REGRESSION UNDER SHAPE CONSTRAINTS IN A A FULL RANK EXPONENTIAL FAMILY

Moulinath Banerjee

The University of Michigan

Abstract

Key words and phrases:

1 Introduction

Function estimation is a ubiquitous, and consequently well–studied problem in nonparametric statistics. The virtue of nonparametric function estimation lies in the fact that restrictive assumptions need not be placed on the function of interest for meaningful inference to be performed. This is often a sensible thing to do, especially when parametric formulations are suspect and can lead quite easily to mis-specification. However, in several scientific problems, qualitative background knowledge about the function may be available, and this if incorporated into the analysis can result in more powerful conclusions being drawn from the data. Shape–restrictions are typical examples of such qualitative knowledge, and appear in a large number of applications. In particular, monotonicity is a shape-restriction that shows up very naturally in different areas of application like reliability, renewal theory, epidemiology and biomedical studies. For example, the monotone function of interest could be a distribution function or a cumulative hazard function (survival analysis), the mean function of a counting processes (demography, reliability, clinical trials), a monotone regression function (dose-response modeling, modeling disease incidence as a function of distance from a toxic source), a monotone density (inference in renewal theory and other applications) or a monotone hazard rate (reliability).

Some of the early work on monotone function estimation goes back to the 1950's. Grenander (1956) derived the MLE of a decreasing density as the slope of the least concave majorant of the empirical distribution function based on i.i.d. observations. The pointwise asymptotic distribution of Grenander's estimator was established by Prakasa Rao

(1969). Brunk (1970) studied the problem of estimating a monotone regression function in a signal plus noise model, with additive homoscedastic errors. A key feature of these monotone function problems is the slower pointwise rate of convergence (usually $n^{1/3}$) of the MLE (under the stipulation that the derivative of the monotone function at the point of interest does not vanish), as compared to the faster $\sqrt{n}$ rate in regular parametric models. Moreover, the pointwise limit distribution of the MLE turns out to be a non-Gaussian one, and was characterized by Groeneboom(1989).

Monotone response models: Many of these monotone function models relate a covariate to the corresponding response through a monotone "link function". We formally describe this as follows. Consider independent and identically distributed observations $\{Y_i, X_i\}_{i=1}^n$, where each (Y_i, X_i) is distributed like (Y, X), and (Y, X) is distributed in the following way: The covariate X is assumed to possess a Lebesgue density p_X (with distribution function F_X). The conditional density of Y given that $X = x$ is given by $p(\cdot, \psi(x))$, where $\{p(\cdot, \theta) : \theta \in \Theta\}$ is a one–parameter exponential family of densities (with respect to some dominating measure) parametrized in the natural or canonical form, and ψ is a smooth (continuously differentiable) monotone increasing function that takes values in Θ. Recall that the density $p(\cdot, \theta)$ can be expressed as:

$$p(y, \theta) = \exp[\theta T(y) - B(\theta)]h(y).$$

Now, it is easy to see that,

$$E[T(Y)|X = x] = B' \circ \psi(x) \equiv \mu(x) \text{ (say)}.$$

Since B is infinitely differentiable on Θ (an open set) and ψ is continuous, $B^{(k)} \circ \psi$ is continuous, for every $k > 0$. Moreover, $B''(\theta) = \mathrm{Var}_\theta(T) > 0$ for any θ implies that B' is a strictly increasing function. It follows that B' is invertible (with inverse function, H, say), so that estimating the regression function μ is equivalent to estimating ψ. The function ψ is called the *link function* and as shown above, is in one–one correspondence with the monotone regression function μ.

We deal with pointwise estimation of μ (equivalently ψ) in the class of monotone response models introduced above. This will involve studying least squares as well as maximum likelihood estimates of the regression function. The asymptotic behavior of these statistics will be studied under the hypothesis that the derivative of the regression function does not vanish at the point of interest. Since, our formulation encompasses a large variety of models, a general approach will enable us to address many similar problems at the same stroke. We provide here some motivating examples of the monotone response models scenario that have been fairly well-studied in the literature.

