Electronic Journal of Statistics
Vol. 11 (2017) 2547–2612
ISSN: 1935-7524
DOI: 10.1214/17-EJS1294

Multinomial and empirical likelihood under convex constraints: Directions of recession, Fenchel duality, the PP algorithm

Marian Grendár

Slovanet a.s., Záhradnícka 151, 821 08 Bratislava, Slovakia

Biomedical Center Martin, Jessenius Faculty of Medicine in Martin, Comenius University in Bratislava, Malá Hora 4C, 036 01 Martin, Slovakia

Institute of Measurement Science, Slovak Academy of Sciences, Dúbravská cesta 9, 841 04 Bratislava, Slovakia
e-mail: marian.grendar@jfmed.uniba.sk

and

Vladimír Špitalský

Slovanet a.s., Záhradnícka 151, 821 08 Bratislava, Slovakia

Department of Mathematics, Faculty of Natural Sciences, Matej Bel University, Tajovského 40, 974 01 Banská Bystrica, Slovakia
e-mail: vladimir.spitalsky@slovanet.net

Abstract: The primal problem of multinomial likelihood maximization restricted to a convex closed subset of the probability simplex is studied. A solution of this problem may assign a positive mass to an outcome with zero count. Such a solution cannot be obtained by the widely used, simplified Lagrange and Fenchel duals. Related flaws in the simplified dual problems, which arise because the recession directions are ignored, are identified and the correct Lagrange and Fenchel duals are developed.

The results permit us to specify linear sets and data such that the empirical likelihood-maximizing distribution exists and is the same as the multinomial likelihood-maximizing distribution. The multinomial likelihood ratio reaches, in general, a different conclusion than the empirical likelihood ratio.

Implications for minimum discrimination information, Lindsay geometry, compositional data analysis, bootstrap with auxiliary information, and Lagrange multiplier test, which explicitly or implicitly ignore information about the support, are discussed.

A solution of the primal problem can be obtained by the PP (perturbed primal) algorithm, that is, as the limit of a sequence of solutions of perturbed primal problems. The PP algorithm may be implemented by the simplified Lagrange or Fenchel dual.

MSC 2010 subject classifications: Primary 62H12, 62H17; secondary 90C46.
Keywords and phrases: Closed multinomial distribution, estimating equation, contingency table, zero cell frequency, El Barmi Dykstra dual, Smith

dual, PP algorithm, epi-convergence, Fisher likelihood, minimum discrimination information.

Received July 2016.

Contents

1. Introduction

Zero counts are a source of difficulties in the maximization of the multinomial likelihood for log-linear models. A considerable literature has been devoted to this issue, culminating in the recent papers by Fienberg and Rinaldo [14] and Geyer [16]. In these studies, convex analysis considerations play a key role.

Less well recognized is that the zero counts also cause difficulties in the maximization of the multinomial likelihood under linear constraints, or, in general, when the cell probabilities are restricted to a convex closed subset of the probability simplex; see Section 2 for a formal statement of the considered primal optimization problem $\mathcal{P}$. Though in this case the nature of the difficulties is different than in the log-linear case, the convex analysis considerations are important here as well, because they permit developing a correct solution of $\mathcal{P}$ – one of the main objectives of the present work.

The problem of finding the maximum multinomial likelihood under linear constraints dates back to, at least, Smith [45], and continues through the work of Aitchison and Silvey [4], Gokhale [18], Klotz [28], Haber [21], Stirling [46], Pelz and Good [36], Little and Wu [32], El Barmi and Dykstra [12, 13], to the recent studies by Agresti and Coull [2], Lang [29], and Bergsma et al. [9], among others. Linear constraints on the cell probabilities appear naturally in marginal homogeneity models, isotonic cone models, mean response models, multinomial-Poisson homogeneous models, and many others; cf. Agresti [1], Bergsma et al. [9]. They also arise in the context of estimating equations.

A solution of $\mathcal{P}$ may assign a positive weight to an outcome with zero count; cf. Theorem 2. This fact affects the Lagrange and Fenchel dual problems to $\mathcal{P}$.

The restricted maximum of the multinomial likelihood defined through the primal problem $\mathcal{P}$ is not amenable to asymptotic analysis, and the primal form is not ideal for numerical optimization. Thus, it is common to consider the Lagrange dual problem instead of the primal problem. This permits the asymptotic analysis (cf. Aitchison and Silvey [4]), and reduces the dimension of the optimization problem, because the number of linear constraints is usually much smaller than the cardinality of the sample space. Smith [45, Sects. 6, 7] has developed a solution of the Lagrange dual problem, under the hidden assumption that every outcome from the sample space appears in the sample at least once; that is, $\nu > 0$, where ν is the vector of the observed relative frequency of outcomes. The same solution was later considered by several authors; see, in particular, Haber [21, p. 3], Little and Wu [32, p. 88], Lang [29, Sect. 7.1], Bergsma et al. [9, p. 65]. It remained unnoticed that, if the assumption $\nu > 0$ is not satisfied, then the solution of Smith's Lagrange dual problem does not necessarily lead to a solution of the primal problem $\mathcal{P}$.

El Barmi and Dykstra [12] studied the maximization of the multinomial likelihood under more general, convex set constraints, where it is natural to replace the Lagrange duality with the Fenchel duality. When the feasible set is defined by the linear constraints, El Barmi and Dykstra's (BD) dual $\mathcal{B}$ reduces to Smith's Lagrange dual. The BD-dual $\mathcal{B}$ leads to a solution of the primal $\mathcal{P}$ if $\nu > 0$. The authors overlooked that this is not necessarily the case if a zero count occurs.

Taken together, the decisions obtained from El Barmi and Dykstra's simplified Fenchel dual $\mathcal{B}$ can be severely compromised. It is thus important to know the correct Fenchel dual $\mathcal{F}$ to $\mathcal{P}$. This is provided by Theorem 6, which also characterizes the solution set of $\mathcal{F}$. It is equally important to know the conditions under which the BD-dual $\mathcal{B}$ leads to a solution of $\mathcal{P}$. The answer is provided by Theorem 16. An analysis of directions of recession is crucial for establishing the theorem.

The findings have implications for the empirical likelihood. Recall that 'in most settings, empirical likelihood is a multinomial likelihood on the sample'; cf. Owen [35, p. 15]. As the empirical likelihood inner problem $\mathcal{E}$ (cf. Section 7) is a convex optimization problem, it has its Fenchel dual formulation. If the feasible set is given by linear equality constraints, then the Fenchel dual to $\mathcal{E}$ is equivalent to El Barmi and Dykstra's dual $\mathcal{B}$ to $\mathcal{E}$. Thanks to this connection, Theorem 16 provides conditions under which the solution set $S_{\mathcal{P}}$ of the multinomial likelihood primal problem $\mathcal{P}$ and the solution set $S_{\mathcal{E}}$ of the empirical likelihood inner problem $\mathcal{E}$ are the same, and the maximum $\hat{L}$ of the multinomial likelihood is equal to the maximum $\hat{L}_{\mathcal{E}}$ of the empirical likelihood. Consequently:

- If C is an H-set or a Z-set with respect to the type ν (for the definition, see Section 4.3), the maximum empirical likelihood does not exist, though the maximum multinomial likelihood exists. The notion of H-set corresponds to the convex hull problem (cf. Owen [34, Sect. 10.4]) and the notion of Z-set corresponds to the zero likelihood problem (cf. Bergsma et al. [8]). By Theorem 16, these are the only ways the empirical likelihood inner problem may fail to have a solution; cf. Section 7. Note, that also the empirical likelihood outer problem may have no solution; cf. the empty set problem, Grendár and Judge [19].
- If any of conditions (i)–(iv) in Theorem 16(b) are not satisfied, then $\hat{L}_{\mathcal{E}} < \hat{L}$, and the empirical likelihood may lead to different inferential and evidential conclusions than those suggested by the multinomial likelihood.

Fisher's [15] original concept of the likelihood carries the discordances between the multinomial and empirical likelihoods also into the continuous *iid* setting; cf. Section 7.1.

The findings also affect other methods, such as the minimum discrimination information, compositional data analysis, Lindsay geometry of multinomial mixtures, bootstrap with auxiliary information, and Lagrange multiplier test, which explicitly or implicitly ignore information about the support and are restricted to the observed outcomes.

Despite its flawed relation to the primal, the BD-dual may be utilized in an algorithm for obtaining a solution of the primal $\mathcal{P}$. The PP algorithm forms a sequence of *p*erturbed *p*rimal problems. Theorem 20 demonstrates that the PP algorithm epi-converges to a solution of $\mathcal{P}$. Even stronger, pointwise convergence can be established for a linear constraint set; see Theorem 21. The convergence theorems imply that the common practice of replacing the zero counts by a small, arbitrary value can be supplanted by a sequence of perturbed primal problems, where the $\delta-$perturbed relative frequency vectors $\nu(\delta) > 0$ are such that $\lim_{\delta \searrow 0} \nu(\delta) = \nu$. Because each $\nu(\delta)$ is strictly positive, the PP algorithm can be implemented through the BD-dual to the perturbed primal. The strict positivity also allows using the Fisher scoring algorithm, Gokhale's algorithm [18], or similar methods, for implementing the PP algorithm.

