Tatra Mt. Math. Publ. **??** (????), 001–009

TATRA MOUNTAINS *Mathematical Publications*

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

Roger Koenker — Ivan Mizera

ABSTRACT. We investigate general schemes relevant for the estimation of a probability density via regularization—their primal and dual versions in the discretized setting. In particular, 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

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

where $\Psi(\mathsf{g})$ indicates the application of a real convex function ψ to the components of g, while $\mathsf{J}(\mathsf{h})$ is rather 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 positive nonzero elements; hereafter, $\succeq$ and $\preceq$ stand for componentwise inequalities.

In some cases, the primal formulation (P) can be simplified. If the function ψ is nondecreasing, then it immediately follows that (P) is equivalent to the unconstrained problem

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

As is customary in convex analysis, we consider convex functions that may attain $+\infty$ as a value; the set where such a function Φ is finite is called its *domain*, $\operatorname{dom}\Phi$. We assume that all convex functions appearing in (P) have domains with

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. We are indebted to Xuming He for valuable discussions and to Erling D. Andersen for creating MOSEK, a convex optimization toolbox for MATLAB.

nonempty interior. Concave functions are handled in an analogous manner, only the rôle of $+\infty$ is played by $-\infty$.

Important examples of convex functions include *indicators* of convex sets; an indicator of a set E is defined to be 0 for all $x \in E$ and $+\infty$ otherwise. A function *conjugate* to Φ is denoted by Φ^* and defined as

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

the latter formulation avoiding the need to compute with infinite values. The conjugate of the function $\lambda\|\cdot\|_p$, related to the ℓ^p norm $\|\cdot\|_p$, is the indicator of the set $\{x \colon \|x\|_q \leq \lambda\}$, a ball in the dual norm; here q is the (Hölder) conjugate of p, that is, $q = \infty$ for $p = 1$, and satisfies the relationship $(1/p) + (1/q) = 1$ for $p > 1$. Conversely, the conjugate of the indicator of the above ball is the multiple of the dual norm. The conjugate of the indicator of the cone $\{x \colon x \succeq \mathsf{0}\}$ is the indicator of the polar cone $\{x \colon x \preceq \mathsf{0}\}$. Finally, the function $(1/2)\|\cdot\|_2^2$ is conjugate to itself; and consequently $\lambda\|\cdot\|_2^2$ to $1/(4\lambda^2)\|\cdot\|_2^2$. Our references for convex analysis are Rockafellar [8], Boyd and Vanderberge [1].

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

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

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. We consider the estimated density to be represented by the vector f consisting 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 i-th row assigns 1 to a gridpoint equal to the i-th 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. In fact, 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 $\{\mathsf{u} \colon \|u\|_p \leq \Lambda\}$. 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 a indeed a probability density can be most conveniently verified through the dual formulation (D).

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

Proof. Since the nonnegativity constraint is directly included in the formulation of (D), it remains to verify that

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

as follows from the assumptions. □

In the simplest case, when matrix L is composed of zeros except for a single 1 in each row corresponding to a datapoint, $\mathsf{w}^{\mathsf{T}}\mathsf{L}$ gets these 1's multiplied by $1/n$, which further multiplying by $\mathsf{1}$ sums 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 *is from* I *and* f *from the image of* I *under* ψ'.

Proof. We start from a formulation equivalent to (P), obtained after rewriting it in terms of new variables u and v,

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

The Lagrange dual of (1) is

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

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

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

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,

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

Minimizing in g is done by expanding into components,

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

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 1 (Maximum Likelihood)**.** The primal formulation of this example was our primary motivation. Its continuous version can be traced back to Silverman [10] and Leonard [6]; in particular, Silverman [10] proposed to estimate the density via the maximum likelihood penalized scheme

$$-\int g\,dP_n + \int \mathrm{e}^g\,dx + \lambda\int \bigl(g^{(k)}\bigr)^2 dx = \min_g ! \tag{6}$$

penalizing the square of the third ($k = 3$) derivative of $g = \log f$, where P_n denotes the empirical probability supported by the datapoints. Other instances involve a possible use of the second derivative instead of third ($k = 2$), championed by Gu [3] and others; the use of the total variation penalty

$$\int |g^{(k)}| dx = \bigvee g^{(k-1)}$$

considered by Koenker and Mizera [4, 5] for $k = 1, 2, 3$; and estimation of a log-concave density studied by Rufibach and Dümbgen [9], which here corresponds to $k = 2$ and the penalty term in the form of the non-positivity constraint on the second derivative (with no tuning parameter λ involved).

