

Math for ML

Amir Hosein Hadian

Probability for Machine learning

A random variable is a real-valued function defined on the set of outcomes of a random experiment.

A process that has random outcomes is called a random experiment

The set of all possible outcomes of a random experiment is called the sample space and is denoted Ω .

A combination of outcomes, a subset of Ω , is called an event.

The set of events is denoted \mathcal{A} . It is assumed that the set \mathcal{A} is a σ -algebra which means that the following properties hold for \mathcal{A} :

- (a) $\Omega \in \mathcal{A}$
- (b) $A^c \in \mathcal{A}$ if $A \in \mathcal{A}$, where A^c is the complement of set A
- (c) $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$ if $A_1, A_2, \dots \in \mathcal{A}$.

The probability P is a set function that maps A into [0, 1] and P is called a probability measure if the following conditions hold:

(d) $P(\Omega) = 1$

(e) $P(A^c) = 1 - P(A)$

(f) $P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i), \text{ if } A_i \cap A_j = \phi \text{ for } i \neq j.$

Example

Flipping a coin twice; generating a σ -algebra set

Number of observations; Poisson distributed

$$P_0(t + \Delta t) = (1 - \lambda \Delta t)P_0(t) + o(\Delta t)$$

$$P_n(t + \Delta t) = (1 - \lambda \Delta t)P_n(t) + \lambda \Delta t P_{n-1}(t) + o(\Delta t)$$

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t), \quad P_0(0) = 1$$

$$\frac{dP_n(t)}{dt} = -\lambda P_n(t) + \lambda P_{n-1}(t), \quad P_n(0) = 1, \quad \text{for } n \geq 1.$$

$$P(\{\omega_n\}) = P_n(t) = \exp(-\lambda t)(\lambda t)^n/n!$$

A random variable X is a real-valued function that assigns the value $X(\omega) \in R$ to each outcome $\omega \in \Omega$. That is, $X : \Omega \rightarrow R$. The function $F(x) = P(\{\omega \in \Omega : X(\omega) \leq x\})$ is called the distribution function or the probability distribution of the random variable X .

$$E(X) = \int_{-\infty}^{\infty} xp(x) dx.$$

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)p(x) dx.$$

$$E(X^k) = \int_{-\infty}^{\infty} x^k p(x) dx.$$

Sum rule

$$p(\mathbf{x}) = \begin{cases} \sum_{\mathbf{y} \in \mathcal{Y}} p(\mathbf{x}, \mathbf{y}) & \text{if } \mathbf{y} \text{ is discrete} \\ \int_{\mathcal{Y}} p(\mathbf{x}, \mathbf{y}) d\mathbf{y} & \text{if } \mathbf{y} \text{ is continuous} \end{cases},$$

Product rule

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y} \mid \mathbf{x})p(\mathbf{x}).$$

Bayes rule

$$\underbrace{p(x | y)}_{\text{posterior}} = \frac{\overbrace{p(y | x) p(x)}^{\text{likelihood prior}}}{\underbrace{p(y)}_{\text{evidence}}}$$

Marginal probability

$$p(\mathbf{y}) := \int p(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \mathbb{E}_{\mathbf{X}}[p(\mathbf{y} \mid \mathbf{x})]$$

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}.$$

$$\begin{aligned} P(X|C_i) &= \prod_{k=1}^n P(x_k|C_i) \\ &= P(x_1|C_i) \times P(x_2|C_i) \times \cdots \times P(x_n|C_i). \end{aligned}$$

age	income	student	credit_rating	com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

$$X = (\text{age} = \text{youth}, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit_rating} = \text{fair})$$

$$P(X|C_i)P(C_i), \text{ for } i = 1, 2.$$

$$P(\text{buys_computer} = \text{yes}) = 9/14 = 0.643$$

$$P(\text{buys_computer} = \text{no}) = 5/14 = 0.357$$

$$P(\text{age} = \text{youth} \mid \text{buys_computer} = \text{yes}) = 2/9 = 0.222$$

$$P(\text{age} = \text{youth} \mid \text{buys_computer} = \text{no}) = 3/5 = 0.600$$

$$P(\text{income} = \text{medium} \mid \text{buys_computer} = \text{yes}) = 4/9 = 0.444$$

$$P(\text{income} = \text{medium} \mid \text{buys_computer} = \text{no}) = 2/5 = 0.400$$

$$P(\text{student} = \text{yes} \mid \text{buys_computer} = \text{yes}) = 6/9 = 0.667$$

$$P(\text{student} = \text{yes} \mid \text{buys_computer} = \text{no}) = 1/5 = 0.200$$

$$P(\text{credit_rating} = \text{fair} \mid \text{buys_computer} = \text{yes}) = 6/9 = 0.667$$

$$P(\text{credit_rating} = \text{fair} \mid \text{buys_computer} = \text{no}) = 2/5 = 0.400$$

$$\begin{aligned} P(X|buys_computer = yes) &= P(age = youth | buys_computer = yes) \\ &\quad \times P(income = medium | buys_computer = yes) \\ &\quad \times P(student = yes | buys_computer = yes) \\ &\quad \times P(credit_rating = fair | buys_computer = yes) \\ &= 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044. \end{aligned}$$

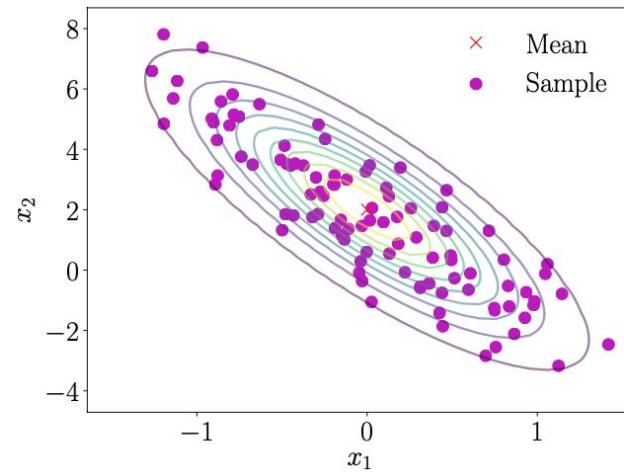
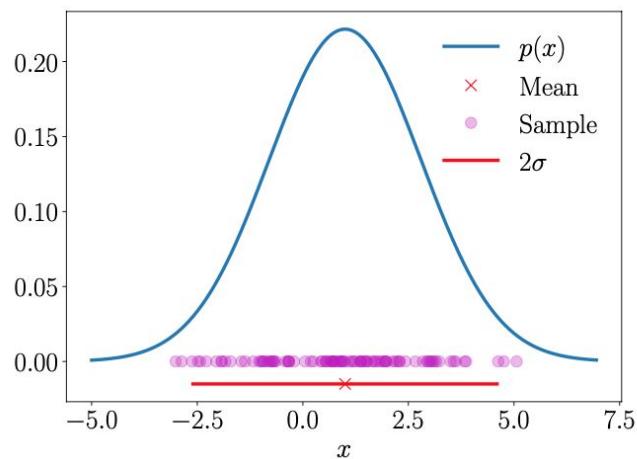
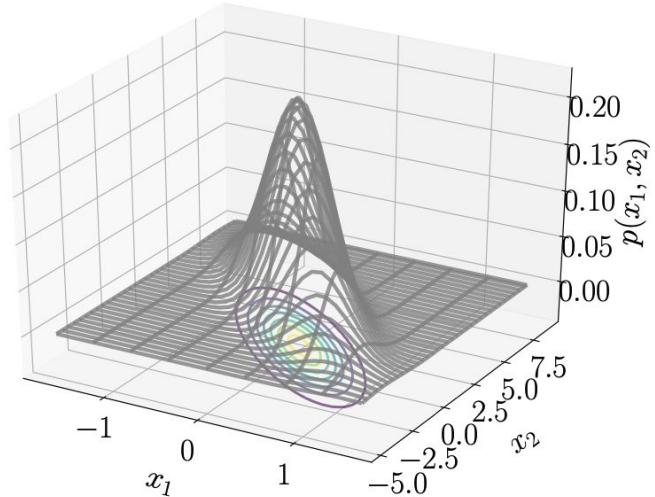
$$P(X|buys_computer = no) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019.$$

$$P(X|buys_computer = yes)P(buys_computer = yes) = 0.044 \times 0.643 = 0.028$$
$$P(X|buys_computer = no)P(buys_computer = no) = 0.019 \times 0.357 = 0.007$$

independence

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}) .$$

- $p(\mathbf{y} \mid \mathbf{x}) = p(\mathbf{y})$
- $p(\mathbf{x} \mid \mathbf{y}) = p(\mathbf{x})$
- $\mathbb{V}_{X,Y}[\mathbf{x} + \mathbf{y}] = \mathbb{V}_X[\mathbf{x}] + \mathbb{V}_Y[\mathbf{y}]$
- $\text{Cov}_{X,Y}[\mathbf{x}, \mathbf{y}] = \mathbf{0}$



Normal distribution

$$p(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right).$$

$$p(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right),$$

$$p(\mathbf{x}, \mathbf{y}) = \mathcal{N} \left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix} \right).$$

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

$$\boldsymbol{\mu}_{x \mid y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} (\mathbf{y} - \boldsymbol{\mu}_y)$$

$$\boldsymbol{\Sigma}_{x \mid y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{yx}.$$

