Basic models in machine learning

Pierre Gaillard

Sept. 26 2022

1 Introduction to supervised learning

Let's start with an example of a practical problem. In order to better optimize its production, a producer is interested in modeling electricity consumption in France as a function of temperature (cf. Figure 1).

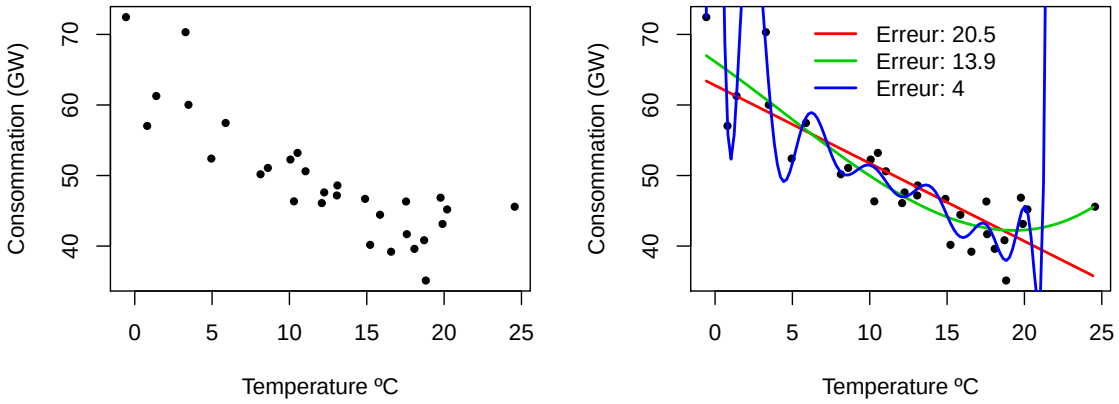


Figure 1: French power consumption (GW) as a function of temperature (°C). To the right are plotted error minimizing functions for polynomial spaces of degrees 1 (red), 3 (green) and 30 (blue).

The objective is to find a function f such that it explains well the power consumption $(y_i)_{1\leqslant i\leqslant n}$ as a function of temperature $(x_i)_{1\leqslant i\leqslant n}$, that is $y_i \approx f(x_i)$. To do this, we can choose a function space $\mathcal{F}$ and solve the empirical risk minimization problem:

$$\widehat{f}_n \in \operatorname*{arg\,min}_{f\in\mathcal{F}} \widehat{\mathcal{R}}(f) := \operatorname*{arg\,min}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 . \tag{1}$$

Care must be taken when selecting the function space to avoid over-fitting (see Figures 1). Although the empirical mean square error decreases when the $\mathcal{F}$ space becomes larger (larger polynomial degrees), the $\widehat{f}_n$ estimator loses its predictive power. The question is: will $\widehat{f}_n$ perform well on new data?

Supervised learning: general setup and notation

Goal. In supervised machine learning, the goal is given some observations $(x_i, y_i) \in \mathcal{X}\times\mathcal{Y}$ of inputs/outputs and given a new input $x \in \mathcal{X}$ to predict well the next output $y \in \mathcal{Y}$. The training

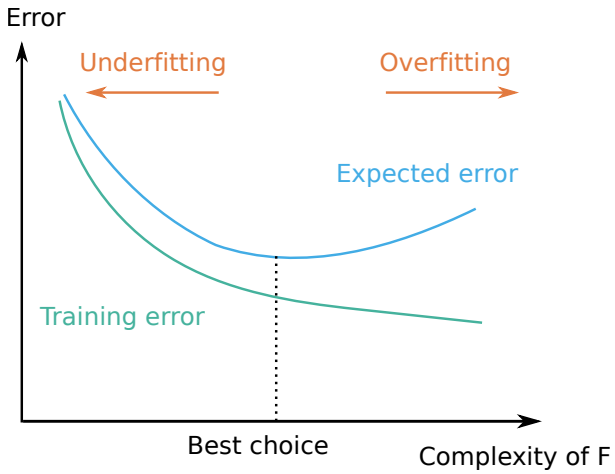


Figure 2: Over-fitting and under-fitting according to the complexity of $\mathcal{F}$. In blue the risk $\mathcal{R}(f)$ which we want to minimize, in green the empirical risk $\widehat{\mathcal{R}}(f)$ that we observe on the training data.

data set will be denoted $D_n := \{(x_i, y_i), i = 1, \dots, n\}$. We will often make the assumption that the observations (x_i, y_i) are realizations of i.i.d. random variables from a distribution ν.

The distribution ν is unknown to the statistician, it's a matter of learning it from the D_n data. A learning rule $\mathcal{A}$ is a function that associates to training data D_n a prediction function $\widehat{f}_n$ (the hat on f indicates that it is an estimator):

$$\begin{array}{rccc} \mathcal{A} : \cup_{n\in\mathbb{N}} (\mathcal{X}\times\mathcal{Y})^n & \to & \mathcal{Y}^{\mathcal{X}} \\ D_n & \mapsto & \widehat{f}_n \end{array} .$$

The estimated function $\widehat{f}_n$ is constructed to predict a new output y from a new x, where (x, y) is a pair of *test data*, i.e. not observed in the training data. The function $\widehat{f}_n$ is an estimator because it depends on the data D_n and not on unobserved parameter (such as ν). If D_n is random, it is a random function.

Risk and empirical risk. The objective is to find an estimator $\widehat{f}_n$ that predicts well new data by minimizing the risk:

$$\mathcal{R}(\widehat{f}_n) := \mathbb{E}\Big[\big(y - \widehat{f}_n(x)\big)^2 \,\Big|\, D_n \Big] \qquad \text{where} \qquad (x, y) \sim \nu \,. \tag{Risk}$$

However, the statistician cannot compute the expectation (and thus the risk) because he does not know ν. A common method in supervised machine learning is therefore to replace the risk with the empirical risk.

$$\widehat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 \,. \tag{Empirical risk}$$

However, one must be careful about over-fitting (case where $\widehat{\mathcal{R}}(f)$ is much lower than $\mathcal{R}(f)$, see Figure 2). In this class, we will study the performance of the least square estimator in the case of the linear model.

2 Linear least-squares regression

In this section, we study the simple but still widely used problem of linear least-squares regression. The linear regression problem can be traced back to Legendre (1805) and Gauss (1809). The word "regression" is said to have been introduced by Galton in the 19th century. By modeling the size of individuals according to that of their fathers, Galton observed a return (regression) towards average height. Larger-than-average fathers tend to have smaller children and vice versa for smaller fathers.

Here, we consider real outputs ($\mathcal{Y} = \mathbb{R}$) and square loss $\ell(y, z) = (y - z)^2$. Given a parametrized family of prediction function $\mathcal{F} := \{f_\theta : \mathcal{X} \to \mathcal{Y}, \theta \in \Theta\}$, we minimize the empirical risk

$$\widehat{\mathcal{R}}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \big(y_i - f_\theta(x_i)\big)^2 .$$

In linear least-square regression, the functions $\theta \mapsto f_\theta(x)$ are assumed to be linear in θ.

