هندسه اطلاعات و کاربردها

نظریه اطلاعات، آمار و یادگیری

تاریخ: ۱۹ تیر ۱۴۰۳

استاد: دکتر یاسایی

دانشگاه: صنعتی شریف

دانشکده: مهندسی برق

اعضای گروه:

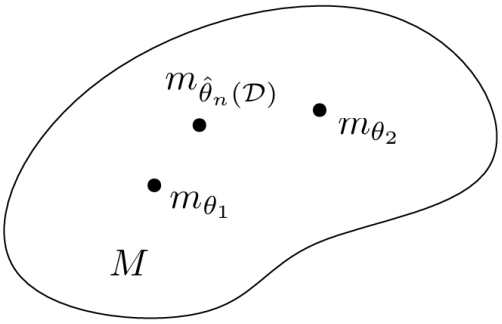
سپهر حیدری ادواری 400109854

برنا خدابنده 400109898

فهرست مطالب

۱ مقدمه

هنگام بررسی توزیع‌های احتمال پارامتری این سوال را می‌توانیم از خود بپرسیم که چه می‌شود اگر این توزیع ها را برحسب پارامتر دیگری توصیف کنیم؟ همواره می‌توان تغییر متغیرهایی داد و پارامتر یک مجموعه از توزیع‌های پارامتری را به پارامتر دیگری تغییر داد. اما خواص بنیادی این توزیع‌ها طبیعتا به پارامتری که ما برای برچسب زدن به توزیع ها استفاده می‌کنیم ربطی ندارد و باید تحت تغییر پارامتر به اصطلاح ناوردا باشند. زمانی که با پارامتر گسسته سر و کار داریم این امر آنچنان نابدیهی نیست و در صورتی که یک تابع یک به یک روی اندیس هایی که نقش پارامتر توزیع را بازی میکنند ایجاد کنیم مجددا به یک مجموعه گسسته از اندیس ها با همان تعداد میرسیم که نقش پارامترهای جدید را بازی می‌کنند و حالا توزیع‌ها را می‌توان برحسب آن‌ها برچسب زد. اما زمانی شرایط جالب تر می‌شود که با پارامترهای پیوسته از جمله اعداد حقیقی یا بردارهای حقیقی(یا به صورت معادل، چندین عدد حقیقی) و غیره سر و کار داریم.



شکل ۱: یک مجموعه پارامتری از توزیع‌ها

برای اینکه مثالی از این موضوع بزنیم می‌توانیم به مجموعه زیر اشاره کنیم:

$$\mathcal{P} = \left\{ p_\lambda(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right),\ \lambda = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_{++} \right\} \qquad (۱)$$

که نشان دهنده مجموعه توزیع‌های گوسی تک متغیره است. در اینجا λ پارامتر توزیع است که شامل میانگین و انحراف معیار توزیع گوسی است. اما به صورت معادل این مجموعه را به صورت‌های دیگر(با پارامترهای دیگر) نیز می‌توان نمایش داد. برای مثال داریم:

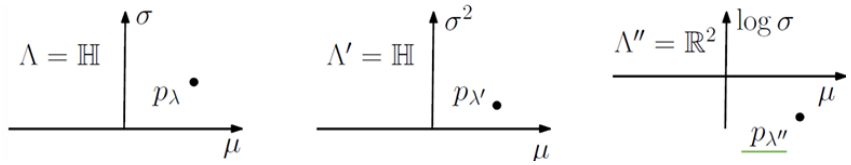
$$\mathcal{P} = \left\{ p_{\lambda'}(x) = \frac{1}{\sqrt{2\pi\lambda_2'}} \exp\left(-\frac{(x-\lambda_1')^2}{2\lambda_2'} \right), \lambda' = (\mu, \sigma^2) \in \mathbb{R}^2 \right\} \qquad (۲)$$

و همینطور:

$$\mathcal{P} = \left\{ p_{\lambda''}(x) = \frac{1}{\sqrt{2\pi} e^{\lambda_2''}} \exp\left(-\frac{(x-\lambda_1'')^2}{2e^{2\lambda_2''}} \right), \lambda'' = (\mu, \log(\sigma)) \in \mathbb{R}^2 \right\} \qquad (۳)$$

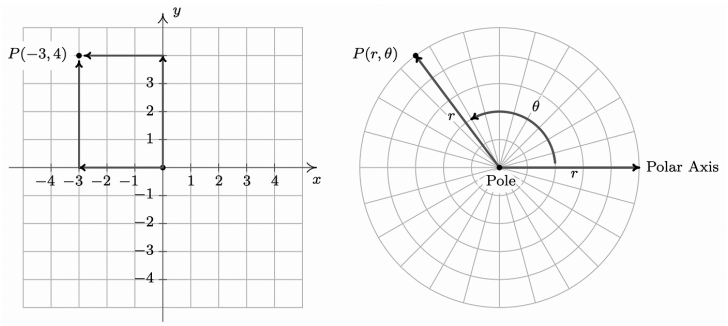
که اگر با پارامترهای گفته شده بخواهیم این مجموعه را نمایش دهیم به صورت زیر می‌باشد:

همانطور که از شکل‌ها مشخص است برای پارامتربندی‌های متقاوت به شکل‌های متفاوت می‌رسیم.



شکل ۲: نمایش مجموعه یکسان با پارامترهای متفاوت(در اینجا $\mathbb{H} = \mathbb{R} \times \mathbb{R}_{++}$)

حالا اگر به هندسه خمینه‌ها(منیفلدها) نگاه کنیم می‌بینیم که با اتفاق مشابهی روبه‌رو هستیم. یک منیفلد خودش یک موجود مجرد است که برای تحلیلی کردن و کار کردن با آن از نگاشت‌هایی از منیفلد به اعداد حقیقی استفاده می‌کنیم و عملا آن را پارامتریزه می‌کنیم تا بتوانیم با پارامترهای آن که اعداد هستند به راحتی کار کنیم. در اینجا نیز به طرق مختلفی می‌توان پارامتریزه کردن را انجام داد و این‌ها ماهیت خود منیفلد را تغییر نمی‌دهند. برای یک مثال ساده می‌توان به منیفلد $\mathbb{R}^2$ نگاه کرد:



شکل ۳: دو پارامتربندی مختلف از یک خمینه

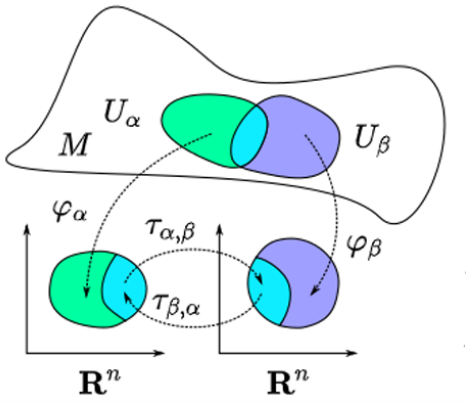
در شکل‌های بالا دو پارامتربندی مختلف از یک خمینه را می‌بینیم. یک بار خمینه $\mathbb{R}^2$ را با مختصات دکارتی پارامتریزه یا مخصته بندی می‌کنیم و یک بار با مختصات قطبی. البته اگر بخواهیم دقیق باشیم استفاده از مختصات قطبی برای پارامتریزه کردن کل منیفلد کمی ایراد دارد از آنجایی که برای هر نقطه که زاویه θ داشته باشد می‌توان $\theta + 2k\pi$ را نیز به آن نقطه نسبت داد که این با تعریف منیفلد که در قسمت بعدی ارایه می‌کنیم مطابقت ندارد. یک راهکاری که برای حل این مشکل مخصات قطبی به نظر می‌رسد مفید باشد این است که بازه زاویه را $[0, 2\pi)$ در نظر بگیریم اما این هم به علت ناپیوستگی‌ای که ایجاد می‌شود مناسب نیست، همچنین در مبدا همچنان با این مشکل مواجه هستیم چون هر زاویه‌ای می‌توانیم به آن نسبت دهیم.

