آنالیز احتماالاتی در ابعاد بالا

گزارش نهایی

مدرس: دکتر یاسایی

آمار بیزی

کیانا عسگری

۱ مقدمه

سال ۱۷۶۳ برای اولین بار فرمول بیز توسط توماس بیز بیان و اثبات شد. پس از آن، پایه و اساس آمار بیزی روی این فرمول بنیان شده و کار روی این موضوع همچنان ادامه دارد.

متدهای بیزی بعلت عمومیت پذیری بالایی که دارند، در بسیاری از مسایل قابل استفاده هستند. استفاده از متد های بیزی این مزیت را دارد که علاوه بر تخیمن جواب، یک بازه اطمینان نیز تولید می‌کند که به خصوص موجب افزایش علاقه به استفاده از آن ها در شبکه های عمیق شده است.

آنالیز متدهای بیزی برای ابعاد بالاتر و مدل های بدون پارامتر به یکی از موضوع های مورد علاقه ی آماردانان در سال های اخیر تبدیل شده ، همچنین از سال ۲۰۰۰ متد های بیزی برای استفاده در مدل های بی پارامتر گسترش یافتند.[۵]

ما در این گزارش در تلاشیم که ابتدا با آمار بیزی آشنا شویم و وارد مدل های بی پارامتر بشویم. در نهایت سعی میکنیم به مشکلات پیش آمده در مدل های بی پارامتر پرداخته و آن‌ها را تحلیل کنیم.

۲ استنتاج بیزی

[۵] در امار بیزی، فرض می کنیم مدل ما با مجموعه اندازه پذیر Θ پارامتریزه شده است؛ هر پارامتر $\theta \in \Theta$ یک توزیع احتمال شرطی $\Pr(.|\theta)$ تعریف میکند.

یک توزیع پیشین بیزی π [۱] یک توزیع اولیه روی مجموعه Θ است که بیانگر باورها و دانش ما در مورد توزیع پارامتر مجهول θ در مجموعه Θ است. ما به دنبال پاسخ دادن به درخواست های مربوط به توزیع Θ بعد از مشاهده نمونه های iid $y_n = (Y_1, ..., Y_n)$ هستیم. توزیع پسین بیزی[۲] با استفاده مستقیم از لایکلیهود نمونه های مشاهده شده و توزیع پسین، با استفاده از فرمول بیز محاسبه میشود:

$$\pi(\theta|y_n) = \frac{\pi(\theta)\mathcal{L}(y_n|\theta)}{\int_\Theta \pi(\nu)\mathcal{L}(y_n|)d\nu} \sim \pi(\theta)\mathcal{L}(y_n|\theta) \qquad (۱)$$

۱.۲ الگوریتم های نمونه برداری

برای بسیاری ازکاربردها، ما حداقل نیازمندیم بتوانیم از توزیع θ نمونه برداری کنیم. نمونه برداری بصورت مستقیم و با استفاده از روش های ابتدایی که تا کنون با آن ها آشنا هستیم بصورت مستقیم از فرمول ۱ ممکن نیست؛ زیرا محاسبه مخرج برای ما ممکن نیست. برای حل این مشکل، روش های زیادی معرفی شده اند که هر کدام مزایا و معایب خودشان را دارند. ما در این بخش تلاش میکنیم یکی از الگوریتم های مهمی که از ایده زنجیره مارکوف استفاده میکند به نام RWHM را بررسی کنیم.

[1]Bayesian prior distribution

[2]Bayesian posterior distributio

۱.۱.۲ MCMC

[۱] ایده الگوریتم های MCMC ساخت یک زنجیره مارکوف از نمونه های θ^k است بطوریکه هر نمونه تنها به نمونه قبلی خود وابستگی داشته باشد و دنباله درنهایت به توزیع خواسته شده میل کند.

متدهای MCMC از بهترین الگوریتم های موجود برای نمونه برداری از توزیع دقیق پسین بیزی در ابعاد پایین به شمار می‌آیند. الگوریتم RWHM یکی از ابتدایی ترین الگوریتم های نمونه برداری با ایده MCMC است که در آمار بیزی استفاده می‌شوند. در ادامه به بررسی این الگوریتم میپردازیم.