⚠ Being linear in θ or x is different. Nothing forces $f_\theta(x)$ to be linear in x. Typically,

$$f_\theta(x) = \langle \theta, \varphi(x) \rangle$$

for some feature map $\varphi(x) \in \mathbb{R}^d$. For example, affine functions may be obtained with $\varphi(x) = (x^\top, 1)^\top$ and polynomials with $\varphi(x) = \big(1, x_1, x_2, x_1^2, x_2^2, x_1 x_2, \dots\big)^\top$. In Figure 1, we have in this way minimized the empirical risk on polynomial spaces of degree 1 (linear model), 3 and 30. We can see that we must be careful not to consider spaces that are too large, at the risk that the model is badly posed (design matrix non injective as seen thereafter). Conversely, for the statistical analysis that we will see next to be verified, one must be in the true model $y = \langle \varphi(x), \theta^* \rangle +$ centered noise. We must therefore make sure that $\varphi(x)$ contains enough descriptors so that the dependency between y and $\varphi(x)$ is indeed linear. Otherwise we pay an additional bias term.

Why should we study linear regression?

- It captures many concepts of learning theory: bias-variance trade-off, need of regularization,...
- It is simple: the analysis can be done in basics maths (linear algebra).
- Using non-linear features, it can be extended to non-linear predictions $\mapsto$ kernel methods.

Matrix notation The empirical risk can be rewritten in matrix notation. Let $y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ be the vector of outputs and $\Phi \in \mathbb{R}^{n \times d}$ the matrix of inputs (also called design matrix or data matrix), which rows are $\varphi(x_i)^\top$:

$$\Phi = \Big(\varphi(x_1), \varphi(x_2), \dots, \varphi(x_n)\Big)^\top \in \mathbb{R}^{n \times d} .$$

The empirical risk is then

$$\widehat{\mathcal{R}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - \langle \theta, \varphi(x_i) \rangle\big)^2 = \frac{1}{n} \big\| y - \Phi\theta \big\|_2^2 . \tag{2}$$

⚠ The matrix notation is very useful to simplify calculation.

2.1 Ordinary Least Squares Estimator (OLS)

In the following, we assume that the design matrix Φ is injective (i.e., the rank of Φ is d). In particular, $d \leqslant n$.

Definition 2.1. *If Φ is injective, the minimizer of the empirical risk*

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} \|y - \Phi\theta\|_2^2,$$

is called the Ordinary Least Squares (OLS) estimator.

Proposition 2.1 (Closed form solution)**.** *If Φ is injective, the OLS exists and is unique. It is given by*

$$\widehat{\theta} = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top y\,.$$

Proof. Since $\widehat{\mathcal{R}}$ is coercive (goes to infinity in infinity) and continuous, it admits at least a minimizer. Furthermore, we have

$$\widehat{\mathcal{R}}(\theta) = \frac{1}{n}\|y - \Phi\theta\|_2^2 = \frac{1}{n}\big(\theta^\top \left(\Phi^\top \Phi\right)\theta - 2\theta^\top \Phi^\top y + \|y\|^2\big).$$

Since $\widehat{\mathcal{R}}$ is differentiable any minimizer should cancel the gradient:

$$\nabla\widehat{\mathcal{R}}(\widehat{\theta}) = \frac{1}{n}\big(\widehat{\theta}^\top(\Phi^\top\Phi) + (\Phi^\top\Phi)\widehat{\theta} - 2\Phi^\top y\big) = \frac{2}{n}\big((\Phi^\top\Phi)\widehat{\theta} - \Phi^\top y\big).$$

where the last equality is because $\Phi^\top\Phi \in \mathbb{R}^{d\times d}$ is symmetric. Since Φ is injective, $\Phi^\top\Phi$ is invertible (Exercise: show this implication). Therefore, a solution of $\nabla\widehat{\mathcal{R}}(\widehat{\theta}) = 0$ satisfies

$$\widehat{\theta} = \left(\Phi^\top\Phi\right)^{-1}\Phi^\top y.$$

However, it remains to check that this is indeed a minimum and therefore that the Hessian is defined as positive, which is the case because: $\nabla^2\widehat{\mathcal{R}}(\widehat{\theta}) = \frac{2}{n}(\Phi^\top\Phi)$. □

Geometric interpretation The linear model seeks to model the output vector $y \in \mathbb{R}^n$ by a linear combination of the form $\Phi\theta \in \mathbb{R}^n$. The image of Φ is the solution space, denoted $\mathrm{Im}(\varphi) = \{z \in \mathbb{R}^n : \exists\theta \in \mathbb{R}^d \text{ s.t. } z = \Phi\theta\} \subseteq \mathbb{R}^n$. This is the vector subspace of $\mathbb{R}^n$ generated by the $d < n$ columns of the design matrix. As $\mathrm{rg}(\Phi) = d$, it is of dimension d.

By minimizing $\|y - \Phi\theta\|$ (cf. Definition 2.1), we thus look for the element of $\mathrm{Im}(\Phi)$ closest to y. This is the orthogonal projection of y on $\mathrm{Im}(\Phi)$, denoted $\widehat{y}$. By definition of the OLS and by the Proposition 2.1, we have:

$$\widehat{y} \overset{\text{Def 2.1}}{=} \Phi\widehat{\theta} \overset{\text{Prop. 2.1}}{=} \Phi(\Phi^\top\Phi)^{-1}\Phi^\top y.$$

In particular, $P_\Phi := \Phi(\Phi^\top\Phi)^{-1}\Phi^\top \in \mathbb{R}^{n\times n}$ is the projection matrix on $\mathrm{Im}(\Phi)$.

Numerical resolution

The closed form formula $\widehat{\theta} = \left(\Phi^\top\Phi\right)^{-1}\Phi^\top y$ from the OLS is useful in analyzing it. However, calculating it naively can be prohibitively expensive. Especially when d is large, one prefers to avoid inverting the design matrix $\Phi^\top\Phi$ which costs $\mathcal{O}(d^3)$ by the Gauss-Jordan method and can be very unstable when the matrix is badly conditioned. The following methods are usually preferred.

QR factorization To improve stability, QR decomposition can be used. Recall that $\widehat{\theta}$ is the solution to the equation:

$$(\Phi^\top \Phi)\widehat{\theta} = \Phi^\top y\,,$$

We write $\Phi \in \mathbb{R}^{n\times d}$ of the form $\Phi = QR$, where $Q \in \mathbb{R}^{n\times d}$ is an orthogonal matrix (i.e., $Q^\top Q = I_d$) and $R \in \mathbb{R}^{d\times d}$ is upper triangular. Upper triangular matrices are very useful for solving linear systems. Substituting in the previous equation, we get:

$$\begin{aligned} R^\top (Q^\top Q) R\widehat{\theta} = R^\top Q^\top y \quad &\Leftrightarrow \quad R^\top R\widehat{\theta} = R^\top Q^\top y \\ &\Leftarrow \quad R\widehat{\theta} = Q^\top y\,. \end{aligned}$$

Then all that remains is to solve a linear system with a triangular upper matrix, which is easy.

