

## Spis treści

0.0.1	Definicja przestrzeni probabilistycznej . . . . .	2
0.0.2	Prawdopodobieństwo warunkowe . . . . .	2
0.0.3	Zmienne losowe . . . . .	2
0.0.4	Rozkłady brzegowe . . . . .	3
0.0.5	Zmienne losowe niezależne . . . . .	4
0.0.6	Rozkłady warunkowe . . . . .	4
0.0.7	Transformacja zmiennych wielowymiarowych . . . . .	4
0.0.8	Macierz kowariancji . . . . .	4
0.0.9	Wielowymiarowy rozkład normalny . . . . .	5
0.0.10	Zbieżność w rachunku prawdopodobieństwa . . . . .	5
0.0.11	Wnioskowanie statystyczne . . . . .	6
0.0.12	Silne prawo wielkich liczb . . . . .	7
0.0.13	Centralne Twierdzenie Graniczne . . . . .	7
0.0.14	Estymatory punktowe MLE i MAP . . . . .	7
0.0.15	Wnioskowanie Bayesowskie . . . . .	9
0.0.16	Modele Gaussowskie . . . . .	9
0.0.17	Liniowe modele Gaussowskie . . . . .	10
0.0.18	Regresja liniowa . . . . .	11
0.0.19	Regularyzacja . . . . .	13
0.0.20	Procesy Gaussowskie . . . . .	14
0.0.21	Wieloklasowa regresja logistyczna . . . . .	17
0.0.22	Metoda spadku wzdłuż gradientu . . . . .	18
0.0.23	Wnioskowanie metodami Monte Carlo . . . . .	20
0.0.24	Algorytm Importance Sampling (IS) . . . . .	21
0.0.25	Algorytm Metropolisa–Hastingsa . . . . .	22
0.0.26	Elementarz teorii informacji . . . . .	24

### 0.0.1 Definicja przestrzeni probabilistycznej

Rozkładem prawdopodobieństwa  $P$  w pewnym zbiorze zdarzeń elementarnych  $\Omega \neq \emptyset$  nazywamy odwzorowanie

$$P : \Sigma \mapsto [0; 1],$$

gdzie  $\Sigma$  jest rodziną podzbiorów  $\Omega$  (inaczej rodziną zdarzeń) taką, że

$$\Omega \in \Sigma, \quad A \in \Sigma \implies A' \in \Sigma, \quad \forall A_1, A_2, \dots \in \Sigma : \bigcup_i A_i \in \Sigma,$$

które spełnia:  $P(\Omega) = 1$  oraz dla dowolnych parami rozłącznych zdarzeń  $A_1, A_2, \dots \in \Sigma$  zachodzi

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i).$$

Trójkę  $(\Omega, \Sigma, P)$  nazywamy przestrzenią probabilistyczną. Z powyższej definicji wynikają znane własności prawdopodobieństwa tj.  $P(A') = 1 - P(A)$  oraz  $P(A \cup B) = P(A) + P(B) - P(A, B)$ .

### 0.0.2 Prawdopodobieństwo warunkowe

Definiujemy również prawdopodobieństwo warunkowe zdarzenia  $A$  pod warunkiem zdarzenia  $B$  o dodatnim prawdopodobieństwie

$$P(A | B) := \frac{P(A, B)}{P(B)}.$$

Na podstawie powyższej definicji definiujemy niezależność zdarzeń  $A, B$  jako własność  $P(A, B) = P(A)P(B)$ , co dla zdarzenia  $B$  o dodatnim prawdopodobieństwie jest równoważne z  $P(A | B) = P(A)$ . Ponadto jeśli zdarzenia  $A_1, A_2, \dots \in \Sigma$  są parami rozłączne i zachodzi  $\bigcup_i A_i = \Omega$  to dla dowolnego zdarzenia  $B \in \Sigma$  możemy zapisać

$$P(B) = \sum_i P(B | A_i)P(A_i).$$

Z definicji prawdopodobieństwa warunkowego trywialnie udowodnić twierdzenie Bayesa

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}.$$

### 0.0.3 Zmienne losowe

W uczeniu maszynowym będą interesować nas zmienne o wartościach w  $\mathbb{R}^n$ . Zmienne takie nazywamy zmiennymi losowymi wielowymiarowymi i definiujemy jako odwzorowania

$$X : \Omega \mapsto \mathbb{R}^n$$

takie, że dla każdego  $A \subseteq \mathbb{R}^n$  zbiór  $\{\omega \in \Omega \mid X(\omega) \in A\}$  należy do rodziny zdarzeń  $\Sigma$ . Przy takiej definicji prawdopodobieństwo, iż zmienna  $X$  ma wartość należącą do pewnego przedziału  $A$  wynosi

$$P(X \in A) = P(\{\omega \in \Omega \mid X(\omega) \in A\}).$$

Dowolny rozkład prawdopodobieństwa zmiennej losowej  $n$ -wymiarowej  $X = (X_1, X_2, \dots, X_n)$  jest wyznaczony jednoznacznie przez zadanie funkcji  $F(\mathbf{x}) : \mathbb{R}^n \mapsto [0; 1]$  zwanej dystrybuantą zdefiniowanej jako

$$F(\mathbf{x}) = F(x_1, \dots, x_n) := P(X_1 \leq x_1, \dots, X_n \leq x_n).$$

Zasadniczo będą nas interesować jednak dwa przypadki rozkładów prawdopodobieństwa zmiennych losowych: rozkłady dyskretne i rozkłady ciągłe. W przypadku rozkładu dyskretnego istnieje pewien przeliczalny zbiór  $S \subset \mathbb{R}^n$  taki, że  $P(X \in S) = 1$ . Rozkład ten jest zadany jednoznacznie przez podanie  $|S|$  liczb  $p_i > 0$  określających prawdopodobieństwa  $p_i = P(X = \mathbf{x}_i)$  dla wszystkich  $\mathbf{x}_i \in S$ . W przypadku rozkładu ciągłego istnieje z kolei funkcja  $p(\mathbf{x}) : \mathbb{R}^n \mapsto [0; \infty)$  taka, że

$$P(X_1 \in [a_1; b_1], \dots, X_n \in [a_n; b_n]) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} p(\mathbf{x}) d^n \mathbf{x}.$$

Funkcje  $p(\mathbf{x})$  nazywamy gęstością prawdopodobieństwa. W obu przypadkach musi być spełniony warunek unormowania postaci odpowiednio

$$\sum_i p_i = 1, \quad \int_{\mathbb{R}^n} p(\mathbf{x}) d^n \mathbf{x} = 1.$$

Będziemy często wykorzystywać wartość oczekiwaną pewnej funkcji  $f(\mathbf{x})$  zmiennej losowej  $X$  zdefiniowaną odpowiednio dla rozkładu  $p$  – dyskretnego lub ciągłego jako

$$\mathbb{E}[f(\mathbf{x})] := \sum_{\mathbf{x}_i \in S} f(\mathbf{x}_i) p_i \cong \int_{\mathbb{R}^n} f(\mathbf{x}) p(\mathbf{x}) d^n \mathbf{x}.$$

Zauważmy przy tym, iż funkcja  $f(\mathbf{x})$  może być zupełnie dowolna, np. dla funkcji charakterystycznej (indykatorowej) zbioru  $A \subset \mathbb{R}^n$   $f(\mathbf{x}) = \mathcal{I}_A$  mamy  $\mathbb{E}[\mathcal{I}_A(\mathbf{x})] = P(X \in A)$  lub dla iloczynu funkcji Heaviside'a  $f(\mathbf{x}) = \theta(t_1 - x_1) \cdots \theta(t_n - x_n)$  mamy  $\mathbb{E}[f(\mathbf{x})] = F(t_1, \dots, t_n)$ .

#### 0.0.4 Rozkłady brzegowe

Niech  $X = (X_1, \dots, X_n)$  będzie  $n$ -wymiarową zmienną losową o dystrybucie  $F(\mathbf{x})$ . Rozkład brzegowy względem  $k$  zmiennych  $X_{\sigma(1)}, \dots, X_{\sigma(k)}$  definiujemy jako rozkład wyznaczony przez dystrybuantę

$$F_{X_{\sigma(1)}, \dots, X_{\sigma(k)}}(x_{\sigma(1)}, \dots, x_{\sigma(k)}) := \lim_{x_{\sigma(k+1)} \rightarrow \infty, \dots, x_{\sigma(n)} \rightarrow \infty} F(x_1, \dots, x_n).$$

### 0.0.5 Zmienne losowe niezależne

Niech  $X = (X_1, \dots, X_k)$  będzie  $n$ -wymiarową zmienną losową o rozkładzie wyznaczonym przez dystrybuantę  $F(\mathbf{x})$ . Powiemy, iż zmienne losowe  $n_1, \dots, n_k$  - wymiarowych ( $n_1 + \dots + n_k = n$ )  $X_1, \dots, X_k$  są niezależne iff dla dowolnych  $\mathbf{x}_1 \in \mathbb{R}^{n_1}, \dots, \mathbf{x}_k \in \mathbb{R}^{n_k}$  zachodzi

$$F(\mathbf{x}_1, \dots, \mathbf{x}_k) = F_{X_1}(\mathbf{x}_1) \cdot \dots \cdot F_{X_k}(\mathbf{x}_k).$$

### 0.0.6 Rozkłady warunkowe

W ogólnym przypadku zmiennej losowej  $n$  - wymiarowej  $Z = (Z_1, \dots, Z_n)$  o ciągłym rozkładzie  $p(\mathbf{z})$  jeśli wydzielimy zmienne  $k$  i  $n - k$  - wymiarowe  $X = (Z_{\sigma(1)}, \dots, Z_{\sigma(k)})$ ,  $Y = (Z_{\sigma(k+1)}, \dots, Z_{\sigma(n)})$  to rozkład warunkowy zmiennej  $X | Y$  definiujemy jako rozkład zadany przez gęstość prawdopodobieństwa

$$p(\mathbf{x} | \mathbf{y}) := \frac{p(\mathbf{z})}{p_Y(\mathbf{y})} = \frac{p(\mathbf{x}, \mathbf{y})}{p_Y(\mathbf{y})}.$$

### 0.0.7 Transformacja zmiennych wielowymiarowych

Niech  $X = (X_1, \dots, X_n)$  będzie zmienną losową wielowymiarową o rozkładzie ciągłym o gęstości  $p_X(\mathbf{x})$ . Rozważmy bijekcję  $(X_1, \dots, X_n) \mapsto (Y_1, \dots, Y_n)$ . Chcemy znaleźć wyrażenie na gęstość  $p_Y(\mathbf{y})$  w nowych zmiennych. Ponieważ infinitesimalne prawdopodobieństwo jest niezmiennicze względem zmiany współrzędnych więc zachodzi

$$p_X(x_1, \dots, x_n) dx_1 \dots dx_n = p_Y(y_1, \dots, y_n) dy_1 \dots dy_n,$$

skąd

$$p_Y(y_1, \dots, y_n) = \left| \frac{\partial(x_1, \dots, x_n)}{\partial(y_1, \dots, y_n)} \right| p_X(x_1(\mathbf{y}), \dots, x_n(\mathbf{y})).$$

