Pure and Applied Functional Analysis
Volume 5, Number 3, 2020, 505–531

Yokohama Publishers
ISSN 2189-3764 ONLINE JOURNAL

PROXIMAL AVERAGES FOR MINIMIZATION OF ENTROPY FUNCTIONALS

HEINZ H. BAUSCHKE AND SCOTT B. LINDSTROM

Abstract. In their 2016 article "Meetings with Lambert $\mathcal{W}$ and Other Special Functions in Optimization and Analysis", Borwein and Lindstrom considered the minimization of an entropy functional utilizing the weighted average of the negative Boltzmann–Shannon entropy $x \mapsto x \log x - x$ and the energy $x \mapsto x^2/2$; the solution employed the Fenchel-Moreau conjugate of the sum. However, in place of a traditional arithmetic average of two convex functions, it is also, perhaps even more, natural to use their *proximal* average. We explain the advantages and illustrate them by computing the analogous proximal averages for the negative entropy and energy. We use the proximal averages to solve entropy functional minimization problems similar to those considered by Borwein and Lindstrom, illustrating the benefits of using a true homotopy. Through experimentation, we discover computational barriers to obtaining solutions when computing with proximal averages, and we demonstrate a method that appears to remedy them.

1. Introduction

Computer-assisted discovery has changed the way in which research is conducted, both in optimization and elsewhere. As noted by Borwein and Lindstrom [9] in 2016:

> In the current mathematical world, it matters less what you know about a given function than whether your computer package of choice (say *Maple*, *Mathematica* or *SAGE*) or online source, say Wikipedia [19] does.

They considered, in particular, that many occurrences of the Lambert $\mathcal{W}$ function in convex analysis may be naturally discovered with the use of the *Symbolic Convex Analysis Tools (SCAT)* package for *Maple* [17, 6], notwithstanding one's possible naivety of special functions. The *SCAT* package for *Maple* was created by Borwein and Chris Hamilton; it grew out of Bauschke and von Mohrenschildt's *Maple* package *fenchel*, which is described in their 2006 article [5]. *SCAT* continues to be developed by D.R. Luke and others. F. Lauster, D.R. Luke, and M.K. Tam have recently illuminated symbolic computation with monotone operators [14].

We will continue the work of Borwein and Lindstrom, [9], by considering *proximal* averages where they originally considered *weighted* averages: for the minimization of entropy functionals. In human-machine collaboration with our computer algebra system (CAS) of choice, *Maple*, we discover closed forms of proximal averages for the

Key words and phrases. convex conjugate, convex optimization, entropy optimization, Fenchel duality, Fenchel-Moreau-Rockafellar-conjugate, Lambert W function, proximal average, special functions, subdifferential, symbolic convex analysis tools.

Boltzmann–Shannon entropy and energy for specific parameters before conjecturing and eventually proving their general form with the computer assistance. Armed with closed forms, we will consider how the minimization of an entropy function changes when the weighted average is replaced with a true homotopy.

The structure of this paper is as follows. In Subsection 1.1, we recall the properties of the Lambert $\mathcal{W}$ function that will prove instrumental in our analysis, and in Subsection 1.2 we recall preliminaries on convex analysis. In Section 2, we recall the basic properties of the proximal average. In Section 3 we consider proximal averages which employ $\mathcal{W}$, first the energy and Boltzmann–Shannon entropy in Subsection 3.1, and then the energy with the exponential in Subsection 3.2; the two are related, importantly, through duality. In Section 4, we introduce the problem of minimizing an entropy functional subject to linear constraints, and in Subsection 4.1 we provide examples. We conclude in Section 5.

1.1. **Lambert $\mathcal{W}$ preliminaries.** Of particular interest to us is the Lambert $\mathcal{W}$ function, which we take to be the real analytic inverse of $x \mapsto xe^x$. The real inverse is two-valued, and, for the sake of our exposition, we consider $\mathcal{W}$ to refer always to the principal branch, shown in Figure 1.2. We will make use of the following elementary identities.

Proposition 1.1. *For any y in the appropriate respective domains, the following identities hold:*

(i) $\mathcal{W}(y)e^{\mathcal{W}(y)} = y$;
(ii) $e^{\mathcal{W}(y)} = \frac{y}{\mathcal{W}(y)}$;
(iii) $\mathcal{W}(y) = \log\left(\frac{y}{\mathcal{W}(y)}\right)$;
(iv) $\log(\mathcal{W}(y)) = \log(y) - \mathcal{W}(y)$.
(v) $\log(\mathcal{W}(e^y)) = y - \mathcal{W}(e^y)$.

Proof. (i): This is true from the fact that $\mathcal{W}$ is the inverse of $x \mapsto xe^x$.
(ii): Divide both sides of 1.1 by $\mathcal{W}(y)$.
(iii): Take the log of both sides of 1.1.
(iv): Since $\log\left(\frac{y}{\mathcal{W}(y)}\right) = \log(y) - \log(\mathcal{W}(y))$, this follows from 1.1.
(v): Apply 1.1, substituting e^y for y. □

An excellent overview of the methods used for symbolic differentiation and antidifferentiation — and their history — is given by R.M. Corless, G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey, and D.E. Knuth [13]. We have, in particular, the following characterization of the derivatives and antiderivative.

Proposition 1.2. *The derivative of $\mathcal{W}$ is given by*

$$\begin{aligned}\mathcal{W}'(x) &= \frac{1}{(1+\mathcal{W}(x))\exp(\mathcal{W}(x))}\\ &= \frac{\mathcal{W}(x)}{x(1+\mathcal{W}(x))}, \quad \text{if } x \neq 0.\end{aligned}$$

Moreover, the nth derivative of $\mathcal{W}$ may be characterized as

$$\frac{d^n\mathcal{W}(x)}{dx^n} = \frac{e^{-n\mathcal{W}(x)}p_n(\mathcal{W}(x))}{(1+\mathcal{W}(x))^{2n-1}} \quad \text{for } n \geq 1.$$

where $p_n(w)$ are polynomials that satisfy the recurrence relation given by

$$p_{n+1}(w) = -(nw+3n-1)\,p_n(w) + (1+w)p_n'(w), \quad \text{for } n \geq 1.$$

For details, see, for example, [13, Section 3].

Proposition 1.3. *The antiderivative of $\mathcal{W}$ may be characterized as*

$$\begin{aligned}\int \mathcal{W}(x)dx &= \left(\mathcal{W}(x)^2 - \mathcal{W}(x) + 1\right)e^{\mathcal{W}(x)} + C \\ &= x\left(\mathcal{W}(x) - 1 + 1/\mathcal{W}(x)\right) + C.\end{aligned}$$

For details, see, for example, [13, Section 3].

Using Proposition 1.2, we also have the following.

Proposition 1.4. *The following hold:*

(i) $\frac{d}{dx}\mathcal{W}(e^x) = \frac{\mathcal{W}(e^x)}{1+\mathcal{W}(e^x)}$;
(ii) $\frac{d}{dx}\left(\mathcal{W}(e^x) + \frac{1}{2}\mathcal{W}(e^x)^2\right) = \mathcal{W}(e^x)$;
(iii) $\frac{d}{dx}e^{\mathcal{W}(x)} = \frac{1}{1+\mathcal{W}(x)}$.

Proof. (i): Apply the chain rule along with the identity from Proposition 1.2 to differentiate $\mathcal{W}(e^x)$.

(ii): Apply the chain rule along with the identity from Proposition 1.2 to differentiate $\left(\mathcal{W}(e^x) + \frac{1}{2}\mathcal{W}(e^x)^2\right)$.

(iii): Apply the chain rule along with the identity from Proposition 1.2 to differentiate $e^{\mathcal{W}(x)}$. □

1.2. **Preliminaries on Convex Analysis.** Throughout, X is a Hilbert space.

Definition 1.5. As in [3], we will work with the following set of functions:

$$\mathcal{F} := \{f : X \to]-\infty,\infty] \mid f \text{ is convex, lower semicontinuous, and proper}\}.$$

Definition 1.6 (Fenchel Conjugate)**.** The Fenchel conjugate f^* of a function $f : X \to [-\infty,\infty]$ is defined as follows:

$$\begin{aligned}f^* : X^* &\to [-\infty,\infty] \\ f^* : x &\mapsto \sup_{y\in X}\{\langle x,y\rangle - f(y)\}.\end{aligned}$$

This is also often referred to as a *convex conjugate* or *Fenchel-Moreau conjugate.*

The function f^* is always convex, i.e. its epigraph is convex. Moreover, we have the following.

Proposition 1.7. [1, Proposition 16.4] *Let $f \in \mathcal{F}$ and $x \in \operatorname{dom}\partial f$. Then $f^{**} = f$ and $\partial f^{**}(x) = \partial f(x)$.*

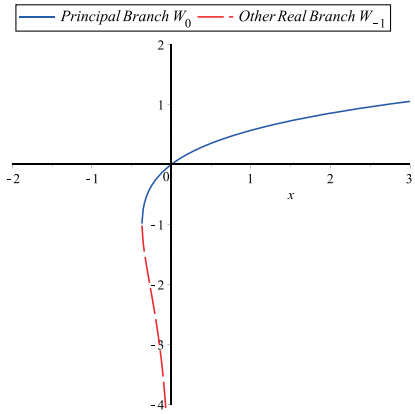

FIGURE 1. The two real branches of Lambert $\mathcal{W}$.

Definition 1.8 (Argmin operator p for Fenchel conjugates)**.** Let f be a proper convex function and f^* its conjugate. We define p_f to be a selection operator satisfying

$$\mathrm{p}_f(x) \in \operatorname*{argmin}_{y\in X} \{\langle x, y\rangle - f(y)\}. \tag{1.1}$$

so that we may express the closed form for f^* as

$$f^*(x) = \sup_{y\in X}\{\langle x, y\rangle - f(y)\} = \langle x, \mathrm{p}_f(x)\rangle - f(\mathrm{p}_f(x)). \tag{1.2}$$

2. Proximal averages

The systematic investigation of the proximal average started in 2008 [3], relying crucially on an important result of Bauschke, E. Matoušková, and S. Reich [4, Theorem 6.1]. Research on the topic continues to grow. One noteworthy recent application is Y.L. Yu's 2013 employment of the proximal average to analyse a novel proximal gradient algorithm [18].

Definition 2.1 (Proximal Average)**.** The proximal average operator is

$$\mathcal{P} : \mathcal{F} \times [0,1] \times \mathcal{F} \to \{f | f : X \to [-\infty, +\infty]\}$$
$$(f_0, \lambda, f_1) \mapsto \left((1-\lambda)\left(f_0 + \tfrac{1}{2}\|\cdot\|^2\right)^* + \lambda\left(f_1 + \tfrac{1}{2}\|\cdot\|^2\right)^*\right)^* - \tfrac{1}{2}\|\cdot\|^2.$$

See, for example, [3, Definition 4.1].

