Topic 5
Self-normalization I

Victor H. de la Peña

Professor of Statistics, Columbia University

Artificial Intelligence Institute for Advances in Optimization
Georgia Institute of Technology 2024

Topics Preview

Topics Preview

Self-normalized statistic, literally, takes the form $\frac{A_n}{B_n}$ (resp., $\frac{A_t}{B_t}$ for continuous cases), where both A_n and B_n are functions of your observations $X_1, ..., X_n$ (resp., A_t, B_t the function of $(X_s)_{0\leq s\leq t}$). One of the bonuses of self-normalization is that you can obtain a statistic, which, with sample size n increasing, maintains a bounded tail probability. Some self-normalized statistics you may encounter are, for example, sample Gini coefficient $\hat{G} = \frac{1}{2}\frac{\frac{1}{n(n-1)}\sum 1\leq i,j\leq n|X_i - X_j|}{\bar{X}_n}$ of positively supported random variables, which almost surely takes value between 0 and 1, and sample squared coefficient of variation $\widehat{cv}^2 := \frac{s_n^2}{\bar{X}_n^2}$, which cancels the scaling.

A celebrated self-normalized statistic you learned in your undergraduate time is Student's t-statistic, by W. Gosset. Recall that you have $\{X_i\}$ i.i.d normal $\mathcal{N}(\mu, \sigma^2)$, and the sample mean and sample variance, respectively defined by

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} \qquad s_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n-1},$$

then the t-statistic is constructed: $T_n = \frac{\bar{X}_n - \mu}{s_n/\sqrt{n}} \sim t_{n-1}$, the t-distribution with degree of freedom n-1. In addition, if you denote by $Y_i = X_i - \mu$, $A_n = \sum_{i=1}^n Y_i$, $B_n^2 = \sum_{i=1}^n Y_i^2$, then you will find that,

$$T_n = \frac{\frac{A_n}{B_n}}{\sqrt{(n - (A_n/B_n)^2)/(n-1)}},$$

a function of the self-normalized statistic $\frac{A_n}{B_n}$.

Notably, W. GOSSET's "Student" t-statistic allowed statistical inference about the value of the mean of a (Gaussian) distribution without knowledge of the actual value of the variance, provided one has a random sample from the target population.

Several years later, B. EFRON, the founder of the bootstrap, developed a self-normalized inequality for independent symmetric variables. And after EFRON, there were some more developments in self-normalization for independent variables, and, for martingales or dependent variables.

Figure: William S. Gosset, June 13, 1876 - October 16, 1937

Topics Preview

Bernstein's inequality

The first time I studied Self-normalization was when I tried to generalize Bernstein's inequality for self-normalized martingales. Let us first recall this inequality:

Theorem (Bernstein's inequality)

Let $\{x_i\}$ be a sequence of independent variables. Assume that $\mathbb{E}(x_i) = 0$ and $\mathbb{E}(x_i^2) = \sigma_i^2 < \infty$ and set $x_n^2 = \sum_{i=1}^n \sigma_i^2$. Furthermore, assume that there exists a constant $0 < c < \infty$ such that, almost surely, $\mathbb{E}(|x_i|^k) \leq (k!/2)\sigma_i^2 c^{k-2}$ for all $k > 2$ (satisfied by subexponential random variables). Then for all $x > 0$.

$$\mathbb{P}\left(\sum_{i=1}^n x_i > x\right) \leq \exp\left(-\frac{x^2}{2(v_n^2 + cx)}\right).$$

Bernstein's inequality for self-normalized martingales

Theorem (de la Peña, 1999, [2])

Let $\{d_i\}$ be a martingale difference sequence w.r.t. filtration $\{\mathcal{F}_i\}$. Assume that $\mathbb{E}(d_j|\mathcal{F}_{j-1}) = 0$ and $\mathbb{E}(d_j^2|\mathcal{F}_{j-1}) = \sigma_j^2 < \infty$ and set $V_n^2 = \sum_{j=1}^n \sigma_j^2$. Furthermore, assume that there exists a constant $0 < c < \infty$ such that, almost surely, $\mathbb{E}(|d_j|^k \Big| \mathcal{F}_{j-1}) \leq (k!/2)\sigma_j^2 c^{k-2}$ for all $k > 2$. Then for all $x, y > 0$.

