Solution 1: Risk Minimization and Gradient Descent

(a)

- Hypothesis space $\mathcal{H}$ is defined as:

$$\mathcal{H} = \{f(\mathbf{x}) = \boldsymbol{x}^\top \boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\}$$

- We fit a linear model, ergo using the $L2$ loss makes sense (e.g., because of the link to Gaussian MLE):

$$L\left(y^{(i)}, f\left(\boldsymbol{x}^{(i)} | \boldsymbol{\beta}\right)\right) = L\left(y^{(i)}, {\boldsymbol{x}^{(i)}}^\top \boldsymbol{\beta}\right) = 0.5\left(y^{(i)} - {\boldsymbol{x}^{(i)}}^\top \boldsymbol{\beta}\right)^2$$

  and the theoretical risk is

$$\mathcal{R}(f) = \mathcal{R}(\boldsymbol{\beta}) = \int L\left(y, f(\mathbf{x})\right) \mathrm{d}\mathbb{P}_{xy} = 0.5 \int (y - f(\mathbf{x})^2) \, \mathrm{d}\mathbb{P}_{xy} = 0.5 \int (y - \boldsymbol{x}^\top \boldsymbol{\beta})^2 \, \mathrm{d}\mathbb{P}_{xy}.$$

(b) The Bayes regret is $\mathcal{R}_L(\hat{f}) - \mathcal{R}_L^*$ and can be decomposed into an estimation error $\left[\mathcal{R}_L(\hat{f}) - \inf_{f \in \mathcal{H}} \mathcal{R}_L(f)\right]$ and an approximation error $[\inf_{f \in \mathcal{H}} \mathcal{R}_L(f) - \mathcal{R}_L^*]$.

(i) If $f^* \in \mathcal{H}$, $\mathcal{R}_L^* = \inf_{f \in \mathcal{H}} \mathcal{R}_L(f)$, i.e., the approximation error is 0 and for $n \to \infty$ the Bayes regret → 0.

(ii) If $f^* \notin \mathcal{H}$, the Bayes regret typically consists of both parts, but as $n \to \infty$, we are left with the approximation error.

(c)

- The empirical risk is

$$\mathcal{R}_{emp}(\boldsymbol{\beta}) = 0.5 \sum_{i=1}^{n} \left(y^{(i)} - {\boldsymbol{x}^{(i)}}^\top \boldsymbol{\beta}\right)^2 = 0.5||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2.$$

- Optimization = minimization of the empirical risk can either be done analytically (the preferred solution in this case!) or using, e.g., gradient descent.

$$\nabla_{\boldsymbol{\beta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\beta}) = 0.5 \nabla_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = -\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

(d) For convex objectives, every local minimum corresponds to a global minimum. To show convexity, calculate the second derivatives:

$$\nabla_{\boldsymbol{\beta}\boldsymbol{\beta}^\top} \mathcal{R}_{\text{emp}}(\boldsymbol{\beta}) = \boldsymbol{X}^\top \boldsymbol{X}.$$

Since $\boldsymbol{z}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{z}$ is the inner product of a vector $\tilde{\boldsymbol{z}} = \boldsymbol{X}\boldsymbol{z}$ with itself, i.e.

$$\boldsymbol{z}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{z} = \tilde{\boldsymbol{z}}^\top \tilde{\boldsymbol{z}} = \sum_{i=1}^{n} \tilde{z}_i^2$$

it is ≥ 0 and hence $\boldsymbol{X}^\top \boldsymbol{X}$ psd and therefore $\mathcal{R}_{\text{emp}}(\boldsymbol{\beta})$ convex.