Remark 2.2 (Symmetric and convex properties of proximal averages)**.** Let $f_0, f_1 \in \mathcal{F}$ and $\lambda \in [0,1]$. Then we have that

$$\mathcal{P}(f_0, 0, f_1) = f_0, \quad \mathcal{P}(f_0, 1, f_1) = f_1, \quad \text{and } \mathcal{P}(f_0, \lambda, f_1) = \mathcal{P}(f_1, 1-\lambda, f_0). \tag{2.1}$$

We also have that $\mathcal{P}(f_0, \lambda, f_1)$ is convex. See, for example, [3, Proposition 4.2].

Remark 2.3 (Conjugacy of proximal averages)**.** When $f_0, f_1 \in \mathcal{F}$ and $\lambda \in [0,1]$ we have that

$$(\mathcal{P}(f_0, \lambda, f_1))^* = \mathcal{P}(f_0^*, \lambda, f_1^*). \tag{2.2}$$

See, for example, [3, Theorem 4.3] or [4, Theorem 6.1].

Definition 2.4 (Simplified notation for proximal averages)**.** We will follow a convenient convention from [3]. Let $f_0, f_1 \in \mathcal{F}$ and $\lambda \in [0,1]$. Let

$$f_\lambda := \mathcal{P}(f_0, \lambda, f_1) \quad \text{and} \quad f_\lambda^* := \mathcal{P}(f_0^*, \lambda, f_1^*).$$

From Remark 2.3 we have that $(f_\lambda)^* = (f^*)_\lambda$, which shows that f_λ^* is not ambiguous.

Definition 2.5 (epi-convergence and epi-topology)**.** Let f and $(f_n)_{n\in\mathbb{N}}$ be functions from X to $]-\infty,+\infty]$. Then $(f_n)_{n\in\mathbb{N}}$ epi-converges to f if for every $x \in X$ the following hold.

(i) For every sequence $(x_n)_{n\in\mathbb{N}}$ in X converging to x, one has $f(x) \leq \liminf f_n(x_n)$.
(ii) There exists a sequence $(y_n)_{n\in\mathbb{N}}$ in X converging to x such that $\limsup f_n(y_n) \leq f(x)$.

In this case we write $f_n \xrightarrow{\text{e}} f$. The *epi-topology* is the topology induced by epi-convergence. See, for example, [3, Definition 5.1]. For greater detail, see [16].

Remark 2.6 (Continuity of $\mathcal{P}$)**.** Suppose that $\mathcal{F}$ is equipped with the epi-topology. Then the proximal average operator $\mathcal{P} : \mathcal{F} \times [0,1] \times \mathcal{F} \to \mathcal{F}$ is continuous. In other words, where $(f_n)_{n\in\mathbb{N}}, (g_n)_{n\in\mathbb{N}}$ are sequences in $\mathcal{F}$ and $(\lambda_n)_{n\in\mathbb{N}}$ is a sequence in $[0,1]$ such that $f_n \xrightarrow{\text{e}} f, g_n \xrightarrow{\text{e}} g$, and $\lambda_n \to \lambda$, then we have that:

$$\mathcal{P}(f_n, \lambda_n, g_n) \xrightarrow{\text{e}} \mathcal{P}(f, \lambda, g) \quad \text{as} \quad n \to \infty. \tag{2.3}$$

For a proof, see, for example, [3, Theorem 5.4].

3. Proximal averages employing Lambert $\mathcal{W}$

Definition 3.1. We define the negative Boltzmann–Shannon entropy as follows:

$$\text{ent} : \mathbb{R} \to \mathbb{R} \cup \{\infty\} : \quad x \mapsto \begin{cases} x \log x - x & x \in]0, \infty]; \\ 0 & x = 0; \\ \infty & \text{otherwise.} \end{cases} \tag{3.1}$$

In [9] the authors considered the average (*not* the proximal average) given by

$$f_t(x) = (1-t)\text{ent}(x) + t\frac{x^2}{2} \tag{3.2}$$

for $0 \leq t \leq 1$ so that f_0 is the Boltzmann–Shannon entropy and f_1 is the energy. For clarity, we will refer to such an average as a *weighted* average, in order to distinguish it from the *proximal* average, and we will consistently use t for the former and λ for the latter.

Borwein and Lindstrom then obtained the conjugate as follows:

$$f_t^*(y) = \frac{(1-t)^2}{2t}\left(\mathcal{W}\left(\frac{t}{1-t}e^{\frac{y}{1-t}}\right) + 2\right)\mathcal{W}\left(\frac{t}{1-t}e^{\frac{y}{1-t}}\right). \tag{3.3}$$

Remark 3.2 (Limiting Cases for Weighted Average)**.** In (3.2), if one considers the *limit* for f_t as $t \to 1$ we obtain the *positive energy*, which is infinite at negative points. In the limit as $t \to 0$ we recover $\text{ent}(x)$. For its conjugate in f_t^* in (3.3), if

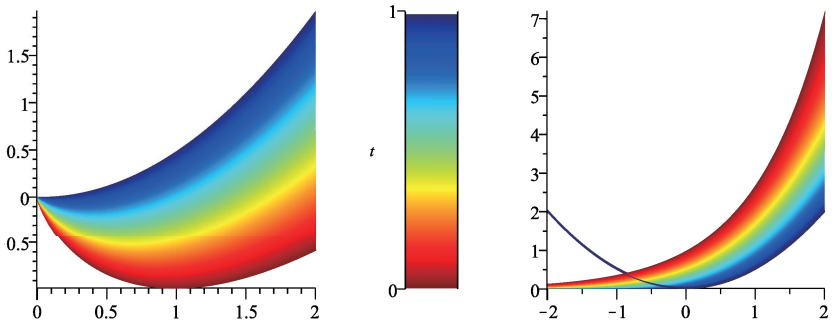

FIGURE 2. f_t from (3.2) (left) and f_t^* from (3.3) (right).

one considers the limit as $t \to 0$ we recover $\exp(x)$ which is the conjugate of $\operatorname{ent}(x)$. In the limit as $t \to 1$ we obtain

$$x \mapsto \begin{cases} \frac{x^2}{2} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

We would expect this, given that $\frac{(\cdot)^2}{2}$ is self-conjugate while $\operatorname{ent}(x)$ is infinite for $x < 0$. Notice, however, that $f_1^* = \frac{1}{2}|\cdot|^2$, and so we do not reobtain f_1^* in the limiting case as $t \to 1$.

Both f_t and f_t^* may be seen in Figure 2.

However, instead of (3.2), it is more natural to consider $f_\lambda = \mathcal{P}(\operatorname{ent}, \lambda, \frac{(\cdot)^2}{2})$. We will compute f_λ and its conjugate f_λ^*, which is the natural analogue to (3.3).

3.1. **Form and proof for f_λ.** Throughout the following $\lambda \in]0,1[$. In each case where we compute a conjugate, we will be seeking the argmax of a concave function that attains its maximum, and so it suffices to find a critical point.

Lemma 3.3. *Let $f_0 := \operatorname{ent}$ and $f_1 := \frac{1}{2}(\cdot)^2$. Then we have the following:*

(i) $\left(f_0 + \frac{1}{2}(\cdot)^2\right)^* = \frac{1}{2}\mathcal{W}(e^x)\,(\mathcal{W}(e^x) + 2)$

(ii) $\left(f_1 + \frac{1}{2}(\cdot)^2\right)^* = \frac{1}{4}(\cdot)^2$.

Proof. (i): By definition,

$$\left(f_0 + \frac{1}{2}(\cdot)^2\right)^*(x) = \sup_{y \in \mathbb{R}} \left\{ xy - f_0(y) - \frac{1}{2}y^2 \right\}.$$

Differentiating the inner term with respect to y and setting equal to zero, we have that the supremum is obtained when y satisfies $x - \log(y) - y = 0$. Solving for y, we obtain

$$\begin{aligned} \log(y) &= x - y, \\ \text{and so} \quad y &= \frac{e^x}{e^y}, \\ \text{which simplifies to} \quad ye^y &= e^x, \end{aligned}$$

and so $y = \mathcal{W}(e^x)$. Substituting this value back into $xy - f_0(y) - \frac{1}{2}y^2$, we have that

$$\left(f_0 + \frac{1}{2}(\cdot)^2\right)^*(x) = x\mathcal{W}(e^x) - \mathcal{W}(e^x)\log\left(\mathcal{W}(e^x)\right) + \mathcal{W}(e^x) - \frac{1}{2}\mathcal{W}(e^x)^2.$$

Factoring and employing the fact that $(\log \circ \mathcal{W})(z) = \log(z) - \mathcal{W}(z)$ we obtain

$$\left(f_0 + \frac{1}{2}(\cdot)^2\right)^*(x) = \frac{1}{2}\mathcal{W}(e^x)\left(2x + \mathcal{W}(e^x) - 2\log(e^x) + 2\right).$$

Because x is real, the right-hand side further simplifies to $\frac{1}{2}\mathcal{W}(e^x)\left(\mathcal{W}(e^x) + 2\right)$, completing the proof of 3.3.

(ii): This is a well-known result and may be obtained by simple arithmetic. □

Note that we may recognize the term $\frac{1}{2}\mathcal{W}(e^x)(\mathcal{W}(e^x) + 2)$ as an antiderivative of $\mathcal{W}(e^x)$ (see Proposition 1.4), a fact we will exploit in the following lemma.