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}).$$

$$p(y \mid \boldsymbol{x}) = \mathcal{N}(y \mid f(\boldsymbol{x}), \sigma^2) .$$

$$y = f(\boldsymbol{x}) + \epsilon ,$$

$$\begin{aligned} p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) &= \mathcal{N}(y \mid \boldsymbol{x}^\top \boldsymbol{\theta}, \sigma^2) \\ \iff y &= \boldsymbol{x}^\top \boldsymbol{\theta} + \epsilon , \quad \epsilon \sim \mathcal{N}(0, \sigma^2) , \end{aligned}$$

$$\begin{aligned}
p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) &= p(y_1, \dots, y_N \mid \mathbf{x}_1, \dots, \mathbf{x}_N, \boldsymbol{\theta}) \\
&= \prod_{n=1}^N p(y_n \mid \mathbf{x}_n, \boldsymbol{\theta}) = \prod_{n=1}^N \mathcal{N}(y_n \mid \mathbf{x}_n^\top \boldsymbol{\theta}, \sigma^2),
\end{aligned}$$

$$\mathcal{X} := \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \text{ and } \mathcal{Y} := \{y_1, \dots, y_N\}$$

$$p(y_* \mid \boldsymbol{x}_*, \boldsymbol{\theta}^*) = \mathcal{N}\!\left(y_* \mid \boldsymbol{x}_*^\top \boldsymbol{\theta}^*, \, \sigma^2\right).$$

$$\boldsymbol{\theta}_{\text{ML}} = \arg \max_{\boldsymbol{\theta}} p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) \, .$$

$$-\log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) = -\log \prod_{n=1}^N p(y_n \mid \mathbf{x}_n, \boldsymbol{\theta}) = -\sum_{n=1}^N \log p(y_n \mid \mathbf{x}_n, \boldsymbol{\theta}),$$

$$\log p(y_n \mid \mathbf{x}_n, \boldsymbol{\theta}) = -\frac{1}{2\sigma^2}(y_n - \mathbf{x}_n^\top \boldsymbol{\theta})^2 + \text{const},$$

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &:= \frac{1}{2\sigma^2} \sum_{n=1}^N (y_n - \mathbf{x}_n^\top \boldsymbol{\theta})^2 \\
&= \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2,
\end{aligned}$$

$$\begin{aligned}
\frac{d\mathcal{L}}{d\boldsymbol{\theta}} &= \frac{d}{d\boldsymbol{\theta}} \left(\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \right) \\
&= \frac{1}{2\sigma^2} \frac{d}{d\boldsymbol{\theta}} \left(\mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\theta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\theta} \right) \\
&= \frac{1}{\sigma^2} (-\mathbf{y}^\top \mathbf{X} + \boldsymbol{\theta}^\top \mathbf{X}^\top \mathbf{X}) \in \mathbb{R}^{1 \times D}.
\end{aligned}$$

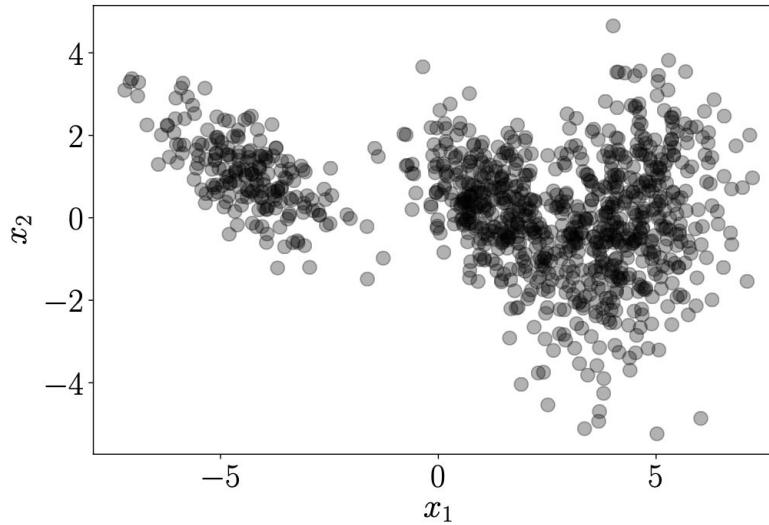
$$\begin{aligned}
\frac{d\mathcal{L}}{d\theta} = \mathbf{0}^\top &\stackrel{(9.11c)}{\iff} \boldsymbol{\theta}_{\text{ML}}^\top \mathbf{X}^\top \mathbf{X} = \mathbf{y}^\top \mathbf{X} \\
&\iff \boldsymbol{\theta}_{\text{ML}}^\top = \mathbf{y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\
&\iff \boldsymbol{\theta}_{\text{ML}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} .
\end{aligned}$$

$$\Phi := \begin{bmatrix} \phi^\top(\mathbf{x}_1) \\ \vdots \\ \phi^\top(\mathbf{x}_N) \end{bmatrix} = \begin{bmatrix} \phi_0(\mathbf{x}_1) & \cdots & \phi_{K-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \cdots & \phi_{K-1}(\mathbf{x}_2) \\ \vdots & & \vdots \\ \phi_0(\mathbf{x}_N) & \cdots & \phi_{K-1}(\mathbf{x}_N) \end{bmatrix} \in \mathbb{R}^{N \times K}$$

where $\Phi_{ij} = \phi_j(\mathbf{x}_i)$ and $\phi_j : \mathbb{R}^D \rightarrow \mathbb{R}$.

$$-\log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) = \frac{1}{2\sigma^2} (\mathbf{y} - \Phi\boldsymbol{\theta})^\top (\mathbf{y} - \Phi\boldsymbol{\theta}) + \text{const.}$$

$$\boldsymbol{\theta}_{\text{ML}} = (\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y}$$

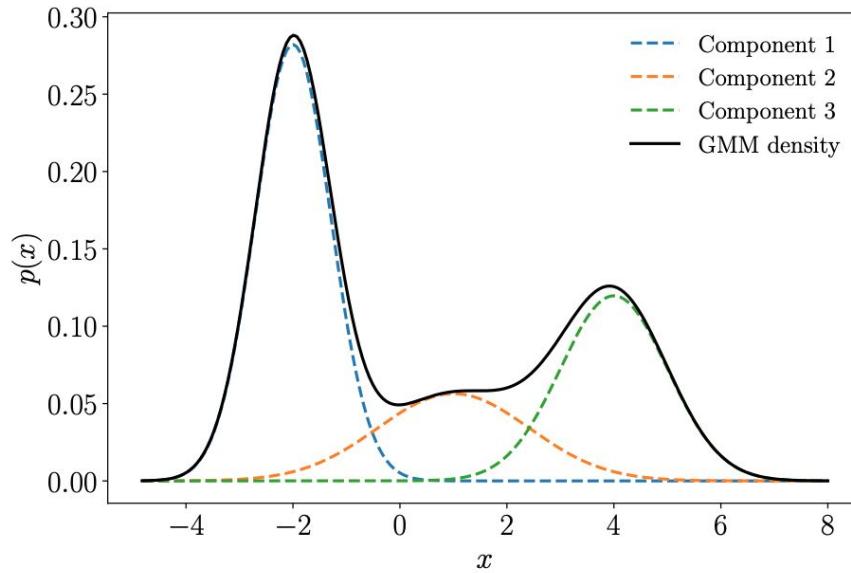


$$p(\boldsymbol{x}) = \sum_{k=1}^K \pi_k p_k(\boldsymbol{x})$$

$$0\leqslant \pi_k\leqslant 1\,,\quad \sum_{k=1}^K \pi_k = 1\,,$$

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$0 \leqslant \pi_k \leqslant 1, \quad \sum_{k=1}^K \pi_k = 1,$$



$$p(x | \theta) = 0.5\mathcal{N}(x | -2, \frac{1}{2}) + 0.2\mathcal{N}(x | 1, 2) + 0.3\mathcal{N}(x | 4, 1)$$

$$p(\mathcal{X} \mid \boldsymbol{\theta}) = \prod_{n=1}^N p(\mathbf{x}_n \mid \boldsymbol{\theta}), \quad p(\mathbf{x}_n \mid \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

$$\log p(\mathcal{X} \mid \boldsymbol{\theta}) = \sum_{n=1}^N \log p(\mathbf{x}_n \mid \boldsymbol{\theta}) = \underbrace{\sum_{n=1}^N \log \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{=: \mathcal{L}}.$$

$$\log \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{D}{2} \log(2\pi) - \frac{1}{2} \log \det(\boldsymbol{\Sigma}) - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}).$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_k} = \mathbf{0}^\top \iff \sum_{n=1}^N \frac{\partial \log p(\mathbf{x}_n | \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} = \mathbf{0}^\top ,$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_k} = \mathbf{0} \iff \sum_{n=1}^N \frac{\partial \log p(\mathbf{x}_n | \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} = \mathbf{0} ,$$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = 0 \iff \sum_{n=1}^N \frac{\partial \log p(\mathbf{x}_n | \boldsymbol{\theta})}{\partial \pi_k} = 0 .$$

$$\frac{\partial \log p(\mathbf{x}_n | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{p(\mathbf{x}_n | \boldsymbol{\theta})} \frac{\partial p(\mathbf{x}_n | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

$$\frac{1}{p(\mathbf{x}_n | \boldsymbol{\theta})} = \frac{1}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

$$r_{nk} := \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

$$p(\boldsymbol{x}_n | \pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\boldsymbol{r}_n := [r_{n1}, \dots, r_{nK}]^\top$$

$$\boldsymbol{\mu}_k^{new} = \frac{\sum_{n=1}^N r_{nk} \boldsymbol{x}_n}{\sum_{n=1}^N r_{nk}},$$