(a) **Monotone Regression Model:** Consider the model

$$Y_i = \mu(X_i) + \epsilon_i \,,$$

where $\{(\epsilon_i, X_i)\}_{i=1}^n$ are i.i.d. random variables, ϵ_i is independent of X_i, each ϵ_i has normal distribution with mean 0 and variance σ^2, each X_i has a Lebesgue density $p_X(\cdot)$ and μ is a monotone function. The above model and its variants have been fairly well–studied in the literature on isotonic regression. Note, in particular, the reference to Brunk (1970) above. Here, $X \sim p_X(\cdot)$ and $Y \mid X = x \sim N(\mu(x), \sigma^2)$. This conditional density comes from the one–parameter exponential family $N(\eta, \sigma^2)$ (for fixed σ^2), η varying. To make the correspondence to the above framework, where we express the conditional density of Y in terms of the natural parameter, we take the natural sufficient statistic $T(Y) = Y$, the natural parameter $\theta = \eta/\sigma^2$, whence $B(\theta) = \theta^2\,\sigma^2/2$ and the link function $\psi(x) = \mu(x)/\sigma^2$. The function $\tilde{B}(x) = E(T(X) \mid Z = z) = \mu(z)$.

(b) **Binary Choice Model:** Here we have a dichotomous response variable $\Delta = 1$ or 0 and a continuous covariate X with a Lebesgue density $p_X(\cdot)$ such that $P(\Delta = 1 \mid X) \equiv G(X)$ is a smooth function of X. Thus, conditional on X, Δ has a Bernoulli distribution with parameter $G(X)$. To cast this model in terms of the natural parameter, we take the link function as $\psi(x) = \log\left(G(x)/(1 - G(x))\right)$ and $p(\delta, \theta) = \delta\theta - \log(1 + e^\theta)$ for $\theta \in \mathbb{R}$, whence the natural sufficient statistic $T(\delta) = \delta$, $B(\theta) = \log(1 + e^\theta)$and $\mu(x) = G(x)$. Models of this kind have been quite broadly studied in econometrics and statistics (see, for example, Dunson (2004), Newton, Czado and Chappell (1996), Salanti and Ulm (2003)). In a biomedical context one could think of Δ as representing the indicator of a disease/infection and Z the level of exposure to a toxin, or the measured level of a bio-marker that is predictive of the disease/infection. In such cases it is often natural to impose a monotonicity assumption on G.

An important special case of the above model (that will appear in our data–analysis section) is the "Current Status Model" from survival analysis that is used extensively in biomedical studies and epidemiological contexts and has received much attention among bio-statisticians and statisticians (see, for example, Sun and Kalbfleisch (1993), Sun (1999), Shiboski (1998), Huang (1996), Banerjee and Wellner (2001)). Consider n individuals who are checked for infection at independent random times $X_1, X_2, \ldots, X_n$; we set $\Delta_i = 1$ if individual i is infected by time X_i and 0 otherwise. We can think of Δ_i as $1\{T_i \leq X_i\}$ where T_i is the (random) time to infection (measured from some baseline period). We refer to X_i as the survival time for the i'th individual. The T_i's are assumed to be independent and also independent of the X_i's and are unknown. We are interested in making

inference on the increasing function F, the common distribution function of the T_i's. We note that $\{\Delta_i, X_i\}_{i=1}^n$ is an i.i.d. sample from the distribution of (Δ, X) where $X \sim h(\cdot)$ for some Lebesgue density h and $\Delta \mid X \sim \text{Bernoulli}(F(X))$. This is precisely the (monotone) binary choice model considered above, with F playing the role of G.