1.1. Organization of the paper

The multinomial likelihood primal problem $\mathcal{P}$ and its characterization (cf. Theorem 2) are presented in Section 2. The Fenchel dual problem $\mathcal{F}$ to $\mathcal{P}$ is introduced in Section 3. A Lagrange dual formulation of the convex conjugate (cf. Theorem 5) serves as a ground for Theorem 6, one of the main results, which provides a relation between the solutions of $\mathcal{P}$ and $\mathcal{F}$. If the feasible set C is polyhedral, a solution of $\mathcal{F}$ can be obtained also from a different Lagrange dual to $\mathcal{P}$; cf. Section 3.1. A special case of the single inequality constraint is discussed in detail in Section 3.2, where a flaw in Klotz's [28] Theorem 1 is noted. In Section 4.1, El Barmi and Dykstra's [12] dual $\mathcal{B}$ is recalled; Section 4.1.1 introduces its special case, the Smith dual problem. Theorem 2.1 of El Barmi and Dykstra [12], and its flaws are presented in Section 4.2, where they are also illustrated by simple examples. Section 4.3 studies the scope of validity of the BD-dual $\mathcal{B}$; cf. Theorem 16. Sequential, active-passive dualization is proposed and analyzed in Section 5. Perturbed primal problem $\mathcal{P}_\delta$ is introduced in Section 6, where the epi-convergence of a sequence of the perturbed primals for a general, convex C, and the pointwise convergence for the linear C are formulated (cf. Theorems 20, 21) and illustrated. Implications of the results for the empirical likelihood method are discussed in Section 7. A brief discussion of implications of the findings for the minimum discrimination information, compositional data analysis, Lindsay geometry of multinomial mixtures, bootstrap with auxiliary information and Lagrange multiplier test is contained in Section 8. Some of the computational and applied aspects of the presented results are summarized in Section 9. Finally, Section 10 comprises detailed proofs of the results.

An `R` code and data to reproduce the numerical examples can be found in [20].

2. Multinomial likelihood primal problem $\mathcal{P}$

Annotation. The primal problem $\mathcal{P}$ is formulated and a basic characterization of its solution is given. The primal has always a solution. Its active coordinates are unique. A solution of $\mathcal{P}$ may assign positive mass to passive letter(s).

Let $\mathcal{X}$ denote a finite *alphabet* (sample space) consisting of m *letters* (outcomes) and $\Delta_{\mathcal{X}}$ denote the *probability simplex*

$$\Delta_{\mathcal{X}} \triangleq \left\{ q \in \mathbb{R}^m : \ q \geq 0, \sum q = 1 \right\};$$

identify $\mathbb{R}^m$ with $\mathbb{R}^{\mathcal{X}}$. Suppose that $(n_i)_{i \in \mathcal{X}}$ is a *realization* of the *closed multinomial distribution* $\Pr((n_i)_{i \in \mathcal{X}}; n, q) = n! \prod q_i^{n_i}/n_i!$ with parameters $n \in \mathbb{N}$ and

$q = (q_i)_{i\in\mathcal{X}} \in \Delta_{\mathcal{X}}$. Then the *multinomial likelihood kernel* $L(q) = L_\nu(q) \triangleq e^{-n\,\ell(q)}$, where *Kerridge's inaccuracy* [27] $\ell = \ell_\nu : \Delta_{\mathcal{X}} \to \bar{\mathbb{R}}$, is

$$\ell(q) \triangleq -\langle \nu, \log q\rangle\,, \tag{2.1}$$

and $\nu \triangleq (n_i/n)_{i\in\mathcal{X}}$ is the *type* (the vector of the relative frequency of outcomes). The conventions $\log 0 = -\infty$, $0\cdot(-\infty) = 0$ apply; $\bar{\mathbb{R}}$ denotes the extended real line $[-\infty,\infty]$ and $\langle a, b\rangle$ is the *scalar product* of a, $b \in \mathbb{R}^m$. Functions and relations on vectors are taken component-wise; for example, $\log q = (\log q_i)_{i\in\mathcal{X}}$. For $x \in \mathbb{R}^m$, $\sum x$ is a shorthand for $\sum_{i\in\mathcal{X}} x_i$.

Consider the problem $\mathcal{P}$ of minimization of ℓ, restricted to a convex closed set $C \subseteq \Delta_{\mathcal{X}}$:

$$\hat{\ell}_{\mathcal{P}} \triangleq \inf_{q\in C} \ell(q), \qquad S_{\mathcal{P}} \triangleq \{\hat{q} \in C :\ \ell(\hat{q}) = \hat{\ell}_{\mathcal{P}}\}. \tag{$\mathcal{P}$}$$

The goal is to find the *solution set* $S_{\mathcal{P}}$ as well as the infimum $\hat{\ell}_{\mathcal{P}}$ of the objective function ℓ over C. The problem $\mathcal{P}$ will be called the *multinomial likelihood primal problem*, or *primal*, for short.

Special attention is paid to the class of *polyhedral* feasible sets C

$$C = \{q \in \Delta_{\mathcal{X}} : \langle q, u_h\rangle \le 0 \text{ for } h = 1, 2, \dots, r\}, \tag{2.2}$$

or to its subclass of sets C given by (a finite number of) linear equality constraints

$$C = \{q \in \Delta_{\mathcal{X}} : \langle q, u_h\rangle = 0 \text{ for } h = 1, 2, \dots, r\}, \tag{2.3}$$

where u_h are vectors from $\mathbb{R}^m$. These feasible sets are particularly interesting from the applied point of view and permit to establish stronger results.

Without loss of generality it is assumed that $\mathcal{X}$ is the *support* $\operatorname{supp}(C)$ of C, that is, for every $i \in \mathcal{X}$ there is $q \in C$ with $q_i > 0$; in other words, the structural zeros (cf. Baker et al. [6, p. 34]) are excluded. Due to the convexity of C this is equivalent to the existence of $q \in C$ with $q > 0$. Under this assumption, Theorem 2 gives a basic characterization of the solution set of $\mathcal{P}$. Before stating it, some useful notions are introduced.

Definition 1. *For a type ν (or, more generally, for any $\nu \in \Delta_{\mathcal{X}}$), the* active *and* passive alphabets *are*

$$\mathcal{X}^a = \mathcal{X}^a_\nu \triangleq \{i \in \mathcal{X} : \nu_i > 0\} \qquad \textit{and} \qquad \mathcal{X}^p = \mathcal{X}^p_\nu \triangleq \{i \in \mathcal{X} : \nu_i = 0\}.$$

The elements of $\mathcal{X}^a, \mathcal{X}^p$ are called active, passive letters*, respectively.*

Put $m_a \triangleq \operatorname{card}\mathcal{X}^a > 0$, $m_p \triangleq \operatorname{card}\mathcal{X}^p \ge 0$. Let $\pi^a : \mathbb{R}^m \to \mathbb{R}^{m_a}$, $\pi^p : \mathbb{R}^m \to \mathbb{R}^{m_p}$ be the natural projections; identify $\mathbb{R}^{m_a}$ with $\mathbb{R}^{\mathcal{X}^a}$ and $\mathbb{R}^{m_p}$ with $\mathbb{R}^{\mathcal{X}^p}$. Note that if $\mathcal{X}^p = \emptyset$ then $m_p = 0$ and $\mathbb{R}^{m_p} = \{0\}$. For $x \in \mathbb{R}^m$, x^a and x^p are the shorthands for $\pi^a(x)$ and $\pi^p(x)$, respectively. (If no ambiguity can occur, the elements of $\mathbb{R}^{m_a}, \mathbb{R}^{m_p}$ will be denoted also by x^a, x^p.) Identify $\mathbb{R}^m$

with $\mathbb{R}^{m_a} \times \mathbb{R}^{m_p}$, so that it is possible to write $x = (x^a, x^p)$ for every $x \in \mathbb{R}^m$. Finally, for a subset M of $\mathbb{R}^m$ and $x \in M$ let

$$M^a \triangleq \pi^a(M) \qquad \text{and} \qquad M^a(x^p) \triangleq \{x^a \in \mathbb{R}^{m_a} : (x^a, x^p) \in M\}$$

be the *active projection* and the *x^p-slice* of M; analogously define M^p and $M^p(x^a)$.

Theorem 2 (Primal problem)**.** *Let $\nu \geq 0$ be from $\Delta_{\mathcal{X}}$. Let C be a convex closed subset of $\Delta_{\mathcal{X}}$ with support $\mathcal{X}$. Then $\hat{\ell}_{\mathcal{P}}$ is finite, $S_{\mathcal{P}}$ is compact, and there is $\hat{q}^a_{\mathcal{P}} \in C^a$, $\hat{q}^a_{\mathcal{P}} > 0$, such that*

$$S_{\mathcal{P}} = \{\hat{q}^a_{\mathcal{P}}\} \times C^p(\hat{q}^a_{\mathcal{P}}).$$

Moreover,

(a) If $\sum \hat{q}^a_{\mathcal{P}} = 1$ then $C^p(\hat{q}^a_{\mathcal{P}}) = \{0^p\}$ and $S_{\mathcal{P}} = \{(\hat{q}^a_{\mathcal{P}}, 0^p)\}$ is a singleton.
(b) If $\sum \hat{q}^a_{\mathcal{P}} < 1$ then $0^p \notin C^p(\hat{q}^a_{\mathcal{P}})$, and $S_{\mathcal{P}}$ is a singleton if and only if the $\hat{q}^a_{\mathcal{P}}$-slice $C^p(\hat{q}^a_{\mathcal{P}})$ of C is a singleton.

Thus the primal has always a solution $\hat{q}_{\mathcal{P}}$. Its active coordinates $\hat{q}^a_{\mathcal{P}}$ are unique and the passive coordinates $\hat{q}^p_{\mathcal{P}}$ are arbitrary such that $\hat{q}_{\mathcal{P}} \in C$. It is worth stressing that $C^p(\hat{q}^a_{\mathcal{P}})$ need not be equal to $\{0^p\}$, that is, a solution of $\mathcal{P}$ may assign positive mass to passive letter(s). The following couple of simple examples illustrates the points; see also the examples in Section 4.2. In the first example $S_{\mathcal{P}}$ is a segment, in the second one $S_{\mathcal{P}}$ is a singleton. Hereafter X denotes a random variable supported on $\mathcal{X}$.

Example 3. Take $\mathcal{X} = \{-1, 0, 1\}$ and $C = \{q \in \Delta_{\mathcal{X}} : \ \mathrm{E}_q(X^2) = \sum_{i \in \mathcal{X}} i^2 q_i = 1/2\}$. Let $\nu = (0, 1, 0)$, so that $\mathcal{X}^a = \{0\}$ and $\mathcal{X}^p = \{-1, 1\}$. Then $C = \{q \in \Delta_{\mathcal{X}} : \ q_0 = 1/2\}$, the minimum of ℓ over C is $\hat{\ell}_{\mathcal{P}} = \log 2$, and $S_{\mathcal{P}} = C$. Here, $\hat{q}^a_{\mathcal{P}} = 1/2$ is (trivially) unique and $C^p(\hat{q}^a_{\mathcal{P}}) = \{(a, 1/2 - a) : a \in [0, 1/2]\}$.