### 0.0.8 Macierz kowariancji

Macierz kowariancji funkcji  $f(\mathbf{x})$  zmiennej losowej  $X$  definiujemy jako

$$\mathbf{\Sigma}[f(\mathbf{x})] := \mathbb{E}[(f(\mathbf{x}) - \boldsymbol{\mu}_f)(f(\mathbf{x}) - \boldsymbol{\mu}_f)^\top],$$

gdzie  $\boldsymbol{\mu}_f = \mathbb{E}[f(\mathbf{x})]$ . Elementy diagonalne  $\Sigma_{ii}$  tej macierzy nazywamy wariancjami zmiennych  $X_i$ , natomiast elementy pozadiagonalne  $\Sigma_{ij}$  nazywamy kowariancjami zmiennych  $X_i$  i  $X_j$ . Oczywiście  $\mathbf{\Sigma}$  jest macierzą symetryczną. Nadto jeśli  $f$  jest funkcją identycznościową tj.  $f(\mathbf{x}) = \mathbf{x}$  to  $\mathbf{\Sigma}$  jest macierzą nieujemnie określoną, gdyż dla dowolnego  $\mathbf{v} \in \mathbb{R}^n$  mamy

$$\mathbf{v}^\top \mathbf{\Sigma} \mathbf{v} = \mathbb{E}[\mathbf{v}^\top (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{v}] = \mathbb{E}[z^2] \geq 0,$$

gdzie  $z = \mathbf{v}^\top (\mathbf{x} - \boldsymbol{\mu}) \in \mathbb{R}$ . Jeśli  $X_1, \dots, X_n$  są niezależne i  $f$  jest funkcją identycznościową to  $\mathbf{\Sigma}$  jest macierzą diagonalną.

### 0.0.9 Wielowymiarowy rozkład normalny

Jeśli zmienna wielowymiarowa  $X = (X_1, \dots, X_n)$  ma wielowymiarowy rozkład normalny (z ang. *Multivariate Normal distribution – MVN*) z wartością oczekiwaną  $\boldsymbol{\mu}$  i macierzą kowariancji  $\boldsymbol{\Sigma}$ , co oznaczamy jako  $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , to gęstość prawdopodobieństwa jest dana

$$\phi(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det \boldsymbol{\Sigma}}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\}$$

Macierz  $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$  nazywamy macierzą precyzji. Jeśli  $\mathbf{v}_i$  są unormowanymi wektorami własnymi macierzy  $\boldsymbol{\Sigma}$ , a  $\lambda_i$  odpowiadającymi im wartościami własnymi i zakładając, iż widmo  $\{\lambda_i\}$  jest niezdegenerowane mamy z twierdzenia spektralnego

$$\boldsymbol{\Lambda} = \sum_{i=1}^n \frac{1}{\lambda_i} \mathbf{v}_i \mathbf{v}_i^\top$$

oraz wiemy, iż wektory  $\{\mathbf{v}_i\}$  tworzą bazę ortonormalną przestrzeni  $\mathbb{R}^n$ . Z powyższego możemy zatem wyrazić wektor  $\mathbf{x} - \boldsymbol{\mu}$  jako kombinację liniową wektorów  $\{\mathbf{v}_i\}$  tj.

$$\mathbf{x} - \boldsymbol{\mu} = \sum_{i=1}^n t_i \mathbf{v}_i,$$

co pozwala zapisać gęstość prawdopodobieństwa jako

$$\phi(t_1, \dots, t_n) \cong \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \frac{t_i^2}{\lambda_i} \right\}.$$

Z powyższego wzoru widać, iż poziomice gęstości są wielowymiarowymi elipsoidami, których półosie są skierowane wzdłuż wektorów własnych  $\boldsymbol{\Sigma}$  i mają długości proporcjonalne do  $\sqrt{\lambda_i}$ .

Powiemy, iż wielowymiarowa zmienna losowa  $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  ma standardowy wielowymiarowy rozkład normalny jeśli  $\boldsymbol{\mu} = \mathbf{0}$  i  $\boldsymbol{\Sigma} = \mathbf{1}$ . Wówczas

$$\phi(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n}} \exp \left\{ -\frac{1}{2} \mathbf{x}^\top \mathbf{x} \right\}.$$

Można wykazać, iż jeśli  $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  dla  $\boldsymbol{\Sigma}$  o niezdegenerowanym widmie to wszystkie rozkłady brzegowe i warunkowe  $X$  są rozkładami normalnymi.

### 0.0.10 Zbieżność w rachunku prawdopodobieństwa

W rachunku prawdopodobieństwa definiujemy trzy zasadnicze rodzaje zbieżności ciągu zmiennych losowych  $(X_n)$ .

- Ciąg  $(X_n)$  jest zbieżny do  $X$  stochastycznie iff

$$\forall \epsilon > 0 : \lim_{n \rightarrow \infty} P(|X_n - X| < \epsilon) = 1.$$

- Ciąg  $(X_n)$  jest zbieżny do  $X$  z prawdopodobieństwem 1 iff

$$P\left(\lim_{n \rightarrow \infty} X_n = X\right) = 1.$$

- Ciąg  $(X_n)$   $n$ -wymiarowych zmiennych losowych jest zbieżny do  $X$  według dystrybuant iff

$$\forall \mathbf{x} \in \mathbb{R}^n, F_X(\mathbf{x}) - \text{ciągła w } \mathbf{x} : \lim_{n \rightarrow \infty} F_{X_n}(\mathbf{x}) = F_X(\mathbf{x})$$

Pomiędzy tak zdefiniowanymi rodzajami zbieżności zachodzą następujące implikacje:

1.  $X_n \rightarrow X$  z prawdopodobieństwem 1  $\implies X_n \rightarrow X$  stochastycznie
2.  $X_n \rightarrow X$  stochastycznie  $\implies X_n \rightarrow X$  według dystrybuant
3.  $X_n \rightarrow X$  stochastycznie  $\implies$  istnieje podciąg  $(X_{n_k})$  zbieżny do  $X$  z prawdopodobieństwem 1

#### 0.0.11 Wnioskowanie statystyczne

Modelem statystycznym nazwiemy parę  $(\chi, \mathcal{P})$ , gdzie  $\mathcal{P}$  jest rodziną rozkładów prawdopodobieństwa w zbiorze  $\chi$ , przy czym będziemy zakładać  $\chi = \mathbb{R}^n$

$$\mathcal{P} := \{p(\mathbf{x} \mid \theta) \mid \theta \in \Theta\},$$

gdzie  $\Theta$  jest zbiorem parametrów modelu  $\mathcal{P}$ . Prostą próbą losową w modelu  $\mathcal{P}$  nazwiemy ciąg niezależnych zmiennych losowych  $X_1, \dots, X_n$  o wartościach w  $\mathbb{R}^n$  i pochodzących z tego samego rozkładu  $p(\mathbf{x} \mid \theta) \in \mathcal{P}$  (w angielskiej terminologii taki ciąg zmiennych losowych nazwiemy *i.i.d.* tj. *independent and identically distributed*). Statystyką z kolei nazwiemy zmienną losową  $T$  będącą funkcją prostej próby losowej tj.  $T = T(X_1, \dots, X_n)$ . Być może najważniejszym przykładem statystyki jest średnia oznaczana jako  $\bar{X}$

$$\bar{X}(X_1, \dots, X_n) := \frac{X_1 + \dots + X_n}{n}.$$

Wartość oczekiwana statystyki średniej  $\bar{X}(X_1, \dots, X_n)$  dla  $X_i$  z rozkładu  $X \sim \mathcal{D}$  o gęstości  $p$  wynosi

$$\mathbb{E}[\bar{X}] = \int \dots \int \frac{1}{n} \left( \sum_{i=1}^n X_i \right) p(X_1) \dots p(X_n) dX_1 \dots dX_n = \mathbb{E}[X].$$

Wariancja statystyki średniej wynosi z kolei

$$\begin{aligned}
\text{Var}[\bar{X}] &= \mathbb{E}[\bar{X}^2] - \mathbb{E}[\bar{X}]^2 \\
&= \int \cdots \int \frac{1}{n^2} \left( \sum_{i=1}^n X_i^2 + \underbrace{\sum_{i \neq j} X_i X_j}_{n(n-1)} \right) p(X_1) \cdots p(X_n) dX_1 \cdots dX_n - \mathbb{E}[X]^2 \\
&= \frac{1}{n} \mathbb{E}[X^2] + \frac{n(n-1)}{n^2} \mathbb{E}[X]^2 - \mathbb{E}[X]^2 = \frac{1}{n} [\mathbb{E}[X^2] - \mathbb{E}[X]^2] = \frac{1}{n} \text{Var}[X].
\end{aligned}$$

### 0.0.12 Silne prawo wielkich liczb

Niech  $(X_n)$  będzie ciągiem zmiennych losowych i.i.d. z pewnego rozkładu  $X \sim \mathcal{D}$ . Przez  $(\bar{X}_n)$  oznaczmy ciąg średnich częściowych tj.

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Wówczas zachodzi silne prawo wielkich liczb

$$P\left(\lim_{n \rightarrow \infty} \bar{X}_n = \mathbb{E}[X]\right) = 1,$$

czyli średnia próbek zbiega do wartości oczekiwanej z prawdopodobieństwem 1.

Silne prawo wielkich liczb daje nam potężne narzędzie do szacowania wartości oczekiwanych, gdyż możemy je przybliżać średnią z dużej liczby próbek losowych, a dokładność tego przybliżenia zależy jedynie od liczby próbek i wariancji  $X$ . Jeśli  $X$  jest zmienną wielowymiarową to dokładność przybliżenia nie zależy wprost od liczby wymiarów i unikamy tzw. *curse of dimensionality*.

### 0.0.13 Centralne Twierdzenie Graniczne

Niech  $(X_n)$  będzie ciągiem  $k$ -wymiarowych zmiennych losowych i.i.d. z dowolnego rozkładu  $X \sim \mathcal{D}$  o wartości oczekiwanej  $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$  i odwracalnej macierzy kowariancji  $\boldsymbol{\Sigma}$ . Oznaczając przez  $(\bar{X}_n)$  ciąg średnich częściowych ciągu  $(X_n)$  zachodzi

$$\sqrt{n}(\bar{X}_n - \boldsymbol{\mu}) \rightarrow Z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

Oznacza to, iż dla ciągu  $X_1, \dots, X_n$  zmiennych losowych i.i.d. z praktycznie dowolnego rozkładu  $X \sim \mathcal{D}$  dla odpowiednio dużych  $n$  średnią z próbek możemy traktować jako zmienną losową o rozkładzie normalnym  $\mathcal{N}(\boldsymbol{\mu}, n^{-1/2}\boldsymbol{\Sigma})$ .

