Tatra Mt. Math. Publ. **39** (2008), 255–264

TATRA MOUNTAINS
Mathematical Publications

PRIMAL AND DUAL FORMULATIONS RELEVANT FOR THE NUMERICAL ESTIMATION OF A PROBABILITY DENSITY VIA REGULARIZATION

Roger Koenker — Ivan Mizera

ABSTRACT. General schemes relevant for the estimation of a probability density via regularization—primal and dual versions in the discretized setting—are investigated. Conditions for the dual solution to be a probability density are given, and a strong duality theorem is proved.

We study various instances of the problem

$$-\mathsf{w}^\mathsf{T}\mathsf{L}\mathsf{h} + \mathsf{s}^\mathsf{T}\Psi(\mathsf{g}) + \mathsf{J}(-\mathsf{P}\mathsf{h}) = \min_{\mathsf{g},\mathsf{h}}\,! \,, \qquad \text{subject to } \mathsf{h} \preceq \mathsf{g}\,, \tag{P}$$

where L and w are evaluation operator and averaging functional described later in the text; $\Psi(\mathsf{g})$ indicates the application of a real convex function ψ to the components of g, while $\mathsf{J}(\mathsf{h})$ is a general convex function applied to the whole vector $-\mathsf{P}\mathsf{h}$, the negative of the result of a linear operator P applied on h. We assume that vectors w and s have nonnegative elements; hereafter, $\succeq$ and $\preceq$ denote componentwise inequalities. If ψ is nondecreasing, the primal formulation (P) can be simplified—it is equivalent to the unconstrained problem

$$-\mathsf{w}^\mathsf{T}\mathsf{L}\mathsf{g} + \mathsf{s}^\mathsf{T}\Psi(\mathsf{g}) + \mathsf{J}(-\mathsf{P}\mathsf{g}) = \min_{\mathsf{g}}\,! \,. \tag{U}$$

Convex functions are allowed to attain $+\infty$ as a value; the *domain*, $\operatorname{dom}\Phi$, is the set where Φ is finite. We assume that all convex functions in (P) and (U) have domains with nonempty interiors. Concave functions are handled in an analogous manner, only the role of $+\infty$ is played by $-\infty$.

A *conjugate* of a convex function Φ is

$$\Phi^*(\mathsf{y}) = \sup_x \big(\mathsf{y}^\mathsf{T}\mathsf{x} - \Phi(\mathsf{x})\big) = \sup_{\mathsf{x}\in\operatorname{dom}\Phi} \big(\mathsf{y}^\mathsf{T}\mathsf{x} - \Phi(\mathsf{x})\big),$$

2000 Mathematics Subject Classification: Primary 62G07; Secondary 94A17, 90C46, 65J20.

Keywords: density estimation, penalty methods, duality, Rényi entropies.

This research was partially supported by NSF grant SES-05-44673 and by the Natural Sciences and Engineering Research Council of Canada.

the latter formulation avoiding the need to compute with infinite values. The conjugate of the function $\lambda\|\cdot\|_p$, where $\|\cdot\|_p$ stands for the ℓ^p norm, is the indicator of the ball in the dual norm, $\{\mathsf{x}\colon \|\mathsf{x}\|_q \le \lambda\}$, where q is the (Hölder) conjugate of p (that is, $q = \infty$ for $p = 1$, and $(1/p) + (1/q) = 1$ for $p > 1$) and the *indicator* of a convex set E is defined to be 0 for all $\mathsf{x} \in E$ and $+\infty$ otherwise. The conjugate of the indicator of the cone $\{\mathsf{x}\colon \mathsf{x} \succeq 0\}$ is the indicator of the polar cone $\{\mathsf{x}\colon \mathsf{x} \preceq 0\}$. Finally, the function $(1/2)\|\cdot\|_2^2$ is conjugate to itself; and consequently, $\lambda\|\cdot\|_2^2$ to $1/(4\lambda)\|\cdot\|_2^2$. Our references for convex analysis are Rockafellar [10], Boyd and Vandenberghe [1].