Lemma 3.4. *Let φ be defined as follows*

$$\varphi := \left((1-\lambda)\left(\frac{1}{2}\mathcal{W}(e^{(\cdot)})(\mathcal{W}(e^{(\cdot)}) + 2)\right) + \lambda\left(\frac{1}{4}(\cdot)^2\right)\right)^*.$$

Then it holds that

$$\mathrm{p}_\varphi(x) = -\frac{(\frac{2}{\lambda} - 2)\mathcal{W}\left(\left(\frac{2}{\lambda} - 1\right)e^{\frac{2x}{\lambda}}\right)}{\frac{2}{\lambda} - 1} + \frac{2x}{\lambda} \tag{3.4}$$

so that we may explicitly write

$$\varphi(x) = x\mathrm{p}_\varphi(x) - (1-\lambda)\left(\frac{1}{2}\mathcal{W}(e^{\mathrm{p}_\varphi(x)})(\mathcal{W}(e^{\mathrm{p}_\varphi(x)}) + 2)\right) - \frac{\lambda}{4}\mathrm{p}_\varphi(x)^2.$$

Proof. By definition,

$$\varphi(x) = \sup_{y\in\mathbb{R}}\left\{xy - (1-\lambda)\left(\frac{1}{2}\mathcal{W}(e^y)(\mathcal{W}(e^y) + 2)\right) - \frac{\lambda}{4}y^2\right\}.$$

Differentiating the inner term with respect to y and setting equal to zero, we obtain

$$x - (1-\lambda)\mathcal{W}(e^y) - \frac{\lambda}{2}y = 0. \tag{3.5}$$

We will show that (3.5) is true if $y = \mathrm{p}_\varphi(x)$. First we will rewrite (3.5) using the fact that $\mathcal{W}(a) = b$ if and only if $be^b = a$, which allows us to remove the $\mathcal{W}(e^y)$ term as follows:

$$\begin{aligned}
\mathcal{W}(e^y) &= \frac{x - \frac{\lambda}{2}y}{1-\lambda},\\
\left(\frac{x - \frac{\lambda}{2}y}{1-\lambda}\right)e^{\left(\frac{x-\frac{\lambda}{2}y}{1-\lambda}\right)} &= e^y,\\
\left(x - \frac{\lambda}{2}y\right)e^{\frac{y\lambda-2x}{2(\lambda-1)}} &= (1-\lambda)e^y.
\end{aligned}$$

This is equivalent to the form returned by *Maple*,

$$e^{\frac{y\lambda-2x}{2(\lambda-1)}}y\lambda - 2e^y\lambda - 2xe^{\frac{y\lambda-2x}{2(\lambda-1)}} + 2e^y = 0, \tag{3.6}$$

and so again naivety need not inhibit the discovery. We will use *Maple*'s form. We need only to show that

$$e^{\frac{\mathrm{p}_\varphi(x)\lambda-2x}{2(\lambda-1)}}\mathrm{p}(x)\lambda-2e^{\mathrm{p}_\varphi(x)}\lambda-2xe^{\frac{\mathrm{p}_\varphi(x)\lambda-2x}{2(\lambda-1)}}+2e^{\mathrm{p}_\varphi(x)}=0. \tag{3.7}$$

First consider the term $e^{\mathrm{p}_\varphi(x)}$. Since for any a,b,z we have that

$$e^{aW(z)+b}=(e^{W(z)})^a e^b=\left(\frac{z}{W(z)}\right)^a e^b,$$

we may let

$$a:=-\left(\frac{\frac{2}{\lambda}-2}{\frac{2}{\lambda}-1}\right),\quad b:=\frac{2x}{\lambda},\quad z:=\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}, \tag{3.8}$$

and thusly rewrite

$$e^{\mathrm{p}_\varphi(x)}=\left(\frac{\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}}{\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)}\right)^{-\frac{\frac{2}{\lambda}-2}{\frac{2}{\lambda}-1}}e^{\frac{2x}{\lambda}}. \tag{3.9}$$

Next consider the term $e^{\frac{\mathrm{p}_\varphi(x)\lambda-2x}{2(\lambda-1)}}$. Using (3.9), we may rewrite it thusly:

$$\begin{aligned}
e^{\frac{\mathrm{p}_\varphi(x)\lambda-2x}{2(\lambda-1)}}&=e^{\frac{-2x}{2(\lambda-1)}}\left(e^{\mathrm{p}_\varphi(x)}\right)^{\frac{\lambda}{2(\lambda-1)}}\\
&=e^{\frac{-2x}{2(\lambda-1)}}\left(\left(\frac{\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}}{\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)}\right)^{-\frac{\frac{2}{\lambda}-2}{\frac{2}{\lambda}-1}}e^{\frac{2x}{\lambda}}\right)^{\frac{\lambda}{2(\lambda-1)}}\\
&=e^{\frac{-2x}{2(\lambda-1)}}\left(\frac{\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}}{\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)}\right)^{-\frac{\lambda}{\lambda-2}}e^{\frac{2x}{2(\lambda-1)}}\\
&=\left(\frac{\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}}{\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)}\right)^{-\frac{\lambda}{\lambda-2}}=\frac{\left(\frac{2}{\lambda}-1\right)^{-\frac{\lambda}{\lambda-2}}e^{-\frac{2x}{\lambda-2}}}{\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)^{-\frac{\lambda}{\lambda-2}}}.
\end{aligned} \tag{3.10}$$

Next consider the term $e^{\frac{\mathrm{p}_\varphi(x)\lambda-2x}{2(\lambda-1)}}\mathrm{p}_\varphi(x)\lambda$. Using (3.4) and (3.10), we may rewrite it as follows:

$$\begin{aligned}
e^{\frac{\mathrm{p}_\varphi(x)\lambda-2x}{2(\lambda-1)}}\mathrm{p}_\varphi(x)\lambda&=\frac{\left(\frac{2}{\lambda}-1\right)^{\frac{-\lambda}{\lambda-2}}e^{\frac{-2x}{\lambda-2}}}{\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)^{\frac{-\lambda}{\lambda-2}}}\left(-\frac{(\frac{2}{\lambda}-2)\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)}{\frac{2}{\lambda}-1}+\frac{2x}{\lambda}\right)\lambda\\
&=-\left(\frac{2}{\lambda}-1\right)^{\frac{-\lambda}{\lambda-2}-1}\left(\frac{2}{\lambda}-2\right)e^{-\frac{2x}{\lambda-2}}\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)^{\left(1+\frac{\lambda}{\lambda-2}\right)}\lambda
\end{aligned}$$

$$+\frac{2x\left(\frac{2}{\lambda}-1\right)^{-\frac{\lambda}{\lambda-2}}e^{-\frac{2x}{\lambda-2}}}{\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)^{-\frac{\lambda}{\lambda-2}}}. \tag{3.11}$$

Using (3.9), (3.10), and (3.11), we may have that the statement (3.7) — which we want to show — is equivalent to:

$$\begin{aligned}0 =& -\left(\frac{2}{\lambda}-1\right)^{\frac{-\lambda}{\lambda-2}-1}\left(\frac{2}{\lambda}-2\right)e^{-\frac{2x}{\lambda-2}}\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)^{\left(1+\frac{\lambda}{\lambda-2}\right)}\lambda\\ &+\frac{2x\left(\frac{2}{\lambda}-1\right)^{-\frac{\lambda}{\lambda-2}}e^{-\frac{2x}{\lambda-2}}}{\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)^{-\frac{\lambda}{\lambda-2}}}+2(1-\lambda)\left(\frac{\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}}{\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)}\right)^{-\frac{\frac{2}{\lambda}-2}{\frac{2}{\lambda}-1}}e^{\frac{2x}{\lambda}}\\ &-2x\frac{\left(\frac{2}{\lambda}-1\right)^{-\frac{\lambda}{\lambda-2}}e^{-\frac{2x}{\lambda-2}}}{\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)^{-\frac{\lambda}{\lambda-2}}}.\end{aligned}$$

Now the positive and negative terms of the form

$$2x\frac{\left(\frac{2}{\lambda}-1\right)^{-\frac{\lambda}{\lambda-2}}e^{-\frac{2x}{\lambda-2}}}{\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)^{-\frac{\lambda}{\lambda-2}}}$$

cancel each other out, leaving us with

$$\begin{aligned}0 =& -\left(\frac{2}{\lambda}-1\right)^{\frac{-\lambda}{\lambda-2}-1}\left(\frac{2}{\lambda}-2\right)e^{-\frac{2x}{\lambda-2}}\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)^{\left(1+\frac{\lambda}{\lambda-2}\right)}\lambda\\ &+2(1-\lambda)\left(\frac{\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}}{\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)}\right)^{-\frac{\frac{2}{\lambda}-2}{\frac{2}{\lambda}-1}}e^{\frac{2x}{\lambda}}.\end{aligned}$$

Rewriting and simplifying, we obtain

$$\begin{aligned}0 =& -\left(\frac{2}{\lambda}-1\right)^{\frac{2(1-\lambda)}{\lambda-2}}2\left(1-\lambda\right)e^{-\frac{2x}{\lambda-2}}\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)^{\left(1+\frac{\lambda}{\lambda-2}\right)}\\ &+2(1-\lambda)\left(\frac{2}{\lambda}-1\right)^{\frac{2(1-\lambda)}{\lambda-2}}e^{\frac{-2x}{\lambda-2}}\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)^{1+\frac{\lambda}{\lambda-2}},\end{aligned}$$

which is true, completing the result. □

Theorem 3.5. *Let f_0, f_1 be defined as in Lemma 3.3 and let* p *be defined as in Lemma 3.4. Then*

$$f_\lambda(x) = \frac{\lambda-1}{2}\mathcal{W}\left(e^{\left(\frac{2\lambda-2}{2-\lambda}\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)+\frac{2x}{\lambda}\right)}\right)^2-\frac{x^2(\lambda-2)}{2\lambda} \tag{3.12}$$

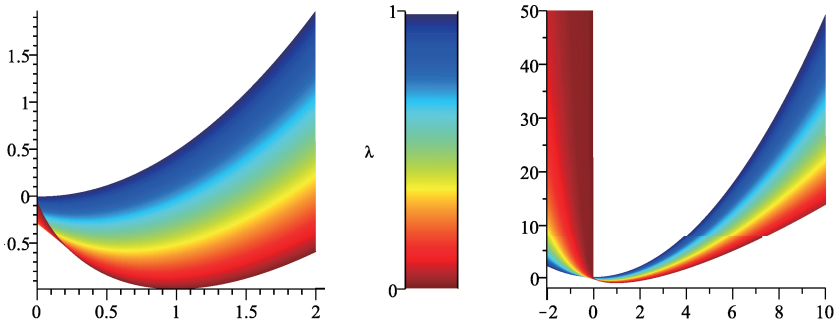

FIGURE 3. f_λ from Theorem 3.5

$$+(\lambda-1)\mathcal{W}\left(e^{\left(\frac{2\lambda-2}{2-\lambda}\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)+\frac{2x}{\lambda}\right)}\right)-\frac{\lambda(\lambda-1)^2}{(\lambda-2)^2}\mathcal{W}\left(\left(\frac{2}{\lambda}-1\right)e^{\frac{2x}{\lambda}}\right)^2.$$

Proof. Using Definition 2.1 together with Lemma 3.3 we have that

$$f_\lambda=\left((1-\lambda)\left(\frac{1}{2}\mathcal{W}(e^{(\cdot)})(\mathcal{W}(e^{(\cdot)})+2)\right)+\lambda\left(\frac{1}{4}(\cdot)^2\right)\right)^*-\frac{1}{2}(\cdot)^2.$$

This is just

$$f_\lambda=\varphi-\frac{1}{2}(\cdot)^2$$

where $\varphi, \mathrm{p}_\varphi$ are as defined as in Lemma 3.4. From this, we have that

$$f_\lambda(x)=x\mathrm{p}_\varphi(x)-(1-\lambda)\left(\frac{1}{2}\mathcal{W}(e^{\mathrm{p}_\varphi(x)})(\mathcal{W}(e^{\mathrm{p}_\varphi(x)})+2)\right)-\frac{\lambda}{4}\mathrm{p}_\varphi(x)^2-\frac{1}{2}x^2,$$

which simplifies, by a great deal of arithmetic, to the form we see in (3.12), completing the result. □

Worthy of note is that this result (in particular, Lemma 3.4) could not be computed by the *SCAT*, nor could *Maple* find the root of (3.6) on its own. The solution was discovered by choosing specific values for λ, solving (3.6), observing, and finally deducing the more general pattern. This serves as an example of the kind of fruitful human-machine collaboration Borwein & Lindstrom sought to emphasize in [9].

