Math 6262: Introduction to Mathematical Statistics
Homework 3 (due on April 30)

Disclaimer: M-estimators. Let $\mathcal{P} := \{P_\theta | \theta \in \Theta \subseteq \mathbb{R}^d\}$ be a family of distributions supported on $\mathcal{Z} \subseteq \mathbb{R}^d$.[1] In the general paradigm of M-estimation, one treats the problem of estimating the true parameter $\theta^* \in \Theta$ that generated the i.i.d. observations $Z_{1:n} := (Z_1, ..., Z_n)$ as that of minimizing some loss function $\ell(\theta, Z)$. If $\ell(\theta, Z)$ is the negative log-likelihood for the p.d.f. of $P_\theta \in \mathcal{P}$, then we return to the maximum-likelihood paradigm, and recover MLE as the corresponding M-estimator:

$$\hat{\theta}_n \in \operatorname*{Argmin}_{\theta \in \Theta} \left\{ L_n(\theta) := \frac{1}{n} \sum_{i \in [n]} \ell(\theta, Z_i) \right\}. \tag{1}$$

In particular, we recover least-squares for $\ell(\theta, x) = \frac{1}{2}\|\theta - x\|_2^2$, corresponding to the full Gaussian location family $\{\mathcal{N}(\theta, \mathbf{I}_d) | \theta \in \mathbb{R}^d\}$. However, sometimes it makes sense to use a loss function $\ell(\theta, x)$ that is not the negative log-likelihood for P_θ – and possibly not even a log-likelihood at all. Such as:

(a) **Mean/location estimation:** here $\ell(\theta, z) = \varphi(\theta - z)$ for some $\varphi : \mathbb{R}^d \to \mathbb{R}_+$, often assumed convex, centrally symmetric ($\varphi(\mathbf{u}) = \varphi(-\mathbf{u})$), and minimized in the origin. Some examples are

ℓ_p-loss (for $p \geq 1$):

$$\ell_p(\theta, z) := \|\theta - z\|_p^p$$

Note that $p = 2$ corresponds to the quadratic loss (and the sample average estimator).

Huber loss ($d = 1$): $\ell(\theta, x) = h(\theta - x)$ where $h : \mathbb{R} \to \mathbb{R}$ is a convex and C^2 function

$$h(u) = \begin{cases} \frac{1}{2}u^2, & |u| \leq 1, \\ |u| - \frac{1}{2}, & |u| > 1. \end{cases}$$

(b) **Regression:** here $Z = (X, Y) \in \mathbb{R}^d \times \mathbb{R}$, and loss functions are $\rho(y - x^\top \theta)$ where $y - x^\top \theta$ is the *residual*; $\rho : \mathbb{R} \to \mathbb{R}_+$ is a *contrast function* (usually convex, even, nondecreasing on $\mathbb{R}_+$). E.g., one can take ρ from a $1d$-location estimation problem.

(c) **Classification:** here $Z = (X, Y) \in \mathbb{R}^d \times \{-1, 1\}$, and loss functions are $\ell(\theta, z) = \phi(-yx^\top \theta)$ where $yx^\top \theta$ is *the margin*, and $\phi : \mathbb{R} \to \mathbb{R}$ is a nondecreasing *cost function*. E.g.: (plot these!)
   - *logistic loss:* $\phi(u) = -\log_2(\frac{e^{-u}}{2\cosh(u)}) = \log_2(1 + e^{2u})$, corresponding to *logistic regression*;
   - *ReLU loss:* $\phi(u) = \max(u, 0)$, often used in neural networks.
   - *hinge loss:* $\phi(u) = \max\{u + 1, 0\}$, corresponding to the *"support vector machine"* (SVM).

[1] We assumed that the parameter θ and observation Z have the same dimensionl; this can be made more general.

In particular, this framework allows to treat various statistical problems (mean estimation, regression, classification, testing) in a unified way. The next several problems concern M-estimators.

1.a^o: Unbiased location estimation. Assume the actual data-generating distribution reads

$$Z = \theta^* + \xi$$

where $\xi \in \mathbb{R}^d$ has a centrally-symmetric p.d.f. $f(\cdot)$. Show that M-estimator (1) with $\ell(\theta, z) = \varphi(\theta - z)$ is unbiased (i.e. $\mathbb{E}_*[\hat{\theta}_n] = \theta^*$) when φ is centrally symmetric. (*Hint: what can you say about* $\nabla\varphi$*?*)

1.b^o: From regression to classification. In *regression* or in *one-dimensional location estimation*, one may want to go with a contrast function $\rho : \mathbb{R} \to \mathbb{R}_+$ that satisfies the following:

> (i) ρ is convex, even (thus minimized at 0), such that $\rho(0) = 0$ and $\rho''(0) = 1$, is 1-Lipschitz over $\mathbb{R}$ (i.e. $|\rho'| \le 1$), and such that $\rho(u) \ge |u| - C$ for some constant $C \ge 0$. (Think of Huber's function.)