[۵] الگوریتم ۱ یک گام از الگوریتم RWHM را نشان می‌دهد. در این الگوریتم، از روی نمونه کنونی θ^k نمونه جدید $\theta^* \sim N(\theta^k, \Sigma)$ انتخاب می‌شود؛ این نمونه نزدیک به نمونه قبلی از روی توزیع گوسی انتخاب می‌شود که ماتریس کواریانس Σ ورودی الگوریتم است. سپس در صورتی که نمونه جدید θ^* نسبت به نمونه θ^k، به توزیع $\pi(\theta|y_n)$ نزدیک تر باشد این نمونه را بعنوان θ^{k+1} پذیرفته، در غیر اینصورت به گام قبلی باز میگردیم. با توجه به الگوریتم ۱ مشاهده میشود که محاسبه مخرج فرمول ۱ دیگر برای ما مشکل ساز نخواهد بود.

Algorithm 1 (Gaussian) Random Walk Hastings-Metropolis (RWHM) algorithm

---

Input: $\boldsymbol{\theta}^k$, Σ (resp. a point in $\mathbb{R}^d$, and a $d \times d$ symmetric positive matrix)
Output: $\boldsymbol{\theta}^{k+1}$ (a vector in $\mathbb{R}^d$).
1: Simulate $\boldsymbol{\theta}^\star \sim N(\boldsymbol{\theta}^k, \Sigma)$.
2: With probability $1 \wedge r$ where

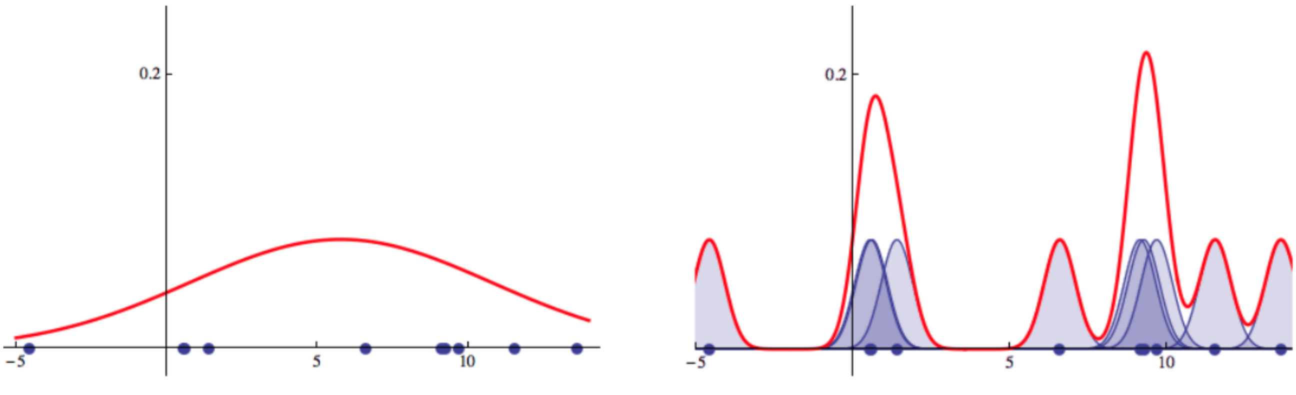
$$r = \frac{\pi(\boldsymbol{\theta}^\star)\ell(\mathbf{y_n}|\boldsymbol{\theta}^\star)}{\pi(\boldsymbol{\theta}^k)\ell(\mathbf{y_n}|\boldsymbol{\theta}^k)}$$

take $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^\star$; otherwise keep the parameter unchanged: $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k$.

وجود ویژگی های خوب الگوریم های MCMC ، این متد ها همچنان نیازمند محاسبه لایکلیهود نمونه های ورودی هستند که در ابعاد بالاتر از لحاظ محاسباتی میتواند مشکل ساز شود. بررسی متد های بیزی در مدل های با بی‌نهایت پارامتر یکی از موضعات مورد علاقه محققان آماری است. در ادامه به بررسی متد های بی‌پارامتر می‌پردازیم.

۳ بیز بی پارامتر

[۳] در مدل های پارامتریک، مجموعه Θ از بعد متناهی است. مدل هایی که در آنها Θ از بعد بی‌نهایت است، مدل های بی‌پارامتر نامیده میشوند. برای مقایسه مدل های پارامتریک با مدل های بدون پارامتر، به مثال زیر توجه کنید:



شکل ۱: تخمین توزیع؛ شکل سمت چپ از مدل با دو پارامتر و شکل سمت راست از مدل بی‌پارامتر بهره می‌برد.

