Journal of Convex Analysis
Volume 29 (2022), No. 4, 1193–1224

A Tight Smooth Approximation of the Maximum Function and its Applications

Ke Yin
Center for Mathematical Sciences, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China
kyin@hust.edu.cn

Kewei Zhang
School of Mathematical Sciences, University of Nottingham, United Kingdom
kewei.zhang@nottingham.ac.uk

Received: December 23, 2021
Accepted: April 22, 2022

We analyse the $C^{1,1}$ tight approximations of the finite maximum function defined by the upper compensated convex transform introduced in a previous paper of the second author [*Compensated convexity and its applications*, Ann. Inst. H. Poincaré (C), Non Linear Analysis 25/4 (2008) 743–771]. We present the precise geometric structure, the tightness property, the sharp error estimates and the asymptotic properties of our approximation. We compare our method with the well-known 'log-sum-exp' smooth approximation by showing that our approximation is geometrically much sharper than the 'log-sum-exp' approximation. We apply our results to smooth approximations for functions defined by the maximum of finitely many smooth functions in $\mathbb{R}^n$ arising from finite and semi-infinite minimax optimization problems.

Keywords: Maximum function, compensated convex transforms, tight smooth approximation, convex optimization, minimax problem.

2010 Mathematics Subject Classification: 90C25, 90C30, 90C34, 90C47, 65K05, 65K10, 49M20, 49M37.

1. Introduction

In this paper we introduce a computable tight (or blending) locally $C^{1,1}$ approximation method for

$$g^{(m)}(x) = \max\{g_k(x) : 1 \le k \le m\}, \quad x \in \mathbb{R}^n \tag{1}$$

with $g_k : \mathbb{R}^n \to \mathbb{R}$ a C^2 smooth function for $k = 1, 2, \dots, m$. Our approximations are based on the upper compensated convex transform $C^u_\lambda(f_m)(y)$ (see [30, 31, 32, 33, 34, 35, 36]) of the maximum function

$$f_m(y) := \max\{y_j : 1 \le j \le m\} \tag{2}$$

with $y = (y_1, y_2, \dots, y_m) \in \mathbb{R}^m$.

K.Y. is partially supported by National Science Foundation of China (NSFC) through grants 11801200 and 12141107. This work was support in part by UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/K032208/1.

ISSN 0944-6532 / \$ 2.50 © Heldermann Verlag

We define our smooth approximations for $g^{(m)}(\cdot)$ by the following composite function

$$g_\lambda^{(m)}(x) = C_\lambda^u(f_m)(G_m(x)), \quad x \in \mathbb{R}^n \tag{3}$$

where the vector-valued function $G_m : \mathbb{R}^n \to \mathbb{R}^m$ is defined by

$$G_m(x) = (g_1(x), g_2(x), \ldots, g_m(x))$$

for $x \in \mathbb{R}^n$. It turns out that $g_\lambda^{(m)}(\cdot)$ is a convexity preserving smooth approximation in the sense that if every g_k is smooth and convex in $\mathbb{R}^n$, then the approximation $g_\lambda^{(m)}(\cdot)$ is a convex, locally $C^{1,1}$ and 'tight' approximation of $g^{(m)}(\cdot)$ in the sense that if $g^{(m)}(x) = g_1(x)$ and $g_1(x) \geq \max_{2\leq j\leq m} g_j(x) + 1/(2\lambda)$, then $g_\lambda^{(m)}(x) = g^{(m)}(x)$.

Since many optimization problems are non-smooth, for which gradient based methods do not apply directly, various smoothing methods have been introduced to approximate the original non-smooth functions (see e.g., [2, 3, 8, 9, 16, 20, 29] and references therein). The widely used finite maximum function $f_m(x) = \max_{1\leq j\leq m} x_j$ for $x = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$ is one of the basic non-smooth functions arising from optimization problems. The study of smooth approximations of the finite maximum function can be traced back at least to Bertsekas' work in 1975 [3] and the follow-up work by Zang [29] using smooth approximations of the function $f_+(t) = \max\{0, t\}$ for $t \in \mathbb{R}$ (also see [8]). The methods used in [3, 8, 29] are very involved if the dimension m is large. A widely used approach is to use the 'log-sum-exp' formula ([2, 9, 26, 27], see also (13) below). The log-sum-exp approximation is not tight and the error estimate depends on the dimension. These make it difficult for us to keep the balance between the descretisation of the function and the choice of the approximation parameter in, say, semi-infinite minimax problems [19, 22, 23, 25, 28]. Therefore it is natural for us to seek alternative, more accurate, tight and simple approximations for which gradient based methods such as those introduced by Nesterov [20, 21] and generalized Newton's methods apply.

The standard unconstrained finite minimax problem is in the form

$$\min_{x\in\mathbb{R}^n} \max_{1\leq j\leq m} g_j(x),$$

where g_j's are smooth functions [20, 22, 25, 26, 27]. Usually finite minimax problems are non-smooth. One approach for solving such problems is to approximate the maximum function by the 'log-sum-exp' smooth approximation [22, 25, 26, 27]. The semi-infinite minimax problem is to solve $\min_{x\in\mathbb{R}^n} \max_{y\in K} f(x, y)$, where $K \subset \mathbb{R}^m$ is a compact set. In many practical examples of semi-infinite minimax problems in optimal engineering design, $K \subset \mathbb{R}$ is an interval on the real line. By discretisation, this type of semi-infinite minimax problems are approximated by finite minimax problems [22, 25]. We will compare our method with the 'log-sum-exp' smooth approximation in [22, 26, 27] in Section 4.

Another type of basic problems in optimization is the feasibility problem (and more generally the split feasibility problem). In many practical situations, the feasibility problem is to find a point in the set $\{x \in \mathbb{R}^n : \ g_j(x) \leq 0, \ j = 1, 2, \ldots m\}$, the intersection of the sublevel sets of constraint functions $g_j : \mathbb{R}^n \to \mathbb{R}$, $j = 1, 2, \ldots, m$.

Many popular methods are based on alternating projections which are particularly successful for linear constraints. The linear methods are also generalised to non-linear problems using projections to subdifferentials [1, 5, 6, 7, 10, 12, 13]. One may view the feasibility problem as a finite minimax problem for which we consider $\min_{x\in\mathbb{R}^n} g_+^{(m)}(x)$, where $g_+(x) = \max\{0, g_1(x), \dots, g_m(x)\}$. We can also approximate $g_+(\cdot)$ directly using a more involved upper compensated convex transform formula than (10) below. We will study tight smooth approximations of $g_+(\cdot)$ elsewhere. Instead, in this paper our approaches are either to use a gradient based method to minimise the smooth approximation $g_\lambda^{(m)}(\cdot)$, at each iteration we compare the value $g_\lambda^{(m)}(x_n)$ against 0, similar to some projection based methods, or to apply the upper compensated convex transform to the $(m+1)$-dimensional maximum function at $(x_1, x_2, \dots, x_m, x_{m+1})$, then to set the last component to be zero, that is, to set $x_{m+1} = 0$.

Next we describe our smooth approximation tools. The notions of compensated convex transforms were introduced in [30]. Approximations by the upper compensated convex transform of maximum-like functions were studied in [30, 31]. General properties of compensated convex transforms and applications to singularity extraction and to approximation and interpolation for sampled functions can be found in [30, 31, 32, 33, 34, 35, 36].

Suppose $f : \mathbb{R}^n \mapsto \mathbb{R}$ satisfies

$$f(x) \geq -C_1 - C_2|x|^2, \quad x \in \mathbb{R}^n, \tag{4}$$

where C_1, $C_2 > 0$ are constants. Then the (quadratic) lower compensated convex transform [30] (lower transform for short) is defined for each $\lambda > C_2$ by

$$C_\lambda^l(f)(x) = \mathrm{co}[\lambda|\cdot|^2 + f](x) - \lambda|x|^2, \quad x \in \mathbb{R}^n, \tag{5}$$

where $|x|$ is the standard Euclidean norm of $x \in \mathbb{R}^n$ and $\mathrm{co}[g]$ denotes the convex envelope [14, 24] of $g : \mathbb{R}^n \mapsto \mathbb{R}$ that is bounded below.

Similarly, if f satisfies

$$f(x) \leq C_1 + C_2|x|^2, \quad x \in \mathbb{R}^n, \tag{6}$$

then the (quadratic) upper compensated convex transform [30] (upper transform for short) is defined for each $\lambda > C_2$ by

$$C_\lambda^u(f)(x) = \lambda|x|^2 - \mathrm{co}[\lambda|\cdot|^2 - f](x), \quad x \in \mathbb{R}^n. \tag{7}$$

If f satisfies the quadratic growth condition

$$|f(x)| \leq C_1 + C_2|x|^2, \quad x \in \mathbb{R}^n, \tag{8}$$

then both the lower and the upper transforms can be defined. Furthermore, in this case we have

$$C_\lambda^u(f)(x) = -C_\lambda^l(-f)(x), \quad x \in \mathbb{R}^n. \tag{9}$$

In this paper we focus on the approximation of the maximum function $f_m(\cdot)$ by its upper transform $C_\lambda^u(f_m)(\cdot)$ given by (10) below. It was shown in [30, 31] that the upper convex transform $C^u_\lambda(f_m)(y)$ of the maximum function $f_m(y)$ can be calculated explicitly for $\lambda > 0$ as

$$C^u_\lambda(f_m)(y) = \lambda|y|^2 - \lambda \operatorname{dist}^2\left(y,\, \frac{\Delta_m}{2\lambda}\right) + \frac{1}{4\lambda}, \quad y \in \mathbb{R}^m, \tag{10}$$

where $|y|$ is the Euclidean norm of $y \in \mathbb{R}^m$, $\Delta_m = \mathrm{co}(K_m)$ is the $(m-1)$-dimensional simplex defined by the convex hull of the standard basis $K_m = \{e_1, e_2, \ldots, e_m\}$ for $\mathbb{R}^m$ and $\operatorname{dist}^2(y,\, K)$ is the Euclidean squared-distance function to a closed set $K \subseteq \mathbb{R}^m$. Later we also define $K_j = \{e_1,\, e_2,\, \ldots,\, e_j\} \subset \mathbb{R}^m$ as the first j elements of K_m $(1 \le j \le m)$ and denote by $\Delta_j = \mathrm{co}(K_j) \subset \mathbb{R}^m$ the convex hull of $K_j \subset K_m$.

If we denote by $P_{\Delta_m/(2\lambda)}(y)$ the convex projection of y to the scaled simplex $\Delta_m/(2\lambda)$ [14, 24], then we see that $C^u_\lambda(f_m)(y)$ is of linear growth in $y \in \mathbb{R}^m$ and is $C^{1,1}$ with Lipschitz constant of the gradient bounded by 2λ. We have

$$\begin{aligned} C^u_\lambda(f_m)(y) &= \lambda|y|^2 - \lambda|y - P_{\Delta_m/(2\lambda)}(y)|^2 + \frac{1}{4\lambda} \\ &= 2\lambda y \cdot P_{\Delta_m/(2\lambda)}(y) - \lambda|P_{\Delta_m/(2\lambda)}(y)|^2 + \frac{1}{4\lambda}, \end{aligned}$$

so that

$$|C^u_\lambda(f_m)(y)| \le 2\lambda|y||P_{\Delta_m/(2\lambda)}(y)| + \lambda|P_{\Delta_m/(2\lambda)}(y)|^2 + \frac{1}{4\lambda} \le |y| + \frac{1}{2\lambda},$$

as $|P_{\Delta_m/(2\lambda)}(y)| \le 1/(2\lambda)$. We also have

$$\nabla C^u_\lambda(f_m)(y) = 2\lambda y - 2\lambda(y - P_{\Delta_m/(2\lambda)}(y)) = 2\lambda P_{\Delta_m/(2\lambda)}(y), \tag{11}$$

so that
$$|\nabla C^u_\lambda(f_m)(y) - \nabla C^u_\lambda(f_m)(z)| \le 2\lambda|y - z|$$

for $y,\, z \in \mathbb{R}^m$. Here we have used the fact [14] that the convex projection to compact convex set $K \subset \mathbb{R}^m$ is a non-expanding map, that is, $|P_K(y) - P_K(z)| \le |y - z|$.

$f_m(\cdot)$ is a piecewise affine function. Geometrically, the approximation $C^u_\lambda(f_m)(\cdot)$ is a piecewise affine-quadratic function which is built by smoothing the the edges and corners of the graph of f_m by automatically generated quadratic pieces. More precisely, the affine pieces of $C^u_\lambda(f_m)(\cdot)$ coincide with the graph of $f_m(\cdot)$, and the quadratic pieces lie above $f_m(\cdot)$ around the points where $f_m(\cdot)$ is non-differentiable.

In general upper transforms for convex functions are convex [30]. Also compensated convex transforms are monotonicity preserving (see [30] Theorem 2.2). A function $f : \mathbb{R}^m \to \mathbb{R}$ is called a non-decreasing function if for every $y \in \mathbb{R}^m$ and $h = (h_1, h_2, \ldots, h_m) \in \mathbb{R}^m$ with $h_k \ge 0$ for $k = 1, 2, \ldots, m$, we have $f(y + h) \ge f(y)$. It is easy to verify that f_m is a non-decreasing function. Therefore $C^u_\lambda(f_m)$ is a non-decreasing convex function.

A classical approach to smooth approximations of the maximum function is based on a composition formula introduced by Bertsekas [3] given by

$$f_m(x) = x_1 + q[x_2 - x_1 + q[\cdots + q[x_{m-1} - x_{m-2} + q[x_m - x_{m-1}]]\cdots]], \tag{12}$$

where $q(t) = \max\{0,\, t\}$ for $t \in \mathbb{R}$.

The focus then was to design smooth approximations of $q(\cdot)$ (see [3, 8, 29]. Theoretically, approximations based on the smoothing of $q(\cdot)$ can be made 'tight' [29]. However, computations and analysis of such approximations for large dimension is a challenge.

There is a more widely used smooth approximation - the 'log-sum-exp' smoothing (also called *exponential penalty function* [26]) or *entropic regularization* (see, e.g.,[18]) for $\lambda > 0$ given by

$$H_\lambda^{(m)}(y) = \frac{1}{\lambda}\log\left(\sum_{j=1}^{m} e^{\lambda y_j}\right), \quad y = (y_1,\, y_2,\, \ldots,\, y_m) \in \mathbb{R}^m. \tag{13}$$

For C^∞ functions $g_j : \mathbb{R}^n \to \mathbb{R}$, $j = 1, 2, \ldots, m$, the corresponding C^∞-smooth approximation of $g^{(m)}(x) = \max\{g_j(x) :\ j = 1, 2, \ldots, m\}$ is defined by $H_\lambda^{(m)}(G_m(x))$. We will compare this approximation with $g_\lambda^{(m)}(x) := C_\lambda^u(f_m)(G_m(x))$ defined by the upper transform in Section 4, showing that $C_\lambda^u(f_m)$ is a point-wise sharper approximation of f_m than $H_{4\lambda}^{(m)}$. Moreover, in the average sense, $C_\lambda^u(f_m)$ approaches f_m asymptotically faster than $H_{\alpha\lambda\log\lambda}^{(m)}$ with $0 < \alpha < 2$.

Before we present the plan of the rest of the paper, let us clarify the terms 'blending' and 'tightness' of approximations for functions. 'Blending' is an operation in computer-aided geometric design (CAGD) [15]) which smoothes edges of a geometric object in a neighbourhood of singularities and keep the rest of the object intact. The result of such an operation is called a 'blend'. Compensated convex transforms are 'tight' approximations for general functions $f : \mathbb{R}^n \to \mathbb{R}$ with quadratic growth in the sense that if in a neighbourhood of a point $x_0 \in \mathbb{R}^m$, $f(\cdot)$ is $C^{1,1}$, then for $\lambda > 0$ sufficiently large, we have $C_\lambda^u(f)(x_0) = f(x_0)$ [30]. Thus we see that the locally $C^{1,1}$ approximation $C_\lambda^u(f)$ of a convex function with quadratic growth [30] can be viewed as a simple blend of the graph of f from above. It was noted in [29] that the approximations of the maximum function obtained by repeatedly using the smoothing $f_+(t) = \{0,\, t\}$ can be made tight, hence can be a smooth blending operation.