Within minutes of choosing correctly we "knew" the answer, because we could visually read off the functions f_0 and f_1 at left in Figure 3, even though a proof took much longer.

3.2. **Form and Proof for f_λ^*.** While the complicated nature of f_λ precludes computing its conjugate in the usual way, we can still compute it using the convenient identity (2.2) found in Remark 2.3. Specifically, since $f_\lambda^*=\mathcal{P}(f_0^*,\lambda,f_1^*)$, we can forget, for the moment, about f_λ and instead compute $\mathcal{P}(f_0^*,\lambda,f_1^*)$ in the same way that we computed f_λ, directly from Definition 2.1.

Remark 3.6. Let $f_0=\mathrm{ent}$ and $f_1=\frac{1}{2}|\cdot|^2$. Then

$$f_0^*=\exp\quad\text{and}\quad f_1^*=\frac{1}{2}|\cdot|^2=f_1.$$

These are both well-known results.

Lemma 3.7. *Let* $f_0 := \text{ent}$ *and* $f_1 := \frac{1}{2}|\cdot|^2$. *Then we have the following*

(i) $\left(f_0^* + \frac{1}{2}|\cdot|^2\right)^* = \frac{1}{2}|\cdot|^2 - \mathcal{W}(e^{(\cdot)}) - \frac{1}{2}\mathcal{W}(e^{(\cdot)})^2$

(ii) $\left(f_1^* + \frac{1}{2}|\cdot|^2\right)^* = \frac{1}{4}|\cdot|^2$.

Proof. (i): By definition,

$$\left(f_0^* + \frac{1}{2}|\cdot|^2\right)^*(x) = \sup_{y\in\mathbb{R}}\left\{xy - f_0^*(y) - \frac{1}{2}y^2\right\} \tag{3.13}$$

Differentiating the inner term with respect to y and setting equal to zero, we have that the supremum is obtained when y satisfies $e^y = x - y$. We will solve for y. Here the Wikipedia page about Lambert $\mathcal{W}$ suggests a handy method [19]. Let $\gamma = x - y$. Then $e^y = \gamma$ and so

$$\gamma e^\gamma = e^y e^{x-y} = e^x$$

and so we have $\gamma = \mathcal{W}(e^x)$. Thus we have $e^y = \mathcal{W}(e^x)$. Taking the log of both sides,

$$y = \log(\mathcal{W}(e^x)) = \log(e^x) - \mathcal{W}(e^x) = x - \mathcal{W}(e^x).$$

Using this as the y value for the inner term in (3.13), we obtain

$$\left(f_0^* + \frac{1}{2}|\cdot|^2\right)^*(x) = x(x - \mathcal{W}(e^x)) - \exp(x - \mathcal{W}(e^x)) - \frac{1}{2}\left(x - \mathcal{W}(e^x)\right)^2,$$

which simplifies to the form in 3.7.

(ii): This is a well-known result and may be obtained by simple arithmetic. □

Lemma 3.8. *Let* θ *be defined as follows*

$$\theta := \left((1-\lambda)\left(\frac{1}{2}(\cdot)^2 - \mathcal{W}(e^{(\cdot)}) - \frac{1}{2}\mathcal{W}(e^{(\cdot)})^2\right) + \lambda\left(\frac{1}{4}(\cdot)^2\right)\right)^*.$$

Then it holds that

$$\mathrm{p}_\theta(x) = \left(\frac{2}{\lambda} - 2\right)\mathcal{W}\left(\frac{\lambda e^{\frac{2x}{2-\lambda}}}{2-\lambda}\right) + \frac{2x}{2-\lambda} \tag{3.14}$$

and so we may write

$$\theta(x) = x\mathrm{p}_\theta(x) - (1-\lambda)\left(\frac{1}{2}\mathrm{p}_\theta(x)^2 - \mathcal{W}(e^{\mathrm{p}_\theta(x)}) - \frac{1}{2}\mathcal{W}(e^{\mathrm{p}_\theta(x)})^2\right) - \frac{\lambda}{4}\mathrm{p}_\theta(x)^2.$$

Proof. Now by definition

$$\theta(x) = \sup_{y\in\mathbb{R}}\left\{xy - (1-\lambda)\left(\frac{1}{2}y^2 - \mathcal{W}(e^y) - \frac{1}{2}\mathcal{W}(e^y)^2\right) - \frac{\lambda}{4}y^2\right\},$$

which simplifies to

$$\theta(x) = \sup_{y\in\mathbb{R}}\left\{xy + \frac{1}{2}(1-\lambda)\mathcal{W}(e^y)^2 + (1-\lambda)\mathcal{W}(e^y) + \frac{1}{4}(\lambda-2)y^2\right\}.$$

Differentiating the inner term with respect to y and setting equal to zero, we obtain

$$(1-\lambda)\mathcal{W}\left(e^{y}\right)+\left(\frac{1}{2}\lambda-1\right)y+x=0 \tag{3.15}$$

We will show that (3.15) is true if $y=\mathrm{p}_\theta(x)$. First we will rewrite (3.15) using the fact that $\mathcal{W}(a)=b$ if and only if $be^b=a$ which allows us to remove the $\mathcal{W}(e^y)$ term as follows:

$$\mathcal{W}(e^y)=\frac{\left(1-\frac{\lambda}{2}\right)y-x}{1-\lambda}$$

$$\text{and so}\quad e^y=\left(\frac{\left(1-\frac{\lambda}{2}\right)y-x}{1-\lambda}\right)e^{\left(\frac{\left(1-\frac{\lambda}{2}\right)y-x}{1-\lambda}\right)},$$

$$\text{which simplifies to}\quad 0=(\lambda y+2x-2y)e^{\left(\frac{(\lambda-2)y+2x}{2\lambda-2}\right)}-2e^y(\lambda-1).$$

This is the form returned by *Maple*. We further consolidate y terms as follows,

$$0=\left((\lambda-2)y+2x\right)\left(e^y\right)^{\left(\frac{\lambda-2}{2\lambda-2}\right)}e^{\left(\frac{x}{\lambda-1}\right)}-2(\lambda-1)e^y, \tag{3.16}$$

which is the form we will use. We need only to show that

$$((\lambda-2)\mathrm{p}_\theta(x)+2x)\,e^{\left(\frac{x}{\lambda-1}\right)}\left(e^{\mathrm{p}_\theta(x)}\right)^{\left(\frac{\lambda-2}{2\lambda-2}\right)}-2(\lambda-1)e^{\mathrm{p}_\theta(x)}=0. \tag{3.17}$$

First consider the term $e^{\mathrm{p}_\theta(x)}$. Since for any a,b,z we have that

$$e^{aW(z)+b}=(e^{W(z)})^a e^b=\left(\frac{z}{W(z)}\right)^a e^b,$$

we may let

$$a:=\left(\frac{2}{\lambda}-2\right),\quad b:=\frac{2x}{2-\lambda},\quad z:=\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{2-\lambda}, \tag{3.18}$$

and thusly rewrite

$$e^{\mathrm{p}_\theta(x)}=\left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{(2-\lambda)\mathcal{W}\left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{2-\lambda}\right)}\right)^{\left(\frac{2}{\lambda}-2\right)}e^{\left(\frac{2x}{2-\lambda}\right)}$$

$$=(2-\lambda)^{\left(2-\frac{2}{\lambda}\right)}\lambda^{\left(\frac{2}{\lambda}-2\right)}e^{\left(\frac{2x}{\lambda}\right)}\mathcal{W}\left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{2-\lambda}\right)^{\left(2-\frac{2}{\lambda}\right)}. \tag{3.19}$$

From this, we have that

$$e^{\left(\frac{x}{\lambda-1}\right)}\left(e^{\mathrm{p}(x)}\right)^{\left(\frac{\lambda-2}{2\lambda-2}\right)}=e^{\left(\frac{x}{\lambda-1}\right)}\left(\left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{(2-\lambda)\mathcal{W}\left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{2-\lambda}\right)}\right)^{\left(\frac{2}{\lambda}-2\right)}e^{\left(\frac{2x}{2-\lambda}\right)}\right)^{\frac{\lambda-2}{2\lambda-2}}$$

$$= e^{\left(\frac{x}{\lambda-1}\right)} \left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{(2-\lambda)\mathcal{W}\left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{2-\lambda}\right)} \right)^{\left(\frac{2-\lambda}{\lambda}\right)} e^{\left(-\frac{x}{\lambda-1}\right)}$$

$$= \left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{(2-\lambda)\mathcal{W}\left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{2-\lambda}\right)} \right)^{\left(\frac{2-\lambda}{\lambda}\right)}$$

$$= (2-\lambda)^{\left(\frac{\lambda-2}{\lambda}\right)} \lambda^{\left(\frac{2-\lambda}{\lambda}\right)} e^{\frac{2x}{\lambda}} \mathcal{W}\left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{2-\lambda} \right)^{\left(\frac{\lambda-2}{\lambda}\right)}. \tag{3.20}$$

Using (3.14), (3.19), and (3.20), we may rewrite (3.17) as follows:

$$\begin{aligned} 0 = {} & (\lambda-2)\left(\left(\tfrac{2}{\lambda}-2\right)\mathcal{W}\left(\frac{\lambda e^{\frac{2x}{2-\lambda}}}{2-\lambda}\right) + \frac{2x}{2-\lambda} \right)(2-\lambda)^{\left(\frac{\lambda-2}{\lambda}\right)} \lambda^{\left(\frac{2-\lambda}{\lambda}\right)} e^{\frac{2x}{\lambda}} \mathcal{W}\left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{2-\lambda} \right)^{\left(\frac{\lambda-2}{\lambda}\right)} \\ & + 2x(2-\lambda)^{\left(\frac{\lambda-2}{\lambda}\right)} \lambda^{\left(\frac{2-\lambda}{\lambda}\right)} e^{\frac{2x}{\lambda}} \mathcal{W}\left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{2-\lambda} \right)^{\left(\frac{\lambda-2}{\lambda}\right)} \\ & - 2(\lambda-1)\,(2-\lambda)^{\left(2-\frac{2}{\lambda}\right)} \lambda^{\left(\frac{2}{\lambda}-2\right)} e^{\left(\frac{2x}{\lambda}\right)} \mathcal{W}\left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{2-\lambda} \right)^{\left(2-\frac{2}{\lambda}\right)}. \end{aligned}$$