We claim that the dual of (P), or when equivalent, (U) is the problem

$$\begin{aligned} -\mathsf{s}^\mathsf{T}\Psi^*(\mathsf{f}) - \mathsf{J}^*(\mathsf{e}) &= \max_{\mathsf{f},\mathsf{e}}!\,, \\ \text{subject to} \quad \mathsf{S}\mathsf{f} = \mathsf{L}^\mathsf{T}\mathsf{w} + \mathsf{P}^\mathsf{T}\mathsf{e} \quad &\text{and} \quad \mathsf{f} \succeq 0\,, \end{aligned} \tag{D}$$

where $\mathsf{S} = \operatorname{diag}(\mathsf{s})$ and $\Psi^*(\mathsf{f})$ indicates the componentwise application of ψ^*.

Both (P)–(U) and (D) are relevant in the study of discretized, numerical formulations of regularized density estimation. The estimated density is represented by the vector f of its values on some collection of points, referred to as a *grid*. The evaluation operator L then expresses the position of n datapoints with respect to the grid via interpolation; for instance, if the datapoints are among gridpoints, then the ith row assigns 1 to a gridpoint equal to the ith datapoint and zero otherwise. The vector w assigns weights to the datapoints—as a rule, $1/n$ to each. Finally, s is the vector of integration weights attached to gridpoints: the identity $\mathsf{s}^\mathsf{T}\mathsf{f} = 1$ expresses the fact that the estimated density integrates to 1. Estimated probability densities are approximated by the densities with respect to the dominating measure on the grid whose atoms are given by s.

As for the penalization term, a typical P is a discretized version of a differential operator appearing in the continuous formulation of the regularization proposal. Typical J involves an ℓ^p norm and a tuning constant, λ, customary in this context: say, $\mathsf{J}(\mathsf{u}) = \lambda\|\mathsf{u}\|_1$ or $\mathsf{J}(\mathsf{u}) = \lambda\|\mathsf{u}\|_2^2$.

Regularization may be also expressed in a constrained form, in which J is the indicator of a set $\{\mathsf{u}\colon \|\mathsf{u}\|_p \le \Lambda\}$.

All these examples are symmetric: $\mathsf{J}(-\mathsf{u}) = \mathsf{J}(\mathsf{u})$. An asymmetric example is provided by J equal to the indicator of $\{\mathsf{u}\colon \mathsf{u} \preceq 0\}$, the style of penalization used in density estimation under monotonicity or convexity constraints.

The fact that the estimated f is indeed a probability density can be most conveniently verified through the dual formulation (D).

Theorem 1. *Suppose that* $\mathsf{w}^\mathsf{T}\mathsf{L}1 = 1$ *and* $\mathsf{P}1 = 0$. *Then the solution* f *of* (D) *satisfies* $\sum_j s_j f_j = 1$ *and* $f_j \ge 0$ *for every* j.

P r o o f. The nonnegativity constraint is directly included in the formulation of (D); it remains to verify that

$$\mathsf{s}^{\mathsf{T}}\mathsf{f} = 1^{\mathsf{T}}\mathsf{S}\mathsf{f} = 1^{\mathsf{T}}(\mathsf{L}^{\mathsf{T}}\mathsf{w} + \mathsf{P}^{\mathsf{T}}\mathsf{e}) = \mathsf{w}^{\mathsf{T}}\mathsf{L}1 + \mathsf{e}^{\mathsf{T}}\mathsf{P}1 = 1.$$

□

In the simplest case, when matrix L is composed of zeros, except for a single 1 in each row corresponding to a datapoint, $\mathsf{w}^{\mathsf{T}}\mathsf{L}$ makes these 1's multiplied by $1/n$; further multiplying by $\mathsf{1}$ makes them sum to 1. More generally, common interpolation schemes yield evaluation operators satisfying the assumption of Theorem 1. As far as potential operators P are concerned, they are discrete, difference versions of differential operators; as such, they annihilate constants—as can be directly verified for difference operators acting on sequences.

Compared to the dual (D), the relationship of the variables appearing in the primal formulations (P) or (U) to the estimated density is not explicit. However, once a strong duality of (P) and (D) is demonstrated true, then the relationship of g to f for qualified ψ is given by

$$\mathsf{f} = \Psi'(\mathsf{g})\,, \tag{E}$$

where $\Psi'(\mathsf{g})$ indicates the componentwise application of ψ', the derivative of ψ.

