Multivariate Verfahren

5. Distance and Similarity Measures

Hannah Schulz-Kümpel

Institut für Statistik, LMU München

Summer Semester 2024

We recall that

- For some space $\mathcal{S}$, we call a function

$$d : \mathcal{S} \times \mathcal{S} \longrightarrow \mathbb{R}$$

a **distance**, if it fulfils the following three requirements $\forall a, b \in \mathcal{S}$:

1. $d(a,b) = 0 \Leftrightarrow a = b$
2. $d(a,b) \geq 0$
3. $d(a,b) = d(b,a)$

- The definition of the euclidean distance is, for some $p \in \mathbb{N}$,

$$d_{\text{euclid}} : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}, \quad (\boldsymbol{a}, \boldsymbol{b}) \longmapsto \sqrt{\sum_{i=1}^{p} (a_i - b_i)^2}\,.$$

Metric (spaces) I

- To be more precise, the previous definition is *sometimes* given for metric spaces, other times *distance* is used as a synonym for *metric*.

- Let's say that in our previous definition, we also required d to fulfill the triangle inequality (which is how we get a *metric*), meaning that for $a, b, c \in \mathcal{S}$
$$d(a,c) \leq d(a,b) + d(b,c)\,. \tag{1}$$

- Then, the positivity assumption becomes redundant, because
$$0 \overset{\text{by (1)}}{=} d(a,a) \leq d(a,b) + d(b,a) \overset{\text{by (3)}}{=} 2d(a,b)\,,$$
so it has to hold that $d(a,b) \geq 0 \; \forall a, b \in \mathcal{S}$!

Metric (spaces) II

Definition (Metric (space))

Given some space $\mathcal{S}$, we call a function $d : \mathcal{S} \times \mathcal{S} \longrightarrow \mathbb{R}$ **metric** (or sometimes **distance**), if it fulfils the following three requirements $\forall a, b, c \in \mathcal{S}$:

1. $d(a, b) = 0 \Leftrightarrow a = b$
2. $d(a, b) = d(b, a)$ (Symmetry)
3. $d(a, c) \leq d(a, b) + d(b, c)$ (Triangle Inequality).

The pair $(\mathcal{S}, d)$ is then called a **metric space**.

Metric (spaces) III

- Fun fact: Metric spaces generalize the concept of the "real line" $\mathbb{R}$ in calculus!
- In all kinds of settings, we will use them as the basis to formulate a mathematical question.
- Another, somewhat similar general concept are **normed vector spaces**, where the *norm* gives the length of a vector.

Connection between norms and distances/metrics I

Definition (credited to John K. Hunter)

A normed vector space $(\mathcal{S}, \|\cdot\|)$ is a vector space X (which we assume to be real) together with a function

$$\|\cdot\| : \mathcal{S} \to \mathbb{R}\,,$$

called a **norm on $\mathcal{S}$**, such that for all $x, y \in \mathcal{S}$ and $k \in \mathbb{R}$:

1. $0 \leq \|x\| < \infty$ and $\|x\| = 0$ if and only if $x = 0$;
2. $\|kx\| = |k|\|x\|$
3. $\|x + y\| \leq \|x\| + \|y\|$.

Connection between norms and distances/metrics II

A norm on a vector space will always give rise to a metric on the same vector space by taking the norm of the difference between two vectors.

Proposition

If $(\mathcal{S}, \|\cdot\|)$ is a normed vector space, then

$$d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}\,, \quad (x, y) \mapsto \|x - y\|$$

is a metric on $\mathcal{S}$.

Proof.

The metric-properties of $\|x - y\|$ follow immediately from the norm-properties - check it yourself :) □

Some examples of norms for $\mathcal{S} = \mathbb{R}^m$ I

- p-**norm**:

$$\|x\|_p := (|x_1|^p + |x_2|^p + \cdots + |x_m|^p)^{\frac{1}{p}}$$

- **Euclidean norm**, which is equal to the 2-norm:

$$\|x\| := \sqrt{x_1^2 + x_2^2 + \cdots + x_m^2}$$

  Note: The euclidean norm clearly gives rise to the euclidean distance d_{euclid}, which is also a metric, using the previous proposition.