The positive and negative terms of the form

$$2x(2-\lambda)^{\left(\frac{\lambda-2}{\lambda}\right)} \lambda^{\left(\frac{2-\lambda}{\lambda}\right)} e^{\frac{2x}{\lambda}} \mathcal{W}\left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{2-\lambda} \right)^{\left(\frac{\lambda-2}{\lambda}\right)}$$

cancel each other out, leaving

$$\begin{aligned} 0 = {} & (\lambda-2)\left(\left(\tfrac{2}{\lambda}-2\right)\mathcal{W}\left(\frac{\lambda e^{\frac{2x}{2-\lambda}}}{2-\lambda}\right) \right)(2-\lambda)^{\left(\frac{\lambda-2}{\lambda}\right)} \lambda^{\left(\frac{2-\lambda}{\lambda}\right)} e^{\frac{2x}{\lambda}} \mathcal{W}\left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{2-\lambda} \right)^{\left(\frac{\lambda-2}{\lambda}\right)} \\ & - 2(\lambda-1)\,(2-\lambda)^{\left(2-\frac{2}{\lambda}\right)} \lambda^{\left(\frac{2}{\lambda}-2\right)} e^{\left(\frac{2x}{\lambda}\right)} \mathcal{W}\left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{2-\lambda} \right)^{\left(2-\frac{2}{\lambda}\right)}, \end{aligned}$$

which further simplifies to

$$\begin{aligned} 0 = {} & 2(\lambda-2)\left(\frac{1}{\lambda}-1\right)\left(\frac{2-\lambda}{\lambda}\right)^{\left(\frac{\lambda-2}{\lambda}\right)} e^{\left(\frac{2x}{\lambda}\right)} \mathcal{W}\left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{2-\lambda} \right)^{\left(2-\frac{2}{\lambda}\right)} \\ & - 2(\lambda-1)\left(\frac{2-\lambda}{\lambda}\right)^{\left(2-\frac{2}{\lambda}\right)} e^{\left(\frac{2x}{\lambda}\right)} \mathcal{W}\left(\frac{\lambda e^{\left(\frac{2x}{2-\lambda}\right)}}{2-\lambda} \right)^{\left(2-\frac{2}{\lambda}\right)}. \end{aligned}$$

Finally, $(\lambda-2)\left(\frac{1}{\lambda}-1\right)=(\lambda-1)\left(\frac{2-\lambda}{\lambda}\right)$ and so the above equation is true, completing the result. □

Theorem 3.9. *Let f_0, f_1 be defined as in Lemma 3.7. Then*

$$f_\lambda^*(x)=(1-\lambda)\mathcal{W}\left(e^{\left(\left(\frac{2}{\lambda}-2\right)\mathcal{W}\left(\frac{\lambda}{2-\lambda}e^{\left(\frac{2x}{2-\lambda}\right)}\right)+\frac{2x}{2-\lambda}\right)}\right)+\frac{\lambda x^2}{4-2\lambda} \tag{3.21}$$

$$+\frac{1}{2}(1-\lambda)\mathcal{W}\left(e^{\left(\left(\frac{2}{\lambda}-2\right)\mathcal{W}\left(\frac{\lambda}{2-\lambda}e^{\left(\frac{2x}{2-\lambda}\right)}\right)+\frac{2x}{2-\lambda}\right)}\right)^2 \tag{3.22}$$

$$+\frac{(\lambda-1)^2(\lambda-2)}{\lambda^2}\mathcal{W}\left(\frac{\lambda}{2-\lambda}e^{\left(\frac{2x}{2-\lambda}\right)}\right)^2.$$

Proof. Using Definition 2.1 together with Lemma 3.7 we have that

$$f_\lambda^*=\left((1-\lambda)\left(\frac{1}{2}|\cdot|^2-\mathcal{W}(e^{(\cdot)})-\frac{1}{2}\mathcal{W}(e^{(\cdot)})^2\right)+\lambda\left(\frac{1}{4}(\cdot)^2\right)\right)^*-\frac{1}{2}(\cdot)^2.$$

This is just

$$f_\lambda^*=\theta-\frac{1}{2}(\cdot)^2,$$

where $\theta, \mathrm{p}_\theta$ are as in Lemma 3.8. From this, we obtain

$$f_\lambda^*(x)=x\mathrm{p}_\theta(x)-(1-\lambda)\left(\frac{1}{2}\mathrm{p}_\theta(x)^2-\mathcal{W}(e^{\mathrm{p}_\theta(x)})-\frac{1}{2}\mathcal{W}(e^{\mathrm{p}_\theta(x)})^2\right)$$
$$-\frac{\lambda}{4}\mathrm{p}_\theta(x)^2-\frac{1}{2}x^2.$$

This simplifies, by a great deal of arithmetic, to the form we see in (3.21), completing the result. □

Similarly to Theorem 3.5, the results admitting Theorem 3.9 (in particular, Lemma 3.8) could not be obtained through the use of *SCAT* or *Maple* alone because these packages cannot invert (3.16). The solution was again discovered with a method similar to that of Theorem 3.5.

Again within minutes of choosing correctly we "knew" the answer, because we could visually read off the functions f_0^* and f_1^* in Figure 4, even though a proof took much longer. Figures 3 and 4 highlight an advantageous characteristic of the proximal average, which we provide in the following remark.

Remark 3.10. Let $f_0, f_1 \in \mathcal{F}$ and $\lambda \in]0,1[$. Let $f_\lambda := \mathcal{P}(f_0, \lambda, f_1)$. Suppose that f_0 or f_1 has full domain and that f_0^* or f_1^* has full domain. Then the following hold:

(1) Both f_λ and f_λ^* have full domain.
(2) If f_0 or f_1 is differentiable everywhere, then so is f_λ.
(3) If f_0 or f_1 is strictly convex and its Fenchel conjugate has full domain, then f_λ is strictly convex.

For a proof, see [3, Theorem 6.2].

Figures 3 and 4 also illustrate another important difference between the behaviour of limiting cases for the proximal average and for the ordinary average.

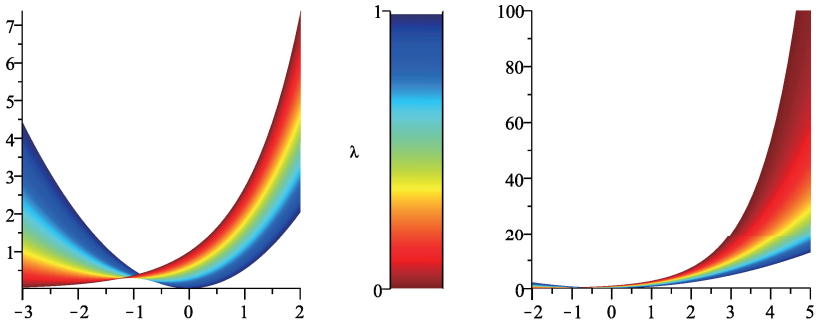

FIGURE 4. f_λ^* from Theorem 3.9

Remark 3.11 (Limiting Cases for Proximal Average)**.** In juxtaposition with Remark 3.2, we obtain different limiting cases for f_λ (3.12) and f_λ^* (3.21). For f_λ, in the limits at $0, 1$, we reobtain f_0 and f_1 respectively. For f_λ^* in the limits at $0, 1$, we reobtain exp and $\frac{1}{2}|\cdot|^2$. This is more natural, because these are f_0^* and f_1^* respectively, and so our continuous transformation of our functions has corresponded with a continuous transformation of their conjugates.

The juxtaposition in Remark 3.11 is both an immediate consequence and an excellent illustration of Remark 2.6. Where $f_0, f_1 \in \mathcal{F}$ and $(\lambda_n)_{n\in\mathbb{N}}$ is a sequence in $[0, 1]$, we have from Remark 2.6 that

$$\begin{aligned}
\text{if } \lambda_n \to 0 \quad \text{then} \quad & \mathcal{P}(f_0, \lambda_n, f_1) \xrightarrow{\text{e}} \mathcal{P}(f_0, 0, f_1) = f_0 \\
\text{and} \quad & \mathcal{P}(f_0^*, \lambda_n, f_1^*) \xrightarrow{\text{e}} \mathcal{P}(f_0^*, 0, f_1^*) = f_0^* \\
\text{and if } \lambda_n \to 1 \quad \text{then} \quad & \mathcal{P}(f_0, \lambda_n, f_1) \xrightarrow{\text{e}} \mathcal{P}(f_0, 1, f_1) = f_1 \\
\text{and} \quad & \mathcal{P}(f_0^*, \lambda_n, f_1^*) \xrightarrow{\text{e}} \mathcal{P}(f_0^*, 1, f_1^*) = f_1^*,
\end{aligned}$$

which is both elegant and convenient.

4. Minimizing an entropy functional

In their 2016 paper [9] Borwein & Lindstrom illustrated the utility of the Lambert $\mathcal{W}$ function by showing how it naturally arises in the problem of minimizing an entropy functional of the form

$$\begin{aligned}
I_f &: L^1([0, 1]) \to \mathbb{R} \\
\text{by} \quad I_f &: x \mapsto \int_0^1 f(x(s))\mathrm{d}s,
\end{aligned}$$

where f is a proper, closed convex function. The problem is to minimize I_f subject to finitely many continuous linear constraints of the form

$$\langle a_k, x\rangle = \int_0^1 a_k(s)x(s)\mathrm{d}s = b_k,$$

for $1 \leq k \leq n$. We may write this linear equality constraint concisely as

$$A : L^1([0,1]) \to \mathbb{R}^n$$
$$\text{by} \quad A : x \mapsto \left(\int_0^1 a_1(s)x(s)\mathrm{d}s, \ldots, \int_0^1 a_n(s)x(s) \right) = b$$
$$\text{where} \quad b := A\rho$$

where $\rho, a_k \in L^\infty([0,1])$ and ρ is a given function used to generate the data vector b. When f^* is smooth and everywhere finite on the real line, the problem

$$\inf_{x \in L^1} \{ I_f(x) | Ax = b \} \tag{4.1}$$

reduces to solving a finite nonlinear equation

$$\int_0^1 (f^*)' \left(\sum_{j=1}^n \mu_j a_j(s) \right) a_k(s)\mathrm{d}s = b_k \quad (1 \leq k \leq n). \tag{4.2}$$

A discussion of why this is the case is given in [9, Section 7], which employs results from Jonathan Borwein's works co-authored with Adrian Lewis [8], Qiji Zhu [12], and Jon Vanderwerff [10], and Liangjin Yao [11]. The matters of primal attainment and constraint qualification are addressed in Borwein's and Lewis' article [7], and an augmented discussion of strong duality is given in Lindstrom's PhD dissertation [15].

As was also true in the setting of [9], this problem and methods discussed in this section are informed by methods found in all of these works, to which we refer the reader for additional information about any underlying theory.