Theorem 2. *Problem* (D) *is a strong dual of the problem* (P). *If ψ is differentiable on the interior I of its domain, then the corresponding solutions of* (D) *and* (P) *satisfy* (E), *whenever* g *and* f *are componentwise from I and the image of I under ψ', respectively.*

P r o o f. We take a formulation equivalent to (P), obtained by rewriting it in terms of new variables u and v,

$$\begin{aligned} &-\mathsf{w}^{\mathsf{T}}\mathsf{L}(\mathsf{g}-\mathsf{v}) + \mathsf{s}^{\mathsf{T}}\Psi(\mathsf{g}) + \mathsf{J}(\mathsf{u}) = \min_{\mathsf{g},\mathsf{u},\mathsf{v}}!\,, \\ &\text{subject to} \quad \mathsf{v} \succeq 0 \quad \text{and} \quad -\mathsf{P}(\mathsf{g}-\mathsf{v}) = \mathsf{u}\,. \end{aligned} \tag{1}$$

The Lagrange dual of (1) is

$$\inf_{\mathsf{g},\mathsf{u},\mathsf{v}} \mathcal{L}(\mathsf{p},\mathsf{e};\mathsf{g},\mathsf{u},\mathsf{v}) = \max_{\mathsf{p},\mathsf{e}}!\,, \qquad \text{subject to} \quad \mathsf{p} \succeq 0\,, \tag{2}$$

where (p used here has no relationship to the parameter p used elsewhere)

$$\mathcal{L}(\mathsf{p},\mathsf{e};\mathsf{g},\mathsf{u},\mathsf{v}) = -\mathsf{w}^{\mathsf{T}}\mathsf{L}(\mathsf{g}-\mathsf{v}) + \mathsf{s}^{\mathsf{T}}\Psi(\mathsf{g}) + \mathsf{J}(\mathsf{u}) + \mathsf{p}^{\mathsf{T}}(-\mathsf{v}) + \mathsf{e}^{\mathsf{T}}\big[-\mathsf{u} - \mathsf{P}(\mathsf{g}-\mathsf{v})\big]$$

is the Lagrangean of (1). The linear part of $\mathcal{L}$, in v, leads to a feasibility constraint

$$\mathsf{L}^{\mathsf{T}}\mathsf{w} + \mathsf{P}^{\mathsf{T}}\mathsf{e} = \mathsf{p}\,, \tag{3}$$

preventing the objective function of (2) from becoming $-\infty$. Under (3), the minimization of the simplified Lagrangean can be done separately in g and u,

$$\begin{aligned}\inf_{\mathsf{g},\mathsf{u}} \mathcal{L}(\mathsf{p},\mathsf{e};\mathsf{g},\mathsf{u}) &= \inf_{\mathsf{g},\mathsf{u}}\big(-\mathsf{w}^\mathsf{T}\mathsf{L}\mathsf{g} - \mathsf{e}^\mathsf{T}\mathsf{P}\mathsf{g} + \mathsf{s}^\mathsf{T}\Psi(\mathsf{g}) - \mathsf{e}^\mathsf{T}\mathsf{u} + \mathsf{J}(\mathsf{u})\big) \\ &= \inf_{\mathsf{g}}\big(-(\mathsf{L}^\mathsf{T}\mathsf{w} + \mathsf{P}^\mathsf{T}\mathsf{e})^\mathsf{T}\mathsf{g} + \mathsf{s}^\mathsf{T}\Psi(\mathsf{g})\big) + \inf_{\mathsf{u}}\big(-\mathsf{e}^\mathsf{T}\mathsf{u} + \mathsf{J}(\mathsf{u})\big), \\ &= \inf_{\mathsf{g}}\big(-\mathsf{p}^\mathsf{T}\mathsf{g} + \mathsf{s}^\mathsf{T}\Psi(\mathsf{g})\big) - \mathsf{J}^*(\mathsf{e})\,.\end{aligned} \tag{4}$$

Minimizing in g is done by expanding into components,

$$\begin{aligned}\inf_{\mathsf{g}}\big(-\mathsf{p}^\mathsf{T}\mathsf{g} + \mathsf{s}^\mathsf{T}\Psi(\mathsf{g})\big) &= \inf_{\mathsf{g}}\bigg(-\sum_j \mathsf{p}_j\mathsf{g}_j + \sum_j \mathsf{s}_j\psi(\mathsf{g}_j)\bigg) \\ &= \sum_j \mathsf{s}_j \inf_{\mathsf{g}_j}\bigg(-\frac{\mathsf{p}_j}{\mathsf{s}_j}\,\mathsf{g}_j + \psi(\mathsf{g}_j)\bigg) = -\sum_j \mathsf{s}_j\psi^*\bigg(\frac{\mathsf{p}_j}{\mathsf{s}_j}\bigg).\end{aligned} \tag{5}$$

The dual formulation (D) is obtained as the summary of (2)–(5), rewritten in terms of $\mathsf{f}_j = \mathsf{p}_j/\mathsf{s}_j$. Finally, (1) satisfies the Slater constraint qualification condition; therefore strong duality holds.