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{n=1}^N r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top,$$

$$\pi_k^{new} = \frac{N_k}{N}\,,\quad k=1,\ldots,K\,,$$

$$\mu_k^{new} = \frac{\sum_{n=1}^N r_{nk} x_n}{\sum_{n=1}^N r_{nk}},$$

$$\begin{aligned}
\frac{\partial p(\mathbf{x}_n \mid \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} &= \sum_{j=1}^K \pi_j \frac{\partial \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\partial \boldsymbol{\mu}_k} = \pi_k \frac{\partial \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\mu}_k} \\
&= \pi_k (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_k} &= \sum_{n=1}^N \frac{\partial \log p(\mathbf{x}_n | \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^N \frac{1}{p(\mathbf{x}_n | \boldsymbol{\theta})} \frac{\partial p(\mathbf{x}_n | \boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} \\
&= \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{=r_{nk}} \\
&= \sum_{n=1}^N r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}.
\end{aligned}$$

$$\sum_{n=1}^N r_{nk} \boldsymbol{x}_n = \sum_{n=1}^N r_{nk} \boldsymbol{\mu}_k^{\text{new}} \iff \boldsymbol{\mu}_k^{\text{new}} = \frac{\sum_{n=1}^N r_{nk} \boldsymbol{x}_n}{\boxed{\sum_{n=1}^N r_{nk}}} = \frac{1}{\boxed{N_k}} \sum_{n=1}^N r_{nk} \boldsymbol{x}_n,$$

$$N_k := \sum_{n=1}^N r_{nk}$$

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{n=1}^N r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top,$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_k} = \sum_{n=1}^N \frac{\partial \log p(\mathbf{x}_n | \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} = \sum_{n=1}^N \frac{1}{p(\mathbf{x}_n | \boldsymbol{\theta})} \frac{\partial p(\mathbf{x}_n | \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k}.$$

$$\frac{\partial p(\mathbf{x}_n | \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} \quad (11.32a)$$

$$= \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \left(\pi_k (2\pi)^{-\frac{D}{2}} \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \right) \right) \quad (11.32b)$$

$$= \pi_k (2\pi)^{-\frac{D}{2}} \left[\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \right) + \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \exp \left(-\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \right) \right]. \quad (11.32c)$$

$$\frac{\partial}{\partial \Sigma_k} \det(\Sigma_k)^{-\frac{1}{2}} \stackrel{(5.101)}{=} -\frac{1}{2} \det(\Sigma_k)^{-\frac{1}{2}} \Sigma_k^{-1}, \quad (11.33)$$

$$\frac{\partial}{\partial \Sigma_k} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) \stackrel{(5.106)}{=} -\Sigma_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1} \quad (11.34)$$

$$\begin{aligned} \frac{\partial p(\boldsymbol{x}_n \mid \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} &= \pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ &\cdot \left[-\frac{1}{2} (\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}) \right]. \end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}_k} &= \sum_{n=1}^N \frac{\partial \log p(\mathbf{x}_n | \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} = \sum_{n=1}^N \frac{1}{p(\mathbf{x}_n | \boldsymbol{\theta})} \frac{\partial p(\mathbf{x}_n | \boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_k} \\
&= \sum_{n=1}^N \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{=r_{nk}} \\
&\quad \cdot \left[-\frac{1}{2} (\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}) \right]
\end{aligned}$$

$$= -\frac{1}{2} \sum_{n=1}^N r_{nk} (\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}) \quad (11.36c)$$

$$= -\frac{1}{2} \boldsymbol{\Sigma}_k^{-1} \underbrace{\sum_{n=1}^N r_{nk}}_{=N_k} + \frac{1}{2} \boldsymbol{\Sigma}_k^{-1} \left(\sum_{n=1}^N r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \right) \boldsymbol{\Sigma}_k^{-1}.$$

$$\begin{aligned}
N_k \boldsymbol{\Sigma}_k^{-1} &= \boldsymbol{\Sigma}_k^{-1} \left(\sum_{n=1}^N r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \right) \boldsymbol{\Sigma}_k^{-1} \\
\iff N_k \boldsymbol{I} &= \left(\sum_{n=1}^N r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \right) \boldsymbol{\Sigma}_k^{-1}.
\end{aligned}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top,$$

$$\pi_k^{new} = \frac{N_k}{N}\,,\quad k=1,\ldots,K\,,$$

$$\mathfrak{L} = \mathcal{L} + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right)$$

$$= \sum_{n=1}^N \log \sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right),$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \pi_k} &= \sum_{n=1}^N \frac{\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda \\
&= \frac{1}{\pi_k} \underbrace{\sum_{n=1}^N \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{= N_k} + \lambda = \frac{N_k}{\pi_k} + \lambda,
\end{aligned}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{k=1}^K \pi_k - 1.$$

$$\pi_k = -\frac{N_k}{\lambda},$$

$$1 = \sum_{k=1}^K \pi_k.$$

$$\sum_{k=1}^K \pi_k = 1 \iff -\sum_{k=1}^K \frac{N_k}{\lambda} = 1 \iff -\frac{N}{\lambda} = 1 \iff \lambda = -N.$$

1. Initialize $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$.
2. *E-step*: Evaluate responsibilities r_{nk} for every data point \mathbf{x}_n using current parameters $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$:

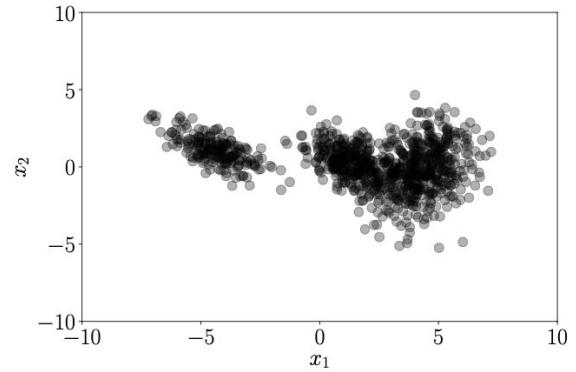
$$r_{nk} = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \quad (11.53)$$

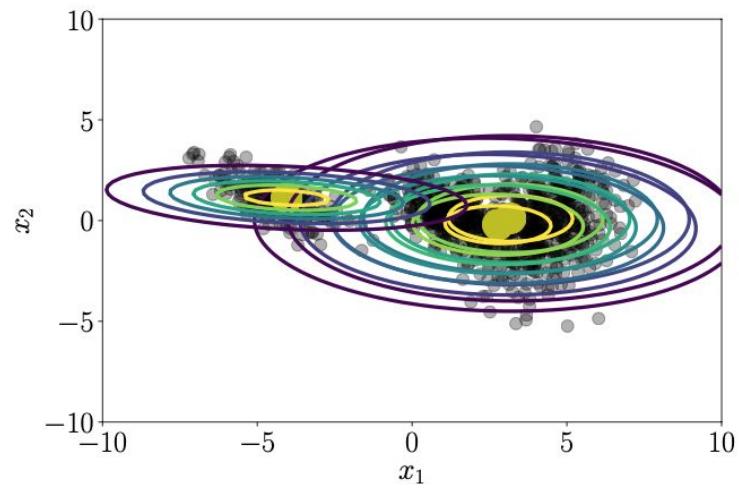
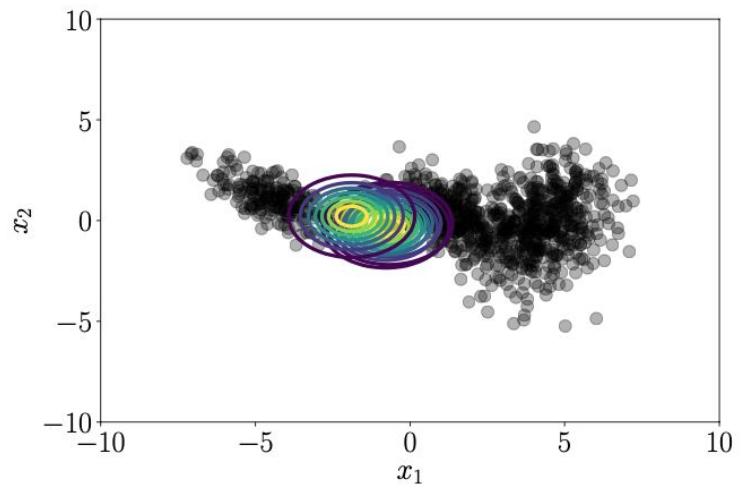
3. *M-step*: Reestimate parameters $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ using the current responsibilities r_{nk} (from E-step):