(c) **Poisson Regression Model:** Suppose that $X \sim p_X(\cdot)$ and $Y \mid X = x \sim \text{Poisson}(\psi(x))$ where ψ is a monotone function. We have n i.i.d. observations from this model. Here, we have $T(x) = x$, $\psi(x) = \log \lambda(x)$, and $p(y, \theta) = -e^\theta + y\,\theta - \log y!$, for $\theta \in \mathbb{R}$. The function $\mu(x) = \lambda(x)$. Here one can think of X as the distribution of a region from a hazardous point source (for example, a nuclear processing plant or a mine) and Y the number of cases of disease incidence at distance X (say, cancer occurrences due to radioactive exposure in the case of the nuclear processing plant, or Silicosis in the case of the mine). Given $X = x$, the number of cases of disease incidence Y at distance x from the source is assumed to follow a Poisson distribution with mean $\lambda(x)$ where λ can be expected to be monotonically decreasing in z, since the harmful effect should decay as we move further out from the source. Variants of this model have received considerable attention in epidemiological contexts (Stone (1988), Diggle, Morris and Morton–Jones (1999), Morton–Jones, Diggle and Elliott (1999)).

Let $\hat{\mu}$ denote the least squares estimate of μ based on the available data, and let $\hat{\mu}^0$ denote the constrained least squares estimate of μ, computed under the null hypothesis that $\mu(x_0) = \eta_0$ (equivalently $\psi(x_0) = \theta_0 \equiv H(\eta_0)$), for some interior point x_0 in the domain of X. Our goal will be to characterize these estimates (and as we will see shortly, this will lead directly to a characterization of the (unconstrained and constrained M.L.E.'s of the function ψ. We first present some background on solving least squares problems under monotonicity constraints.

Cumulative sum diagram and greatest convex minorant: Consider a set of points $\{(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_0 = y_0 = 0$ and $x_0 < x_1 < \ldots < x_n$; consider the left-continuous function $P(x)$ such that $P(x_i) = y_i$ and such that $P(x)$ is constant on (x_{i-1}, x_i). We will denote the vector of slopes (left–derivatives) of the greatest convex minorant (henceforth GCM) of $P(x)$ computed at the points $(x_1, x_2, \ldots, x_n)$ by $\text{slogcm}\,\{(x_i, y_i)\}_{i=0}^n$. The GCM of $P(x)$ is, of course, also the GCM of the function that one obtains by connecting the points $\{x_i, y_i\}_{i=0}^n$ successively, by means of straight lines. The slope of the convex minorant plays an important role in the characterization of solutions to least squares problems under monotonicity constraints.

Let $\mathcal{X} = \{x_1 < x_2 < \ldots < x_k\}$ be a linearly ordered set and let w be a positive (weight) function defined on this set. Let g be an arbitrary real-valued function defined

on this set. We seek to characterize $g^\star$, the least squares projection of g onto $\mathcal{F}$ (with respect to the weight function w) where $\mathcal{F}$ is the set of all functions f defined on $\mathcal{X}$ that are increasing with respect to the ordering on $\mathcal{X}$; i.e. $f(x_1) \leq f(x_2) \leq \ldots \leq f(x_k)$. In other words,

$$g^\star = \text{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} (f(x_i) - g(x_i))^2 \, w(x_i) \quad (\star) \ .$$

We show that there is indeed a unique vector that minimizes the above expression and that it is given by:

$$\{g^\star(x_i)\}_{i=1}^n = \text{slogcm}\, \{(W_i, G_i)\}_{i=0}^n$$

where $W_i = \sum_{j=1}^{i} w(x_j)$ and $G_i = \sum_{j=1}^{i} g(x_i)\, w(x_i)$. Note that $g^\star$ defined in this manner is an increasing function. It is also useful to keep in mind that the integers 1 through n can be split up into a number of (ordered) blocks $B_1, B_2, \ldots, B_r$, such that $g^\star(x_j) = c_i$ for all $j \in B_i$ and $c_1 < c_2 < \ldots < c_r$. If $n_1, n_2, \ldots, n_r \equiv n$ are the end-points of these blocks, we have $G_{n_i} = G^\star(W_{n_i})$ for $i = 1, 2, \ldots, r$, with $G^\star$ defined below. On each block B_i, we have that $c_i = (\sum_{j \in B_i} g(x_i)\, w(x_i) / (\sum_{j \in B_i} w(x_i))$.