Note that the square loss $\rho(u) = \frac{1}{2}u^2$ satisfies all these properties but the last one; enforcing the Lipschitz property (i.e., a global bound on $|\rho'|$) allows to ensure *robustness* of an estimator.[2] On the other hand, in *classification* it is desired to use a cost function $\phi : \mathbb{R} \to \mathbb{R}_+$ satisfying the following:

> (ii) ϕ is convex, 1-Lipschitz, and is an upper bound for the step function $\theta(u) := \mathbb{1}\{u \ge 0\}$ tight at 0 (i.e. $\phi(0) = 1$ and $\phi(u) \ge \theta(u)$ for all $u \in \mathbb{R}$).

1. Show that any contrast ρ satisfying (i) generates a cost function ϕ satisfying (ii), in the form:

$$\phi(u) = a\rho(u) + bu + c \tag{2}$$

   for some universal constants a, b, c independent of ρ. *Hint: to understand why this form, plot the* <u>*derivatives*</u> *of the Huber and logistic loss: both have a sigmoid shape, but different ranges.*

2. Apply the above rule to a "pseudo-Huber" function $\rho(u) = \log(\cosh(u))$ and sketch the graph of the resulting $\phi(u)$. Verify that $\rho(-\infty) > 0$ (strictly). Explain why (2) does not allow to ensure $\rho(-\infty) = 0$ (as for the logistic and hinge losses). *Hint: how many conditions to fit?*

1.c^o: Bayes estimator. In the Bayesian paradigm of statistical inference, instead of fixing θ at some (unknown) value θ^*, one allows θ to be a *random variable* with a *known* distribution Π over Θ. Let $L(\theta', \theta)$ be the *loss function* of inference θ' on the parameter value θ. This can be any nonnegative function on $\Theta \times \Theta$, but usually one would take $L(\theta', \theta^*) = L(\theta')$, the negative log-likelihood corresponding to the population $\mathbb{P}_\theta$ and evaluated at θ'.[3] Next, one defines the *risk*

$$R(g|\theta) := \mathbb{E}_{Z_{1:n} \sim \mathbb{P}_\theta^{\otimes n}}[L(\hat{\theta}(Z_{1:n}), \theta)]$$

[2] Intuitively, we do not want to be perturbed "too much" by a "counterfeit" datapoint that might be far from the bulk of datapoints in the sample. This was the key idea in Huber's 1964 seminal paper [Hub64]: he modified the square loss to enforce Lipschitzness, then established a minimax property of the corresponding M-estimator in the *contamination model*, where the data comes from a mixture of a normal distribution with (arbitrary) "parasitic" one.

[3] For example, $L(\theta', \theta) = \frac{1}{2}\|\theta' - \theta\|_2^2$ for the Gaussian location family $\{\mathcal{N}(\theta, \mathbf{I}_d), \theta \in \mathbb{R}\}$.

of an estimator $g := \hat{\theta}(\cdot)$; note that $R(\cdot|\theta)$ is a functional—called *risk functional*—over estimators, i.e. measurable functions $g : \mathcal{Z}^n \to \Theta$. Finally, the *Bayes risk* of estimator g w.r.t. Π is the functional

$$R_\Pi(g) := \mathop{\mathbb{E}}_{\theta\sim\Pi}[R(g|\theta)].$$

Clearly, R_Π depends on the prior Π, which has to be chosen "reasonably:" we put a priori weight $\pi(\theta)$ on each distribution $\mathbb{P}_\theta$ in the family $\mathcal{P}$, so we are now biased towards distributions that we consider "a priori more likely." On the positive side, once Π is fixed, the Bayes risk can be computed for any estimator, so we can compare estimators according to their Bayes risks – and construct the best one.

Definition 1. *Any estimator $g_\Pi : \mathcal{Z}^n \to \Theta$ minimizing the Bayes risk is called a* **Bayes estimator***.*

Show the following explicit characterization of Bayes estimators. (We can assume $n = 1$ – why?)

Theorem 1. *Any Bayes estimator $g_\Pi = \hat{\theta}_\Pi(\cdot)$ can be characterized as follows: for each possible observed value $z_{1:n} \in \mathcal{Z}^n$, the value $\hat{\theta}_\Pi(z_{1:n})$ minimizes the* posterior loss *given that $Z_{1:n} = z_{1:n}$, i.e.*

$$\hat{\theta}_\Pi(z_{1:n}) \in \mathop{\mathrm{Argmin}}_{\theta'\in\Theta} \int_\Theta L(\theta',\theta)\,\pi(\theta|z_{1:n})\,d\theta,$$

where $\pi(\theta|z_{1:n})$ is the posterior density (denoting with $f_\theta^{\otimes n}$ the p.d.f. of the product distribution $\mathbb{P}_\theta^{\otimes n}$):

$$\pi(\theta|z_{1:n}) = \frac{f_\theta^{\otimes n}(z_{1:n})\,\pi(\theta)}{\int_\Theta f_{\theta'}^{\otimes n}(z_{1:n})\,\pi(\theta')\,d\theta'}.$$

Hint: write $R_\Pi(g)$ explicitly as a double integral in θ and $z_{1:n}$. Then, treating $\hat{\theta}(z_{1:n})$ as a "continuum-vector" with "entries" indexed by $z_{1:n} \in \mathcal{Z}^n$, take partial derivative w.r.t. $\hat{\theta}(z_{1:n})$.