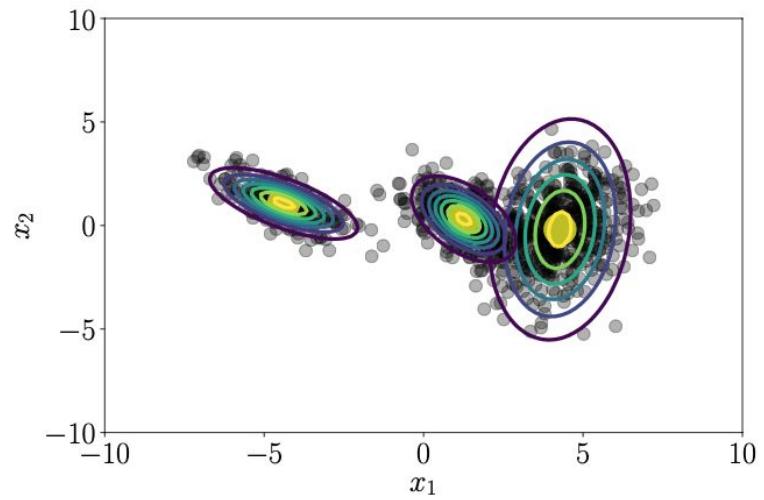
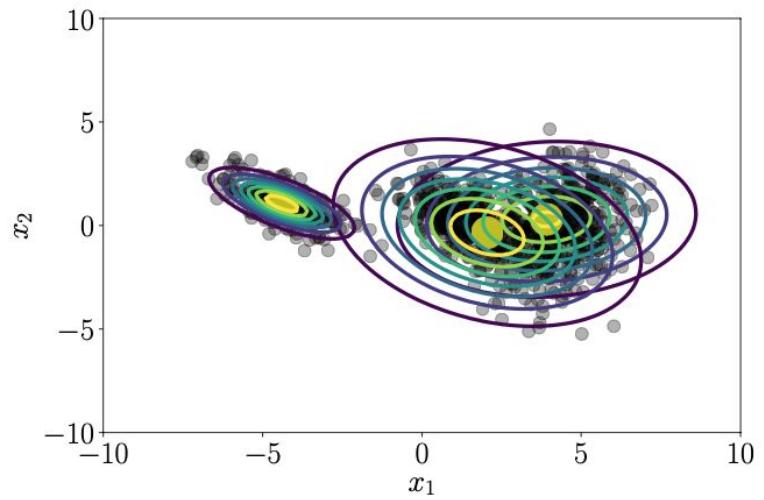
$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^N r_{nk} \mathbf{x}_n, \quad (11.54)$$

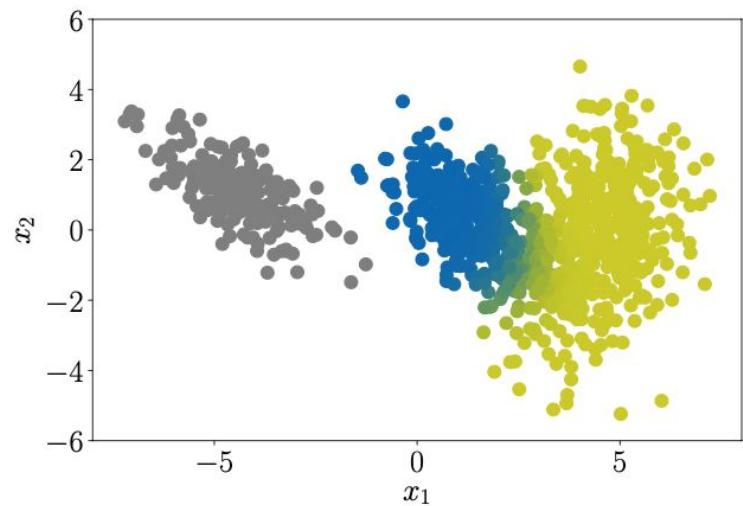
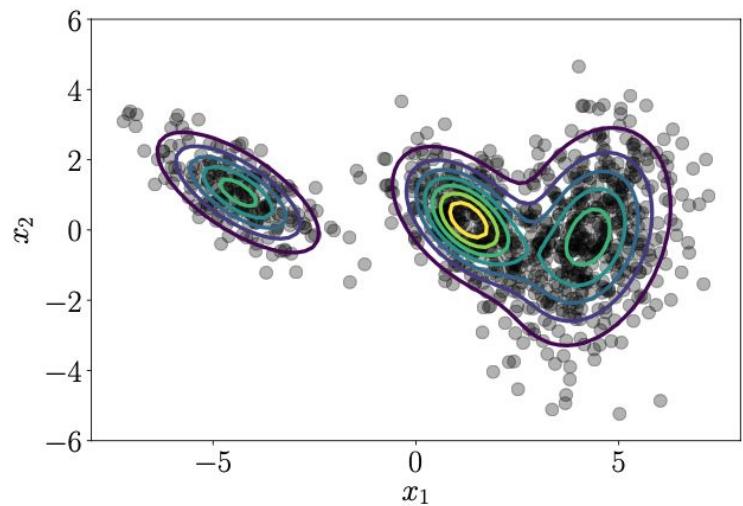
$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^N r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top, \quad (11.55)$$

$$\pi_k = \frac{N_k}{N}. \quad (11.56)$$

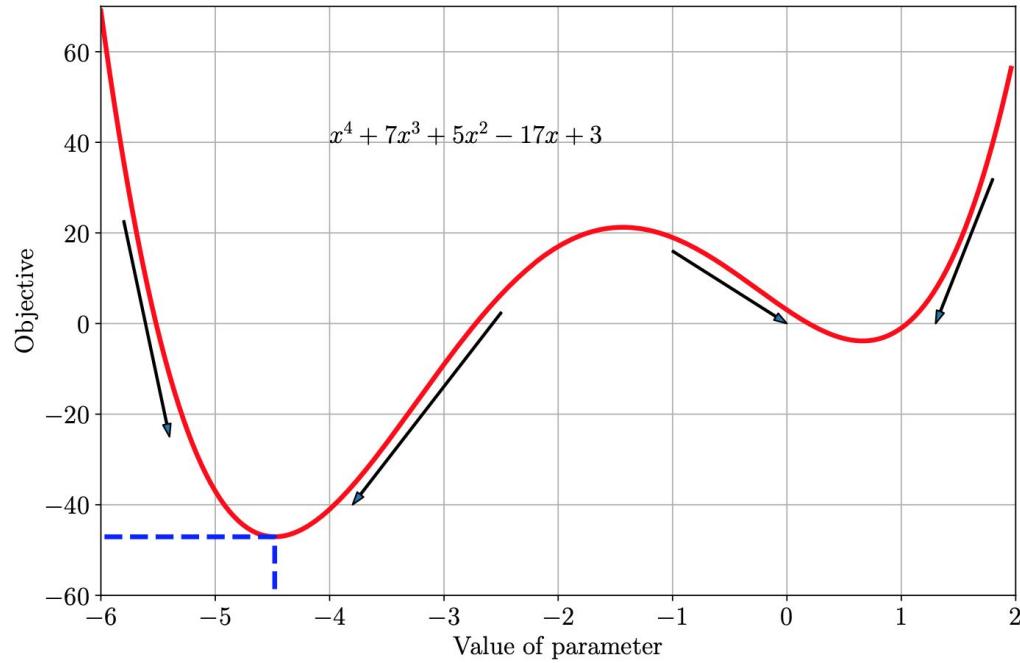








Optimization for Machine learning



$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \,,$$

$$f\,:\,\mathbb{R}^d\,\rightarrow\,\mathbb{R}$$

$$\boldsymbol{x}_1 = \boldsymbol{x}_0 - \gamma((\nabla f)(\boldsymbol{x}_0))^\top$$

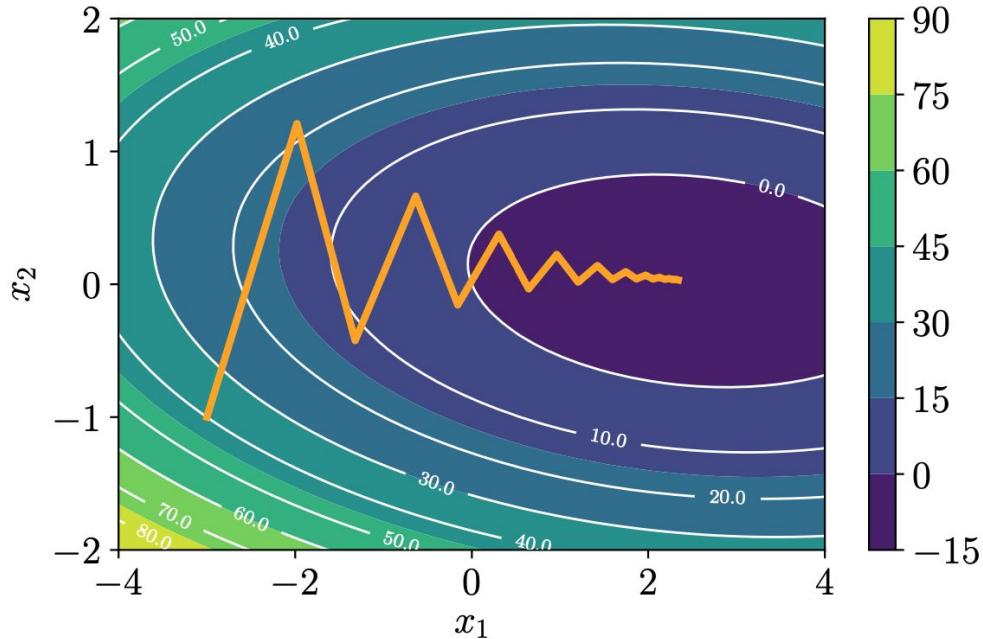
for a small *step-size* $\gamma \geq 0$, then $f(\boldsymbol{x}_1) \leq f(\boldsymbol{x}_0)$

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \gamma_i((\nabla f)(\boldsymbol{x}_i))^\top .$$

$$f\left(\begin{bmatrix}x_1 \\ x_2\end{bmatrix}\right) = \frac{1}{2}\begin{bmatrix}x_1 \\ x_2\end{bmatrix}^\top \begin{bmatrix}2 & 1 \\ 1 & 20\end{bmatrix} \begin{bmatrix}x_1 \\ x_2\end{bmatrix} - \begin{bmatrix}5 \\ 3\end{bmatrix}^\top \begin{bmatrix}x_1 \\ x_2\end{bmatrix}$$

$$\nabla f\left(\begin{bmatrix}x_1 \\ x_2\end{bmatrix}\right) = \begin{bmatrix}x_1 \\ x_2\end{bmatrix}^\top \begin{bmatrix}2 & 1 \\ 1 & 20\end{bmatrix} - \begin{bmatrix}5 \\ 3\end{bmatrix}^\top.$$

$$\boldsymbol{x}_0 = [-3, -1]^\top \qquad \qquad \gamma = 0.085$$



- When the function value increases after a gradient step, the step-size was too large. Undo the step and decrease the step-size.
- When the function value decreases the step could have been larger. Try to increase the step-size.

