence Set
Gaussian	$\mu \in \mathbb{R}$ $\sigma \in \mathbb{R}_+$	$\frac{1}{2} Z_n(\mu,\sigma) - \frac{n+c+3}{2} \log\left(\frac{n}{n+c} Z_n(\widehat{\mu}_n,\sigma) + \frac{c}{n+c} Z_n(\mu,\sigma) + c\right)$ $\leqslant \log\frac{1}{\delta} - \frac{n}{2}\log 2 - \left(\frac{c}{2}+2\right)\log c + \frac{1}{2}\log(n+c) + \log\frac{\Gamma\left(\frac{c+3}{2}\right)}{\Gamma\left(\frac{n+c+3}{2}\right)}$
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	$\mu \in \mathbb{R}_+$	$\frac{S_n}{\mu} - (n+c+1)\log\left(\frac{S_n}{\mu} + c\right)$ $\leqslant \log\frac{1}{\delta} + \log\frac{\Gamma(c)}{\Gamma(n+c)} - \log(n+c) - c\log c$
Pareto	$\alpha \in \mathbb{R}$	$\alpha L_n - (n+c+1)\log(\alpha L_n + c)$ $\leqslant \log\frac{1}{\delta} + \log\frac{\Gamma(c)}{\Gamma(n+c)} - \log(n+c) - c\log c$
Chi-square	$k \in \mathbb{N}$	$n\log\Gamma\left(\frac{k}{2}\right) - \frac{k}{2}K_n - \log J\left(c, c\psi_0\left(\frac{k}{2}\right)\right)$ $+ \log J\left(n+c, K_n + c\psi_0\left(\frac{k}{2}\right)\right) \leqslant \log\frac{1}{\delta}$

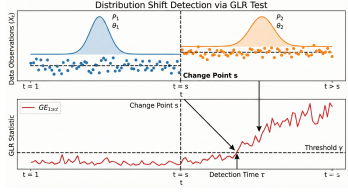
Application: Generalized Likelihood Ratio (GLR)

- We observe a sequence $X_1, X_2, \ldots$ and aim to detect any change in its generating distribution. The time-uniform Bregman deviation bounds enable tight control of the **false alarm rate** of the GLR test used for this purpose.
- In an exponential family model $\mathcal{E}$, the GLR stopping time for threshold $\beta > 0$ is

$$\tau(\beta; \mathcal{E}) = \min \left\{ t \in \mathbb{N} : \max_{0 \leq s < t} G^{\mathcal{E}}_{1:s:t} \geq \beta \right\}.$$

- The GLR statistic is defined as

$$G^{\mathcal{E}}_{1:s:t} = \inf_{\theta} \sup_{\theta_1, \theta_2} \left[\log \frac{\prod_{i=1}^{s} p_{\theta_1}(X_i)}{\prod_{i=1}^{t} p_{\theta}(X_i)} + \log \frac{\prod_{i=s+1}^{t} p_{\theta_2}(X_i)}{1} \right].$$



Application: Generalized Likelihood Ratio (GLR)

- Assume the null hypothesis (no change): $X_t \sim p_\theta$ for all t.
- The false alarm probability is

$$\Pr\left(\tau(\beta;\mathcal{E}) < \infty\right) = \Pr\left(\exists\, s < t : G^{\mathcal{E}}_{1:s:t} \geq \beta\right).$$

- Solving the optimizations over θ_1 and θ_2, the GLR statistic becomes

$$G^{\mathcal{E}}_{1:s:t} = \inf_{\theta} \Big[s\, B_L(\theta,\, \theta_{1:s})1 + (t-s)\, B_L(\theta,\, \theta_{s+1:t}) \Big].$$

where the segment estimators satisfy

$$\nabla L(\theta_{1:s}) = \frac{1}{s}\sum_{i=1}^{s} F(X_i), \qquad \nabla L(\theta_{s+1:t}) = \frac{1}{t-s}\sum_{i=s+1}^{t} F(X_i).$$

- Thus,

$$\Pr(\tau(\beta;\mathcal{E}) < \infty) \leq \Pr(\exists s : sB_L(\theta_{1:s},\theta) \geq \beta_1) + \Pr(\exists s < t : (t-s)B_L(\theta_{s+1:t},\theta) \geq \beta_2),$$

with $\beta_1 + \beta_2 = \beta$.