The plan of the rest of the paper is as follows. In Section 2, we derive further fine properties of the upper transform $C_\lambda^u(f_m)$ of the maximum function f_m. Without loss of generality, for $x = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$, we assume that $x_1 \geq x_2 \geq \cdots \geq x_m$ by a 'sorting' process. Our main result is on the geometric structure of the upper transform $C_\lambda^u(f_m)(x)$ which is reduced to the geometric structure of the convex projection $P_{\frac{\Delta_m}{2\lambda}}x$. In Theorem 2.2, we give a necessary and sufficient condition for the convex projection $P_{\frac{\Delta_m}{2\lambda}}x$ to lie in the relative interior of a face of $\frac{\Delta_m}{2\lambda}$, which is a $(k-1)$-dimensional simplex. As a result we obtain the tightness (or blending) property of $C_\lambda^u(f_m)$ that $C_\lambda^u(f_m)(x) = f_m(x)$ if and only if $x_1 - x_2 \geq 1/(2\lambda)$. As a corollary of the geometric structure of the convex projection $P_{\frac{\Delta_m}{2\lambda}}x$, we derive the sharp error estimate

$$0 \leq C_\lambda^u(f_m)(x) - f_m(x) \leq \frac{m-1}{4m\lambda}, \quad x \in \mathbb{R}^m.$$

We also give a sufficient condition for $C_\lambda^u(f_m)(\cdot)$ to be twice differentiable at a given point with its Hessian in the simple form given by (27).

Then, in Section 2 we establish the asymptotic properties for $C^u_\lambda(f_m)(x)$, $\nabla C^u_\lambda(f_m)(x)$ and $\nabla^2 C^u_\lambda(f_m)(x)$ respectively at each fixed $x \in \mathbb{R}^m$ when $\lambda > 0$ is sufficiently large but finite.

In Section 3 we study the composite function

$$g^{(m)}(x) = f_m(G_m(x)) = \max\{g_j(x) : \; j = 1, 2, \dots, m\}$$

with $g_j : \mathbb{R}^n \to \mathbb{R}$ smooth functions, where $G_m(x) := (g_1(x), \dots, g_m(x))$. We show that if every g_j is convex and locally $C^{1,1}$ in $\mathbb{R}^n$, e.g., $g_j \in C^2(\mathbb{R}^n)$, then $C^u_\lambda(f_m)(G_m(\cdot))$ is also convex and locally $C^{1,1}$. Also for C^2-functions g_j for $j = 1, 2, \dots, m$, at a given point $x \in \mathbb{R}^n$, after a 'sorting' operation we may assume that $g_1(x) \geq g_2(x) \geq \cdots \geq g_m(x)$. If $P_{\Delta_m/(2\lambda)} G_m(x)$ is a relative interior point of some $\Delta_k/(2\lambda)$ with $k \leq m$ and certain non-degenerate conditions are satisfied, then $C^u_\lambda(f_m)(G_m(\cdot))$ is twice differentiable at x with its Hessian given by (30) which only involves the functions g_j for $1 \leq j \leq k$, their gradients and Hessians. This is quite different from the 'log-sum-exp' approximation $H^{(m)}_\lambda(G_m(x))$ whose Hessian depends on all g_j's, their gradients and Hessians at any given $x \in \mathbb{R}^m$. In case $g_j(y) = \ell_j(y) = a_j \cdot y + b_j$ are affine functions, the Hessian of $C^u_\lambda(f_m)(G_m(x))$ is given by the simple formula (see (31))

$$\nabla^2 C^u_\lambda(f_m)(G_m(x)) = 2\lambda \left[\sum_{i=1}^k a_i \otimes a_i - \frac{1}{k} \left(\sum_{j=1}^k a_j \right) \otimes \left(\sum_{s=1}^k a_s \right) \right].$$

We also present examples for some commonly used convex functions in applications such as the semi-infinite minimax problems and its discretization.

In Section 4, we compare properties and asymptotic behavior of the smooth approximation $C^u_\lambda(f_m)(\cdot)$ with those obtained by the 'log-sum-exp' approximation given by $H^{(m)}_\mu(\cdot)$. We establish a point-wise comparison in Theorem 4.1 saying that

$$f_m(y) \leq C^u_\lambda(f_m)(y) < H^{(m)}_{4\lambda}(y).$$

We also show in Theorem 4.4 that in the average sense, $C^u_\lambda(f_m)(\cdot)$ approaches $f_m(\cdot)$ faster than $H^{(m)}_{\alpha\lambda\log\lambda}(\cdot)$ with $0 < \alpha < 2$. We show that for any closed ball $\bar{B}_r(0) \subseteq \mathbb{R}^m$ centred at 0 with radius $r > 0$,

$$\lim_{\lambda \to +\infty} \int_{\bar{B}_r(0)} \frac{C^u_\lambda(f_m)(y) - f_m(y)}{H^{(m)}_{\alpha\lambda\log\lambda}(y) - f_m(y)} \, dy = 0.$$

Under the strong convexity assumption that $y^T \nabla^2 g_j(x) y \geq c|y|^2$ for all $j = 1, 2, ..., m$ and for all x, $y \in \mathbb{R}^n$ with $c > 0$ a constant, we give estimates between the minimiser of the original function $g^{(m)}(\cdot)$ with the minimisers obtained by our approximation $C^u_\lambda(f_m)(G_m(\cdot))$ and those obtained by using the 'log-sum-exp' smoothing respectively.

In the Appendix we present proofs of two technical results needed in Section 4.

In this paper we denote by $|a|$ and $a \cdot b$ the standard Euclidean norm of $a \in \mathbb{R}^m$ and the inner product of a, $b \in \mathbb{R}^m$ for some $m \geq 1$. We also denote by $\mathrm{co}(f)$ the convex envelope of a function $f : \mathbb{R}^m \to \mathbb{R}$ which is bounded below and $\mathrm{co}(K)$ the convex hull of a bounded closed set $K \subset \mathbb{R}^m$ [14]. For $t > 0$ we define $K/t := \{x/t : \; x \in K\}$. It is easy to verify for a compact set K that $\mathrm{co}(K/t) = \mathrm{co}(K)/t$.

2. The upper transform of the maximum function

In this section some fine properties of the upper transform $C^u_\lambda(f_m)$ of the maximum function f_m are studied. We present the precise tightness (blending) properties, the sharp error estimate, the geometric structure of $C^u_\lambda(f_m)$ based on the convex projection $P_{\Delta_m/(2\lambda)}x$ from $\mathbb{R}^m$ to the simplex $\Delta_m/(2\lambda)$ for a fixed $\lambda > 0$ and the asymptotic properties for $C^u_\lambda(f_m)(x)$, $\nabla C^u_\lambda(f_m)(x)$ and $\nabla^2 C^u_\lambda(f_m)(x)$ respectively for fixed $x \in \mathbb{R}^m$ and for sufficiently large $\lambda > 0$. In order to prove our tightness (blending) result for $C^u_\lambda(f_m)$ (Corollary 2.3), we need the following explicitly calculated upper transform for $f_2(x)$ in $\mathbb{R}^2$ which can be easily derived from its definition.

Lemma 2.1. *Let $f_2(x) = \max\{x_1, x_2\}$ for $x = (x_1, x_2) \in \mathbb{R}^2$. We have*

$$\begin{aligned} C^u_\lambda(f_2)(x) &= \lambda|x|^2 - \lambda \operatorname{dist}^2(x,\, \Delta_2/(2\lambda)) + \frac{1}{4\lambda} \\ &= \begin{cases} \dfrac{x_1+x_2}{2} + \dfrac{\lambda}{2}(x_1-x_2)^2 + \dfrac{1}{8\lambda} & |x_1-x_2| \le \dfrac{1}{2\lambda}, \\ f_2(x) & |x_1-x_2| \ge \dfrac{1}{2\lambda}. \end{cases} \end{aligned} \tag{14}$$

Furthermore, we have

$$\nabla C^u_\lambda(f_2)(x) = \begin{cases} \left(\dfrac{1}{2} + \lambda(x_1-x_2),\ \dfrac{1}{2} - \lambda(x_1-x_2)\right) & |x_1-x_2| \le \dfrac{1}{2\lambda}, \\ (0,\, 1) & x_2 \ge x_1 + \dfrac{1}{2\lambda}, \\ (1,\, 0) & x_1 \ge x_2 + \dfrac{1}{2\lambda}; \end{cases}$$

$$\nabla^2 C^u_\lambda(f_2)(x) = \begin{cases} \lambda \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} & |x_1-x_2| < \dfrac{1}{2\lambda}, \\ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} & |x_1-x_2| > \dfrac{1}{2\lambda}. \end{cases}$$

It is known that $C^u_\lambda(f_m)(x)$ is a globally $C^{1,1}$ and convex function [30]. Furthermore, we have $f_m(x) \le C^u_\lambda(f_m)(x)$ for all $x \in \mathbb{R}^m$. By definition, we have, for example,

$$\operatorname{co}(\lambda|\cdot|^2 - f_m(\cdot))(x) \le \lambda|x|^2 - f_m(x), \tag{15}$$

for all $x \in \mathbb{R}^m$. If we have $C^u_\lambda(f_m)(x^{(0)}) = f_m(x^{(0)})$, then by a lemma in [17], both $C^u_\lambda(f_m)$ and f_m are differentiable at $x^{(0)}$ and $\nabla C^u_\lambda(f_m)(x^{(0)}) = \nabla f_m(x^{(0)})$ as the right hand side of (15) is superdifferentiable and the left hand side is convex [17].

Next we focus on the geometric structure of the convex projection $P_{\Delta_m/(2\lambda)}$ which leads to the precise 'tightness' property of the approximation $C^u_\lambda(f_m)$ to f_m.

Observe that an equivalent statement to that a point $y \in \mathbb{R}^m$ satisfying $C^u_\lambda(f_m)(y) = f_m(y)$ is that the convex projection $P_{\Delta_m/(2\lambda)}(y) = e_j/(2\lambda)$ for some $1 \le j \le m$. Therefore we see that $f_m(x) < C^u_\lambda(f_m)(x)$ if and only if $P_{\Delta_m/(2\lambda)}x$ lies in the relative interior of a $(k-1)$-dimensional face of the simplex $\Delta_m/(2\lambda)$ for some $2 \le k \le m$ with k the largest among all such faces. Thus

$$P_{\frac{\Delta_m}{2\lambda}}x = \sum_{i=1}^k \lambda_i \frac{e_{j_i}}{2\lambda}, \quad \lambda_i > 0, \quad \sum_{i=1}^k \lambda_i = 1, \quad 1 \le i \le k.$$

Without loss of generality and for simplicity of the notation, we may assume that $x_1 \geq x_2 \geq \cdots \geq x_m$. Note that the simplex Δ_m is the disjoint union of all of the relative interiors of its faces and we may also call the vertices as their relative interiors of these zero-dimensional faces. We have

Theorem 2.2. *Suppose $x_1 \geq x_2 \geq \cdots \geq x_m$. Then the convex projection $P_{\frac{\Delta_m}{2\lambda}}x$ from x to $\Delta_m/(2\lambda)$ is in the relative interior of the $(k-1)$-dimensional face $\Delta^{(k)}/(2\lambda)$ with $1 \leq k \leq m$, where*

$$\frac{\Delta^{(k)}}{2\lambda} = \mathrm{co}\left[\frac{K^{(k)}}{2\lambda}\right] \quad \textit{with} \quad K^{(k)} = \{e_{j_1},\, e_{j_2},\, \ldots,\, e_{j_k}\} \subset K_m \tag{16}$$

if and only if $k \in \{1, 2, \ldots, m\}$ is the largest integer such that

$$\begin{aligned}
&\text{(i)} \qquad 0 < x_j - \frac{1}{k}\sum_{i=1}^{k} x_i + \frac{1}{2k\lambda} \leq \frac{1}{2\lambda}, \quad j = 1,\, 2,\, \ldots,\, k;\\
&\text{(ii)} \qquad x_s - \frac{1}{k}\sum_{i=1}^{k} x_i + \frac{1}{2k\lambda} \leq 0, \quad s = k+1,\, k+2,\, \ldots,\, m.
\end{aligned} \tag{17}$$

Furthermore we have

$$P_{\frac{\Delta_m}{2\lambda}}x = P_{\frac{\Delta_k}{2\lambda}}x = \sum_{j=1}^{k}\left(x_j - \frac{1}{k}\sum_{i=1}^{k} x_i + \frac{1}{2k\lambda}\right) e_j, \tag{18}$$

and $\dfrac{\Delta^{(k)}}{2\lambda} = \dfrac{\Delta_k}{2\lambda} := \mathrm{co}\left[\dfrac{K_k}{2\lambda}\right]$, *with* $K_k = \{e_1,\, e_2,\, \ldots,\, e_k\} \subset K_m$.
As a consequence, if $x_1 - x_s \geq \frac{1}{2\lambda}$, *then* $s > k$.

Proof. Suppose $P_{\frac{\Delta_m}{2\lambda}}x = P_{\frac{\Delta^{(k)}}{2\lambda}}x$. Clearly

$$P_{\frac{\Delta^{(k)}}{2\lambda}}x = \sum_{j=1}^{k} \lambda_j e_{i_j},\ \Delta^{(k)}/(2\lambda) = \mathrm{co}\left[\frac{K^{(k)}}{2\lambda}\right]$$

where $K^{(k)} = \{e_{i_1},\, e_{i_2}, \ldots, e_{i_k}\} \subset K_m$, and without loss of generality, we may assume

$$0 < \lambda_k \leq \lambda_{k-1} \leq \cdots \leq \lambda_1 \leq \frac{1}{2\lambda}, \qquad \sum_{i=1}^{k} \lambda_i = \frac{1}{2\lambda}. \tag{19}$$

On the other hand, the $k-1$ dimensional simplex $\Delta^{(k)}/(2\lambda)$ is contained in the subspace $E^{(k)}$ spanned by $\{e_{i_j}\}_{j=1}^{k}$. Since $P_{\Delta_m/(2\lambda)}x$ is the nearest point to x belonging to the relative interior of $\Delta^{(k)}/(2\lambda)$, it is the orthogonal projection of x to the affine plane $L^{(k)} \subset E^{(k)}$ spanned by $\{e_{i_j}\}_{j=1}^{k}$. Let $\hat{e} = \frac{1}{2k\lambda}\sum_{j=1}^{k} e_{i_j} \in \Delta^{(k)}/(2\lambda)$, then $\hat{e}$ is perpendicular to $L^{(k)}$. If we let $P_{E^{(k)}}x$ be the orthogonal projection of x to $E^{(k)}$, we have

$$P_{\frac{\Delta^{(k)}}{2\lambda}}x = P_{E^{(k)}}x - (\hat{e} \cdot P_{E^{(k)}}x)\frac{\hat{e}}{|\hat{e}|^2} + \hat{e} = \sum_{j=1}^{k}\left(x_{i_j} - \frac{1}{k}\sum_{s=1}^{k} x_{i_s} + \frac{1}{2k\lambda}\right) e_{i_j}.$$

Thus we have $$0 < \lambda_j = x_{i_j} - \frac{1}{k}\sum_{s=1}^{k} x_{i_s} + \frac{1}{2k\lambda} \leq \frac{1}{2\lambda}.$$

By (19) we see that $x_{i_1} \geq x_{i_2} \geq \cdots \geq x_{i_k}$ which implies that $i_1 < i_2 < \cdots < i_k$ if we arrange components which are equal by increasing subscripts if necessary.

Since $\mathrm{co}[\mathrm{dist}^2(\cdot,\, K_m/(2\lambda))] = \mathrm{dist}^2(\cdot,\, \Delta_m/(2\lambda))$ (see [30]), by the definition of the upper transform of $C^u_\lambda(f_m)(\cdot)$, we have

$$\mathrm{co}\left[\lambda|\cdot|^2 - f_m\right](y) = \lambda\,\mathrm{dist}^2\left(y,\, \frac{\Delta_m}{2\lambda}\right) - \frac{1}{4\lambda} \leq \lambda|y|^2 - f_m(y) \leq \lambda|y|^2 - y_s \tag{20}$$

for all $y \in \mathbb{R}^m$ and for every fixed $s = 1, 2, \ldots, m$.

At x above, (20) implies that the tangent plane of $\mathrm{co}\left[\lambda|\cdot|^2 - f_m\right](y)$ at $y = x$ lies below the graph of the function $y \to \lambda|y|^2 - y_s$. As

$$\mathrm{dist}^2\left(x,\, \frac{\Delta_m}{2\lambda}\right) = \left|x - P_{\frac{\Delta^{(k)}}{2\lambda}} x\right|^2,$$

we have $$\lambda\left|x - P_{\frac{\Delta^{(k)}}{2\lambda}} x\right|^2 - \frac{1}{4\lambda} + 2\lambda\left(x - P_{\frac{\Delta^{(k)}}{2\lambda}} x\right)\cdot(y - x) \leq \lambda|y|^2 - y_s, \tag{21}$$

for all $y \in \mathbb{R}^m$ and for every fixed $s \in \{1, 2, \ldots, m\}$.

We move the left hand side of this inequality to the right hand side and define

$$q_s(y) = \lambda|y|^2 - y_s - \lambda\left|x - P_{\frac{\Delta^{(k)}}{2\lambda}} x\right|^2 + \frac{1}{4\lambda} - 2\lambda\left(x - P_{\frac{\Delta^{(k)}}{2\lambda}} x\right)\cdot(y - x), \quad y \in \mathbb{R}^m$$

and require that its minimum is non-negative. We consider two different cases. If $s \in \{i_1,\, i_2,\, \ldots,\, i_k\}$, by solving $\nabla q_s(y) = 0$ we see that the minimum value is 0. So (17) holds in this case.