فرض کنید n نمونه $x_1, ..., x_n \in \mathcal{R}$ را مشاهده کرده ایم و میخواهیم توزیع آن‌ها را تخمین بزنیم. در روش اول، داده ها را روی یک توزیع گوسی با استفاده از MLE، فیت میکنیم. در این روش درجه آزادی ۲ است.
در روش بدون پارامتر، به ازای هر نمونه $X_i = x_i$، یک توزیع گوسی به میانگین x_i و واریانس مشخص به تخمین گر اضافه میکنیم. توزیع تخمین زده شده در نهایت بصورت $p_n(x) := \frac{1}{n} g(x|x_i, \sigma)$ محاسبه میشود. درجه ازادی در این روش، برابر $n+1$ است. در بسیاری از شرایط، استفاده از روش دوم به روش اول ارجحیت دارد؛ بعنوان مثال اگر توزیع اصلی داده ها تعداد نامشخص قله داشته باشد، توزیع بدست آمده توسط روش اول فاصله زیادی از توزیع اصلی خواهد داشت و استفاده از مدل بدون پارامتر عملکرد بهتری از خود نشان خواهد داد.
استفاده از متدهای بیزی در چهارچوب مدل های بدون پارامتر، چالش های جدیدی ایجاد می‌کند[۴]:

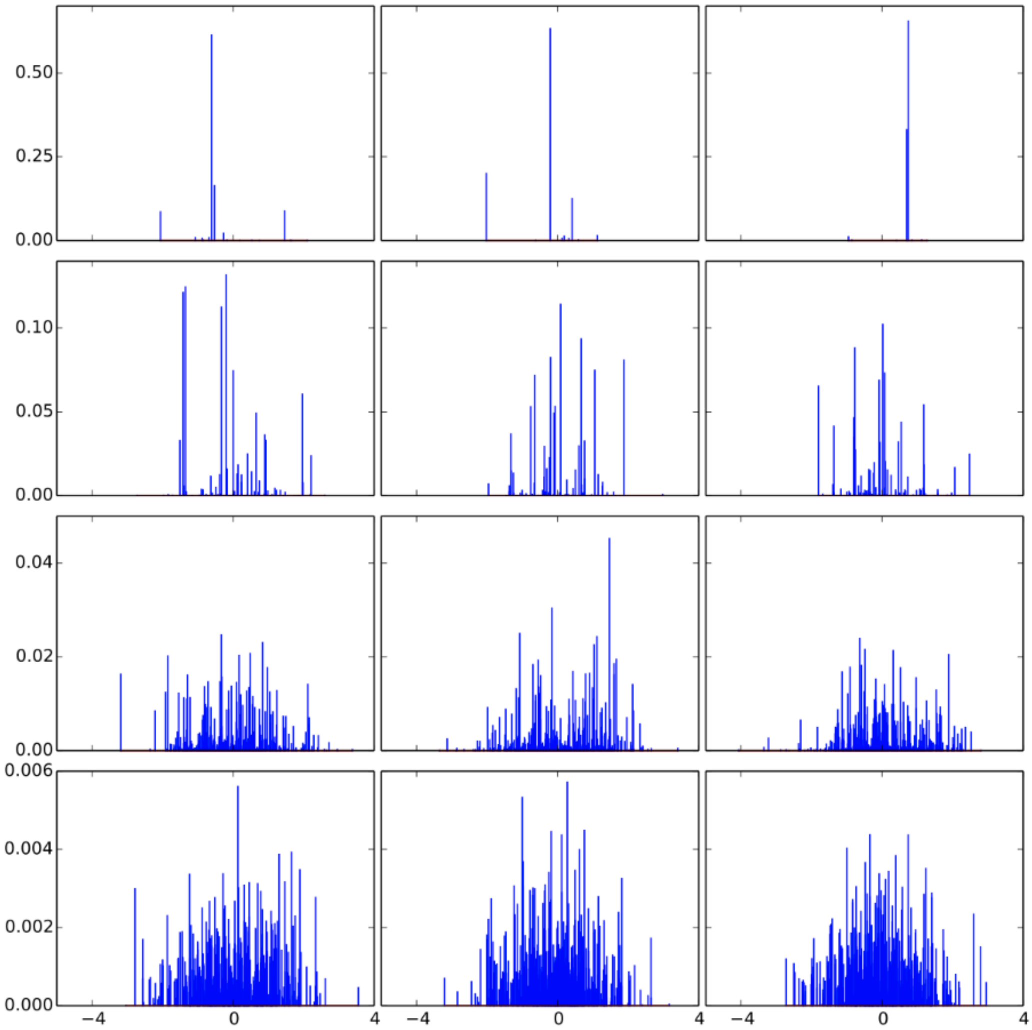
- چگونه یک توزیع پیشین مناسب روی فضای با بی‌نهایت بعد انتخاب کنیم؟
- آیا توزیع پسین بدست آمده، عملکرد مناسبی دارد؟