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \gamma_i ((\nabla f)(\boldsymbol{x}_i))^\top + \alpha \Delta \boldsymbol{x}_i$$

$$\Delta \boldsymbol{x}_i = \boldsymbol{x}_i - \boldsymbol{x}_{i-1} = \alpha \Delta \boldsymbol{x}_{i-1} - \gamma_{i-1} ((\nabla f)(\boldsymbol{x}_{i-1}))^\top,$$

$$\alpha \in [0, 1]$$

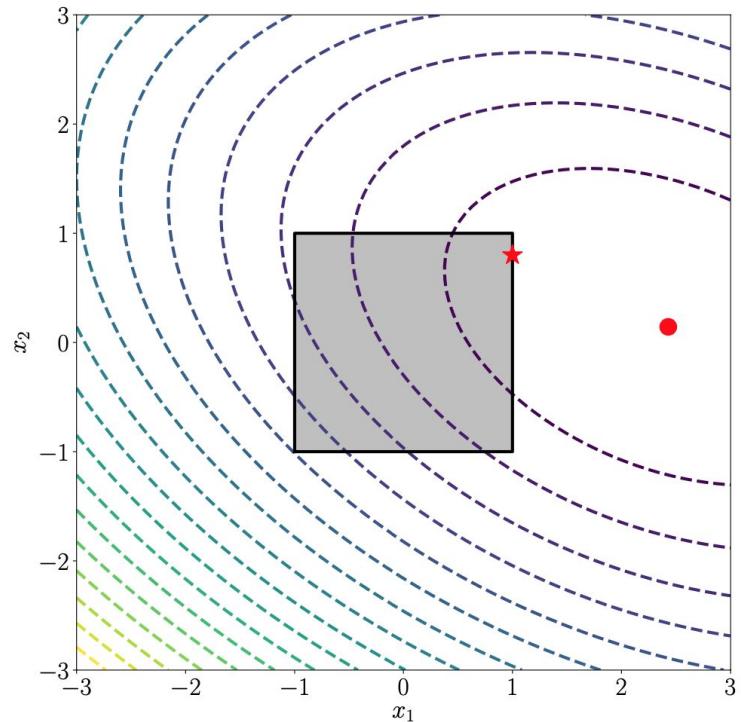
stochastic gradient descent (often shortened as SGD) is a stochastic approximation of the gradient descent method for minimizing an objective function that is written as a sum of differentiable functions.

$$L(\boldsymbol{\theta}) = \sum_{n=1}^N L_n(\boldsymbol{\theta}),$$

$$L(\boldsymbol{\theta}) = -\sum_{n=1}^N \log p(y_n | \boldsymbol{x}_n, \boldsymbol{\theta}),$$

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \gamma_i (\nabla L(\boldsymbol{\theta}_i))^\top = \boldsymbol{\theta}_i - \gamma_i \sum_{n=1}^N (\nabla L_n(\boldsymbol{\theta}_i))^\top$$

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & g_i(\boldsymbol{x}) \leq 0 \quad \text{for all} \quad i = 1, \dots, m. \end{aligned}$$



$$J(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^m \mathbf{1}(g_i(\mathbf{x})) ,$$

$$\mathbf{1}(z) = \begin{cases} 0 & \text{if } z \leq 0 \\ \infty & \text{otherwise} \end{cases} .$$

$$\begin{aligned}
\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) &= f(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i g_i(\boldsymbol{x}) \\
&= f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}),
\end{aligned}$$

$$\begin{aligned} & \min_{\boldsymbol{x}} \quad f(\boldsymbol{x}) \\ \text{subject to} \quad & g_i(\boldsymbol{x}) \leq 0 \quad \text{for all} \quad i = 1, \dots, m \end{aligned}$$

$$\begin{aligned} & \max_{\boldsymbol{\lambda} \in \mathbb{R}^m} \quad \mathfrak{D}(\boldsymbol{\lambda}) \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq \mathbf{0}, \end{aligned}$$

where $\boldsymbol{\lambda}$ are the dual variables and $\mathfrak{D}(\boldsymbol{\lambda}) = \min_{\boldsymbol{x} \in \mathbb{R}^d} \mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda})$.

$$\max_y \min_x \varphi(x, y) \leq \min_x \max_y \varphi(x, y).$$

For all x, y $\min_x \varphi(x, y) \leq \max_y \varphi(x, y).$

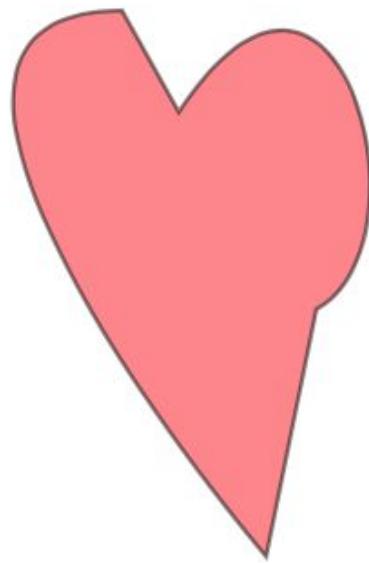
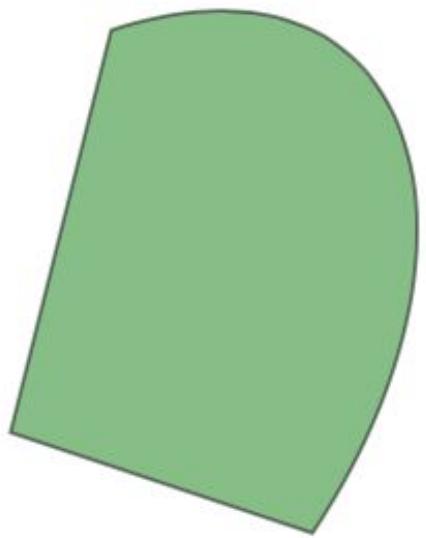
$$\chi \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$

$$J(\boldsymbol{x}) = \max_{\boldsymbol{\lambda} \geqslant \mathbf{0}} \mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \, .$$

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \max_{\boldsymbol{\lambda} \geqslant \mathbf{0}} \mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \, .$$

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \max_{\boldsymbol{\lambda} \geqslant \mathbf{0}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \geqslant \max_{\boldsymbol{\lambda} \geqslant \mathbf{0}} \min_{\boldsymbol{x} \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}).$$

$$\begin{aligned} & \min_{\boldsymbol{x}} \quad f(\boldsymbol{x}) \\ \text{subject to} \quad & g_i(\boldsymbol{x}) \leq 0 \quad \text{for all } i = 1, \dots, m \\ & h_j(\boldsymbol{x}) = 0 \quad \text{for all } j = 1, \dots, n. \end{aligned}$$



Definition 7.2. A set \mathcal{C} is a *convex set* if for any $x, y \in \mathcal{C}$ and for any scalar θ with $0 \leq \theta \leq 1$, we have

$$\theta x + (1 - \theta)y \in \mathcal{C}. \quad (7.29)$$

Definition 7.3. Let function $f : \mathbb{R}^D \rightarrow \mathbb{R}$ be a function whose domain is a convex set. The function f is a *convex function* if for all \mathbf{x}, \mathbf{y} in the domain of f , and for any scalar θ with $0 \leq \theta \leq 1$, we have

$$f(\theta \mathbf{x} + (1 - \theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1 - \theta)f(\mathbf{y}). \quad (7.30)$$

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) \, .$$

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^d} \quad & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \end{aligned}$$

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \boldsymbol{c}^\top \boldsymbol{x} + \boldsymbol{\lambda}^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}),$$

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = (\boldsymbol{c} + \boldsymbol{A}^\top \boldsymbol{\lambda})^\top \boldsymbol{x} - \boldsymbol{\lambda}^\top \boldsymbol{b}.$$

$$\mathfrak{D}(\boldsymbol{\lambda}) = -\boldsymbol{\lambda}^\top \mathbf{b}.$$

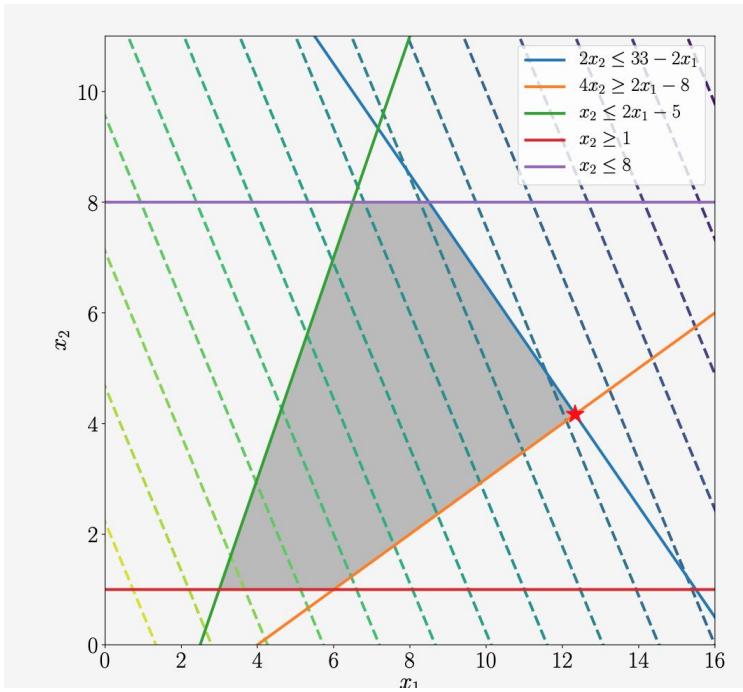
.