Let $G^\star$ denote the GCM of the points $\{(W_i, G_i)\}_{i=0}^n$. Then, note that $G^\star(0) = G_0 = 0$ and $G^\star(W_n) = G_n$. Also, $G^\star(W_i) \leq G(W_i)$ for all i. Furthermore,

$$g^\star(x_l) < g^\star(x_{l+1}) \Rightarrow G^\star(W_l) = G_l \, . \tag{1}$$

Note that $G^\star$ is a piecewise linear function. Now, straightforward algebra shows that:

$$\sum_{i=1}^{n} (g(x_i) - f(x_i))^2 \, w(x_i) = \sum_{i=1}^{n} (g(x_i) - g^\star(x_i))^2 \, w(x_i) + \sum_{i=1}^{n} (g^\star(x_i) - f(x_i))^2 \, w(x_i)$$

$$+2 \sum_{i=1}^{n} (g(x_i) - g^\star(x_i))(g^\star(x_i) - f(x_i))\, w(x_i) \, .$$

We will show that the last term in the above display is non-negative, whence it will follow that

$$\sum_{i=1}^{n} (g(x_i) - f(x_i))^2 \, w(x_i) \geq \sum_{i=1}^{n} (g(x_i) - g^\star(x_i))^2 \, w(x_i) + \sum_{i=1}^{n} (g^\star(x_i) - f(x_i))^2 \, w(x_i) \quad (\star\star) \ ;$$

it is clear from the above display that $g^\star$ is definitely *a* minimizer of the sum of squares criterion (displayed on the right side of $(\star)$) over $f \in \mathcal{F}$. Uniqueness follows from the fact that if $\tilde{f}$ is some minimizer of the sum of squares criterion, then

$$\sum_{i=1}^{n} (g(x_i) - \tilde{f}(x_i))^2 \, w(x_i) = \sum_{i=1}^{n} (g(x_i) - g^\star(x_i))^2 \, w(x_i) \, ,$$

whence display $(\star\star)$ implies that

$$0 \geq \sum_{i=1}^{n} (g^\star(x_i) - \tilde{f}(x_i))^2 \, w(x_i) \, ,$$

and this, by the positivity of w is only possible if $\tilde{f}(x_i) = g^\star(x_i)$ for all i. Having taken care of the uniqueness, we focus now on establishing that

$$\sum_{i=1}^{n} (g(x_i) - g^\star(x_i))(g^\star(x_i) - f(x_i)) \, w(x_i) \geq 0 \, . \tag{2}$$

We claim that:

$$\sum_{i=1}^{n} (g(x_i) - g^\star(x_i)) \, (g^\star(x_i) - f(x_i)) \, w(x_i) = \sum_{i=2}^{n} \left[(f(x_i) - f(x_{i-1})) - (g^\star(x_i) - g^\star(x_{i-1})) \right] \left[G_{i-1} - G^\star(W_{i-1}) \right] \, . \tag{3}$$

We will establish this representation later. Given that this holds, we can conclude by (1) above, that

$$\sum_{i=2}^{n} \left[g^\star(x_i) - g^\star(x_{i-1}) \right] \left[G_{i-1} - G^\star(W_{i-1}) \right] = 0 \, ,$$

whence

$$\sum_{i=2}^{n} \left[(f(x_i) - f(x_{i-1})) - (g^\star(x_i) - g^\star(x_{i-1})) \right] \left[G_{i-1} - G^\star(W_{i-1}) \right]$$