If $s \notin \{i_1,\, i_2,\, ...,\, i_k\}$, we see that the minimum point $y^* = (y_1^*, y_2^*, ..., y_m^*)$ is given by

$$\begin{cases} y_\ell^* = x_\ell & \ell \notin \{s,\, i_1,\, i_2,\, \ldots,\, i_k\}; \\ y_{i_j}^* = \dfrac{1}{k}\displaystyle\sum_{t=1}^{k} x_{i_t} - \dfrac{1}{2k\lambda} & j = 1, 2, \ldots, k; \\ y_s^* = \dfrac{1}{2\lambda} + \dfrac{1}{k}\displaystyle\sum_{t=1}^{k} x_{i_t} - \dfrac{1}{2k\lambda}. & \end{cases}$$

The minimum value satisfies $q_s(y^*) \geq 0$ if and only if

$$x_s - \frac{1}{k}\sum_{t=1}^{k} x_{i_t} + \frac{1}{2k\lambda} \leq 0.$$

Thus we have $x_{i_j} = x_j$, $x_1 \geq x_2 \geq \cdots x_k > x_{k+1}$, $\Delta^{(k)} = \Delta_k$ and both (i) and (ii) in Theorem 2.2 are satisfied. Now we claim that $\Delta_k/(2\lambda)$ has the largest dimension among faces $\Delta_\ell/(2\lambda)$ for which $P_{\Delta_m/(2\lambda)}(x)$ is a relative interior point of $\Delta_\ell/(2\lambda)$.

If this is not the case, we see that $P_{\Delta_m/(2\lambda)}(x)$ is also a relative interior point of some $(\ell-1)$-dimensional face $\Delta^{(\ell)}/(2\lambda)$ with $\ell > k$. By repeating the argument earlier, we see that $\Delta^{(\ell)}/(2\lambda) = \Delta_\ell/(2\lambda)$ with $\Delta_k/(2\lambda)$ lies on the relative boundary of $\Delta_\ell/(2\lambda)$. Thus

$$x - P_{\frac{\Delta_k}{2\lambda}}x = x - P_{\frac{\Delta_\ell}{2\lambda}}x + P_{\frac{\Delta_\ell}{2\lambda}}x - P_{\frac{\Delta_k}{2\lambda}}x, \text{ and } \; x - P_{\frac{\Delta_\ell}{2\lambda}}x \perp P_{\frac{\Delta_\ell}{2\lambda}}x - P_{\frac{\Delta_k}{2\lambda}}x.$$

As $P_{\frac{\Delta_\ell}{2\lambda}}x \neq P_{\frac{\Delta_k}{2\lambda}}x$, by Pythagoras Theorem, we have

$$\left|x - P_{\frac{\Delta_\ell}{2\lambda}}x\right| < \left|x - P_{\frac{\Delta_k}{2\lambda}}x\right|,$$

hence $P_{\frac{\Delta_k}{2\lambda}}x$ cannot be the nearest point of x in $\Delta_m/(2\lambda)$. This is a contradiction.

It is easy to see that the converse statement is also true as $P_{\frac{\Delta_k}{2\lambda}}x$ is given by (18) and k is the largest integer such that (i) and (ii) hold. Thus if $P_{\frac{\Delta_m}{2\lambda}}x = P_{\frac{\Delta^{(\ell)}}{2\lambda}}x$, then $\Delta^{(\ell)} = \Delta_\ell$ and $l = k$.

To show the last argument in the theorem. Suppose that $x_1 - x_s \geq \frac{1}{2\lambda}$ and $s \leq k$. Then by (i)

$$0 < x_s - \frac{1}{k}\sum_{i=1}^{n} x_i + \frac{1}{2k\lambda} \leq x_1 - \frac{1}{2\lambda} - \frac{1}{k}\sum_{i=1}^{n} x_i + \frac{1}{2k\lambda} \leq 0, \tag{22}$$

which is a contradiction. This completes the proof of the theorem. □

If the maximum face is a vertex $(k = 1)$, Theorem 2.2 provides a criterion for $C_\lambda^u(f_m)(x) = f_m(x)$. This is a tightness result for $C_\lambda^u(f_m)(\cdot)$ to approximate $f_m(\cdot)$. If we examine the proof of Theorem 2.2 for this case, we have the following necessary and sufficient condition for $C_\lambda^u(f_m)(x) = f_m(x)$.

Corollary 2.3. *Let $m \geq 2$ and suppose that $x = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$ satisfies $x_1 \geq x_2 \geq \cdots \geq x_m$. Then*

$$C_\lambda^u(f_m)(x) = f_m(x) \quad \textit{if and only if} \quad x_1 - x_2 \geq \frac{1}{2\lambda}.$$

Proof. If $x_1 - x_2 \geq 1/(2\lambda)$, then for any $j > 1$ we have

$$x_j - x_1 + 1/(2\lambda) \leq x_2 - x_1 + 1/(2\lambda) \leq 0.$$

Therefore items (i) and (ii) in Theorem 2.2 are satisfied for $k = 1$. Next we show that the zero dimensional face - the vertex $\{e_1/(2\lambda)\}$ is the face with largest dimension for which (i) and (ii) holds. Otherwise we have $k > 1$ such that both (i) and (ii) hold. If (i) holds, we have

$$0 < \frac{x_k - x_1}{k} - \frac{1}{k}\sum_{j=1}^{k} x_j + \frac{1}{2k\lambda} = \frac{x_k - x_1}{k} + \frac{1}{k}\sum_{j=2}^{k-1}(x_k - x_j) + \frac{1}{2k\lambda} \leq \frac{x_k - x_1}{k} + \frac{1}{2k\lambda} \leq 0,$$

which is a contradiction as we have assumed the conditions that $x_1 - x_2 \geq 1/(2\lambda)$ and $x_2 \geq x_3 \geq \cdots \geq x_m$.

If $f_m(x) = C_\lambda^u(f_m(x))$, we define $f^{(2)}(y) = \max\{y_1, y_2\}$ for $y = (y_1, y_2, \ldots, y_m)$.

Clearly $f^{(2)}(y) \leq f_m(y)$ for $y \in \mathbb{R}^m$. Thus $C_\lambda^u(f^{(2)})(y) \leq C_\lambda^u(f_m(y))$ for $y \in \mathbb{R}^m$.

At x we have $f^{(2)}(x) \leq C_\lambda^u(f^{(2)})(x) \leq C_\lambda^u(f_m(x)) = f_m(x) = f^{(2)}(x)$.

Thus $C^u_\lambda(f^{(2)})(x) = f^{(2)}(x)$. We also observe that $C^u_\lambda(f^{(2)})(y) = C^u_\lambda(f_2)(\hat{y})$ where $f_2(\cdot)$ is the two-dimensional maximum function and $\hat{y} := (y_1, y_2) \in \mathbb{R}^2$ for $y \in \mathbb{R}^m$. From the explicit formula (14) in Lemma 2.1, we see that $C^u_\lambda(f_2)(\hat{y}) = f_2(\hat{y})$ if and only if $|y_1 - y_2| \geq 1/(2\lambda)$. Thus at x we have $C^u_\lambda(f_2)(\hat{x}) = f_2(\hat{x})$ so that we obtain $|x_1 - x_2| \geq 1/(2\lambda)$. Since we have also assumed that $x_1 \geq x_2$, hence $x_1 - x_2 \geq 1/(2\lambda)$. □

Remark 2.4. The proof of Corollary 2.3 was first established in an unpublished work by us. Here we present a new proof based on Theorem 2.2. If we do not assume the ordering $x_1 \geq x_2 \geq \cdots \geq x_m$, then the second largest component of $x = (x_1, x_2, \ldots, x_m)$ can be represented by

$$\hat{f}_m(x) = \min_{1\leq i\leq m} \max\{x_j : \quad 1 \leq j \leq m,\ j \neq i\}, \quad x \in \mathbb{R}^m.$$

The tightness property stated in Corollary 2.3 for the case $k = 1$ is equivalent to

$$C^u_\lambda(f_m)(x) = f_m(x) \quad \text{if and only if} \quad f_m(x) - \hat{f}_m(x) \geq \frac{1}{2\lambda}.$$

Note that the set on which $f_m(x) = \hat{f}_m(x)$ is contained in the union of finitely many hyperplanes $V := \cup_{1\leq i<j\leq m} L_{i,j}$, where $L_{i,j} = \{x \in \mathbb{R}^m :\ x_i = x_j\}$. Hence for every point x in the open set $\mathbb{R}^m \setminus V$, there exists λ such that $f_m(x) \geq \hat{f}_m(x) + \frac{1}{2\lambda}$ and $C^u_\lambda(f_m)(x) = f_m(x)$. In this sense we may say that $C^u_\lambda(f_m)$ is a tight approximation of f_m.

Corollary 2.5. *Suppose $m \geq 2$. We have the following sharp error estimate*

$$0 \leq C^u_\lambda(f_m)(x) - f_m(x) \leq \frac{m-1}{4m\lambda}, \quad x \in \mathbb{R}^m. \tag{23}$$

Remark 2.6. We can easily see that for $m = 2$, as in Lemma 2.1 Equation (14), the constant $1/(8\lambda)$ agrees with that for the explicitly calculated upper transform $C^u_\lambda(f_2)(x)$. In the proof of Corollary 2.5 we also observe that if $P_{\frac{\Delta_m}{2\lambda}} x = P_{\frac{\Delta_k}{2\lambda}} x$, where $\Delta_k = \mathrm{co}(\{e_1,\, e_2,\, \ldots,\, e_k\})$, we have a slightly sharper estimate than (23) that

$$0 \leq C^u_\lambda(f_m)(x) - f_m(x) \leq \frac{k-1}{4k\lambda}.$$

When $k = 1$ we have $C^u_\lambda(f_m)(x) = f_m(x)$.

We need the following Lemma for the proof of Corollary 2.5.

Lemma 2.7. *Suppose $k \geq 2$, then*

$$\sup\left\{(1-\lambda_1)^2 + \sum_{j=2}^{k} \lambda_j^2 : \quad \lambda_1 \geq \lambda_2,\, \ldots,\, \lambda_k > 0,\ \sum_{j=1}^{k} \lambda_j = 1\right\} = 1 - \frac{1}{k}, \tag{24}$$

and the maximum is reached at $\lambda_1 = \lambda_2 = \cdots = \lambda_k = 1/k$.

Proof. If $k = 2$, the function to be minimised is $2(1-\lambda_1)^2 \leq 1/2$ as $\lambda_1 \geq \lambda_2$ and $\lambda_1 \geq 2$ must be satisfied at the same time.

Clearly we must have $1/k \le \lambda_1 < 1$. Now fix such a λ_1 and consider the maximum of

$$F(\lambda_2,\,\dots,\,\lambda_k) := (1-\lambda_1)^2+\sum_{j=2}^{k}\lambda_j^2 \quad \text{subject to} \quad 0<\lambda_j\le\lambda_1 \ \text{ and } \sum_{j=2}^{k}\lambda_j = 1-\lambda_1.$$

It is well-known that this convex maximization problem over a polygon reaches its maximum at a vertex point. If $\lambda_1 \ge 1/2$, then one of the maximum point is $\lambda_2 = 1-\lambda_1$ and $\lambda_j = 0$ for $3 \le j \le k$. In this case the maximum of $F(\cdot)$ is $2(1-\lambda_1)^2$. This quantity reaches its maximum when $\lambda_1 = 1/2$.

If $1/k \le \lambda_1 \le 1/2$, then $1-\lambda_1 = j\lambda_1 + r\lambda_1$ with $j \ge 1$ an integer and $0 \le r < 1$. In this case, after a 'sorting' operation, a vertex must be in the form

$$\lambda_2 = \cdots = \lambda_{j+1} = \lambda_1,\ \lambda_{j+2} = r\lambda_1 \text{ and } \lambda_i = 0 \text{ for } i > j+2.$$

Note that here we must have $j \le k-1$. For such a point we have

$$\begin{aligned} F(\lambda_2,\,\dots,\,\lambda_k) &= (1-\lambda_1)^2 + j\lambda_1^2 + r\lambda_1^2 = (1-\lambda_1)^2 + (1-\lambda_1)\lambda_1 \\ &= 1-\lambda_1 \le 1-\frac{1}{k}. \end{aligned}$$

The proof is complete as we have assumed $k > 2$. □

Proof of Corollary 2.5. Without loss of generality we may assume that we have $x_1 \ge x_2 \ge \cdots \ge x_m$. Suppose $P_{\Delta_m/(2\lambda)}x$ is contained in the relative interior of face $\Delta_k/(2\lambda)$ with the largest dimension. By Theorem 2.2, we can write

$$P_{\Delta_m/(2\lambda)}x = \sum_{j=1}^{k}\lambda_j e_j/(2\lambda)$$

with $\sum_{j=1}^{k}\lambda_j = 1$ and $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_k > 0$. Since $e_1/(2\lambda) \in \Delta_k/(2\lambda)$ and $P_{\Delta_m/(2\lambda)}x$ is the orthogonal projection to the plane spanned by $\Delta_k/(2\lambda)$, we have

$$x - P_{\frac{\Delta_m}{2\lambda}}x \perp \frac{e_1}{2\lambda} - P_{\frac{\Delta_m}{2\lambda}}x.$$

We also have

$$f_m(y) = \lambda|y|^2 - \lambda\,\mathrm{dist}^2\left(y,\,\frac{K_m}{2\lambda}\right) + \frac{1}{4\lambda}, \quad y \in \mathbb{R}^m$$

with $K_m = \{e_1,\, e_2,\, \dots,\, e_m\}$. Hence

$$\begin{aligned} &0 \le C_\lambda^u(f_m)(x) - f_m(x) = \lambda\left(\mathrm{dist}^2\left(x,\,\frac{K_m}{2\lambda}\right) - \mathrm{dist}^2\left(x,\,\frac{\Delta_m}{2\lambda}\right)\right) \\ &= \lambda\left(\left|x-\tfrac{e_1}{2\lambda}\right|^2 - \left|x - P_{\frac{\Delta_m}{2\lambda}}x\right|^2\right) = \lambda\left|\tfrac{e_1}{2\lambda} - \textstyle\sum_{j=1}^{k}\tfrac{\lambda_j}{2\lambda}e_j\right|^2 \\ &= \tfrac{1}{4\lambda}\left(1-\lambda_1)^2 + \textstyle\sum_{j=2}^{k}\lambda_j^2\right) \le \tfrac{k-1}{4k\lambda} \le \tfrac{m-1}{4m\lambda}. \end{aligned}$$

The last inequality above follows from Lemma 2.7 and the fact $m \ge k$. □

We need the following simple result for further calculations of the gradient and Hessian for $C_\lambda^u(f_m)(\cdot)$.

Corollary 2.8. *Suppose $m \geq 2$ and that the point $x = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$ satisfies $x_1 \geq x_2 \geq \cdots \geq x_m$. If $P_{\frac{\Delta_m}{2\lambda}}(x)$ is a relative interior point of face $\frac{\Delta_k}{2\lambda}$ for some $1 \leq k \leq m$ with the largest dimension $k-1$ and satisfies assumptions* (i) *and* (ii) *in Theorem* 2.2*, then*

$$\nabla C^u_\lambda(f_m)(x) = 2\lambda P_{\frac{\Delta_k}{2\lambda}}(x) = 2\lambda \sum_{j=1}^{k} \left(x_j + \frac{1}{2k\lambda} - \frac{1}{k}\sum_{i=1}^{k} x_i \right) e_j. \tag{25}$$

If we assume (i) *in Theorem* 2.2 *and*

$$\text{(ii)}^* \qquad x_{k+1} - \frac{1}{k}\sum_{i=1}^{k} x_i + \frac{1}{2k\lambda} < 0, \tag{26}$$

then $\nabla C^u_\lambda(f_m)(y) = 2\lambda P_{\frac{\Delta_k}{2\lambda}}(y)$ in a neighbourhood of x and $C^u_\lambda(f_m)(\cdot)$ is twice differentiable at x with the Hessian matrix given by

$$\nabla^2 C^u_\lambda(f_m)(x) = 2\lambda \begin{pmatrix} \mathbf{A}_k & 0 \\ 0 & 0 \end{pmatrix}, \qquad \text{where} \quad \mathbf{A}_k = \mathbf{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T, \tag{27}$$

with $\mathbf{I}_k$ the $k \times k$ identity matrix and $\mathbf{1}_k = (1, 1, \ldots, 1)^T \in \mathbb{R}^k$ is the column vector with entries 1*.*

If we assume (i) *in Theorem* 2.2 *and*

$$\text{(ii)}^{**} \qquad x_s - \frac{1}{k}\sum_{i=1}^{k} x_i + \frac{1}{2k\lambda} = 0, \; s = k+1, \ldots, k+r \quad (r \leq m-k),$$

then $C^u_\lambda(f_m)(\cdot)$ is not twice differentiable at x.