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^m} -\mathbf{b}^\top \boldsymbol{\lambda}$$

subject to $\mathbf{c} + \mathbf{A}^\top \boldsymbol{\lambda} = \mathbf{0}$

$$\boldsymbol{\lambda} \geqslant \mathbf{0}.$$

$$\begin{aligned}
 & \min_{\boldsymbol{x} \in \mathbb{R}^2} - \begin{bmatrix} 5 \\ 3 \end{bmatrix}^\top \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\
 & \text{subject to } \begin{bmatrix} 2 & 2 \\ 2 & -4 \\ -2 & 1 \\ 0 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq \begin{bmatrix} 33 \\ 8 \\ 5 \\ -1 \\ 8 \end{bmatrix}
 \end{aligned}$$



$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \quad \frac{1}{2} \boldsymbol{x}^\top \boldsymbol{Q} \boldsymbol{x} + \boldsymbol{c}^\top \boldsymbol{x}$$

subject to $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b},$

$$\begin{aligned}
\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) &= \frac{1}{2} \boldsymbol{x}^\top \boldsymbol{Q} \boldsymbol{x} + \boldsymbol{c}^\top \boldsymbol{x} + \boldsymbol{\lambda}^\top (\boldsymbol{A} \boldsymbol{x} - \boldsymbol{b}) \\
&= \frac{1}{2} \boldsymbol{x}^\top \boldsymbol{Q} \boldsymbol{x} + (\boldsymbol{c} + \boldsymbol{A}^\top \boldsymbol{\lambda})^\top \boldsymbol{x} - \boldsymbol{\lambda}^\top \boldsymbol{b},
\end{aligned}$$

$$\boldsymbol{Q} \boldsymbol{x} + (\boldsymbol{c} + \boldsymbol{A}^\top \boldsymbol{\lambda}) = \mathbf{0}.$$

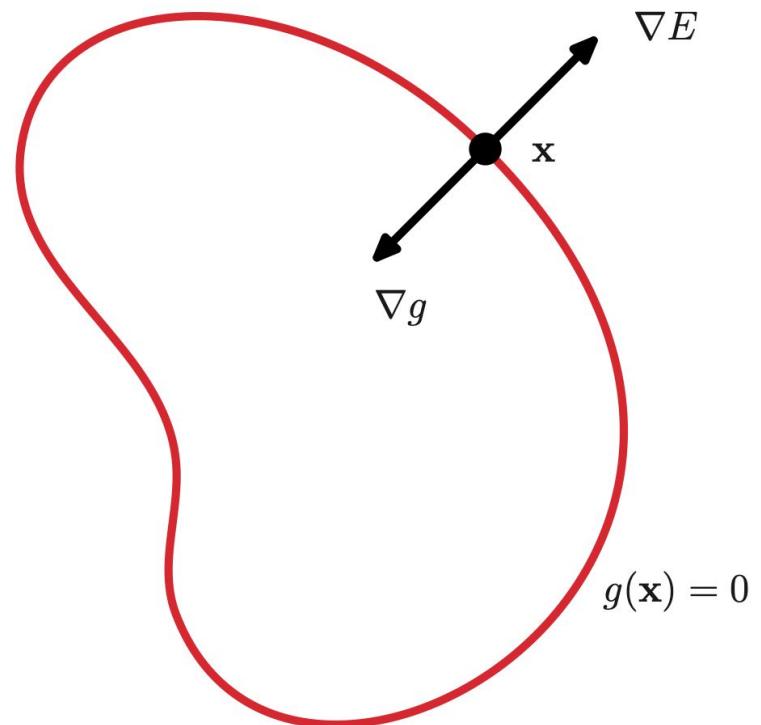
$$\boldsymbol{x} = -\boldsymbol{Q}^{-1}(\boldsymbol{c} + \boldsymbol{A}^\top \boldsymbol{\lambda}).$$

$$\mathfrak{D}(\boldsymbol{\lambda}) = -\frac{1}{2}(\mathbf{c} + \mathbf{A}^\top \boldsymbol{\lambda})^\top \mathbf{Q}^{-1}(\mathbf{c} + \mathbf{A}^\top \boldsymbol{\lambda}) - \boldsymbol{\lambda}^\top \mathbf{b}.$$

$$\begin{aligned} \max_{\boldsymbol{\lambda} \in \mathbb{R}^m} \quad & -\frac{1}{2}(\mathbf{c} + \mathbf{A}^\top \boldsymbol{\lambda})^\top \mathbf{Q}^{-1}(\mathbf{c} + \mathbf{A}^\top \boldsymbol{\lambda}) - \boldsymbol{\lambda}^\top \mathbf{b} \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq \mathbf{0}. \end{aligned}$$

$$E(\mathbf{x}) = E(x_1, x_2, \dots x_D)$$

$$g(\mathbf{x}) = 0$$



$$E(\mathbf{x}) = E(x_1, x_2, \dots x_D) \quad \nabla E = 0$$

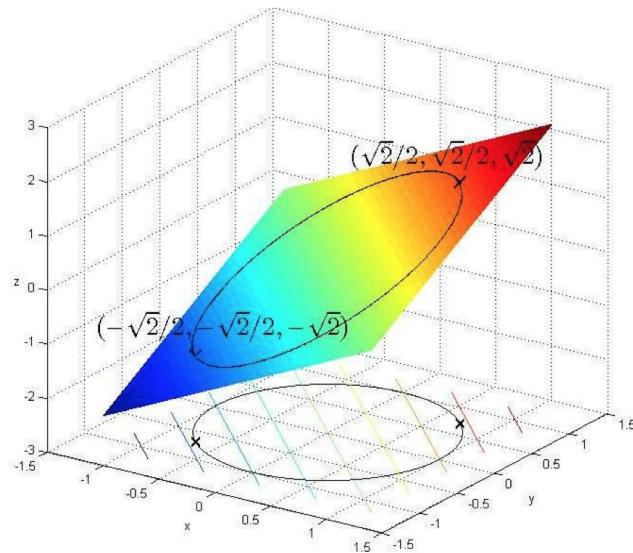
$$g(\mathbf{x}) = 0 \quad \nabla E + \lambda \nabla g = 0$$

$$L(\mathbf{x}, \lambda) = E(\mathbf{x}) + \lambda g(\mathbf{x})$$

$$\frac{dL}{d\lambda} = g(\mathbf{x}) = 0$$

$$\frac{dL}{d\mathbf{x}} = \nabla E + \lambda \nabla g = 0$$

$$\begin{aligned} & \arg \min_{x,y} x + y \\ \text{subject to } & x^2 + y^2 = 1 \end{aligned}$$



$$L(x, y, \lambda) = x + y + \lambda(x^2 + y^2 - 1)$$

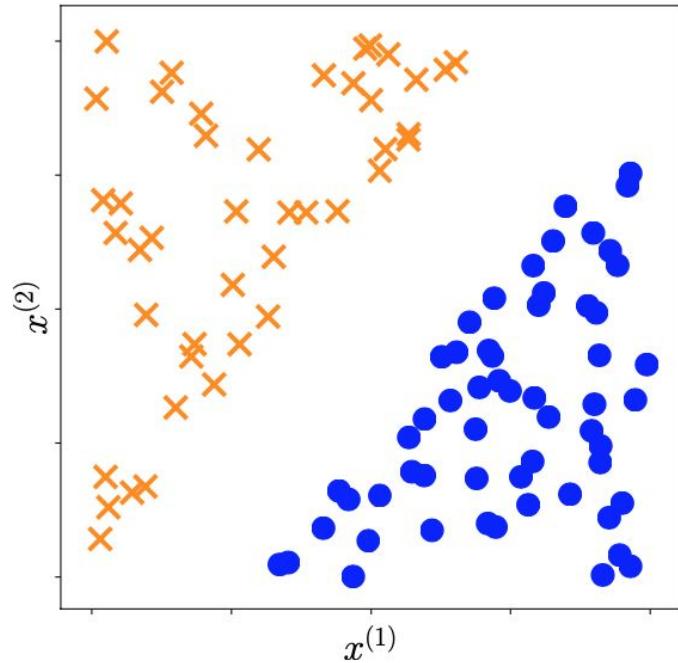
$$\frac{dL}{dx} = 1 + 2\lambda x = 0$$

$$\frac{dL}{dy} = 1 + 2\lambda y = 0$$

$$\frac{dL}{d\lambda} = x^2 + y^2 - 1 = 0$$

$$x = y = \pm \frac{1}{\sqrt{2}}.$$

$$x = y = -\frac{1}{\sqrt{2}}$$



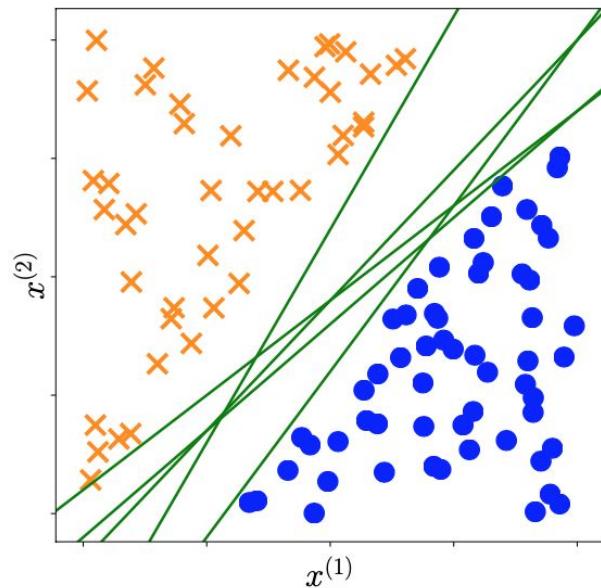
$$\begin{aligned}f : \mathbb{R}^D &\rightarrow \mathbb{R}\\ \boldsymbol{x} &\mapsto \langle \boldsymbol{w}, \boldsymbol{x}\rangle + b\,,\end{aligned}$$