equals

$$\sum_{i=2}^{n} \left[(f(x_i) - f(x_{i-1})) \right] \left[G_{i-1} - G^\star(W_{i-1}) \right] \, ,$$

and this is non-negative by dint of the facts that $f(x_i) \geq f(x_{i-1})$ for $i \geq 2$, that $G_{i-1} - G^\star(W_{i-1}) \geq 0$. This establishes (2). It only remains to establish (3). To this end, recall that $G_k = \sum_{i=1}^{k} g(x_i) \, w(x_i)$ and set $G_k^\star = G^\star(W_k) = \sum_{i=1}^{k} g^\star(x_i) \, w(x_i)$, and interpret

G_0 and $G_0^\star$ as 0. Also, note that $G_n = G_n^\star$. Then, we can write:

$$\begin{aligned}
\sum_{i=1}^{n} (g^\star(x_i) - f(x_i))(g(x_i) - g^\star(x_i))\, w(x_i) &= \sum_{i=1}^{n} (g^\star(x_i) - f(x_i))\, \left[G_i - G_{i-1} - (G_i^\star - G_{i-1}^\star)\right] \\
&= \sum_{i=1}^{n} (g^\star(x_i) - f(x_i))\, [G_i - G_i^\star] \\
&\qquad - \sum_{i=1}^{n} (g^\star(x_i) - f(x_i))\, \left[G_{i-1} - G_{i-1}^\star\right] \\
&= \sum_{1}^{n-1} (g^\star(x_i) - f(x_i))\, [G_i - G_i^\star] \\
&\qquad - \sum_{i=2}^{n} (g^\star(x_i) - f(x_i))\, \left[G_{i-1} - G_{i-1}^\star\right] \\
&= \sum_{1}^{n-1} (g^\star(x_i) - f(x_i))\, [G_i - G_i^\star] \\
&\qquad - \sum_{i=1}^{n-1} (g^\star(x_{i+1}) - f(x_{i+1}))\, [G_i - G_i^\star] \\
&= \sum_{i=1}^{n-1} [f(x_{i+1}) - f(x_i) - (g^\star(x_{i+1}) - g^\star(x_i))]\, (G_i - G_i^\star) \\
&= \sum_{i=2}^{n} [f(x_i) - f(x_{i-1}) - (g^\star(x_i) - g^\star(x_{i-1}))]\, (G_{i-1} - G_{i-1}^\star)\,.
\end{aligned}$$

This completes the derivation.

Least squares estimates of μ: We are now in a position to characterize the least squares estimate of μ in the general monotone response model. The unconstrained least squares estimate $\hat{\mu}$ is given by:

$$\hat{\mu} = \text{argmin}_{\mu \text{ increasing}} \sum_{i=1}^{n} (T(Y_i) - \mu(X_i))^2\,.$$

Let $\{X_{(i)}\}_{i=1}^n$ denote the ordered values of the X_i's and let $Y_{(i)}$ denote the response value corresponding to $X_{(i)}$. Since μ is increasing, the minimization problem is readily seen to reduce to one of minimizing $\sum_{i=1}^n (T(Y_{(i)}) - \mu_i)^2$ over all $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_n$ (where $\mu_i = \mu(X_{(i)})$). The solution to this problem is known, by the characterization of the least squares projection, that was derived above. It follows directly from that discussion (the

g_i's are the $T(Y_{(i)})$'s, the ordered set is $\{1, 2, \ldots, n\}$ and take the weight function to be identically equal to $1/n$) that $\{\hat{\mu}_i\}_{i=1}^n$, the minimizer over all $\mu_1 \leq \ldots \leq \mu_n$ is given by:

$$\{\hat{\mu}_i\}_{i=1}^n = \text{slogcm}\,\{G_n(X_{(i)}), V_n(X_{(i)})\}_{i=0}^n$$

where

$$G_n(x) = \frac{1}{n}\sum_{i=1}^n 1(X_i \leq x) \equiv \mathbb{P}_n\, 1(X \leq x) \;\text{ and }\; V_n(x) = \frac{1}{n}\sum_{i=1}^n T(Y_i)\, 1(X_i \leq x) \equiv \mathbb{P}_n\,(T(Y)\, 1(X \leq x))\,.$$