### 0.0.14 Estymatory punktowe MLE i MAP

Rozważamy model statystyczny  $\mathcal{P} = \{p(\mathbf{x} | \theta) | \theta \in \Theta\}$ . Estymatorem parametru  $\theta$  nazwiemy statystykę  $\hat{\theta}(X_1, \dots, X_n)$  służącą do oszacowania wartości tego

parametru. Wartość tej statystyki dla konkretnej realizacji prostej próby losowej  $\theta(\mathbf{x}_1, \dots, \mathbf{x}_n)$  nazwiemy estymatą parametru  $\theta$ . Dodatkowo definiujemy obciążenie (z ang. *bias*) estymatora jako wielkość

$$\mathbb{B}[\hat{\theta}] := \mathbb{E}[\hat{\theta}] - \theta.$$

Zasadniczo będą nas interesować dwa rodzaje estymat: MLE i MAP. W przypadku estymaty MLE (z ang. *Maximum Likelihood Estimate*) definiujemy funkcję wiarygodności (*likelihood*) dla modelu  $\mathcal{P} = \{p(\mathbf{x} | \theta) | \theta \in \Theta\}$  i realizacji prostej próby losowej (którą nazwiemy również danymi lub obserwacjami)  $D = (\mathbf{x}_1, \dots, \mathbf{x}_n)$  jako

$$p(D | \theta) = \prod_{i=1}^n p(\mathbf{x}_i | \theta).$$

Estymatą MLE nazywamy taką wartość parametru  $\theta_{\text{MLE}} \in \Theta$ , że

$$p(D | \theta_{\text{MLE}}) = \max_{\theta \in \Theta} p(D | \theta).$$

Ponieważ znajdowanie maksimum funkcji będącej iloczynem nie jest zadaniem przyjemnym (choćby obliczanie pochodnych iloczynu funkcji jest trudniejsze od sumy), więc wprowadzamy zanegowaną logarytmiczną funkcję wiarygodności

$$\ell(D | \theta) = -\log p(D | \theta) = -\sum_{i=1}^n \log p(\mathbf{x}_i | \theta),$$

wówczas ze względu na fakt, iż funkcja  $\log x$  jest ściśle rosnąca estymatę MLE możemy równoważnie wyznaczyć jako

$$\ell(D | \theta_{\text{MLE}}) = \min_{\theta \in \Theta} \ell(D | \theta).$$

Funkcję  $\ell$  będziemy również nazywać funkcją kosztu.

W przypadku estymaty MAP (z ang. *Maximum a posteriori estimate*) wprowadzamy gęstość rozkładu a posteriori jako

$$p(\theta | D) = \frac{1}{Z} p(D | \theta) \pi(\theta),$$

gdzie  $Z$  jest stałą wynikającą z warunku unormowania, a  $\pi(\theta)$  to gęstość prawdopodobieństwa opisująca rozkład a priori parametru  $\theta$ . Estymatą MAP nazywamy taką wartość parametru  $\theta_{\text{MAP}} \in \Theta$ , że

$$p(\theta_{\text{MAP}} | D) = \max_{\theta \in \Theta} p(\theta | D).$$

Zauważmy przy tym iż liczba  $Z$  nie jest nam potrzebna, gdyż wystarczy zmaksymalizować licznik tj.

$$\theta_{\text{MAP}} = \arg \max_{\theta \in \Theta} p(D | \theta) \pi(\theta).$$



### 0.0.15 Wnioskowanie Bayesowskie

Zajmiemy się teraz wnioskowaniem opartym na twierdzeniu Bayesa. Rozpatrujemy model statystyczny  $\mathcal{P} = \{p(\mathbf{x} \mid \theta) \mid \theta \in \Theta\}$ . Załóżmy, iż mamy obserwacje  $D = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ , wówczas twierdzenie Bayesa możemy zapisać jako

$$p(\theta \mid D) = \frac{p(D \mid \theta)\pi(\theta)}{p_D(D)} = \frac{p(D \mid \theta)\pi(\theta)}{\int_{\Theta} p(D \mid \theta)\pi(\theta) d\theta},$$

gdzie  $p(\theta \mid D)$  nazywamy rozkładem a posteriori (posteriorem),  $p(D \mid \theta)$  – wiarygodnością (likelihood), a  $\pi(\theta)$  – rozkładem a priori (priorem).

Całe wnioskowanie Bayesowskie opiera się na wyznaczeniu rozkładu a posteriori, który wyraża całą naszą wiedzę o estymowanym parametrze  $\theta$ . Na podstawie tego rozkładu możemy wyznaczyć estymatę punktową MAP, jak również niepewność związaną z wyznaczeniem tej estymaty np. poprzez wyznaczenie przedziału wiarygodności  $C_{1-\alpha}(\theta \mid D) = [\theta_l; \theta_u]$  takiego, że

$$P(\theta \in [\theta_l; \theta_u] \mid D) = 1 - \alpha,$$

dla ustalonego  $0 < \alpha < 1$ . Możemy również skonstruować rozkład predykcyjny (z ang. *posterior predictive distribution*) określający prawdopodobieństwo zaobserwowania nowej obserwacji  $\mathbf{x}$

$$p(\mathbf{x} \mid D) = \int_{\Theta} p(\mathbf{x} \mid \theta)p(\theta \mid D) d\theta.$$

Znając rozkład a posteriori estymowanego parametru  $\theta$  możemy nie tylko wyznaczyć estymaty punktowe, wartości oczekiwane i przedziały wiarygodności, ale również znaleźć estymator Bayesa (z ang. *Bayes estimator*), który minimalizuje wartość oczekiwaną pewnej funkcji kosztu (z ang. *loss/cost function*)  $L(\theta, \hat{\theta})$  po wszystkich estymatorach  $\hat{\theta}$

$$\theta_{\text{Bayes}} = \arg \min_{\hat{\theta}} \int_{\Theta} L(\theta, \hat{\theta})p(\theta \mid D) d\theta.$$

### 0.0.16 Modele Gaussowskie

Jak już wspomnieliśmy w przypadku gdy zmienna losowa ma wielowymiarowy rozkład normalny  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  wszystkie rozkłady brzegowe i warunkowe są również rozkładami normalnymi. W szczególnym przypadku gdy zmienne  $k$  i  $n - k$  – wymiarowe  $\mathbf{x}$  i  $\mathbf{y}$  mają łącznie rozkład normalny

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

gdzie

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\mu}_{\mathbf{y}} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}} \\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}} \end{bmatrix}$$

można pokazać iż

$$\mathbf{x} \mid \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}), \quad \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{y}\mathbf{y}}),$$

gdzie

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}} &= \boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\Sigma}_{\mathbf{xy}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}) \\ \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} &= \boldsymbol{\Sigma}_{\mathbf{xx}} - \boldsymbol{\Sigma}_{\mathbf{xy}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \boldsymbol{\Sigma}_{\mathbf{yx}} \end{aligned}.$$

### 0.0.17 Liniowe modele Gaussowskie

Powyższe własności rozkładów łącznych pozwalają jawnie wnioskować w tzw. liniowych modelach Gaussowskich (z ang. *Linear Gaussian Models*). Załóżmy, iż nasze obserwacje są modelowane przez  $n$ -wymiarową zmienną losową  $\mathbf{y}$  o rozkładzie normalnym z estymowanym parametrem  $\mathbf{x}$  i znanymi parametrami  $\mathbf{A}, \mathbf{b}, \boldsymbol{\Sigma}_{\mathbf{y}}$  tak, że wiarygodność ma postać

$$\mathbf{y} \mid \mathbf{x} \sim \mathcal{N}(\mathbf{Ax} + \mathbf{b}, \boldsymbol{\Sigma}_{\mathbf{y}}),$$

gdzie  $\mathbf{A}$  jest macierzą wymiaru  $n \times k$ . Jako prior na parametr  $\mathbf{x}$  przyjmujemy również rozkład normalny o pewnych zadanych parametrach  $\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}$  (taki wybór rozkładu a priori nazywamy rozkładem sprzężonym do wiarygodności)

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}).$$

Wówczas łatwo pokazać, iż rozkład a posteriori jest rozkładem normalnym

$$\mathbf{x} \mid \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}})$$

z parametrami

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} &= \left[ \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} + \mathbf{A}^{\top} \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{A} \right]^{-1} \\ \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}} &= \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} \left[ \mathbf{A}^{\top} \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} (\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \boldsymbol{\mu}_{\mathbf{x}} \right] \end{aligned}.$$

Założmy teraz, iż mamy ciąg obserwacji  $(\mathbf{y}_1, \dots, \mathbf{y}_m)$ . Wnioskowanie Bayesowskie możemy wówczas stosować iteracyjnie tzn. na początku dla 0 obserwacji rozkład estymowanego parametru jest opisany przez prior  $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ . Po zaobserwowaniu jednego  $\mathbf{y}_1$  aktualizujemy nasze przekonania co do parametru  $\mathbf{x}$  zgodnie z powyższym wzorem i otrzymujemy rozkład normalny o parametrach

$$\begin{aligned} \boldsymbol{\Sigma}_1 &= \left[ \boldsymbol{\Sigma}_0^{-1} + \mathbf{A}^{\top} \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{A} \right]^{-1} \\ \boldsymbol{\mu}_1 &= \boldsymbol{\Sigma}_1 \left[ \mathbf{A}^{\top} \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} (\mathbf{y}_1 - \mathbf{b}) + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right] \end{aligned}.$$

Po zaobserwowaniu kolejnego  $\mathbf{y}_2$  ponownie wykorzystujemy powyższe wzory ale jako prior wykorzystując rozkład w poprzedniej iteracji. W ogólności możemy

zapisać wzór rekurencyjny na  $m + 1$  rozkład jako

$$\begin{aligned}\boldsymbol{\Sigma}_{m+1} &= \left[ \boldsymbol{\Sigma}_m^{-1} + \mathbf{A}^\top \boldsymbol{\Sigma}_y^{-1} \mathbf{A} \right]^{-1} \\ \boldsymbol{\mu}_{m+1} &= \boldsymbol{\Sigma}_{m+1} \left[ \mathbf{A}^\top \boldsymbol{\Sigma}_y^{-1} (\mathbf{y}_{m+1} - \mathbf{b}) + \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\mu}_m \right],\end{aligned}$$

skąd możemy od razu podać wzór na parametry  $m$ -tego rozkładu

$$\begin{aligned}\boldsymbol{\Sigma}_m &= \left[ \boldsymbol{\Sigma}_0^{-1} + m \mathbf{A}^\top \boldsymbol{\Sigma}_y^{-1} \mathbf{A} \right]^{-1} \\ \boldsymbol{\mu}_m &= \boldsymbol{\Sigma}_m \left[ \mathbf{A}^\top \boldsymbol{\Sigma}_y^{-1} \left( \sum_{i=1}^m \mathbf{y}_i - m \mathbf{b} \right) + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right].\end{aligned}$$

Taki sam wynik można by uzyskać rozpatrując łączny rozkład a posteriori dla obserwacji  $D = (\mathbf{y}_1, \dots, \mathbf{y}_m)$  tj.