Proof. From (10) and Theorem 2.2 it is easy to see that

$$\nabla C^u_\lambda(f_m)(x) = 2\lambda P_{\frac{\Delta_m}{2\lambda}} x = 2\lambda \sum_{j=1}^{k} \left(x_j - \frac{1}{k}\sum_{i=1}^{k} x_i + \frac{1}{2k\lambda} \right) e_j.$$

If we further assume (ii)*, we see that for y in a small neighbourhood of x, (i) and (ii)* still hold, hence

$$\nabla C^u_\lambda(f_m)(y) = 2\lambda P_{\frac{\Delta_m}{2\lambda}} y = 2\lambda \sum_{j=1}^{k} \left(y_j - \frac{1}{k}\sum_{i=1}^{k} y_i + \frac{1}{2k\lambda} \right) e_j.$$

This affine mapping is clearly differentiable for y close to x. So the Hessian at x is given by (27). □

Remark 2.9. We observe that the Hessian matrix $\nabla C^u_\lambda(f_m)(\cdot)$ is piecewise linear and absolutely continuous, so $\nabla^2 C^u_\lambda(f_m)(\cdot)$ exists almost every where. However, $\nabla^2 C^u_\lambda(f_m)(\cdot)$ is not invertible, with eigenvalues 2λ with multiplicity $k-1$ and 0 with multiplicity $m-k+1$. All non-zero vectors perpendicular to $\mathbf{1}_k$ are eigenvectors corresponding to the eigenvalue 2λ.

Next we study the asymptotic properties of $C^u_\lambda(f_m)$ and $\nabla C^u_\lambda(f_m)$ when $\lambda > 0$ is sufficiently large but finite. Such properties will be used later for our smooth approximations.

Definition 2.10. Let $x = (x_1, x_2, \dots, x_m) \in \mathbb{R}^m$ satisfy $x_1 \geq x_2 \geq \cdots \geq x_m$. We say that x is a level -$(k-1)$ singular point of f_m if $x_1 = x_2 = \cdots = x_k > x_{k+1}$. A point $x \in \mathbb{R}^m$ is called a regular point of f_m if x is a level-0 singular point, that is, $k = 1$. In this case we assume $x_1 > x_2$. A level-$(m-1)$ singular point x satisfies $x_1 = x_2 = \cdots = x_m$.

Note that in general we can always re-arrange components of a vector by 'sorting' so that they are in non-increasing order.

Theorem 2.11. *Assume that* $x = (x_1, x_2, ..., x_m) \in \mathbb{R}^m$ *satisfies* $x_1 \geq x_2 \geq ... \geq x_m$.

(i) *If* $x \in \mathbb{R}^m$ *is a level-*$(k-1)$ *singular point of* f_m *for some* $2 \leq k \leq m$, *then for* $\lambda > 0$ *sufficiently large,* $\nabla^2 C^u_\lambda(f_m)(x)$ *exists and*

$$\begin{aligned} &C^u_\lambda(f_m)(x) - f_m(x) = \frac{k-1}{4k\lambda}, \quad \nabla C^u_\lambda(f_m)(x) = \frac{1}{k}\sum_{i=1}^{k} e_i, \\ &\nabla^2 C^u_\lambda(f_m)(x) = 2\lambda \begin{pmatrix} \mathbf{A}_k & 0 \\ 0 & 0 \end{pmatrix}, \quad \textit{where } \mathbf{A}_k \textit{ is given by } (27). \end{aligned} \tag{28}$$

(ii) *If* y *is a regular point of* f_m, *that is,* y *is a level-*0 *singular point in the sense that* $x_1 > x_2$, *then for* $\lambda > 0$ *sufficiently large, we have*

$$C^u_\lambda(f_m)(x) = f_m(x) = x_1, \quad \nabla C^u_\lambda(f_m)(x) = \nabla f_m(x) = e_1, \quad \nabla^2 C^u_\lambda(f_m)(x) = 0.$$

Proof. (i) Suppose $x = (x_1, x_2, \dots, x_m) \in \mathbb{R}^m$ is a level-$(k-1)$ singular point with $k \geq 2$. We have $x_1 = x_2 = \cdots = x_k > x_{k+1}$. We prove that when $\lambda > 0$ is sufficiently large, the face with the largest dimension such that $P_{\frac{\Delta_m}{2\lambda}} x$ is a relative interior point is $\Delta_k/(2\lambda)$, assumptions (i) in Theorem 2.2 and (ii)* in Corollary 2.8 hold, and

$$P_{\frac{\Delta_m}{2\lambda}} x = \sum_{j=1}^{k} \frac{1}{2k\lambda} e_j.$$

For (i) to hold on $\Delta_k/(2\lambda)$, we see that

$$x_j - \frac{1}{k}\sum_{j=1}^{k} x_j + \frac{1}{2k\lambda} = x_1 - x_1 + \frac{1}{2k\lambda} = \frac{1}{2k\lambda} > 0$$

as $x_1 = x_2 = \cdots = x_k$. To satisfy (ii)* for $k+1 \leq s \leq m$ in our case, we need

$$x_s < \frac{1}{k}\sum_{j=1}^{k} x_j - \frac{1}{2k\lambda} = x_1 - \frac{1}{2k\lambda} \Longleftrightarrow x_1 - x_{k+1} > \frac{1}{2k\lambda} \Longleftrightarrow \lambda > \frac{1}{2k(x_1 - x_k)}.$$

Next we show that $\Delta_k/(2\lambda)$ is the face with the largest dimension on which (i) and (ii)* hold. To see this let us consider $\ell > k$, we only need to show that (i) cannot hold by showing that

$$x_\ell < \frac{1}{\ell}\sum_{j=1}^{\ell} x_j - \frac{1}{2\ell\lambda}.$$

We have

$$\begin{aligned}\frac{1}{\ell}\sum_{j=1}^{\ell} x_j - \frac{1}{2\ell\lambda} - x_\ell &= \left(\frac{kx_1}{\ell} + \sum_{j=k+1}^{\ell} x_j\right) - x_\ell - \frac{1}{2\ell\lambda}\\ &= \frac{k(x_1 - x_\ell)}{\ell} + \frac{1}{\ell}\sum_{j=k+1}^{\ell}(x_j - x_\ell) - \frac{1}{2\ell\lambda}\\ &\geq \frac{k(x_1 - x_{k+1})}{\ell} - \frac{1}{2\ell\lambda} = \frac{k}{\ell}\left(x_1 - x_{k+1} - \frac{1}{2k\lambda}\right) > 0\end{aligned}$$

when $\lambda > \frac{1}{2k(x_1 - x_{k+1})}$. Here we have used the fact that

$$x_1 = x_2 = \cdots = x_k > x_{k+1} \geq x_{k+2} \geq \cdots \geq x_m.$$

Thus we have $\quad P_{\frac{\Delta_m}{2\lambda}} x = \displaystyle\sum_{j=1}^{k}\left(x_j + \frac{1}{2k\lambda} - \frac{1}{k}\sum_{i=1}^{k} x_i\right) e_j = \sum_{j=1}^{k} \frac{1}{2k\lambda} e_j.$

We also have that for y in a small neighbourhood of x and for $\lambda > \frac{1}{2k(x_1 - x_{k+1})}$,

$$P_{\frac{\Delta_m}{2\lambda}} y = \sum_{j=1}^{k}\left(y_j + \frac{1}{2k\lambda} - \frac{1}{k}\sum_{i=1}^{k} y_i\right) e_j.$$

Thus $C_\lambda^u(f_m)(\cdot)$ is twice differentiable near x hence for $\lambda > \frac{1}{2k(x_1 - x_{k+1})}$, we have

$$\begin{aligned}&C_\lambda^u(f_m)(x) - f_m(x) = \frac{1}{4\lambda}\left|e_1 - \frac{1}{k}\sum_{j=1}^{k} e_j\right|^2 = \frac{k-1}{4k\lambda},\\ &\nabla C_\lambda^u(f_m)(x) = 2\lambda P_{\frac{\Delta_m}{2\lambda}} x = \sum_{j=1}^{k} \frac{1}{2k} e_j,\\ &\nabla^2 C_\lambda^u(f_m)(x) = 2\lambda \begin{pmatrix} \mathbf{A}_k & 0 \\ 0 & 0 \end{pmatrix},\end{aligned}$$

where $\mathbf{A}_k$ is given by (27). The proof of (i) is complete.

(ii) Since $x_1 > x_2 \geq x_3 \geq \cdots \geq x_m$, we have $x_1 - x_2 > 1/(2\lambda)$ if $\lambda > \frac{1}{2k(x_1 - x_2)}$. By Corollary 2.3, we see that in this case $C_\lambda^u(f_m)(x) = f_m(x)$ and in a neighbourhood of x we have $\nabla C_\lambda^u(f_m)(y) = e_1$, so we obtain $\nabla^2 C_\lambda^u(f_m)(x) = 0$. This completes the proof of (ii). □

3. Smooth approximations for maximum of smooth functions

As described in Section 1, in this section we consider smooth approximations of the function

$$g^{(m)}(x) = \max\{g_1(x), \ldots, g_m(x)\}, \quad \text{for} \quad x \in \mathbb{R}^n$$

with $g_j : \mathbb{R}^n \to \mathbb{R}$ a smooth (C^2) function for $j = 1, 2, \ldots, m$. Let

$$G_m(x) = (g_1(x), \ldots, g_m(x)) \in \mathbb{R}^m, \quad \text{for} \quad x \in \mathbb{R}^n.$$

As in previous sections we define $f_m(y) = \max\{y_i : i = 1, 2, \ldots, m\}$ for $y \in \mathbb{R}^m$ and denote by $C^u_\lambda(f_m)(y)$ the upper compensated convex transform of f_m at $y \in \mathbb{R}^m$.

First we calculate the gradient for the smoothed function $C^u_\lambda(f_m)(G_m(x))$ and its Hessian when it exists. For vectors a, $b \in \mathbb{R}^n$, we define $a \otimes b = ab^T$. Here we treat a and b as column vectors.

Theorem 3.1. *Suppose $g_j \in C^2(\mathbb{R}^n)$ for $j = 1, 2, \ldots, m$ with $m \geq 2$. Let $g^{(m)}(x)$ and $G_m(x)$ be as defined previously for $x \in \mathbb{R}^n$. Then for $\lambda > 0$ we have*

$$\nabla C^u_\lambda(f_m)(G_m(x)) = 2\lambda \nabla G_m(x)^T P_{\frac{\Delta_m}{2\lambda}}(G_m(x)) = 2\lambda \sum_{j=1}^m \lambda_j(G_m(x)) \nabla g_j(x), \tag{29}$$

where $P_{\frac{\Delta_m}{2\lambda}}(G_m(x)) = \sum_{j=1}^m \lambda_j(G_m(x)) e_j$ is the convex projection of $G_m(x)$ to $\Delta_m/(2\lambda)$ with $\lambda_j(G_m(x)) \geq 0$ for $j = 1, 2, \ldots, m$ and $\sum_{j=1}^m \lambda_j(G_m(x)) = 1/(2\lambda)$.

If we further assume that $g_1(x) \geq g_2(x) \geq \cdots \geq g_m(x)$ and $\Delta_k/(2\lambda)$ is the face of $\Delta_m/(2\lambda)$ with the largest dimension with $1 \leq k \leq m$ such that $P_{\Delta_m/(2\lambda)} G_m(x)$ is a relative interior of $\Delta_k/(2\lambda)$ and $G_m(x)$ satisfies conditions (i) *in Theorem* 2.2 *and* (ii)* *in Corollary* 2.8*, then $\nabla^2 C^u_\lambda(f_m)(G_m(x))$ exists and is given by*

$$\begin{aligned}\nabla^2 C^u_\lambda(f_m)(G_m(x)) = \; & 2\lambda \sum_{j=1}^k \lambda_j(G_m(x)) \nabla^2 g_j(x) + 2\lambda \sum_{i=1}^k \nabla g_i(x) \otimes \nabla g_i(x) \\ & - \frac{2\lambda}{k} \left(\sum_{\ell=1}^k \nabla g_\ell(x) \right) \otimes \left(\sum_{s=1}^k \nabla g_s(x) \right),\end{aligned} \tag{30}$$

where $\lambda_i(G_m(x)) > 0$ for $i = 1, 2, \ldots, k$ and $\sum_{i=1}^k \lambda_i(G_m(x)) = 1/(2\lambda)$ satisfying

$$P_{\frac{\Delta_k}{2\lambda}}(G_m(x)) = \sum_{j=1}^k \lambda_j(G_m(x)) e_j.$$

In the above statement, if $k = 1$, then by Corollary 2.3*, we have*

$$g_1(x) - g_2(x) > 1/(2\lambda).$$

Thus for the case $k = 1$ we have $\nabla^2 C^u_\lambda(f_m)(G_m(x)) = \nabla^2 g_1(x)$.

Proof. By the chain rule we see that formula (29) for $\nabla C^u_\lambda(f_m)(G_m(x))$ holds.

Now we establish (30) under the assumptions (i) in Theorem 2.2 and (ii)* in Corollary 2.8. Since $G_m(\cdot)$ is C^2-smooth hence is continuous, by our assumption, we see that for y in a neighbourhood of x,

$$P_{\frac{\Delta_m}{2\lambda}} G_m(y) = P_{\frac{\Delta_k}{2\lambda}} G_m(y) = \sum_{j=1}^k \lambda_j(G_m(y)) e_j = \sum_{j=1}^k \left(g_j(y) - \frac{1}{k} \sum_{i=1}^k g_i(y) + \frac{1}{2k\lambda} \right) e_j$$

with $$\lambda_j(G_m(y)) = g_j(y) - \frac{1}{k} \sum_{i=1}^k g_i(y) + \frac{1}{2k\lambda} > 0, \quad j = 1, 2, \ldots, k,$$

and $$g_s(y) + \frac{1}{2k\lambda} < 0, \quad k+1 \leq s \leq m.$$

Therefore $\lambda_j(G_m(y))$ is differentiable at x for $j = 1, 2, \ldots, k$ and

$$\frac{\partial \lambda_j(G_m(x))}{\partial x_s} = \frac{\partial g_j(x)}{\partial x_s} - \frac{1}{k}\sum_{i=1}\frac{\partial g_i(x)}{\partial x_s},$$

so that we have

$$\frac{\partial^2 C^u_\lambda(f_m)(G_m(x))}{\partial x_s \partial x_\ell} = 2\lambda \frac{\partial}{\partial x_s}\left(\sum_{j=1}^k \lambda_j(G_m(y))\frac{\partial g_j(x)}{\partial x_\ell}\right)$$

$$= 2\lambda\sum_{j=1}^k \lambda_j(G_m(y))\frac{\partial^2 g_j(x)}{\partial x_\ell \partial x_s} + 2\lambda\sum_{j=1}^k \frac{\partial \lambda_j(G_m(y))}{\partial x_s}\frac{\partial g_j(x)}{\partial x_\ell}$$

$$= 2\lambda\sum_{j=1}^k \lambda_j(G_m(y))\frac{\partial^2 g_j(x)}{\partial x_\ell \partial x_s} + 2\lambda\sum_{j=1}^k \left(\frac{\partial g_j(x)}{\partial x_s} - \frac{1}{k}\sum_{i=1}\frac{\partial g_i(x)}{\partial x_s}\right)\frac{\partial g_j(x)}{\partial x_\ell}$$

$$= 2\lambda\sum_{j=1}^k \left(\lambda_j(G_m(y))\frac{\partial^2 g_j(x)}{\partial x_\ell \partial x_s} + \frac{\partial g_j(x)}{\partial x_s}\frac{\partial g_j(x)}{\partial x_\ell}\right) - \frac{2\lambda}{k}\left(\sum_{j=1}^k \frac{\partial g_j(x)}{\partial x_s}\right)\left(\sum_{i=1}^k \frac{\partial g_i(x)}{\partial x_\ell}\right).$$

Thus we have derived (30):

$$\begin{aligned}\nabla^2 C^u_\lambda(f_m)(G_m(x)) = \;& 2\lambda\sum_{j=1}^k \lambda_j(G_m(y))\nabla^2 g_j(x) + 2\lambda\sum_{j=1}^k \nabla g_j(x)\otimes\nabla g_j(x)\\ & -\frac{2\lambda}{k}\left(\sum_{i=1}^k \nabla g_i(x)\right)\otimes\left(\sum_{j=1}^k \nabla g_j(x)\right).\end{aligned}$$

If we apply Theorem 3.1 to affine functions, we have

Corollary 3.2. *Let $\ell_i(y) = a_i \cdot y + b_i$ for $y \in \mathbb{R}^n$ be affine functions with $a_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$ for $i = 1, 2, \ldots, m$. Let the functions $L_m : \mathbb{R}^n \to \mathbb{R}^m$ be defined by $L_m(y) = (\ell_1(y), \ell_2(y), \ldots, \ell_m(y))$ for $y \in \mathbb{R}^n$. If we assume that $\ell_1(x) \geq \ell_2(x) \geq \cdots \geq \ell_m(x)$ and $\Delta_k/(2\lambda)$ is the face of $\Delta_m/(2\lambda)$ with the largest dimension $1 \leq k \leq m$ such that $P_{\Delta_m/(2\lambda)}L_m(x)$ is a relative interior of $\Delta_k/(2\lambda)$ and $L_m(x)$ satisfies* (i) *in Theorem* 2.2 *and* (ii)* *in Corollary* 2.8*, then $\nabla^2 C^u_\lambda(f_m)(L_m(x))$ exists and is given by*

$$\nabla^2 C^u_\lambda(f_m)(L_m(x)) = 2\lambda\left[\sum_{i=1}^k a_i\otimes a_i - \frac{1}{k}\left(\sum_{j=1}^k a_j\right)\otimes\left(\sum_{s=1}^k a_s\right)\right]. \qquad (31)$$

Proof. If we let $g_i(y) := \ell_i(y)$ for $y \in \mathbb{R}^n$. Then $\nabla g_i(y) = a_i$ and $\nabla^2 g_i(y) = 0$ for $y \in \mathbb{R}^n$. If we substitute these in (30), the conclusion then follows. □

Remark 3.3. There are some obvious cases for which $\nabla^2 C^u_\lambda(f_m)(L_m(x))$ is not invertible.