Gradient descent We can completely bypass the need of matrix inversion or factorization using gradient descent. It consists in solving the minimization problem step by step by approaching the minimum through gradient steps. For example, we initialize $\widehat{\theta}_0 = 0$, then update:

$$\begin{aligned} \widehat{\theta}_{i+1} &= \widehat{\theta}_i - \eta \nabla \widehat{\mathcal{R}}(\widehat{\theta}_i) \\ &= \widehat{\theta}_i - \frac{2\eta}{n}\big((\Phi^\top \Phi)\widehat{\theta}_i - y^\top \Phi\big)\,, \end{aligned}$$

where $\eta > 0$ is a learning parameter. We see that if the algorithm converges, then it converges to a point canceling the gradient, thus to the OLS solution. To have convergence, the η parameter must be well calibrated, but this is beyond the scope of these notes.

If the data set is much too big, $n \gg 1$. It can also be prohibitively expensive to load all the data to make the $\nabla \widehat{\mathcal{R}}(\widehat{\theta}_i)$ calculation. The common solution is then to do Stochastic Gradient Descent, where gradient steps are made only on estimates of $\nabla \widehat{\mathcal{R}}(\widehat{\theta}_i)$, calculated on a random subset of the data.

2.2 Statistical analysis

In this section, we will provide theoretical guarantees for the OLS. To do so, we will need some probabilistic assumptions.

2.2.1 Stochastic assumptions

Any kind of guarantees requires assumption about how the data is generated. In this section, we consider a stochastic framework that will allow us to analyze the performance of OLS.

Assumption 1 (Linear model)**.** *We assume that there exists a vector $\theta^* \in \mathbb{R}^d$ such that for all $1 \leqslant i \leqslant n$*

$$y_i = \langle \varphi(x_i), \theta^* \rangle + \varepsilon_i\,, \tag{3}$$

where $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top \in \mathbb{R}^n$ is a vector of errors (or noise). The ε_i are assumed to be centered independent variables $\mathbb{E}[\varepsilon_i] = 0$ and with variance $\mathbb{E}[\varepsilon_i] = 0$.

Recall that x_i, y_i and ε_i (from now on) are random variables. The noise ε_i comes from the fact that in practice the observation y_i never completely fits the linear forecast. This is due to noise or unobserved explanatory variables. The Equation (3) can be rewritten in matrix form:

$$y = \Phi\theta^* + \varepsilon$$

where $y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$, $\Phi = (\varphi(x_1), \dots, \varphi(x_n))^\top \in \mathbb{R}^{n \times d}$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top \in \mathbb{R}^n$.
From here, there are two settings of analysis for least squares:

- *Fixed design.* In this setting, the design matrix Φ is not random but deterministic and the features $\varphi(x_1), \dots, \varphi(x_n)$ are fixed. The expectations are thus only with respect to ε_i and y_i and the goal is to minimize the risk

$$\mathcal{R}_\Phi(\theta) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \varphi(x_i)^\top\theta\right)^2\right] = \mathbb{E}\left[\frac{1}{n}\left\|y - \Phi\theta\right\|_2^2\right], \tag{4}$$

  for new random observations y_i (different from the ones observed in the dataset) but on the same inputs.

- *Random design.* Here, both the inputs and the outputs are random. This is the most standard setting of supervised machine learning. The goal is to minimize the risk (sometimes called the generalization error) defined in Equation (Risk).

In this class, we consider the fixed design setting because it eases the notation and the calculation (we only need simple linear algebra).

2.2.2 Bias/variance decomposition

Before analyzing the statistical properties of OLS, we state a general result under the linear model which illustrate the trade-off between estimation and approximation (or bias and variance).

Proposition 2.2 (Risk decomposition)**.** *Under the linear model (Assumption 1) with fixed design, for any $\theta \in \mathbb{R}^d$ it holds*

$$\mathbb{E}\big[\mathcal{R}_\Phi(\theta) - \mathcal{R}_\Phi(\theta^*)\big] = \|\theta - \theta^*\|_\Sigma^2$$

where $\Sigma = \frac{1}{n}\Phi^\top\Phi \in \mathbb{R}^{d\times d}$ and $\|\theta\|_\Sigma^2 = \theta^\top\Sigma\theta$. If θ is a random variable (because it depends on a random data set) then

$$\mathbb{E}\big[\mathcal{R}_\Phi(\theta)] - \mathcal{R}_\Phi(\theta^*) = \underbrace{\left\|\mathbb{E}[\theta] - \theta^*\right\|_\Sigma^2}_{Bias} + \underbrace{\mathbb{E}\Big[\left\|\theta - \mathbb{E}[\theta]\right\|_\Sigma^2\Big]}_{Variance}.$$

Proof. Now, let $\theta \in \mathbb{R}^d$. Then,

$$\begin{aligned}
\mathcal{R}_\Phi(\theta) &= \mathbb{E}\Big[\frac{1}{n}\left\|y - \Phi\theta\right\|_2^2\Big] \\
&= \mathbb{E}\Big[\frac{1}{n}\left\|y - \Phi\theta^* + \Phi(\theta^* - \theta)\right\|_2^2\Big] \\
&= \mathbb{E}\Big[\frac{1}{n}\left\|y - \Phi\theta^*\right\|^2\Big] + \frac{1}{n}\mathbb{E}\Big[\cancel{\left(y - \Phi\theta^*\right)^\top}\Big]\Phi(\theta^* - \theta) + \frac{1}{n}\left\|\Phi(\theta^* - \theta)\right\|_2^2 \\
&= \mathcal{R}_\Phi(\theta^*) + \left\|\theta - \theta^*\right\|_\Sigma^2.
\end{aligned}$$

If θ is random, we have the following bias-variance decomposition

$$\begin{aligned}
\mathbb{E}\big[\mathcal{R}_\Phi(\theta)] - \mathcal{R}_\Phi(\theta^*) &= \mathbb{E}\Big[\left\|\theta - \mathbb{E}[\theta] + \mathbb{E}[\theta] - \theta^*\right\|_\Sigma^2\Big] \\
&= \mathbb{E}\Big[\left\|\theta - \mathbb{E}[\theta]\right\|_\Sigma^2\Big] + \mathbb{E}\Big[(\theta - \mathbb{E}[\theta])^\top\Sigma(\mathbb{E}[\theta] - \theta^*)\Big] + \mathbb{E}\Big[\left\|\mathbb{E}[\theta] - \theta^*\right\|_\Sigma^2\Big] \\
&= \mathbb{E}\Big[\left\|\theta - \mathbb{E}[\theta]\right\|_\Sigma^2\Big] + \cancel{\mathbb{E}\big[(\theta - \mathbb{E}[\theta])\big]^\top}\Sigma(\mathbb{E}[\theta] - \theta^*) + \left\|\mathbb{E}[\theta] - \theta^*\right\|_\Sigma^2 \\
&= \mathbb{E}\Big[\left\|\theta - \mathbb{E}[\theta]\right\|_\Sigma^2\Big] + \mathbb{E}\Big[\left\|\mathbb{E}[\theta] - \theta^*\right\|_\Sigma^2\Big].
\end{aligned}$$

□

It is worth to note that the optimal risk satisfies

$$\mathcal{R}_\Phi(\theta^*) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \big(y_i - \varphi(x_i)^\top\theta^*\big)^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\varepsilon_i^2\big] = \sigma^2\,.$$

2.2.3 Statistical properties of OLS

We now show some guarantees for the OLS estimator.