$$\mathbb{P}\left(\frac{\sum_{i=1}^n d_i}{V_n^2} > x, \frac{1}{V_n^2} \leq y\right) \leq \exp\left(-\frac{x^2}{2(y+cx)}\right).$$

The inequality is a sharp extension for Bernstein's inequality, since when $V_n^2 = v_n^2$ (nonrandom) the two inequalities are equivalent. The key steps in obtaining this result involve the use of Markov's inequality followed by the decoupling inequality in Lecture 4 (see de la Peña [1] and [2])

$$\mathbb{E}\left[g\exp(\lambda\sum_{i=1}^{n} d_i)\right] \leq \sqrt{\mathbb{E}\left[g^2\exp\left(2\lambda\sum_{i=1}^{n} y_i\right)\right]}.$$

We then (conditionally) apply the standard results for sums of independent random variables to complete the proof. The entire proof is too long to present here, and you may find it in [2].

Topics Preview

Auto-Regressive Processes

An example of self-normalized processes in dependent variables arose in the context of Maximum Likelihood Estimators (MLEs) for the parameter in auto-regressive (AR) processes. I will use this example to summon the **canonical assumption**, which will be a pivotal topic in the next two days. Let us consider the following Auto-Regressive Gaussian process $(Y_i)_0^\infty$, such that

$$Y_i = \alpha Y_{i-1} + \epsilon_i, \qquad Y_0 = 0, \tag{1}$$

where $\alpha \neq 0$ is a fixed, unknown parameter and ϵ_i are independent standard normal random variables $\mathcal{N}(0,1)$.

To obtain the MLE of α, we establish our log-likelihood function

$$\begin{aligned} l(\alpha; Y_1, ..., Y_n) &= \log_\alpha f(Y_1, ..., Y_n) \\ &= \sum_{j=1}^{n} (Y_j - \alpha Y_{j-1})^2/2 - n \log(\sqrt{2\pi}). \end{aligned}$$

Taking the derivative w.r.t. α, equating to zero and solving for α, we obtain the MLE for α,

$$\hat{\alpha} = \frac{\sum_{j=1}^{n} Y_{j-1} Y_j}{\sum_{j=1}^{n} Y_{j-1}^2} = \frac{\sum_{j=1}^{n} Y_{j-1}(\alpha Y_{j-1} + \epsilon_j)}{\sum_{j=1}^{n} Y_{j-1}^2} = \alpha + \frac{\sum_{j=1}^{n} Y_{j-1}\epsilon_j}{\sum_{j=1}^{n} Y_{j-1}^2}. \quad (2)$$

You now find, without much surprise, that this MLE is a self-normalization, and so is $\hat{\alpha} - \alpha$:

$$\hat{\alpha} - \alpha = \frac{\sum_{j=1}^{n} Y_{j-1}\epsilon_j}{\sum_{j=1}^{n} Y_{j-1}^2}. \quad (3)$$

We now construct the filtration $\mathcal{F} := \sigma(Y_1, ..., Y_n; \epsilon_1, ...\epsilon_n)$, and therefore the numerator $\sum_{j=1}^{n} Y_{j-1}\epsilon_j =: A_n$ is a martingale w.r.t. $\mathcal{F}$. And the denominator

$$\sum_{j=1}^{n} Y_{j-1}^2 = \sum_{j=1}^{n} \mathbb{E}[Y_{j-1}^2 \epsilon_j^2 | \mathcal{F}_{j-1}] =: B_n^2 \tag{4}$$

