تمرین ۱ (اثبات رابطه واریانس مجموع دو متغیر تصادفی) *می‌خواهیم نشان دهیم:*

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$$

جواب:

گام ۱: تعریف واریانس

$$\mathrm{Var}(Z) = \mathbb{E}[(Z - \mathbb{E}[Z])^2]$$

گام ۲: جایگذاری $Z = X + Y$

$$\mathrm{Var}(X+Y) = \mathbb{E}[(X + Y - \mathbb{E}[X+Y])^2]$$

گام ۳: خطی بودن امید ریاضی

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

پس داریم:

$$\mathrm{Var}(X+Y) = \mathbb{E}[(X - \mathbb{E}[X] + Y - \mathbb{E}[Y])^2]$$

گام ۴: باز کردن مربع

$$(X - \mathbb{E}[X] + Y - \mathbb{E}[Y])^2 = (X - \mathbb{E}[X])^2 + (Y - \mathbb{E}[Y])^2 + 2(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])$$

گام ۵: گرفتن امید ریاضی هر جمله

$$\mathrm{Var}(X+Y) = \mathbb{E}[(X - \mathbb{E}[X])^2] + \mathbb{E}[(Y - \mathbb{E}[Y])^2] + 2\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

گام ۶: استفاده از تعریف واریانس و کوواریانس

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mathbb{E}[X])^2], \quad \mathrm{Var}(Y) = \mathbb{E}[(Y-\mathbb{E}[Y])^2], \quad \mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$$

در نتیجه:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$$

رابطه اثبات شد.

تمرین ۲ (کاهش مسئله الاستیک نت به لاسو) *مسئله الاستیک نت به شکل زیر تعریف شده است:*

$$J_1(w) = \|y - Xw\|_2^2 + \lambda_2\|w\|_2^2 + \lambda_1\|w\|_1$$

برای کاهش به مسئله لاسو، داده‌های تغییر یافته را به صورت زیر تعریف می‌کنیم:

$$\tilde{X} = \begin{pmatrix} X \\ \sqrt{\lambda_2} I_d \end{pmatrix}, \quad \tilde{y} = \begin{pmatrix} y \\ 0_{d\times 1} \end{pmatrix}.$$

تابع هزینه لاسو متناظر با داده‌های تغییر یافته به شکل زیر نوشته می‌شود:

$$J_2(\tilde{w}) = \|\tilde{y} - \tilde{X}\tilde{w}\|_2^2 + \lambda_1\|\tilde{w}\|_1$$

اثبات رابطه بین مینیمایزرها:

۱. **گام اول: جایگذاری در تابع هزینه** فرض کنیم $w = \tilde{w}$. آنگاه داریم:

$$J_1(\tilde{w}) = \|y - X\tilde{w}\|_2^2 + \lambda_2\|\tilde{w}\|_2^2 + \lambda_1\|\tilde{w}\|_1$$

۲. **گام دوم: بازنویسی قسمت نُرم دو با داده‌های تغییر یافته** با توجه به تعریف $\tilde{X}$ و $\tilde{y}$:

$$\tilde{y} - \tilde{X}\tilde{w} = \begin{pmatrix} y - X\tilde{w} \\ 0 - \sqrt{\lambda_2}\tilde{w} \end{pmatrix} = \begin{pmatrix} y - X\tilde{w} \\ -\sqrt{\lambda_2}\tilde{w} \end{pmatrix}.$$

بنابراین نُرم دو آن برابر است با:

$$\|\tilde{y} - \tilde{X}\tilde{w}\|_2^2 = \|y - X\tilde{w}\|_2^2 + \|\sqrt{\lambda_2}\tilde{w}\|_2^2 = \|y - X\tilde{w}\|_2^2 + \lambda_2\|\tilde{w}\|_2^2.$$

۳. **گام سوم: بازنویسی تابع هزینه** با جایگذاری نتیجه بالا در J_2:

$$J_2(\tilde{w}) = \|\tilde{y} - \tilde{X}\tilde{w}\|_2^2 + \lambda_1\|\tilde{w}\|_1 = \|y - X\tilde{w}\|_2^2 + \lambda_2\|\tilde{w}\|_2^2 + \lambda_1\|\tilde{w}\|_1 = J_1(\tilde{w}).$$

۴. **گام چهارم: نتیجه‌گیری درباره مینیمایزرها** از آنجا که تابع‌های هزینه برای هر w برابرند، مینیمایزرها نیز برابر خواهند بود:

$$\tilde{w}^* = \arg\min J_2(\tilde{w}) \quad \Rightarrow \quad w^* = \arg\min J_1(w) = \tilde{w}^*.$$

.

تمرین ۳ (مدل برنولی–گوسی و نُرم صفر در منظم سازی) فرض کنید می‌خواهیم وزن‌های $w = (w_1, \ldots, w_d)$ یک مدل خطی را به صورت sparse پیدا کنیم. یکی از روش‌های Bayesian برای ایجاد sparsity، استفاده از prior Gaussian–Bernoulli است:

$$w_i \sim \textit{Bernoulli}(\pi) \cdot \mathcal{N}(0, \sigma^2),$$

که در آن:

- با احتمال π، $w_i \neq 0$ و از توزیع گوسی $\mathcal{N}(0, \sigma^2)$ نمونه می‌شود،
- با احتمال $1 - \pi$، $w_i = 0$ است.