Proposition 2.3. *Under the linear model (i.e., Assumption 1) with fixed design, the OLS estimator $\widehat{\theta}$ defined in Definition 2.1 satisfies:*

- *it is unbiased* $\mathbb{E}\big[\widehat{\theta}\big] = \theta^*$.
- *its variance is* $\mathrm{Var}(\widehat{\theta}) = \frac{\sigma^2}{n}\Sigma^{-1}$.

We can even show that the OLS satisfies the Gauss-Markov property. It is optimal among unbiased estimators of θ, in the sense that it has a minimal variance-covariance matrix.

Proof. Using $\mathbb{E}[\varepsilon_i] = 0$ and $y = \Phi\theta^* + \varepsilon$, we have

$$\mathbb{E}[\widehat{\theta}] = \mathbb{E}\big[(\Phi^\top\Phi)^{-1}\Phi^\top y\big] = \mathbb{E}\big[(\Phi^\top\Phi)^{-1}\Phi^\top\Phi\theta^* + \cancel{(\Phi^\top\Phi)^{-1}\Phi^\top\varepsilon}\big] = \theta^*\,.$$

Furthermore, using $\mathrm{Var}(y) = \mathrm{Var}(\varepsilon) = \sigma^2 I_n$, we have

$$\mathrm{Var}(\widehat{\theta}) = \mathrm{Var}\big((\Phi^\top\Phi)^{-1}\Phi^\top y\big) = (\Phi^\top\Phi)^{-1}\Phi^\top\mathrm{Var}(y)\Phi(\Phi^\top\Phi)^{-1} = \sigma^2(\Phi^\top\Phi)^{-1} = \frac{\sigma^2}{n}\Sigma^{-1}\,.$$

□

Corollary 2.4 (Excess risk of OLS)**.** *Under the linear model with fixed design, the excess risk of the OLS satisfy*

$$\mathbb{E}\big[\mathcal{R}_\Phi(\widehat{\theta})\big] - \mathcal{R}_\Phi(\theta^*) = \frac{\sigma^2 d}{n}\,.$$

Proof. Using the bias-variance decomposition and the fact that θ^* is unbiased (i.e., $\mathbb{E}[\widehat{\theta}] = \theta^*$), we have

$$\begin{aligned}
\mathbb{E}\big[\mathcal{R}_\Phi(\widehat{\theta})] - \mathcal{R}_\Phi(\theta^*) &= \cancel{\big\|\mathbb{E}[\widehat{\theta}] - \theta^*\big\|_\Sigma^2} + \mathbb{E}\Big[\big\|\widehat{\theta} - \mathbb{E}[\widehat{\theta}]\big\|_\Sigma^2\Big] = \mathbb{E}\Big[\big\|\widehat{\theta} - \theta^*\big\|_\Sigma^2\Big] \\
&= \mathbb{E}\left[(\widehat{\theta} - \theta^*)^\top\Sigma(\widehat{\theta} - \theta^*)\right] \\
&= \frac{1}{n}\mathbb{E}\big[(\widehat{\theta} - \theta^*)^\top\Phi^\top\Phi(\widehat{\theta} - \theta^*)\big] \\
&= \frac{1}{n}\mathbb{E}\big[\,\mathrm{Tr}\left((\widehat{\theta} - \theta^*)^\top\Phi^\top\Phi(\widehat{\theta} - \theta^*)\right)\big] \\
&= \frac{1}{n}\mathbb{E}\big[\,\mathrm{Tr}\left(\Phi(\widehat{\theta} - \theta^*)(\widehat{\theta} - \theta^*)^\top\Phi^\top\right)\big] \qquad \leftarrow \text{because } \mathrm{Tr}(AB) = \mathrm{Tr}(BA) \\
&= \frac{1}{n}\,\mathrm{Tr}\left(\Phi\mathbb{E}\big[(\widehat{\theta} - \theta^*)(\widehat{\theta} - \theta^*)^\top\big]\Phi^\top\right) \qquad \leftarrow \text{because } \mathbb{E} \text{ and Tr are linear operators} \\
&= \frac{1}{n}\,\mathrm{Tr}\left(\Phi\mathrm{Var}(\widehat{\theta})\Phi^\top\right) \\
&= \frac{\sigma^2}{n}\,\mathrm{Tr}\left(\Phi(\Phi^\top\Phi)^{-1}\Phi^\top\right) = \frac{\sigma^2 d}{n}\,,
\end{aligned}$$

where the last equality is because $\Phi(\Phi^\top\Phi)^{-1}\Phi^\top = P_\Phi$ is the orthogonal projection matrix onto the d-dimensional subspace $\mathrm{Im}(\Phi)$. □

Exercise. show that the expected risk $\mathbb{E}\big[\widehat{\mathcal{R}}_\Phi(\widehat{\theta})\big] = \frac{n-d}{n}\sigma^2$. In particular, an unbiased estimator of the noise σ^2 is

$$\widehat{\sigma}^2 = \frac{\|y - \Phi\widehat{\theta}\|^2}{n-d}\,.$$

Gaussian noise model A very considered special case is Gaussian noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. This choice comes not only from the fact that it allows to compute many additional statistical properties on $\widehat{\theta}$ and to perform tests (confidence intervals, significance of variables, ...). In practice, it is also motivated by the central limit theorem and the fact that noise is often an addition of many phenomena not explained by the linear combination of the explanatory variables.

Proposition 2.5. *In the linear model with Gaussian noise, the maximum likelihood estimators of θ and σ satisfy respectively:*

$$\widehat{\theta}_{MV} = \big(\Phi^\top\Phi\big)^{-1}\Phi y \qquad \textit{and} \qquad \widehat{\sigma}^2_{MV} = \frac{\|y - \Phi\widehat{\theta}\|^2}{n}\,.$$

We will prove more formally this proposition in the maximum likelihood section (Section 4. We therefore find the least-squares estimator obtained by minimizing the empirical risk. The variance estimator is biased.