Example 4 (E4). Motivated by Wets [47, p. 88], let $\mathcal{X} = \{1, 2, 3\}$, $C = \{q \in \Delta_{\mathcal{X}} : q_1 \leq q_2 \leq q_3\}$ and $\nu = (0, 1, 0)$. Then $\mathcal{X}^a = \{2\}$, $\mathcal{X}^p = \{1, 3\}$. Since $S_{\mathcal{P}} = \{(0, 1, 1)/2\}$, the positive weight $1/2$ is assigned to the passive, unobserved letter 3.

The Fisher scoring algorithm which is commonly used to solve $\mathcal{P}$ when C is a linear set may fail to converge when the zero counts are present; cf. Stirling [46]. Other numerical methods, such as the augmented Lagrange multiplier methods, which are used to solve the convex optimization problem under polyhedral and/or linear C may have difficulties to cope with large m. Moreover, the primal is not amenable for asymptotic analysis. Thus, it is desirable to approach $\mathcal{P}$ from the side of the convex duality.

3. Fenchel dual problem $\mathcal{F}$ to $\mathcal{P}$

Annotation. The Fenchel dual problem $\mathcal{F}$ to $\mathcal{P}$ is introduced. A Lagrange dual formulation of the convex conjugate (cf. Theorem 5) serves as a ground for

Theorem 6 which provides a relation between the solutions of $\mathcal{P}$ and $\mathcal{F}$. If the feasible set C is polyhedral, a solution of $\mathcal{F}$ can be obtained also from a different Lagrange dual to $\mathcal{P}$; cf. Section 3.1. A special case of the single inequality constraint is discussed in detail in Section 3.2 where also the concept of the base solution is introduced. As a minor point, a flaw in Klotz's [28] Theorem 1 is also noted there.

Consider the *Fenchel dual problem* $\mathcal{F}$ to the primal $\mathcal{P}$:

$$\hat{\ell}_{\mathcal{F}} \triangleq \inf_{y\in C^*} \ell^*(-y), \qquad S_{\mathcal{F}} \triangleq \{\hat{y} \in C^* :\ \ell^*(-\hat{y}) = \hat{\ell}_{\mathcal{F}}\}, \tag{$\mathcal{F}$}$$

where

$$C^* \triangleq \{y \in \mathbb{R}^m : \langle y, q\rangle \le 0 \text{ for every } q \in C\}$$

is the *polar cone* of C and

$$\ell^* : \mathbb{R}^m \to \bar{\mathbb{R}}, \qquad \ell^*(z) \triangleq \sup_{q\in\Delta_{\mathcal{X}}} (\langle q, z\rangle - \ell(q))$$

is the *convex conjugate* of ℓ (in fact, the convex conjugate of $\tilde{\ell} : \mathbb{R}^m \to \bar{\mathbb{R}}$ given by $\tilde{\ell}(x) = \ell(x)$ for $x \in \Delta_{\mathcal{X}}$ and $\tilde{\ell}(x) = \infty$ otherwise).

The Fenchel dual $\mathcal{F}$ is often more tractable than the primal $\mathcal{P}$, in particular when C is given by linear equality and/or inequality constraints. This is the case of the models for contingency tables, mentioned in Introduction. Also an *estimating equations* model C_Θ leads to a linear feasible set C_θ, when $\theta \in \Theta$ is fixed. The model is $C_\Theta \triangleq \bigcup_{\theta\in\Theta} C_\theta$, where

$$C_\theta \triangleq \bigcap_{h=1}^{r} \{q \in \Delta_{\mathcal{X}} : \langle q, u_h(\theta)\rangle = 0\}, \tag{3.1}$$

and $u_h : \Theta \to \mathbb{R}^m$, $h = 1, 2, \ldots, r$, are the *estimating functions*. There $\theta \in \Theta \subseteq \mathbb{R}^d$ and d need not be equal to r. Since r is usually much smaller than m, $\mathcal{F}$ may be easier to solve numerically than $\mathcal{P}$.

Observe that the convex conjugate itself is defined through an optimization problem, the *convex conjugate primal problem* (*cc-primal*, for short), whose solution set is

$$S_{cc}(z) \triangleq \{q \in \Delta_{\mathcal{X}} : \langle q, z\rangle - \ell(q) = \ell^*(z)\}. \tag{3.2}$$

The structure of $S_{cc}(z)$ is described by Proposition 36. The conjugate ℓ^* can be evaluated by means of the Lagrange duality, where the Lagrange function is

$$K_z(x, \mu) = \langle x, z\rangle - \ell(x) - \mu \left(\sum x - 1\right).$$

It holds that (cf. Lemma 34)

$$\ell^*(z) = \inf_{\mu\in\mathbb{R}} k_z(\mu), \qquad \text{where} \quad k_z(\mu) = \sup_{x\ge 0} K_z(x, \mu).$$

For every $\mu \in \mathbb{R}$ and $a, b \in \mathbb{R}^{m_a}$, $a > 0$, $b > -\mu$, define

$$I_\mu(a \,\|\, b) \triangleq \left\langle a, \log \frac{a}{\mu + b} \right\rangle.$$

Theorem 5 (Convex conjugate by Lagrange duality)**.** *Let $\nu \in \Delta_{\mathcal{X}}$ and $z \in \mathbb{R}^m$. Then*

$$\ell^*(z) = -1 + \hat{\mu}(z) + I_{\hat{\mu}(z)}(\nu^a \,\|\, -z^a),$$

where

$$\hat{\mu}(z) \triangleq \max\{\bar{\mu}(z^a), \max(z^p)\}, \tag{3.3}$$

and $\bar{\mu}(z^a)$ is the unique solution of

$$\sum \frac{\nu^a}{\mu - z^a} = 1, \qquad \mu \in (\max(z^a), \infty). \tag{3.4}$$

The key point is that the $\bar{\mu}(z^a)$, which solves (3.4), is not always the $\hat{\mu}(z)$ which minimizes $k_z(\mu)$. They are the same if and only if $\bar{\mu}(z^a) \geq \max(z^p)$. As it will be seen in Theorem 6, if $z \in S_{\mathcal{F}}$ then this inequality decides whether the solution of primal $\mathcal{P}$ is supported only on the active letters, or some probability mass is placed also to the passive letter(s).

Theorem 5 serves as a foundation for Theorem 6, which states the relation between the Fenchel dual and the primal.

Theorem 6 (Relation between $\mathcal{F}$ and $\mathcal{P}$)**.** *Let $\nu, C, \hat{q}^a_{\mathcal{P}}$ be as in Theorem 2. Then*

$$\hat{\ell}_{\mathcal{F}} = -\hat{\ell}_{\mathcal{P}},$$

$S_{\mathcal{F}}$ is a nonempty convex compact set, $S_{\mathcal{F}} \perp S_{\mathcal{P}}$, and $\hat{\mu}(-\hat{y}_{\mathcal{F}}) = 1$ for every $\hat{y}_{\mathcal{F}} \in S_{\mathcal{F}}$. Moreover, if we put

$$\hat{y}^a_{\mathcal{F}} \triangleq \frac{\nu^a}{\hat{q}^a_{\mathcal{P}}} - 1^a, \tag{3.5}$$

then $(\hat{y}^a_{\mathcal{F}}, -1^p) \in S_{\mathcal{F}}$ and the following hold:

(a) If $\sum \hat{q}^a_{\mathcal{P}} = 1$ then $\bar{\mu}(-\hat{y}^a_{\mathcal{F}}) = 1$ and

$$S_{\mathcal{F}} = \{\hat{y}^a_{\mathcal{F}}\} \times \{y^p \in C^{*p}(\hat{y}^a_{\mathcal{F}}) : \min(y^p) \geq -1\}.$$

(b) If $\sum \hat{q}^a_{\mathcal{P}} < 1$ then $\bar{\mu}(-\hat{y}^a_{\mathcal{F}}) < 1$ and

$$S_{\mathcal{F}} = \{\hat{y}^a_{\mathcal{F}}\} \times \{y^p \in C^{*p}(\hat{y}^a_{\mathcal{F}}) : \min(y^p) = -1\}.$$

Thus, in the active coordinates, the solution of $\mathcal{F}$ is unique and is related to $\hat{q}^a_{\mathcal{P}}$ by (3.5). Together with Theorem 2 this yields

$$S_{\mathcal{P}} = \left\{\frac{\nu^a}{1 + \hat{y}^a_{\mathcal{F}}}\right\} \times C^p\left(\frac{\nu^a}{1 + \hat{y}^a_{\mathcal{F}}}\right).$$

The structure of the solution set of $\mathcal{F}$ in the passive letters is determined by the relation of $\bar{\mu}(-\hat{y}^a_{\mathcal{F}})$ to $1 = \hat{\mu}(-\hat{y}_{\mathcal{F}})$.

In the case (a), $\bar{\mu}(-\hat{y}^a_{\mathcal{F}}) = \hat{\mu}(-\hat{y}_{\mathcal{F}})$, and the passive projections $\hat{y}^p_{\mathcal{F}}$ of the solutions satisfy $\hat{y}^p_{\mathcal{F}} \geq -1$. Then it suffices to solve the primal solely in the active letters and $S_{\mathcal{P}}$ is a singleton. This happens for instance if $\nu > 0$.

In the case (b), $\bar{\mu}(-\hat{y}^a_{\mathcal{F}}) < \hat{\mu}(-\hat{y}_{\mathcal{F}})$, and $\hat{y}^p_{\mathcal{F}}$ satisfy $\min(\hat{y}^p_{\mathcal{F}}) = -1$. Then every solution of the primal assigns the positive probability to at least one passive letter.

To sum up, the Fenchel dual problem, once solved, permits to find $\hat{q}^a_{\mathcal{P}}$ through (3.5) and this way it incorporates C^* into $\hat{q}^a_{\mathcal{P}}$. In the case of linear or polyhedral C the dual may reduce dimensionality of the optimization problem, yet the numerical solution of $\mathcal{F}$ is somewhat hampered by the need to obey (3.3). Moreover, $\hat{q}^p_{\mathcal{P}}$ remains to be found; in this respect see Proposition 47.