For fixed y, the domain of the concave function $\varphi(x) = yx - \psi(x)$ is the same as the domain of ψ. If ψ has a derivative on I, so does φ; if y belongs to a range of I under ψ', then there is x^* in I, depending on y, such that $y = \psi'(x^*)$. That is, $\varphi'(x^*) = 0$, and consequently φ attains its global maximum at x^*, because φ is concave. Hence, the conjugate is $\psi^*(y) = yx^* - \psi(x^*)$ and can be obtained via taking the derivative of ψ and setting it equal to zero. Applying this procedure componentwise in (5) yields $\psi'(\mathsf{g}_j) = \mathsf{p}_j/\mathsf{s}_\mathsf{j} = \mathsf{f}_j$, whenever the additional assumptions of the theorem are satisfied. □

Example (Maximum Likelihood). In continuous version, this primal formulation can be traced back to Leonard [7] and Silverman [12]. The latter proposed

$$-\int g dP_n + \int \mathrm{e}^g\,\mathrm{d}x + \lambda\int \big(g^{(k)}\big)^2\,\mathrm{d}x = \min_g !\tag{6}$$

using the third ($k = 3$) derivative to estimate the logarithm, g, of a density f, with the symbol P_n denoting the empirical probability supported by the datapoints; Gu [3] and others championed second ($k = 2$) derivative instead. The total variation penalty

$$\int |g^{(k)}|\,\mathrm{d}x = \bigvee g^{(k-1)}$$

was considered by Koenker and Mizera [5, 6] for $k = 1, 2, 3$. Rufibach and Dümbgen [11] investigated maximum likelihood estimation of a log-concave density, which in our setting corresponds to $k = 2$ and the penalty

in the form of the non-positivity constraint on the second derivative (with no tuning parameter λ).

In the discrete setting, the kth derivative operator is replaced by an appropriate difference operator P, and the evaluation operator L and vector of weights w by their typical instances described above. Since $\psi(x) = \mathrm{e}^x$ is nondecreasing, (P) is equivalent to the unconstrained formulation (U), whose specific form is, for symmetric $\mathsf{J}(\mathsf{u}) = \lambda\|\mathsf{u}\|_p^p$ and $p = 1, 2$,

$$-\mathsf{w}^{\mathsf{T}}\mathsf{L}\mathsf{g} + \mathsf{s}^{\mathsf{T}}\mathrm{e}^{\mathsf{g}} + \lambda\|\mathsf{P}\mathsf{g}\|_p^p = \min_{\mathsf{g}}!\,, \tag{7}$$

where e^{g} is understood componentwise. The additional assumptions of Theorem 2 are satisfied, so that indeed $\mathsf{f} = \mathrm{e}^{\mathsf{g}}$, and

$$\psi^*(y) = \begin{cases} y\log y - y\,, & \text{for } y > 0\,, \\ 0\,, & \text{for } y = 0\,, \\ +\infty\,, & \text{otherwise}\,. \end{cases}$$

The feasibility requirement related to the fact that $\operatorname{dom}\psi^* = [0, +\infty)$ independently enforces the nonnegativity constraint on f. Silverman [12] showed, via an argument based on the specific properties of the exponential function, that the result of (6) is a probability density; the same conclusion follows, in the discrete setting, from our Theorems 1 and 2 for all formulations of the type (7). If the assumptions of Theorem 1 regarding P, L, and w are satisfied, then the dual objective function

$$-\sum_j \mathsf{s}_j\mathsf{f}_j\log\mathsf{f}_j + \sum_j \mathsf{s}_j\mathsf{f}_j\,,$$

can be further simplified, because the second sum is equal to 1, a constant. The resulting dual of (7), cast in the minimization form, is, for $p = 1$,