۲ پیشنیازها

حالا کمی به پیشنیازهای هندسه دیفرانسیل(هندسه خمینه‌ها یا منیفلدها) و هندسه ریمانی(منیفلدهایی که به تانسور متریک مجهز هستند) می‌پردازیم. اولین چیز تعریف خود منیفلد است. شاید تعریف دقیق ریاضی و مجرد آن خیلی برای هدف ما لازم نباشد، با این وجود اشاره کوتاهی به آن می‌کنیم.

۲.آ منیفلد چیست؟

در این بخش تعریف منیفلد را میاوریم و روی خاصیت‌هایی از آن که برای درک کلی مفهوم واجب هستند تاکید بیشتری می‌کنیم.



شکل ۴: تصویر شماتیک از یک منیفلد با دو تا از چارت‌هایش

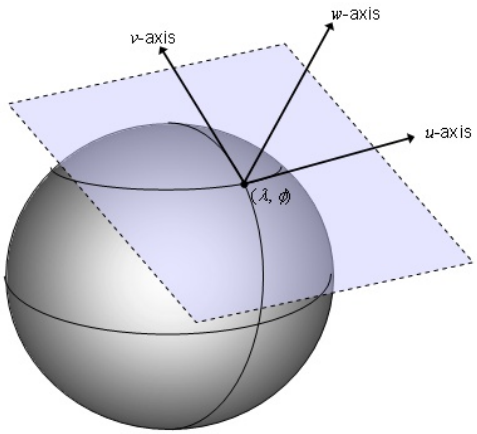
منیفلد یک فضای توپولوژیک مجهز به یک توپولوژی هاسدورف است که توسط تعدادی مجموعه باز(اعضای توپولوژی) پوشانده می‌شود که به این مجموعه‌های باز که منیفلد را می‌پوشانند چارت گفته می‌شود. هر کدام از این چارت‌ها با زیرمجموعه بازی از $\mathbb{R}^N$ (که در اینجا به N بعد منیفلد گفته می‌شود) همسان ریخت(Homeomorphic) است. در اشتراک بین چارت‌ها باید دیفیومورفیزم(Diffeomorphism) بین آنها برقرار باشد.

این تعریف به صورت شهودی به این معنی است که منیفلد موجودی است که به صورت موضعی با تعداد ثابتی مختصه که اعداد حقیقی هستند توصیف می‌شود. در نواحی‌ای که چارت‌ها با هم اشتراک دارند باید تبدیل از یک مختصات به مختصات دیگر به صورت وارون پذیر، پیوسته و مشتق‌پذیر انجام شود. با رعایت کردن این جزییات متوجه ایرادهایی که در قسمت قبل از مختصات قطبی گرفتیم می‌شویم. برای مثال اگر منیفلد را یک حلقه یا S^1 در نظر بگیریم متوجه می‌شویم که اگر این حلقه را با یک چارت بخواهیم بپوشانیم دچار مشکل می‌شویم چون که یا به یک نقطه بیش از یک مختصه نسبت می‌دهیم یا ناپیوستگی ایجاد می‌کنیم. و برای پوشاندن این حلقه به دست کم دو چارت نیاز داریم.

در عمل برای هندسه اطلاعات این موضوع که به بیشتر از یک چارت برای پوشاندن منیفلد نیاز داشته باشیم کم پیش میاید و در اکثر مواقع می‌توانیم فرض کنیم که برای منیفلدهایی که ما با آن‌ها کار می‌کنیم می‌توانیم کل آن را با یک چارت بپوشانیم.

۲.ب چگونه بردار و هم‌بردار تعریف کنیم؟

موجود ریاضی دیگری که باید تعریف کنیم بردارها و هم‌بردارها هستند. برای اینکار میتوانیم به صورت شهودی تصور کنیم که منیفلد در یک فضای $\mathbb{R}^M$ که در آن M بزرگتر از بعد منیفلد است نشانده(Embede) شده است. حالا ما برای تعریف بردار میتوانیم بردارهای مماس بر هر نقطه منیفلد را در نظر بگیریم. یعنی به عبارتی در هر نقطه از منیفلد یک فضای برداری جداگانه تعریف کنیم که شامل بردارهای مماس از فضای بزرگتر در آن نقطه است.



شکل ۵: تعریف شهودی فضای مماس

اما برای این موضوع نیز تعریف دقیق تری وجود دارد که بدون نشاندن منیفلد در فضای بزرگتر ممکن است و به جهاتی بسیار کاربردی‌تر هم هست. به دو روش بردارها یا فضای مماس را تعریف می‌کنند. این دو روش با هم معادل‌اند.

روش اول: در نظر گرفتن خم‌هایی که از نقطه مدنظر منیفلد عبور می‌کنند. آن دسته از خم‌هایی که پس از اعمال نگاشت نقشه(نگاشت از منیفلد به فضای پارامترها) شیب یکسانی در فضای پارامترها دارند را در نظر میگیریم. به کلاس هم‌ارزی این خم‌ها می‌گوییم یک بردار. برای اینکه این موجودهای ساخته شده واقعا فضای برداری تشکیل دهند باید جمع بردارها و ضرب عدد در بردار را تعریف کنیم. یک تعریف طبیعی برای این عمل‌ها می‌توان انجام این عمل‌ها روی شیب‌ها در فضای پارامترها باشد. سپس باید نشان دهیم که این جمع بردارها و ضرب عدد در بردار که تعریف کرده‌ایم خوش تعریف است یعنی به انتخاب نماینده از کلاس هم‌ارزی بستگی ندارد. پس از آن باید اثبات کنیم که این اعمال تعریف شده خواص فضای برداری را برقرار می‌کنند. چک کردن این امرها کار ساده‌ای است که به راحتی می‌توان انجام داد.

روش دوم: تعریف بردارها به صورت عملگرهای دیفرانسیلی مرتبه اول که روی توابع مشتق پذیر از منیفلد به اعداد حقیقی اثر می‌کنند. منظور از مرتبه اول بودن این است که علاوه بر خطی بودن باید خاصیت لایبنیتزی داشته باشند. یعنی اگر منیفلد $\mathcal{M}$ را داشته باشیم داریم:

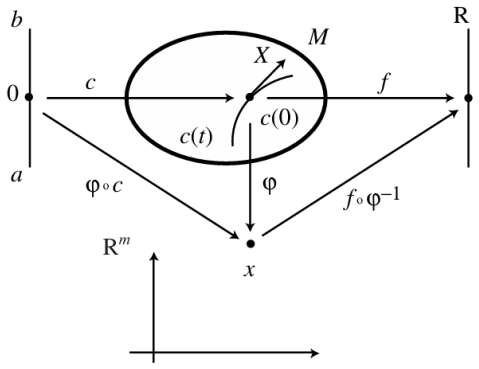
$$f, g : \mathcal{M} \longrightarrow \mathbb{R}, \quad a, b \in \mathbb{R}, \quad X \in T_P\mathcal{M}, \tag{۴}$$

$$\Longrightarrow X(a\ f + b\ g) = a\ X(f) + b\ X(g), \quad X(fg) = X(f)g(P) + f(P)X(g) \tag{۵}$$

این تعریف در اکثر مواقع کاربردی تر است. همچنین فضای دوگان که آن را با $T_P^*\mathcal{M}$ نشان می‌دهیم به صورت فضای دوگان $T_P\mathcal{M}$ تعریف می‌شود.