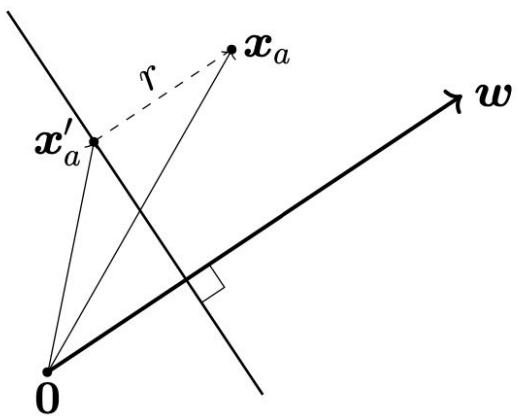
$$\left\{\boldsymbol{x}\in\mathbb{R}^D:f(\boldsymbol{x})=0\right\}\;.$$

$$\langle \mathbf{w}, \mathbf{x}_n \rangle + b \geq 0 \quad \text{when} \quad y_n = +1$$

$$\langle \mathbf{w}, \mathbf{x}_n \rangle + b < 0 \quad \text{when} \quad y_n = -1.$$

$$y_n(\langle \mathbf{w}, \mathbf{x}_n \rangle + b) \geq 0.$$

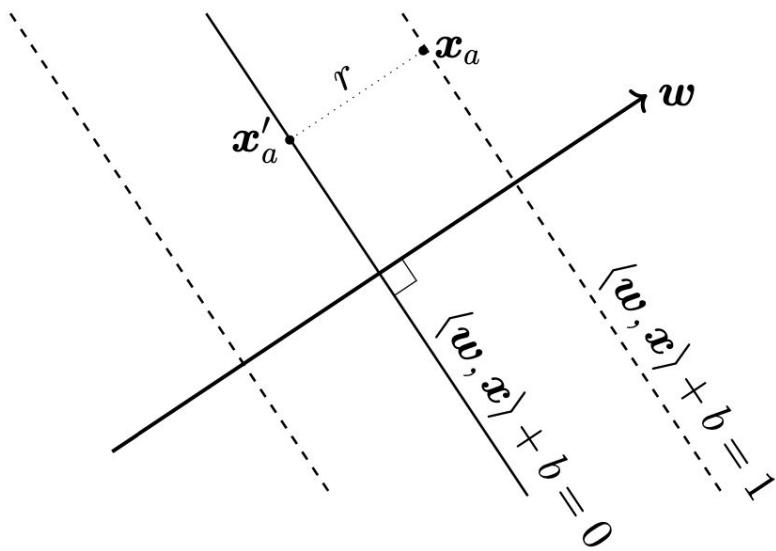




$$\mathbf{x}_a = \mathbf{x}'_a + r \frac{\mathbf{w}}{\|\mathbf{w}\|}.$$

$$y_n(\langle \pmb{w},\pmb{x}_n\rangle+b)\geqslant r\,.$$

$$\|\pmb{w}\| = 1,$$



$$\max_{\mathbf{w}, b, r} \underbrace{r}_{\text{margin}}$$

subject to $\underbrace{y_n(\langle \mathbf{w}, \mathbf{x}_n \rangle + b) \geq r}_{\text{data fitting}}, \underbrace{\|\mathbf{w}\| = 1}_{\text{normalization}}, r > 0,$

$$\langle \pmb{w},\pmb{x}'_a\rangle+b=0\,.$$

$$\left\langle \pmb{w},\pmb{x}_a-r\frac{\pmb{w}}{\|\pmb{w}\|}\right\rangle+b=0\,.$$

$$\langle \pmb{w},\pmb{x}_a\rangle+b-r\frac{\langle \pmb{w},\pmb{w}\rangle}{\|\pmb{w}\|}=0\,.$$

$$r=\frac{1}{\|\pmb{w}\|}\,.$$

$$\max_{\mathbf{w}, b} \quad \frac{1}{\|\mathbf{w}\|}$$

subject to $y_n(\langle \mathbf{w}, \mathbf{x}_n \rangle + b) \geq 1 \quad \text{for all } n = 1, \dots, N.$

$$\min_{\mathbf{w}, b} \quad \frac{1}{2} \|\mathbf{w}\|^2$$

subject to $y_n(\langle \mathbf{w}, \mathbf{x}_n \rangle + b) \geq 1 \quad \text{for all } n = 1, \dots, N.$

$$\begin{aligned} \max_{\mathbf{w}, b, r} \quad & \underbrace{r}_{margin} \\ \text{subject to} \quad & \underbrace{y_n(\langle \mathbf{w}, \mathbf{x}_n \rangle + b)}_{data fitting} \geq r, \quad \underbrace{\|\mathbf{w}\| = 1}_{normalization}, \quad r > 0, \end{aligned} \tag{12.20}$$

is equivalent to scaling the data, such that the margin is unity:

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \underbrace{\frac{1}{2} \|\mathbf{w}\|^2}_{margin} \\ \text{subject to} \quad & \underbrace{y_n(\langle \mathbf{w}, \mathbf{x}_n \rangle + b)}_{data fitting} \geq 1. \end{aligned} \tag{12.21}$$

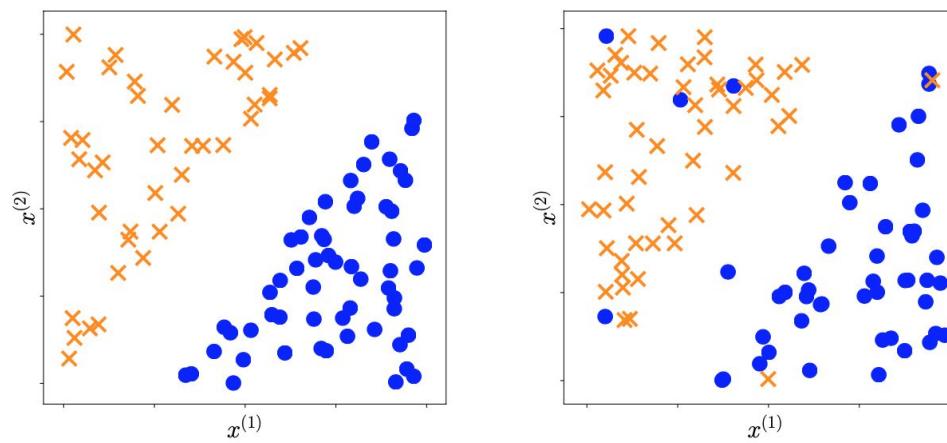
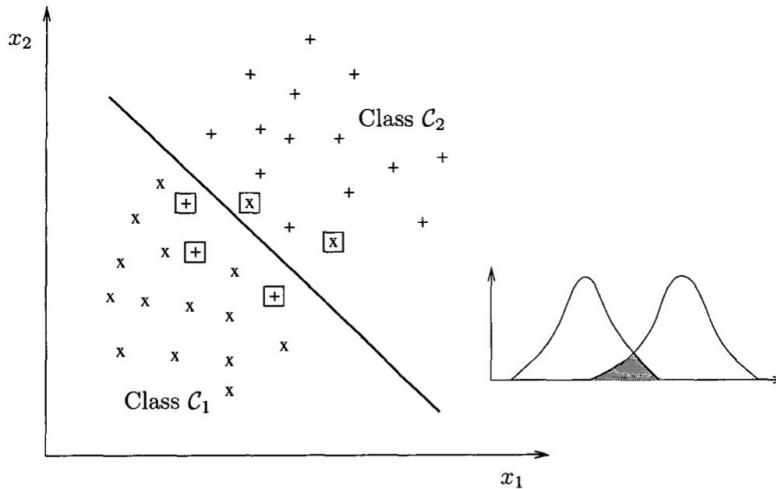
$$\begin{aligned} & \max_{\mathbf{w}', b, r} \quad r^2 \\ \text{subject to} \quad & y_n \left(\left\langle \frac{\mathbf{w}'}{\|\mathbf{w}'\|}, \mathbf{x}_n \right\rangle + b \right) \geq r, \quad r > 0. \end{aligned} \tag{12.22}$$