در ادامه سعی میکنیم به این دو سوال پاسخ دهیم.

۱.۳ تخمین cdf

فرض کنید $X_1, .., X_n$ نمونه های iid از یک cdf ناشناخته مثل F باشد. میخواهیم با استفاده از بیز بدون پارامتر، F راتخمین بزنیم. برای اینکار، توزیع پسین π را روی تمام cdf های ممکن مانند F تعریف میکنیم. برای این کار از فرآیند دیریکله کمک میگیریم.
توزیع π با دو پارامتر F_0 و α مشخص میشود که F_0یک cdf دلخواه و α یم پارامتر مثبت و بیانگر فشردگی توزیع حول f_0 است. برای تعریف توزیع پسین، نحوه نمونه برداری از آن را بیان میکنیم:

۱. نمونه های $s_1, ...$ را بصورت مستقل از F_0 بردار.
۲. نمونه های $V_1, ...$ را بصورت مستقل از $Beta(1, \alpha)$ بردار.
۳. قرار بده $w_1 = V_1, w_j = V_j \Pi_{j=i}^{j-1}(1 - V_i)$
۴. تابع F را بصورت یک توزیع گسسته با قرار دادن وزن w_i روی s_i بدست بیاور: $F = \sum_{j=1}^{\infty} w_j \delta_{s_j}$ شکل ۲ نحوه وابستگی توزیع بدست آمده به پارامتر α را نشان میدهد.



شکل ۲: چهار سطر نشان دهنده چهار مقدار α از پایین به بالا ۱ و ۱۰ و ۱۰۰ و ۱۰۰۰ است. هر سطر شامل سه نمونه از توزیع $DP(N(0,1),\alpha)$ است. توزیع های بدست آمده از طریق فرآیند گوسی، توزیع های گسسته ای هستند که با توجه به میزان پارامتر α فشرده هستند.

بعد از تعریف توزیع پیشین، سوال بعدی که پیش می‌اید نحوه نمونه برداری از توزیع حاشیه ای است. یک روش برای انجام این کار نمونه گیری $F \sim \pi$ طبق الگوریتم قبی و سپس نمونه گیری $X_1, ..., X_n$ از F است. به این الگوریتم فرآیند رستوران چینی[۳] گفته میشود:

۱. نمونه X_1 را از توزیع F_0 بردار.

۲. برای $i = 2, ..., n$ نمونه های X_i را مشابه رابطه زیر تولید کن:

$$X_i|X_1, ..., X_{i-1} = \begin{cases} X \sim F_{i-1} & \text{with probability } \frac{i-1}{i+\alpha-1} \\ X \sim F_0 & \text{with probability } \frac{\alpha}{i+\alpha-1} \end{cases}$$

که در آن F_{i-1} توزیع تجربی نمونه های $X-1, ..., X_{i-1}$ است

از نظر اسم گذاری، مساله بیان شده در ابتدا به این صورت بود که در یک رستوران چینی، وقتی nامین مشتری وارد رستوران میشود با احتمال $n_j/(n+\alpha-1)$ روی صندلی j ام مینشیند که در آن $n_j = |\{i : c_i = j\}|$ و $c_i = j$ به معنای این است که X_i مقدار مشخص X_j^* را میگیرد. آنگاه

$$X_n = \begin{cases} X_j^* & \text{with probability } \frac{n_j}{i+\alpha-1} \\ X \sim F_0 & \text{with probability } \frac{\alpha}{i+\alpha-1} \end{cases}$$

قضیه ۱ فرض کنید $X_1, ..., X_n \sim F$ و فرض کنید روی F توزیع پیشین $\pi = DP(\alpha, F_0)$ قرار داده شده باشد. آنگاه توزیع پسین بدست آمده دارای توزیع $\pi(.|X_1, ..., X_n) \sim DP(\alpha+n, \bar{F_n})$ خواهد بود که در آن

$$\bar{F_n} = \frac{n}{n+\alpha}F_n + \frac{\alpha}{n+\alpha}F_0$$

[۴]

پس نمونه برداری از توزیع پیشین مشابه نمونه برداری از توزیع پسین خواهد بود.