همچنین می‌توانیم میدان‌های برداری و هم‌برداری تعریف کنیم. برای میدان‌های برداری که آن‌ها را با $\mathfrak{X}$ نشان می‌دهیم دقیقا مانند خود بردارها تعریف می‌شوند با این تفاوت که به ازای هر نقطه یک بردار داریم بنابراین باید در این دو خاصیت صدق کند:

$$f, g : \mathcal{M} \longrightarrow \mathbb{R}, \quad a, b \in \mathbb{R}, \quad X \in \mathfrak{X}, \tag{۶}$$



شکل ۶: تعریف اول از بردارها

$$\Longrightarrow X(a\ f + b\ g) = a\ X(f) + b\ X(g), \quad X(fg) = X(f)\ g + f\ X(g) \tag{۷}$$

میدان‌های هم‌برداری نیز به صورت مشابه با اثر کردن روی میدان‌های هم برداری و بدست آوردن یک میدان اسکالر تعریف می‌شوند.

حالا که فضاهای برداری و هم‌برداری را تعریف کردیم می‌توانیم با ضرب تانسوری آن‌ها فضاهای حاصل ضربی را نیز بسازیم:

$$T^*_P\mathcal{M} \otimes T_P\mathcal{M} \otimes T_P\mathcal{M} \otimes T^*_P\mathcal{M} \otimes T_P\mathcal{M} \otimes \dots$$

همچنین برای میدان‌های تانسوری به صورت مشابه می‌توانیم عمل کنیم.

۲.ج هندسه ریمانی

حالا با معرفی تانسور متریک، سرفصل جدیدی از هندسه به نام هندسه ریمانی را معرفی می‌کنیم. تانسور متریک تعریف ضرب داخلی بردارها است. به این معنی که ضرب داخلی دو بردار $X, Y \in T_P\mathcal{M}$ را با صورت زیر تعریف میکنیم:

$$\langle X, Y\rangle := g(X, Y) \tag{۸}$$

که در آن $g \in T^*_P\mathcal{M} \otimes T^*_P\mathcal{M}$ تانسور متریک است. از آنجایی که این تانسور نشان دهنده ضرب داخلی است یک تانسور متقارن است و اگر به صور ت ماتریسی به آن نگاه کنیم یک ماتریس مثبت معین است. از تانسور متریک می‌توان برای تعریف فاصله بین نقاط نزدیک به هم استفاده کرد طوری که اگر دو نقطه اختلاف مختصه‌هایشان به صورت $d\theta^i$ باشد آن وقت فاصله آن دو نقطه به توان دو به صورت زیر می‌باشد:

$$ds^2 = g_{ij} d\theta^i d\theta^j \tag{۹}$$

که در اینجا از نمادگذاری جمع انیشتین استفاده شده یعنی روی اندیس‌های تکراری جمع زده می‌شود.
حالا که فضاهای ضرب تانسوری و میدان‌های تانسوری را معرفی کردیم می‌توانیم فرم‌های دیفرانسیلی را معرفی کنیم. فرم‌های دیفرانسیلی خود یک بحث بسیار مفصل در هندسه دیفرانسیل هست که توضیح آن‌ها در اینجا شاید خیلی مناسب نباشد. تنها چیزی که به آن نیز داریم تا در یکی از کاربردهای هندسه اطلاعات از آن استفاده کنیم فرم حجم ناوردا است که در اینجا رابطه‌اش را می‌آوریم:

$$\omega = \sqrt{\det g}\ d\theta^1 \wedge d\theta^2 \wedge \dots \wedge d\theta^N \tag{۱۰}$$

این فرم دیفرانسیلی نقش عنصر حجم را هنگام انتگرال گیری روی منیفلد بازی می‌کند. ویژگی خاص آن این است که تحت تغییر مختصات ناوردا است. که این ویژگی از همان عبارت $\sqrt{\det g}$ بدست میاید. اگر بخواهیم به صورت خیلی ساده به آن فکر کنیم میتوانیم به تغییر متغیر عادی در انتگرال‌ها و دترمینان ژاکوبی فکر کنیم. عبارت رادیکال دترمینان متریک نقش یک بر روی دترمینان ژاکوبی را بازی میکند. برای همین است که این عنصر حجم ناوردا است.

موضوع دیگری که در هندسه ریمانی مطرح می‌شود Affine Connection و مشتق هموردا(Covariant Derivative) است. که یک جور تعمیم مشتق جهتدار است. البته تعمیم‌های دیگری نیز مانند مشتق لی وجود دارند که تفاوت‌های عمیقی با مشتق هموردا دارد. به صورت کلی مشتق هموردا را به صورت زیر نمایش می‌دهیم:

$$\nabla_X Y \tag{۱۱}$$

که مشتق نسبت به Y است در جهت X ویژگی های این مشتق به صورت زیر اند:

$$\nabla_X(Y_1+Y_2)=\nabla_X Y_1+\nabla_X Y_2,\quad \nabla_X(fY)=X(f)Y+f\nabla_X Y \tag{۱۲}$$

$$\nabla_{X_1+X_2}Y=\nabla_{X_1}Y+\nabla_{X_2}Y,\quad \nabla_{fX}Y=f\nabla_X Y \tag{۱۳}$$

از روی این مشتق دو مفهوم دیگر که اهمیت زیادی دارند ظاهر می‌شوند. مفاهیم منحنی‌های ژئودزی و انتقال موازی. به صورت خیلی ساده منحنی های ژئودزی تعمیم خط راست به فضاهای خمیده هستند. Optimal Transport یکی از موضوع‌های مهمی است که از طریق مفهوم ژئودزی به هندسه اطلاعات ارتباط پیدا میکند. از آنجایی که منحنی های ژئودزی تحت شرایطی منحنی های با کمترین طول بین دو نقطه از منیفلد هستند. به صورت ساده اگر به ژئودزی فکر کنم می توانیم معادله حاکم بر آن را بدست بیاوری. از آنجایی که قرار است تعمیمی از خط راست باشد داریم:

$$\nabla_{\dot\gamma(t)}\dot\gamma(t)\propto\dot\gamma(t) \tag{۱۴}$$

ضریب تناسب به پارامتر خم یعنی t بستگی دارد. اگر پارامتر خم را متناسب با طول یا به اصطلاح پارامتر آفین انتخاب کنیم ضریب تناسب صفر می‌شود و داریم:

$$\nabla_{\dot\gamma(t)}\dot\gamma(t)=0 \tag{۱۵}$$

اگر معادله بالا را برحسب مولفه‌ها باز کنیم داریم:

$$\ddot\theta^i+\Gamma^i{}_{jk}\dot\theta^i\dot\theta^j=0 \tag{۱۶}$$

که در آن $\Gamma^i{}_{jk}$ ضرایب کریستوفل هستند که به صورت زیر تعریف می‌شوند:

$$\Gamma^i{}_{jk}=d\theta^i(\nabla_{\partial_j}\partial_k) \tag{۱۷}$$

در اینجا و در قبل تر از نمادهای ∂_i و $d\theta^i$ به ترتیب برای بردارهای پایه وابسته به مختصات و بردارهای پایه فضای دوگان وابسته به مختصات استفاده کرده‌ایم(البته این نمادگذاری گاهی باعث اشتباه می‌شود از آنجایی که بعضی اوقات به اختلاف کوچک در پارامترها نیز $d\theta^i$ نسبت می‌دهیم. مانند رابطه‌ای که برای فاصله نقاط نزدیک به هم نوشتیم$ds^2=g_{ij}d\theta^i d\theta^j$ در اینجا $d\theta^i$ به معنی هم‌بردار نیست.). در حالت کلی هر ∇ ای که ویژگی‌های گفته شده را برقرار کند یک Affine Connection معتبر است. اما یک Affine Connection خاص که از روی متریک بدست میاید وجود دارد که به آن Levi-Civita Connection گفته می‌شود. ویژگی آن این است که متریک را ثابت نگه میدارد و مشتق همووردای متریک طبق این کانکشن صفر می‌شود $\nabla_X g=0$، که به راحتی میتوان رابطه‌ای برای ضرایب کریستوفل این کانکشن بر حسب مشتقات متریک حساب کرد.