is the conditional variance of A_n. Thus $\hat{\alpha} - \alpha = \frac{A_n}{B_n^2}$ is a process self-normalized by the conditional variance. Since ϵ_i's are $\mathcal{N}(0, 1)$, then we have that for any $\lambda \in \mathbb{R}$,

$$M_n := exp\left(\lambda A_n - \frac{\lambda^2 B_n^2}{2}\right) \tag{5}$$

is an exponential martingale w.r.t. $\mathcal{F}_n$ (I leave this claim as an exercise left for you to verify). With optimal stopping theorem, you have that $\mathbb{E}(M_n) = \mathbb{E}(M_1) = 1$, for all $n \geq 1$, which leads that

$$\mathbb{E}exp\left(\lambda A_n - \frac{\lambda^2 B_n^2}{2}\right) \leq 1 \tag{6}$$

the canonical assumption in the next section.

Topics Preview

Canonical Assumption

So far, we have found an inequality (6). And such inequality, for a pair of random variables A, B with $B > 0$, taking the general form

$$\mathbb{E} \exp(\lambda A - \lambda^2 B^2/2) \leq 1, \tag{7}$$

frequently appears in probability theory and stochastic analysis. There are three regimes of interest: (7) holds

- for all real λ;
- for all $\lambda \geq 0$;
- for all $0 \leq \lambda < \lambda_0$, where $0 < \lambda < \infty$.

Gaussian Bounds

In this lecture, we only focus on the first case, i.e.,

$$\mathbb{E}\exp(\lambda A - \lambda^2 B^2/2) \leq 1,$$

holds for all real λ. One theorem that bounds the tail probability of A/B^2, with a constraint for B^2 is given below.

Theorem (de la Peña, Klass and Lai, 2004 [3])

Under the canonical assumption for all real λ,

$$\mathbb{P}\left(\frac{A}{B^2} > x, \frac{1}{B^2} \leq y\right) \leq \exp\left(-\frac{x^2}{2y}\right) \tag{8}$$

for all $x, y > 0$.

Proof

The key here is to **"keep" the indicator** when using MARKOV Inequality. In fact, for all measurable set $\boldsymbol{S}$,

$$\begin{aligned}
\mathbb{P}(\frac{A}{B^2} > x, \boldsymbol{S}) &= \mathbb{P}(\exp(A) > \exp(xB^2), \boldsymbol{S}) \\
&\leq \inf_{\lambda>0} \mathbb{E}\left[\exp(\frac{\lambda}{2}A - \frac{\lambda}{2}xB^2)\mathbb{I}_{\{A/B^2>x,\boldsymbol{S}\}}\right] \\
&= \inf_{\lambda>0} \mathbb{E}\left[\exp(\frac{\lambda}{2}A - \frac{\lambda^2}{4}B^2 - (\frac{\lambda}{2}x - \frac{\lambda^2}{4})B^2)\mathbb{I}_{\{A/B^2>x,\boldsymbol{S}\}}\right] \\
&\leq \inf_{\lambda>0} \sqrt{\mathbb{E}\left[\exp(\lambda A - \frac{\lambda^2}{2}B^2\right]} \sqrt{\mathbb{E}\left[-(\lambda x - \frac{\lambda^2}{2})B^2)\mathbb{I}_{\{A/B^2>x,\boldsymbol{S}\}}\right]}
\end{aligned}$$

by the CAUCHY-SCHWARZ inequality.

Proof Cont'd

So far, we have

$$\mathbb{P}(\frac{A}{B^2} > x, \boldsymbol{S}) \leq$$
$$\inf_{\lambda>0} \sqrt{\mathbb{E}\left[\exp(\lambda A - \frac{\lambda^2}{2}B^2)\right]} \sqrt{\mathbb{E}\left[-(\lambda x - \frac{\lambda^2}{2}B^2)\mathbb{I}_{\{A/B^2>x,\boldsymbol{S}\}}\right]}$$

The first term in the last inequality is bounded by 1, by the canonical assumption. The value minimizing the second term is $\lambda = x$, and therefore

$$\mathbb{P}(A/B^2 > x, \boldsymbol{S}) \leq \sqrt{\mathbb{E}\left[\frac{-x^2B^2}{2}\right]\mathbb{I}_{\{A/B^2>x,\boldsymbol{S}\}}}.$$

Let us set $\boldsymbol{S} = \{\frac{1}{B^2} < y\}$ and our theorem follows as we claimed.

