Problems are labeled (*) for easy, (**) for medium, and (*** *) for hard. For homework 3, solve (at least) five stars worth of questions. Recommended exercises: 1, 2, 5, 6.

1. (*) Let \( y_i = x + e_i \) for \( i = 1, \ldots, n \), with \( x \in \mathbb{R} \), and \( e_i \sim N(0, \sigma^2) \) i.i.d. The MLE for \( x \) is the sample mean (empirical average) \( \hat{x} = \frac{1}{n} \sum_{i} y_i \). In the frequentist framework, \( x \) is fixed, and \( \hat{x} \) has a distribution \( p(\hat{x}|x) \). In the Bayesian framework with a uniform prior, \( x \) is random, and has a distribution \( p(x|y) \). Even though \( p(\hat{x}|x) \) and \( p(x|y) \) are philosophically different, show that their expressions coincide in this particular example. Generalize your argument to a general linear model of the form \( y_i = a^T_i x + e_i \), where \( x \in \mathbb{R}^m \).

2. (** Poisson noise) Suppose that a gamma-ray detector counts photons in \( n \) energy bins, yielding integer measurements \( y_i \), with \( i = 1, \ldots, n \). The photon counts \( y_i \) are independent and modeled by Poisson distributions \( P(y_i = x) = \frac{\lambda^x e^{-\lambda}}{x!} \), but the intensities \( \lambda_i \) that give rise to those counts are assumed to follow a power law \( \lambda_i = \mu \epsilon_i \), for some unknown \( \mu \), but known numbers \( \epsilon_i \) in geometric progression. Find the maximum likelihood estimator (MLE) of \( \mu \). [Hint: assume that \( \lambda_i = \mu \epsilon_i \) goes in the forward model, not in the prior.]

3. (***) In the setting of the previous question, find or characterize the maximum likelihood estimator of \( \lambda_i \) and \( \mu \), when a relaxed form of \( \lambda_i = \mu \epsilon_i \) is imposed via the prior \( p(\lambda_1, \ldots, \lambda_n, \mu) \sim \exp \left[ -\frac{1}{2\pi} \sum_i (\lambda_i - \mu \epsilon_i)^2 \right] \), for some small \( \delta > 0 \), rather than through the forward model.

4. (*) Jeffreys parameters, continued. In homework 1, we encountered positive parameters, whose distance is better measured via \( |\log \sigma_1 - \log \sigma_2| \) than via \( |\sigma_1 - \sigma_2| \). Consider a Bayesian prior \( p(\sigma) \) which is objective for such (Jeffreys) parameters, in the sense that it is a (unnormalized) uniform probability distribution for \( \log \sigma \).
   (a) What prior distribution does this give rise to, for \( \sigma \) itself?
   (b) What prior distribution does this give rise to, for \( \sigma^a \) with \( a > 0 \)?

5. (*) Consider \( y = Ax_0 + e \) with \( e_i \sim N(0, \sigma^2) \) i.i.d., and the Tikhonov-regularized least-squares problem
   \[
   \min_{x} \| Ax - y \|_2^2 + \lambda^2 \| x \|_2^2.
   \]
   In this exercise, suppose that \( \sigma \) and \( \| x_0 \| \) are known, and assume that \( A \) is square and a multiple of an isometry i.e., \( A^T A = A A^T = \sigma^2 I \). Find \( \lambda \) for which the MSE \( \mathbb{E} \| x - x_0 \|_2^2 \) is minimum. (The solution to this exercise illustrates the following useful heuristic: \( \lambda \) should be chosen so that the misfit and regularization terms are of comparable size.)

6. (**) Regularization by terminating the iterations. Consider solving \( Ax = y \) with square \( A \), by gradient descent:
   \[
   x_{n+1} = x_n + \alpha A^T (y - Ax), \quad x_0 = 0,
   \]
   for sufficiently small \( \alpha > 0 \). Let \( A = U \Sigma V^T \) be the singular value decomposition of \( A \); at convergence,
   \[
   x_{\infty} = \sum_i v_i \sigma_i^{-1} u_i^T y.
   \]
   (a) Find a function \( f_n(\sigma) \) for which
   \[
   x_n = \sum_i v_i f_n(\sigma_i) \sigma_i^{-1} u_i^T y.
   \]
   (b) In the limit of large \( n \), show that \( x_n \) is close to the solution of a Tikhonov-regularized least-squares problem, i.e., find \( \lambda(n) \) for which \( x_n \) is well-approximated by the solution of
   \[
   \min \| Ax - y \|_2^2 + \lambda(n)^2 \| x \|_2^2.
   \]
   (In practice, coupling the formula that you obtain for \( \lambda(n) \), with the conclusion of the previous exercise, shows a useful way to choose \( n \) for inverse problems.)