موضوع دیگر انتقال موازی است. انتقال موازی یک بردار به معنی حرکت یک بردار اولیه از یک فضای مماس اولیه

روی یک خمی از منیفلد به نقطه نهایی است طوری که تاجای ممکن بردار دست نخورد. این را میتوان اینطوری فرمول بندی کرد که مشتق هموردای بردار در راستای خم صفر باشد. ویژگی دیگر کانکشن لوی چویتا این است که اگر دو بردار را با این کانکشن انتقال موازی دهیم ضرب داخلی دوبردار نهایی با ضرب داخلی دو بردار اولیه برابر است. به عبارتی اگر انتقال موازی تحت خم $c(t)$ و کانکشن آفین $\prod_{c(t)}^{\nabla}$ نشان دهیم داریم:

$$\left\langle \prod_{c(t)}^{\nabla} X, \prod_{c(t)}^{\nabla} Y \right\rangle = \langle X, Y \rangle \tag{۱۸}$$

۲.د ساختارهای هندسی دوگان

در قسمت قبل با کانکشن لوی چویتا آشنا شدیم که ضرب داخلی را تحت انتقال موازی حفظ میکرد. حالا اگر کانکشنی غیر از لوی چویتا داشته باشیم میتوانیم یکه کانکشن دیگر که مزدوج آن باشد به این صورت تعریف کنیم:

$$\left\langle \prod_{c(t)}^{\nabla^*} X, \prod_{c(t)}^{\nabla} Y \right\rangle = \langle X, Y \rangle \tag{۱۹}$$

که در آن ∇^* کانکشن مزدوج با کانکشن اولیه ∇ است. به راحتی می‌توان دید که اگر میانگین این دو کانکشن را در نظر بگیریم به همان کانکشن لوی چویتا میرسیم. این ساختار که کمی از اصل هندسه ریمانی فاصله دارد در هندسه اطلاعات بسار مفید واقع می‌شود.
تا اینجا صحبتی از انحنای منیفلد نکردیم. برای محاسبه انحنا نیاز به تعریف تانسور ریمان داریم که در اینجا نمی گنجد. از روی تانسور ریمان می‌توان معیار مختلف کوچکتری برای نشان دادن انحنا پیدا کرد. قضیه ای به نام قضیه اساسی هندسه اطلاعات وجود دارد که بیان می‌کند که اگر یک منیفلد طبق کانکشن ∇ تخت یا بدون انحنا باشد طبق کانکشن مزدوج آن یعنی ∇^* نیز تخت خواهد بود. در واقع قضیه کلی تر از این است و بیان میکند که اگر انحنای ثابت α داشته باشیم آنگاه با کانکشن مزدوج نیز همان انحنای ثابت را در منیفلد داریم، اما ما با همان حالت تخت سر و کار خواهیم داشت چون که همانطور که جلوتر می‌بینیم برای دو خانواده مهم توزیع ها یعنی خانواده نمایی و میکسچر ها این ویژگی برقرار است.

۳ انحراف ها

یک مفهوم که بار ها داخل درس و در کل در مقایسه ها دیده ایم، مفهوم انحراف بوده، این انحراف حالا میتواند بین دو توزیع باشد، مانند انحراف کولبک–لیبر(KL Divergence)، فاصله TV یا باقی معیار ها مانند f–انحراف ها که در درس بار ها با آن ها برخورد کردیم.

حال در هندسه، همانطور که گفتیم حداقل در هندسه ریمانی، نیاز به تعریف هندسه داریم به صورت $(\mathcal{M}, g)$، که در تمامی این هندسه ها معیار فاصله یک معیار مهم است، که g یا متریک بر آن دلالت دارد.

با الهام از این موضوع، ما در قدم اول برای تعریف یک هندسه، به تعریف فاصله یا به طور کلی تر به تعریف یک انحراف میپردازیم.

۳.آ متریک فیشر

قبل از بررسی کامل اینکه یک انحراف چیست، ببینیم چگونه میتوانیم با استفاده از یک انحراف، به خواص هندسی رویه اطلاعات خود می پردازیم. در واقع به صورت تاریخی، این مشاهده شروع اصلی هندسه اطلاعات است.

قبل تر در درس دیدیم که فاصله بین دو توزیع مشابه، با پارامتر های نزدیک رفتار توان دویی دارد. در تعریف متریک نیز دیده ایم که با استفاده از متریک، فاصله نقاط نزدیک را میتوانیم به صورت زیر حساب کنیم.

$$D(p_\theta, p_{\theta+\delta}) = \delta^T (I(\theta) + o(1))\delta, \quad ds^2 = g_{ij} dx^i dx^j$$

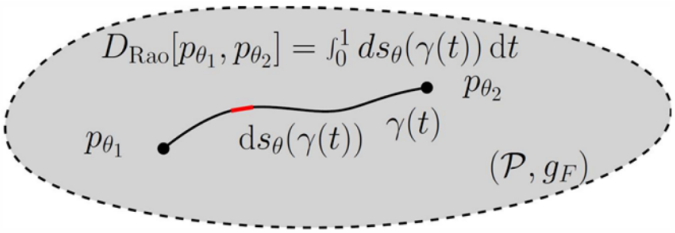
همچنین، تبدیل ماتریس فیشر را دیده بودیم، و تبدیل یک متریک بین مختصات را نیز از هندسه داشتیم. که در آن A ماتریس ژاکوبین تبدیل بین θ, θ' است.

$$I(\theta') = A^T I(\theta) A, \quad g' = A^T g A$$

همانطور که واضح است، ماتریس اطلاعات فیشر بسیار مشابه به یک متریک عمل میکند، میتوانیم از این ایده استفاده کنیم تا هندسه و در واقع خم فیشر–رائو را تشکیل دهیم.

$$\begin{aligned} g_F(u, v) &= \mathbb{E}[u(x), v(x)] = \mathbf{Cov}(u, v) \\ g_F(\partial_i, \partial_j) &= \mathbb{E}[\partial_i l_x(\theta) \partial_j l_x(\theta)] = I_{ij}(\theta) \\ g_F(u, v) &= [u]_B^T I(\theta) [v]_B \end{aligned}$$

حال با استفاده از این میتوانیم فاصله فیشر–رائو بین دو توزیع را تعریف کنیم، به صورت طول کمینه خمینه ای که بین این دو توزیع کشیده میشود، با توجه به متریک رائو.



شکل ۷: فاصله فیشر–رائو

به طور کلی محاسبه این فاصله به شدت سخت است زیرا نیاز به حل معادله ژئودزی دارد، حتی برا متغیر های گوسی چند متغیره رابطه بسته ای وجود ندارد.

یک نکته مهم این متر، این است که حال این فاصله دیگر به مختصه بستگی ندارد، و به عبارتی ناوردا است.

۳.ب ارتباط متمم از یک انحراف: $(\mathcal{M}, D) = (\mathcal{M}, {}^{D}g, {}^{D}\nabla, {}^{D}\nabla^{*} = {}^{D^{*}}\nabla)$

اگر به صورت ساده به مسئله نگاه کنیم، یک انحراف در این ادبیات به صورت یک تابع فاصله پیوسته و مشتق پذیر است، که میتواند نامتقارن هم باشد.