We interpret $X_{(0)}$ as $-\infty$, so that $G_n(0) = V_n(0) = 0$. The (unconstrained) least squares estimate of μ is formally taken to be the piecewise constant right continuous function, such that $\hat{\mu}(X_{(i)}) = \hat{\mu}_i$ for $i = 1, 2, \ldots, n$.

We next consider the problem of determining the constrained least squares estimator, where the constraint is given by $H_0 : \mu(x_0) = \eta_0$. It is easy to see that this amounts to solving two separate optimization problems: (a) Minimize $\sum_{i=1}^m (T(Y_{(i)} - \mu_i)^2$ over all $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_m \leq \eta_0$ and (b) Minimize $\sum_{i=m+1}^n (T(Y_{(i)}) - \mu_i)^2$ over all $\eta_0 \leq \mu_{m+1} \leq \mu_{m+2} \leq \ldots \leq \mu_n$; here m is that integer for which $X_{(m)} < x_0 < X_{(m+1)}$. The vector that solves (a) (say $\{\hat{\mu}_i^0\}_{i=1}^m$) is given by:

$$\{\hat{\mu}_i^0\}_{i=1}^m = \text{slogcm}\,\{G_n(X_{(i)}), V_n(X_{(i)})\}_{i=0}^m \;\wedge \eta_0\,,$$

where the minimum is interpreted as being componentwise. On the other hand, the vector that solves (b) (say $\{\hat{\mu}_i^0\}_{i=m+1}^n$) is given by:

$$\{\hat{\mu}_i^0\}_{i=m+1}^n = \text{slogcm}\,\{G_n(X_{(i)}) - G_n(X_{(m)}), V_n(X_{(i)} - V_n(X_{(m)})\}_{i=m}^n \;\vee \eta_0\,,$$

where the maximum is also interpreted as being componentwise. The constrained MLE $\hat{\mu}^0$ is then taken to be the piecewise constant right–continuous function, such that $\hat{\mu}^0(X_{(j)}) = \hat{\mu}_j^0$ for $j = 1, 2, \ldots, n$ and $\hat{\mu}^0(x_0) = \eta_0$, and such that $\hat{\mu}^0$ has no jumps outside the set $\{X_{(j)}\}_{j=1}^n \cup \{x_0\}$.

The characterization of the constrained least squares estimator follows from the following proposition which is left as an exercise.

Proposition: Let $\mathcal{X} = \{x_1 < x_2 < \ldots < x_n\}$ be an ordered set and let w be a positive weight function defined on $\mathcal{X}$. Let g be a real-valued function defined on $\mathcal{X}$ and (i) $\mathcal{F}_{u,\eta}$ denote the set of increasing functions defined on $\mathcal{X}$ that are bounded above by η, (ii) $\mathcal{F}_{l,\eta}$ denote the set of increasing functions defined on $\mathcal{X}$ that are bounded below by η. Show that the least squares projection of g on $\mathcal{F}_{u,\eta}$ is given by:

$$g_{u,\eta}^\star \equiv \text{slogcm}\,\{(W_i, G_i)\}_{i=0}^n \wedge \eta\,,$$

and the least squares projection of g on $\mathcal{F}_{l,\eta}$ is given by:

$$g^{\star}_{l,\eta} \equiv \text{slogcm}\left\{(W_i, G_i)\right\}_{i=0}^{n} \vee \eta\,.$$

While this proposition (as well as the characterization of $g^{\star}$ above) can be deduced using the Kuhn-Tucker theorem to be introduced shortly, I would like you to use an embellishment of the arguments that I used to derive $g^{\star}$ as the solution to the unconstrained least squares problem under monotonicity constraints.