$$\begin{aligned}p(\mathbf{x} \mid D) &\cong \pi(\mathbf{x}) \prod_{i=1}^m p(\mathbf{y}_i \mid \mathbf{x}) \cong \\ &\exp \left\{ -\frac{1}{2} \left[ (\mathbf{x} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\mathbf{x} - \boldsymbol{\mu}_0) + \sum_{i=1}^m (\mathbf{y}_i - \mathbf{A}\mathbf{x} - \mathbf{b})^\top \boldsymbol{\Sigma}_y^{-1} (\mathbf{y}_i - \mathbf{A}\mathbf{x} - \mathbf{b}) \right] \right\}.\end{aligned}$$

### 0.0.18 Regresja liniowa

Założmy, iż modelujemy obserwacje postaci  $(y, \mathbf{x})$  gdzie  $y$  to skalar zwany zmienną objaśnianą, którego wartość obserwujemy, a  $\mathbf{x}$  to wektor zmiennych objaśniających, który kontrolujemy tj. zakładamy, iż wektor  $\mathbf{x}$  dla danego pomiaru  $y$  znamy dokładnie. Dodatkowo zakładamy, iż  $y$  zależy liniowo od  $\mathbf{x}$  tj.

$$y = \mathbf{w}^\top \mathbf{x} + \epsilon,$$

gdzie  $\epsilon \sim \mathcal{N}(0, \sigma^2)$  dla znanego  $\sigma$  jest tzw. błędem losowym, a  $\mathbf{w}$  jest estymowanym przez nas parametrem. Możemy zatem zapisać

$$y \mid \mathbf{w} \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}, \sigma^2).$$

Powiedzmy, iż zaobserwowaliśmy ciąg obserwacji  $D = (y_1, \dots, y_m)$  dla zadanych (lub dokładnie znanych) przez nas  $(\mathbf{x}_1, \dots, \mathbf{x}_m)$ . Wiarygodność ma zatem postać

$$p(D \mid \mathbf{w}) \cong \prod_{i=1}^m \exp \left\{ -\frac{1}{2\sigma^2} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \right\}.$$

W przypadku regresji liniowej zamiast pełnego wnioskowania Bayesowskiego o parametrze  $\mathbf{w}$  często stosuje się prostsze podejście polegające na znalezieniu estymaty punktowej MLE. Zanegowana logarytmiczna funkcja wiarygodności ma postać

$$\ell(D \mid \mathbf{w}) = \frac{1}{2\sigma^2} \sum_{i=1}^m (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \text{const.}$$

Człon stały możemy oczywiście pominąć i zapisać

$$\ell(D \mid \mathbf{w}) \cong \sum_{i=1}^m (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 = (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) ,$$

gdzie

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_m^\top \end{bmatrix} .$$

Ponieważ otrzymana funkcja  $\ell$  ma postać formy kwadratowej, więc problem optymalizacyjny polegający na znalezieniu minimum  $\ell$  nazywa się metodą najmniejszych kwadratów (z ang. *OLS – Ordinary Least Squares*). Aby wyznaczyć estymatę  $\mathbf{w}_{\text{MLE}}$  musimy rozwiązać równanie

$$\frac{\partial \ell}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} [\mathbf{y}^\top \mathbf{y} + \mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w} - 2\mathbf{y}^\top \mathbf{X} \mathbf{w}] = \mathbf{0} ,$$

skąd

$$2\mathbf{X}^\top \mathbf{X} \mathbf{w} - 2\mathbf{X}^\top \mathbf{y} = \mathbf{0} ,$$

zatem

$$\mathbf{w}_{\text{MLE}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} .$$

Pełniejszą informację o parametrze  $\mathbf{w}$  możemy uzyskać rozpatrując rozkład a posteriori  $p(\mathbf{w} \mid D)$ . Jeśli jako prior przyjmiemy rozkład normalny z pewnymi parametrami  $\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$  to zauważmy, iż otrzymujemy instancję liniowego modelu Gaussowskiego

$$\begin{aligned} \mathbf{y} \mid \mathbf{w} &\sim \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{1}) \\ \mathbf{w} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \end{aligned} ,$$

skąd rozkład a posteriori jest rozkładem normalnym

$$\mathbf{w} \mid \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

o parametrach

$$\begin{aligned} \boldsymbol{\Sigma}_m &= [\boldsymbol{\Sigma}_0^{-1} + \sigma^{-2} \mathbf{X}^\top \mathbf{X}]^{-1} \\ \boldsymbol{\mu}_m &= \boldsymbol{\Sigma}_m [\sigma^{-2} \mathbf{X}^\top \mathbf{y} + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0] \end{aligned} .$$

W powyższych wzorach nazwy parametrów nie są przykładowe: po zaobserwowaniu 0 przykładów rozkład parametru  $\mathbf{w}$  jest rozkładem a priori  $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ ; po zaobserwowaniu po jednej wartości  $y_i$  w  $m$  zadanych (znanych dokładnie) punktach  $\mathbf{x}_i$  otrzymujemy rozkład a posteriori  $\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ . Gdybyśmy w każdym z  $m$  punktów  $\mathbf{x}_i$  dokonywali pomiaru  $y_i$   $s$ -krotnie to wtedy wykorzystując

wzory wyprowadzone przy iteracyjnym stosowaniu wnioskowania w liniowym modelu Gaussowskim otrzymujemy rozkład normalny o parametrach

$$\begin{aligned}\boldsymbol{\Sigma}_{m;s} &= \left[ \boldsymbol{\Sigma}_0^{-1} + \frac{s}{\sigma^2} \mathbf{X}^\top \mathbf{X} \right]^{-1} \\ \boldsymbol{\mu}_{m;s} &= \boldsymbol{\Sigma}_{m;s} \left[ \sigma^{-2} \mathbf{X}^\top \sum_{i=1}^s \mathbf{y}_i + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right] .\end{aligned}$$

Rozkład predykcyjny dla nowej obserwacji  $y$  poczynionej w punkcie  $\mathbf{x}$  jest dany przez

$$p(y | \mathbf{y}) = \int_{\mathbb{R}^n} p(y | \mathbf{w}) p(\mathbf{w} | \mathbf{y}) d^n \mathbf{w} .$$

Nietrudno zauważyć, iż będzie to rozkład normalny o parametrach

$$\begin{aligned}\mu_{y|\mathbf{y}} = \mathbb{E}[y | \mathbf{y}] &= \int_{\mathbb{R}} y p(y | \mathbf{y}) dy = \int_{\mathbb{R}^n} d^n \mathbf{w} p(\mathbf{w} | \mathbf{y}) \int_{\mathbb{R}} dy y p(y | \mathbf{w}) \\ &= \int_{\mathbb{R}^n} d^n \mathbf{w} p(\mathbf{w} | \mathbf{y}) \mathbf{x}^\top \mathbf{w} = \mathbf{x}^\top \boldsymbol{\mu}_m .\end{aligned}$$

oraz

$$\begin{aligned}\sigma_{y|\mathbf{y}}^2 = \mathbb{E}[(y - \mu_{y|\mathbf{y}})^2 | \mathbf{y}] &= \int_{\mathbb{R}^n} d^n \mathbf{w} p(\mathbf{w} | \mathbf{y}) \int_{\mathbb{R}} dy (y - \mu_{y|\mathbf{y}})^2 p(y | \mathbf{w}) \\ &= \int_{\mathbb{R}^n} d^n \mathbf{w} p(\mathbf{w} | \mathbf{y}) \int_{\mathbb{R}} dy (y^2 + \mu_{y|\mathbf{y}}^2 - 2\mu_{y|\mathbf{y}} y) p(y | \mathbf{w}) \\ &= \int_{\mathbb{R}^n} d^n \mathbf{w} p(\mathbf{w} | \mathbf{y}) (\sigma^2 + (\mathbf{x}^\top \mathbf{w})^2 + \mu_{y|\mathbf{y}}^2 - 2\mu_{y|\mathbf{y}} \mathbf{x}^\top \mathbf{w}) \\ &= \sigma^2 + \int_{\mathbb{R}^n} d^n \mathbf{w} p(\mathbf{w} | \mathbf{y}) (\mathbf{x}^\top \mathbf{w} - \mathbf{x}^\top \boldsymbol{\mu}_m)^2 \\ &= \sigma^2 + \mathbf{x}^\top \mathbb{E}[(\mathbf{w} - \boldsymbol{\mu}_m)(\mathbf{w} - \boldsymbol{\mu}_m)^\top | \mathbf{y}] \mathbf{x} = \sigma^2 + \mathbf{x}^\top \boldsymbol{\Sigma}_m \mathbf{x} .\end{aligned}$$

Powyżej skorzystaliśmy ze znanego faktu, iż dla jednowymiarowej zmiennej losowej zachodzi  $\sigma^2 = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}[X^2] - \mu_X^2$ , skąd  $\mathbb{E}[X^2] = \sigma^2 + \mu_X^2$ . Podsumowując rozkład predykcyjny ma postać

$$y | \mathbf{y} \sim \mathcal{N}(\mathbf{x}^\top \boldsymbol{\mu}_m, \sigma^2 + \mathbf{x}^\top \boldsymbol{\Sigma}_m \mathbf{x}) .$$

### 0.0.19 Regularyzacja

Regularyzacją nazywamy proces polegający na wprowadzeniu ad hoc do zagadnienia optymalizacji dodatkowych członów tak, aby rozwiązanie było „reguluarne” (prostsze, nieosobliwe, jednoznaczne ...). W przypadku funkcji kosztu  $\ell$

najczęściej dodajemy człon penalizujący rozwiązania o dużej normie estymowanego parametru postaci

$$\gamma \|\theta\|$$

dla pewnej normy  $\|\cdot\|$  i hiper-parametru  $\gamma$  określającego siłę regularyzacji. W kontekście Bayesowskim regularyzację można również rozumieć jako pewną niechęć („tłumienie”, zachowawczość) modelu do zmiany rozkładu a priori estymowanego parametru po pojawieniu się kolejnych obserwacji.

Przykładowo jeśli w zagadnieniu Bayesowskiej regresji liniowej jako prior przyjmujemy rozkład normalny

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{1})$$

to rozkład a posteriori jest rozkładem normalnym o parametrach

$$\begin{aligned} \Sigma_m &= \sigma^2 \left[ \gamma \mathbf{1} + \mathbf{X}^\top \mathbf{X} \right]^{-1} \\ \mu_m &= \left[ \gamma \mathbf{1} + \mathbf{X}^\top \mathbf{X} \right]^{-1} \mathbf{X}^\top \mathbf{y} \end{aligned}$$

gdzie  $\gamma = \sigma^2/\tau^2$  jest hiper-parametrem określającym siłę regularyzacji. Zauważmy, że im większa jest wartość  $\gamma$  (mniejsza niepewność związana z rozkładem a priori) tym drugi człon w nawiasie staje się mniej istotny. Taki sam wynik możemy uzyskać metodą OLS jeśli do funkcji kosztu dodamy człon regularyzujący dla zwykłej normy euklidesowej. Zagadnienie minimalizacji funkcji kosztu będącej formą kwadratową z dodanym członem regularyzującym nazywamy również regresją grzbietową.