Example 4 is used to illustrate the results.

Example 4 (cont'd). Since the set C is polyhedral, a solution of $\mathcal{F}$ has the form $\hat{y}_{\mathcal{F}} = \sum_h \hat{\alpha}_h u_h$, $h = 1, 2$, where $u_1 = (1, -1, 0)$, $u_2 = (0, 1, -1)$, and $\hat{\alpha} \in \mathbb{R}^2_+$ minimizes $-1 + \hat{\mu}(-\sum_h \alpha_h u_h) + I_{\hat{\mu}(-\sum_h \alpha_h u_h)}(\nu^a \parallel \sum_h \alpha_h u^a_h)$; cf. Thm. 5. Recall that $\hat{\mu}(\cdot)$ is defined by (3.3) and note that the equation (3.4) involves a one-dimensional zero finding.

Taken together, finding a solution of $\mathcal{F}$ for a polyhedral set is not numerically demanding. It is even simpler for a linear C, as it involves unconstrained minimization over $\alpha \in \mathbb{R}^r$.

Here, $\alpha = (0, 1)$, and $\hat{\ell}_{\mathcal{F}} = -\log(2)$. Thus, $\hat{\ell}_{\mathcal{F}} = -\hat{\ell}_{\mathcal{P}}$. Using (3.5), $\hat{q}^a = 1/2$. Consequently, a positive weight must be assigned to at least one of the passive letters. By Propositions 47 and 36, if $\hat{y}^p_i \neq 1$, then the passive letter is assigned the zero weight. Here, $\hat{y}^p = (0, 1)$, so $\hat{q}_1 = 0$. Thus, $\hat{q}_3 = 1/2$.

3.1. Polyhedral and linear C

In the polyhedral case (2.2)

$$C = \{q \in \Delta_{\mathcal{X}} : \langle q, u_h \rangle \leq 0 \text{ for } h = 1, 2, \ldots, r\},$$

a solution of the Fenchel dual $\mathcal{F}$ can also be obtained from the saddle points of the following Lagrange function

$$L(q, \alpha) \triangleq -\sum_h \alpha_h \langle q, u_h \rangle - \ell(q) \qquad (q \in \Delta_{\mathcal{X}}, \alpha \in \mathbb{R}^r_+).$$

Indeed, it holds that

$$\hat{\ell}_{\mathcal{F}} = \inf_{\alpha \geq 0} \sup_{q \in \Delta_{\mathcal{X}}} L(q, \alpha) = \sup_{q \in \Delta_{\mathcal{X}}} \inf_{\alpha \geq 0} L(q, \alpha) = -\hat{\ell}_{\mathcal{P}}.$$

Hence $(\hat{q}, \hat{\alpha})$ is a saddle point of $L(q, \alpha)$ if and only if $\hat{q} \in S_{\mathcal{P}}$ and $\sum_h \hat{\alpha}_h u_h \in S_{\mathcal{F}}$; cf. Bertsekas [10, Prop. 2.6.1].

The vectors from $S_{\mathcal{F}}$ of the form $\sum_h \hat{\alpha}_h u_h$ will be called the *base solutions* of $\mathcal{F}$. From the Farkas lemma (cf. Bertsekas [10, Prop. 3.2.1]) and the monotonicity of ℓ^* (Lemma 37) it follows that, for a polyhedral C, there always exists a base solution, and every solution of $\mathcal{F}$ is a sum of a base solution and a vector from $\mathbb{R}^m_- \triangleq \{z \in \mathbb{R}^m : z \leq 0\}$; that is,

$$S_{\mathcal{F}} = \{\hat{y}_{\mathcal{F}} + (0^a, z^p) : \hat{y}_{\mathcal{F}} \text{ is a base solution}, \ z^p \leq 0, \ \hat{y}^p_{\mathcal{F}} + z^p \geq -1\}.$$

There may exist many base solutions. However, if $u_1^a, \dots, u_r^a$ are linearly independent then the base solution is unique, since then the system of equations $\sum_h \hat{\alpha}_h u_h^a = \hat{y}_{\mathcal{F}}^a$ has a unique solution $\hat{\alpha}$.

Analogous claim holds for a linear C (cf. (2.3)), but in this case $\alpha \in \mathbb{R}^r$, by the Farkas lemma.

3.2. Single inequality constraint and Klotz's Theorem 1

To illustrate the base solution in connection with Theorem 6, consider

$$C = \{q \in \Delta_{\mathcal{X}} : \langle q, u \rangle \leq 0\},$$

given by a single inequality constraint. By the Farkas lemma, $C^* = \{\alpha u : \alpha \geq 0\} + \mathbb{R}_-^m$. If $u^p \geq 0$ the case (a) of Theorem 6 applies (for if not, there is a base solution $\hat{\alpha}_{\mathcal{F}} u$ and, by Theorem 6(b), $\min(\hat{\alpha}_{\mathcal{F}} u^p)$ should be -1; this is not possible since $\min(\hat{\alpha}_{\mathcal{F}} u^p) \geq 0$).

Assume that $\min(u^p) < 0$ and the case (b) of Theorem 6 applies. Take any base solution $\hat{\alpha}_{\mathcal{F}} u$. Then, by Theorem 6(b), $\min(\hat{\alpha}_{\mathcal{F}} u^p) = -1$; so

$$\hat{\alpha}_{\mathcal{F}} = -\frac{1}{\min(u^p)} \qquad \text{and} \qquad \hat{q}_{\mathcal{P}}^a = \frac{\nu^a}{1 - \frac{u^a}{\min(u^p)}}\,. \tag{3.6}$$

Further, by Theorem 2, $\hat{q}_{\mathcal{P}}^p$ is arbitrary from $C^p(\hat{q}_{\mathcal{P}}^a)$. If u^p attains the minimum at a single letter, then the solution of the primal $\mathcal{P}$ is always unique.

To sum up, the case (b) of Theorem 6 happens if and only if

$$\min(u^p) < 0 \qquad \text{and} \qquad \sum \frac{\nu^a}{1 - \frac{1}{\min(u^p)} u^a} < 1. \tag{3.7}$$

The primal problem $\mathcal{P}$ with this C was considered by Klotz [28]. Theorem 1 of [28] asserts that the solution of $\mathcal{P}$ takes the form

$$\hat{q}_K \triangleq \frac{\nu}{1 + \hat{\alpha}_K u}$$

if $\langle \nu, u \rangle > 0$ and $\min(u^a) < 0$; see Klotz's condition (3.1b). There, $\hat{\alpha}_K$ is the unique root of

$$\sum \frac{\nu}{1 + \alpha u} = 1 \qquad \text{such that} \quad \alpha \in (0, -1/\min(u^a)).$$

Thus, under Klotz's condition (3.1b), the solution of $\mathcal{P}$ should assign zero weight to any passive letter. This is not the case, as the following example demonstrates.

Example 7 (Base solution of $\mathcal{F}$, one inequality constraint). Take $\mathcal{X} = \{-2, -1, 0, 1, 2\}$, $u = (-2, -1, 0, 1, 2)$ and $\nu = (0, 3, 0, 0, 7)/10$. Here Klotz's condition (3.1b) is satisfied and an easy computation yields $\hat{\alpha}_K = 11/20$; thus $\hat{q}_K = (0, 2, 0, 0, 1)/3$ and $\ell(\hat{q}_K) = 0.890668$. However, the primal is solved by $\hat{q}_{\mathcal{P}} = (1, 12, 0, 0, 7)/20$, so a positive mass is placed also to the passive letter -2; $\ell(\hat{q}_{\mathcal{P}}) = 0.888123$. Since $\min(u^p) = -2$ and $\sum \hat{q}_{\mathcal{P}}^a < 1$, the condition (3.7) is satisfied and $\hat{q}_{\mathcal{P}}^a$ is just that given by (3.6).

In a similar way it is possible to analyze a single equality constraint. For more than one equality/inequality constraint the condition guaranteeing that $\sum \hat{q}^a_{\mathcal{P}} < 1$ cannot be given in such a simple form as (3.7).

4. El Barmi-Dykstra dual problem $\mathcal{B}$ to $\mathcal{P}$

Annotation. The simplified Fenchel dual developed by El Barmi and Dykstra is recalled in Section 4.1. Its shortcomings are illustrated by simple examples in Section 4.2. Finally, the scope of validity of El Barmi and Dykstra's dual is studied in Section 4.3.

4.1. BD-dual $\mathcal{B}$ to $\mathcal{P}$

El Barmi and Dykstra [12] consider a *simplified Fenchel dual problem* $\mathcal{B}$ (the *BD-dual*, for short)

$$\hat{\ell}_{\mathcal{B}} \triangleq \inf_{y \in C^*} \ell^*_{\mathcal{B}}(y), \qquad S_{\mathcal{B}} \triangleq \{\hat{y} \in C^* : \ell^*_{\mathcal{B}}(\hat{y}) = \hat{\ell}_{\mathcal{B}}\}, \tag{$\mathcal{B}$}$$

where

$$\ell^*_{\mathcal{B}} : \mathbb{R}^m \to \bar{\mathbb{R}}, \qquad \ell^*_{\mathcal{B}}(y) \triangleq \begin{cases} I_1(\nu^a \,\|\, y^a) = \left\langle \nu^a, \log \dfrac{\nu^a}{1+y^a} \right\rangle & \text{if } y^a > -1, \\ \infty & \text{otherwise,} \end{cases}$$

will be called the *BD conjugate* of ℓ. Note that the BD-dual $\mathcal{B}$ is easier to solve than $\mathcal{F}$, as in the former $\hat{\mu}$ is fixed to 1.

If C is polyhedral then, in analogy with the concept of the base solution of $\mathcal{F}$, the vectors from $S_{\mathcal{B}}$ of the form $\sum \hat{\alpha}_{\mathcal{B},h} u_h$ are called the *base solutions* of $\mathcal{B}$. As above, every solution of $\mathcal{B}$ can be written as a sum of a base solution and a vector from $\mathbb{R}^m_-$.