For the function f in the construction of I_f, Borwein et al. opted to use f_t from (3.2), for which the corresponding f_t^* has the form in (3.3) for $0 < t < 1$ and $f_t^* = \exp$, $\frac{1}{2}|\cdot|^2$ for $t = 0, 1$, respectively. For this choice:

$$(f_t^*)'(x) = \begin{cases} \frac{1-t}{t} \mathcal{W}\left(\frac{t}{1-t} \exp\left(\frac{x}{1-t} \right) \right) & \text{if} \quad t \in]0,1[\\ \exp(x) & \text{if} \quad t = 0 \\ x & \text{if} \quad t = 1. \end{cases}$$

In the limiting case as t approaches 0, (f_t^*) approaches exp, while in the limiting case as t approaches 1 we obtain $\max\{0, x\}$, given the discussion of the limiting cases of f_t^* in Remark 3.2.

Remark 4.1. Let $f : X \to]-\infty, +\infty[$ be proper. Then $f^* : X \to]-\infty, +\infty]$ is proper. Let $x, u \in X$. Then

$$u \in \partial f^*(x) \iff f(u) + f^*(x) = \langle x, u \rangle \iff x \in \partial f(u).$$

For details, see [1, proposition 16.9]. Thus we have that

$$\mathrm{ran}(\partial f^*) \subset \mathrm{dom}(\partial f) \subset \mathrm{dom}(f).$$

Consequently, for all $x \in X$ we have that:

$$(\ \forall t \in [0,1[\ ) \quad f_t^*(x) \in \mathrm{dom}(f_t) = [0, \infty[\,.$$

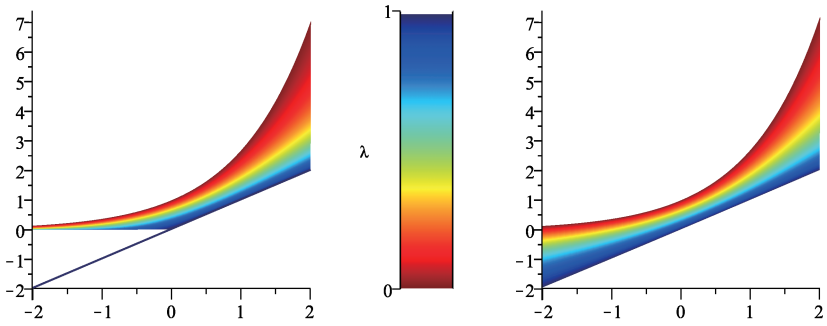

FIGURE 5. $(f_t^*)'$ (left) and $(f_\lambda^*)'$ (right)

For the function f in the construction of I_f, we consider f_λ from (3.12), for which the corresponding f_λ^* has the form in (3.21) for $0 < \lambda < 1$ and $f_\lambda^* = \exp$, $\frac{1}{2}|\cdot|^2$ for $\lambda = 0, 1$ respectively as explained in Remark 3.11 and as follows from Theorem 3.9 by differentiation. For this choice:

$$
\begin{aligned}
&(f_0^*)'(x) = \exp(x)\\
&(f_1^*)'(x) = x\\
&\text{and for } \ 0 < \lambda < 1,\\
&(f_\lambda^*)'(x) = \frac{1}{1+\omega(x)}\left(\frac{2(1-\lambda)}{\lambda}\left(\omega(x) - \frac{\lambda}{\lambda-2}\right)\mathcal{W}\left(e^{\left(\frac{2}{\lambda}-2\right)\omega(x)+\frac{2x}{2-\lambda}}\right)\right.\\
&\qquad\left. - \frac{4(\lambda-1)^2}{\lambda^2}\omega(x)^2 + \frac{\lambda x}{2-\lambda}\omega(x) + \frac{x\lambda}{2-\lambda}\right)\\
&\text{where} \quad \omega(x) = \mathcal{W}\left(\frac{\lambda}{2-\lambda}e^{\frac{2x}{2-\lambda}}\right).
\end{aligned}
$$

The functions $(f_t^*)'$ and $(f_\lambda^*)'$ may be seen in Figure 5. Figure 5 also serves to highlight one of the consequences of Remark 3.10 in our case.

In juxtaposition with f_t, which takes the value infinity for all negative real values, f_λ has full domain because f_1 has full domain. Consequently the conjugate f_λ^* of f_λ decreases on part of its domain for values of $\lambda \in]0, 1]$; this is in contrast with the conjugate f_t^* of f_t, which is nondecreasing except for the case $t = 1$. As a result, the image of $(f_\lambda^*)'$ contains negative numbers for $\lambda \in]0, 1]$ while the image of $(f_t^*)'$ contains negative numbers only for $t = 1$. In terms of Remark 4.1, f_λ differs from f_t in the sense that

$$(\forall x \in \mathcal{H}) \quad (\forall \lambda \in]0, 1]) \quad f_\lambda^*(x) \in \operatorname{dom}(f_\lambda) =]-\infty, \infty[\,.$$

Remark 4.2. In their original article [9], the authors have labelled solutions computed for the *limiting* case $\lim_{t\to 1}(f_t^*)' = \max\{\cdot, 0\}$ with the label $t = 1$; however, $(f_0^*)'$ is actually just the identity $x \mapsto x$. This labelling confusion does not change any of the key results of the paper; it affects only computed examples.

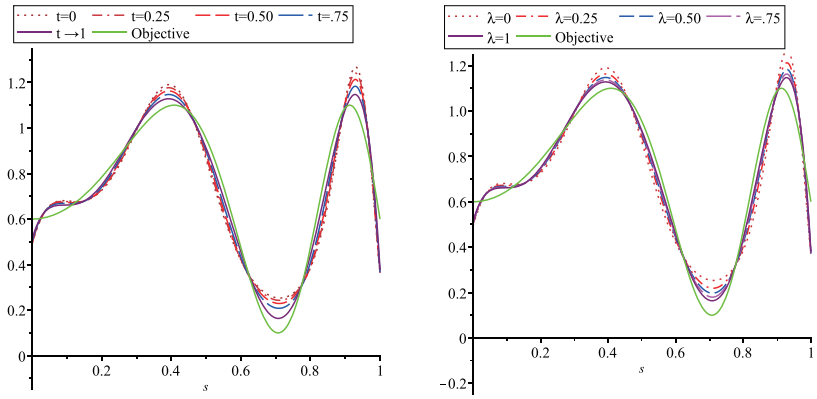

FIGURE 6. Primal solutions from Example 4.3 appear quite similar.

Where $\mu_1, \ldots, \mu_n$ are the optimal multipliers in (4.2), the primal solution x_λ to the primal problem (4.1) is then given by

$$x_\lambda(s) = (f_\lambda^*)' \left(\sum_{j=1}^{n} \mu_j a_j(s) \right).$$

A key difference between our setting and that of [9] is then immediately apparent: for $t \neq 1$, the primal solutions when optimizing with the conventional average f_t could not take on negative values. When instead using the proximal average, f_λ, the primal solutions may take on negative values so long as $\lambda \neq 0$. The hard barrier (or lack of hard barrier) against negative values may be considered either an advantage or disadvantage depending upon one's intentions.

4.1. **Computed Examples.** For all examples where we solve (4.1), we compute with 8 moments ($n = 8$), and we follow the lead of Borwein & Lindstrom [9], employing a Gaussian quadrature with 20 abscissas for the numerical integration necessary to solve the system (4.2). One may consult Borwein & Lindstrom [9] for an index on computation which explains a simple implementation with Newton's method. When reporting solutions for the weighted average f_t, instead of the case where $t = 1$, we choose to plot the limiting case:

$$\lim_{t \to 1} (f_t^*)' = \max\{\cdot, 0\}.$$

The first reason for this is that the exact cases $t = 1$ and $\lambda = 1$ coincide (see Remark 3.11), and so comparing them is not as interesting. The second reason is to be consistent with the method of reporting employed in [9] (see Remark 4.2).

We compute with vertical translations of the function we wish to reconstruct, the function used by Borwein & Lindstrom,

$$\rho : s \mapsto \frac{3}{5} + \frac{1}{2} \sin\left(3\pi s^2\right), \tag{4.3}$$

with which we compute in Example 4.3.

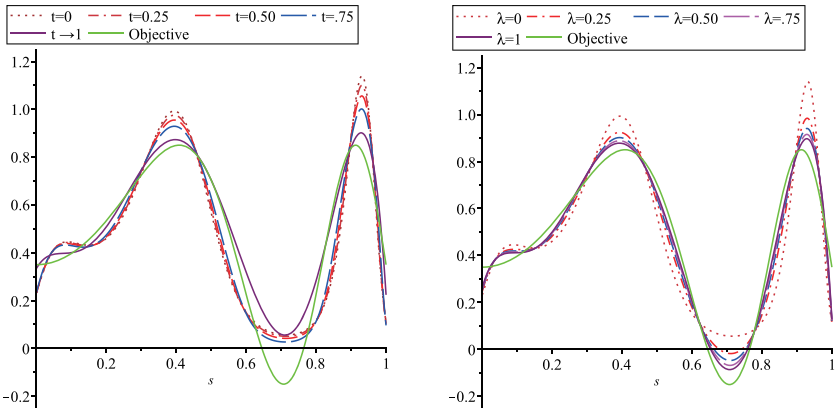

FIGURE 7. Primal solutions from Example 4.4 are noticeably different, particularly where the objective function is negative-valued.

Example 4.3 (Similarities between weighted average and proximal average)**.** Figure 6 shows similar-looking primal solutions obtained by computing with the weighted average f_t and the proximal average f_λ where the objective function is as in (4.3). Importantly, in this case $\rho(s) \geq 0 \; \forall s \in [0,1]$.

The advantage of homotopy becomes apparent when the objective function has negative output values, as it does in the next example.

Example 4.4 (Differences between weighted average and proximal average)**.** For the second example, we compute with a negative translation of the previous objective function:

$$\rho : s \mapsto \frac{7}{20} + \frac{1}{2}\sin\left(3\pi s^2\right).$$

Figure 7 shows the primal solutions for the weighted average f_t at left and for the proximal average f_λ at right.

The presence of negative values for the objective function ρ illuminates an important advantage of the proximal average f_λ. Because $(f_\lambda^*)'$ is allowed to have negative range values for $\lambda > 0$ (as shown in Figure 5), the primal solutions in the proximal average case are able to have negative range values for $\lambda > 0$. As a result, the primal solutions corresponding to the proximal average with $\lambda > 0$ are a better fit for our objective function ρ than the primal solutions corresponding to the weighted average.

For the advantage of homotopy—that primal solutions may take on negative values when $\lambda \neq 0$—there is a price to pay computationally. Namely, in contradistinction with the case of the weighted average f_t, Newton's method no longer reliably solves the problem for the proximal average f_λ when the objective function is permitted to take values below or near zero. This is shown in Example 4.5.