Now let us go back to our AR model and recall

$$\hat{\alpha} - \alpha = \frac{\sum_{j=1}^{n} Y_{j-1}\epsilon_j}{\sum_{j=1}^{n} Y_{j-1}^2}.$$

We apply this bound with $y = \frac{1}{z}$ to (3), and yield that

$$\mathbb{P}\left(|\hat{\alpha} - \alpha| > x, \sum_{j=1}^{n} Y_{j-1}^2 \geq z\right) \leq 2\exp(\frac{-x^2 z}{2}).$$

Examples of Self-Normalized Statistic

Here I list some self-normalized statistics $\frac{A}{B}$ that satisfy (7) for some λ.

Lemma

Let W_t be a standard Brownian motion. Assume that T is a stopping time such that $T < \infty$ a.s.. Then for all $\lambda \in \mathbb{R}$,

$$\mathbb{E} \exp(\lambda W_T - \lambda^2 T/2) \leq 1.$$

Lemma

Let M_t be a continuous, square-integrable martingale, with $M_0 = 0$. Then $\exp\{\lambda M_t - \lambda^2 \langle M \rangle_t / 2\}$ is a supermartingale for all $\lambda \in \mathbb{R}$, and therefore

$$\mathbb{E} \exp(\lambda M_t - \lambda^2 \langle M \rangle_t / 2) \leq 1.$$

The inequality is also valid if M_t is only assumed to be a continuous local martingale (by application of Fatou's lemma).

Lemma (de la Peña [2])

Let $\{d_i\}$ be a sequence of variables adapted to an increasing sequence of $\sigma-$fields $\{\mathcal{F}_i\}$. Assume that the d_i's are conditionally symmetric (i.e., $\mathcal{L}(d_i|\mathcal{F}_{i-1}) = \mathcal{L}(-d_i|\mathcal{F}_{i-1})$). Then $\exp(\lambda \sum_{i=1}^n d_i - \lambda^2 \sum_{i=1}^n d_i^2/2)$, $n \geq 1$, is a supermartingale with mean ≤ 1, for all $\lambda \in \mathbb{R}$.

Remark: There is no integrability assumption made in this example.

Lemma

Let $\{d_n\}$ *be a sequence of random variables adapted to an increasing sequence of* $\sigma-$*fields* $\{\mathcal{F}_n\}$ *such that* $\mathbb{E}(d_n|\mathcal{F}_{n-1}) \leq 0$ *and* $\dot{n} \leq M$ *a.s. for all* n *and some random positive constant* M. *Let* $0 < \lambda_0 \leq M^{-1}$, $A_n = \sum_{i=1}^n d_i$, $B_n^2 = (1 + \frac{1}{2}\lambda_0 M)\sum_{i=1}^n \mathbb{E}(d_i^2|\mathcal{F}_{i-1})$, $A_0 = B_0 = 0$. *then* $\{\exp(\lambda A_n - \frac{1}{2}B_n^2), \mathcal{F}_n, n \geq 0\}$ *is a supermartingale for every* $0 \leq \lambda \leq \lambda_0$.

References I

[1] V. H. de la Peña. "A bound on the moment generating function of a sum of dependent variables with an application to simple random sampling without replacement". In: *Annales de l'IHP Probabilités et statistiques*. Vol. 30. 2. 1994, pp. 197–211.

[2] V. H. de la Peña. "A general class of exponential inequalities for martingales and ratios". In: *The Annals of Probability* 27.1 (1999), pp. 537–564.

[3] V. H. de la Peña, M. J. Klass, and T. L. Lai. "Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws". In: (2004).