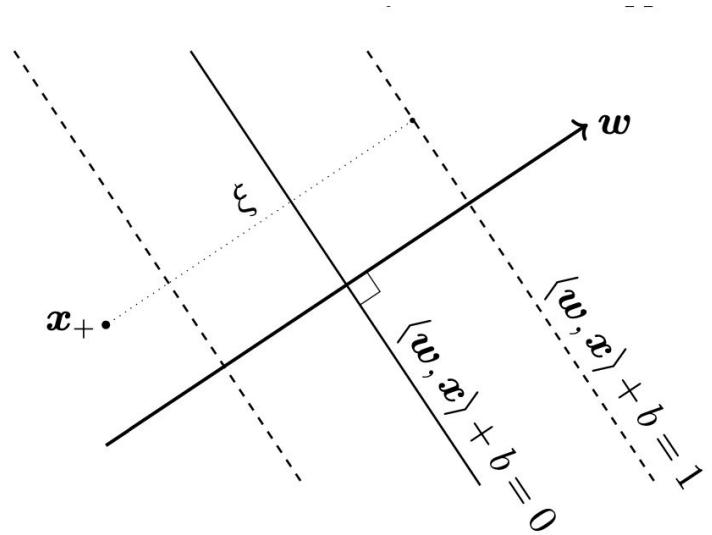
$$\begin{aligned} & \max_{\mathbf{w}', b, r} \quad r^2 \\ \text{subject to} \quad & y_n \left(\underbrace{\left\langle \frac{\mathbf{w}'}{\|\mathbf{w}'\| r}, \mathbf{x}_n \right\rangle}_{\mathbf{w}''} + \underbrace{\frac{b}{r}}_{b''} \right) \geq 1, \quad r > 0 \end{aligned}$$

$$\|\boldsymbol{w}''\| = \left\| \frac{\boldsymbol{w}'}{\|\boldsymbol{w}'\| r} \right\| = \frac{1}{r} \cdot \left\| \frac{\boldsymbol{w}'}{\|\boldsymbol{w}'\|} \right\| = \frac{1}{r}.$$

$$\max_{\boldsymbol{w}'', b''} \quad \frac{1}{\|\boldsymbol{w}''\|^2}$$

subject to $y_n (\langle \boldsymbol{w}'', \boldsymbol{x}_n \rangle + b'') \geq 1.$





$$\min_{\boldsymbol{w}, b, \xi} \quad \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{n=1}^N \xi_n$$

subject to $y_n(\langle \boldsymbol{w}, \boldsymbol{x}_n \rangle + b) \geq 1 - \xi_n$
 $\xi_n \geq 0$

$$\mathfrak{L}(\mathbf{w}, b, \xi, \alpha, \gamma) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi_n \quad (12.34)$$

$$-\underbrace{\sum_{n=1}^N \alpha_n (y_n (\langle \mathbf{w}, \mathbf{x}_n \rangle + b) - 1 + \xi_n)}_{\text{constraint (12.26b)}} - \underbrace{\sum_{n=1}^N \gamma_n \xi_n}_{\text{constraint (12.26c)}}.$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = \boldsymbol{w}^\top - \sum_{n=1}^N \alpha_n y_n {\boldsymbol{x}_n}^\top,$$

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{n=1}^N \alpha_n y_n,$$

$$\frac{\partial \mathcal{L}}{\partial \xi_n} = C - \alpha_n - \gamma_n.$$

$$\boldsymbol{w} = \sum_{n=1}^N \alpha_n y_n \boldsymbol{x}_n,$$

$$\mathfrak{D}(\xi, \alpha, \gamma) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \sum_{i=1}^N y_i \alpha_i \left\langle \sum_{j=1}^N y_j \alpha_j \mathbf{x}_j, \mathbf{x}_i \right\rangle$$

$$+ C \sum_{i=1}^N \xi_i - b \sum_{i=1}^N y_i \alpha_i + \sum_{i=1}^N \alpha_i - \sum_{i=1}^N \alpha_i \xi_i - \sum_{i=1}^N \gamma_i \xi_i .$$

$$\mathfrak{D}(\xi, \alpha, \gamma) = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^N \alpha_i + \sum_{i=1}^N (C - \alpha_i - \gamma_i) \xi_i .$$

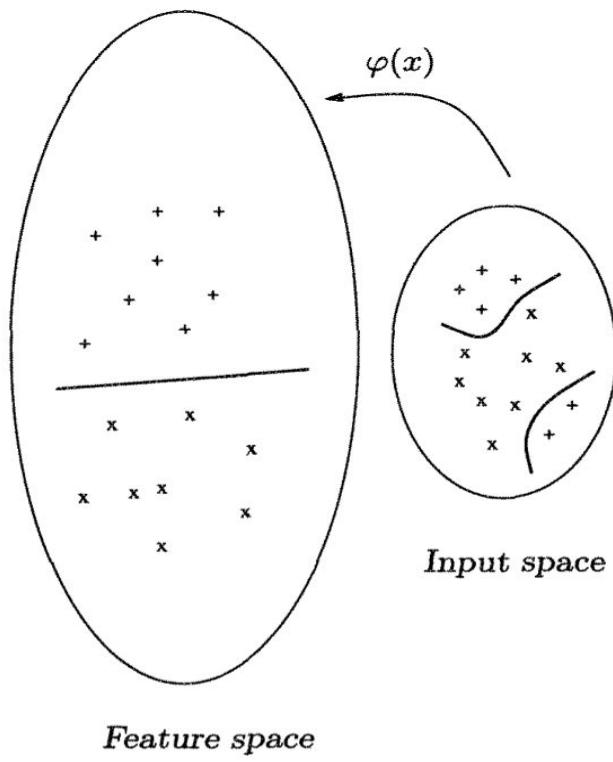
$$\begin{aligned}
\min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \sum_{i=1}^N \alpha_i \\
\text{subject to} \quad & \sum_{i=1}^N y_i \alpha_i = 0 \\
& 0 \leq \alpha_i \leq C \quad \text{for all } i = 1, \dots, N .
\end{aligned} \tag{12.41}$$

$$K(x, z) = \sum_{i=1}^{n_{\mathcal{H}}} \lambda_i \phi_i(x) \phi_i(z),$$

Mercer's condition requires that

$$\int K(x, z) g(x) g(z) dx dz \geq 0$$

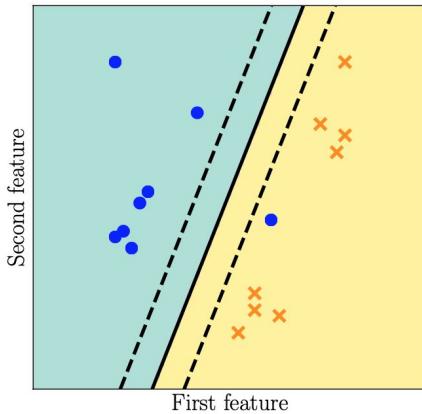
$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{\phi}(\boldsymbol{x}_i), \boldsymbol{\phi}(\boldsymbol{x}_j) \rangle_{\mathcal{H}} .$$



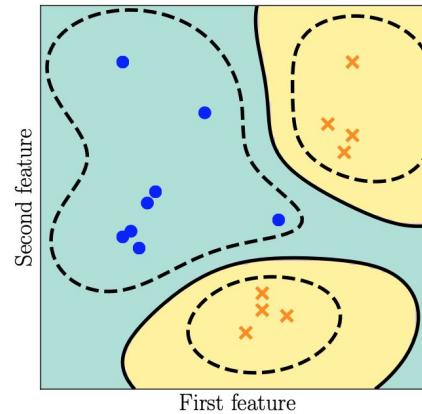
$$y(x) = \text{sign}[w^T \varphi(x) + b]$$

$$\begin{aligned} \min_{w,b,\xi} J_P(w, \xi) = & \frac{1}{2} w^T w + c \sum_{k=1}^N \xi_k \\ \text{such that } & y_k [w^T \varphi(x_k) + b] \geq 1 - \xi_k, \quad k = 1, \dots, N \\ & \xi_k \geq 0, \quad k = 1, \dots, N. \end{aligned}$$

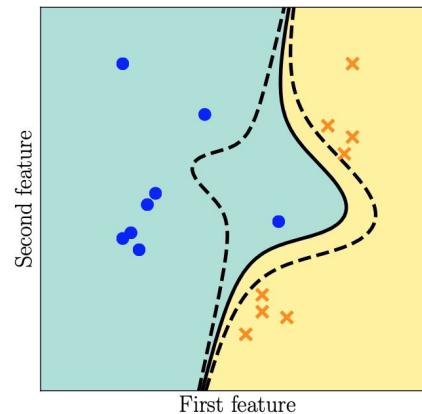
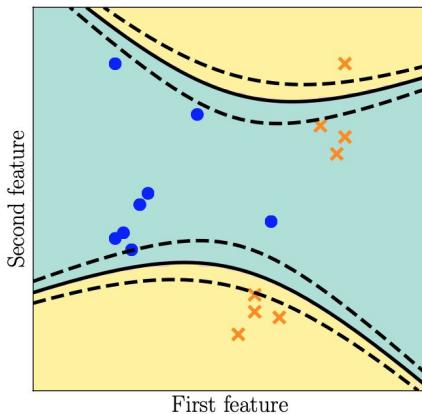
$$\begin{aligned} \max_{\alpha} J_D(\alpha) = & -\frac{1}{2} \sum_{k,l=1}^N y_k y_l K(x_k, x_l) \alpha_k \alpha_l + \sum_{k=1}^N \alpha_k \\ \text{such that } & \sum_{k=1}^N \alpha_k y_k = 0 \\ & 0 \leq \alpha_k \leq c, \quad k = 1, \dots, N. \end{aligned}$$



(a) SVM with linear kernel



(b) SVM with RBF kernel



Thank you for your attention!