Example 4.5 (Computational challenge)**.** To illustrate a computational disadvantage of homotopy, we compute with the data vector b generated by

$$\rho : s \mapsto \frac{1}{5} + \frac{1}{2}\sin\left(3\pi s^2\right),$$

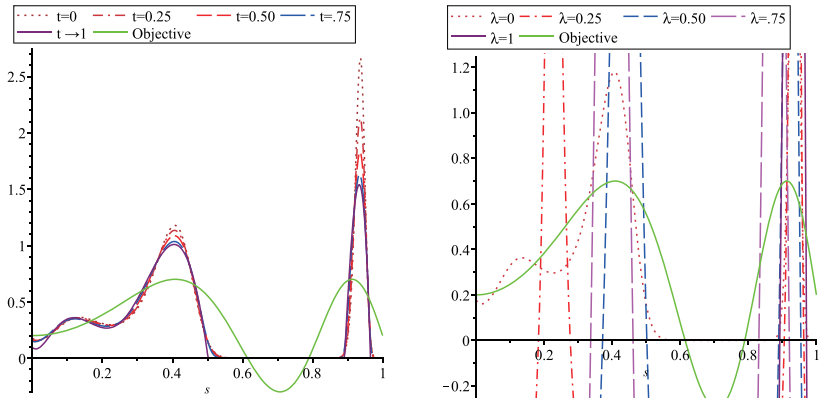

FIGURE 8. Newton's method is less reliable for the proximal average in the case of Example 4.5.

which is another downward translation of the function used to generate the data vector in Example 4.4. With the starting point of $(\frac{1}{2},\ldots,\frac{1}{2}) \in \mathbb{R}^8$, Newton's method fails to find the optimal solution for the homotopy f_λ while it still manages to find the optimal solution for the weighted average f_t. Figure 8 shows the primal solution for f_t at left and the primal output when Newton's method is paused after 400 iterates for f_λ at right.

Rather than using Newton's method, one might instead use gradient descent to solve the system (4.2), either by seeking to

(i) solve the dual problem directly, or
(ii) minimize the sum of the squares of the gradient components.

(i): In the former case, the gradient we use has the n components

$$\int_0^1 (f^*)'\left(\sum_{j=1}^n \mu_j a_j(s)\right) a_k(s)\mathrm{d}s - b_k \quad (1 \leq k \leq n),$$

which is, of course, the system from (4.2).
(ii): In the latter case, the problem becomes:

$$\text{Find } \mu \in \mathbb{R}^n \text{ such that } \mathcal{G}(\mu) := \sum_{k=1}^n \mathcal{G}_k(\mu) = 0 \text{ where}$$

$$\mathcal{G}_k : \mu \mapsto \left(\int_0^1 (f^*)'\left(\sum_{j=1}^n \mu_j a_j(s)\right) a_k(s)\mathrm{d}s - b_k\right)^2, (1 \leq k \leq n).$$

Again using a Gaussian quadrature rule with m abscissas $s_1,\ldots,s_m$ and corresponding weights $w_1,\ldots,w_m$, we let

$$\sum_{i=1}^m w_i (f^*)'\left(\sum_{j=1}^n \mu_j a_j(s_i)\right) a_k(s_i) :\approx \int_0^1 (f^*)'\left(\sum_{j=1}^n \mu_j a_j(s)\right) a_k(s)\mathrm{d}s$$

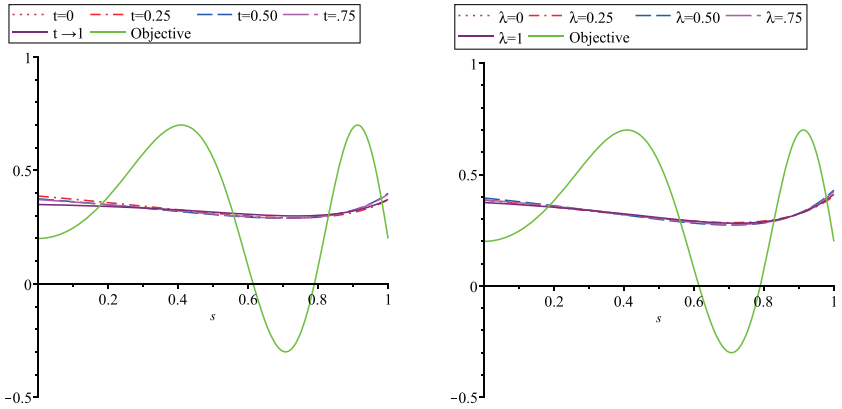

FIGURE 9. Gradient descent may be insufficient as in Example 4.6.

and so (4.2) reduces to finding $\mu \in \mathbb{R}^n$ such that:

$$\mathcal{G}(\mu) = \sum_{k=1}^{n} \left(\left(\sum_{i=1}^{m} w_i (f^*)' \left(\sum_{j=1}^{n} \mu_j a_j(s_i) \right) a_k(s_i) \right) - b_k \right)^2 = 0. \tag{4.4}$$

To solve (4.4), we may use gradient descent where

$$\nabla \mathcal{G}(\mu) = \left(\frac{\partial}{\partial \mu_1} \mathcal{G}(\mu), \ldots, \frac{\partial}{\partial \mu_n} \mathcal{G}(\mu) \right).$$

Example 4.6 (Gradient Descent)**.** When we implement gradient descent for either of the above approaches with the same starting point and objective function from Example 4.5, the method tends to stall. Consequently, the primal solutions yielded do not correspond to the true solution for the problem and only roughly resemble the function ρ used to generate the data. This is shown at right in Figure 9.

4.2. **A homotopy method.** As a remedy for stalling, we may employ a homotopy-type method whereby we solve a *sequence of problems*. Suppose we seek a solution where the objective function is given by

$$\rho : s \mapsto \frac{7}{20} + \frac{1}{2} \sin\left(3\pi s^2\right) - \Delta.$$

Then we let

$$\rho_N : s \mapsto \frac{7}{20} + \frac{1}{2} \sin\left(3\pi s^2\right) - N\delta, \quad N \in \{0, \ldots, \upsilon\} \subset \mathbb{N}, \quad \delta > 0, \quad \upsilon\delta = \Delta.$$

We further define μ^N to be the solution to (4.2) for the problem corresponding to the linear constraint generated by the function ρ_N. We can find μ^0 with Newton's method (and did so in Example 4.4). We may then use μ^0 as our *starting point* for solving the problem corresponding to objective function ρ_1. If we are successful, we may then use the solution, μ^1, as our starting point for finding μ^2. Continuing in this fashion we aim to solve a sequence of problems where the final problem corresponds to the function with which we are concerned. The solution, μ^υ, is the solution we seek.

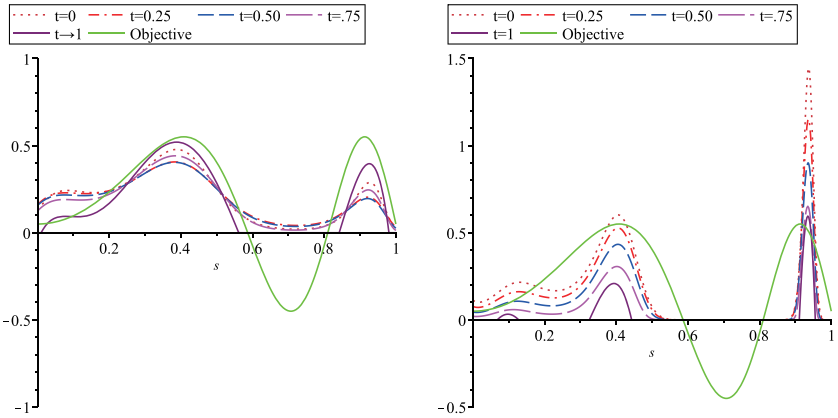

FIGURE 10. Primal values obtained by computing with weighted average in Example 4.7.

This may be thought of as a homotopy method, in the sense that we solve a sequence of problems corresponding to a sequence of perturbed linear constraints $b^0, b^1, b^2, \ldots$ where the solution corresponding to the constraint b^0 is known and the solution corresponding to the constraint b^υ is the one we seek. We illustrate in the following example.

Example 4.7 (Solving a sequence of minimization problems)**.** We desire to solve for the constraint $Ax = b = A\rho$ with

$$\rho : s \mapsto \frac{7}{20} + \frac{1}{2}\sin\left(3\pi s^2\right) - \frac{3}{10} = \frac{1}{20} + \frac{1}{2}\sin\left(3\pi s^2\right).$$

Then, letting $\delta = \frac{1}{10}$, and $\upsilon = 3$, we may consider the sequence of problems corresponding to the objective functions given by

$$\rho_N : s \mapsto \frac{7}{20} + \frac{1}{2}\sin\left(3\pi s^2\right) - N\frac{1}{10}, \quad N \in \{0, .., 3\}.$$

We computed μ^0 using Newton's method in Example 4.4. Using gradient descent with a step size modifier of $\frac{1}{10}$ and taking μ^0 as our starting value, we obtain μ^1. In the same way, we use μ^1 to find μ^2; finally we use μ^2 to find μ^3, which is the solution we seek for the minimization problem induced by the objective function ρ. The corresponding primal values for various λ are shown in Figure 11.

Notice that the translated generating function ρ used to generate the linear constraint in Example 4.7 has actually been translated even further than the version used to generate the linear constraint in both Examples 4.5 and 4.6. This homotopy method appears also to solve the proximal version of the problem from Examples 4.5 and 4.6.

For comparison, we show the resultant primal values obtained by computing with the weighted average in Figure 10. At left we computed with the function $G(\mu)$, and at right we attacked the dual problem directly. The solutions for f_t at left are distinctly different from those at right, which more closely resemble the weighted average solutions in Figure 8 from Example 4.5. In the table below

we compare the errors from the linear constraint where x_t is the primal solution obtained by computing with f_t. For Example 4.5 we used Newton's method. For Example 4.7, we computed 5 iterates for the first subproblems and 100 iterates for the final subproblem. When working with $G(\mu)$, we used a gradient descent step size modifier of 1/10; when attacking the dual problem directly, we used a size of 1.

t value	Example 4.5 Newton's $\|Ax_t - b\|$	Example 4.7 $G(\mu)$ $\|Ax_t - b\|$	Example 4.7 Dual direct $\|Ax_t - b\|$
0	7.46E−11	3.82E−2	7.91E−3
0.25	6.26E−11	3.61E−2	3.37E−2
0.5	2.09E−10	3.55E−2	7.33E−2
0.75	4.42E−10	3.24E−2	1.25E−1
$\to 1$	2.61E−3	2.04E−2	1.72E−1

Computing with the weighted average, we have solutions that do a poorer job of satisfying the linear constraint than in Example 4.5, where ρ has been translated downward by a smaller amount. While the observations we will make about the proximal case below suggest that a better satisfaction of the linear constraint may be possible if we continue to run more iterates, it is also likely that we have reached the limitations of the data for which the weighted average can be successfully used. The reasons are as follows.

Since ρ returns negative values, $\rho \notin \mathrm{dom}(I_{f_t})$. For this reason, it is difficult to verify whether or not the conditions for strong duality hold unless we can find some other $x \in \mathrm{dom} I_{f_t}$ such that $Ax = b$ (for example, our numerically obtained solutions for $t < 1$ from Example 4.5).