۲.۳ ارزیابی توزیع پسین

در این بخش، از سری‌های رندوم برای توزیع پیشین استفاده میکنیم [۲]. فرض کنید J یک متغیر تصادفی باشد که در اعداد طبیعی مقدار میگیرد. برای هر $J \in \mathbb{N}$ آرایه $\xi = (\xi_1, ..., \xi_J)$ از توابع پایه مستقل را داریم. فرض کنید π توزیع پسین تولید شده توسط سری تصادفی با توابع پایه ξ و ضرایب تصادفی θ باشد. ابتدا میخواهیم خوش رفتاری توزیع پسین را تضمین کنیم. فرض میکنیم توزیع پیشین π دو شرط زیر را ارضاع میکند:

(A۱)

$$\exists c_1, c_2 > 0, 0 \le t_2 \le t_1 \le 1 : \ \exp\left(-c_1 j \log^{t_1} j\right) \le \pi(J=j) \le \exp\left(-c_2 j \log^{t_2} j\right)$$

(A۲) با داشتن J، برای توزیع π روی بردار $\theta = (\theta_1, ..., \theta_J)$ داشته باشیم

$$\exists c_3 > 0, \forall \|\theta_0\| \le H : \pi(\|\theta - \theta_0\| \le \epsilon) \ge \exp\left(-c_3 J \log 1/\epsilon\right)$$

که در آن H به اندازه کافی بزرگ و ϵ به اندازه کافی کوچک انتخاب شده است. همچنین فرض میکنیم

$$\exists C, t_3 > 0 : \pi(\theta \notin [-M, M]^J) \le J\exp\left(-Cm^{t_3}\right)$$

قضیه ۲ فرض کنید $\epsilon_n \ge \bar{\epsilon_n}$ دو دنباله از اعداد مثبت باشند بطوریکه $\epsilon_n \to \infty$ و $n\bar{\epsilon_n} \to \infty$ فرض کنید برای تخیمن تابع w_0، توزیع پیشین در نظر گرفته شده شرایط *(A1) (A2)* را ارضا کند. فرض کنید دنباله های $J_n, \bar{J}_n, M_n$ از اعداد مثبت، و تابع $e(.)$ اکیدا نزولی و نامنفی وجود داشته باشند. همچنین فرض کنید $\forall j \in N \exists \theta_{0,j} \in R^j$ بطوریکه روابط زیر برای ثابت های $a_1 > 0, a-2, C_0, H$ برقرار باشد:

[3] Chinese Restaurant Process

[۴] من ثبات از این قضیه که در منبع ۴ معرفی شده نتونستم پیدا کنم و ایده خودم برای اثبات این بود که با نوشتنرابطه نمونه برداری یکسان میشن منتها چون خیلی ایده خامی بود ترجیح دادم اثبات براش نیارم چون در این بخش کلا تلاشم فقط شهود گرفتن به نمونه برداری در ابعاد بالا و مشکلاتش بود

$$\|\boldsymbol{\theta}_{0,j}\|_\infty \leq H, \quad d(w_0, \boldsymbol{\theta}_{0,j}^T \boldsymbol{\xi}) \leq e(j), \tag{2.2}$$

$$J_n \{\log J_n + \log a(J_n) + \log M_n + C_0 \log n\} \leq n\epsilon_n^2, \tag{2.3}$$

$$e(\bar{J}_n) \leq \bar{\epsilon}_n, \quad c_1 \bar{J}_n \log^{t_1} \bar{J}_n + c_3 \bar{J}_n \log(2a(\bar{J}_n)/\bar{\epsilon}_n) \leq a_2 n\bar{\epsilon}_n^2, \tag{2.4}$$

$$n\bar{\epsilon}_n^2 \leq C J_n \log^{t_2} J_n \text{ for any constant } C, \quad J_n \exp\{-CM_n^{t_3}\} \leq (a_1 - 1)\exp\{-n\bar{\epsilon}_n^2\}. \tag{2.5}$$