(i) If the linear span satisfies $S := \mathsf{span}[a_j : \; j = 1, 2, \ldots, k] \neq \mathbb{R}^n$, then $S^\perp \neq \{0\}$ so that the Hessian given by (31) has a nontrivial kernel. A simple example of this case is when $k < n$, that is, when the convex projection is on a lower dimensional simplex $\Delta_k/(2\lambda)$ with $k < n$.

(ii) If there is some $y^* \in \mathbb{R}^n \setminus \{0\}$ such that $a_1 \cdot y^* = a_2 \cdot y^* = \cdots = a_k \cdot y^*$, then 0 is the smallest eigenvalue with y^* an eigenvector. This is a direct consequence of the Cauchy-Schwarz inequality. This condition can be satisfied even when $S = \mathbb{R}^n$ and $k \geq n$ if, say, a basis is given by $a_1, a_2, \dots, a_n$ and $a_j \in \mathrm{co}\{a_1, a_2, \dots, a_n\}$ - the convex hull of the finite set $\{a_1, a_2, \dots, a_n\}$ for $j = n+1, \dots, m$ when $m > n$.

Theorem 3.4. *Suppose $g_k \in C^2(\mathbb{R}^n)$ for $k = 1, 2, \dots, m$ with $m > 1$. Let $g^{(m)}(x) = \max\{g_k(x) : k = 1, 2, \dots, m\}$ and $G_m(x) = (g_1(x), g_2(x), \dots, g_m(x))$ for $x \in \mathbb{R}^n$. Then for $\lambda > 0$ we have*

(i) *If g_i is convex in $\mathbb{R}^n$ for $i = 1, 2, \dots, m$, then $C^u_\lambda(f_m)(G_m(\cdot))$ is convex in $\mathbb{R}^n$.*

(ii) *The function $C^u_\lambda(f_m)(G_m(\cdot)) \in C^{1,1}_{loc}(\mathbb{R}^n)$. More precisely, in any bounded domain $\Omega \subseteq \mathbb{R}^n$, let $L_\Omega \geq 0$ be the maximum of the Lipschitz constants of $G_m(\cdot)$ and $\nabla G_m(\cdot)$ in Ω and $M_\Omega = \sup_{x\in\Omega} |\nabla G_m(x)|$, then, for all $x, y \in \Omega$,*

$$|\nabla C^u_\lambda(f_m)(G_m(y) - \nabla C^u_\lambda(f_m)(G_m(x)| \leq L_\Omega(2\lambda M_\Omega + 1)|y - x|. \tag{32}$$

(iii) $g^{(m)}(x) \leq C^u_\lambda(f_m)(G_m(x)) \leq g^{(m)}(x) + \dfrac{m-1}{4m\lambda}$ *for* $x \in \mathbb{R}^n$.
Furthermore, $C^u_\lambda(f_m)(G_m(x)) = g^{(m)}(x)$ if and only if

$$g^{(m)}(x) \geq \hat{g}^{(m)}(x) + \frac{1}{2\lambda}, \tag{33}$$

where $\hat{g}^{(m)}(x) := \min_{1\leq i\leq m} \max\{g_j(x),\ 1 \leq j \leq m,\ j \neq i\}, \quad x \in \mathbb{R}^n.$

(iv) *If $G_m(x)$ is a level-$(k-1)$ singular point of $f_m(\cdot)$, then*

$$\begin{aligned}
&\lim_{\lambda\to+\infty} \lambda(C^u_\lambda(f_m)(G_m(x)) - g^{(m)}(x)) = \frac{k-1}{4k},\\
&\lim_{\lambda\to+\infty} \nabla C^u_\lambda(f_m)(G_m(x)) = \frac{1}{k}\sum_{s=1}^{k} \nabla g_{i_s}(x), \qquad (34)\\
&\lim_{\lambda\to+\infty} \frac{1}{2\lambda}\nabla^2 C^u_\lambda(f_m)(G_m(x)) = \sum_{i=1}^{k} \nabla g_i(x) \otimes \nabla g_i(x) - \frac{1}{k}(\sum_{\ell=1}^{k} \nabla g_\ell) \otimes (\sum_{s=1}^{k} \nabla g_s).
\end{aligned}$$

Proof. Item (i) is a direct consequence of the fact that $y \to C^u_\lambda(f_m)(y)$ is both convex and monotonically increasing.

To prove (ii) we first recall (30) that

$$\nabla C^u_\lambda(f_m)(G_m(x)) = 2\lambda \nabla G_m(x)^T P_{\frac{\Delta_m}{2\lambda}}(G_m(x)) = 2\lambda \sum_{j=1}^{m} \lambda_j(G_m(x))\nabla g_j(x).$$

Since $g_j \in C^2(\mathbb{R}^n)$ for $j = 1, 2, \dots, m$ and $P_{\frac{\Delta_m}{2\lambda}}(\cdot)$ is an Lipschitz mapping, we see that $\nabla C^u_\lambda(f_m)(G_m(x)) \in C^1(\mathbb{R}^n)$. Next we establish (32). We have, for $x, y \in \Omega$,

$$\begin{aligned}
&|\nabla C^u_\lambda(f_m)(G_m(x)) - \nabla C^u_\lambda(f_m)(G_m(x))|\\
&= 2\lambda|\nabla G_m(x)^T P_{\frac{\Delta_m}{2\lambda}}(G_m(x)) - \nabla G_m(y)^T P_{\frac{\Delta_m}{2\lambda}}(G_m(y))|\\
&\leq 2\lambda|\nabla G_m(x)^T P_{\frac{\Delta_m}{2\lambda}}(G_m(x)) - \nabla G_m(y)^T P_{\frac{\Delta_m}{2\lambda}}(G_m(x))|\\
&\quad +2\lambda|\nabla G_m(y)^T P_{\frac{\Delta_m}{2\lambda}}(G_m(x)) - \nabla G_m(y)^T P_{\frac{\Delta_m}{2\lambda}}(G_m(y))|
\end{aligned}$$

$$\le 2\lambda |P_{\frac{\Delta_m}{2\lambda}}(G_m(x))||\nabla G_m(x) - \nabla G_m(y)| + 2\lambda M_\Omega |P_{\frac{\Delta_m}{2\lambda}}(G_m(x)) - P_{\frac{\Delta_m}{2\lambda}}(G_m(y))|$$
$$\le L_\Omega |x-y| + 2\lambda M_\Omega |G_m(x) - G_m(y)|$$
$$\le L_\Omega (1 + 2\lambda M_\Omega)|x-y|.$$

Item (iii) is a direct consequence of Corollaries 2.3 and 2.5.

Item (iv) follows directly from Theorem 2.11 and Theorem 3.1. □

4. Comparisons between two approximation methods

In this section we compare our smoothing formula $C^u_\lambda(f_m)(\cdot)$ of the maximum function $f_m(\cdot)$ with the log-sum-exp formula $H^{(m)}_\lambda(\cdot)$.

Let us recall that the 'log-sum-exp' smooth approximation for the maximum function $f_m(\cdot)$ defined by (13) is given by

$$H^{(m)}_\lambda(y) = \frac{1}{\lambda}\log\left(\sum_{k=1}^m e^{\lambda y_k}\right), \quad y \in \mathbb{R}^m. \tag{35}$$

Clearly for any $\lambda > 0$, $H^{(m)}_\lambda(\cdot)$ is a C^∞ function. This allows us to take partial derivatives of any order. Also $H^{(m)}_\lambda(\cdot)$ comes with an analytic formula which does not require any sub-routines to calculate its value besides those built-in ones such as 'exp' and 'log'. In order to a avoid the possible numerical overflow or underflow caused by the exponential function, $H^{(m)}_\lambda(\cdot)$ is computed by its shifted version ([20, 22]) as

$$H^{(m)}_\lambda(y) = f_m(y) + \frac{1}{\lambda}\log\left(\sum_{k=1}^m e^{\lambda(y_k - f_m(y))}\right). \tag{36}$$

However, it is shown in [4] that the shifted formula is not more accurate than the unshifted one.

The calculations above by shifted formula requires the computation of the maximum function $f_m(y)$ at y. If we try to find the maximum, the complexity is $\mathcal{O}(m)$. If we sort the whole vector complexity is about $\mathcal{O}(m\log m)$ [11]. Even for the maximum of finitely many affine functions $\ell^{(m)}(x) := \max\{a_i \cdot x + b_i : \ i = 1, 2, \dots, m\}$, the Hessian for the 'log-sum-exp' formula is rather complicated. However the Hessian of $C^u_\lambda(f_m)(L_m(x))$ is given by a rather simple formula (31) and may require fewer components than all of the m components when we calculate $\nabla^2 H^{(m)}_\lambda(L_m(x))$.

The well-known error estimate for the approximation $H^{(m)}_\lambda(\cdot)$ is given by

$$f_m(y) \le H^{(m)}_\lambda(y) \le f_m(y) + \frac{\log m}{\lambda}, \quad y \in \mathbb{R}^m. \tag{37}$$

The error bound $(\log m)/\lambda$ is sharp and can be realised when all components of y are equal. For a finite minimax optimization problem of a fixed size m, this error may not cause serious problems for smoothing optimization.

However if one considers semi-infinite minimax problems, for example, in the form

$$\min_{x\in\mathbb{R}^n}\max_{y\in[0,\,1]} g(x,y),$$

one natural approximation is to descretise [0, 1] and consider, e.g., $y_k = k/m$ for $k = 0,1,2,\ldots,m$. Then the finite minimax problem is considered first by solving $\min_{x\in\mathbb{R}^n}\max_{1\le k\le m} g(x,y_k)$. Under mild conditions [22], one can show that the finite problem converges to the semi-infinite problem as $m\to\infty$. Due to the error bound (37) this makes it difficult to consider a sequence of finite minimax problems for a fixed smoothing papameter $\lambda > 0$. If we use the upper transform to approximate $g^{(m)}(x) := \max_{1\le k\le m} g(x,y_k)$, we have, from (23) that the error estimate satisfies

$$g^{(m)}(x) \le C^u_\lambda(f_m)(G_m(x)) \le g^{(m)}(x) + \frac{1}{4\lambda}$$

which is independent of m.

We also observe that $H^{(m)}_\lambda(\cdot)$ is a nowhere tight approximation of $f_m(\cdot)$ in the sense that $f_m(y) < H^{(m)}_\lambda(y)$ for all $y\in\mathbb{R}^m$. This makes it more challenging to be used for nonlinear feasibility and problems where precise geometric information of approximations is needed. If we use the upper transform and assume $G_m(x)$ is 'sorted', we have $C^u_\lambda(f_m)(G_m(x)) = g^{(m)}(x)$ if $g_1(x) - g_2(x) \ge 1/(2\lambda)$. In this case a point is in the feasible set if $C^u_\lambda(f_m)(G_m(x)) \le 0$. In general we have, based on the inequalities above that

$$\begin{aligned}&\big\{x\in\mathbb{R}^n:\ C^u_\lambda(f_m)(G_m(x))\le 0\big\}\subset\big\{x\in\mathbb{R}^n:\ g_j(x)\le 0,\ j=1,2,\ldots,m\big\}\\ &\subset\big\{x\in\mathbb{R}^n:\ C^u_\lambda(f_m)(G_m(x))\le \frac{1}{4\lambda}\big\}.\end{aligned}$$

In case $K = \{x\in\mathbb{R}^n:\ g_j(x)\le 0,\ j=1,2,\ldots,m\} = \emptyset$, that is, if K is not feasible, we see that under the coercivity assumption that $g^{(m)}(x)\to+\infty$ as $|x|\to\infty$, we have $g^{(m)}(x)\ge c_0 > 0$ for a constant c_0 and for all $x\in\mathbb{R}^n$. In this case, we also have $C^u_\lambda(f_m)(G_m(x))\ge c_0$ for all $x\in\mathbb{R}^n$.

Next we state our main comparison results for $C^u_\lambda(f_m)(\cdot)$ and $H^{(m)}_\mu(\cdot)$. The first result (Theorem 4.1) says that pointwisely $C^u_\lambda(f_m)(\cdot)$ is a much sharper approximation of $f_m(\cdot)$ than $H^{(m)}_{4\lambda}(\cdot)$. The second result is a 'weak' generalisation of the last part of Theorem 4.1 concerning an average asymptotic comparison of the speed of convergence to $f_m(\cdot)$ by $C^u_\lambda(f_m)(\cdot)$ and $H^{(m)}_{\mu(\lambda)}(\cdot)$ respectively in the sense that

$$\lim_{\lambda\to+\infty}\int_{\bar B_r(0)}\frac{C^u_\lambda(f_m)(x) - f_m(x)}{H^{(m)}_{\mu(\lambda)}(x) - f_m(x)}\,dx = 0,$$

where $\mu(\lambda)$ is a positive increasing function of $\lambda > 1$ and $\bar B_r(0) = \{y\in\mathbb{R}^m:\ |y|\le r\}$ is the closed ball centred at 0 with radius $r > 0$. In Theorem 4.4 we claim that the above result holds for $m\ge 2$ if $\mu(\lambda) = \alpha\lambda\log(\lambda)$ with $0<\alpha<2$. In particular, for $m = 2$ the same result is true for $\mu(\lambda) = 2\lambda\log(\lambda)$. The third result (Theorem 4.7) says that under a strong convexity assumption, we can derive error estimates between the minimiser of the original function $g^{(m)}(\cdot)$ and those obtained by the two different smoothing methods above respectively. We further establish that both the minimum value and the minimiser obtained from $C^u_\lambda(f_m)(G_m(\cdot))(\cdot)$ are closer to those obtained from $g^{(m)}(\cdot)$ than those obtained from $H^{(m)}_{4\lambda}(G_m(\cdot))$.

Theorem 4.1. *For $\lambda > 0$ and $m \geq 2$, we have*

$$f_m(y) \leq C^u_\lambda(f_m)(y) < H^{(m)}_{4\lambda}(y), \qquad y \in \mathbb{R}^m. \tag{38}$$

We need the following lemma (Lemma 4.2) for the proof of Theorem 4.1. We put the proof of the lemma in the Appendix. It is elementary but we have not been able to find a reference for it.

Lemma 4.2. *Suppose $m \geq 2$. Then the following equality holds.*

$$\max\left\{\sum_{j=1}^m \lambda_j x_j^2 - \left(\sum_{i=1}^m \lambda_i x_i\right)^2 : 0 \leq \lambda_j \leq 1,\ x_j \in \mathbb{R}, \sum_{j=1}^m \lambda_j = 1,\ \sum_{\ell=1}^m x_\ell^2 = 1\right\} = \frac{1}{2}. \tag{39}$$

The maximum is reached, for example, at the point with $x_1 = 1/\sqrt{2}$, $x_2 = -1/\sqrt{2}$, $\lambda_1 = \lambda_2 = 1/2$ and $x_j = 0$, $\lambda_j = 0$ for $3 \leq j \leq m$.

An obvious consequence of Lemma 4.2 is that

$$\sum_{j=1}^m \lambda_j x_j^2 - \left(\sum_{i=1}^m \lambda_i x_i\right)^2 \leq \frac{|x|^2}{2} \tag{40}$$

for all $x = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$ and $0 \leq \lambda_j \leq 1$, satisfying $\sum_{j=1}^m \lambda_j = 1$.