### 0.0.20 Procesy Gaussowskie

Jak już wspomnieliśmy macierz kowariancji  $n$ -wymiarowej zmiennej losowej  $\mathbf{x}$  o wartości oczekiwanej  $\mu$  jest zdefiniowana jako

$$\Sigma = \mathbb{E} [(\mathbf{x} - \mu)(\mathbf{x} - \mu)^\top] .$$

Pokazaliśmy również, iż macierz ta jest nieujemnie określona. Dodatkowo pokażemy, iż dla każdej nieujemnie określonej macierzy symetrycznej  $\mathbf{K}$  wymiaru  $n \times n$  istnieje  $n$ -wymiarowa zmienna losowa o wielowymiarowym rozkładzie normalnym dla której  $\mathbf{K}$  jest macierzą kowariancji. Istotnie dla każdej nieujemnie określonej macierzy symetrycznej istnieje macierz  $\mathbf{L}$  taka, że

$$\mathbf{K} = \mathbf{L}\mathbf{L}^\top ,$$

jest to tzw. dekompozycja Choleskiego. Niech  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ , wówczas zmienna losowa  $\mathbf{Lz}$  ma rozkład o zerowej wartości oczekiwanej i macierzy kowariancji

$$\mathbb{E} [(\mathbf{Lz})(\mathbf{Lz})^\top] = \mathbb{E} [\mathbf{Lzz}^\top \mathbf{L}^\top] = \mathbf{L} \mathbb{E} [\mathbf{zz}^\top] \mathbf{L}^\top = \mathbf{L}\mathbf{1}\mathbf{L}^\top = \mathbf{K} .$$

Powyższe własności wskazują, iż macierze kowariancji można w pewnym sensie utożsamiać z nieujemnie określonymi macierzami symetrycznymi.

Zdefiniujemy teraz funkcję kowariancji  $k : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$  taką, że  $\forall m \in \mathbb{N} :$   
 $\forall X = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset \mathbb{R}^n$  macierz

$$k(X, X) = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \cdots & k(\mathbf{x}_1, \mathbf{x}_m) \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \cdots & k(\mathbf{x}_2, \mathbf{x}_m) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_m, \mathbf{x}_1) & k(\mathbf{x}_m, \mathbf{x}_2) & \cdots & k(\mathbf{x}_m, \mathbf{x}_m) \end{bmatrix}$$

jest dodatnio określoną macierzą symetryczną. Funkcję  $k$  nazywamy również jądrem dodatnio określonym (z ang. *positive definite kernel*) lub jądrem Mercera. Dla dwóch zbiorów punktów  $X = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset \mathbb{R}^n$  i  $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_s\} \subset \mathbb{R}^n$  i funkcji kowariancji  $k$  wprowadzimy oznaczenie

$$k(X, Y) := \begin{bmatrix} k(\mathbf{x}_1, \mathbf{y}_1) & k(\mathbf{x}_1, \mathbf{y}_2) & \cdots & k(\mathbf{x}_1, \mathbf{y}_s) \\ k(\mathbf{x}_2, \mathbf{y}_1) & k(\mathbf{x}_2, \mathbf{y}_2) & \cdots & k(\mathbf{x}_2, \mathbf{y}_s) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_m, \mathbf{y}_1) & k(\mathbf{x}_m, \mathbf{y}_2) & \cdots & k(\mathbf{x}_m, \mathbf{y}_s) \end{bmatrix}.$$

Poniżej podajemy kilka przykładów funkcji kowariancji

- *Gaussian kernel* dla normy  $\|\cdot\|$  i hiper-parametru  $l$

$$k(\mathbf{x}, \mathbf{y}) = \exp \left\{ -\frac{1}{2l^2} \|\mathbf{x} - \mathbf{y}\|^2 \right\}$$

- *Periodic kernel* dla normy  $\|\cdot\|$  i hiper-parametrów  $l, p$

$$k(\mathbf{x}, \mathbf{y}) = \exp \left\{ -\frac{2}{l^2} \sin^2 \left( \frac{\pi}{p} \|\mathbf{x} - \mathbf{y}\| \right) \right\}$$

- *White noise kernel* dla hiper-parametru  $\sigma$

$$k(\mathbf{x}, \mathbf{y}) = \sigma^2 \delta_{\mathbf{x}, \mathbf{y}}$$

- *Matérn kernel* dla normy  $\|\cdot\|$  i hiper-parametrów  $l, \nu$

$$k(\mathbf{x}, \mathbf{y}) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{\sqrt{2\nu}}{l} \|\mathbf{x} - \mathbf{y}\| \right)^\nu K_\nu \left( \frac{\sqrt{2\nu}}{l} \|\mathbf{x} - \mathbf{y}\| \right),$$

gdzie  $\Gamma(x)$  to funkcja gamma Eulera, a  $K_\nu(x)$  to zmodyfikowana funkcja Bessela 2-go rodzaju rzędu  $\nu$ .

Dodatkowo suma lub iloczyn dwóch funkcji kowariancji oraz złożenie funkcji kowariancji z wielomianem o nieujemnych współczynnikach jest również funkcją kowariancji.

Procesem Gaussowskim (z ang. *Gaussian Process*) nazywamy rodzinę skalarnych zmiennych losowych indeksowanych przez punkty  $\mathbf{x} \in \mathbb{R}^n$

$$\mathcal{GP} = \{f_{\mathbf{x}} \mid \mathbf{x} \in \mathbb{R}^n\}$$

taką że każdy skończony podzbiór  $\mathcal{GP}$  ma łącznie wielowymiarowy rozkład normalny tj. dla dowolnego zbioru  $X = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset \mathbb{R}^n$  zachodzi

$$\begin{bmatrix} f_{\mathbf{x}_1} \\ \vdots \\ f_{\mathbf{x}_m} \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X).$$

Zauważmy, iż process Gaussowski możemy jednoznacznie zdefiniować podając „przepisy” na parametry  $\boldsymbol{\mu}_X$  i  $\boldsymbol{\Sigma}_X$  dla dowolnego zbioru  $X$ . W praktyce często przyjmujemy  $\boldsymbol{\mu}_X = \mathbf{0}$ , natomiast przepisem na macierz kowariancji może być zdefiniowana wyżej funkcja kowariancji  $k(X, X)$  tj.

$$\begin{bmatrix} f_{\mathbf{x}_1} \\ \vdots \\ f_{\mathbf{x}_m} \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, k(X, X)).$$

Process Gaussowski daje nam w praktyce rozkład prawdopodobieństwa nad funkcjami  $f : \mathbb{R}^n \mapsto \mathbb{R}$ , których charakter jest określony przez jądro  $k$  (np. funkcja gładka dla jądra Gaussowskiego, okresowa dla jądra periodycznego, itp.). Zauważmy, że nie wnioskujemy tu o parametrach konkretnej rodziny funkcji (jak w przypadku regresji liniowej); interesuje nas jedynie rozkład predykcyjny. Załóżmy, iż w zadanych (lub dokładnie znanych) przez nas punktach  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$  zaobserwowaliśmy wartości pewnej funkcji, o których zakładamy, iż pochodzą z procesu Gaussowskiego zadanego jądrem  $k$ , które wyraża nasze założenia a priori co do charakteru badanej funkcji

$$\mathbf{f}_X = \begin{bmatrix} f_{\mathbf{x}_1} \\ \vdots \\ f_{\mathbf{x}_m} \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, k(X, X)).$$

Powiedzmy, iż chcemy znać wartości  $\mathbf{f}_Y$  tej funkcji w zadanych punktach  $Y = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_s\}$ . Ponieważ założyliśmy, iż wartości funkcji pochodzą z procesu Gaussowskiego, więc rozkład łączny  $\mathbf{f}_X$  i  $\mathbf{f}_Y$  jest rozkładem normalnym

$$\begin{bmatrix} \mathbf{f}_X \\ \mathbf{f}_Y \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} k(X, X) & k(X, Y) \\ k(Y, X) & k(Y, Y) \end{bmatrix}\right).$$

Zauważmy, iż jest to instancja modelu Gaussowskiego, więc rozkład warunkowy  $\mathbf{f}_Y \mid \mathbf{f}_X$  jest również rozkładem normalnym o parametrach

$$\begin{aligned} \boldsymbol{\mu} &= k(Y, X)k^{-1}(X, X)\mathbf{f}_X \\ \boldsymbol{\Sigma} &= k(Y, Y) - k(Y, X)k^{-1}(X, X)k(X, Y) \end{aligned}$$



Dodatkową niepewność związaną z pomiarem wartości  $\mathbf{f}_X$  możemy uchwycić zmieniając postać jądra

$$k(\mathbf{x}, \mathbf{y}) \leftarrow k(\mathbf{x}, \mathbf{y}) + \mathcal{I}_X(\mathbf{x}) \sigma^2 \delta_{\mathbf{x}, \mathbf{y}},$$

gdzie  $\sigma$  jest hiper-parametrem określającym precyzję pomiaru. Oczywiście  $k$  jest dalej funkcją kowariancji, gdyż takie podstawienie powoduje jedynie dodanie dodatnich członów do pewnych elementów diagonalnych macierzy kowariancji, więc macierz ta jest nadal symetryczna i dodatnio określona. Wówczas rozkład predykcyjny ma parametry

$$\begin{aligned} \boldsymbol{\mu} &= k(Y, X) [k(X, X) + \sigma^2 \mathbf{1}]^{-1} \mathbf{f}_X \\ \boldsymbol{\Sigma} &= k(Y, Y) - k(Y, X) [k(X, X) + \sigma^2 \mathbf{1}]^{-1} k(X, Y) \end{aligned}.$$

#### 0.0.21 Wieloklasowa regresja logistyczna

Założmy, iż modelujemy obserwacje postaci  $(t, \mathbf{x})$ , gdzie  $t \in \{\tau_1, \tau_2, \dots, \tau_s\}$  to etykieta określająca przynależność do jednej z  $s$  klas, a  $\mathbf{x} \in \mathbb{R}^n$  jest znanym (lub zadany) przez nas dokładnie wektorem cech obiektu dla których zaobserwowaną klasą jest  $t$ . Zakładamy ponadto, iż prawdopodobieństwo przynależności do klasy  $\tau_j$  (jednej z  $s$  klas) dla wektora cech  $\mathbf{x}$  ma postać tzw. funkcji softmax

$$\pi_j(\mathbf{x}) = \frac{1}{Z(\mathbf{x})} e^{\mathbf{w}_j^\top \mathbf{x}},$$

gdzie  $\mathbf{w}_j$  są estymowanymi przez nas parametrami. Ze względu na warunek unormowania musimy mieć

$$\sum_{j=1}^s \pi_j = 1,$$

skąd stała normalizacyjna  $Z(\mathbf{x})$  ma postać

$$Z(\mathbf{x}) = \sum_{j=1}^s e^{\mathbf{w}_j^\top \mathbf{x}}.$$

Rozkład zmiennej losowej  $t$  jest w takim razie dyskretnym rozkładem wielopunktowym (z ang. *categorical distribution*) postaci

$$t \mid \mathbf{w}_1, \dots, \mathbf{w}_s \sim \text{Cat}(\pi_1(\mathbf{x}), \dots, \pi_s(\mathbf{x})).$$