In the discrete setting of (P), the k-th derivative operator is usually 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 in this example, for symmetric $\mathsf{J}(\mathsf{u}) = \lambda \|\mathsf{u}\|_p^p$ and $p = 1, 2$, is

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

where e^g has to be understood componentwise. An elementary calculation determines that 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 fact that $\operatorname{dom} \psi^* = [0, +\infty)$ independently enforces the nonnegativity constraint on f through the objective function, as a feasibility requirement. Silverman [10] 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 version, 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 objective function in the dual of (7),

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

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 then, for $p = 1$,

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

and for $p = 2$,

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

The dual of the penalty-constrained version,

$$-\mathsf{w}^\mathsf{T}\mathsf{Lg} + \mathsf{s}^\mathsf{T}\mathsf{e}^\mathsf{g} = \min_{\mathsf{g}}! \quad \text{subject to } \|\mathsf{Pg}\|_p \leq \Lambda, \tag{10}$$

of the primal (7), is (p and q being conjugate)

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

Finally, the dual of the shape-constrained formulation,

$$-\mathsf{w}^\mathsf{T}\mathsf{Lg} + \mathsf{s}^\mathsf{T}\Psi(\mathsf{g}) = \min_{\mathsf{g}}! \quad \text{subject to } \mathsf{Pg} \preceq 0 \tag{12}$$

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

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

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

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

where $s^{-1} = (\sum_j \mathsf{s}_j)^{-1}$ is the uniform probability mass function on the grid.

The dual formulation of the penalized likelihood problem as a maximum entropy problem offers a possibility of generalization by replacing the Shannon entropy term by, say, one from the system of Rényi entropies indexed by a parameter $\alpha > 0$; similarly to the Kullback-Leibler case, this entails the appropriate minimum divergence interpretations. Formally, Rényi's entropies include that of Example 1 for $\alpha = 1$; the Rényi entropy with exponent $\alpha \neq 1$ is defined as $(1-\alpha)^{-1} \log(\mathsf{s}^\mathsf{T}\mathsf{f}^\alpha)$, where f^α is hereafter interpreted componentwise. See Rényi [7]. The maximization of this function is equivalent to the maximization of $-\operatorname{sign}(\alpha - 1)\mathsf{s}^\mathsf{T}\mathsf{f}^\alpha$ or, equivalently, $-\operatorname{sign}(\alpha - 1)\mathsf{s}^\mathsf{T}\mathsf{f}^\alpha/\alpha$.

Let us denote by ψ_p a function equal to x^p/p for $x \geq 0$ and to 0 for $x < 0$. The conjugate of ψ_p for $p > 1$ is the function ψ_p^* equal to y^q/q for $y \geq 0$ (with 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 2 (Minimum χ^2)**.** This example is the special case of the Rényi formulation, with $\alpha = 2$; that is, $\psi(x) = \psi_2$. The conjugate of ψ_2 is equal to $y^2/2$ for $y \geq 0$. The dual in this example replaces 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 $\mathsf{s}^\mathsf{T}\mathsf{f}^2$ (after the elimination of the redundant constant in the objective). The corresponding primal is obtained by replacing $\sum_j \mathrm{e}^{\mathsf{g}_j}$ 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 function is equivalent to minimizing the χ^2-divergence

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

If instead of ψ_2 we consider $\psi(x)$ equal to $(1/2)x^2$ 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

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

In all these variants, both primal and dual estimate directly the density f, due to the fact that $\psi'(x) = x$.

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

$$-\mathsf{w}^\mathsf{T}\mathsf{L}\mathsf{h} + \tfrac{2}{3}\mathsf{s}^\mathsf{T}\mathsf{g}^{3/2} + \lambda \|\mathsf{P}\mathsf{h}\|_p^p = \min_{\mathsf{g}}! \qquad \text{subject to } \mathsf{h} \preceq \mathsf{g}.$$

Due to the fact that in any of these variants $\mathsf{f} = \mathsf{g}^2$, this example could be nicknamed “Silverman for Good”. Apart from the additional middle term, the objective functions differs from the original proposal of Good [2] also in the first term, which is not based on the logarithm of the square root of the estimated density, but instead directly 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 terminology.