- **Maximum norm**:

$$\|x\|_{\text{max}} := \max\left\{|x_1|, |x_2|, \ldots, |x_m|\right\}$$

Some examples of norms for $\mathcal{S} = \mathbb{R}^m$ II

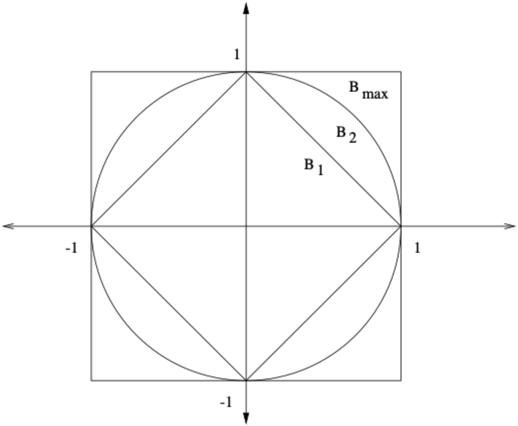


Figure: The unit balls in $\mathbb{R}^2$ for the Euclidean norm (B_2), the 1-norm (B_1) and the maximum norm (B_{max}). Source: https://www.math.ucdavis.edu/ hunter/book/ch1.pdf

What do Statisticians use metrics for?

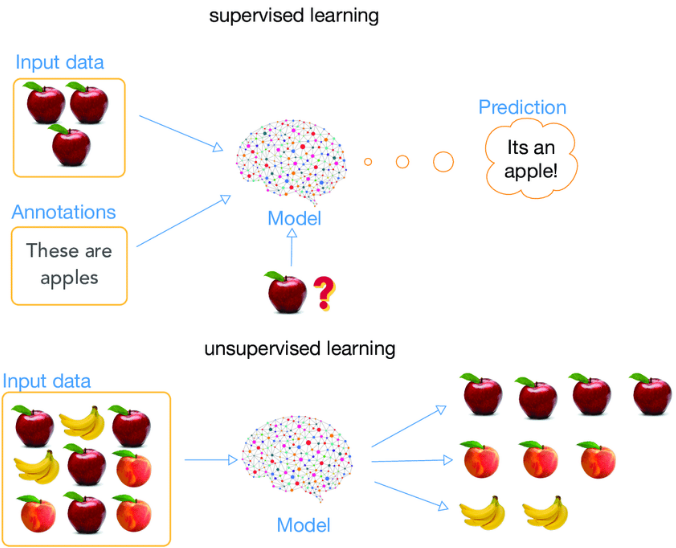
- Clearly, metrics quantify some notion of distance between mathematical objects - which we basically need everywhere, all the time!

- Just some examples:

  1. We recall that we may view estimates of *variance* as *inertia* or *spread around the center of gravity* - for which we require a definition of distance between our points!

  2. Anytime we want to optimize something w.r.t. *distance/loss* - both in supervised and unsupervised learning!

Supervised vs. unsupervised learning: an overview



Examples of using loss in (un)supervised learning

- **Parameter estimation using OLS!** Here, we want to minimize over the squared loss (which is clearly a metric) to estimate a parameter $\longrightarrow$ *supervised learning*.

- **Clustering according to some loss (metric)** in *unsupervised learning* - we need to know how close points are to, e.g., "means".

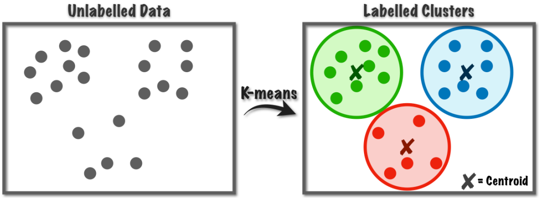


Figure: Example of Clustering; Source: Towards Data Science

Relevant spaces to consider metrics on

- **Spaces of "data points".** Here, there are actually different cases we need to consider separately :
  1. *metric data points* taking values in $\mathbb{R}^m$, $m \in \mathbb{N}$.
  2. *categorical data points* taking values in $\Omega_1 \times \cdots \times \Omega_m$, $m \in \mathbb{N}$ - where $\Omega_i = \{\omega_1^{(i)}, \omega_2^{(i)} \dots\}$ denotes the set of values the ith variable/entry of the data point could take.
  3. *mixed data points*, where some entries/variables are metric and others categorical.
- **Function spaces.** These are, for example, relevant in nonparametric statistics, where the parameters are functions.
- **Spaces of probability measures.** These have a host of applications, from parameter estimation to proof of convergences etc.