$$\begin{gathered} \sum \mathsf{s}_j\mathsf{f}_j\log\mathsf{f}_j = \min_{\mathsf{f},\mathsf{e}}!\,, \\ \text{subject to} \quad \mathsf{S}\mathsf{f} = \mathsf{L}^{\mathsf{T}}\mathsf{w} + \mathsf{P}^{\mathsf{T}}\mathsf{e}, \quad \mathsf{f} \succeq 0, \quad \text{and} \quad \|\mathsf{e}\|_\infty \le \lambda\,, \end{gathered} \tag{8}$$

and for $p = 2$,

$$\begin{gathered} \sum \mathsf{s}_j\mathsf{f}_j\log\mathsf{f}_j + \frac{1}{4\lambda}\|\mathsf{e}\|_2^2 = \min_{\mathsf{f},\mathsf{e}}!\,, \\ \text{subject to} \quad \mathsf{S}\mathsf{f} = \mathsf{L}^{\mathsf{T}}\mathsf{w} + \mathsf{P}^{\mathsf{T}}\mathsf{e}\,, \quad \text{and} \quad \mathsf{f} \succeq 0\,. \end{gathered} \tag{9}$$

The dual of the penalty-constrained version of the primal (7),

$$-\mathsf{w}^{\mathsf{T}}\mathsf{L}\mathsf{g} + \mathsf{s}^{\mathsf{T}}\mathrm{e}^{\mathsf{g}} = \min_{\mathsf{g}}!\,, \qquad \text{subject to} \quad \|\mathsf{P}\mathsf{g}\|_p \le \Lambda\,, \tag{10}$$

is (p and q being conjugate)

$$\begin{gathered}\sum_j \mathsf{s}_j \mathsf{f}_j \log \mathsf{f}_j + \Lambda \|\mathsf{e}\|_q = \min_{\mathsf{f},\mathsf{e}} !\,, \\ \text{subject to} \quad \mathsf{S}\mathsf{f} = \mathsf{L}^\intercal \mathsf{w} + \mathsf{P}^\intercal \mathsf{e}\,, \quad \text{and} \quad \mathsf{f} \succeq 0\,.\end{gathered} \tag{11}$$

Finally, the dual of the shape-constrained formulation,

$$-\mathsf{w}^\intercal \mathsf{L}\mathsf{g} + \mathsf{s}^\intercal \Psi(\mathsf{g}) = \min_{\mathsf{g}} !\,, \qquad \text{subject to} \quad \mathsf{P}\mathsf{g} \preceq 0 \tag{12}$$

(yielding log-concave f when P is a second-order difference operator), is

$$\begin{gathered}\sum_j \mathsf{s}_j \mathsf{f}_j \log \mathsf{f}_j = \min_{\mathsf{f},\mathsf{e}} !\,, \\ \text{subject to} \quad \mathsf{S}\mathsf{f} = \mathsf{L}^\intercal \mathsf{w} + \mathsf{P}^\intercal \mathsf{e}\,, \quad \mathsf{f} \succeq 0\,, \quad \text{and} \quad \mathsf{e} \preceq 0\,.\end{gathered} \tag{13}$$

The essence of all the dual variants is the maximization of the Shannon entropy of f, or, equivalently, the minimization of the Kullback-Leibler divergence

$$\mathcal{K}(\mathsf{f}, \sigma^{-1}) = \sum_j \mathsf{s}_j \mathsf{f}_j \log \frac{\mathsf{f}_j}{\sigma^{-1}} = \sum_j \mathsf{s}_j \mathsf{f}_j \log \frac{\mathsf{s}_j \mathsf{f}_j}{\mathsf{s}_j \sigma^{-1}} = \sum_j \mathsf{s}_j \mathsf{f}_j \log \mathsf{f}_j + \log \sigma\,,$$

where $\sigma^{-1} = (\sum_j \mathsf{s}_j)^{-1}$ can be viewed as a discretization of the uniform density.