2^o: Confidence-boosted testing via voting. Let $X_1, ..., X_n$ be an i.i.d. sample from $\mathbb{P}_\theta$, $\theta \in \Theta$. Assume also that $n = 2mk$ for some $m, k \in \mathbb{N}$, and there is a deterministic test $\phi(x_{1:k})$ that, using k observations, distinguishes between the two hypotheses H_0, H_1[4] with confidence 2/3, that is

$$\max\left\{\sup_{\theta\in\Theta_0} \mathbb{E}_\theta[\phi(X_{1:k})],\ \sup_{\theta\in\Theta_1} \mathbb{E}_\theta[1-\phi(X_{1:k})]\right\} \le \frac{1}{3}.$$

Now, consider the following simple procedure:

1. Split $X_{1:n}$ into $2m$ batches $X^{(1)}, ..., X^{(2m)}$ of k observations each, i.e. $X^{(j)} := X_{k(j-1)+1\,:\,k(j-1)+k}$.
2. Let $Z_j := \phi(X^{(j)})$, and consider the test

$$\varphi = \varphi(X_{1:n}) = \mathbb{1}\left[\sum_{j\in[2m]} Z_j > m\right]$$

—in other words, accept/reject H_0 by aggregating the "basic" tests via the majority-vote rule.

[4]Corresponding to some partition $\Theta = \Theta_0 \sqcup \Theta_1$, but this is not important in the context of this problem.

(a) Working with the normal approximation for the binomial distribution (neglecting the error of this approximation), and using that $\mathbb{P}[U \ge u] \le e^{\frac{-u^2}{2}}$ where $U \sim \mathcal{N}(0,1)$, show the following:

$$\text{Risk}_m(\varphi) \;\; \text{“} \le \text{”} \;\; e^{-cm}.$$

Here $c > 0$ is a constant; $\text{Risk}_m(\varphi)$ is the worst-case error (of either type) for test $\varphi(X_{1:n})$ with $n = 2mk$; finally, "$\le$" means the following: *the inequality would be valid if the distribution of the appropriate asymptotically normal statistic were simply* replaced *with* $\mathcal{N}(0,1)$.

(b) Your next task is to show that the above inequality is actually valid and, moreover, valid in finite sample. To this end, justify (in English) the inequalities

$$\text{Risk}_m(\varphi) \le \sum_{j=m+1}^{2m} \binom{2m}{j} \left(\frac{1}{3}\right)^j \left(\frac{2}{3}\right)^{2m-j} \le m \binom{2m}{m} \left(\frac{1}{3}\right)^m \left(\frac{2}{3}\right)^m,$$

then bound the right-hand side. *Hint: you can use that* $\binom{2m}{m} \le 2^{2m}$ *(explain why this is true).*

3^o: Local behavior of f-divergences. In this exercise, you are invited to show that f-divergence with a *strictly convex* function f locally behaves as the χ^2-divergence. Namely, assume that $f : \mathbb{R}_{++} \to \mathbb{R}$ (where $\mathbb{R}_{++}$ is the set of all positive reals) satisfies the following assumptions:

- $f(1) = 0$;
- uniformly bounded third derivative on $\mathbb{R}_{++}$, that is f''' exists on $\mathbb{R}_{++}$ and $\sup_{r>0} |f'''(r)| < \infty$;
- f is strictly convex (and thus by the previous assumption $f''(r) > 0$ for any $r > 0$).

Recall that the associated f-divergence between two distributions P, Q with the same sample space, with densities p, q with respest to a dominating measure μ, is

$$D_f(P||Q) := \mathbb{E}_Q \left[f\left(\frac{dP}{dQ}\right) \right] \quad = \int_{\mathcal{X}} f\left(r(x)\right) q(x) d\mu(x),$$

where $r(x) := \frac{p(x)}{q(x)}$ is the likelihood ratio and $\mathcal{X}$ is the support of μ. Fixing some P and Q, consider the "segment" between them, that is, the family of distributions $P_t := (1-t)Q + tP$ for $0 \le t \le 1$.

1. Show that as $t \to 0$,

   $$D_f(P_t||Q) = (1 + o(1)) \frac{f''(1)}{2} \chi^2(P_t||Q)$$

   where $o(1) \to 0$ and $\chi^2(P||Q)$ is the chi-square divergence, i.e. $D_h(P||Q)$ with $h(r) = (1-r)^2$.
2. Verify that $\chi^2(P_t||Q) = t^2 \chi^2(P||Q)$ and conclude that $D_f(P_t||Q)$ is locally quadratic in t.

Hint: consider the 3rd-order Taylor expansion of $f(r)$ at $r = 1$; the 1st-order term must vanish.

References

[Hub64] Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964.