Kuhn-Tucker theorem: Let ϕ be a strictly convex function defined on $\mathbb{R}^n$ and potentially assuming values in the extended real line. Define $R = \phi^{-1}(\mathbb{R})$ and consider the problem of minimizing ϕ on R, subject to a number of (convex) inequality and equality constraints that may be written as $g_i(x) \leq 0$ for $i = 1, 2, \ldots, k$ and $g_i(x) = 0$ for $i = k+1, \ldots, m$. Here, the g_i's are convex functions. Then $\hat{x} \in R$ uniquely minimizes ϕ subject to the m constraints if and only if there exist non-negative (Lagrange multipliers) $\lambda_1, \lambda_2, \ldots, \lambda_m$ such that (a) $\sum_{i=1}^{m} \lambda_i\, g_i(\hat{x}) = 0$ and (b) $\bigtriangledown \phi(\hat{x}) + G^{T}_{n\times m}\, \lambda_{m\times 1} = 0$, where $G_{m\times n}$ is the total derivative of the function $(g_1, g_2, \ldots, g_m)^T$ at the point $\hat{x}$.

Maximum likelihood estimators of ψ: The likelihood function for ψ, up to a multiplicative factor not depending on ψ, is given by:

$$L_n(\psi, \{Y_i, X_i\}_{i=1}^{n}) = \prod_{i=1}^{n} \exp(\psi(X_i)\, T(Y_i) - B(\psi(X_i)))\, h(Y_i)\,,$$

whence the log-likelihood function for ψ, up to an additive factor that does not depend on ψ, is given by:

$$l_n(\psi, \{Y_i, X_i\}_{i=1}^{n}) = \sum_{i=1}^{n} \left[\psi(X_i)\, T(Y_i) - B(\psi(X_i))\right] \equiv \sum_{i=1}^{n} \left[\psi(X_{(i)})\, T(Y_{(i)}) - B(\psi(X_{(i)}))\right].$$

Writing $\psi(X_{(i)}) = \psi_i$, it is seen that the problem of computing $\hat{\psi}_n$, the unconstrained MLE reduces to minimizing $\phi(\psi_1, \psi_2, \ldots, \psi_n) \equiv \sum_{i=1}^{n} \left[-\psi_i\, T(Y_{(i)}) + B(\psi_i)\right]$ over $\psi_1 \leq \psi_2 \leq \ldots \leq \psi_n$. The strict convexity of B implies the strict convexity of ϕ and the Kuhn-Tucker theorem may be invoked with $g_i(\tilde{\psi}) = \psi_i - \psi_{i+1}$ for $i = 1, 2, \ldots, n-1$ (here $\tilde{\psi}$ denotes the vector $(\psi_1, \psi_2, \ldots, \psi_n)$). Denoting the minimizer of ϕ by $\hat{\psi} = (\hat{\psi}_1, \hat{\psi}_2, \ldots, \hat{\psi}_n)$, the conditions (a) and (b) of that theorem translate to: $\lambda_i = \sum_{j=1}^{i} \left(T(Y_{(j)}) - B'(\hat{\psi}_j)\right) \geq 0$ for $i = 1, 2, \ldots, n-1$, and $\sum_{j=1}^{n} \left(T(Y_{(j)}) - B'(\hat{\psi}_j)\right) = 0$.