Application: Generalized Likelihood Ratio (GLR)

- The first term is controlled directly by the time-uniform deviation bound (Theorem 3) using a regularized estimator of $\theta_{1:s}$.
- For the second term we need a **doubly time-uniform** bound: uniform over both window start s and window end t.
- For any $s < t$, $c > 0$, and reference θ_0, define the regularized window estimator

$$\theta_{s+1:t,c}(\theta_0) = (\nabla L)^{-1}\left(\frac{\sum_{i=s+1}^{t} F(X_i) + c\,\nabla L(\theta_0)}{t-s+c}\right).$$

- The corresponding window information gain $\gamma_{s+1:t,c}(\theta_0)$ is defined analogously to Definition 2.
- **Theorem 5 (Doubly Time-Uniform Concentration).** Let $\delta \in (0,1]$ and $g : \mathbb{N} \to \mathbb{R}_+$ satisfy $\sum_{t=1}^{\infty} \frac{1}{g(t)} \le 1$ (e.g. $g(t) = (1+t)\log^2(1+t)$). Then

$$\Pr(\exists\, s < t : (t-s+c)\, B_L(\theta_{s+1:t,c}(\theta_0), \theta) > \log g(t) + \gamma_{s+1:t,c}(\theta_0)) \le \delta.$$

- This yields uniform control over *all* possible windows $[s+1:t]$ and therefore controls the false alarm rate of the GLR test.

Application: Bandits and Online Learning

- Time uniform concentration enables analysis in fixed error BAI.
- PAC Bayes bounds for fixed tolerance.
- Active Learning in Bandit problems.
- Concentration of Laplace distribution in Differential Privacy.
- Heavy tailed distributions in risk averse/ corrupted bandits.
- Confidence sequences enable anytime upper confidence bounds (UCB) for arm means in parametric bandits.

Application: Sequential Testing

- Sequential tests (e.g., SPRT-like procedures) can use these time-uniform bounds to control error probabilities while stopping adaptively.
- Works for composite hypotheses in exponential families.

Limitations

- Requires smoothness and regularity (invertible Hessian) in parameter region.
- Laplace approximation introduces constants; worst-case higher-order terms need careful control.
- Complexity of computing $\gamma_{n,c}$ in large d may be nontrivial (but often approximable).

Future directions

- Exploring alternative proof strategies to match the bound given by the Law of Iterated Logarithm.
- Extensions to dependent data other than non-iid data (Markov, mixing processes).
- Non-smooth exponential families or constrained parameter spaces.
- Efficient computation/approximation of $\gamma_{n,c}$ in high dimension.

Related Work

This paper has been cited in the following peer-reviewed works:

- Marc Jourdan *et al.*, *Dealing with Unknown Variances in Best-Arm Identification*, ALT 2023. PDF
- Mohamed-Hicham Leghettas *et al.*, *Learning Bregman Divergences with Application to Robustness*, NeurIPS 2023. PDF
- Shubhanshu Shekhar *et al.*, *Sequential Changepoint Detection via Backward Confidence Sequences*, ICML 2023. PDF
- Nicolas Emmenegger *et al.*, *Likelihood Ratio Confidence Sets for Sequential Decision Making*, NeurIPS 2023. PDF
- Junghyun Lee *et al.*, *Improved Regret Bounds of (Multinomial) Logistic Bandits via Regret-to-Confidence-Set Conversion*, AISTATS 2024. PDF

Conclusion

- Introduced time-uniform, dimension-aware Bregman deviation bounds for general exponential families.
- Central concept: **Bregman information gain** $\gamma_{n,c}$.
- Broad applicability: confidence sequences, GLR, bandits, sequential testing.
- Balances statistical tightness with geometric insight coming from L.

References

- Chowdhury, Sayak Ray, Patrick Saux, Odalric Maillard, and Aditya Gopalan. *"Bregman deviations of generic exponential families."* COLT 2023. PDF

Thanks!

Thank you!

Questions?