فرض کنید $W_{J_n,M_n} = \{w = \theta^T \xi : \theta \int R^j, j \leq J_n, \|\theta\|_\infty \leq M_n\}$ آنگاه روابط زیر برقرار هستند:

$$\log D(n^{-C_0}, \mathcal{W}_{J_n,M_n}, d) \leq n\epsilon_n^2, \tag{2.6}$$

$$\Pi(W \notin \mathcal{W}_{J_n,M_n}) \leq a_1 \exp\{-bn\bar{\epsilon}_n^2\}, \tag{2.7}$$

$$-\log \Pi\{w = \boldsymbol{\theta}^T \boldsymbol{\xi} : d(w_0, w) \leq \bar{\epsilon}_n\} \leq a_2 n\bar{\epsilon}_n^2. \tag{2.8}$$

برهان. ابتدا رابطه ۶.۲ را با استفاده از عدد پوششی اثبات میکنیم. با استفاده از رابطه ۱.۲ داریم:

$$\begin{aligned}
&\log D(n^{-C_0}, \mathcal{W}_{J_n,M_n}, d) \\
&\leq \log \left\{ \sum_{j=1}^{J_n} D(n^{-C_0}/a(j), \{\boldsymbol{\theta} \in \mathbb{R}^j, \|\boldsymbol{\theta}\|_\infty \leq M_n\}, \|\cdot\|_2) \right\} \\
&\leq \log \left[J_n \left\{ \sqrt{J_n} M_n a(J_n) n^{C_0} \right\}^{J_n} \right] \\
&\leq J_n(\log J_n + \log M_n + \log a(J_n) + C_0 \log n) \leq n\epsilon_n^2.
\end{aligned}$$

سپس به اثبات رابطه ۷.۲ میپردازیم. دقت کنید برای $c_2' > 0$

$$\pi(w \notin W_{J_n,M_n}) \leq \pi(J > J_n) + \sum_{j=1}^{J_n} \pi(\theta \notin [-M_n, M_n]^j)\pi(J = j)$$

$$\leq exp(-c_2' J_n \log^{t_2} J_n) + J_n exp(-CM_n^{t_3})$$

$$\leq a_1 exp(-n\bar{\epsilon}_n^2)$$

برای اثبات رابطه ۸.۲، از ۲.۲ استفاده میکنیم. چون $d(w_0, \theta_{0,j}^T \xi) \leq e(j) \leq \bar{\epsilon_n}$ آنگاه

$$\pi(w : d(w_0, \theta^T \xi) \leq 2\bar{\epsilon_n}) \geq \pi(J = \bar{J}_n)\pi(\|\theta - \theta_0\| \leq \bar{\epsilon_n}/a(\bar{J}_n))$$

$$\geq exp(-c_1 \bar{J}_n \log^{t_1} \bar{J}_n) exp(-c_3 \bar{J}_n \log(\frac{a(\bar{(J_n)})}{\epsilon_n}))$$

با لگاریتم منفی گرفتن از طرفین به رابطه خواسته شده میرسیم و حکم اثبات میشود. ■

میتوانیم قضیه ۲ را ساده تر کرده بطوریکه نتیجه آن برای توزیع پسین قابل استفاده باشد:

قضیه ۳ فرض کنید مشاهده های X_i بصورت iid مشاهده کرده ایم. فرض کنید w_0 مقدار واقعی تخمینگر، یک تابع پیوسته با مشتقات کراندار تا مرتبه α_0 باشد بطوریکه مشتق α_0 آن در شرط هولدر صدق کند:

$$|w_0^{\alpha_0}(x) - w_0^{\alpha_0}(y)| \leq C|x-y|^{\alpha-\alpha_0}$$

فرض کنید $r = 2, \infty$
$\epsilon_n \geq \bar{\epsilon_n}$ دو دنباله از اعداد مثبت باشند بطوریکه $\epsilon_n \to \infty$. همچنین فرض کنید $\exists \theta_0 \in R^j, \|\theta_0\|_\infty \leq H$ بطوریکه
فرض کنید توزیع پیشین در نظر گرفته شده شرایط *(A1) (A2)* را ارضا کند. فرض کنید دنباله های $J_n, \bar{J}_n, M_n$ از اعداد

$$\|w_0 - \boldsymbol{\theta}_0^T \boldsymbol{\xi}\|_r \leq C_1 J^{-\alpha}, \tag{2.12}$$

$$\|\boldsymbol{\theta}_1^T \boldsymbol{\xi} - \boldsymbol{\theta}_2^T \boldsymbol{\xi}\|_r \leq C_2 J^{K_0} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2, \ \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^J. \tag{2.13}$$