تعریف: به ریاضی، یک انحراف، $D : M \times M \to [0, \infty)$ بر روی خمینه $\mathcal{M}$، با توجه به مختصه Θ یک تابع ۳ بار مشتق پذیر است که خواص زیر را ارضا میکند.

۱. $D(\theta : \theta') \geq 0$ برای تمامی $\theta, \theta' \in \Theta$ و که تساوی تنها برای $\theta = \theta'$ رخ میدهد.

۲. $\frac{\partial D(\theta:\theta')}{\partial \theta^i} = \frac{\partial D(\theta:\theta')}{\partial \theta^j} = 0$ برای تمامی جهت ها.

۳. $-\frac{\partial D(\theta:\theta')}{\partial \theta'^j \partial \theta^i}$ یک ماتریس مثبت معین است.

به سادگی میتوان نشان داد که اکثر انحراف هایی که در درس دیدیم، این خواص را ارضا میکنند.

دو خاصیت آخر به ما القا میکنند که این انحراف ها، برای ما در واقع نوعی [2]فاصله هستند و اینگونه رفتار میکنند.

داخل درس به این مفهوم چندین بار برخورد کردیم، رفتار همسایگی نزدیک انحراف هایی که دیده ایم مانند انحراف χ^2 و KL را دیدیم که به صورت $\delta\theta^T I(\theta)\delta\theta$ رفتار میکند، که فاصله توان ۲ ای است، قضیه فیثاغورس برای KL را دیدیم(به این ها در قسمت های بعدی باز میگردیم)، و حتی به صورت ساده تر، دیده ایم که برای دو توزیع نورمال گوسی، $D_{KL}(\mathcal{N}(\mu_1, \sigma)||\mathcal{N}(\mu_2, \sigma)) = \frac{\|\mu_1 - \mu_2\|_2^2}{2\sigma^2}$ که به طور واضح به توان دویی بودن این انحراف ها اشاره میکند.

همانطورکه اندکی پیش اشاره کردیم، با بررسی رفتار انحرافی مانند KL، در همسایگی آن، میبینیم که $D_{KL}(p_\theta||p_{\theta+\delta}) = \delta^T(I(\theta) + o(1))\delta$، که به هندسه ای مانند چیزی که دیده بودیم در $ds^2 = g_{ij}dx^i dx^j$ اشاره میکند، با الهام از این، ما هندسه خود را میتوانیم از یک واگرایی تعریف کنیم.

$$
\begin{aligned}
{}^{D}g_{ij} &= -\frac{\partial^2 D(\theta : \theta')}{\partial\theta^i \partial\theta'^j}\Big|_{\theta=\theta'} \\
\Gamma^{k}_{ij} &= -\frac{\partial^3 D(\theta : \theta')}{\partial\theta^i \partial\theta^j \partial\theta'^k}\Big|_{\theta=\theta'} \\
\Gamma^{*k}_{ij} &= -\frac{\partial^3 D(\theta : \theta')}{\partial\theta'^i \partial\theta'^j \partial\theta^k}\Big|_{\theta=\theta'} \\
{}^{D}C_{ijk} &= \Gamma^{*k}_{ij} - \Gamma^{k}_{ij}
\end{aligned}
$$

طبیعتا اینجا نیز میتوانیم به ازای $\alpha \in \mathbb{R}$ خانواده ای از خمینه های آماری بسازیم.

$$\left\{(\mathcal{M}, {}^{D}g, {}^{D}C^{\alpha}) = (\mathcal{M}, {}^{D}g, {}^{D}\nabla^{-\alpha}, {}^{D}\nabla^{\alpha})\right\}_{\alpha\in\mathbb{R}}$$

۳.ج انحراف برگمن

در این قسمت، به یک قاعده کلی برای تعریف یک انحراف میپردازیم، (که در هندسه اطلاعات بسیار مفید واقع شده اند [۱])، یک تابع محدب و مشتق پذیر را در نظر بگیرید، به صورت $F(\theta)$ که به آن تابع پتانسیل میگوییم.

برای تولید یک انحراف، در مرحله اول میتوانیم صرفا از $F(\theta) - F(\theta')$ استفاده کنیم، ولی این انحراف به علت وجود تغییرات خطی در همسایگی، خواص قبلی را ارزا نمیکند، پس به سادگی، انحراف را به شکل زیر در نظر میگیرم.

$$B_F(\theta : \theta') = F(\theta) - F(\theta') - (\theta - \theta')^T \nabla F(\theta')$$

نمونه ای از چند تابع پتانسیل معروف:

- تابع پتانسیل درجه دو:
$$F(x) = \frac{1}{2}x^T Q x \Rightarrow B_F(\theta : \theta') = \frac{1}{2}(\theta - \theta')'^T Q(\theta - \theta')$$
- تابع پتانسیل آنتروپی منفی:
$$F(p) = \sum_i p_i \log(p_i) \Rightarrow B_F(p : q) = D_{KL}(p||q)$$
- انرژی آزاد، به این مورد برمیگردیم:

$$\mathcal{E} = \left\{ p_\theta(x) = \exp\left(\sum_i t_i(x)\theta_i - F(\theta + k(x)) \right) | \theta \in \Theta \right\} \Rightarrow B_F(\theta : \theta') = D_{KL}(p_\theta||p_{\theta'})$$

این انحراف، نوعی هندسه ساده را برای ما تعریف میکنند.

$$\begin{aligned} {}^F g &= \nabla^2 F(\theta) \\ {}^F\Gamma &= 0 \Rightarrow {}^F\nabla - flat \\ {}^F C_{ijk} &= \partial_i \partial_j \partial_k F(\theta) \end{aligned}$$

همچنین، با استفاده از تبدیل دوگان محدب، متوانیم تابع پتانسیل دوگان را بدست آوریم.

$$F^*(\eta) := \sup\left\{ \theta^T \eta - F(\theta) \right\}$$

از بهینه سازی، میدانیم که $\nabla^2 F^*(\eta)\nabla^2 F(\theta) = \mathbf{I}$، و که این ماکسیمم در نقطه $\eta = \nabla F(\theta), \theta = \nabla F^*(\eta)$ رخ میدهد، که با استفاده از آن باقی المان های هندسه خود را تشکیل میدهیم.

$$\begin{aligned} {}^F g^{ij}(\eta) &= \partial^i \partial^j F^*(\eta) \\ {}^F\Gamma^{*ijk} &= 0 \Rightarrow {}^F\nabla^* - flat \\ {}^F C^{ijk} &= \partial^i \partial^j \partial^k F^*(\eta) \end{aligned}$$

در اینجا به این نتیجه بنیادی میرسیم که این خمینه اطلاعات، یک خاصیت دوگانه تخت بودنی دارد، به این معنا که هم ∇-flat و هم ∇^*-flat است.