گام ۱: نوشتن log-prior احتمال joint برای تمام وزن‌ها به شکل زیر است:

$$p(w) = \prod_{i=1}^{d} \left[\pi\,\mathcal{N}(w_i; 0, \sigma^2) + (1 - \pi)\delta(w_i)\right],$$

که $\delta(w_i)$ دلتای دیراک است. با گرفتن لگاریتم:

$$\log p(w) = \sum_{i:w_i \neq 0} \left(\log \pi + \log \mathcal{N}(w_i; 0, \sigma^2)\right) + \sum_{i:w_i = 0} \log(1 - \pi).$$

گام ۲: ارتباط با نُرم صفر (ℓ_0) اگر تعداد وزن‌های غیرصفر را $\|w\|_0$ بنامیم، می‌توان نوشت:

$$\log p(w) = \underbrace{\|w\|_0 \cdot \log \frac{\pi}{1 - \pi}}_{\text{جمله sparsity}} + \sum_{i:w_i \neq 0} \log \mathcal{N}(w_i; 0, \sigma^2) + \textit{const}.$$

بنابراین، فرض prior Gaussian–Bernoulli باعث می‌شود که یک جمله متناسب با $\|w\|_0$ در تابع هزینه ظاهر شود.

گام ۳: تعریف تابع هزینه MAP در بهینه‌سازی MAP برای یافتن وزن‌ها:

$$\hat{w}_{MAP} = \arg\min_{w} -\log p(y|X,w) - \log p(w),$$

جمله $-\log p(w)$ به عنوان regularizer عمل می‌کند. با جایگذاری log-prior داریم:

$$-\log p(w) \sim const + \underbrace{-\|w\|_0 \log \frac{\pi}{1-\pi}}_{\text{نُرم صفر}} - \sum_{i:w_i \neq 0} \log \mathcal{N}(w_i; 0, \sigma^2)$$

که نشان می‌دهد هر وزن غیرصفر یک «هزینه» به تابع هزینه اضافه می‌کند و مدل sparse می‌شود.

گام ۴: جمع‌بندی و نتیجه

- prior Gaussian–Bernoulli به صورت طبیعی sparsity را القا می‌کند.
- تعداد وزن‌های غیرصفر $\|w\|_0$ مستقیماً وارد تابع هزینه می‌شود و عمل regularization انجام می‌دهد.
- بنابراین فرض اینکه هر وزن با احتمال π غیرصفر و با احتمال $1-\pi$ صفر است، منجر به ظاهر شدن جمله‌ای متناسب با $\|w\|_0$ در تابع هزینه می‌شود.
- این همان **نُرم صفر** (ℓ_0) **regularizer** است که به ما کمک می‌کند وزن‌های غیرضروری را حذف کنیم و مدل sparse به دست آوریم.

تمرین ۴ (تفاوت منظم‌سازهای ℓ_1 و ℓ_2 در لاسو و ریج) در مسائل رگرسیون منظم‌شده، دو نوع منظم‌ساز رایج داریم:

۱. لاسو (ℓ_1)

$$J(w) = \|y - Xw\|_2^2 + \lambda\|w\|_1 = \|y - Xw\|_2^2 + \lambda \sum_i |w_i|$$

۲. ریج (ℓ_2)

$$J(w) = \|y - Xw\|_2^2 + \lambda\|w\|_2^2 = \|y - Xw\|_2^2 + \lambda \sum_i w_i^2$$

دیدگاه هندسی:

- ℓ_1 ناحیه مجاز ضرایب به شکل یک **الماس (diamond)** است. گوشه‌های الماس روی محورهای مختصات قرار دارند، بنابراین وقتی سطح خطای مربعی (بیضوی) با این ناحیه برخورد می‌کند، اغلب نقطه‌ی برخورد روی یکی از گوشه‌ها قرار می‌گیرد. این نقاط متناظر با وزن‌های صفر هستند.
- ℓ_2 ناحیه مجاز به شکل **دایره (sphere)** است. دایره گوشه ندارد، بنابراین برخورد سطح خطا معمولاً در نقاط داخلی رخ می‌دهد و ضرایب کوچک اما **غیرصفر** باقی می‌مانند.

دیدگاه احتمالاتی: (Bayesian)

- لاسو متناظر با پیشین **لاپلاس (Laplace prior)** برای ضرایب است:

$$p(w_i) \propto \exp(-\lambda|w_i|)$$

نوک تیز در صفر باعث می‌شود که بسیاری از ضرایب دقیقاً صفر شوند، و مدل sparse تولید شود.

- ریج متناظر با **prior Gaussian** است:

$$p(w_i) \propto \exp(-\lambda w_i^2)$$

که پیوسته‌تر است و احتمال صفر دقیقاً خیلی کمتر است، بنابراین ضرایب کوچک اما غیرصفر باقی می‌مانند.

کاربردها:

- منظم‌ساز ℓ_1 برای **انتخاب ویژگی** ، داده‌های با ابعاد بالا ، مدل‌های sparse و جلوگیری از overfitting در حضور تعداد زیادی متغیر استفاده می‌شود.
- منظم‌ساز ℓ_2 برای **پایداری مدل**، کاهش حساسیت به نویز و جلوگیری از overfitting مناسب است، اما وزن‌ها معمولاً غیرصفر باقی می‌مانند و مدل sparse تولید نمی‌کند.