The dual formulation of the penalized likelihood problem as a maximum entropy problem can be generalized by replacing the Shannon entropy term by some of the Rényi entropies, indexed by a parameter $\alpha > 0$; similarly to the Kullback-Leibler case, the appropriate minimum divergence interpretations follow. Formally, Rényi's entropies include the Shannon one for $\alpha = 1$; the R é n y i [9] entropy with exponent $\alpha \neq 1$ is defined as $(1-\alpha)^{-1} \log\bigl(\mathsf{s}^\intercal \mathsf{f}^\alpha\bigr)$, where f^α is interpreted componentwise. The maximization of this function is equivalent to the maximization of $-\operatorname{sign}(\alpha-1)\mathsf{s}^\intercal \mathsf{f}^\alpha$ or, equivalently, $-\operatorname{sign}(\alpha-1)\mathsf{s}^\intercal \mathsf{f}^\alpha/\alpha$.

Let ψ_p be a function equal to x^p/p for $x \geq 0$ and to 0 for $x < 0$. The conjugate, ψ_p^*, of ψ_p is for $p > 1$ equal to y^q/q for $y \geq 0$ (p and q conjugate), and to $+\infty$ otherwise. Note that ψ_p is nondecreasing, hence (P) is equivalent to (U) whenever $\psi = \psi_p$.

Example (Minimum Pearson χ^2). The special case of the Rényi system for $\alpha = 2$ yields $\psi(x) = \psi_2$ and $\psi_2^* = y^2/2$ for $y \geq 0$. The dual is obtained by replacing the entropy term $\sum_j \mathsf{s}_j \mathsf{f}_j \log \mathsf{f}_j$ in the objective function of (8), (9), (11), and (13) by $\mathsf{s}^\intercal \mathsf{f}^2$, and eliminating the redundant constant in the objective.

The corresponding primal results from replacing $\sum_j \mathsf{s}_j e^{\mathsf{g}_j}$ in (7), (10), and (12) by $\sum_j \mathsf{s}_j \psi_2(\mathsf{g}_j)$. Minimizing the dual (and in this case also primal) objective is

equivalent to minimizing the χ^2-divergence

$$\chi^2(\mathsf{f},\sigma^{-1}) = \sum_j s_j \frac{(\mathsf{f}_j - \sigma^{-1})^2}{\sigma^{-1}} = \sum_j \frac{(\mathsf{s}_j\mathsf{f}_j - \mathsf{s}_j\sigma^{-1})^2}{\mathsf{s}_j\sigma^{-1}} = \sigma\left(\sum_j \mathsf{s}_j\mathsf{f}_j^2\right) - 1\,.$$

If instead of ψ_2 we consider $\psi(x) = (1/2)x^2$ for all x, we can cast both primal and dual in a quadratic programming form. However, the correct primal has to be written in the constrained form (P) now, because ψ is no longer monotone. In particular, the correct formulation for the setting corresponding to (7) is

$$-\mathsf{w}^\mathsf{T}\mathsf{L}\mathsf{h} + \tfrac{1}{2}\mathsf{s}^\mathsf{T}\mathsf{g}^2 + \lambda\|\mathsf{P}\mathsf{h}\|_p^p = \min_{\mathsf{g},\mathsf{h}}!\,, \qquad \text{subject to} \quad \mathsf{h} \preceq \mathsf{g}\,.$$

In all variants, both primal and dual yield directly $\mathsf{f} = \mathsf{g}$, because $\psi'(x) = x$.

Example. Another special case of the Rényi scheme, with $\alpha = 3/2$, puts $\mathsf{s}^\mathsf{T}\mathsf{f}^{3/2}$ into the objective function of (8), (9), (11), and (13). For the primal, we may take either $\sum_j \mathsf{s}_j\psi^3(\mathsf{g}_j)$ in (7), (10), and (12); or we may use $\psi(x) = (1/3)|x|^3$ instead, leaving the dual unchanged, but making the primal constrained; for instance, the formulation (7) becomes

$$-\mathsf{w}^\mathsf{T}\mathsf{L}\mathsf{h} + \tfrac{1}{3}\mathsf{s}^\mathsf{T}\mathsf{g}^3 + \lambda\|\mathsf{P}\mathsf{h}\|_p^p = \min_{\mathsf{g},\mathsf{h}}!\,, \qquad \text{subject to } \mathsf{h} \preceq \mathsf{g}\,.$$

Due to the fact that $\mathsf{f} = \mathsf{g}^2$ in any of these variants, we could nickname this example "Silverman for Good". Apart from the additional middle term, the objective function differs from the original proposal of Good [2] also in the first term; ours is not based on the logarithm of the square root of the estimated density, but on the square root itself. It would be interesting to know whether there is any Bayesian justification for such an approach, whether in "mufti" or "full regalia". In any case, the primal formulation yields a square root of a probability density, a "rootogram" in Tukey's terminology.