۳.د فضیه فیثاغورس تعمیم یافته

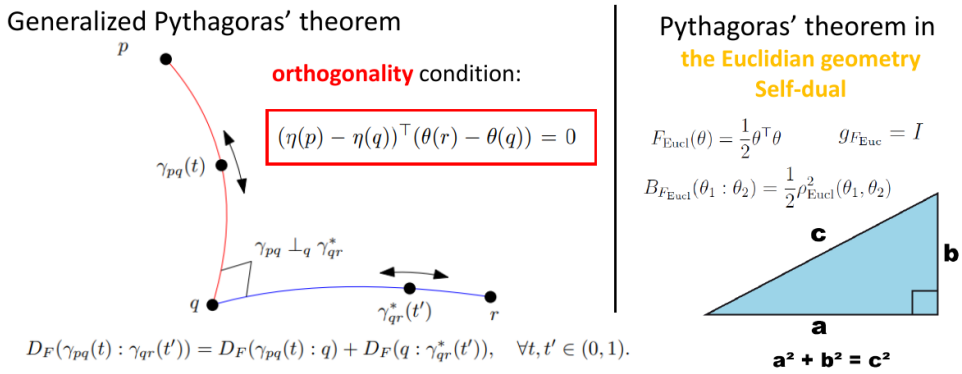
به سادگی میتوان دید که رابطه زیر برقرار است.

$$B_F(\theta_1 : \theta_2) = B_F(\theta_1 : \theta_3) + B_F(\theta_3 : \theta_2) - (\theta_1 - \theta_3)^T(\nabla F(\theta_2) - \nabla F(\theta_3))$$

و در مختصات دگان، داریم که $\eta = \nabla F(\theta)$، در نتیجه رابطه زیر را میتوانیم ببینیم.

$$B_F(\theta_1 : \theta_2) = B_F(\theta_1 : \theta_3) + B_F(\theta_3 : \theta_2) \iff (\theta_1 - \theta_3) \perp (\eta_1 - \eta_3)$$

که با توجه به شهود ما بر انحراف ها و رفتار فاصله توان دویی آنها، در واقع جلوه ای از قضیه فیثاغورس است.
با استفاده از این رابطه، و توابع پتانسیل متنوع، میتوانیم هم فیثاغورس اصلی، فیثاغورس برای انحراف KL و ...را اثبات کنیم.
در نگاه اول ممکن است به نظر بیاید که این مشاهده ارزش آنچنانی ندارد، ولی در واقع پایه ای از تعریف ما برای تصویر اطلاعاتی و باقی چیز های مبتنی بر آن است.

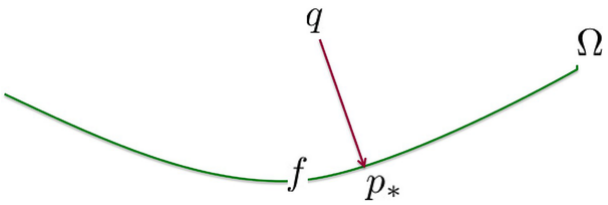


شکل ۸: نمایی از قضیه فیثاغورس تعمیم یافته

۴ رویکرد ها

۴.آ تصویر کردن

با توجه به قضیه فیثاغورس، و مثبت بودن انحراف ها، میتوانیم به یک مفهوم بنیادی دست پیدا کنیم.



اگر فرض کنیم p^* نقطه کمینه کننده انحراف بین یک خمینه ∇^*-flat با نقطه p باشد، برای باقی نقاط بر روی این خمینه رابطه زیر را داریم.

$$D(p:p')=D(p:p^*)+D(p^*:p')>D(p:p^*)$$

در نتیجه کمینه کننده انحراف، یک نقطه یکتا است، این نیز با نتیجه هایی که در کلاس از آن استفاده میکردیم مطابقت دارد، برای مثال اگر از انحراف KL استفاده کنیم، این نقطه در برخی از حالت ها به ما یکتا بودن تخمینگر بیشترین درست نمایی(MLE) را نشان میدهد! در برخی از حالت ها هم به ما یکتا بودن پاسخ قاعده بیشینه آنتروپی را نشان میدهد!
با استفاده از مفهوم یکتا بودن کمینه کننده انحراف، تصویر کردن را تعریف میکنیم.

$$S:\nabla^*flat\Rightarrow\nabla projection:P_s=\arg\min_{Q\in S}D(\theta(P):\theta(Q))$$
$$S:\nabla^flat\Rightarrow\nabla^* projection:P_s=\arg\min_{Q\in S}D(\theta(Q):\theta(P))$$

همچنین با الهام گرفتن از این، میتوانیم انحراف بین دو رویه را محسابه کنیم.

$$D(\mathcal{P}:\mathcal{Q})=\min_{q\in\mathcal{Q},p\in\mathcal{P}}D(p:q)$$

در حالتی که جفت رویه ها تخت بودن مناسبی داشته باشند، جواب یکتا است و با الگوریتم های چند مرحله ای میتوان به آن رسید. الگوریتم هایی مانند الگوریتم EM و مراحل یادگیری VAE را میتوان با این مفهوم معادل سازی کرد.

۴.ب تصویر-m و تصویر-e

ابتدا خانواده نمایی و خانواده ترکیبی توزیع ها را در نظر بگیرید.

$$\mathcal{E} = \{p_\theta(x) = \exp\big(\theta^T t(x) - F(\theta)\big) | \theta \in \Theta\}$$
$$\mathcal{M} = \left\{ p_\theta(x) = \sum_{i=1}^{D} \theta_i p_i(x) + (1 - \sum_{i=1}^{D} \theta_i) p_0(x) | \theta \in \Theta \right\}$$

که در آن $F(\theta)$ تابع انرژی آزاد است.

$$F(\theta) = \log \int \exp\big(\theta^T t(x)\big) d\mu(x) \Rightarrow F^*(\eta) = \int p(x) \log(p(x)) d\mu(x)$$

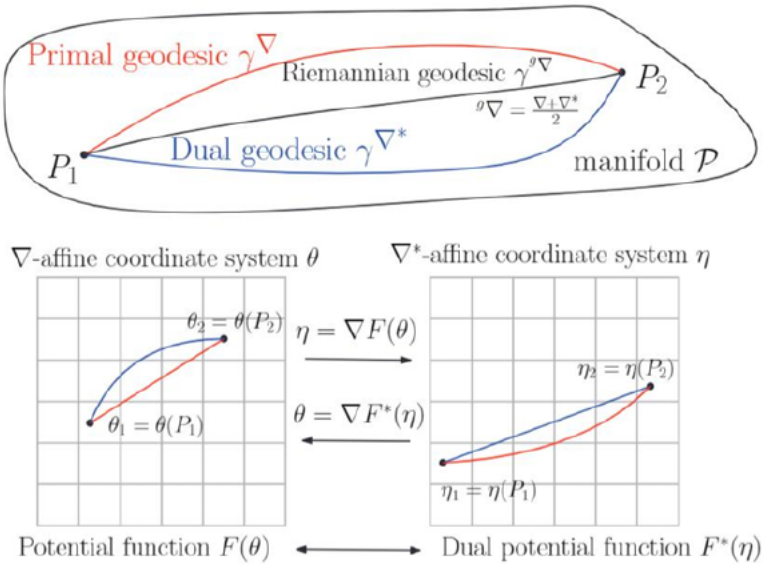
و همچنین برای مختصه η داریم که:

$$\eta = \nabla_\theta F(\theta) = \frac{\int (t(x) \exp\big(\theta^T t(x)\big) d\mu(x))}{\int \exp(\theta^T t(x)) d\mu(x)} = \mathbb{E}_\theta[t(x)]$$

در نتیجه، به مختصه دگان η در اینجا مختصه ممان ها میگویند، و به مختصه θ مختصه پارامتر کانونیک میگویند.
میتوان دید که در این فضا، روابط زیر برقرار است.

$$\nabla^m = (\nabla^e)^*, \nabla^e = (\nabla^m)^*, {}^M_e\Gamma = 0, {}^e_M\Gamma = 0, {}^m_M\Gamma = 0, {}^e_E\Gamma = 0$$

که در واقع یعنی در یک هندسه دگان تخت هستیم، و که توزیع های ترکیبی و خانواده نمایی دگان هم هستند.
و میتوانیم در این مختصه ها ژئودزی تعریف کنیم به صورت زیر.