مثبت وجود داشته باشند بطوریکه شرایط زیر برقرار باشند.

$$bn\bar{\epsilon}_n^2 \leq J_n \log^{t_2} J_n, \quad \log J_n + n\bar{\epsilon}_n^2 \leq M_n^{t_3}, \tag{2.14}$$

$$J_n\{(K_0+1)\log J_n + \log M_n + C_0 \log n\} \leq n\epsilon_n^2, \tag{2.15}$$

$$\bar{J}_n^{-\alpha} \leq \bar{\epsilon}_n, \quad \bar{J}_n\{c_1 \log^{t_1} \bar{J}_n + c_3 K_0 \log(\bar{J}_n) + c_3 \log(1/\bar{\epsilon}_n)\} \leq 2n\bar{\epsilon}_n^2, \tag{2.16}$$

$$\rho_n(w_1, w_2) \lesssim n^{a_3} \|w_1 - w_2\|_r^{a_4} \ \textit{for any}\ w_1, w_2 \in \mathcal{W}_{J_n, M_n}, \tag{2.17}$$

$$\max\left\{n^{-1}\sum_{i=1}^n K(p_{i,w_0}, p_{i,w}), n^{-1}\sum_{i=1}^n V(p_{i,w_0}, p_{i,w})\right\} \leq C\|w_1 - w_2\|_r^2, \tag{2.18}$$

آنگاه توزیع پسیین حول w_0 فشرده میشود.

اثبات این قضیه با استفاده از قضیه ۲ و دقت به این نکته که ϵ_n شعاع همسایگی θ_0 برای همگرایی است نتیجه میشود.
این قضیه به ما میگوید به منظور بدست آوردن فشردگی در توزیع پسین، انتخاب دنباله های $J_n, \bar{J}_n, \epsilon_n, M_n$ بطوریکه در شرایط قضیه صدق کنند و همچنین این که فاصله KL با نرم اقلیدسی $\|.\|_r$ کران بخورد شرایط مهم و ضروری هستند. نحوه انتخاب دنباله های مدنظر در اکثر چهارچوب ها کاری ساده و مشابه است اما حفظ شرط کرانداری KL، بین مسآله‌های مختلف به شدت متفاوت و در بسیاری از موارد کاری دشوار است.

۴ خلاصه

در این گزارش دیدیم که در چهارچوب های ساده‌تر مدل های ابعاد پایین، میتوان از متد های بیزی برای حل مساعل متعدد بدون مواجه به فشار محساباتی زیاد،استفاده کرد. همچنین با الگوریتم نمونه برداری RWHM که از ایده زنجیره های مارکوف استفاده میکردند آشنا شدیم. اخیرا متد های بیزی در شبکه های عصبی عمیق نیز به شدت مورد استقبال قرار گرفته اند. برای اطلاعات بیشتر در این حوزه به [۱] مراجعه کنید.

قضیه کلی ۳ نشان داد که در چهارچوب کلی استفاده از سری های تصادفی در مدل های بدون پارامتر، انتظار عملکرد مناسبی از توزیع پسین داریم. اراعه الگوریتم های نمونه برداری در مدل های بدون پارامتر همچنان موضوع مورد بحث است.

مراجع

[۱] LAURENT VALENTIN JOSPIN: Hands-on Bayesian Neural Networks - a Tutorial for Deep Learning Users.

[۲] Weining Shen: Adaptive Bayesian procedures using random series priors.

[۳] Peter Orbanz: Lecture Notes on Bayesian Nonparametrics.

[۴] Han Liu, and Larry Wasserman: Nonparametric Bayesian Methods.

[۵] Nicolas Chopin, ON SOME RECENT ADVANCES ON HIGH DIMENSIONAL BAYESIAN STATISTICS.