Basic metrics for metric data (on $\mathbb{R}^m$) I

Using the proposition from slide 7 and the norms from slide 9, we immediately get the following metrics:

- p-**metric**:

$$d_p(x,y) := \|x-y\|_p = \left(\sum_{i=1}^{m} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

- **Euclidean distance**, which is equal to the 2-metric:

$$d_{\text{euclid}}(x,y) := \|x-y\| = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

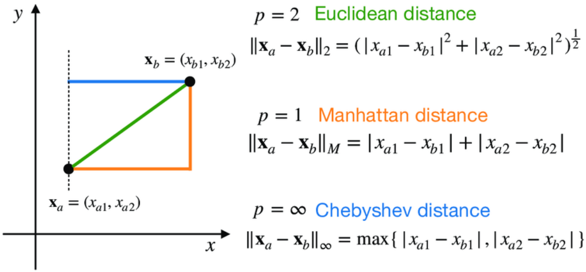
- **Chebyshev distance** (induced by the Maximum norm):

$$d_{\text{Chebyshev}}(x,y) := \|x-y\|_{\max} = \max\left\{|x_1-y_1|, |x_2-y_2|, \ldots, |x_m-y_m|\right\}$$

Basic metrics for metric data (on $\mathbb{R}^m$) II

- Furthermore, the 1-metric is referred to as **Manhattan distance**:

$$d_{\text{Manhattan}}(x, y) := \sum_{i=1}^{m} |x_i - y_i|$$



From: Fu, Chen & Yang, Jianhua. (2021). Granular Classification for Imbalanced Datasets: A Minkowski Distance-Based Method. Algorithms. 14. 54. 10.3390/a14020054.

Distances on $\mathbb{R}^m$ that are not quite metrics I

- Sometimes to quantify standardized distances between points, the following *similarity approach* is used:

  1. Define a similarity measure $s : \mathcal{S} \times \mathcal{S} \longrightarrow \mathbb{R}$, s.t.

$$\delta(a, a) = 1 \; \forall a \in \mathcal{S} \quad \& \quad \delta(a, b) = \delta(b, a) \; \forall a, b \in \mathcal{S}$$

  2. Define the distance function

$$d : \mathcal{S} \times \mathcal{S} \longrightarrow \mathbb{R}, \quad (a, b) \longmapsto 1 - \delta(a, b)\,.$$

- One example would be using the *Pearson correlation coefficient* as δ to get the **Pearson correlation distance**.

Distances on $\mathbb{R}^m$ that are not quite metrics II

- Another, quite popular choice for similarity measure is the **cosine similarity**

$$\delta_{\mathsf{cos}} : \mathbb{R}^m \times \mathbb{R}^m \longrightarrow [-1, 1], \quad (x, y) \longmapsto \frac{\sum_{i=1}^{n} x_i y_i}{\|x\| \|y\|} \,,$$

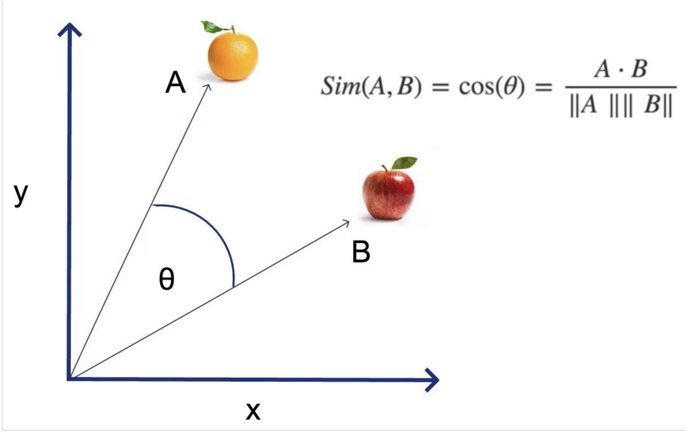
which gives rise to the **cosine distance**

$$d_{\mathsf{cos}}(x, y) := 1 - \delta_{\mathsf{cos}}(x, y) \,.$$

- The cosine distance is often used in the context of data mining; for instance in information retrieval and text mining, where each word is assigned a unique coordinate.

