1. **Monotonicity of loss and regularizer as the regularization parameter changes.** We choose the parameter $\theta \in \mathbb{R}^d$ to minimize the regularized empirical risk, $\mathcal{L}(\theta) + \lambda r(\theta)$, where $\mathcal{L}(\theta) = \frac{1}{n} \sum_i \ell(\theta^T x^i, y^i)$ is the empirical loss, $r(\theta)$ is the regularizer, and $\lambda \geq 0$ is the regularization parameter. (The exact form of $\mathcal{L}$ does not matter in this problem.) You might think that as $\lambda$ increases, $r(\theta)$ decreases while $\mathcal{L}(\theta)$ increases. In this exercise we verify that this is the case.

Suppose $0 < \lambda \leq \tilde{\lambda}$. Let $\theta^\ast$ minimize $\mathcal{L}(\theta) + \lambda r(\theta)$, and $\tilde{\theta}^\ast$ minimize $\mathcal{L}(\theta) + \tilde{\lambda} r(\theta)$.

(a) Show that $r(\theta^\ast) \geq r(\tilde{\theta}^\ast)$. In other words, increasing $\lambda$ will never make our regularization error larger. **Hint.** Use the fact that $\theta^\ast$ is the minimizer of $\mathcal{L}(\theta) + \lambda r(\theta)$ and similarly for $\tilde{\theta}^\ast$. Compare losses for $\theta^\ast$, $\tilde{\theta}^\ast$, $\lambda$ and $\tilde{\lambda}$.

(b) Show that $\mathcal{L}(\theta^\ast) \leq \mathcal{L}(\tilde{\theta}^\ast)$. That is, increasing $\lambda$ will never decrease our training error.

2. **Common loss functions.** You are given the following choices of penalty functions $p : \mathbb{R} \to \mathbb{R}$.

- **Square penalty.** Set $p^{\text{sqr}}(r) = r^2$.
- **Huber penalty.** For some parameter $\alpha > 0$,
  \[ p^{\text{hub}}(r) = \begin{cases} r^2 & \text{if } |r| \leq \alpha \\ \alpha(2|r| - \alpha) & \text{if } |r| > \alpha \end{cases}. \]

- **Log Huber penalty.** For some parameter $\alpha > 0$,
  \[ p^{\text{dh}}(r) = \begin{cases} r^2 & \text{if } |r| \leq \alpha \\ \alpha^2(1 - 2 \log(\alpha) + \log(r^2)) & \text{if } |r| > \alpha \end{cases}. \]

(a) Compute the gradient of each resulting empirical risk $\mathcal{L}$ obtained by setting $\ell(\hat{y}, y) = p(\hat{y} - y)$.

(b) State whether $\mathcal{L}$, formed under the given penalty function, is convex. Briefly justify your answers.

(c) Write Julia functions for each of the penalty functions $p$ which require a prediction $\hat{y}$ and label $y$, and return the resulting empirical risk $\mathcal{L}$.

(d) You are given a dataset `loss.json` with features $x_i \in \mathbb{R}$ represented in an $n$-vector $x$ and labels $y_i \in \mathbb{R}$ represented in an $n$-vector $y$. You are also given three predictors, shown below. For each predictor evaluate all three loss functions. Compare those values and pick the best performing predictor for the given dataset for each of the losses. Comment the results.
\[
\hat{y} = 15.5x + 5.9 \\
\hat{y} = 10x + 6.5 \\
\hat{y} = 5x + 7.9
\]

3. **Gradient descent.** You are given a dataset gd.json with features \(x_i^T \in \mathbb{R}^d\) represented in a \(n \times d\) matrix \(X\) and labels \(y_i \in \mathbb{R}\) represented in a \(n\)-vector \(y\). In this question you will be looking for an affine predictor using the Huber penalty shown below.

For some parameter \(\alpha > 0\),

\[
p^{\text{hub}}(r) = \begin{cases} 
    r^2 & \text{if } |r| \leq \alpha \\
    \alpha(2|r| - \alpha) & \text{if } |r| > \alpha
\end{cases}
\]

(a) Write a Julia function \(\theta = \text{GD}(X,y)\) which returns the optimal \(\theta^*\) by performing gradient descent for a Huber loss function with an update rule shown below. *Hint.* Question 2a.

\[
\theta^{k+1} = \theta^k - h^k \nabla f(\theta^k).
\]

(b) Perform a regression fitting using the newly created \(\text{GD}\) function. Experiment with different values of learning rate \(h^k\). Report the optimal \(\theta^*\) which minimizes the loss.

(c) Is the value of \(\theta\) optimal? Give examples when this approach succeeds and fails to identify the optimal value. Explain your answer.