Zauważmy, iż prawdopodobieństwo wylosowania etykiety  $t$  dla parametrów  $\mathbf{w}_j$  możemy zapisać jako

$$p(t \mid \mathbf{w}_1, \dots, \mathbf{w}_s) = \prod_{j=1}^s \pi_j(\mathbf{x})^{\delta(t, \tau_j)}.$$

Powiedzmy, że mamy obserwacje  $D = (t_1, \dots, t_m)$  dla znanych (lub zadanych) przez nas dokładnie wektorów cech  $(\mathbf{x}_1, \dots, \mathbf{x}_m)$ . Funkcja wiarygodności ma wówczas postać

$$p(D \mid \mathbf{w}_1, \dots, \mathbf{w}_s) = \prod_{i=1}^m p(t_i \mid \mathbf{w}_1, \dots, \mathbf{w}_s) = \prod_{i=1}^m \prod_{j=1}^s \pi_j(\mathbf{x}_i)^{\delta(t_i, \tau_j)}.$$

Jako prior dla parametrów  $\mathbf{w}_j$  przyjmujemy rozkład normalny z pewnym hiperparametrem  $\gamma$

$$\forall j \in \{1, \dots, s\} : \mathbf{w}_j \sim \mathcal{N}(\mathbf{0}, \gamma^{-1} \mathbf{1}).$$

W przypadku regresji logistycznej ograniczymy się do znalezienia estymaty MAP parametrów  $\mathbf{w}_j$  tak, aby w przyszłości do nowego wektora cech  $\mathbf{x}$  przyporządkować klasę o największym prawdopodobieństwie  $\pi_j(\mathbf{x})$ . Znalezienie estymaty MAP sprowadza się do znalezienia minimum zregulowanej funkcji kosztu

$$\begin{aligned} \ell^*(D \mid \mathbf{w}_1, \dots, \mathbf{w}_s) &= -\log[p(D \mid \mathbf{w}_1, \dots, \mathbf{w}_s) \pi(\mathbf{w}_1, \dots, \mathbf{w}_s)] \\ &= -\log \left[ \prod_{k=1}^s e^{-\frac{\gamma}{2} \mathbf{w}_k^\top \mathbf{w}_k} \prod_{i=1}^m \prod_{j=1}^s \pi_j(\mathbf{x}_i)^{\delta(t_i, \tau_j)} \right] \\ &= \frac{\gamma}{2} \sum_{j=1}^s \mathbf{w}_j^\top \mathbf{w}_j - \sum_{i=1}^m \sum_{j=1}^s \delta(t_i, \tau_j) \log \pi_j(\mathbf{x}_i). \end{aligned}$$

Niestety dla tak zdefiniowanej funkcji kosztu nie można znaleźć wzoru na minimum w postaci analitycznej, dlatego wykorzystamy numeryczny algorytm optymalizacji zwany spadkiem wzdłuż gradientu.

### 0.0.22 Metoda spadku wzdłuż gradientu

Algorytm spadku wzdłuż gradientu (z ang. *gradient descent*) dla funkcji  $f(\mathbf{x}_1, \dots, \mathbf{x}_m)$  ma postać

1. Wybierz parametry początkowe  $\mathbf{x}_1^{(0)}, \dots, \mathbf{x}_m^{(0)}$
2. Powtarzaj

$$\begin{aligned} \mathbf{x}_1^{(t+1)} &= \mathbf{x}_1^{(t)} - \epsilon_1 \frac{\partial f}{\partial \mathbf{x}_1} \Big|_{\mathbf{x}_1^{(t)}, \dots, \mathbf{x}_m^{(t)}} \\ &\vdots \\ \mathbf{x}_m^{(t+1)} &= \mathbf{x}_m^{(t)} - \epsilon_m \frac{\partial f}{\partial \mathbf{x}_m} \Big|_{\mathbf{x}_1^{(t)}, \dots, \mathbf{x}_m^{(t)}} \end{aligned}$$

gdzie  $\epsilon_1, \dots, \epsilon_m$  to hiperparametry zwane stałymi uczącymi (z ang. *learning rate*).

Zakładając  $\epsilon_1 = \dots = \epsilon_m = \epsilon$  i wprowadzając

$$\mathbf{X} := \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_m^\top \end{bmatrix}, \quad \frac{\partial f}{\partial \mathbf{X}} := \begin{bmatrix} \frac{\partial f}{\partial \mathbf{x}_1}^\top \\ \vdots \\ \frac{\partial f}{\partial \mathbf{x}_m}^\top \end{bmatrix}$$

możemy zapisać powyższe równania w kompaktowej formie

$$\mathbf{X}^{(t+1)} = \mathbf{X}^{(t)} - \epsilon \frac{\partial f}{\partial \mathbf{X}} \Big|_{\mathbf{X}^{(t)}}.$$

---

Aby zminimalizować numerycznie funkcję kosztu  $\ell^*$  stosując metodę spadku wzdłuż gradientu musimy obliczyć pochodne funkcji kosztu po parametrach  $\mathbf{w}_j$

$$\frac{\partial \ell^*}{\partial \mathbf{w}_k} = \gamma \mathbf{w}_k - \sum_{i=1}^m \sum_{j=1}^s \delta(t_i, \tau_j) \frac{\partial}{\partial \mathbf{w}_k} \log \pi_j(\mathbf{x}_i),$$

ale

$$\begin{aligned} \frac{\partial}{\partial \mathbf{w}_k} \log \pi_j(\mathbf{x}_i) &= \frac{1}{\pi_j(\mathbf{x}_i)} \frac{Z(\mathbf{x}_i) \frac{\partial e^{\mathbf{x}_i^\top \mathbf{w}_j}}{\partial \mathbf{w}_k} - e^{\mathbf{x}_i^\top \mathbf{w}_j} \frac{\partial Z(\mathbf{x}_i)}{\partial \mathbf{w}_k}}{Z^2(\mathbf{x}_i)} \\ &= \frac{Z(\mathbf{x}_i)}{e^{\mathbf{x}_i^\top \mathbf{w}_j}} \frac{Z(\mathbf{x}_i) \mathbf{x}_i e^{\mathbf{x}_i^\top \mathbf{w}_k} \delta_{jk} - e^{\mathbf{x}_i^\top \mathbf{w}_j} e^{\mathbf{x}_i^\top \mathbf{w}_k} \mathbf{x}_i}{Z^2(\mathbf{x}_i)} \\ &= \mathbf{x}_i \delta_{jk} - \mathbf{x}_i \pi_k(\mathbf{x}_i) \end{aligned}$$

zatem

$$\frac{\partial \ell^*}{\partial \mathbf{w}_k} = \gamma \mathbf{w}_k - \sum_{i=1}^m \mathbf{x}_i \sum_{j=1}^s \delta(t_i, \tau_j) \delta_{jk} + \sum_{i=1}^m \mathbf{x}_i \pi_k(\mathbf{x}_i) \sum_{j=1}^s \delta(t_i, \tau_j).$$

Zauważmy jednak, iż

$$\sum_{j=1}^s \delta(t_i, \tau_j) = 1, \quad \sum_{j=1}^s \delta(t_i, \tau_j) \delta_{jk} = \delta(t_i, \tau_k),$$

zatem ostatecznie

$$\frac{\partial \ell^*}{\partial \mathbf{w}_k} = \gamma \mathbf{w}_k + \sum_{i=1}^m \mathbf{x}_i [\pi_k(\mathbf{x}_i) - \delta(t_i, \tau_k)].$$

Wprowadzając macierze

$$\begin{aligned} \mathbf{X} &= \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_m^\top \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} \mathbf{w}_1^\top \\ \vdots \\ \mathbf{w}_s^\top \end{bmatrix}, \\ \mathbf{S} &= \begin{bmatrix} \pi_1(\mathbf{x}_1) & \cdots & \pi_s(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \pi_1(\mathbf{x}_m) & \cdots & \pi_s(\mathbf{x}_m) \end{bmatrix}, \quad \mathbf{T} = \begin{bmatrix} \delta(t_1, \tau_1) & \cdots & \delta(t_1, \tau_s) \\ \vdots & \ddots & \vdots \\ \delta(t_m, \tau_1) & \cdots & \delta(t_m, \tau_s) \end{bmatrix} \end{aligned}$$

możemy w takim razie zapisać zdefiniowaną wyżej macierz pochodnych wymaganych do algorytmu spadku wzdłuż gradient w kompaktowej formie jako

$$\frac{\partial \ell^*}{\partial \mathbf{W}} = (\mathbf{S} - \mathbf{T})^\top \mathbf{X}.$$

Zauważmy, iż zregularyzowana funkcja kosztu rośnie wraz ze wzrostem liczby obserwacji  $m$ . Wynika z tego, iż stała ucząca musi być zależna od liczby przykładów. Możemy na przykład stwierdzić, iż  $\epsilon \leftarrow m^{-1}\epsilon$  i wówczas minimalizujemy tak naprawdę średni koszt  $\ell^*/m$ .

### 0.0.23 Wnioskowanie metodami Monte Carlo

Całe wnioskowanie Bayesowskie opiera się na wyznaczaniu rozkładów a posteriori, które wyrażają naszą wiedzę o estymowanym parametrze. Do tej pory rozważaliśmy modele Bayesowskie dla których prior i wiarygodność były dane przez rozkłady normalne. Dzięki temu mogliśmy wyprowadzić analityczne wzory na parametry rozkładu a posteriori, który również był rozkładem normalnym. Dla wielu interesujących modeli nie jesteśmy jednak w stanie tego zrobić (np. w zagadnieniu regresji logistycznej ograniczyliśmy się jedynie do estymaty punktowej), gdyż obliczenie stałej normalizującej dla rozkładu  $p(\theta \mid D)$  może wymagać obliczenia całki, której nie jesteśmy w stanie wyrazić w sposób jawny lub sumy po wykładniczo wielu elementach. Wnioskowanie Bayesowskie można jednak prowadzić w modelach, w których nie dysponujemy jawnym wzorem na gęstość prawdopodobieństwa rozkładu a posteriori. Okazuje się, iż do generowania próbek z rozkładu  $p(\theta \mid D)$  wystarczy znajomość tego rozkładu z dokładnością do stałej normalizującej, a zatem wystarczy znać rozkład łączny  $p(\theta, D) = p(D \mid \theta)\pi(\theta)$ . Generowanie próbek z kolei wystarczy natomiast, na mocy silnego prawa wielkich liczb, do szacowania wartości średnich dowolnych funkcji estymowanego parametru  $\theta$ . Przypomnijmy, iż na mocy silnego prawa wielkich liczb ciąg średnich częściowych  $(\bar{X}_n)$  ciągu zmiennych losowych  $(X_n)$  i.i.d. z rozkładu  $X \sim \mathcal{D}$  jest zbieżny z prawdopodobieństwem 1 do wartości oczekiwanej  $\mathbb{E}[X]$  tj.

$$P\left(\lim_{n \rightarrow \infty} \bar{X}_n = \mathbb{E}[X]\right) = 1.$$

Wartość oczekiwaną  $\mathbb{E}[X]$  możemy zatem przybliżyć średnią  $\bar{X}_n$  z dużej ilości próbek.

