Exercise 1:

Imagine you work in industry and have a data set $\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \dots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)\right)$. You train a model $\hat{f}(\mathbf{x})$ on this data set and now you want to bring it into production. Your customer wants to know which performance they can expect from your model, when using it from now on. As an answer, you want to provide an estimate for the generalization error of this model, i.e., $\mathrm{GE}\left(\hat{f}, L\right)$.

Since you have no data left to test your model on, you try to estimate, as a proxy for $\mathrm{GE}\left(\hat{f}, L\right)$, how good a model could be that would have been learned on n data points, i.e., $\mathrm{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}}, \rho)$ with $n_{\text{train}} = n$. But, since also in this case you would have no data points left to test your model on, you try the next best thing:

For a learner $\mathcal{I}$, n_{train} training observations and a performance measure ρ, the **generalization error** can be formally expressed as:

$$\mathrm{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}}, \rho) = \lim_{n_{\text{test}} \to \infty} \mathbb{E}_{\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}} \sim \mathbb{P}_{xy}} \left[\rho\left(\mathbf{y}, \boldsymbol{F}_{\mathcal{D}_{\text{test}}, \mathcal{I}(\mathcal{D}_{\text{train}}, \boldsymbol{\lambda})}\right)\right], \tag{1}$$

where for now we assume that $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ can be independently sampled from $\mathbb{P}_{xy}$.

1) **What is the generalization error? Describe the formula above in your own words.**

In practice, the data generating process $\mathbb{P}_{xy}$ is usually unknown and we cannot directly sample observations from it (instead, we typically use the available data $\mathcal{D}$ as a proxy). However, let's for now assume we can sample as many times as we like from $\mathbb{P}_{xy}$.

2) **Explain how you could empirically estimate the generalization error $\mathrm{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}}, \rho)$ with $n_{\text{train}} = 100$ of a learner $\mathcal{I}$ with configuration $\boldsymbol{\lambda}$ trained on $n_{\text{train}} = 100$ observations and evaluated on performance measure ρ, given that you can sample from $\mathbb{P}_{xy}$ as often as you like.**

In addition to an unknown data-generating process $\mathbb{P}_{xy}$, supervised learning is often restricted to a data set $\mathcal{D}$ of fixed size n. Therefore, the true generalization error $\mathrm{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}}, \rho)$ (with $n_{\text{train}} = n$ referring to all available data) remains unknown. In this case, hold-out splitting is a simple procedure that can be used to estimate the generalization error:

$$\widehat{\mathrm{GE}}(\mathcal{I}, \boldsymbol{\lambda}, (J_{\text{train}}, J_{\text{test}}), \rho) = \rho\left(\mathbf{y}_{J_{\text{test}}}, \boldsymbol{F}_{J_{\text{test}}, \mathcal{I}(\mathcal{D}_{\text{train}}, \boldsymbol{\lambda})}\right), \tag{2}$$

where $(J_{\text{train}}, J_{\text{test}})$ with $J_{\text{train}} \in \{1, \dots, n\}^{n_{\text{train}}}$ and $J_{\text{test}} \in \{1, \dots, n\}^{n_{\text{test}}}$ are index vectors that specify the subset of $\mathcal{D}$ the learner $\mathcal{I}$ is trained on, with $|J_{\text{train}}| < n$ (we train our model on less data as we have at hand) and $|J_{\text{train}}| + |J_{\text{test}}| = n$. Note the change in notation compared to the theoretical generalization error $\mathrm{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}}, \rho) = \mathrm{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}}, \rho)$ above:

- $\mathrm{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}}, \rho)$ is the *theoretical* generalization error (the estimand). It is a single number that depends only on the learner, its hyperparameters, the training size n_{train} (typically all available data as we are interested in the generalization error of a model trained on all available data) and the performance measure ρ, and averages over all possible training and test samples from $\mathbb{P}_{xy}$.
- $\widehat{\mathrm{GE}}(\mathcal{I}, \boldsymbol{\lambda}, (J_{\text{train}}, J_{\text{test}}), \rho)$ is an *empirical estimator* based on a concrete split $(J_{\text{train}}, J_{\text{test}})$ of the fixed data set $\mathcal{D}$. Here the effective training size is $|J_{\text{train}}| < n_{\text{train}}$ (as we need also some data points for testing). For a fixed $|J_{\text{train}}|$, different choices of $(J_{\text{train}}, J_{\text{test}})$ generally lead to different values of $\widehat{\mathrm{GE}}$.

3) **Explain how the choice of training size $|J_{\text{train}}|$ may influence the bias of $\widehat{\mathrm{GE}}(\mathcal{I}, \boldsymbol{\lambda}, (J_{\text{train}}, J_{\text{test}}), \rho)$ wrt $\mathrm{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}}, \rho)$.**

4) **Explain how the choice of training size $|J_{\text{train}}|$ may influence the variance of $\widehat{\mathrm{GE}}(\mathcal{I}, \boldsymbol{\lambda}, (J_{\text{train}}, J_{\text{test}}), \rho)$.**