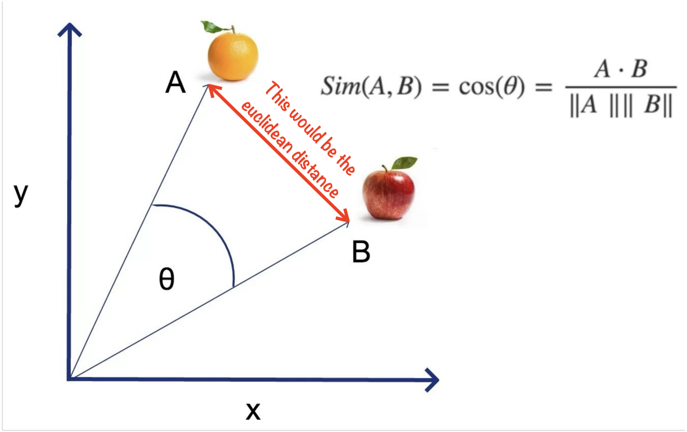
Distances on $\mathbb{R}^m$ that are not quite metrics III

- Then, the distance depends not on the length of the vectors, but on the the angle between them w.r.t. the center of the coordinate space.



Distances on $\mathbb{R}^m$ that are not quite metrics III

- Then, the distance depends not on the length of the vectors, but on the the angle between them w.r.t. the center of the coordinate space.



Basic approach to metrics for categorical data points

- For categorical data points taking values in $\Omega = \Omega_1 \times \cdots \times \Omega_m$, $m \in \mathbb{N}$, the simplest and most popular metric is what is often referred to as **0-1 loss**:

$$L : \Omega \times \Omega \longrightarrow \{0, 1\}, \quad (x, y) \longmapsto \begin{cases} 0, & \text{if } x = y, \\ 1, & \text{otherwise.} \end{cases}$$

- For *ordered* categorical data points, i.e. when $\omega_1 \ll \omega_2 \ll \cdots \ll \omega_m$ we could expand the concept of $0 - 1$ loss and replace "1, otherwise" with a different distance value for each pair of possible values, as long as the pair (ω_1, ω_m) gets assigned the largest distance value and so on.

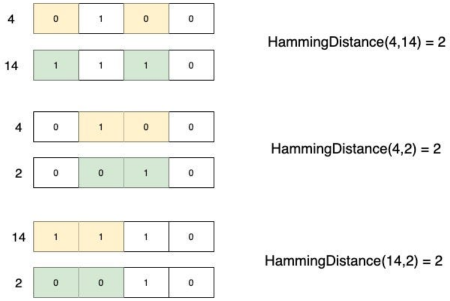
- Additionally, if a distance instead of a metric suffices, we may use the *similarity approach* from slide 17, taking, e.g., a correlation coefficient for categorical variables as δ.

Pairwise comparison:Hamming and Levenshtein distances I

- Another approach to comparing categorical data points is to *pairwise compare each element*.
- This approach is especially popular in information theory, linguistics, and computer science.
- In this context, the **Hamming** and **Levenshtein distances** are especially popular.
- While we will not consider the exact definitions, the following provides an intuition for both:

Pairwise comparison:Hamming and Levenshtein distances II

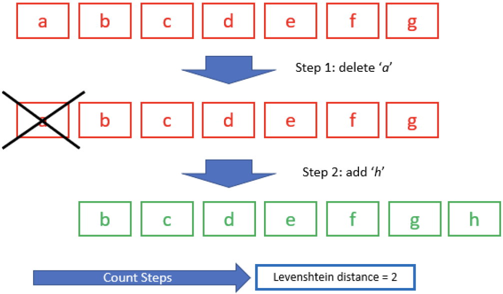
- The *Hamming distance* quantifies the the number of positions at which the elements of our categorical points are different.
- Here is an example from Medium:



Pairwise comparison:Hamming and Levenshtein distances III

- Meanwhile, the *Levenshtein distance* can even compare points of different lengths! It quantifies the minimum of changes (including deletions) necessary to change one point into the other.

- Here is an example from Towards Data Science:



What if we have mixed data points?

- Metrics defined specifically for mixed data are not a focus of this class, but here is one exemplary suggestions:

- **Use a weighted sum of distances:** Let's say that we can "divide" a mixed data point into L parts, for each of which we have a suitable distance function at hand. Then we may define

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{l=1}^{L} d_l(\boldsymbol{x}[l], \boldsymbol{y}[l]) \cdot w_l \,,$$

  but mind the scaling of variables within our parts and define the weights carefully to achieve a meaningful result.