4.1.1. Smith dual problem

By the Farkas lemma, for the feasible set $C = \{q \in \Delta_{\mathcal{X}} : \langle q, u_h \rangle = 0, h = 1, 2, \dots, r\}$ given by r linear equality constraints, the polar cone is $C^* = \{y = \sum_h \alpha_h u_h : \alpha \in \mathbb{R}^r\} + \mathbb{R}^m_-$. Then the BD-dual becomes

$$\inf_{\alpha \in \mathbb{R}^r} I_1\left(\nu^a \,\middle\|\, \sum_h \alpha_h u^a_h\right), \tag{4.1}$$

which is equivalent to the *simplified Lagrange dual problem* (the *Smith dual*, for short) considered by Smith [45, Sect. 6] and many other authors; see, in particular, Haber [21, p. 3], Little and Wu [32, p. 88], Lang [29, Sect. 7.1], and Bergsma et al. [9, p. 65]. It is worth noting that (4.1) is an unconstrained optimization problem.

4.2. Flaws of BD-dual $\mathcal{B}$

In [12, Thm. 2.1] El Barmi and Dykstra state the following relationship between $\mathcal{P}$ and $\mathcal{B}$.

Theorem 8 (Theorem 2.1 of [12]). *Let C be a convex closed subset of $\Delta_{\mathcal{X}}$.*

- *($\mathcal{B} \to \mathcal{P}$) If $\hat{\ell}_{\mathcal{B}}$ is finite, then $S_{\mathcal{B}}$ is nonempty and $\hat{\ell}_{\mathcal{B}} = -\hat{\ell}_{\mathcal{P}}$. Moreover, for every $\hat{y}_{\mathcal{B}} \in S_{\mathcal{B}}$, $\hat{q}_{\mathcal{B}} \triangleq \frac{\nu}{1+\hat{y}_{\mathcal{B}}}$ belongs to $S_{\mathcal{P}}$.*
- *($\mathcal{P} \to \mathcal{B}$) If $\hat{\ell}_{\mathcal{P}}$ is finite, then $\hat{\ell}_{\mathcal{B}} = -\hat{\ell}_{\mathcal{P}}$. Moreover, if $\hat{q}_{\mathcal{P}} \in S_{\mathcal{P}}$, then $\hat{y}_{\mathcal{B}} \triangleq \frac{\nu}{\hat{q}_{\mathcal{P}}} - 1$ belongs to $S_{\mathcal{B}}$. (There, $0/0 = 0$ convention applies.)*

Though $\hat{\ell}_{\mathcal{P}}$ is always finite, $\hat{\ell}_{\mathcal{B}}$ may be infinite, leaving the solution set $S_{\mathcal{P}}$ inaccessible through ($\mathcal{B} \to \mathcal{P}$) of [12, Thm. 2.1]. Moreover, the claims ($\mathcal{B} \to \mathcal{P}$) and ($\mathcal{P} \to \mathcal{B}$) of the theorem are not always true. In fact, there are three possibilities (cf. Theorem 16):

1. $\hat{\ell}_{\mathcal{B}} = -\infty$;
2. $\hat{\ell}_{\mathcal{B}}$ is finite, but there is a *BD-duality gap*, that is, $\hat{\ell}_{\mathcal{B}} < -\hat{\ell}_{\mathcal{P}}$;
3. $\hat{\ell}_{\mathcal{B}} = -\hat{\ell}_{\mathcal{P}}$ (*no BD-duality gap*).

To illustrate them, we give below six simple examples: one where the BD-duality works and the other five where it fails. In Examples 9, 10, 13, 14 the set C is linear, whereas Example 11 presents a nonlinear C, and in Example 12 the set C is defined by linear inequalities.

Example 9 (No BD-duality gap). Let $\mathcal{X} = \{-1, 0, 1\}$ and $C = \{q \in \Delta_{\mathcal{X}} : \mathrm{E}_q X = 0\}$; that is, C is given by (2.3) with $u = (-1, 0, 1)$. Let $\nu = (1, 0, 1)/2$. Thus $\mathcal{X}^a = \{-1, 1\}$, $\mathcal{X}^p = \{0\}$, and $C^a(0^p) = \{(1, 1)/2\}$. By the Farkas lemma, $C^* = \{y = \alpha u : \alpha \in \mathbb{R}\} + \mathbb{R}^3_-$. For the considered optimization problems, the following hold:

- $\mathcal{P}$: $S_{\mathcal{P}} = \{(1, 0, 1)/2\}$ and $\hat{\ell}_{\mathcal{P}} = \log 2$.
- $\mathcal{F}$: Since $\hat{\alpha}_{\mathcal{F}} = 0$, the base solution of $\mathcal{F}$ is $(0, 0, 0)$ and $\hat{\ell}_{\mathcal{F}} = -\log 2$. Further, (3.5) implies that $\hat{q}^a_{\mathcal{P}} = \nu^a/(1 + \hat{y}^a_{\mathcal{F}})$; thus, $\hat{q}^a_{\mathcal{P}} = \nu^a$.

In this setting, no BD-duality gap occurs:

- $\mathcal{B} \to \mathcal{P}$: $\ell^*_{\mathcal{B}}(\alpha u) \propto -1/2[\log(1-\alpha) + \log(1+\alpha)]$. Thus $\hat{\alpha}_{\mathcal{B}} = 0$, $\hat{\ell}_{\mathcal{B}} = -\log 2 = -\hat{\ell}_{\mathcal{P}}$. Since $\hat{q}_{\mathcal{B}} = (1, 0, 1)/2$, it indeed solves $\mathcal{P}$.
- $\mathcal{P} \to \mathcal{B}$: $\hat{\ell}_{\mathcal{B}} = -\hat{\ell}_{\mathcal{P}}$ by the previous case. Since the base solution of $\mathcal{B}$ is $(0, 0, 0)$ and $\hat{y}_{\mathcal{B}} = (0, 0, 0)$, $\hat{y}_{\mathcal{B}} \in S_{\mathcal{B}}$.

Example 10 (H-set). Let $\mathcal{X}$, C be as in Example 9 and $\nu = (1, 0, 0)$. Thus $\mathcal{X}^a = \{-1\}$, $\mathcal{X}^p = \{0, 1\}$, and $C^a(0^p) = \emptyset$.

- $\mathcal{P}$: $S_{\mathcal{P}} = \{(1, 0, 1)/2\}$ and $\hat{\ell}_{\mathcal{P}} = \log 2$.
- $\mathcal{F}$: Since $\hat{\alpha}_{\mathcal{F}} = -1$, the base solution of $\mathcal{F}$ is $(1, 0, -1)$ and $\hat{\ell}_{\mathcal{F}} = -\log 2$. Thus, from (3.5) it follows that $\hat{q}^a_{\mathcal{P}} = (1/2)$.

- $\mathcal{B}\to\mathcal{P}$: $\ell^*_{\mathcal{B}}(\alpha u) \propto -\log(1-\alpha)$, which does not have finite infimum.
- $\mathcal{P}\to\mathcal{B}$: $\hat{\ell}_{\mathcal{B}}$ is not finite and $S_{\mathcal{B}} = \emptyset$. Thus $\hat{y}_{\mathcal{B}} = (1,-1,-1) \notin S_{\mathcal{B}}$.

Example 11 (nonlinear C, H-set). Let $\mathcal{X} = \{2,3,4\}$, $C = \{q \in \Delta_{\mathcal{X}} : \sum_{i\in\mathcal{X}} i\, q_i^2 \le 1\}$, and $\nu = (1,0,0)$. Thus $\mathcal{X}^a = \{2\}$, $\mathcal{X}^p = \{3,4\}$, and $C^a(0^p) = \emptyset$.

- $\mathcal{P}$: Write $C = \bigcup_{k\in[0,1]} C_k$, where $C_k \triangleq \{q \in \Delta_{\mathcal{X}} : \sum_{i\in\mathcal{X}} i\, q_i^2 = k\}$. On C_k, $7q_3 = 4 - 4q_2 \pm \sqrt{-12 + 24q_2 - 26q_2^2 + 7k}$ and the nonnegativity of the term under the square-root implies that q_2 should belong to the interval $[6/13 - 1/26\sqrt{-168+182k}, 6/13 + 1/26\sqrt{-168+182k})]$. This in turn implies that $\hat{q}_2(k) = 6/13 + 1/26\sqrt{-168+182k}$. The function $\hat{q}_2(k)$ attains its maximum at $k = 1$, for which $\hat{q}_2 = 0.6054$. The other two elements of $\hat{q}_{\mathcal{P}}$ are determined uniquely. Thus, $\hat{q}_{\mathcal{P}} = (0.6054, 0.2255, 0.1691)$.
- $\mathcal{B}$: As shown above, $q_2 \le (12+\sqrt{24})/26 \triangleq c < 1$ for every $q \in C$. Put $d = c/(1-c)$. Then, for every $\alpha \ge 0$, $y \triangleq \alpha(1,-d,-d) \in C^*$ and $\ell^*_{\mathcal{B}}(y) = -\log(1+\alpha)$. Hence the BD-dual problem has $\hat{\ell}_{\mathcal{B}} = -\infty$.

Example 12 (monotonicity, H-set). Let $\mathcal{X} = \{1,2,3\}$, $C = \{q \in \Delta_{\mathcal{X}} : q_1 \le q_2 \le q_3\}$, and $\nu = (0,1,0)$, as in Example 4. Thus $\mathcal{X}^a = \{2\}$, $\mathcal{X}^p = \{1,3\}$, and $C^a(0^p) = \emptyset$. The polyhedral set C is given by (2.2) with $u_1 = (1,-1,0)$, $u_2 = (0,1,-1)$.

- $\mathcal{P}$: $S_{\mathcal{P}} = \{(0,1,1)/2\}$ and $\hat{\ell}_{\mathcal{P}} = \log 2$.
- $\mathcal{F}$: Since $\hat{\alpha}_{\mathcal{F}} = (0,1)$, the base solution is $\hat{y}_{\mathcal{F}} = (0,1,-1)$. Thus $\hat{q}^a_{\mathcal{P}} = 1/2$ by (3.5).
- $\mathcal{B}\to\mathcal{P}$: For every $\alpha \ge 0$, $\alpha u_2 \in C^*$ and $\ell^*_{\mathcal{B}}(\alpha u_2) = -\log(1+\alpha)$. Thus $\hat{\ell}_{\mathcal{B}} = -\infty$.
- $\mathcal{P}\to\mathcal{B}$: Since $S_{\mathcal{B}} = \emptyset$, $\hat{y}_{\mathcal{B}} = (-1,1,-1) \notin S_{\mathcal{B}}$.