In fact, ρ may have been translated so far downward that it is no longer possible to satisfy the linear constraint. This occurs if there does not exist an $x \in \mathrm{dom}(I_{f_t})$ such that $Ax = b$. Since b still lies in the positive orthant, it is difficult to verify whether this has occurred for the present example. However, further translations downward will eventually yield a data vector b that does not lie in the non-negative orthant. Since the monomials $a_1(s), \ldots, a_n(s)$ are non-negative on $[0, 1]$, Ax may only lie outside of the non-negative orthant if $x(s)$ takes on negative values in $[0, 1]$. However, such an $x(s)$ is not in the domain of I_{f_t} unless $t = 1$. In such a case, the linear constraint cannot possibly be satisfied.

In other words, if $\rho(s) \in L^1([0, 1])$ is non-negative, the constraint definitely can be satisfied (indeed, it is satisfied by ρ). If $A\rho$ lies outside of the non-negative orthant, the constraint definitely cannot be satisfied. If $\rho(s)$ takes on negative values in $[0, 1]$ but $A\rho$ is still in the positive orthant, determining whether or not the linear constraint can be satisfied may be more difficult.

From a numerical standpoint, we may attempt to check by taking the linear system $Mx = b$ — where M is the (#moments) × (#abscissas) matrix representing a discretization of A, where cell i in a row j consists of the ith weight multiplied by the values of a_j evaluated at the ith abscissas — for x with the requirement that x lie in the positive orthant. Decreasing the number of abscissas to match the number of moments eliminates free variables, although we pay the price of having possibly eliminated some feasible solutions (solutions lying in the positive orthant).

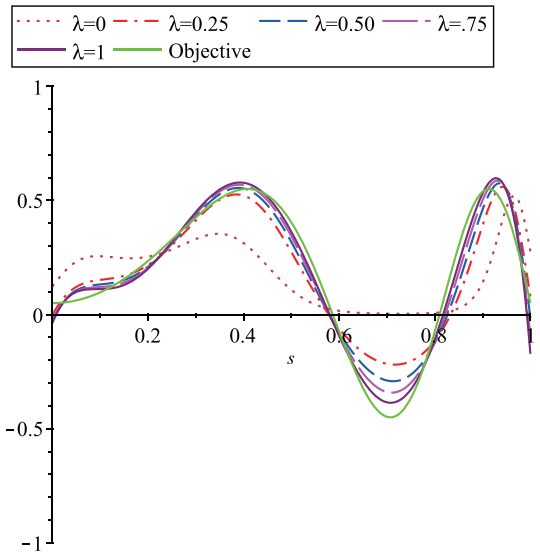

FIGURE 11. Example 4.7 demonstrates that solving a sequence of problems makes more solutions accessible.

With 8 moments and 8 abscissas, the unique solution x for Example 4.4 lies in the positive orthant. For Example 4.5, x lies just outside of the positive orthant, but the unit precision distance we obtained from the linear constraint with Newton's method indicates that by increasing the number of moments to 20 we have recovered a feasible solution. For Example 4.7 with 8 moments and 8 abscissas, our uniquely determined x lies twice as far from the positive orthant. Thus we have a certificate of feasibility for Example 4.5, experimental evidence of feasibility for Example 4.6, and reasonable doubt that feasibility is possible for Example 4.7.

The proximal average, by contrast, does not entail such theoretical problems. After running the first subproblems to 5 iterates, with a gradient descent step size modifier of $1/10$ for minimizing $G(\mu)$ we record the errors from the linear constraint after varying numbers of iterates for the final subproblem as follows.

λ value	100 iterates $\|Ax_\lambda - b\|$	1100 iterates $\|Ax_\lambda - b\|$	2100 iterates $\|Ax_\lambda - b\|$
0	3.82E−2	1.96E−2	1.34E−2
0.25	2.97E−2	4.43E−3	4.35E−3
0.5	1.85E−2	1.77E−3	1.72E−3
0.75	1.12E−2	6.75E−4	6.53E−4
1	8.85E−3	1.90E−3	1.66E−3

For $\lambda > 0$, conditions for strong duality are still satisfied by ρ, and the problem is still feasible. This, combined with the apparent visual fit for $\lambda = 0.25, 0.5, 0.75, 1$, suggests that the homotopy method is working, albeit slowly.

When we attack the dual problem directly, the performance improves. We find that we are able to obtain solutions with only two subproblems, solving first with $N = 0$ and then with $N = 3$. After solving the $N = 0$ case with Newton's method,

we record the errors from the linear constraint after varying numbers of iterates of gradient descent (with no step size modification) for the second subproblem as follows.

λ value	100 iterates $\|Ax_\lambda - b\|$	1100 iterates $\|Ax_\lambda - b\|$	2100 iterates $\|Ax_\lambda - b\|$
0	9.12E−3	4.01E−3	3.37E−3
0.25	2.35E−3	9.95E−4	5.59E−4
0.5	1.07E−3	4.00E−4	2.03E−4
0.75	4.79E−4	1.49E−4	7.89E−5
1	1.57E−4	7.58E−6	1.77E−6

The apparent necessity of homotopy methods when computing with proximal averages when ρ returns lower negative values, particularly for λ nearer to 0, may be related to the penalty for negative values becoming more and more extreme as $\lambda \to 0$, finally achieving a hard barrier at $\lambda = 0$.

5. Conclusion

In this paper, we have catalogued advantages and disadvantages of computing with entropy functionals constructed from proximal averages instead of weighted averages. The weighted average affords ease of computation with hard barriers, but fewer problems may be solvable. In contrast, the proximal average allows us to choose graphically a selection from the net of primal solutions which may afford a better visual fit by being flexible with the enforcement of the barrier. We have explained from a theoretical standpoint why this is the case, and have illustrated it in practice with our examples, giving special attention to the computational challenges one may encounter when working with steep penalties. We have also shown how the Lambert $\mathcal{W}$ function is instrumental in both the weighted averages and proximal averages. In so doing, we have shown how the human-machine collaboration so frequently championed by Borwein may be used to compute hard proximal averages.

We suggest several possibilities for continued investigation.

(i) It is natural to consider also proximal averages employing the Fermi–Dirac entropy, which admits hard barriers on *both* sides of a closed interval $[0, 1]$. The net produced by the proximal average of the Fermi–Dirac entropy with the Boltzmann–Shannon entropy should admit a hard barrier against negative numbers and a flexible barrier against numbers greater than 1. One could also consider the net produced by the Fermi–Dirac entropy with the energy.

(ii) One may also consider the proximal average of two log barriers with empty intersection of their domains.

(iii) It is quite natural to investigate the case where one replaces the energy (as the proximal term in the construction of the proximal average) with another supercoercive function.

(iv) Another natural question is: what might we say about the epigraphs of the net of primal solutions for the entropy minimization problem when proximal averages are employed? May we obtain results on some form of continuous transformation of the primal solutions?

Such investigations are likely to prove interesting, and will almost certainly demand the use of similar human-machine collaboration techniques. This present work is a step in that direction and is a natural template for such future investigation. We conclude by noting that the visualization of the entire family f_λ of functions admitted by the proximal average illustrate epi-continuity in a beautiful and natural way.

Acknowledgement. The authors wish to thank an anonymous referee for their careful reading and constructive comments.

Dedication. This paper is dedicated to the fond memory of Jonathan M. Borwein, our adviser, mentor, and friend. Jon's guiding philosophy and inspiration underpin not only this work, but so much of everything we do — and who we are are — as mathematicians.

References

[1] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Mathematics/Ouvrages de Mathematiques de la SMC, Second edition, Springer, 2017.
[2] H. H. Bauschke, R. Goebel, Y. Lucet and X. Wang, *The proximal average: basic theory*, SIAM Journal on Optimization **19** (2008), 766–785.
[3] H. H. Bauschke, Y. Lucet and M. Trienis, *How to transform one convex function continuously into another*, SIAM Review **50** (2008), 115–132.
[4] H. H. Bauschke, E. Matoušková and S. Reich, *Projection and proximal point methods: Convergence results and counterexamples*, Nonlinear Analysis **56** (2004), 715-738.
[5] H. H. Bauschke and M. von Mohrenschildt, *Symbolic computation of Fenchel conjugates*, ACM SIGSAM Bulletin **40** (2006), 18–28.
[6] J. M. Borwein and C. Hamilton, *Symbolic convex analysis: algorithms and examples*, Mathematical Programming **116** (2009), 17–35.
[7] J. M. Borwein and A. S. Lewis, *Duality relationships for entropy-like minimization problems*, SIAM Control and Optimization **29** (1991), 325–338.
[8] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*, Springer, 2000 (2nd Edition, 2006).
[9] J. M. Borwein and S. B. Lindstrom, *meetings with Lambert $\mathcal{W}$ and other special functions in optimization and analysis*, Pure and Applied Functional Analysis **1** (2016), 361-396.
[10] J. M. Borwein and J. D . Vanderwerff. *Convex Functions: Constructions, Characterizations and Counterexamples*, Cambridge University Press, 2010.
[11] J. M. Borwein and L. Yao, *Legendre-type integrands and convex integral functions*, Journal of Convex Analysis **21** (2014), 264–288.
[12] J. M. Borwein and Q. J. Zhu, *Techniques of Variational Analysis*, CMS/Springer-Verlag, 2005.
[13] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey and D.E. Knuth, *On the LambertW function*, Advances in Computational Mathematics **5** (1996), 329–359.
[14] F. Lauster, D. R. Luke and M. K. Tam, *Symbolic computation with monotone operators*, Set-Valued and Variational Analysis **26** (2018), 353–368.
[15] S.B. Lindstrom, *Proximal point algorithms, dynamical systems, and associated operators: modern perspectives from experimental mathematics*, University of Newcastle, 2018.
[16] R. T. Rockafellar and R.J-B Wets, *Variational Analysis*, Springer Science & Business Media, 2009.
[17] "Symbolic Convex Analysis Tools (SCAT)" package for *Maple*. Available at `http://num.math.uni-goettingen.de/~r.luke/work/research.html` and `http://carma.newcastle.edu.au/ConvexFunctions/SCAT.ZIP` .

[18] Y. Yu, *Better approximation and faster algorithm using the proximal average*, Advances in Neural Information Processing Systems **1** (2013), 458–466.
[19] Wikipedia contributors. "Lambert W function," *Wikipedia, The Free Encyclopedia.* `https://en.wikipedia.org/wiki/Lambert_W_function#Example_1`

Manuscript received July 24 2019
revised September 10 2019

H. H. Bauschke
Mathematics, University of British Columbia Okanagan, Kelowna, B.C., V1V 1V7, Canada
E-mail address: `heinz.bauschke@ubc.ca`

S. B. Lindstrom
CARMA, University of Newcastle, Callaghan, Australia, 2308
E-mail address: `scott.b.lindstrom@polyu.edu.hk`