Wnioskowanie Monte Carlo pozwala nam szacować różne wielkości w tzw. hierarchicznych modelach Bayesowskich (z ang. *Bayesian hierarchical modeling*). Rozważmy jeszcze raz przykład regresji liniowej w ujęciu Bayesowskim, ale rozważmy teraz model postaci

$$\begin{aligned} \sigma^2 &\sim \mathcal{D}(\lambda) \\ \mathbf{w} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\ y \mid \mathbf{w}, \sigma^2 &\sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}, \sigma^2) \end{aligned},$$

gdzie  $\lambda, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$  są pewnymi hiper-parametrami. Dla takiego modelu nie możemy w ogólności znaleźć jawnej postaci rozkładu a posteriori. Jeśli jednak umiemy generować próbki z rozkładu łącznego

$$Z \cdot p(\mathbf{w}, \sigma^2 \mid D) = p(D, \mathbf{w}, \sigma^2) = p(D \mid \mathbf{w}, \sigma^2) \pi(\mathbf{w}) \pi(\sigma^2)$$

to wszystkie interesujące wielkości możemy oszacować jako odpowiednie średnie. Pozostaje pytanie w jaki sposób generować próbki ze skomplikowanych rozkładów prawdopodobieństwa, których gęstości znamy jedynie z dokładnością do stałej normalizującej. Poniżej przedstawimy dwa algorytmy próbkowania: algorytm IS oraz Metropolisa–Hastingsa będący szczególną realizacją całej rodziny algorytmów próbkowania zwanych Markov Chain Monte Carlo (MCMC).

#### 0.0.24 Algorytm Importance Sampling (IS)

Założmy, iż chcemy obliczyć wartość oczekiwaną pewnej funkcji zmiennej losowej  $\mathbf{x}$  względem skomplikowanego rozkładu prawdopodobieństwa  $p(\mathbf{x})$ , który znamy jedynie z dokładnością do stałej normalizującej

$$p(\mathbf{x}) = \frac{1}{Z_p} \tilde{p}(\mathbf{x})$$

tj. szukamy

$$\mathbb{E}_p[f(\mathbf{x})] = \int f(\mathbf{x}) p(\mathbf{x}) d^n \mathbf{x}.$$

Jeśli umiemy generować próbki  $\mathbf{x}$  z innego (prostszego) rozkładu  $q(\mathbf{x})$ , który nazywamy rozkładem proponującym kandydatów (z ang. *proposal distribution*) to możemy zapisać

$$\begin{aligned} \mathbb{E}_p[f(\mathbf{x})] &= \int_{\mathbb{R}^n} f(\mathbf{x}) p(\mathbf{x}) d^n \mathbf{x} = \int_{\mathbb{R}^n} f(\mathbf{x}) \frac{p(\mathbf{x})}{q(\mathbf{x})} q(\mathbf{x}) d^n \mathbf{x} \\ &= \mathbb{E}_q \left[ f(\mathbf{x}) \frac{p(\mathbf{x})}{q(\mathbf{x})} \right] = \frac{Z_q}{Z_p} \mathbb{E}_q \left[ f(\mathbf{x}) \frac{\tilde{p}(\mathbf{x})}{\tilde{q}(\mathbf{x})} \right]. \end{aligned}$$

Zakładamy tutaj, iż nośnik rozkładu  $p$  zawiera się w nośniku  $q$  tj.  $\text{supp } p \subseteq \text{supp } q$ . Stosunek stałych  $Z_p/Z_q$  również możemy oszacować z próbek z  $q$ , gdyż mamy

$$Z_p = \int_{\mathbb{R}^n} \tilde{p}(\mathbf{x}) d^n \mathbf{x} = Z_q \int_{\mathbb{R}^n} \frac{\tilde{p}(\mathbf{x})}{\tilde{q}(\mathbf{x})} q(\mathbf{x}) d^n \mathbf{x} = Z_q \mathbb{E}_q \left[ \frac{\tilde{p}(\mathbf{x})}{\tilde{q}(\mathbf{x})} \right],$$

skąd ostatecznie

$$\mathbb{E}_p[f(\mathbf{x})] = \frac{\mathbb{E}_q \left[ f(\mathbf{x}) \frac{\tilde{p}(\mathbf{x})}{\tilde{q}(\mathbf{x})} \right]}{\mathbb{E}_q \left[ \frac{\tilde{p}(\mathbf{x})}{\tilde{q}(\mathbf{x})} \right]}.$$

Jeśli z rozkładu  $q$  wygenerowaliśmy próbki  $X = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$  to na mocy silnego prawa wielkich liczb mamy

$$\mathbb{E}_p[f(\mathbf{x})] \approx \frac{\sum_{i=1}^m f(\mathbf{x}_i) \frac{\tilde{p}(\mathbf{x}_i)}{\tilde{q}(\mathbf{x}_i)}}{\sum_{i=1}^m \frac{\tilde{p}(\mathbf{x}_i)}{\tilde{q}(\mathbf{x}_i)}} = \sum_{i=1}^m \lambda_i f(\mathbf{x}_i),$$

gdzie

$$\lambda_i = \frac{\tilde{p}(\mathbf{x}_i)/\tilde{q}(\mathbf{x}_i)}{\sum_{j=1}^m \tilde{p}(\mathbf{x}_j)/\tilde{q}(\mathbf{x}_j)}.$$

Algorytm Importance Sampling jest prostym algorytmem Monte Carlo, który ma jeden zasadniczy problem. W jaki sposób mamy wybrać rozkład proponujący kandydatów  $q$ ? Pewną odpowiedź na to pytanie sugeruje analiza wariancji statystyki

$$\bar{f}_m(\mathbf{x}_1, \dots, \mathbf{x}_m) = \frac{1}{m} \sum_{i=1}^m \frac{f(\mathbf{x}_i)p(\mathbf{x}_i)}{q(\mathbf{x}_i)}$$

dla  $\mathbf{x}_i \sim q$  i zakładając dla uproszczenia, iż  $f$  jest funkcją skalarną mamy

$$\text{Var}[\bar{f}_m] = \frac{1}{m} \text{Var}_q \left[ f(\mathbf{x}) \frac{p(\mathbf{x})}{q(\mathbf{x})} \right] = \frac{1}{m} \int_{\mathbb{R}^n} \frac{(f(\mathbf{x})p(\mathbf{x}) - \mu_f q(\mathbf{x}))^2}{q(\mathbf{x})} d^n \mathbf{x}.$$

Chcemy oczywiście, aby wariancja była jak najmniejsza, gdyż wówczas mała liczba próbek da dobre przybliżenie wartości oczekiwanej. Rozkład proponujący kandydatów powinien być zatem proporcjonalny do  $f(\mathbf{x})p(\mathbf{x})$ , co może być trudne do praktycznego zrealizowania.

### 0.0.25 Algorytm Metropolisa–Hastingsa

Cała klasa algorytmów próbkowania MCMC opiera się na idei wyrażenia generowania próbek jako ewolucji pewnego łańcucha Markowa. Łańcuchem Markowa nazywamy ciąg zmiennych losowych  $(X_t)$  o wartościach w  $\mathbb{R}^n$  taki, że spełnione jest kryterium Markowa

$$\forall A \subset \mathbb{R}^n : P(X_t \in A \mid X_{t-1} = \mathbf{x}_{t-1}, \dots, X_0 = \mathbf{x}_0) = P(X_t \in A \mid X_{t-1} = \mathbf{x}_{t-1}).$$

Elementy ciągu nazywamy stanami łańcucha Markowa. Dany łańcuch jest zadany jednoznacznie przez podanie gęstości prawdopodobieństwa przejścia łańcucha ze stanu  $\mathbf{x} \rightarrow \mathbf{y}$ , którą będziemy oznaczać przez  $\pi(\mathbf{y} \mid \mathbf{x})$  (zakładamy, iż prawdopodobieństwo przejścia jest niezależne od chwili  $t$  – łańcuch taki nazywamy jednorodnym). Funkcja  $\pi$  spełnia oczywiście warunek unormowania

$$\int_{\mathbb{R}^n} \pi(\mathbf{y} \mid \mathbf{x}) d^n \mathbf{y},$$

istotnie prawdopodobieństwo przejścia gdziekolwiek ze stanu  $\mathbf{x}$  jest równe 1. Będziemy zakładać dodatkowo, iż  $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n : \pi(\mathbf{y} \mid \mathbf{x}) > 0$ . Rozkład  $p(\mathbf{x})$

łańcucha Markowa (tj. rozkład prawdopodobieństwa z którego losujemy stan łańcucha w danej chwili  $t$ ) z daną funkcją przejścia  $\pi$  nazwiemy rozkładem stacjonarnym tego łańcucha iff

$$p(\mathbf{y}) = \int_{\mathbb{R}^n} \pi(\mathbf{y} | \mathbf{x}) p(\mathbf{x}) d^n \mathbf{x} .$$

Rozkład stacjonarny danego łańcucha oznaczmy przez  $p^*(\mathbf{x})$ . Zauważmy, iż jeśli stan początkowy łańcucha  $X_0$  pochodzi z rozkładu stacjonarnego  $p^*$  to każdy kolejny stan  $X_t$  również pochodzi z rozkładu stacjonarnego. Jeśli z kolei stan początkowy pochodzi z jakiegoś innego rozkładu  $p_0$  to rozkład łańcucha w chwili  $t$  jest dany przez relację rekurencyjną

$$p_t(\mathbf{y}) = \int_{\mathbb{R}^n} \pi(\mathbf{y} | \mathbf{x}) p_{t-1}(\mathbf{x}) d^n \mathbf{x} , \quad \text{dla } t > 1.$$

Rozkładem granicznym łańcucha Markowa nazwiemy granicę w sensie zbieżności punktowej

$$\lim_{t \rightarrow \infty} p_t(\mathbf{x}) .$$

Przy podanych wyżej założeniach istnieje twierdzenie, które mówi iż taki łańcuch Markowa posiada jednoznaczny rozkład stacjonarny tożsamy z rozkładem granicznym. Ponadto warunkiem wystarczającym, aby dany rozkład  $p(\mathbf{x})$  był rozkładem stacjonarnym łańcucha Markowa jest

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n : \pi(\mathbf{y} | \mathbf{x}) p(\mathbf{x}) = \pi(\mathbf{x} | \mathbf{y}) p(\mathbf{y}) ,$$

co wynika z scałkowania powyższego równania

$$\int_{\mathbb{R}^n} \pi(\mathbf{y} | \mathbf{x}) p(\mathbf{x}) d^n \mathbf{x} = \int_{\mathbb{R}^n} \pi(\mathbf{x} | \mathbf{y}) p(\mathbf{y}) d^n \mathbf{x} = p(\mathbf{y}) \int_{\mathbb{R}^n} \pi(\mathbf{x} | \mathbf{y}) d^n \mathbf{x} = p(\mathbf{y}) .$$

Kryterium to nazywamy kryterium lokalnego balansu (z ang. *detailed balance condition*).