2.3 Ridge regression

If Φ is not injective (i.e., $\mathrm{rg}(\Phi) \neq d$), the matrix $\Sigma := \Phi^\top\Phi$ is no longer invertible and the OLS optimization problem admits several solutions. The problem is said to be poorly posed or unidentifiable.

The Proposition 2.3 reminds us that the variance of $\widehat{\theta}$ depends on the conditioning of the matrix $\Sigma^{-1} = (\Phi^\top\Phi)^{-1}$. The more the columns of the latter are likely to be dependent, the less stable $\widehat{\theta}$ will be. Several solutions allow to deal with the case where $\mathrm{rg}(\Phi) < d$:

- *explicit complexity control* by reducing the solution space $\mathrm{Im}(\Phi)$. This can be done by removing columns from the Φ matrix until it becomes injective (for example, by reducing the degree of polynomials). One can also set identifiability constraints of the form $\theta \in V$ a vector subspace of $\mathbb{R}^d$ such that any element $y \in \mathrm{Im}(\Phi)$ has a unique antecedent $\theta \in V$ with $y = \Phi\theta$. For example, we could choose $V = \mathrm{Ker}(\Phi)^\perp$.
- *implicit complexity control* by regularizing the empirical risk minimization problem. The most common is to regularize by adding $\|\theta\|_2^2$ (Ridge regression, which we see below) or $\|\theta\|_1$ (Lasso regression).

Definition 2.2. *For a regularization parameter λ, the Ridge regression estimator is defined as*

$$\widehat{\theta}_\lambda \in \operatorname*{arg\,min}_{\theta\mathbb{R}^d} \left\{ \frac{1}{n}\|y - \Phi\theta\|_2^2 + \lambda\|\theta\|_2^2 \right\}.$$

The regularization parameter $\lambda > 0$ regulates the trade-off between the variance of $\widehat{\theta}$ and its bias.

Proposition 2.6. *The Ridge regression estimator is unique (even if Φ is not injective) and satisfies*

$$\widehat{\theta}_\lambda = \left(\Phi^\top \Phi + n\lambda I_n\right)^{-1} \Phi^\top y \,.$$

The proof is similar to the one of OLS and left as exercise. We can see that there is no longer the problem of inverting $\Phi^\top \Phi$ since the Ridge regression amounts to replacing $\left(\Phi^\top \Phi\right)^{-1}$ by $\left(\Phi^\top \Phi + n\lambda I_n\right)^{-1}$ in the OLS solution.

Proposition 2.7 (Risk of Ridge regression)**.** *Under the linear model (Assumption 1), the Ridge regression estimator satisfies*

$$\mathbb{E}\big[\mathcal{R}_\Phi(\widehat{\theta}_\lambda)\big] - \mathcal{R}_\Phi(\theta^*) = \sum_{j=1}^{d} (\theta_j^*)^2 \frac{\lambda_j}{(1+\lambda_j/\lambda)^2} + \frac{\sigma^2}{n} \sum_{j=1}^{d} \frac{\lambda_j^2}{(\lambda_j+\lambda)^2} \,,$$

where λ_j is the j-th eigenvalue of $\Sigma = \frac{1}{n}\Phi^\top \Phi$. In particular, the choice $\lambda^* = \frac{\sigma\sqrt{\operatorname{Tr}(\Sigma)}}{\|\theta^*\|_2 \sqrt{n}}$ yields

$$\mathbb{E}\big[\mathcal{R}_\Phi(\widehat{\theta}_{\lambda^*})\big] - \mathcal{R}_\Phi(\theta^*) \leqslant \frac{\sigma\sqrt{2\operatorname{Tr}(\Sigma)\|\theta^*\|_2}}{\sqrt{n}} \,.$$

The proof, which follows from the bias-variance decomposition (Proposition 2.2) is left as exercise.

Note that as $\lambda \to 0$, its risk converges to the one of OLS. The first term corresponds to the bias of the Ridge estimator. Thus, on the downside the Ridge estimator is biased in contrast to the OLS. But on the positive side, its variance does not involve the inverse of Σ but of $\Sigma + \lambda I_d$ which is better conditioned. It has therefore a lower variance. The parameter λ controls this trade-off.

We can compare the excess risk bound obtained by $\widehat{\theta}_{\lambda^*}$ with the one of OLS which was $\sigma^2 d/n$:

- First, the one of OLS decreases in $O(1/n)$ while this one converges slower in $O(1/\sqrt{n})$ which could seem worse. Yet Ridge has a milder dependence on the noise σ instead of σ^2.
- Furthermore, since $\operatorname{Tr}(\Sigma) \leqslant \max_{1\leqslant i\leqslant n} \|\varphi(x_i)\|^2$, if the input norms are bounded by R, the excess risk of Ridge does not depend on the dimension d, which can even be infinite. It is called a *dimension free* bound.

The calibration of the regularization parameter is essential in practice. It can for example be done analytically as in the proposition (but some quantities are unknown σ^2, $\|\theta^*\|$,...). In practice one resorts to train/validation set or *cross-validation (generalized)*.

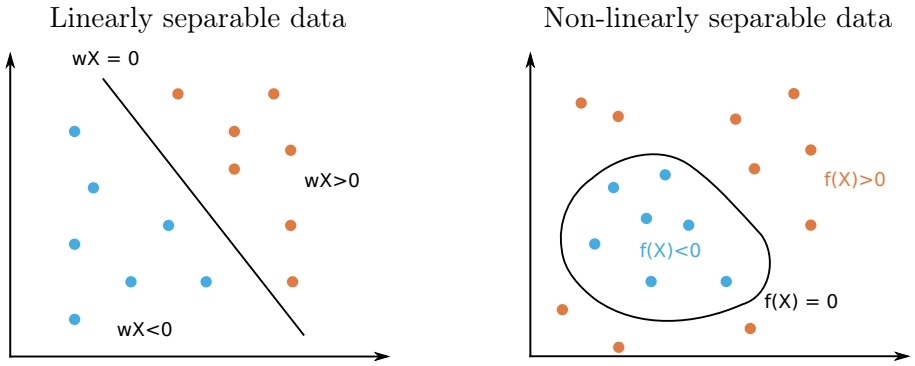
3 Logistic regression

We will consider the binary classification problem in which one wants to predict outputs in $\{0,1\}$ from inputs in $\mathbb{R}^d$. We consider a training set $D_n := \{(x_i, y_i)\}_{1 \leqslant i \leqslant n}$. The data points (x_i, y_i) are i.i.d. random variables and follow a distribution $\mathcal{P}$ in $\mathcal{X} \times \mathcal{Y}$. Here, $\mathcal{Y} = \{0,1\}$ but it is also common to consider $\{-1,1\}$.