Example (Minimum Hellinger). Another example from the Rényi system, with $\alpha = 1/2$, sets $\psi(x) = -1/x$, for $x < 0$ and $+\infty$ elsewhere. The conjugate is $\psi^*(y) = -2\sqrt{y}$, for $y \geq 0$, and ∞ elsewhere. The dual (for $p = 1$) has, in the minimization form and after the elimination of the redundant constant, $-\mathsf{s}^\mathsf{T}\sqrt{\mathsf{f}}$ in the objective of (8), (9), (11), and (13); $\sqrt{\mathsf{f}}$ is again applied componentwise. The dual objective minimizes the Hellinger distance

$$\begin{aligned}\mathcal{H}(\mathsf{f},\sigma^{-1}) &= \sum_j \mathsf{s}_j\left(\sqrt{\mathsf{f}_j} - \sqrt{\sigma^{-1}}\right)^2 \\ &= \sum_j \left(\sqrt{\mathsf{s}_j\mathsf{f}_j} - \sqrt{\mathsf{s}_j\sigma^{-1}}\right)^2 \\ &= 2 - 2\sqrt{\sigma^{-1}}\sum \mathsf{s}_j\sqrt{\mathsf{f}_j}\,.\end{aligned}$$

Since ψ is nondecreasing, the primal can be cast in its unconstrained version (U), just replacing the $\sum_j \mathsf{s}_j e^{\mathsf{g}_j}$ term in (7), (10), or (12) by $-\mathsf{s}^\mathsf{T}\mathsf{g}^{-1}$, where g^{-1} is the componentwise reciprocal value of g; however, the domain restriction for ψ has to be included as a feasibility constraint. The resulting primal analog of (7) is

$$-\mathsf{w}^\mathsf{T}\mathsf{L}\mathsf{g} - \mathsf{s}^\mathsf{T}\mathsf{g}^{-1} + \lambda\|\mathsf{P}\mathsf{g}\|_1 = \min_{\mathsf{g}}!\,, \qquad \text{subject to} \quad \mathsf{g} \preceq 0\,.$$

For symmetric penalties, it is more convenient to recast the primal in terms of $\mathsf{h} = -\mathsf{g}$:

$$\mathsf{w}^\mathsf{T}\mathsf{L}\mathsf{h} + \mathsf{s}^\mathsf{T}\mathsf{h}^{-1} + \lambda\|\mathsf{P}\mathsf{h}\|_1 = \min_{\mathsf{h}}!\,, \qquad \text{subject to} \quad \mathsf{h} \succeq 0\,.$$

The estimated density $\mathsf{f} = 1/\mathsf{g}^2 = 1/\mathsf{h}^2$; hence h could be called, in the Tukey spirit, a "rootosparsity", and g, being negative, a "hanging rootosparsity".

In our implementations, we observed that numerical performance may be improved by adding the (theoretically redundant) nonnegativity constraint $\mathsf{f} \succeq 0$ also in the primal formulation. However, this is rather an unimportant detail, because dual formulations always ran significantly faster and were more numerically stable than their primal counterparts.

Example (Maximum empirical likelihood). The limiting variant of the Rényi system for $\alpha = 0$ is $\psi(x) = -1/2 - \log(-x)$ for $x < 0$, and $+\infty$ otherwise. The dual puts $-\mathsf{s}^\mathsf{T}\log\mathsf{f}$ into the objective function of (8), (9), (11), and (13), while the primal (unconstrained, but with a feasibility constraint) puts $-\mathsf{s}^\mathsf{T}\log(-\mathsf{g})$ into in (7), (10), or (12). For instance, recasting (7) in terms of $\mathsf{h} = -\mathsf{g}$ gives

$$\mathsf{w}^\mathsf{T}\mathsf{L}\mathsf{h} - \mathsf{s}^\mathsf{T}\log\mathsf{h} + \lambda\|\mathsf{P}\mathsf{h}\|_1 = \min_{\mathsf{h}}!\,, \qquad \text{subject to} \quad \mathsf{h} \succ 0\,.$$

