

# Stat243: Problem Set 8, Due Friday December 3 (5 pm)

November 19, 2021

This covers Unit 11 (plus a problem on importance sampling from Unit 9).

It's due **as PDF submitted to Gradescope** and submitted via GitHub at 5 pm on December 3.

Comments:

1. The formatting requirements are the same as previous problem sets.
2. Please note my comments in the syllabus about when to ask for help and about working together. In particular, **please give the names of any other students that you worked with on the problem set and indicate in comments any ideas or code you borrowed from another student.**

## Problems

1. Let's consider importance sampling and explore the need to have the sampling density have heavier tails than the density of interest. Assume that we want to estimate  $\phi = EX$  and  $\phi = E(X^2)$  with respect to a density,  $f$ . We'll make use of the Pareto distribution, which has the pdf  $p(x) = \frac{\beta\alpha^\beta}{x^{\beta+1}}$  for  $\alpha < x < \infty$ ,  $\alpha > 0$ ,  $\beta > 0$ . The mean is  $\frac{\beta\alpha}{\beta-1}$  for  $\beta > 1$  and non-existent for  $\beta \leq 1$  and the variance is  $\frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)}$  for  $\beta > 2$  and non-existent otherwise.
  - (a) Does the tail of the Pareto decay more quickly or more slowly than that of an exponential distribution?
  - (b) Suppose  $f$  is an exponential density with parameter value equal to 1, shifted by two to the right so that  $f(x) = 0$  for  $x < 2$ . Pretend that you can't sample from  $f$  and use importance sampling where our sampling density,  $g$ , is a Pareto distribution with  $\alpha = 2$  and  $\beta = 3$ . Use  $m = 10000$  to estimate  $EX$  and  $E(X^2)$  and compare to the known expectations for the shifted exponential. Recall that  $\text{Var}(\hat{\phi}) \propto \text{Var}(h(X)f(X)/g(X))$ . Create histograms of  $h(x)f(x)/g(x)$  and of the weights  $f(x)/g(x)$  to get an idea for whether  $\text{Var}(\hat{\phi})$  is large. Note if there are any extreme weights that would have a very strong influence on  $\hat{\phi}$ .
  - (c) Now suppose  $f$  is the Pareto distribution described above and pretend you can't sample from  $f$  and use importance sampling where our sampling density,  $g$ , is the exponential described above. Respond to the same questions as for part (b), comparing to the known values for the Pareto.
2. Consider the "helical valley" function (see the *ps8.R* file in the repository). Plot slices of the function to get a sense for how it behaves (i.e., for a constant value of one of the inputs, plot as a 2-d function of the other two). Syntax for *image()*, *contour()* or *persp()* (or the *ggplot2* equivalents) from the R bootcamp materials will be helpful. Now try out *optim()* and *nlm()* for finding the minimum of this function (or use *optimx()*). Explore the possibility of multiple local minima by using different starting points.

3. Consider a censored regression problem. We assume a simple linear regression model,  $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$ . Suppose we have an iid sample, but that for any observation with  $Y > \tau$ , all we are told is that  $Y$  exceeded the threshold and not its actual value. In a given sample,  $n_c$  of the  $n$  observations will (in a stochastic fashion) be censored, depending on how many exceed the fixed  $\tau$ . A real world example (but with censoring in the left tail) is in measuring pollutants, for which values below a threshold are reported as below the limit of detection. Another real world example is US tax revenue data where the incomes of wealthy taxpayers may be reported as simply exceeding, say, 1 million dollars.

- (a) Design an EM algorithm to estimate the three parameters,  $\theta = (\beta_0, \beta_1, \sigma^2)$ , taking the complete data to be the available data plus the actual values of the censored observations. You'll need to make use of  $E(W|W > \tau)$  and  $\text{Var}(W|W > \tau)$  where  $W$  is normally distributed. Be careful that you carefully distinguish  $\theta$  from the current value at iteration  $t$ ,  $\theta_t$ , in writing out the expected log-likelihood and computing the expectation and that your maximization be with respect to  $\theta$ . A few hints:

- i. Considering the notation we used in class when discussing EM, it's natural to think of  $Z$  as the (unobserved) values of the censored observations. You can think of  $n_c$  as being part of  $X$  (in the sense of the meaning of  $X$  in the class notes, not the  $x_i$  covariate values), along with the uncensored observations.
- ii. When you write out the complete data log-likelihood, it's helpful to set it up so there is a term involving the observed values and a term involving the censored values. When you condition the latter term, you'll be conditioning on the censored values being bigger than threshold.
- iii. From the Johnson and Kotz bibles on distributions, the mean and variance of the truncated normal distribution,  $f(W) \propto \mathcal{N}(\mu, \sigma^2)I(W > \tau)$ , are:

$$\begin{aligned} E(W|W > \tau) &= \mu + \sigma \rho(\tau^*) \\ V(W|W > \tau) &= \sigma^2 \left( 1 + \tau^* \rho(\tau^*) - \rho(\tau^*)^2 \right) \\ \rho(\tau^*) &= \frac{\phi(\tau^*)}{1 - \Phi(\tau^*)} \\ \tau^* &= (\tau - \mu)/\sigma, \end{aligned}$$

where  $\phi(\cdot)$  is the standard normal density and  $\Phi(\cdot)$  is the standard normal CDF.

- iv. You should recognize that your expected log-likelihood can be expressed as a regression of  $\{Y_{obs}, m_t\}$  on  $\{x\}$  where  $Y_{obs}$  are the non-censored data and  $\{m_{i,t}\}$ ,  $i = 1, \dots, n_c$  are used in place of the censored observations. Note that  $\{m_{i,t}\}$  will be functions of  $\theta_t$  and thus constant in terms of the maximization step. Your estimator for  $\sigma^2$  should involve a ratio where the numerator involves the usual sum of squares for the non-censored data plus an additional term that you should interpret statistically.
  - v. You should be able to analytically maximize the expected log likelihood.
- (b) Propose reasonable starting values for the three parameters as functions of the observations.
- (c) Write an R function, with auxiliary functions as needed, to estimate the parameters. Make use of the initialization from part (b). You may use `lm()` for updating  $\beta$ . You'll need to include criteria for deciding when to stop the optimization. Test your function using data simulated based on the code in `ps8.R` with (a) a modest proportion of exceedances expected, say 20%, and (b) a high proportion, say 80%.

- (d) A different approach to this problem just directly maximizes the log-likelihood of the observed data, which for the censored observations just involves the likelihood terms,  $P(Y_i > \tau)$ . Estimate the parameters (and standard errors) for your test cases using *optim()* with the BFGS option in R. You will want to consider reparameterization, and possibly use of the *parscale* argument. Compare how many iterations EM and BFGS take. Note that parts (c) and (d) together provide a nice test of your EM derivation and code, since you should get the same results from the two optimization approaches.