Proof of Theorem 4.1. It is easy to see that $f_m(y) \leq C^u_\lambda(f_m)(y)$ (see [30] and (15)). We only need to show that $C^u_\lambda(f_m)(y) < H^{(m)}_{4\lambda}(y)$. By the definition of the upper transform, it is written as

$$\lambda|y|^2 - \mathrm{co}[\lambda|\cdot|^2 - f_m(\cdot)](y) < H^{(m)}_{4\lambda}(y),$$

which is obviously equivalent to

$$\lambda|y|^2 - H^{(m)}_{4\lambda}(y) < \mathrm{co}[\lambda|\cdot|^2 - f_m(\cdot)](y). \tag{41}$$

Now we establish (41). We show that the left hand side of (41) is a convex function by computing the quadratic form defined by its Hessian. We have, for $x = (x_1, x_2, \ldots, x_m)^T \in \mathbb{R}^m$ that

$$\begin{aligned} &x^T \nabla^2(\lambda|y|^2 - H^{(m)}_{4\lambda}(y))x \\ &= 2\lambda\left(|x|^2 - 2\left(\frac{\sum_{k=1}^m x_k^2 e^{4\lambda y_k}}{\sum_{j=1}^m e^{4\lambda y_j}} - \left(\frac{\sum_{j=1}^m x_j e^{4\lambda y_j}}{\sum_{k=1}^m e^{4\lambda y_k}}\right)^2\right)\right) \geq 0 \end{aligned} \tag{42}$$

by Lemma 4.2 (see (40)) if we set $\lambda_j = e^{4\lambda y_j}/(\sum_{k=1}^m e^{4\lambda y_k})$ and observe that we have $0 < \lambda_j < 1$ and $\sum_{k=1}^m \lambda_k = 1$. Here we view $x \in \mathbb{R}^m$ as a column vector.

Since $m \geq 2$, we also have that $f_m(z) < H^{(m)}_{4\lambda}(z)$ for every $z \in \mathbb{R}^m$. Hence

$$\lambda|z|^2 - H^{(m)}_{4\lambda}(z) < \lambda|z|^2 - f_m(z). \tag{43}$$

For every $y \in \mathbb{R}^m$, there exist $0 \leq \lambda_1, \lambda_2, \ldots, \lambda_{m+1} \leq 1$ and $x_1, x_2, \ldots, x_{m+1} \in \mathbb{R}^m$ satisfying $\sum_{j=1}^{m+1} \lambda_j = 1$ and $\sum_{j=1}^{m+1} \lambda_j x_j = y$ such that

$$\mathrm{co}[\lambda|\cdot|^2 - f_m(\cdot)](y) = \sum_{j=1}^{m+1} \lambda_j(\lambda|x_j|^2 - f_m(x_j)).$$

By (43) and the convexity of $\lambda|y|^2 - H_{4\lambda}^{(m)}(y)$,

$$\begin{aligned}\sum_{j=1}^{m+1} \lambda_j(\lambda|x_j|^2 - f_m(x_j)) &> \sum_{j=1}^{m+1} \lambda_j(\lambda|x_j|^2 - H_{4\lambda}^{(m)}(x_j)) \\ &\geq \mathrm{co}[\lambda|\cdot|^2 - H_{4\lambda}^{(m)}(\cdot)](y) = \lambda|y|^2 - H_{4\lambda}^{(m)}(y).\end{aligned}$$

This shows the strict inequality (41) holds. □

Remark 4.3. Theorem 4.1 implies that pointwisely $C_\lambda^u(f_m)$ is a better approximation to f_m than $H_{4\lambda}^{(m)}$. A direct consequence of (38) is that for every Lebesgue measurable set $\Omega \subset \mathbb{R}^m$ with finite Lebesgue measure, we have

$$\lim_{\lambda\to+\infty} \int_\Omega \frac{C_\lambda^u(f_m)(y) - f_m(y)}{H_{4\lambda}^{(m)}(y) - f_m(y)}\, \mathrm{d}y = 0.$$

Therefore it is natural to ask whether in the average (integral) sense, we can find a function $\mu(\lambda)$ for $\lambda > 1$ such that $\mu(\lambda)/\lambda \to +\infty$, such that the above integral asymptotic result still holds if we replace 4λ by $\mu(\lambda)$.

Theorem 4.4. *Suppose $m \geq 2$. Let $\mu_\alpha(\lambda) = \alpha\lambda\log(\lambda)$ with $\lambda > 1$ and $\alpha > 0$. Let*

$$A(\lambda,\, \alpha,\, r,\, m) := \int_{\bar{B}_r(0)} \frac{C_\lambda^u(f_m)(x) - f_m(x)}{H_{\mu_\alpha(\lambda)}^{(m)}(x) - f_m(x)}\, \mathrm{d}x \tag{44}$$

with $\bar{B}_r(0) \subseteq \mathbb{R}^m$ the closed ball centred at 0 with radius $r > 0$. Then

(i) *for every $m \geq 2$, for every $0 < \alpha < 2$, we have*

$$\lim_{\lambda\to+\infty} A(\lambda,\, \alpha,\, r,\, m) = 0; \tag{45}$$

(ii) *in particular, for $m = 2$ and $\alpha = 2$, we have*

$$\lim_{\lambda\to+\infty} A(\lambda,\, 2,\, r,\, 2) = 0; \tag{46}$$

(iii) *for $m = 2$, the parameter $\alpha = 2$ is sharp for (46) to hold, that is, for any $\alpha > 2$,*

$$\liminf_{\lambda\to+\infty} A(\lambda,\, \alpha,\, r,\, 2) = +\infty. \tag{47}$$

The proof of Theorem 4.4 is rather technical. We put its proof at the end of the Appendix.

Remark 4.5. Theorem 4.4 implies that in the average sense $C^u_\lambda(f_m)(\cdot)$ approaches $f_m(\cdot)$ faster than $H^{(m)}_{\alpha\lambda\log(\lambda)}(\cdot)$. The ratio $\mu_\sigma(\lambda)$ is superlinear with respect to $\lambda > 1$. The parameter $0 < \alpha < 2$ is nearly sharp when $m \geq 3$ in (i) above as at least in the case $m = 2$ we see that $\alpha = 2$ is sharp.

However, by considering points with equal components, say $x_t = (t, t, \dots, t) \in \mathbb{R}^m$ we observe that when $\lambda > 1$ is sufficiently large, we have, by Theorem 2.11, formula (28) and the fact that $H^{(m)}_\mu(x_t) - f_m(x_t) = \frac{\log m}{\mu}$, that the integrand in (44) is not pointwise bounded above in any $\bar{B}_r(0)$ with $r > 0$ when $\lambda \to \infty$ as

$$\frac{C^u_\lambda(f_m)(x_t) - f_m(x_t)}{H^{(m)}_{\mu_\alpha(\lambda)}(x_t) - f_m(x_t)} = \frac{\frac{m-1}{4m\lambda}}{\frac{\log m}{\alpha\lambda\log(\lambda)}} = \frac{\alpha(m-1)\log(\lambda)}{4m\log(m)} \to \infty.$$

We need the following technical lemma for the proof of Theorem 4.4 whose proof can be found in the Appendix.

Lemma 4.6. *Let $\sigma > 0$ be a constant. Then*

$$\begin{aligned}
&\lim_{\lambda\to+\infty} \frac{\log\lambda}{\lambda}\int_0^1 \frac{1}{\log\left(1+\frac{1}{\lambda^{\sigma t}}\right)}\,\mathrm{d}t = 0 \quad \text{if } 0<\sigma<1;\\
&\liminf_{\lambda\to+\infty} \frac{\log\lambda}{\lambda}\int_0^1 \frac{1}{\log\left(1+\frac{1}{\lambda^{t}}\right)}\,\mathrm{d}t \geq \frac{1}{e};
\end{aligned} \tag{48}$$

$$\begin{aligned}
&\lim_{\lambda\to+\infty} \frac{\log\lambda}{\lambda}\int_0^1 \frac{(1-t)^2}{\log\left(1+\frac{1}{\lambda^{t}}\right)}\,\mathrm{d}t = 0;\\
&\lim_{\lambda\to+\infty} \frac{\log\lambda}{\lambda}\int_0^1 \frac{(1-t)^2}{\log\left(1+\frac{1}{\lambda^{\sigma t}}\right)}\,\mathrm{d}t = +\infty \quad \text{if } \sigma>1.
\end{aligned} \tag{49}$$

The following theorem gives an error estimate of the minimisers of the two approximations of $g^{(m)}(x) = \max\{g_j(x) : j = 1, 2, \dots, m\}$.

Theorem 4.7. *Suppose $g_j : \mathbb{R}^m \to \mathbb{R}$ is a C^2 function for $j = 1, 2, \dots, m$ and there is a constant $c > 0$ such that $y^T\nabla^2 g_j(x)y \geq c|x|^2$ for all $x, y \in \mathbb{R}^m$ and $j = 1, 2, \dots, m$. Let*

$$\begin{aligned}
g^{(m)}(x) &= \max\{g_j(x) : 1 \leq j \leq m\},\\
g^{(m)}_\lambda(x) &= C^u_\lambda(f_m)(G_m(x)),\\
h^{(m)}_\lambda(x) &= H^{(m)}_{4\lambda}(G_m(x)),\ x \in \mathbb{R}^n,
\end{aligned}$$

where $G_m(x) = (g_1(x), g_2(x), \dots, g_m(x))$ and $\lambda > 0$. Suppose the global minima

$$\begin{aligned}
argmin\{g^{(m)}(x) : x \in \mathbb{R}^n\} &= x_\mu \in \mathbb{R}^m,\\
argmin\{g^{(m)}_\lambda(x) : x \in \mathbb{R}^n\} &= x^c_\lambda \in \mathbb{R}^m,\\
argmin\{h^{(m)}_\lambda(x) : x \in \mathbb{R}^n\} &= x^h_\lambda \in \mathbb{R}^m,
\end{aligned}$$

exist, then we have

$$|x_\mu - x^c_\lambda| \le \frac{1}{\sqrt{2c\lambda}}, \quad |x_\mu - h^{(m)}_\lambda| \le \frac{\sqrt{\log m}}{\sqrt{2c\lambda}}, \tag{50}$$

and
$$g^{(m)}(x_\mu) \le g^{(m)}_\lambda(x^c_\lambda) \le h^{(m)}_\lambda(x^h_\lambda). \tag{51}$$

Proof. From the error estimate (23) in Corollary 2.5, we have

$$g^{(m)}(x_\mu) \le g^{(m)}_\lambda(x_\mu) \le g^{(m)}(x_\mu) + \frac{1}{4\lambda}, \quad g^{(m)}(x^c_\lambda) \le g^{(m)}_\lambda(x^c_\lambda) \le g^{(m)}(x^c_\lambda) + \frac{1}{4\lambda}.$$

Thus we obtain

$$0 \le g^{(m)}_\lambda(x_\mu) - g^{(m)}_\lambda(x^c_\lambda) \le g^{(m)}(x_\mu) - g^{(m)}(x^c_\lambda) + \frac{1}{4\lambda} \le \frac{1}{4\lambda} \tag{52}$$

as x_μ is the global minimiser of $g^{(m)}(x)$ in $\mathbb{R}^n$.

Since x^c_λ is a stationary point of $g^{(m)}_\lambda(x) = C^u_\lambda(f_m)(x)$, by Corollary 2.8 and the chain rule we have

$$\begin{aligned}\nabla g^{(m)}_\lambda(x^c_\lambda) &= \nabla C^u_\lambda(f_m)(G_m(x^c_\lambda)) = \nabla_x G_m(x^c_\lambda)^T \nabla_y C^u_\lambda(f_m)(G_m(x^c_\lambda))\\ &= 2\lambda \nabla_x G_m(x^c_\lambda)^T P_{\frac{\Delta_m}{2\lambda}}(G_m(x^c_\lambda)) = 0,\end{aligned} \tag{53}$$

where $\nabla_y C^u_\lambda(f_m)(G_m(x^c_\lambda))$ is the gradient of $C^u_\lambda(f_m)(y)$ at $y = G_m(x^c_\lambda) \in \mathbb{R}^m$.

Now since $C^u_\lambda(f_m)(\cdot)$ is convex and C^1 in $\mathbb{R}^m$, by the definition of $g^{(m)}_\lambda$ and (52) (53) we have,

$$\begin{aligned}\frac{1}{4\lambda} &\ge C^u_\lambda(f_m)(G_m(x_\mu)) - C^u_\lambda(f_m)(G_m(x^c_\lambda))\\ &\ge \nabla_y C^u_\lambda(f_m)(G_m(x^c_\lambda)) \cdot (G_m(x_\mu) - G_m(x^c_\lambda))\\ &= \nabla_y C^u_\lambda(f_m)(G_m(x^c_\lambda)) \cdot \Big(\nabla G_m(x^c_\lambda) \cdot (x_\mu - x^c_\lambda) + \frac{1}{2}(x_\mu - x^c_\lambda)^T \nabla^2 G_m(\xi)(x_\mu - x^c_\lambda)\Big)\\ &\overset{(53)}{=} \frac{1}{2}\nabla_y C^u_\lambda(f_m)(G_m(x^c_\lambda)) \cdot \big((x_\mu - x^c_\lambda)^T \nabla^2 G_m(\xi)(x_\mu - x^c_\lambda)\big)\\ &= \lambda P_{\frac{\Delta_m}{2\lambda}}(G_m(x^c_\lambda)) \cdot \big((x_\mu - x^c_\lambda)^T \nabla^2 G_m(\xi)(x_\mu - x^c_\lambda)\big)\\ &= \lambda \sum_{j=1}^m \lambda_j(G_m(x^c_\lambda))(x_\mu - x^c_\lambda)^T \nabla^2 g_j(\xi)(x_\mu - x^c_\lambda) \ge \frac{c}{2}|x_\mu - x^c_\lambda|^2,\end{aligned}$$

where we interpret the results of the Taylor's expansion in the bracket of the second line above as column vectors with ξ a point in the interval $[x_\mu, x^c_\lambda]$ (line segment). Here we have used the notation $\nabla_y C^u_\lambda(f_m)(G_m(x^c_\lambda))$ as the gradient of $C^u_\lambda(f_m)(y)$ at $y = G_m(x^c_\lambda) \in \mathbb{R}^m$ and the fact that

$$P_{\frac{\Delta_m}{2\lambda}}(G_m(x^c_\lambda)) = \sum_{j=1}^m \lambda_j(G_m(x^c_\lambda))e_j,$$

where $\lambda_j(G_m(x^c_\lambda)) \ge 0, \quad 1 \le j \le m$, and $\sum_{j=1}^m \lambda_j(G_m(x^c_\lambda)) = \frac{1}{2\lambda}$.

Thus the first inequality in (50) holds, that is,

$$|x_m - x^c_\lambda| \le \frac{1}{\sqrt{2c\lambda}}.$$

Similarly for $h^{(m)}(x)$ we can easily see that

$$\begin{aligned}\frac{\log m}{4\lambda} &\geq h_\lambda^{(m)}(x_\mu) - h_\lambda^{(m)}(x_\lambda^h)\\ &= \nabla h_\lambda^{(m)}(x_\lambda^h)\cdot(x_\mu - x_\lambda^h) + \frac{1}{2}(x_\mu - x_\lambda^h)^T\nabla^2 h_\lambda^{(m)}(\eta)(x_\mu - x_\lambda^h)\\ &\geq \frac{1}{2}\sum_{j=1}^m\left(\frac{e^{4\lambda g_j(x_\lambda^h)}}{\sum_{i=1}^m e^{4\lambda g_i(x_\lambda^h)}}\right)(x_\mu - x_\lambda^h)^T\nabla^2 g_j(\eta)(x_\mu - x_\lambda^h)\\ &\geq \frac{c}{2}|x_\mu - x_\lambda^h|^2,\end{aligned}$$

where $\eta \in [x_\mu,\, x_\lambda^h]$ and we have use the fact that $\nabla h_\lambda^{(m)}(x_\lambda^h) = 0$. Thus the second inequality in (50) is also true:

$$|x_\mu - x_\lambda^h| \leq \frac{\sqrt{\log m}}{\sqrt{2c\lambda}}.$$

It is also easy to see that (51) is also true as we have $g^{(m)}(x) \leq g_\lambda^{(m)}(x) \leq h_\lambda^{(m)}(x)$ for all $x \in \mathbb{R}^n$ and the fact that x_μ, x_λ^c and x_λ^h are minimisers of $g^{(m)}(\cdot)$, $g_\lambda^{(m)}(\cdot)$ and $h_\lambda^{(m)}(\cdot)$ respectively. □

Appendix

We establish Theorem 4.4 in this section. The left hand of (40) can be viewed as the variance of some finite distribution in probability and there is a vast literature on estimates of the upper bound for the variance under various assumptions. However we are unable to find a result we can quote while we believe that Lemma 4.2 should be known.