Example 13 (Z-set). Motivated by Example 4 of Bergsma et al. [8], let $\mathcal{X}$, C be as in Example 9 and let $\nu = (1,1,0)/2$. Thus $\mathcal{X}^a = \{-1,0\}$, $\mathcal{X}^p = \{1\}$, and $C^a(0^p) = \{(0,1)\}$.

- $\mathcal{P}$: $S_{\mathcal{P}} = \{(1,2,1)/4\}$ and $\ell_{\mathcal{P}} = (1/2)\log 8$.
- $\mathcal{F}$: Since $\hat{\alpha}_{\mathcal{F}} = -1$, the base solution of $\mathcal{F}$ is $(1,0,-1)$ and $\hat{\ell}_{\mathcal{F}} = -(1/2)\log 8$. Thus, by (3.5), $\hat{q}^a_{\mathcal{P}} = (1,2)/4$.
- $\mathcal{B}\to\mathcal{P}$: $\ell^*_{\mathcal{B}}(\alpha u) \propto -(1/2)\log(1-\alpha)$, hence $\hat{\ell}_{\mathcal{B}} = -\infty$.
- $\mathcal{P}\to\mathcal{B}$: Since $S_{\mathcal{B}} = \emptyset$, $\hat{y}_{\mathcal{B}} = (1,0,-1) \notin S_{\mathcal{B}}$.

Observe that Theorem 2.1 of [12] implies that $\hat{q}^p_{\mathcal{B}} = 0^p$, provided that $\hat{\ell}_{\mathcal{B}}$ is finite. However, $C^p(\hat{q}^a)$ may be different than $\{0^p\}$; in such a case $\hat{q}^p_{\mathcal{P}}$ has a strictly positive coordinate and the BD-duality gap occurs.

Example 14 (BD-duality gap). Let $\mathcal{X} = \{-1,1,10\}$, $C = \{q \in \Delta_{\mathcal{X}} : \mathrm{E}_q X = 0\}$, so that $u = (-1,1,10)$, and $\nu = (3,2,0)/5$. Thus $\mathcal{X}^a = \{-1,1\}$, $\mathcal{X}^p = \{10\}$, and $C^a(0^p) = \{(1,1)/2\}$.

- $\mathcal{P}$: $S_{\mathcal{P}} = \{(54,44,1)/99\}$ and $\hat{\ell}_{\mathcal{P}} = 0.6881$.
- $\mathcal{F}$: Since $\hat{\alpha}_{\mathcal{F}} = -1/10$, the base solution of $\mathcal{F}$ is $(1,-1,-10)/10$ and $\hat{\ell}_{\mathcal{F}} = -0.6881$. From (3.5), it follows that $\hat{q}^a_{\mathcal{P}} = (54,44)/99$.

- $\mathcal{B}\to\mathcal{P}$: $\ell^*_{\mathcal{B}}(\alpha u) \propto -[\nu_{-1}\log(1-\alpha)+\nu_1\log(1+\alpha)]$. Since $\hat{\alpha}_{\mathcal{B}} = -1/5$ and $\hat{y}_{\mathcal{B}} = \hat{\alpha}_{\mathcal{B}} u$, thus $\hat{q}_{\mathcal{B}} = \nu/(1+\hat{y}_{\mathcal{B}}) = (1,1,0)/2$, $\hat{\ell}_{\mathcal{B}} = -0.6931$, and $\hat{\ell}_{\mathcal{B}} < -\hat{\ell}_{\mathcal{P}}$.
- $\mathcal{P}\to\mathcal{B}$: Since $\hat{\ell}_{\mathcal{P}}$ is finite, it should hold that $\hat{\ell}_{\mathcal{P}} = -\hat{\ell}_{\mathcal{B}}$, but it does not. Moreover, $\hat{y}_{\mathcal{B}} = (1,-1,-10)/10 = -(1/10)u \notin S_{\mathcal{B}}$.

4.3. Scope of validity of BD-dual $\mathcal{B}$

The correct relation of the BD-dual $\mathcal{B}$ to the primal $\mathcal{P}$ is stated in Theorem 16. In order to formulate the conditions under which $\hat{\ell}_{\mathcal{B}}$ is infinite (cf. Theorem 16(a)) the notions of H-set and Z-set are introduced. They are implied by the recession cone considerations of $\mathcal{B}$.

Definition 15. *If a nonempty convex closed set $C \subseteq \Delta_{\mathcal{X}}$ and a type ν are such that $C^a(0^p) = \emptyset$, then we say that C is an* H-set with respect to ν. *The set C is called a* Z-set with respect to ν *if $C^a(0^p)$ is nonempty but its support is strictly smaller than $\mathcal{X}^a$.*

Note that $C^a(0^p)$ comprises (active projections of) those $q \in C$ which are supported on the active letters. Thus, C is neither an H-set nor a Z-set if and only if there is $q \in C$ with $q^a > 0$, $q^p = 0$.

Clearly, in Example 10, C is an H-set with respect to the ν. The same set C becomes a Z-set with respect to the ν considered in Example 13. And it is neither an H-set nor a Z-set with respect to the ν studied in Example 9. Further, the feasible set C considered in Example 14 is neither an H-set nor a Z-set with respect to the particular ν. In Examples 11 and 12, C is an H-set with respect to the ν.

Theorem 16 (Relation between $\mathcal{B}$ and $\mathcal{P}$)**.** *Let $\nu, C, \hat{q}^a_{\mathcal{P}}, \hat{y}^a_{\mathcal{F}}$ be as in Theorems 2 and 6.*

(a) If C is either an H-set or a Z-set with respect to ν then

$$\hat{\ell}_{\mathcal{B}} = -\infty \qquad \text{and} \qquad S_{\mathcal{B}} = \emptyset.$$

(b) If C is neither an H-set nor a Z-set then $\hat{\ell}_{\mathcal{B}}$ is finite, $\hat{\ell}_{\mathcal{B}} \le \hat{\ell}_{\mathcal{F}}$, and there is $\hat{y}^a_{\mathcal{B}} \in C^{*a}$ such that $\bar{\mu}(-\hat{y}^a_{\mathcal{B}}) = 1$ and

$$S_{\mathcal{B}} = \{\hat{y}^a_{\mathcal{B}}\} \times C^{*p}(\hat{y}^a_{\mathcal{B}}).$$

Moreover, there is no BD-duality gap, that is,

$$\hat{\ell}_{\mathcal{B}} = -\hat{\ell}_{\mathcal{P}},$$

if and only if any of the following (equivalent) conditions hold:

(i) $\sum \hat{q}^a_{\mathcal{P}} = 1$ (that is, $S_{\mathcal{P}} = \{(\hat{q}^a_{\mathcal{P}}, 0^p)\}$);

(ii) $\hat{y}^a_{\mathcal{B}} = \hat{y}^a_{\mathcal{F}}$ (that is, $\nu^a/(1+\hat{y}^a_{\mathcal{B}}) = \hat{q}^a_{\mathcal{P}}$);

(iii) $\ell(\hat{q}_{\mathcal{B}}) + \ell^*(-\hat{y}_{\mathcal{B}}) = 0$ (extremality relation) *for some* $\hat{y}_{\mathcal{B}} \in S_{\mathcal{B}}$*, where*

$$\hat{q}_{\mathcal{B}} \triangleq \left(\frac{\nu^a}{1+\hat{y}^a_{\mathcal{B}}}, 0^p \right);$$

(iv) $(\hat{y}^a_{\mathcal{B}}, -1^p) \in C^*$.

Informally put, Theorem 16(a) demonstrates that the BD-dual breaks down if C is either an H-set or a Z-set with respect to the observed type ν. Then the $(\mathcal{B} \to \mathcal{P})$ part of Theorem 8 does not apply. At the same time the $(\mathcal{P} \to \mathcal{B})$ part of Theorem 8 does not hold, as $S_{\mathcal{P}} \neq \emptyset$, yet $S_{\mathcal{B}} = \emptyset$. This is illustrated by Examples 10–13.

Part (b) of Theorem 16 captures the other discomforting fact about the BD-dual: if the solution of $\mathcal{B}$ exists, it may not solve the primal problem $\mathcal{P}$. By (i) this happens whenever the solution $\hat{q}_{\mathcal{P}}$ of $\mathcal{P}$ assigns a positive weight to at least one of the passive letters (provided that $S_{\mathcal{B}} \neq \emptyset$). See Example 14.

At least, for $\nu > 0$ the BD-dual works well.

Corollary 17. *If* $\nu > 0$ *then* $S_{\mathcal{P}} = \{\hat{q}_{\mathcal{P}}\}, S_{\mathcal{F}} = \{\hat{y}_{\mathcal{F}}\}, S_{\mathcal{B}} = \{\hat{y}_{\mathcal{B}}\}$ *are singletons,*

$$\hat{\ell}_{\mathcal{B}} = \hat{\ell}_{\mathcal{F}} = -\hat{\ell}_{\mathcal{P}}, \qquad \hat{y}_{\mathcal{B}} = \hat{y}_{\mathcal{F}} = \frac{\nu}{\hat{q}_{\mathcal{P}}} - 1, \qquad \text{and} \qquad \hat{q}_{\mathcal{P}} \perp \hat{y}_{\mathcal{B}}.$$

The corollary justifies the use of the BD-dual for solving $\mathcal{P}$ when $\nu > 0$. Recall that in the case of linear C the BD-dual is just Smith's simplified Lagrange dual problem (4.1), which is an unconstrained optimization problem. It can be solved numerically by standard methods for unconstrained optimization or by El Barmi & Dykstra's [12] cyclic ascent algorithm.