Goal We would like to use a similar algorithm to linear regression. However, since the outputs y_i are binary and belong to $\{0,1\}$ we cannot predict them by linear transformation of the inputs x_i (which belong to $\mathbb{R}^d$). We will thus classify the data thanks to classification rules $f : \mathbb{R}^d \mapsto \mathbb{R}$ such that:

$$f(x_i) \begin{cases} \geqslant 0 & \Rightarrow \quad y_i = +1 \\ < 0 & \Rightarrow \quad y_i = 0 \end{cases} ,$$

to separate the data into two groups. In particular, we will consider linear functions f of the form $f_\theta : x \mapsto x^\top \theta$. This assumes that the data are well-explained by a linear separation (see figure below).



Of course, if the data does not seem to be linearly separable, we can use similar tricks that we mentioned for linear regression (polynomial regression, kernel regression, splines,...). We search a feature map $x \mapsto \varphi(x)$ into a higher dimensional space in which the data are linearly separable. This will be the topic of the class on Kernel methods.

Loss function To minimize the empirical risk, it remains to choose a loss function to assess the performance of a prediction. A natural loss is the *binary loss*: 1 if there is a mistake ($f(x_i) \neq y_i$) and 0 otherwise. The empirical risk is then:

$$\widehat{\mathcal{R}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{y_i \neq \mathbb{1}_{x_i^\top \theta \geqslant 0}} .$$

This loss function is however not convex neither in θ. The minimization problem $\min_\theta \widehat{\mathcal{R}}(\theta)$ is extremely hard to solve. The idea of logistic regression consists in replacing the binary loss with another similar loss function which is convex in θ. This is the case of the *Hinge loss* and of the logistic loss $\ell : \{0,1\} \times \mathbb{R} \to \mathbb{R}_+$. The latter assigns to a linear prediction $z = x^\top \theta$ and an observation $y \in \{0,1\}$ the loss

$$\ell(y, z) := y \log\left(1 + e^{-z}\right) + (1 - y) \log\left(1 + e^{z}\right) . \tag{5}$$

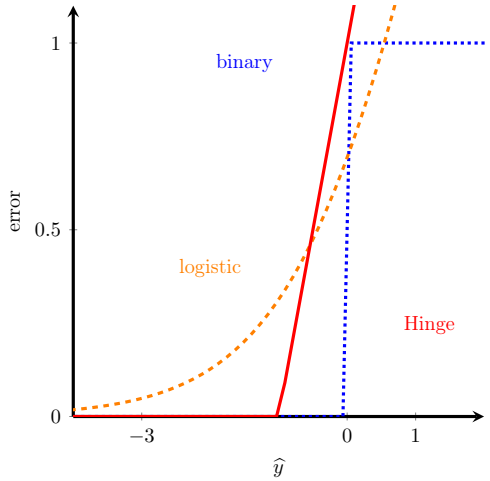


Figure 3: Binary, logistic and Hinge loss incured for a prediction $z := x^\top \theta$ when the true observation is $y = 0$.

The binary loss, Hinge loss and logistic loss are plotted in Figure 3.

⚠ Note that if the output space is $\mathcal{Y} = \{-1, 1\}$, the logistic loss is defined differently: $\ell(y, z) := \log(1 + e^{-zy})$.

Definition 3.1 (Logistic regression estimator)**.** *The logistic regression estimator is the solution of the following minimization problem:*

$$\widehat{\theta}_{(logit)} = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, x_i^\top \theta)\,,$$

where ℓ is the logistic loss defined in Equation (5)*.*

The advantage of the logistic loss with respect to the Hinge loss is that it has a probabilistic interpretation by modeling $\mathbb{P}(y = 1|x)$, where (x, y) is a couple of random variables following the law of (x_i, y_i). We will see more on this in the lecture on Maximum Likelihood.

Computation of $\widehat{\theta}_{(\text{logit})}$ Similarly to OLS, we may try to analytically solve the minimization problem by canceling the gradient of the empirical risk. Since

$$\frac{\partial \ell(y, z)}{\partial z} = \sigma(z) - y, \qquad \text{where} \quad \sigma : z \mapsto \frac{1}{1 + e^{-z}}$$

is the logistic function, we have:

$$\nabla \widehat{\mathcal{R}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} x_i \big(\sigma(x_i^\top \theta) - y_i\big) = \frac{1}{n} x \big(y - \sigma(x\theta)\big)$$

where $x := (x_1, \dots, x_n)^\top$, $y := (y_1, \dots, y_n)$, and $\sigma(x\theta)_i := \sigma(x_i^\top \theta)$ for $1 \leqslant i \leqslant n$. Bad news: the equation $\nabla \widehat{\mathcal{R}}(\theta) = 0$ has no closed-form solution. It needs to be solved through iterative algorithm (gradient descent, Newton's method,...). Fortunately, this is possible because the logistic loss is convex in its first argument. Indeed,

$$\frac{\partial^2 \ell(y, z)}{\partial z} = \sigma(z)\sigma(-z) > 0\,.$$

The loss is strictly convex, the solution is thus unique.

Regularization Similarly to linear regression, logistic regression may over-fit the data (especially when $p > n$). One needs then to add a regularization such as $\lambda\|\theta\|_2^2$ to the logistic loss.

4 Probabilistic models: maximum likelihood estimation

In probabilistic modeling, we are given a set of observations $D_n = (y_1, \dots, y_n)$ in $\mathcal{Y}$ that we assume to be generated from some unknown i.i.d. distribution. The objective is to find a probabilistic model that explains well the data. For instance by estimating the density of the underlying distribution. If possible, we would like the model to predict well new data and to be able to incorporate prior knowledge and assumptions.

Let μ denote some reference measure on the output set $\mathcal{Y}$. Typically, μ is the counting measure if $\mathcal{Y} \subset \mathbb{N}$ or the Lebesgue measure if $\mathcal{Y} \subset \mathbb{R}^p$.

Definition 4.1 (Parametric model)**.** *Let $d \geqslant 1$ and $\Theta \subseteq \mathbb{R}^d$ be a set of parameters. A parametric model $\mathcal{P}$ is a set of probability distributions taking value in $\mathcal{Y}$ with a density with respect to μ and indexed by Θ: $\mathcal{P} := \{p_\theta d\mu | \theta \in \Theta\}$.*

Example 4.1. *Here are a few examples of statistical parametric models based on well known family distributions:*

- *Bernoulli model: $\mathcal{Y} = \{0, 1\}$, $\Theta = [0, 1]$, and $p_\theta(k) = \theta^k(1-\theta)^{1-k}$ for $k \in \mathcal{Y}$.*
- *Binomial model: $\mathcal{Y} = \mathbb{N}$, $\Theta = [0, 1]$ and $p_\theta(k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$;*
- *Gaussian model: $\mathcal{Y} = \mathbb{R}$, $\Theta = \{(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+\}$ and $p_{(\mu,\sigma)}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$*
- *Multidimensional Gaussian model: $\mathcal{Y} = \mathbb{R}^d$, $\Theta = \{(\mu, \Sigma) \in \mathbb{R}^d \times \mathcal{M}_d(\mathbb{R})\}$ and*

$$p_{(\mu,\Sigma)}(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}.$$

- *Exponential model on $\mathcal{Y} = \mathbb{R}_+$,...*

Now, we assume that we are given some model $\mathcal{P}$ indexed by $\theta \in \Theta$ and we assume that the data D_n is generated independently from $p_{\theta^*} \in \mathcal{P}$ for some unknown parameter θ^*. We would like to recover the best parameter θ^* from the data. Note that in practice the data might come from a distribution which is not in $\mathcal{P}$: we call this misspecification but we will not enter into this details in this class.

