Solution 1:

1) **Interpretation of the generalization error** $\text{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}}, \rho)$

- $\text{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}}, \rho)$ is the *expected future performance* of learner $\mathcal{I}$ with configuration $\boldsymbol{\lambda}$ trained on n_{train} observations and evaluated with performance measure ρ.
- In $\text{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}}, \rho) = \lim_{n_{\text{test}} \to \infty} \mathbb{E}_{\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}} \sim \mathbb{P}_{xy}} \left[\rho\left(\mathbf{y}, \boldsymbol{F}_{\mathcal{D}_{\text{test}}, \mathcal{I}(\mathcal{D}_{\text{train}}, \boldsymbol{\lambda})}\right) \right]$, the randomness comes from repeatedly sampling train and test sets $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}$ from $\mathbb{P}_{xy}$.
- For each draw:
  - We train a model $\mathcal{I}(\mathcal{D}_{\text{train}}, \boldsymbol{\lambda})$ on $\mathcal{D}_{\text{train}}$.
  - We evaluate its performance $\rho(\mathbf{y}, \boldsymbol{F}_{\mathcal{D}_{\text{test}}, \mathcal{I}(\mathcal{D}_{\text{train}}, \boldsymbol{\lambda})})$ on an independent test set $\mathcal{D}_{\text{test}}$.
- Taking the expectation over all such draws and the limit $n_{\text{test}} \to \infty$ removes randomness from the particular test set and yields the true expected performance of the learner for training size n_{train}.

2) **Empirical estimation of** $\text{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}} = 100, \rho)$ **when we can sample from** $\mathbb{P}_{xy}$

- Repeat the following for $k = 1, \ldots, K$:
  - Draw a training set $\mathcal{D}_{\text{train},k}$ of size $n_{\text{train}} = 100$ from $\mathbb{P}_{xy}$.
  - Draw an independent test set $\mathcal{D}_{\text{test},k}$ of size n_{test} (large) from $\mathbb{P}_{xy}$.
  - Train the learner: $\hat{f}^{[k]} = \mathcal{I}(\mathcal{D}_{\text{train},k}, \boldsymbol{\lambda})$.
  - Compute the performance on the test set: $\rho\left(\mathbf{y}_{J_{\text{test},k}}, \boldsymbol{F}_{J_{\text{test},k}, \mathcal{I}(\mathcal{D}_{\text{train},k}, \boldsymbol{\lambda})}\right)$.
- Average K values: $\widehat{\text{GE}}_K(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}} = 100, \rho) = \frac{1}{K} \sum_{k=1}^{K} \rho\left(\mathbf{y}_{J_{\text{test},k}}, \boldsymbol{F}_{J_{\text{test},k}, \mathcal{I}(\mathcal{D}_{\text{train},k}, \boldsymbol{\lambda})}\right)$.
- For $K, n_{\text{test}} \to \infty$, the estimator converges to the theoretical quantity $\text{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}} = 100, \rho)$.

3) **Effect of training size** $|J_{\text{train}}|$ **on the bias of the hold-out estimator**

- In practice we only have a fixed data set $\mathcal{D}$ of size n. The target we care about is the generalization error of a learner trained on *all* available data, i.e.:

$$\text{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}}, \rho) = \text{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}} = n, \rho),$$

- Hold-out splitting uses only a subset $\mathcal{D}_{\text{train}}$ of size $|J_{\text{train}}| < n$ for training and a disjoint subset $\mathcal{D}_{\text{test}}$ for testing. The empirical estimator is $\widehat{\text{GE}}(\mathcal{I}, \boldsymbol{\lambda}, (J_{\text{train}}, J_{\text{test}}), \rho) = \rho\left(\mathbf{y}_{J_{\text{test}}}, \boldsymbol{F}_{J_{\text{test}}, \mathcal{I}(\mathcal{D}_{\text{train}}, \boldsymbol{\lambda})}\right)$.
- Because models trained on fewer points are typically worse on average than models trained on all n points, the estimator is *pessimistically biased*:

$$\mathbb{E}_{(J_{\text{train}}, J_{\text{test}})} \left[\widehat{\text{GE}}(\mathcal{I}, \boldsymbol{\lambda}, (J_{\text{train}}, J_{\text{test}}), \rho) \right] \geq \text{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}}, \rho).$$

- For regression tasks with loss-based measures such as MSE or MAE, this inequality means we systematically *overestimate* the true expected loss of a model trained on all n observations.

4) **Effect of training size** $|J_{\text{train}}|$ **on the variance of the hold-out estimator**

- We have the constraint $|J_{\text{train}}| + |J_{\text{test}}| = n$.
- **Large** training size $|J_{\text{train}}|$:
  - $|J_{\text{test}}|$ is small $\Rightarrow$ $\rho(\cdot)$ is computed on few test observations.
  - The estimator $\widehat{\text{GE}}(\mathcal{I}, \boldsymbol{\lambda}, (J_{\text{train}}, J_{\text{test}}), \rho)$ has *high variance* across different splits.
- **Small** training size $|J_{\text{train}}|$:
  - $|J_{\text{test}}|$ is large $\Rightarrow$ variance due to the test set is reduced.
  - But the model is trained on few data points, which increases pessimistic bias as in (3).
- There is a *trade-off* between bias and variance when choosing $|J_{\text{train}}|$: larger $|J_{\text{train}}|$ decreases bias but increases variance; smaller $|J_{\text{train}}|$ decreases variance but increases bias.