Function spaces

- Sometimes, we will want to quantify the distance between functions.
- In applied statistics, this is mostly the case when the parameter of a model is a function instead of a finite-dimensional vector (*Nonparametric inference*).
- How do we define function spaces? Well usually, we say that some function space is *the set of all functions that*
    1. map from the same space (preimage)
    2. to the same space (image)
    3. fulfill some additional characteristics.

Some popular norms and metrics on function spaces I

- For a function $f : X \longrightarrow Y$, the **supremum norm** is defined as

$$\|f\|_\infty := \sup\left\{|f(x)| : x \in X\right\}$$

which, for another function $g : X \longrightarrow Y$ gives rise to the following metric:

$$\|f - g\|_\infty := \sup\left\{|f(x) - g(x)| : x \in X\right\}.$$

- *What would we need from a function space on which this norm and metric are defined?*
Definitely, that for any function in the space $|f(x)| < \infty \; \forall x \in X$.

  - *For example:*
  $B(\mathbb{R}) := \{f : R \longrightarrow \mathbb{R} | \exists M \in \mathbb{R} : \; f(x) \leq M \;\; \forall x \in X\}$.

Some popular norms and metrics on function spaces II

- For a function $f : X \longrightarrow Y$, the **p-norm** is defined as

$$\|f\|_p := \left(\int_X |f(x)|^p dx\right)^{\frac{1}{p}}$$

which, for another function $g : X \longrightarrow Y$ gives rise to the following metric:

$$\|f - g\|_p := \left(\int_X |f(x) - g(x)|^p dx\right)^{\frac{1}{p}} .$$

- *What would we need from a function space on which this norm and metric are defined?* Definitely, that for any function in the space $\int_X |f(x)|^p dx < \infty \; \forall x \in X$.

  ▶ *For example:* $C(K) := \{f : K \longrightarrow \mathbb{R} | f \text{ is continuous}\}$ for some compact set K such as $[a, b]$, $a, b \in \mathbb{R}$.

Metrics for random variables/distributions

- Next we will look at two popular metrics, plus a "distance-like" function (does not satisfy all requirements), for probability measures.

- Note that the elements that we are comparing are probability measures, and we need to define a suitable space for them.

- Let Ω denote some specific sample space equipped with σ-algebra $\mathcal{F}$.

- For the following slides, we will assume that all mentioned probability measures are elements of the following set:

$$\mathcal{S}\big(\Omega, \mathcal{F}\big) = \Big\{\mu : \quad \mu \text{ is a probability measure on } (\Omega, \mathcal{F})\mu\Big\}.$$

Total Variation (TV) distance

- The *total variation (TV) distance* is a metric for probability measures which quantifies the largest absolute difference between the probabilities that the two probability distributions assign to the same event.
- For two probability measure μ and ν defined on the same probability space $(\Omega, \mathcal{F})$, it is defined as

$$D_{\mathsf{TV}}(\mu, \nu) = \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)| \, .$$

- Note that this metric again organically arises from the TV-norm $\|\mu\|_{\mathsf{TV}} := \sup_{A \in \mathcal{F}} |\mu(A)|$ on the space $\mathcal{S}\big(\Omega, \mathcal{F}\big)$ we previously defined.

Wasserstein metric

- The *Wasserstein metric* is a metric for probability measures.
- *Very broadly*, we may interpret this metric as quantifying the "minimum cost" of transforming one probability measure into another.
- Specifically, the p-Wasserstein metric between probability measures μ and ν is defined as

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\Omega \times \Omega} d(x, y)^p d\gamma(x, y) \right)^{1/p},$$

where $\Gamma(\mu, \nu)$ is the set of possible probability measures on $\Omega \times \Omega$, so that μ and ν exist as marginal distributions and d is a suitable metric on Ω.

Kullback–Leibler (KL) divergence I

- The Kullback-Leibler (KL) divergence, also known as relative entropy, is a measure of how one probability distribution diverges from a second probability distribution.
- While it is often used to quantify how different an estimate of a probability distribution is from the probability distribution we theoretically expected (or assume to be true).
- Note, however, that the KL divergence *is not a metric* because it does not fulfill the requirement of symmetry!