We now set $\hat{\psi}_i = H(\hat{\mu}_i)$ (where H is the inverse function of B'), so that $\hat{\mu}_i = B'(\hat{\psi}_i)$ and show that the above conditions are satisifed, whence it will follow immediately that

the unconstrained MLE $\hat{\psi} = H(\hat{\mu})$. Let $B_1, B_2, \ldots, B_r$ denote the (ordered) blocks of indices on which the solution $\hat{\mu}$ is constant, and let c_i denote the constant value of $\hat{\mu}_j$ for $j \in B_i$; let $m_1 < m_2 < \ldots < m_r \equiv n$ denote the end-points of these blocks. Then, from the characterization of the isotonic regression, it follows that for each $1 \leq i \leq r$ (interpret m_0 as 0):

$$\frac{\sum_{j=m_{i-1}+1}^{m_i} T(Y_{(j)})}{m_i - m_{i-1}} = c_i \tag{4}$$

and for $m_{i-1} < s < m_i$

$$\frac{\sum_{j=m_{i-1}+1}^{s} T(Y_{(j)})}{s - m_{i-1}} \geq c_i \,. \tag{5}$$

This shows that

$$\lambda_i \equiv \sum_{j=1}^{i} (T(Y_{(j)}) - \hat{\mu}_j) = 0 \;\text{ for }\; i = m_1, n_2, \ldots, m_{r-1} \,, \tag{6}$$

since

$$\lambda_{m_l} - \lambda_{m_{l-1}} = \sum_{j=m_{l-1}+1}^{m_l} (T(Y_{(j)}) - \hat{\mu}_j) = (m_l - m_{l-1}) \left(\frac{\sum_{j=m_{l-1}+1}^{m_l} T(Y_{(j)})}{m_l - m_{l-1}} - c_i \right) = 0 \,,$$

using (4), for $l = 1, 2, \ldots, r-1$ (interpret λ_0 as 0). Furthermore

$$\sum_{i=1}^{n} (T(Y_{(j)}) - B'(\hat{\psi}_i)) = \sum_{l=1}^{r} (m_l - m_{l-1}) \left(\frac{\sum_{j=m_{l-1}+1}^{m_l} T(Y_{(j)})}{m_l - m_{l-1}} - c_i \right) = 0 \,.$$

Next consider $1 \leq s \leq n-1$ such that s is not the end-point of a block, and suppose that s lies strictly between m_{l-1} and m_l (where $1 \leq l \leq r$). Then,

$$\lambda_s = \lambda_{m_{l-1}} + (s - m_{l-1}) \left(\frac{\sum_{j=m_{i-1}+1}^{s} T(Y_{(j)})}{s - m_{l-1}} - c_l \right) \geq 0 \,,$$

using (6) and (5). This shows that our choice of $\{\hat{\psi}_i\}_{i=1}^n$ does indeed satisfy the Kuhn-Tucker conditions and completes the argument.

Exercise: As in the case of the constrained least squares estimator, show, by splitting the likelihood maximization problem into two parts, and subsequently invoking the Kuhn-Tucker theorem, that the constrained MLE $\hat{\psi}^0$, computed under $H_0 : \psi(x_0) = \theta_0$ (where $\theta_0 = H(\eta_0)$), is given by $\hat{\psi}^0_n = H(\hat{\mu}^0)$.

Exercise – The Grenander estimator: Consider i.i.d. observations $X_1, X_2, \ldots, X_n$

from a density f on $[0, \infty)$ where f is given to be non-increasing. The goal of this exercise is to find the M.L.E. of f under this monotonicity constraint. Thus:

$$\hat{f}_{MLE} = \text{argmax}_{f \text{ nonincreasing density}} \prod_{i=1}^{n} f(X_i)\,.$$

(a) Show that without loss of generality the MLE can be taken to be a piecewise constant non-increasing function.

(b) Consider the left-hand slope of the least concave majorant of the empirical distribution function F_n based on the X_i's. This is a piecewise constant left continuous function. Denote this by $\tilde{f}$. Show that the vector $\{\tilde{f}(X_{(i)})\}_{i=1}^{n}$ minimizes $\sum_{i=1}^{n} (1/(n(X_{(i)} - X_{(i-1)})) - f_i)^2 (X_{(i)} - X_{(i-1)})$ over all $f_1 \geq f_2 \geq \ldots \geq f_n$.

(c) Using the Kuhn-Tucker theorem, show that $\tilde{f}$ is, in fact, the MLE of f.