Finally, it is worth noting that the solution sets $S_{\mathcal{P}}$ and $S_{\mathcal{F}}$ are always compact but $S_{\mathcal{B}}$, if nonempty, is compact if and only if $\nu > 0$, that is, there is no passive letter. If $\nu \not> 0$ and $S_{\mathcal{B}} \neq \emptyset$, then $S_{\mathcal{B}}$ is unbounded from below.

4.3.1. Base solution of $\mathcal{B}$ and no BD-duality gap

The case (iv) of Theorem 16(b) provides a way to find out whether a solution of $\mathcal{P}$ assigns the zero weights to the passive letters or not. First, determine whether C is neither an H-set nor a Z-set with respect to ν. Then, solve the BD-dual $\mathcal{B}$ and find a solution $\hat{y}_{\mathcal{B}}$ of it. Finally, verify that $(\hat{y}^a_{\mathcal{B}}, -1^p)$ belongs to the polar cone C^*. For example, if $\hat{y}^p_{\mathcal{B}} \geq -1$ then this is satisfied automatically.

On the other hand, in order to have $\hat{q}^p_{\mathcal{P}} = 0^p$ (that is, to have no BD-duality gap), $\hat{y}^p_{\mathcal{B}} \geq -1$ must be satisfied by some $\hat{y}^p_{\mathcal{B}} \in S_{\mathcal{B}}$. In the case when C is polyhedral, there must exist a base solution $\hat{y}^p_{\mathcal{B}} = \sum_h \hat{\alpha}_h u_h$ of $\mathcal{B}$ with $\hat{y}^p_{\mathcal{B}} \geq -1$. The next Example illustrates the point.

Example 18 (Base solution of $\mathcal{B}$). Let $\mathcal{X} = \{-1, 1, a, b\}$, where $b > a > 1$. Let $C = \{q \in \Delta_{\mathcal{X}} : \mathrm{E}_q X = 0\}$, so that $u = (-1, 1, a, b)$. Let $\nu = (\nu_{-1}, \nu_1, 0, 0)$; hence $\mathcal{X}^a = \{-1, 1\}$, $\mathcal{X}^p = \{a, b\}$, and $C^a(0^p) = \{(1,1)/2\}$. For what values of ν_1 the solution of $\mathcal{P}$ assigns zero weights to the passive letters? First, $\hat{\alpha}_{\mathcal{B}}$ solves

$\langle \hat{q}^a_{\mathcal{B}}, u^a \rangle = 0$, where $\hat{q}^a_{\mathcal{B}} = \frac{\nu^a}{1+\hat{\alpha}_{\mathcal{B}} u^a}$. This gives $\hat{\alpha}_{\mathcal{B}} = 2\nu_1 - 1$. Thus, $\hat{\alpha}_{\mathcal{B}} < 0$ when $\nu_1 < 1/2$. Then, in order to have $\hat{q}^p_{\mathcal{P}} = 0^p$, the condition $\hat{y}^p_{\mathcal{B}} \geq -1$ gives that $\nu_1 \geq \frac{b-1}{2b}$. In the other case ($\hat{\alpha}_{\mathcal{B}} \geq 0$) the condition is not binding, hence $\hat{q}^p_{\mathcal{P}} = 0^p$ for $\nu_1 \geq 1/2$. Taken together, for $\nu_1 < \frac{b-1}{2b}$ it holds that $\hat{q}_{\mathcal{B}} \neq \hat{q}_{\mathcal{P}}$, since $\hat{q}^p_{\mathcal{P}}$ has a positive coordinate and $\hat{q}^p_{\mathcal{B}} = 0^p$. To give a numeric illustration, let $a = 2$, $b = 5$, and $\nu_1 = 3/10 < 4/10$. Then $\hat{q}_{\mathcal{P}} = (14, 9, 0, 1)/24$, that is, a positive weight is assigned to a passive letter. For $\nu_1 = 9/20$, which is above the threshold, $\hat{q}_{\mathcal{P}} = (1, 1, 0, 0)/2$.

To sum up, El Barmi and Dykstra's dual may fail to lead to the solution of the multinomial likelihood primal $\mathcal{P}$ in different ways. For a particular type ν the feasible set C may be an H-set or a Z-set, and then the BD-dual fails to attain finite infimum. Even if this is not the case $\mathcal{B}$ may fail to provide a solution of $\mathcal{P}$, due to the BD-duality gap. Theorem 16(b) states equivalent conditions under which the BD-dual is in the extremality relation with $\mathcal{P}$ and leads to a solution of $\mathcal{P}$; see also Lemma 58.

In the next two sections other possibilities of solving $\mathcal{P}$ are explored. First, an active-passive dualization is considered. Then, a perturbed primal problem and the PP algorithm are studied. Interestingly, a solution of the perturbed primal problem may be obtained from the BD-dual problem.

5. Active-passive dualization

Annotation. Sequential, active-passive dualization is proposed and analyzed. Its working is illustrated by solving Example 13, where the BD-dual fails due to the Z-set.

The *active-passive dualization* is based on a reformulation of the primal $\mathcal{P}$ as a sequence of partial minimizations

$$\hat{\ell}_{\mathcal{P}} = \inf_{q^p \in C^p} \inf_{q^a \in C^a(q^p)} \ell(q^a, q^p).$$

Assume that q^p is such that the slice $C^a(q^p)$ has support $\mathcal{X}^a$ (this is not a restriction, since otherwise the inner infimum is ∞). Since $\nu^a > 0$, Corollary 17 gives that a solution of the inner (active) primal problem $\mathcal{A}_\kappa$

$$\begin{aligned} \hat{\ell}_{\mathcal{P}}(q^p) &\triangleq \inf_{q^a \in C^a(q^p)} \ell(q^a, q^p), \\ S_{\mathcal{P}}(q^p) &\triangleq \{\hat{q}^a \in C^a(q^p) : \ \ell(\hat{q}^a, q^p) = \hat{\ell}_{\mathcal{P}}(q^p)\}, \end{aligned} \tag{$\mathcal{A}_\kappa$}$$

can be obtained from its BD-dual problem $\mathcal{B}_\kappa$

$$\begin{aligned} \hat{\ell}_{\mathcal{B}}(q^p) &\triangleq \sup_{y^a \in (C^a(q^p))^*} I_\kappa(\nu^a \,\|\, y^a), \\ S_{\mathcal{B}}(q^p) &\triangleq \{\hat{y}^a \in (C^a(q^p))^* : I_\kappa(\nu^a \,\|\, \hat{y}^a) = \hat{\ell}_{\mathcal{B}}(q^p)\}, \end{aligned} \tag{$\mathcal{B}_\kappa$}$$

where $\kappa = \kappa(q^p) \triangleq 1/(1 - \sum q^p)$.

Theorem 19 (Relation between $\mathcal{B}_\kappa$ and $\mathcal{A}_\kappa$)**.** *Let $q^p \in C^p$ be such that the support of $C^a(q^p)$ is $\mathcal{X}^a$. Then there is a unique solution $\hat{y}^a(q^p)$ of $\mathcal{B}_\kappa$, and*

$$\hat{q}^a(q^p) \triangleq \frac{\nu^a}{\kappa(q^p) + \hat{y}^a(q^p)} \tag{5.1}$$

is the unique member of $S_{\mathcal{P}}(q^p)$. Moreover, $\hat{q}^a(q^p) \perp \hat{y}^a(q^p)$ and $\hat{\ell}_{\mathcal{B}}(q^p) = -\hat{\ell}_{\mathcal{P}}(q^p)$.

Thanks to (5.1) and the extremality relation between $\mathcal{A}_\kappa$ and $\mathcal{B}_\kappa$, the *active-passive (AP) dual form* of the active-passive primal is

$$\sup_{q^p \in C^p} \sup_{y^a \in (C^a(q^p))^*} I_\kappa(\nu^a \,\|\, y^a).$$

The active-passive dualization is illustrated by the following example.

Example 13 (cont'd). Here $C^p = [0, 1/2]$ and the AP dual can be written in the form

$$\hat{q}_1^p = \operatorname*{argmax}_{q_1^p \in C^p} \sup_{\alpha \in \mathbb{R}} I_{\kappa(q_1^p)}(\nu^a \,\|\, \alpha v^a),$$

where $v^a = u^a + (\kappa(q_1^p)\, q_1^p u_1^p) 1^a$. The inner optimization gives

$$\hat{\alpha}(q_1^p) = \frac{1 - 3q_1^p}{2q_1^p(2q_1^p - 1)}.$$

This $\hat{\alpha}(q_1^p)$, plugged into $I_{\kappa(q_1^p)}(\nu^a \,\|\, \alpha v^a)$, yields

$$(1/2) \log\left[1 + \hat{\alpha}(q_1^p)(2q_1^p - 1)\right] \;+\; (1/2) \log\left[1 + \hat{\alpha}(q_1^p) q_1^p\right] \;-\; \log[1 - q_1^p],$$

which has to be maximized over $q_1^p \in C^p = [0, 1/2]$. The maximum is attained at $\hat{q}_1^p = 1/4$. Thus $\hat{\alpha}(\hat{q}_1^p) = -1$, $\kappa(\hat{q}_1^p) = 4/3$, $v_{-1}^a = -2/3$, and $v_0^a = 1/3$; since $\hat{q}^a(\hat{q}_1^p) = \nu^a / (\kappa(\hat{q}_1^p) + \hat{\alpha}(\hat{q}_1^p) v^a)$ by (5.1), $\hat{q}_{-1}^a = 1/4$ and $\hat{q}_0^a = 1/2$. Hence, $\hat{q} = (1, 2, 1)/4$.

In the outer, passive optimization, it is possible to exploit the structure of $S_{\mathcal{P}}$ (cf. Theorem 2), and this way reduce the dimension of the problem. This is the case, for instance, when C is polyhedral.

6. Perturbed primal $\mathcal{P}_\delta$ and PP algorithm

Annotation. Perturbed primal problem $\mathcal{P}_\delta$ and the PP algorithm are introduced. The epi-convergence of a sequence of the perturbed primals for a general, convex C, and the pointwise convergence for the linear C are formulated (cf. Theorems 20, 21) and illustrated.