Kullback–Leibler (KL) divergence II

Definition (Kullback–Leibler (KL) divergence)

For two probability measures μ and ν on a space $\mathcal{X}$, the *Kullback–Leibler divergence*is defined as

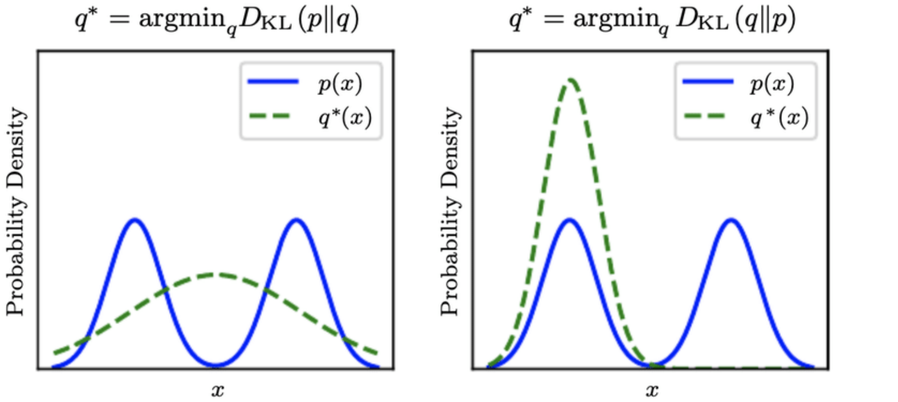
$$D_{\mathsf{KL}}(\mu||\nu) = \int_{\mathcal{X}} \log\left(\frac{\mu(x)}{\nu(x)}\right) d\mu(x)$$

$$= \begin{cases} \int_{-\infty}^{\infty} \mu(x) \cdot \log\left(\frac{\mu(x)}{\nu(x)}\right) dx, & \text{if } \mu \text{ and } \nu \text{ are defined} \\ & \text{by continuous distributions;} \\ \sum_{x \in M} \mu(x) \cdot \log\left(\frac{\mu(x)}{\nu(x)}\right), & \text{if } \mu \text{ and } \nu \text{ are defined} \\ & \text{by discrete distributions.} \end{cases}$$

Kullback–Leibler (KL) divergence III

Visualization

Source: Yang, Xuxi & Duvaud, Werner & Wei, Peng. (2020). Continuous Control for Searching and Planning with a Learned Model.



KL divergence and maximum likelihood I

Interestingly, maximum likelihood estimation (in parametric, i.i.d. settings), asymptotically, amounts to minimizing the KL divergence between the "true" assumed distribution and the estimated distribution.

Per definition, we have that

$$D_{KL}[P(x|\theta_0) \,\|\, P(x|\hat{\theta})] = \mathbb{E}_{x \sim P(x|\theta_0)} \left[\log \frac{P(x|\theta_0)}{P(x|\hat{\theta})} \right]$$

$$= \mathbb{E}_{x \sim P(x|\theta_0)} \left[\log\, P(x|\theta_0) - \log\, P(x|\hat{\theta}) \right]$$

$$= \mathbb{E}_{x \sim P(x|\theta_0)} \left[\log\, P(x|\theta_0) \right] - \mathbb{E}_{x \sim P(x|\theta_0)} \left[\log\, P(x|\hat{\theta}) \right]$$

KL divergence and maximum likelihood II

$$\Longrightarrow \arg\min_{\hat{\theta}} D_{KL}[P(x|\theta_0) \,\|\, P(x|\hat{\theta})] = \arg\max_{\hat{\theta}} \mathbb{E}_{x\sim P(x|\theta_0)}\left[\log P(x|\hat{\theta})\right]$$

The *law of large numbers (LLN)* gives us that

$$\lim_{n\longrightarrow\infty} \frac{1}{n}\sum_{i=1}^{n} \log\, P(x_i|\hat{\theta}) = \mathbb{E}_{x\sim P(x|\theta_0)}\left[\log\, P(x|\hat{\theta})\right] .$$

And, therefore, we have

$$\begin{aligned}\arg\min_{\hat{\theta}} D_{KL}[P(x|\theta_0) \,\|\, P(x|\hat{\theta})] &= \arg\max_{\hat{\theta}} \frac{1}{n}\sum_{i=1}^{n} \log\, P(x_i|\hat{\theta}) \\ &= \arg\max_{\hat{\theta}} \log\, P(x_i|\hat{\theta}) \\ &= \arg\max_{\hat{\theta}} P(x_i|\hat{\theta}) = \hat{\theta}_{\mathsf{ML}}\ !\end{aligned}$$