شکل ۹: ژئودزی در مختصه اصلی و دگان

۱.ب.۴ تخمینگر MLE و تصویر–m

اگر با استفاده از ∇^m انتقال موازی داشته باشیم، میتوان در واقع یک تصویر–m کرده ایم، از استفاده این میتوان شهودی بر تخمینگر MLE را دید.
به سادگی میتوان دید که مسئله ماکسیمم درست نمایی، معادل کمینه کردن انحراف کولبک لیبر بین توزیع پارامتریزه شده و توزیع تجربی است.

$$D(p_e, p_\theta) = \sum_i \frac{n_i}{n} log(\frac{n_i}{np_\theta(x_i)}) = -h(p_e) + \frac{1}{n}\sum_i n_i log(p_\theta(x_i)) = -h(p_e) + \frac{1}{n}l(\theta)$$

در نتیجه میتوانیم ببینیم که این مسئله کمینه کردن انحراف کولبک لیبر، دقیقا همان تعریف ما از تصویر کردن است، با توجه به اینکه توزیع پارامتریزه شده ما یک ناحیه e-flat است.

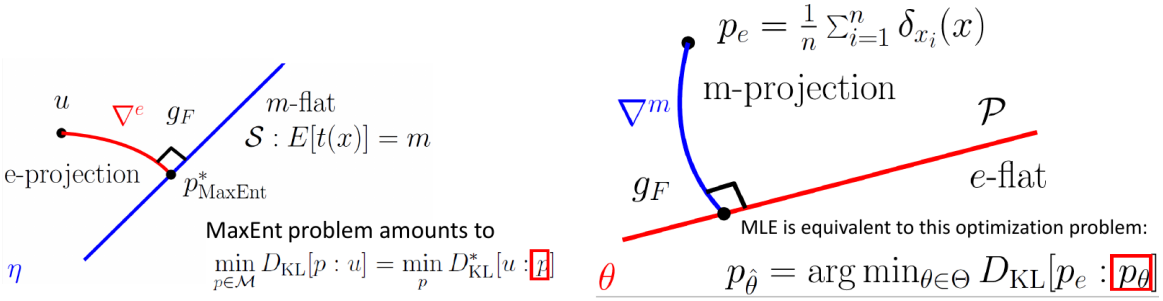
$$\hat{\theta}^{MLE} = \text{Proj}_{\mathcal{P}}^{\nabla^m}(p_e)$$

۲.ب.۴ قائده بیشینه آنتروپی

یک قائده دیگر که استفاده میکنیم، تایین کردن توزیع با استفاده از چند مشاهده است، زمانی که پیشینه خاصی نداریم. معمولا در این مواقع از قائده بیشینه آنتروپی استفاده میکنیم.
در این حالت از مسئله زیر استفاده میکنیم.

$$D(p,u) = \sum_i p_i \log(Np_i) = \log(N) + \sum_i p_i \log(p_i) = \log(N) - h(p)$$

در نتیجه کمینه کردن انحراف مزکور، معادل بیشینه کردن آنتروپی است.
اگر مشاهده های ما به صورت مشاهده های $\mathbb{E}[t_i(x)] = a_i$ باشد، این مشاهده ها ما را به صورت خطی محدود میکند، در واقع این ها ما را به زیرمجموعه $\mathcal{M}_{n-k}$ می برند.
حال چون که امید ریاضی تابعی خطی است، مجموعه ما به توزیع های ترکیبی بسته است، در نتیجه باید یک محموعه m-flat باشد. و نتیجه میگیریم که پاسخ بهینه، در واقع تصویر–m توزیع با بیشینه آنتروپی(یکنواخت)، بر روی زیر رویه $\mathcal{M}_{n-k}$ میباشد.



شکل ۱۰: تخمینگر MLE(چپ)، بیشینه آنتروپی(راست)

۵ کاربردها

۵.آ بررسی اثر گذاری

در قسمت قبلی، رویکرد های جدید، و به کل زاویه دید جدیدی بر روی مسائل پیشین داشتیم، که با استفاده از هندسه خاصه میتوانیم اقدام به بهینه کردن آن روش ها و به معرفی روش ها بپردازیم، برای مثال در [۵] روش جدید بر روی الگگوریتم EM معرفی میکند، [۳] به تعمیم دادن قاعده بیشینه آنتروپی میپردازد، و [۲] به یادگیری VAE ها میپردازد.
به طور کلی، هندسه اطلاعات میتواند زاویه جدیدی به مسائل زیادی در علوم اطلاعات بیاورد، برای مثال:

- **آمار:** الگوریتم هایی مانند ،EM و مدل های ARMA
- **یادگیری ماشین:** مدل های بولتزمن، رویه های عصبی، گرادیان طبیعی(در ادامه بررسی میشود)
- **پردازش سیگنال:** آنالیز منابع مستقل(ICA)، تجزیه نامنفی ماتریس(NMF)
- **برنامه ریزی ریاضی:** روش های مبتنی بر بهینه سازی، برای مثال روش الگوریتم تکاملی طبیعی(NES)
- **تئوری بازی:** توابع امتیاز دهی(Score functions)

در ادامه به برخی از الگوریتم ها، بخصوص آنهایی که مستقیما از هندسه اطلاعات بر می آیند را بررسی میکنیم.

۵.ب گرادیان طبیعی(NGD)

ابتدا به الگوریتم گرادیان نزولی معمولی نگاه میکنیم، و مشکلات آن را زیر نظر قرار میدهیم.

$$\text{GD:} \quad \theta_{t+1} = \theta_t - \alpha_t \nabla_\theta L_\theta(\theta_t)$$

نکته این است که اگر حال ما توزیع خود، و در عمل تابع زیان خود را با پارامتر بندی دیگری مانند $\eta = \eta(\theta)$ پارامتر بندی کنیم، مسیر بهینه سازی به طور کلی متفاوت خواهد بود، زیرا گرادیان های متفاوتی داریم، و در حالت کلی ممکن است به مینیمم های محلی متفاوتی منجر شوند.

$$\eta_{t+1} = \eta_t - \alpha_t \nabla_\eta L_\eta(\eta_t) = \eta_t - \alpha_t \nabla_\eta L_\theta(\theta_t) = \eta_t - \alpha_t \left[\frac{\partial \theta}{\partial \eta}\right] \nabla_\theta L_\theta(\theta_t)$$
$$= \eta(\theta_t) - \alpha_t \left[\frac{\partial \theta}{\partial \eta}\right] \nabla_\theta L_\theta(\theta_t) \approx \eta(\theta_t - \alpha_t \left[\frac{\partial \eta}{\partial \theta}\right]^{-1} \left[\frac{\partial \theta}{\partial \eta}\right] \nabla_\theta L_\theta(\theta_t)) \neq \eta(\theta_{t+1})$$

این اتفاق با انتظار ما از ناوردایی نثبت به پارامتر بندی مطابق نیست، در نتیجه مطلوب نیست. همچنین در نزدیکی نقطه بهینه، به علت محو شدن اطلاعات فیشر، نرخ بهینه شدن به شدت افت میکند.
با استفاده از هندسه، و روش هایی که آنجا استفاده میکردیم، میفهمیم که باید در واقع گرادیان را در جهت بیشینه کاهش ریمانی در نظر گرفت، نه اقلیدسی، که در نهایت به این معنی است که باید از متریک استفاده کنیم.

$$\text{NGD:} \quad \theta_{t+1} = \theta_t - \alpha_t g(\theta)^{-1} \nabla_\theta L_\theta(\theta_t) = \theta_t - \alpha_t I(\theta)^{-1} \nabla_\theta L_\theta(\theta_t)$$