Proof of Lemma 4.2. Let $D_k = \{\tau = (\tau_1, ..., \tau_k) : 0 \leq \tau_j \leq 1,\, \sum_{j=1}^k \tau_j = 1\} \subset \mathbb{R}^k$ for a positive integer $k > 1$ and D_k^o its interior where the inequalities in the definition of D_k are strict, that is, $0 < \tau_j < 1$ for $1 \leq j \leq k$. We denote by $S^{k-1} := \{x \in \mathbb{R}^k : |x| = 1\}$ the unit sphere in $\mathbb{R}^k$. Now we define for $2 \leq k \leq m$

$$I_k(\tau,\, x) = \sum_{j=1}^k \tau_j x_j^2 - \left(\sum_{i=1}^k \tau_j x_j\right)^2, \quad \tau \in D_k,\ x \in S^{k-1}.$$

By Cauchy-Schwarz inequality it is clear that $I_k(\tau,\, x) \geq 0$ and 0 is its minimum value reach when τ or x has exactly one non-zero components or all components of x are equal, that is, $x_1 = x_2 = \cdots = x_k$. We also see that $I_m(\tau,\, x)$ reaches its maximum in the interior of one of the faces of D_m. Without loss of generality we may assume that the maximum $(\tau^{(m)},\, x^{(m)}) \in D_k$ as at all of vertices of D_m we have $I_m(\tau,\, x) = 0$.

Let $J_k(\tau,\, x,\, \lambda,\, \mu) = \lambda\sum_{j=1}^k \tau_j + \mu\sum_{j=1}^k x_j^2$. We only need to compare values of $I_k(\tau,\, x)$ at all interior critical points in D_k^o for $2 \leq k \leq m$ for the function

$$I_k(\tau,\, x) - J_k(\tau,\, x,\, \lambda,\, \mu),$$

$$\text{subject to } 0 < \tau_j < 1,\ 1 \leq j \leq k,\ \textstyle\sum_{j=1}^k \tau_j = 1,\ \sum_{j=1}^k x_j^2 = 1$$

by using the Lagrangian multiplier method.

We need to find solutions of the following system of equations

$$\frac{\partial}{\partial \tau_i}(I_k(\tau,\, x) - J_k(\tau,\, x,\, \lambda,\, \mu)) = x_i^2 - 2(\textstyle\sum_{j=1}^k \tau_j x_j)x_i - \lambda = 0, \tag{54}$$

$$\frac{\partial}{\partial x_i}(I_k(\tau,\, x) - J_k(\tau,\, x,\, \lambda,\, \mu)) = 2\tau_i x_i - 2(\textstyle\sum_{j=1}^k \tau_j x_j)\tau_i - 2\mu x_i = 0 \tag{55}$$

for $i = 1, 2, \ldots, k$, where $0 < \tau_j < 1$, $\sum_{j=1}^k \tau_j = 1$, and $\sum_{j=1}^k x_j^2 = 1$.

We solve (54) directly for $1 \le i \le k$ and we see that x_i can take at most two values

$$x_i = \sum_{i=1}^k \tau_j x_j + \sqrt{\sum_{j=1}^k \tau_j x_j^2 - \left(\sum_{i=1}^k \tau_j x_j\right)^2} := \sigma_1$$

or

$$x_i = \sum_{i=1}^k \tau_j x_j - \sqrt{\sum_{j=1}^k \tau_j x_j^2 - \left(\sum_{i=1}^k \tau_j x_j\right)^2} := \sigma_2.$$

Clearly $\sigma_1 \ge \sigma_2$. Also if we multiply (54) by τ_i and take the sum, we obtain

$$\lambda = \sum_{j=1}^k \tau_j x_j^2 - 2\left(\sum_{i=1}^k \tau_j x_j\right)^2.$$

We consider three different cases: $\lambda = 0$, $\lambda < 0$ and $\lambda > 0$.

Case 1: $\lambda = 0$. In this case we have either $x_i = 0$ or $x_i = c \neq 0$ - a constant. If only one of these two cases happen, then $I_k(\tau, x) = 0$. Otherwise we may assume that $x_1 = x_2 = \cdot x_\ell = c$ and $x_j = 0$ for $\ell + 1 \le j \le k$ with $1 < \ell < k$. Let $\tau_1 + \cdots + \tau_\ell = s$, then $0 < s < 1$ as we are only looking for critical points for $\tau \in D_k^o$. Since $|x|^2 = 1$ we have $c = 1/\sqrt{\ell}$ and $I_k(\tau, x) = s/l - s^2/l = s(1-s)/l \le 1/4 < 1/2$. Thus in this case critical points are not maximum points.

Case 2: $\lambda < 0$. In this case we have either $0 < \sigma_2 \le \sigma_1$ or $\sigma_2 \le \sigma_1 < 0$. If $\sigma_2 = \sigma_1$ or x_i takes only one values of σ_1 and σ_2, we see that $I_k(\tau, x) = 0$. Thus we may assume that $\sigma_2 < \sigma_1$ and that they have the same sign. We may also assume that $x_1 = \cdots = x_\ell = \sigma_1$, $x_{\ell+1} = \cdots = x_k = \sigma_2$ with $1 \le \ell < k$ and $s = \tau_1 + \cdots + \tau_\ell$ satisfies $0 < s < 1$. Thus

$$\sigma_1 = s\sigma_1 + (1-s)\sigma_2 + \sqrt{s(1-s)(\sigma_1 - \sigma_2)^2} = s\sigma_1 + (1-s)\sigma_2 + \sqrt{s(1-s)}(\sigma_1 - \sigma_2),$$

so that $((1-s) - \sqrt{s(1-s)})(\sigma_1 - \sigma_2) = 0$. This implies $s = 1/2$ as $\sigma_1 - \sigma_2 > 0$. Therefore we have $\ell\sigma_1^2 + (k-\ell)\sigma_2 = 1$ and $I_k(\tau, x) = (\sigma_1 - \sigma_2)^2/4 \le 1/4 < 1/2$ as $|\sigma_1| \le 1/\sqrt{\ell} \le 1$, $|\sigma_2| \le 1/\sqrt{k-\ell} \le 1$ and σ_1 and σ_2 have the same sign. Thus in this case critical points are not maximum points.

Case 3: $\lambda > 0$. In this case we have $\sigma_1 > 0 > \sigma_2$. Again if all x_i take only one value, we have $I_k(\tau, x) = 0$. We use the same notation as in Case 2 and assume that $0 < s < 1$ and $1 \le \ell < k$. We then have $\sigma_1 = s\sigma_1 + (1-s)\sigma_2 + \sqrt{s(1-s)}(\sigma_1 - \sigma_2)$, so that $((1-s) - \sqrt{s(1-s)})(\sigma_1 - \sigma_2) = 0$ which implies $s = 1/2$. Thus we obtain $I_k(\tau, x) = (\sigma_1 - \sigma_2)^2/4$.

Now we substitute these to (55). By summing up equations (55) from $i = 1$ to $i = \ell$ and from $i = \ell + 1$ to $i = k$ respectively, then divide both sides by 2, we obtain

$$(1 - 2\mu\ell)\sigma_1 - \sigma_2 = 0, \qquad -\sigma_1 + (1 - 2\mu(k - \ell))\sigma_2 = 0.$$

This system of equations has a non-trivial solution if and only if the determinant of the coefficient matrix equals zero which is equivalent to $2\mu^2\ell(k-\ell)-k\mu=0$. If $\mu=0$, then $\sigma_1=\sigma_2$ which is not possible under our assumptions. If $\mu=k/(2\ell(k-\ell))$, we consider the constraint $|x|=1$ for $x\in\mathbb{R}^k$. We see that the two solutions must be $\sigma_1=\sqrt{(k-\ell)/(\ell k)}$ and $\sigma_2=-\sqrt{\ell/(k-\ell)k}$ so that

$$I_k(\tau,x)=\frac{(\sigma_1-\sigma_2)^2}{4}=\frac{1}{4k}\left(\frac{k-\ell}{\ell}+\frac{\ell}{k-\ell}+2\right).$$

It is easy to see that the right hand side of the above reaches its maximum if $\ell=1$ or $\ell=k-1$. In both cases we have $I_k(\tau,x)=k/(4(k-1))$. The maximum is reached if we take $k=2$ which gives $I_2(\tau,x)=1/2$ at $\tau=(1/2,\,1/2)$ and $x=(1/\sqrt{2},\,-1/\sqrt{2})$. □

Proof of Lemma 4.6. We will use the following simple fact repeatedly below that for $\tau>0$,

$$\frac{1}{1+\tau}\le\log\left(1+\frac{1}{\tau}\right)\le\frac{1}{\tau}.$$

We first prove the first claim in (48) which is a simple calculus question. We have for $\lambda>1$ that

$$\begin{aligned}0\le\frac{\log\lambda}{\lambda}\int_0^1\frac{1}{\log\left(1+\frac{1}{\lambda^{\sigma t}}\right)}\,\mathrm{d}t&\le\frac{\log\lambda}{\lambda}\frac{1}{\log\left(1+\frac{1}{\lambda^{\sigma}}\right)}\\&\le\frac{\log\lambda}{\lambda}(\lambda^\sigma+1)=\frac{\log\lambda}{\lambda}+\frac{\log\lambda}{\lambda^{1-\sigma}}\to0\text{ as }\lambda\to+\infty.\end{aligned}$$

Now we prove the second claim in (48). Let $\epsilon(\lambda)=1/\log\lambda$ for $\lambda>1$, we have

$$\begin{aligned}\frac{\log\lambda}{\lambda}\int_0^1\frac{1}{\log\left(1+\frac{1}{\lambda^t}\right)}\,\mathrm{d}t&\ge\frac{\log\lambda}{\lambda}\int_{1-\epsilon(\lambda)}^1\frac{1}{\log\left(1+\frac{1}{\lambda^t}\right)}\,\mathrm{d}t\\&\ge\frac{\epsilon(\lambda)\log\lambda}{\lambda}\lambda^{1-\epsilon(\lambda)}=\frac{\epsilon(\lambda)\log\lambda}{\lambda^{\epsilon(\lambda)}}=\frac{1}{e}.\end{aligned}$$

The conclusion follows.

Next we establish the first limit in (49). Let $\epsilon(\lambda)=1/\sqrt{\log\lambda}$. We have

$$\frac{\log\lambda}{\lambda}\int_0^1\frac{(1-t)^2}{\log\left(1+\frac{1}{\lambda^t}\right)}\,\mathrm{d}t=I_1+I_2,$$

where

$$\begin{aligned}0\le I_1&:=\frac{\log\lambda}{\lambda}\int_0^{1-\epsilon(\lambda)}\frac{(1-t)^2}{\log\left(1+\frac{1}{\lambda^t}\right)}\,\mathrm{d}t\le\frac{\log\lambda}{\lambda}\frac{1}{\log\left(1+\frac{1}{\lambda^{1-\epsilon(\lambda)}}\right)}\,\mathrm{d}t\\&\le\frac{\log\lambda}{\lambda}(1+\lambda^{1-\epsilon(\lambda)})=\frac{\log\lambda}{\lambda}+\frac{\log\lambda}{\lambda^{\epsilon(\lambda)}}=\frac{\log\lambda}{\lambda}+\frac{\log\lambda}{\exp(\epsilon(\lambda)\log\lambda)}\\&=\frac{\log\lambda}{\lambda}+\frac{\log\lambda}{\exp(\sqrt{\log\lambda})}\to0\quad\text{as }\lambda\to+\infty.\end{aligned}$$

We also have

$$
\begin{aligned}
0 \le I_2 &:= \frac{\log\lambda}{\lambda}\int_{1-\epsilon(\lambda)}^{1}\frac{(1-t)^2}{\log\left(1+\frac{1}{\lambda^t}\right)}\,dt \le \frac{\log\lambda}{\lambda\log\left(1+\frac{1}{\lambda}\right)}\int_{1-\epsilon(\lambda)}^{1}(1-t)^2\,\mathrm{d}t\\
&\le \frac{\log\lambda}{\lambda}(1+\lambda)\int_0^{\epsilon(\lambda)} s^2\,\mathrm{d}s = \frac{(\epsilon(\lambda))^3\log\lambda}{3\lambda}(1+\lambda)\\
&= \frac{1}{3\lambda\sqrt{\log\lambda}}+\frac{1}{3\sqrt{\log\lambda}} \to 0 \quad \text{as } \lambda\to+\infty.
\end{aligned}
$$

The result then follows.

Finally we establish the second limit in (49). Since $\sigma > 1$, we may fix $0 < \epsilon < 1$ such that $(1-\epsilon)\sigma > 1$. Then we have

$$
\begin{aligned}
&\frac{\log\lambda}{\lambda}\int_0^1\frac{(1-t)^2}{\log\left(1+\frac{1}{\lambda^{\sigma t}}\right)}\,\mathrm{d}t \ge \frac{\log\lambda}{\lambda}\int_{1-\epsilon}^1\frac{(1-t)^2}{\log\left(1+\frac{1}{\lambda^{\sigma t}}\right)}\,\mathrm{d}t\\
&\ge \frac{\log\lambda}{\lambda\log\left(1+\frac{1}{\lambda^{\sigma(1-\epsilon)}}\right)}\int_{1-\epsilon}^1(1-t)^2\,\mathrm{d}t \ge \frac{\epsilon^3\lambda^{\sigma(1-\epsilon)}\log\lambda}{3\lambda} \to +\infty \text{ as } \lambda\to+\infty. \qquad \square
\end{aligned}
$$

Proof of Theorem 4.4 We first consider the special case (ii) where we have $m = 2$ with $0 < \alpha \le 2$. By Lemma 2.1 we obtain

$$
C^u_\lambda(f_2)(x) - f_2(x) = \begin{cases}
\frac{\lambda}{2}\left(x_1 - x_2 - \frac{1}{2\lambda}\right)^2 & 0 \le x_1 - x_2 \le \frac{1}{2\lambda},\\
\frac{\lambda}{2}\left(x_2 - x_1 - \frac{1}{2\lambda}\right)^2 & 0 \le x_2 - x_1 \le \frac{1}{2\lambda},\\
0 & |x_1 - x_2| \ge \frac{1}{2\lambda}.
\end{cases}
$$

Thus

$$
\begin{aligned}
\int\limits_{\bar B_r(0)}\frac{C^u_\lambda(f_2)(x) - f_2(x)}{H^{(2)}_{\mu_\alpha(\lambda)}(x) - f_2(x)}\,\mathrm{d}x &= \frac{\lambda}{2}\int\limits_{U_{1,2}}\frac{\left(x_1 - x_2 - \frac{1}{2\lambda}\right)^2}{\frac{1}{\mu_\alpha(\lambda)}\log(1+\exp[-\mu_\alpha(\lambda)(x_1 - x_2)])}\,\mathrm{d}x\\
&+\frac{\lambda}{2}\int\limits_{U_{2,1}}\frac{\left(x_2 - x_1 - \frac{1}{2\lambda}\right)^2}{\frac{1}{\mu_\alpha(\lambda)}\log(1+\exp[-\mu_\alpha(\lambda)(x_2 - x_1)])}\,\mathrm{d}x := I_1 + I_2, \qquad (56)
\end{aligned}
$$

for $U_{1,2} = \left\{x \in \bar B_r(0) : 0 \le x_1 - x_2 \le \frac{1}{2\lambda}\right\}$, $U_{2,1} = \left\{x \in \bar B_r(0) : 0 \le x_2 - x_1 \le \frac{1}{2\lambda}\right\}$. For I_1 in (6.1) we change variables by a rotation $u = (u_1, u_2)$ with $u_1 = \frac{1}{\sqrt{2}}(x_1 - x_2)$, $u_2 = \frac{1}{\sqrt{2}}(x_1 + x_2)$. Thus in $U_{1,2}$ we have $0 \le u_1 \le \frac{1}{2\sqrt{2}\lambda} < r$ and $|u_2| \le r$ when $\lambda > 1$ is sufficiently large. Therefore we have

$$
\begin{aligned}
0 \le I_1 &= \frac{\lambda}{2}\int_{U_{1,2}}\frac{\left(x_1 - x_2 - \frac{1}{2\lambda}\right)^2}{\frac{1}{\mu_\alpha(\lambda)}\log(1+\exp[-\mu_\alpha(\lambda)(x_1 - x_2)])}\,\mathrm{d}x\\
&= \frac{\lambda}{2}\int_{V_{1,2}}\frac{\left(\sqrt{2}u_1 - \frac{1}{2\lambda}\right)^2}{\frac{1}{\mu_\alpha(\lambda)}\log(1+\exp[-\mu_\alpha(\lambda)\sqrt{2}u_1])}\,\mathrm{d}u_1\,\mathrm{d}u_2
\end{aligned}
$$