4.1 Maximum likelihood estimation

The idea behind maximum likelihood estimation is to choose the most probable parameter $\theta \in \Theta$ for the observed data. Assume that $\mathcal{Y}$ is discrete and that $y \sim p_{\theta^*} d\mu$ for some $\theta^* \in \Theta$. Then, given any observation y_i, the probability that y takes the value y_i equals $p_{\theta^*}(y_i)$. Similarly, the probability of observing $(y_1, \dots, y_n) \in \mathcal{Y}^n$ if all the samples were sampled independently from p_θ is $\prod_{i=1}^m p_\theta(y_i)$. Hence, the high level idea of maximum likelihood estimation will be to maximize this probability over $\theta \in \Theta$. This is formalized by the definition of the likelihood which also holds for non-discret set $\mathcal{Y}$.

Definition 4.2 (Likelihood)**.** *Let $\mathcal{P} = \{p_\theta, \theta \in \Theta\}$ a parametric model and $y \in \mathcal{Y}$. The likelihood is the function $\theta \mapsto p_\theta(x)$. The likelihood $L(.|D_n)$ of a data set $D_n = (y_1, \dots, y_n)$ is the function*

$$L(\cdot|D_n) : \theta \mapsto \prod_{i=1}^{m} p_\theta(y_i)\,.$$

The maximum likelihood estimator (MLE) is then the parameter which maximizes the likelihood, i.e.,

$$\widehat{\theta}_n \in \arg\max_{\theta\in\Theta} \left\{ \prod_{i=1}^{n} p_\theta(y_i) \right\}.$$

This principle was proposed by Ronal Fisher in 1922 and was validated since with good theoretical properties. It is worth pointing out that since log is an increasing function, the maximum likelihood estimator can also be obtained by maximizing the log-likelihood:

$$\widehat{\theta}_n \in \arg\max_{\theta\in\Theta} \left\{ \sum_{i=1}^{n} \log(p_\theta(y_i)) \right\}. \tag{MLE}$$

This turns out to be much more convenient in practice because it is easier to maximize a sum than a product. Convince yourself by computing the gradients!

Examples

- Bernoulli model: $\mathcal{Y} = \{0, 1\}$, $\Theta = [0, 1]$, $p_\theta(y) = \theta^y(1-\theta)^{(1-y)}$. We assume that D_n was generated from a Bernoulli distribution of parameter θ^*, then the maximum likelihood estimator is:

  $$\widehat{\theta}_n = \arg\min_{0\leqslant\theta\leqslant 1} \frac{1}{n}\sum_{i=1}^{n} \big(y_i \log\theta + (1-y_i)\log(1-\theta)\big)\,.$$

  Denoting $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$ the empirical average and solving $d\log L(\widehat{\theta}_n|D_n)/d\theta = 0$ yields

  $$\frac{\bar{y}_n}{\widehat{\theta}_n} - \frac{1-\bar{y}_n}{1-\widehat{\theta}_n} = 0 \quad \Rightarrow (1-\bar{y}_n)\widehat{\theta}_n = (1-\widehat{\theta}_n)\bar{y}_n \quad \Rightarrow \quad \widehat{\theta}_n = \bar{y}_n\,.$$

  Therefore the maximum likelihood estimator is in this case the empirical mean.

- As an exercise, compute the maximum likelihood estimator for the models seen in Example 4.1.

Link with empirical risk minimization In density estimation, the goal is to find the density of the distribution which generated the data. Assuming that the density belongs to the model $\mathcal{P}$, the possible densities are p_θ, for $\theta \in \Theta$. A standard loss function in this setting is the negative log-likelihood: $\ell : (\theta, y) \in \Theta \times \mathcal{Y} \mapsto -\log\big(p_\theta(y)\big)$. The risk (or generalization error) is then:

$$\mathcal{R}(\theta) = -\mathbb{E}_y\big[\log(p_\theta(y))\big]\,.$$

In particular, if $y \sim p_{\theta^*}d\mu$ for some $\theta^* \in \Theta$, θ^* minimize the risk and the objective is to recover θ^*. The empirical risk is then by definition

$$\widehat{\mathcal{R}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log\big(p_\theta(y_i)\big)\,.$$

Therefore, the empirical risk minimizer matches the estimator obtained from maximum likelihood in Equation (MLE).

Link with Kullback-Leibler divergence The Kulback-Leibler divergence is a measure of dissimilarity two between probability distributions. It was introduced by Kullback and Leibler in 1951.

Definition 4.3 (Kullback-Leibler divergence)**.** *Let $pd\mu$ and $qd\mu$ be two probability distributions. The Kullback-Leibler divergence from p to q is defined as*

$$KL(p||q) := \mathbb{E}_{y\sim pd\mu}\left[\log\frac{p(y)}{q(y)}\right] = \int_{\mathcal{Y}} p(y)\log\frac{p(y)}{q(y)}d\mu(y)\,.$$

The KL divergence has various interpretations. As we will see now, it can be interpreted as the excess risk of the measure $p_\theta d\mu$ when the data follows distribution $p_{\theta^*}d\mu$ when the loss function is the negative log-likelihood. Assume that the data D_n were generated from p_{θ^*}. Then, the excess risk can be written

$$\begin{aligned}\mathcal{R}(\theta) - \mathcal{R}(\theta^*) &= -\mathbb{E}_{y\sim\theta^*}\big[\log(p_\theta(y))\big] + \mathbb{E}_{y\sim\theta^*}\big[\log(p_{\theta^*}(y))\big] \\ &= \mathbb{E}_{y\sim\theta^*}\Big[\log\Big(\frac{p_{\theta^*}(y)}{p_\theta(y)}\Big)\Big] =: KL\big(p_{\theta^*}||p_\theta\big)\end{aligned}$$

where $\mathbb{E}_{\theta^*}[f(y)]$ denotes $\mathbb{E}_{y\sim p_{\theta^*}d\mu}[f(y)]$ the expectation of $f(y)$ when y follows $p_{\theta^*}d\mu$.

Another interpretation comes from information theory. It can be seen as the difference of bits needed to encode D_n under a code optimized for $p_\theta d\mu$ compared to a code optimized for $p_{\theta^*}d\mu$.