The dual objective is equivalent to the reversed Kullback-Leibler divergence

$$\begin{aligned}\mathcal{K}(\sigma^{-1}, \mathsf{f}) &= \sum_j \mathsf{s}_j \sigma^{-1} \log \frac{\sigma^{-1}}{\mathsf{f}_j} \\ &= \sum_j \mathsf{s}_j \sigma^{-1} \log \frac{\mathsf{s}_j \sigma^{-1}}{\mathsf{s}_j \mathsf{f}_j} \\ &= -\sigma^{-1} \sum_j \mathsf{s}_j \log \mathsf{f}_j - \log \sigma\,,\end{aligned}$$

whose minimizing is known to be equivalent to maximizing the so-called empirical likelihood in the sense of Owen [8]; see Hall and Presnell [4]. The salient feature of this example is that the function h penalized in the primal is "sparsity", the reciprocal of the estimated density f.

EXAMPLE. One can easily come to the idea to employ the popular and simple total variation distance in the minimum divergence formulation, choosing the dual objective to minimize

$$V(\mathsf{f}, \sigma^{-1}) = \sum_j s_j |\mathsf{f}_j - \sigma^{-1}|\,,$$

a formulation leading to a linear programming problem. However, the primal in this case would involve a function $\psi(x)$ equal to x/σ for x in the interval $[-1, 1]$, and $+\infty$ elsewhere. This indicates difficulties and likely explains the strange results we observed in our implementations.

Acknowledgement. We are indebted to Xuming He for valuable discussions, to Erling D. Andersen for creating MOSEK, a convex optimization toolbox for MATLAB, and to the anonymous referee for the careful reading of the manuscript.

REFERENCES

[1] BOYD, S.—VANDENBERGHE, L.: *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[2] GOOD, I. J.: *A nonparametric roughness penalty for probability densities*, Nature **229** (1971), 29–30.

[3] GU, C.: *Smoothing Spline ANOVA Models*. Springer-Verlag, New York, 2002.

[4] HALL, P.—PRESNELL, B.: *Density estimation under constraints*, J. Comput. Graph. Statist. **8** (1999), 259–277.

[5] KOENKER, R.—MIZERA, I.: *The alter egos of the regularized maximum likelihood density estimators: deregularized maximum-entropy, Shannon, Renyi, Simpson, Gini, and stretched strings*. In: Proceedings of 7th Prague Symposium on Asymptotic Statistics and 15th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes—Prague Stochastics '06 (M. Huškova, M. Janžura, eds.), Prague, August 21–25, 2006, Matfyzpress, Prague, 2006, pp. 145–157.

[6] KOENKER, R.—MIZERA, I.: *Density estimation by total variation regularization*. In: Advances in statistical modeling and inference, Essays in honor of Kjell A. Doksum (V. Nair, ed.), World Scientific, Singapore, 2006.

[7] LEONARD, T.: *Density estimation, stochastic processes and prior information*, J. Roy Stat. Soc. **40** (1978), 113–132.

[8] OWEN, A. B.: *Empirical Likelihood*. Chapman & Hall/CRC, Boca Raton, 2001.

[9] RÉNYI, A.: *On measures of entropy and information*. In: Proc. 4th Berkeley Symp. Math. Stat. Probab., Vol. 1: Contributions to the Theory of Statistics (J. Neyman, ed.), University of California June 20–July 30, 1960, University of California Press, Berkeley, 1961, pp. 547–561.

[10] ROCKAFELLAR, R. T.: *Convex Analysis*. Princeton University Press, Princeton, 1970.

[11] RUFIBACH, K.—DÜMBGEN, L.: *Maximum likelihood estimation of a log-concave density: basic properties and uniform consistency*, preprint arXiv:0709.0334.

[12] SILVERMAN, B. W.: *On the estimation of a probability density function by the maximum penalized likelihood method*, Ann. Statist. **10** (1982), 795–810.

Received October 25, 2006

Roger Koenker
University of Illinois at Urbana-Champaign
Departments of Economics and Statistics
Champaign, Illinois, 61620
USA

E-mail: rkoenker@uiuc.edu

Ivan Mizera
University of Alberta
Department of Mathematical and Statistical Sciences
CAB 632, Edmonton, Alberta, T6J 0Z2
CANADA

E-mail: mizera@stat.ualberta.ca