$$
\begin{aligned}
&\le \frac{\lambda\mu_\alpha(\lambda)}{2}\int_{-r}^{r}\mathrm{d}u_2\int_0^{\frac{1}{2\sqrt{2}\lambda}}\frac{\left(\sqrt{2}u_1-\frac{1}{2\lambda}\right)^2}{\log(1+\exp[-\mu_\alpha(\lambda)\sqrt{2}u_1])}\,\mathrm{d}u_1\\
&= r\lambda\mu_\alpha(\lambda)\int_0^{\frac{1}{2\sqrt{2}\lambda}}\frac{\left(\sqrt{2}u_1-\frac{1}{2\lambda}\right)^2}{\log(1+\exp[-\mu_\alpha(\lambda)\sqrt{2}u_1])}\,\mathrm{d}u_1 := J, \qquad (57)
\end{aligned}
$$

where $V_{1,2}=\left\{u\in\bar{B}_r(0)\,:\,0\le u_1\le\dfrac{1}{2\sqrt{2}\lambda}\right\}$. Now in the last integral J in (57) we make a change of variable $u_1=\dfrac{t}{2\sqrt{2}\lambda}$, and we obtain

$$
\begin{aligned}
J &= r\lambda\mu_\alpha(\lambda)\int_0^{\frac{1}{2\sqrt{2}\lambda}}\frac{\left(\sqrt{2}u_1-\frac{1}{2\lambda}\right)^2}{\log(1+\exp[-\mu_\alpha(\lambda)\sqrt{2}u_1])}\,\mathrm{d}u_1\\
&= \frac{r\lambda\mu_\alpha(\lambda)}{4\lambda^2}\int_0^1\frac{(t-1)^2}{\log\left(1+\exp\left(-\frac{\mu_\alpha(\lambda)}{2\lambda}t\right)\right)}\frac{\mathrm{d}t}{2\sqrt{2}\lambda}\\
&= \frac{\alpha r}{8\sqrt{2}}\frac{\log\lambda}{\lambda}\int_0^1\frac{(t-1)^2}{\log\left(1+\frac{1}{\lambda^{\alpha t/2}}\right)}\,\mathrm{d}t\\
&\le \frac{2r}{8\sqrt{2}}\left[\frac{\log\lambda}{\lambda}\int_0^1\frac{(t-1)^2}{\log\left(1+\frac{1}{\lambda^t}\right)}\,\mathrm{d}t\right]\to 0 \quad \text{as } \lambda\to+\infty
\end{aligned}
$$

by the first limit in (48) in Lemma 4.6. Thus $I_1\to 0$ as $\lambda\to+\infty$. Similarly we see that $I_2\to 0$ as $\lambda\to+\infty$. Therefore the proof for the case $m=2$, $0<\alpha\le 2$ is complete.

Next we consider case (iii) when $m=2$ and $\alpha>2$. We define for $\lambda>\max\{1,\frac{\sqrt{2}}{r}\}$ the set $U_+:=\{x\in\mathbb{R}^2\,:\,0\le x_1-x_2\le\frac{1}{2\lambda},\ |x_1|+|x_2|\le r\}\subseteq\bar{B}_r(0)$. Similar to the proof of (ii), we make the same change of variables $u=(u_1,u_2)$ as above in U_+. In our current situation we have $0\le u_1\le\frac{1}{2\sqrt{2}}\lambda$ and $|u_2|\le r/\sqrt{2}$. Thus

$$
\begin{aligned}
&\int\limits_{\bar{B}_r(0)}\frac{C_\lambda^u(f_2)(x)-f_2(x)}{H^{(2)}_{\mu_\alpha(\lambda)}(x)-f_2(x)}\,\mathrm{d}x\\
&\ge \frac{\lambda}{2}\int\limits_{U_+}\frac{\left(x_1-x_2-\frac{1}{2\lambda}\right)^2}{\frac{1}{\mu_\alpha(\lambda)}\log(1+\exp[-\mu_\alpha(\lambda)(x_1-x_2)])}\,\mathrm{d}x\\
&= \frac{\lambda\mu_\alpha(\lambda)}{2}\int\limits_{-r/2}^{r/2}\mathrm{d}u_2\int\limits_0^{\frac{1}{2\sqrt{2}\lambda}}\frac{\left(\sqrt{2}u_1-\frac{1}{2\lambda}\right)^2}{\log(1+\exp[-\mu_\alpha(\lambda)\sqrt{2}u_1])}\,\mathrm{d}u_1\\
&= \frac{r\lambda\mu_\alpha(\lambda)}{2}\int\limits_0^{\frac{1}{2\sqrt{2}\lambda}}\frac{\left(\sqrt{2}u_1-\frac{1}{2\lambda}\right)^2}{\log(1+\exp[-\mu_\alpha(\lambda)\sqrt{2}u_1])}\,\mathrm{d}u_1\\
&= \frac{r\lambda\mu_\alpha(\lambda)}{8\lambda^2}\int\limits_0^1\frac{(t-1)^2}{\log\left(1+\exp\left(-\frac{\mu_\alpha(\lambda)}{2\lambda}t\right)\right)}\frac{\mathrm{d}t}{2\sqrt{2}\lambda}
\end{aligned}
$$

$$= \frac{\alpha r}{16\sqrt{2}}\left[\frac{\log\lambda}{\lambda}\int_0^1 \frac{(t-1)^2}{\log\left(1+\frac{1}{\lambda^{\alpha t/2}}\right)}\,\mathrm{d}t\right] \to +\infty \text{ as } \lambda\to+\infty, \tag{58}$$

following the second limit in (49) as $\alpha/2 > 1$.

Finally we consider case (i): $m \geq 2$ and $0 < \alpha < 2$. From the remark following Corollary 2.3, we observe that $C^u_\lambda(f_m)(x) - f_m(x) > 0$ in $\mathbb{R}^m$ if and only if we have $f_m(x) - \hat{f}_m(x) < \frac{1}{2\lambda}$, that is, the difference between the largest component $f_m(x)$ of $x = (x_1, x_2, \ldots, x_m)$ and the second largest component $\hat{f}_m(x)$ is less than $\frac{1}{2\lambda}$. If we define

$$W_{i,j} := \left\{x \in \bar{B}_r(0) : 0 \leq x_i - x_j \leq \frac{1}{2\lambda},\ x_j \geq x_k \text{ for } 1 \leq k \leq m,\ k \neq i,\ j\right\},$$

then we see that the set $W := \left\{x \in \bar{B}_r(0) : C^u_\lambda(f_m)(x) - f_m(x) > 0\right\} \subseteq \bigcup_{1\leq i\neq j\leq m} W_{i,j}$.

Thus we have

$$\begin{aligned} 0 &\leq \int_{\bar{B}_r(0)} \frac{C^u_\lambda(f_m)(x) - f_m(x)}{H^{(m)}_{\mu_\alpha(\lambda)}(x) - f_m(x)}\,\mathrm{d}x \leq \int_{\bigcup_{1\leq i\neq j\leq m} W_{i,j}} \frac{C^u_\lambda(f_m)(x) - f_m(x)}{H^{(m)}_{\mu_\alpha(\lambda)}(x) - f_m(x)}\,\mathrm{d}x \\ &\leq \sum_{1\leq i\neq j\leq m} \int_{W_{i,j}} \frac{\frac{1}{4\lambda}}{\frac{1}{\mu_\alpha(\lambda)}\log\left(1+\exp[-\mu_\alpha(\lambda)(x_i-x_j)]\right)}\,\mathrm{d}x := \sum_{1\leq i\neq j\leq m} J_{i,j}. \end{aligned} \tag{59}$$

Here we have used the error estimate (23) in Corollary 2.5 in the numerator of the integrand in $W_{i,j}$, and the fact that $x_i \geq x_j \geq x_k$ for $k \neq i,\ j$ in the denominator so that

$$\begin{aligned} & H^{(m)}_{\mu_\alpha(\lambda)}(x) - f_m(x) \\ = & \frac{1}{\mu_\alpha(\lambda)}\log\left(1 + \exp[-\mu_\alpha(\lambda)(x_i - x_j)] + \sum_{1\leq k\leq m,\, k\neq i,\, j}^{m} \exp[-\mu_\alpha(\lambda)(x_i - x_k)]\right) \\ \geq & \frac{1}{\mu_\alpha(\lambda)}\log\left(1+\exp[-\mu_\alpha(\lambda)(x_i - x_j)]\right) \end{aligned}$$

for $x \in W_{i,j}$. Obviously if we can prove that $J_{1,2} \to 0$ as $\lambda \to +\infty$, where $J_{i,j}$'s are defined in (59), the proof is then complete as the only difference between $J_{1,2}$ and other $J_{i,j}$'s is the locations of component indices.

Similar to the proof of the first part, we change variables in the (x_1, x_2)-plane by the same rotation as above. We set $u_1 = \frac{1}{\sqrt{2}}(x_1 - x_2)$, $u_2 = \frac{1}{\sqrt{2}}(x_1 + x_2)$, and $u_k = x_k$ for $3 \leq k \leq m$. We also write $\hat{u} = (u_3, \ldots, u_m) \in \mathbb{R}^{m-2}$ and $\bar{B}_{r,m-2}(0) \subseteq \mathbb{R}^{m-2}$ be the closed ball centred at 0 with radius r. Thus in $W_{1,2}$ we have $0 \leq u_1 \leq \frac{1}{2\sqrt{2}\lambda} < r$ and $|u_2| \leq \sqrt{x_1^2 + x_2^2} \leq r$.

We then have

$$
\begin{aligned}
0 \le J_{1,2} &:= \int\limits_{W_{1,2}} \frac{\frac{1}{4\lambda}}{\frac{1}{\mu_\alpha(\lambda)} \log\left(1+\exp[-\mu_\alpha(\lambda)(x_1-x_2)]\right)}\,\mathrm{d}x \\
&\le \frac{\alpha \log \lambda}{4} \int\limits_{\bar{B}_{r,m-2}(0)} \mathrm{d}\hat{u} \int_{-r}^{r} \mathrm{d}u_2 \int_0^{\frac{1}{2\sqrt{2}\lambda}} \frac{\mathrm{d}u_1}{\log(1+\exp[-\mu_\alpha(\lambda)\sqrt{2}u_1])} \\
&= \frac{\alpha r}{4\sqrt{2}} \mathrm{vol}(\bar{B}_{r,m-2}(0)) \left(\frac{\log \lambda}{\lambda} \int_0^1 \frac{\mathrm{d}t}{\log(1+\frac{1}{\lambda^{\alpha t/2}})} \right) \to 0, \text{ as } \lambda \to +\infty
\end{aligned}
$$

by (48) in Lemma 4.6, where $\mathrm{vol}(\bar{B}_{r,m-2}(0))$ is the volume of the $(m-2)$-dimensional ball. Thus $J_{1,2} \to 0$ as $\lambda \to +\infty$, and the result follows. □

References

[1] H. Bauschke, J. M. Borwein: *On projection algorithms for solving convex feasibility problems*, SIAM Review 38/3 (1996) 367–426.

[2] A. Beck, M. Teboulle: *Smoothing and first order methods: A unified framework*, SIAM J. Optimization 22/2 (2012) 557–580.

[3] D. P. Bertsekas: *Nondifferentiable optimization via approximation*, Math. Programming Study 3 (1975) 1–25.

[4] P. Blanchard, D. J. Higham, N. J. Higham: *Accurately computing the log-sum-exp and softmax functions*, IMA J. Numer. Analysis 41/4 (2021) 2311–2330.

[5] J. P. Boyle, R. L. Dykstra: *A Method for Finding Projections onto the Intersection of Convex Sets in Hilbert Spaces*, in: *Advances in Order Restricted Statistical Inference*, Lecture Notes in Statistics 37, Springer, New York (1986) 28–47.

[6] D. Butnariu, Y. Censor, P. Gurfil, E. Hadar: *On the behavior of subgradient projections methods for convex feasibility problems in Euclidean spaces*, SIAM J. Optimization 19/2 (2008) 786–807.

[7] Y. Censor, A. Lent: *Cyclic subgradient projections*, Math. Programming 24/1 (1982) 233–235.

[8] C. Chen, O. L. Mangasarian: *A class of smoothing functions for nonlinear and mixed complementarity problems*, Comp. Optimization Appl. 5/2 (1996) 97–138.

[9] X. Chen: *Smoothing methods for nonsmooth, nonconvex minimization*, Math. Programming 134/1 (2012) 71–99.

[10] P. L. Combettes: *Hilbertian convex feasibility problem: Convergence of projection methods*, Appl. Math. Optimization 35/3 (1997) 311–330.

[11] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein: *Introduction to Algorithms*, MIT Press, Cambridge (2001).

[12] R. L. Dykstra: *An algorithm for restricted least squares regression*, J. Amer. Statist. Assoc. 78/384 (1983) 837–842.

[13] R. Escalante, M. Raydan: *Alternating Projection Methods*, Fundamentals of Algorithms 8, SIAM, Philadelphia (2011).

[14] J.-B. Hiriart-Urruty, C. Lemaréchal: *Fundamentals of Convex Analysis*, Grundlehren, Springer, Berlin (2001).

[15] J. Hoschek, D. Lasser, L. L. Schumaker: *Fundamentals of Computer Aided Geometric Design*, A. K. Peters, Wellesley (1993).

[16] P. J. Huber: *Robust estimation of a location parameter*, in: *Breakthroughs in Statistics. Volume 2*, Springer Series in Statististics, Springer, New York (1992) 492–518.

[17] B. Kirchheim, J. Kristensen: *Differentiability of convex envelopes*, C. Rend. Acad. Sci., Series I Math. 333/8 (2001) 725–728.

[18] X.-S. Li, S.-C. Fang: *On the entropic regularization method for solving min-max problems with applications*, Math. Meth. Oper. Res. 46/1 (1997) 119–130.

[19] M. López, G. Still: *Semi-infinite programming*, Eur. J. Oper. Res. 180/2 (2007) 491–518.

[20] Yu. Nesterov: *Smooth minimization of non-smooth functions*, Math. Programming 103/1 (2005) 127–152.

[21] Yu. Nesterov: *Gradient methods for minimizing composite functions*, Math. Programming 140/1 (2013) 125–161.

[22] E. Y. Pee, J. O. Royset: *On solving large-scale finite minimax problems using exponential smoothing*, J. Optimization Theory Appl. 148/2 (2011) 390–421.

[23] E. Polak, J. O. Royset, R. S. Womersley: *Algorithms with adaptive smoothing for finite minimax problems*, J. Optimization Theory Appl. 119/3 (2003) 459–484.

[24] R. T. Rockafellar: *Convex Analysis*, Princeton Landmarks in Mathematics 28, Princeton University Press, Princeton (1970).

[25] J. O. Royset, E. Y. Pee: *Rate of convergence analysis of discretization and smoothing algorithms for semiinfinite minimax problems*, J. Optimization Theory Appl. 155/3 (2012) 855–882.

[26] S. Xu: *Smoothing method for minimax problems*, Comp. Optimization Appl. 20/3 (2001) 267–279.

[27] H.-X. Yin: *Error bounds of two smoothing approximations for semi-infinite minimax problems*, Acta Math. Appl. Sinica, Engl. Ser. 25/4 (2009) 685–696.

[28] S. Žaković, B. Rustem: *Semi-infinite programming and applications to minimax problems*, Ann. Oper. Res. 124/1-4 (2003) 81–110.

[29] I. Zang: *A smoothing-out technique for min-max optimization*, Math. Programming 19/1 (1980) 61–77.

[30] K. Zhang: *Compensated convexity and its applications*, Ann. Inst. H. Poincaré (C), Non Linear Analysis 25/4 (2008) 743–771.

[31] K. Zhang: *Convex analysis based smooth approximations of maximum functions and squared-distance functions*, J. Nonlinear Convex Analysis 9 (2008) 379–406.

[32] K. Zhang, E. Crooks, A. Orlando: *Compensated convexity, multiscale medial axis maps and sharp regularity of the squared-distance function*, SIAM J. Math. Analysis 47/6 (2015) 4289–4331.

[33] K. Zhang, E. Crooks, A. Orlando: *Compensated convexity methods for approximations and interpolations of sampled functions in Euclidean spaces: Theoretical foundations*, SIAM J. Math. Analysis 48/6 (2016) 4126–4154.

[34] K. Zhang, E. Crooks, A. Orlando: *Compensated convexity methods for approximations and interpolations of sampled functions in Euclidean spaces: Applications to contour lines, sparse data, and inpainting*, SIAM J. Imaging Sciences 11/4 (2018) 2368–2428.

[35] K. Zhang, A. Orlando, E. Crooks: *Compensated convexity and Hausdorff stable geometric singularity extractions*, Math. Models Meth. Appl. Sci. 25/4 (2015) 747–801.

[36] K. Zhang, A. Orlando, E. Crooks: *Compensated convexity and Hausdorff stable extraction of intersections for smooth manifolds*, Math. Models Meth. Appl. Sci. 25/5 (2015) 839–873.