Properties and remarks about the KL-divergence:

- $KL(P||Q) \geqslant 0$ by Jensen's inequality
- $KL(p||p) = 0$. Therefore, we see that p_{θ^*} minimize the the risk and thus maximize the likelihood.
- If the distributions are discrete and μ is the counting measure, we have in particular $KL(p||q) := \sum_{i\in\mathcal{Y}} p(i)\log\big(\frac{p(i)}{q(i)}\big)$.
- The Kullback–Leibler divergence is defined only if for all $A \subset \mathcal{Y}$, $q(A) = 0$ implies $p(A) = 0$, i.e., if q is absolutely continuous with respect to p.
- Though KL is often seen as a distance, it does not fill the requirements: it is not symmetric and it does not satisfy the triangular inequality.
- With an abuse of notation, we can rewrite the empirical risk minimization for log loss with the KL:

  $$\widehat{\theta}_n \in \operatorname*{arg\,min}_{\theta\in\Theta} KL(\widehat{p}_n||p_\theta)$$

  where $\widehat{p}_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ is the empirical measure (which does not have any density with respect to the Lebesgue measure though).

Conditional modeling

Until now, we considered the problem of density estimation when the data set has only outputs $y_i \in \mathcal{Y}$. However, the principle of maximum likelihood can be extended to couples of input outputs $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ in $\mathcal{X}\times\mathcal{Y}$. We can then distinguish two different modeling:

- generative modeling: we aim at estimating the density of couples of input outputs (x, y) among a family of densities $(x, y) \in \mathcal{X}\times\mathcal{Y} \mapsto p_\theta(x, y)$ on $\mathcal{X}\times\mathcal{Y}$. Then the risk and the empirical risks are:

  $$\mathcal{R}(\theta) = -\mathbb{E}\big[\log(p_\theta(x, y)\big] \qquad \widehat{\mathcal{R}}(\theta) = -\frac{1}{n}\sum_{i=1}^n \log\big(p_\theta(x_i, y_i)\big)\,.$$

This can be useful to generate some new samples (see what is obtained with GANs).

– conditional modeling: we aim at estimating the density of an output y given an input x. The family of densities are now densities $y \in \mathcal{Y} \mapsto p_\theta(.|x)$ on $\mathcal{Y}$ only but that depend on the inputs. The risks are then

$$\mathcal{R}(\theta) = -\mathbb{E}\big[\log(p_\theta(y|x)\big] \qquad \widehat{\mathcal{R}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log\big(p_\theta(y_i|x_i)\big)\,.$$

This is useful if one want to predict the distribution or the value of a new output y given x.

4.2 Probabilistic interpretation of least-squares and logistic regression

4.2.1 Probabilistic insight of linear regression

We consider a data set $D_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ of samples in $\mathcal{X} \times \mathcal{Y}$. We assume that the outputs y_i were independently generated from a Gaussian distribution of mean $w^\top x_i$ and variance σ^2. In other words, we model an output y given an input x as

$$y = {w_*}^\top x + \varepsilon, \qquad \text{where} \quad \varepsilon \sim \mathcal{N}(0, \sigma_*^2)\,.$$

for some unknown $\theta^* = (w_*, \sigma_*^2) \in \mathbb{R}^d \times \mathbb{R}_+$. Our family of possible conditional densities is indexed by parameters $\theta = (w, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+$

$$p_\theta(y|x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - w^\top x)^2}{2\sigma^2}}\,.$$

The empirical risk (or conditional log-likelihood) is then

$$\widehat{\mathcal{R}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log\big(p_\theta(y_i|x_i)\big) = \frac{1}{2n\sigma^2}\sum_{i=1}^{n}\big(y_i - w^\top x_i\big)^2 + \frac{1}{2}\log(2\pi\sigma^2)\,.$$

Therefore, the maximum likelihood estimator $\widehat{w}_n$ of w in a Gaussian model is the estimator obtained by least-squares linear regression. As an exercise, you may show that the maximum likelihood estimator for σ is

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \widehat{w}_n^\top x_i\big)^2\,.$$

4.2.2 Probabilistic insight of logistic regression

The advantage of the logistic loss with respect to the Hinge loss is that it has a probabilistic interpretation by modeling $\mathbb{P}(y = 1|x)$, where (x, y) is a couple of random variables following the law of (x_i, y_i). By Bayes rules, we have

$$\mathbb{P}(y = 1|x) = \frac{\mathbb{P}(x|y=1)\mathbb{P}(y=1)}{\mathbb{P}(x|y=1)\mathbb{P}(y=1) + \mathbb{P}(x|y=0)\mathbb{P}(y=0)} = \frac{1}{1 + \frac{\mathbb{P}(y=0)\mathbb{P}(x|y=0)}{\mathbb{P}(y=1)\mathbb{P}(x|y=1)}}\,.$$

Denote by

$$f(x) := \log\left(\frac{\mathbb{P}(y=1,x)}{\mathbb{P}(y=0,x)}\right) = \log\left(\frac{\mathbb{P}(y=1)\mathbb{P}(x|y=1)}{\mathbb{P}(y=0)\mathbb{P}(x|y=0)}\right) = \log\left(\frac{\mathbb{P}(y=1)}{\mathbb{P}(y=0)}\right) + \log\left(\frac{\mathbb{P}(x|y=1)}{\mathbb{P}(x|y=0)}\right)$$

the logarithmic ratio of the probability of observing x if y equals 0 with the one of observing x if $y = 1$. Then,

$$\mathbb{P}(y = 1|x) = \frac{1}{1 + e^{-f(x)}} =: \sigma(f(x)) \qquad \text{with} \qquad \sigma(z) = \frac{1}{1 + e^{-z}}\,.$$

The function σ is called the logistic function and satisfies $\sigma(-z) = 1-\sigma(z)$ et $\frac{\mathrm{d}\sigma(z)}{\mathrm{d}z} = \sigma(z)\sigma(-z)$. Its interest is that it allows to transform a function f with value in $\mathbb{R}$ into a probability between 0 and 1.

The logistic regression model is in fact the same as assuming that f is linear of the form $f : x \mapsto x^\top\theta$. Recall that as with linear regression, x could be replaced with a feature vector $\varphi(x)$.

Proposition 4.1. *Assuming, that $(x_i, y_i)_{1\leqslant i\leqslant n}$ is an n-sample such that $\mathbb{P}(y_i = 1|x_i) = \sigma(x_i^\top\theta)$, then the estimator of the maximum of θ is $\widehat{\theta}_{(logit)}$.*

Proof. The log-likelihood can be written

$$\begin{aligned}
\log L(\theta|D_n) &= \sum_{i=1}^{n} \log\left(\mathbb{P}_\theta(y_i = 1|x_i)^{y_i}(1 - \mathbb{P}_\theta(y_i = 0|x_i)^{1-y_i}\right) \\
&= \sum_{i=1}^{n} \log\left(\sigma(\theta^\top x_i)^{y_i}\sigma(-\theta^\top x_i)^{1-y_i}\right) \\
&= -\sum_{i=1}^{n} \ell(y_i, \theta^\top x_i)
\end{aligned}$$

where ℓ is the logistic loss. The logistic regression estimator is therefore the maximum likelihood estimator. □