Podstawowa idea wykorzystania łańcuchów Markowa do generowania próbek ze skomplikowanego rozkładu  $p$  jest więc następująca: tworzymy łańcuch Markowa opisany powyżej, dla którego  $p$  jest rozkładem stacjonarnym, wówczas rozpoczynając w dowolnym dopuszczalnym stanie początkowym  $X_0$  po wykonaniu dużej liczby kroków (etap ten nazywamy okresem przejściowym z ang. *burn-in period*) stan  $X_t$  (dla  $t \gg 1$ ) tego łańcucha będzie w przybliżeniu pochodził z rozkładu granicznego  $p$  (nie jest jednak prosto stwierdzić po jak długim okresie przejściowym przybliżenie to jest wystarczająco dobre). Aby otrzymać z takiej procedury próbki prawdziwie i.i.d. każda z próbek musiałaby pochodzić z ponownego uruchomienia takiego łańcucha. Oczywiście jest to nieefektywne, więc w praktyce generujemy próbki z jednego łańcucha po prostu odrzucając pewne z nich tak aby uniknąć znaczących korelacji.

Pozostaje pytanie jak skonstruować funkcję przejścia  $\pi(\mathbf{y} | \mathbf{x})$  dla danego rozkładu granicznego  $p(\mathbf{x})$ . Podstawową konstrukcję podaje algorytm Metropolisa–Hastingsa:

1. Jako stan początkowy przyjmij dowolną dopuszczalną wartość  $\mathbf{x} \leftarrow \mathbf{x}_0$ .
2. Powtarzaj:
  - (a) Będąc w aktualnym stanie  $\mathbf{x}$  z prostego rozkładu proponującego kandydatów  $q(\mathbf{y} | \mathbf{x})$  wylosuj kandydata  $\mathbf{y}$  na wartość łańcucha w kolejnym stanie.
  - (b) Z prawdopodobieństwem

$$r(\mathbf{y} | \mathbf{x}) = \min \left\{ 1, \frac{p(\mathbf{y})q(\mathbf{x} | \mathbf{y})}{p(\mathbf{x})q(\mathbf{y} | \mathbf{x})} \right\}$$

zaakceptuj kandydata jako nowy stan i przejdź do stanu  $\mathbf{y}$ . W przeciwnym razie pozostać w stanie  $\mathbf{x}$

Funkcja przejścia ma zatem postać

$$\pi_{\text{MH}}(\mathbf{y} | \mathbf{x}) = q(\mathbf{y} | \mathbf{x})r(\mathbf{y} | \mathbf{x}).$$

Pozostaje tylko wykazać, iż spełnione jest kryterium lokalnego balansu. Istotnie mamy

$$\begin{aligned} \pi_{\text{MH}}(\mathbf{y} | \mathbf{x})p(\mathbf{x}) &= \min \{q(\mathbf{y} | \mathbf{x})p(\mathbf{x}), q(\mathbf{x} | \mathbf{y})p(\mathbf{y})\} \\ \pi_{\text{MH}}(\mathbf{x} | \mathbf{y})p(\mathbf{y}) &= \min \{q(\mathbf{x} | \mathbf{y})p(\mathbf{y}), q(\mathbf{y} | \mathbf{x})p(\mathbf{x})\} \end{aligned} \quad ,$$

skąd  $\pi_{\text{MH}}(\mathbf{y} | \mathbf{x})p(\mathbf{x}) = \pi_{\text{MH}}(\mathbf{x} | \mathbf{y})p(\mathbf{y})$ . Zauważmy, iż nie musimy znać  $p(\mathbf{x})$  z dokładnością do stałej normalizującej, gdyż

$$\frac{p(\mathbf{y})}{p(\mathbf{x})} = \frac{\tilde{p}(\mathbf{y})/Z_p}{\tilde{p}(\mathbf{x})/Z_p} = \frac{\tilde{p}(\mathbf{y})}{\tilde{p}(\mathbf{x})}.$$

Poza algorytmem Metropolisa–Hastingsa jest wiele innych algorytmów z rodziny MCMC. Większość z nich implementuje konkretny sposób generowania (zostawiając resztę struktury) tak, aby zmniejszyć korelację po okresie przejściowym i przyspieszyć zbieżność. Standardowo wykorzystywanymi algorytmami z tej klasy są algorytmy HMC (*Hamiltonian Monte Carlo*) oraz NUTS (*No U-Turn Sampler*).

### 0.0.26 Elementarz teorii informacji

Mając dany skończony zbiór zdarzeń elementarnych  $\{A_1, \dots, A_n\}$  taki, że wynikiem eksperymentu losowego może być dokładnie jedno z nich oraz prawdopodobieństwa  $p_1, \dots, p_n$ ,  $\sum_i p_i = 1$  każdego z nich powiemy, iż

$$A := \begin{pmatrix} A_1 & A_2 & \cdots & A_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$$

jest *schematem skończonym* (z ang. *finite scheme*). Przykładowo rzut sprawiedliwą, sześcienną kostką do gry jest opisany przez schemat

$$\begin{pmatrix} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}.$$



Zauważmy, że każdy schemat skończony opisuje pewną *niepewność* dotyczącą doświadczenia losowego. Przykładowo jest oczywiste, iż dla schematów

$$\begin{pmatrix} A_1 & A_2 \\ 0.99 & 0.01 \end{pmatrix}, \quad \begin{pmatrix} A_1 & A_2 \\ 0.5 & 0.5 \end{pmatrix}$$

pierwszy z nich opisuje znacznie mniejszą niepewność od drugiego, gdyż prawie z pewnością wynikiem eksperymentu losowego będzie  $A_1$ . Wprowadzimy teraz wielkość, która w sensowny sposób mierzy ilość niepewności w danym schemacie skończonym. Wielkością taką jest *entropia Shannona* zdefiniowana dla schematu

$$A = \begin{pmatrix} A_1 & A_2 & \cdots & A_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$$

jako

$$H(A) = H(p_1, p_2, \dots, p_n) := - \sum_{i=1}^n p_i \lg p_i,$$

gdzie możemy wybrać dowolną ustaloną podstawę logarytmu oraz stwierdzamy, iż jeśli  $p_k = 0$  to  $p_k \lg p_k = 0$ . Jeśli jako podstawę wybierzemy liczbę 2 to entropię mierzymy w *bitach* tj. 1 bit jest to ilość niepewności zawarta w schemacie skończonym o dwóch jednakowo prawdopodobnych wynikach

$$H = -\log_2 \frac{1}{2} = 1.$$

Przekonamy się teraz, iż tak zdefiniowana miara niepewności ma szereg własności, których spodziewalibyśmy się dla sensownej miary niepewności. Zauważmy wpierw, iż  $H(p_1, \dots, p_n) = 0$  iff dokładnie jedno zdarzenie  $A_k \in A$  jest pewne, a pozostałe niemożliwe. Zauważmy dodatkowo, iż z nierówności Jensena mamy dla funkcji wypukłej  $\phi(x) = x \lg x$

$$\phi\left(\sum_{i=1}^n \lambda_i x_i\right) \leq \sum_{i=1}^n \lambda_i \phi(x_i),$$

dla dowolnych  $x_1, \dots, x_n \in \mathbb{R}$  i  $\lambda_1, \dots, \lambda_n \in [0; 1]$ ,  $\sum_i \lambda_i = 1$ , skąd

$$\frac{1}{n} \lg \frac{1}{n} \leq \frac{1}{n} \sum_{i=1}^n p_i \lg p_i = -\frac{1}{n} H(p_1, \dots, p_n),$$

czyli

$$H(p_1, \dots, p_n) \leq -\lg \frac{1}{n} = H(1/n, 1/n, \dots, 1/n),$$

czyli niepewność zawarta w danym schemacie skończonym jest mniejsza lub równa od niepewności zawartej w analogicznym schemacie, w którym wszystkie wyniki są jednakowo prawdopodobne.

Załóżmy teraz, że mamy dwa niezależne schematy skończone

$$A = \begin{pmatrix} A_1 & A_2 & \cdots & A_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}, \quad B = \begin{pmatrix} B_1 & B_2 & \cdots & B_m \\ q_1 & q_2 & \cdots & q_m \end{pmatrix}$$

takie, że dla każdej pary zdarzeń  $A_i, B_j$  prawdopodobieństwo wystąpienia zdarzenia  $A_i B_j$  wynosi  $p_i q_j$ . Zbiór zdarzeń  $A_i B_j$  z prawdopodobieństwami  $r_{ij} = p_i q_j$  reprezentuje nowy schemat skończony  $AB$ . Wówczas

$$\begin{aligned} -H(AB) &= \sum_{i=1}^n \sum_{j=1}^m r_{ij} \lg r_{ij} = \sum_{i=1}^n \sum_{j=1}^m p_i q_j (\lg p_i + \lg q_j) \\ &= \sum_{i=1}^n p_i \lg p_i + \sum_{j=1}^m q_j \lg q_j = -H(A) - H(B), \end{aligned}$$

skąd

$$H(AB) = H(A) + H(B).$$

Rozważmy teraz przypadek gdy schematy  $A, B$  są zależne. Przez  $q_{ij}$  oznaczmy prawdopodobieństwo zajścia zdarzenia  $B_j$  pod warunkiem zdarzenia  $A_i$  tj.  $q_{ij} = p(B_j | A_i)$ . Schemat  $AB$  jest teraz opisany prawdopodobieństwami  $r_{ij} = p_i q_{ij}$  zatem

$$-H(AB) = \sum_{i=1}^n \sum_{j=1}^m p_i q_{ij} (\lg p_i + \lg q_{ij}) = -H(A) + \sum_{i=1}^n p_i \sum_{j=1}^m q_{ij} \lg q_{ij}$$

gdyż  $\sum_j q_{ij} = 1$  (prawdopodobieństwo zajścia dowolnego zdarzenia z  $B$  pod warunkiem wystąpienia zdarzenia  $A_i$  wynosi 1), natomiast wielkość  $-\sum_{j=1}^m q_{ij} \lg q_{ij}$  jest warunkową entropią schematu  $B$  pod warunkiem zajścia zdarzenia  $A_i$ , co oznaczmy jako  $H(B | A = A_i)$

$$H(AB) = H(A) + \sum_{i=1}^n p_i H(B | A = A_i).$$

Ostatni człon jest w takim razie wartością oczekiwaną wielkości  $H(B)$  w schemacie  $A$ , co oznaczmy jako  $H(B | A)$ . Mamy w takim razie

$$H(AB) = H(A) + H(B | A).$$

Z nierówności Jensena można dodatkowo pokazać, że zachodzi  $H(B | A) \leq H(B)$ .