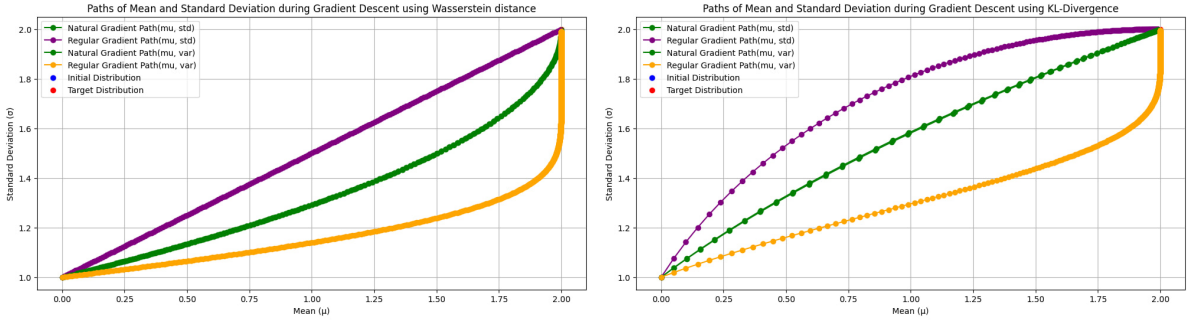
ابتدا، در عمل میبینیم که این مشکل محو شدن گرادیان در نزدیکی نقطه بهینه را رفع میکند، و همچنین ناوردایی را همراه خود میاورد.

$$\eta_{t+1} = \eta_t - \alpha_t I(\eta)^{-1} \nabla_\eta L_\eta(\eta_t) = \eta_t - \alpha_t I(\eta)^{-1} \nabla_\eta L_\theta(\theta_t) = \eta_t - \alpha_t \left[\frac{\partial \eta}{\partial \theta}\right] I(\theta)^{-1} \left[\frac{\partial \eta}{\partial \theta}\right] \left[\frac{\partial \theta}{\partial \eta}\right]$$
$$\cdot \nabla_\theta L_\theta(\theta_t) = \eta(\theta_t) - \alpha_t \left[\frac{\partial \eta}{\partial \theta}\right] \nabla_\theta L_\theta(\theta_t) \approx \eta(\theta_t - \alpha_t I(\theta)^{-1} \nabla_\theta L_\theta(\theta_t)) = \eta(\theta_{t+1})$$

در واقع دیگر بهینه سازی ما وابسته به پارامتر بندی و تعریف مسئله نیست، میتوانیم حتی بگوییم معادل با بهترین پارامتر بندی است!

۱.ب.۵ پیاده سازی

در اینجا، برای دیدن کاربردی بودن این الگوریتم، یک مسئله ساده را در نظر میگیرم.
اینجا، تلاش میکنیم تا بین دو توزیع گوسی، درون یابی کنیم، و از دو نوع تابع زیان متفاوت، یکی انحراف کولبک لیبر، و دیگری فاصله واشراشتاین استفاده میکنیم. در پیاده سازی خود، باری در تعریف از گوسی با پارامتر بندی (μ, σ) و بار دیگر با (μ, σ^2) جلو میرویم، تا تفاوت ها و اثر ناوردایی را مشاهده کنیم.



شکل ۱۱: روند بهینه سازی، تابع زیان کولبک لیبر(راست)، واشراشتاین(چپ)

در پیاده سازی خود میبینیم که، مسیر های بسیار متفاوتی طی میکنند، ولی با استفاده از گرادیان طبیعی، جفت مسیر ها یکسان است. همچنین مشاهده میکنیم که نزدیک به نقطه بهینه، گرادیان تقریبا محو شده و ایتریشن های بسیار زیادی طول میکشد، در جدول زیر عملکرد را میبینیم.

$L(\theta)$	**NGD**	**GD:** (μ, σ)	**GD:** (μ, σ^2)
KL	67	214	1288
Wasserstein	164	223	2595

جدول ۱: مقایسه بین الگوریتم ها، تعداد مرحله طی شده برای بهینه سازی

که نتایج جدول ۱ برای ما واضح میکند که الگوریتم گرادیان نزولی طبیعی روش خوبی است.

۵.ج تایین توزیع پیشین پیوسته

یکی از کاربردهای دیگری که می‌توان به آن اشاره کرد برای زمانی است که می‌خواهیم توزیع پیشین روی پارامترها در یک مساله تخمین در نظر بگیریم. اگر اطلاعاتی در مورد توزیع پیشین نداشته باشیم علاقه داریم که توزیعی انتخاب کنیم که بیشترین ابهام را داشته باشد. اگر تخمین بین چند گزینه باشد یا به عبارتی پارامتر گسسته باشد به راحتی می‌توانیم توزیع

یکنواخت را برای آن ها در نظر بگیریم اما برای وقتی که پارامترها پیوسته اند این سوال پیش می‌آید که توزیع یکنواخت نسبت به چه پارامتربندی ای؟ در اینجا میتوانیم به تعریف عنصر حجم ناورا نگاه کنیم:

$$\omega = \sqrt{\det g}\, d\theta^1 \wedge d\theta^2 \wedge \cdots \wedge d\theta^N \tag{۲۰}$$

اگر بخواهیم مشابه حالت گسسته عمل کنیم میتوانیم جواب را اینگونه در نظر بگیریم که توزیع باید متناسب باشد با $\sqrt{\det g}$ یعنی داریم:

$$\pi(\theta) \propto \sqrt{\det g} \tag{۲۱}$$

در اینجا π توزیع پیشین روی پارامتر است. که ضریب تناسب از بهنجار بودن توزیع احتمال بدست میاید. بنابراین با دید هندسی توانستیم به این سوال نابدیهی که برای مورد پیوسته چه توزیعی را میتوانیم معادل توزیع یکنواخت در نظر گرفت پیدا کرد. همانطور که گفته شده نابدیهی بودن از اینجا میاید که یکنواخت بودن نسبت به یک پارامتر یکنواخت بودن نسبت به پارامتر دیگر را نتیجه نمی دهد. اما با استفاده از فرم حجم ناوردا این مشکل برطرف می‌شود.

۶ تقدیر، نتیجه گیری و منابع

در تهیه این گزارش، از کار های Frank nielsen و به خصوص مقاله [۴] استفاده بسیار شده و در این قسمت از ایشان تقدیر میکنیم، از منابع منبع باز ایشان نیز در بعضی از نتیجه گیری ها و نمودار ها استفاده کرده ایم.

در نهایت، حوزه هندسه اطلاعات، بسیار حوزه بزرگی است و چکیده کردن آن بیشتر از این گزارش، از تمامیت آن میکاهید، این حوزه بسیار حوزه جذابی است و دید های بسیار متفاوتی به علوم اطلاعات، نسبت به روش های کلاسیک میدهد.

ممنون از همراهی شما.

References

[1] Shun-ichi Amari and Andrzej Cichocki. “Information geometry of divergence functions”. In: *BULLETIN OF THE POLISH ACADEMY OF SCIENCES TECHNICAL SCIENCES* 58 (Mar. 2010). DOI: 10.2478/v10175-010-0019-1.

[2] Tian Han, Jun Zhang, and Ying Nian Wu. “From em-Projections to Variational Auto-Encoder”. In: *NeurIPS 2020 Workshop: Deep Learning through Information Geometry*. 2020. URL: https://openreview.net/forum?id=NXbapYR49pg.

[3] Pablo A. Morales and Fernando E. Rosas. “Generalization of the maximum entropy principle for curved statistical manifolds”. In: *Physical Review Research* 3.3 (Sept. 2021). ISSN: 2643-1564. DOI: 10.1103/physrevresearch.3.033216. URL: http://dx.doi.org/10.1103/PhysRevResearch.3.033216.

[4] Frank Nielsen. “An Elementary Introduction to Information Geometry”. In: *Entropy* 22.10 (Sept. 2020), p. 1100. ISSN: 1099-4300. DOI: 10.3390/e22101100. URL: http://dx.doi.org/10.3390/e22101100.

[5] Sammy Suliman. *The EM Algorithm in Information Geometry*. May 2024.