Example 4 (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 function is $\psi^*(y) = -2\sqrt{y}$, for $y \geq 0$, and ∞ elsewhere. The dual (for $p = 1$), in the minimization form and after the elimination of the redundant constant, puts $-\mathsf{s}^{\mathsf{T}}\sqrt{\mathsf{f}}$, where $\sqrt{\mathsf{f}}$ is again applied componentwise, into the objective of (8), (9), (11), and (13). The minimization of the dual objective is equivalent to the minimization of the Hellinger distance

$$\mathcal{H}(\mathsf{f}, s^{-1}) = \sum_j s_j\big(\sqrt{\mathsf{f}_j} - \sqrt{1/s}\big)^2 = \sum_j \big(\sqrt{\mathsf{s}_j\mathsf{f}_j} - \sqrt{\mathsf{s}_j/s}\big)^2 = 2 - 2\sqrt{s}\sum s_j\sqrt{\mathsf{f}_j}.$$

The primal can be cast—because ψ is nondecreasing—in the unconstrained version (U), just replacing the $\sum_j \mathrm{e}^{\mathsf{g}_j}$ term in (7), (10), or (12) by $-\mathsf{s}^{\mathsf{T}}\mathsf{h}^{-1}$, where h^{-1} is the componentwise reciprocal value of h; however, we have to include the domain restriction for ψ as a feasibility constraint. The resulting primal formulation in case of (7) is

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

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

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

The estimated density f is the reciprocal of g^2, hence g could be called, in the Tukey spirit, a “rootosparsity”; and consequently h, being negative, a “hanging rootosparsity”.

In implementations, we observed that the numerical perfomance may be improved by adding the (theoretically redundant) nonnegativity constraint $\mathsf{f} \succeq 0$ also in the primal formulation. This is, however, rather an unimportant detail, because dual formulations always ran significantly faster and were more numerically stable than their primal counterparts. The utility of the latter is rather theoretical—as a guidance in more complex formulations, where density estimation is merely a building block. It seems that the results did not substantially differ for different Rényi exponents; thus, if maximum likelihood formulation turns out to be numerically infeasible, there can be a viable Rényi alternative.

Example 5. The choice $\psi(x) = -1/2 - \log(-x)$ for $x < 0$, and $+\infty$ otherwise, can be viewed as a limiting case of the Rényi system for $\alpha = 0$. It is similar in spirit and shape to Example 4; however, we are unaware about any minimum distance interpretation. After elimination of the redundant constants, 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 $\mathsf{g} = -\mathsf{h}$ gives

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

The salient feature of this example is that the function g penalized in the primal is "sparsity", the reciprocal of the estimated density f.

Example 6. One can easily come to the idea to employ the popular and simple total variation distance in the minimum divergence formulation; in such a case, the dual objective would be chosen to minimize

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

and the appropriate version of (8) would have a computationally appealing form of a linear programming problem. However, the primal in this case would involve a function $\psi(x)$ equal to x/s for x in the interval $[-1, 1]$, and $+\infty$ elsewhere. This indicates difficulties and likely explains the strange results we observed when we implemented this formulation.

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] 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: *Prague Stochastics 2006, Proceedings of the joint session of 7th Prague Symposium on Asymptotic Statistics and 15th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, held in Prague from August 21 to 25, 2006*, M. Huskova and M. Janzura, eds., 145–157. Matfyzpress, Prague, 2006.

[5] 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.

[6] LEONARD, T.: *Density estimation, stochastic processes and prior information*, *Journal of the Royal Statistical Society, Series B (Methodological)* **40** (1978), 113–132.

[7] RÉNYI, A.: *On measures of entropy and information.* In: *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, held at the Statistical Laboratory, University of California, June 20–July 30, 1960, Volume 1: Contributions to the Theory of Statistics*, J. Neyman, ed., University of California Press, Berkeley, 1961.

[8] ROCKAFELLAR, R. T.: *Convex analysis*, Princeton University Press, Princeton, 1970.

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

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

University of Illinois at Urbana-Champaign
Departments of Economics and Statistics
Champaign, Illinois, 61620
USA
E-mail: rkoenker@uiuc.edu

University of Alberta
Department of Mathematical and Statistical Sciences
CAB 632, Edmonton, Alberta, T6J 0Z2
Canada
E-mail: mizera@stat.ualberta.ca