For $\delta > 0$ let $\nu(\delta) \in \Delta_{\mathcal{X}}$ be a perturbation of the type ν; we assume that

$$\nu(\delta) > 0 \qquad \text{and} \qquad \lim_{\delta \searrow 0} \nu(\delta) = \nu. \tag{6.1}$$

The perturbation activates passive, unobserved letters. For every $\delta > 0$ consider the *perturbed primal problem* $\mathcal{P}_\delta$

$$\begin{aligned} \hat{\ell}_{\mathcal{P}}(\delta) &\triangleq \inf_{q \in C} \ell_\delta(q) = \inf_{q \in C} -\langle \nu(\delta), \log q \rangle \,, \\ S_{\mathcal{P}}(\delta) &\triangleq \{\hat{q} \in C :\; \ell_\delta(\hat{q}) = \hat{\ell}_{\mathcal{P}}(\delta)\}, \end{aligned} \tag{$\mathcal{P}_\delta$}$$

where $\ell_\delta \triangleq \ell_{\nu(\delta)}$.

Since the activated type $\nu(\delta)$ has no passive coordinate, the perturbed primal problem $\mathcal{P}_\delta$ can be solved, for instance, via the BD-dualization; recall Corollary 17. Thus, for every $\delta > 0$,

$$S_{\mathcal{P}}(\delta) = \{\hat{q}_{\mathcal{P}}(\delta)\}, \quad S_{\mathcal{F}}(\delta) = S_{\mathcal{B}}(\delta) = \{\hat{y}_{\mathcal{B}}(\delta)\}, \quad \hat{q}_{\mathcal{P}}(\delta) = \frac{\nu(\delta)}{1 + \hat{y}_{\mathcal{B}}(\delta)} \,. \tag{6.2}$$

How is $\hat{q}_{\mathcal{P}}(\delta)$ related to $S_{\mathcal{P}}$, and $\hat{\ell}_{\mathcal{P}}(\delta)$ to $\hat{\ell}_{\mathcal{P}}$? Theorem 20 asserts that ℓ_δ^C epi-converges to ℓ_ν^C when $\delta \searrow 0$. (There, for a map $f : D \to \bar{\mathbb{R}}$ and a set $C \subseteq D$, the map $f^C : \mathbb{R}^m \to \bar{\mathbb{R}}$ is given by $f^C(x) = f(x)$ if $x \in C$, and $f^C(x) = \infty$ if $x \notin C$.) The epi-convergence (cf. Rockafellar and Wets [42, Chap. 7]) is used in convex analysis to study a limiting behavior of perturbed optimization problems. It is an important modification of the uniform convergence (cf. Kall [26], Wets [47]).

Theorem 20 (Epi-convergence of $\mathcal{P}_\delta$ to $\mathcal{P}$)**.** *Assume that C is a convex closed subset of $\Delta_{\mathcal{X}}$ with support $\mathcal{X}$, $\nu \in \Delta_{\mathcal{X}}$, and $(\nu(\delta))_{\delta>0}$ is such that (6.1) is true. Then*

$$\ell_\delta^C \quad \textit{epi-converges to} \quad \ell_\nu^C \qquad \textit{for} \quad \delta \searrow 0.$$

Consequently,

$$\lim_{\delta \searrow 0} \inf_{\hat{q}_{\mathcal{P}} \in S_{\mathcal{P}}} d(\hat{q}_{\mathcal{P}}(\delta), \hat{q}_{\mathcal{P}}) = 0$$

and the active coordinates of solutions of $\mathcal{P}_\delta$ converge to the unique point $\hat{q}_{\mathcal{P}}^a$ of $S_{\mathcal{P}}^a$:

$$\lim_{\delta \searrow 0} \hat{q}_{\mathcal{P}}^a(\delta) = \hat{q}_{\mathcal{P}}^a \in S_{\mathcal{P}}^a.$$

Moreover, if $S_{\mathcal{P}}$ is a singleton *(particularly, if $\nu > 0$) then also the passive coordinates converge and*

$$\lim_{\delta \searrow 0} \hat{q}_{\mathcal{P}}(\delta) \in S_{\mathcal{P}}.$$

Thus, if δ is small enough, the (unique) solution $\hat{q}_{\mathcal{P}}(\delta)$ of the perturbed primal $\mathcal{P}_\delta$ is close to a solution of the primal problem $\mathcal{P}$. Theorem 20 also states that in the active coordinates the convergence is pointwise, to the unique $\hat{q}_{\mathcal{P}}^a$ of $S_{\mathcal{P}}^a$. The next theorem demonstrates that if C is given by linear constraints and $\nu(\delta)$ is defined in a 'uniform' way in δ, also the passive coordinates of $\hat{q}_{\mathcal{P}}(\delta)$ converge pointwise.

Theorem 21 (Convergence of $\mathcal{P}_\delta$ to $\mathcal{P}$, linear C)**.** *Let C be given by (2.3), $\nu \in \Delta_{\mathcal{X}}$, and $(\nu(\delta))_{\delta>0}$ be such that (6.1) is true. Further, assume that $\nu(\cdot)$*

is continuously differentiable and that there is a constant $c > 0$ such that, for every $i \in \mathcal{X}^p$,

$$|\vartheta_i'(\delta)| \le c\,\vartheta_i(\delta), \qquad \textit{where} \quad \vartheta_i(\delta) \triangleq \frac{\nu_i(\delta)}{\sum_{j\in\mathcal{X}^p} \nu_j(\delta)}\,. \tag{6.3}$$

Then

$$\lim_{\delta\searrow 0} \hat{q}_{\mathcal{P}}(\delta) \quad \textit{exists and belongs to } S_{\mathcal{P}}.$$

Notice that the condition (6.3) is satisfied if $\nu_i(\delta) = \nu_j(\delta)$ for every $i, j \in \mathcal{X}^p$; for example if

$$\nu(\delta) = \frac{1}{1+\delta m_p}(\nu + \delta\xi), \tag{6.4}$$

where $\xi \triangleq (0^a, 1^p)$ is the vector with $\xi_i = 1$ if $i \in \mathcal{X}^p$ and $\xi_i = 0$ if $i \in \mathcal{X}^a$. This corresponds to the case when every passive coordinate is 'activated' by equal weight.

The following example demonstrates that without the assumption (6.3) the convergence in passive letters need not occur.

Example 22 (Divergent $\hat{q}_{\mathcal{P}}(\delta)$). Consider the setting of Example 3. That is, $\mathcal{X} = \{-1, 0, 1\}$ and

$$C = \{q \in \Delta_{\mathcal{X}} : \langle q, u\rangle = 0\} = \{(a, 1/2, 1/2 - a) : a \in [0, 1/2]\},$$

where $u = (1, -1, 1)$. For $\nu = (0, 1, 0)$, $S_{\mathcal{P}} = C$. Define the perturbed types $\nu(\delta)$ in such a way that

$$\nu(\delta) = \begin{cases} (\delta, 1-3\delta, 2\delta) & \text{if } \delta \in \{1/(2n) : n \in \mathbb{N}\}, \\ (2\delta, 1-3\delta, \delta) & \text{if } \delta \in \{1/(2n+1) : n \in \mathbb{N}\}, \end{cases}$$

and $\nu(\cdot)$ is C^1 on $(0, 1)$. Then, for every $n \in \mathbb{N}$,

$$\hat{q}_{\mathcal{P}}(1/(2n)) = (1/6, 1/2, 1/3) \qquad \text{and} \qquad \hat{q}_{\mathcal{P}}(1/(2n+1)) = (1/3, 1/2, 1/6);$$

so the limit $\lim \hat{q}_{\mathcal{P}}(\delta)$ does not exist. Note that in this case the condition (6.3) is violated. Indeed, $\vartheta_1(1/(2n)) = 2/3$ and $\vartheta_1(1/(2n+1)) = 1/3$ for every n; hence, by the mean value theorem, for every n there is $\zeta_n \in (1/(2n+1), 1/(2n))$ such that $\vartheta_1'(\zeta_n) = 2n(n+1)/3$. Since $0 < \vartheta_1(\zeta_n) \le 1$, $\lim_n \vartheta_1'(\zeta_n)/\vartheta_1(\zeta_n) = \infty$.

In Examples 10, 13, 14, with $\nu(\delta)$ given by (6.4), the pointwise convergence can be demonstrated analytically.

Example 10 (cont'd). Here $\hat{y}_{\mathcal{B}}(\delta) = \hat{\alpha}_{\mathcal{B}}(\delta)u$, where $\hat{\alpha}_{\mathcal{B}}(\delta) = \frac{\delta-1}{\delta+1}$. So, $\lim \hat{\alpha}_{\mathcal{B}}(\delta) = -1$ and $\lim \hat{q}_{\mathcal{P}}(\delta) = (1, 0, 1)/2 \in S_{\mathcal{P}}$.

Example 13 (cont'd). First, $\hat{y}_{\mathcal{B}}(\delta) = \hat{\alpha}_{\mathcal{B}}(\delta)u$ and

$$\hat{\alpha}_{\mathcal{B}}(\delta) = \operatorname*{argmax}_{\alpha\in\mathbb{R}} \left[\nu_{-1}(\delta)\log(1-\alpha) + \nu_0(\delta)\log 1 + \nu_1(\delta)\log(1+\alpha)\right];$$

this leads to $\hat{\alpha}_{\mathcal{B}}(\delta) = \frac{2\delta-1}{2\delta+1}$. Thus, $\lim \hat{\alpha}_{\mathcal{B}}(\delta) = -1$ and, since $\hat{q}_{\mathcal{P}}(\delta) = \frac{\nu(\delta)}{1+\hat{\alpha}_{\mathcal{B}}(\delta)u}$, $\lim \hat{q}_{\mathcal{P}}(\delta) = (1, 2, 1)/4 \in S_{\mathcal{P}}$.

Example 14 (cont'd). Here $\hat{y}_{\mathcal{B}}(\delta) = \hat{\alpha}_{\mathcal{B}}(\delta)u$ with $\hat{\alpha}_{\mathcal{B}}(\delta) = -\frac{3-\sqrt{1+392\delta+400\delta^2}}{20(1+\delta)}$. Thus, $\lim \hat{\alpha}_{\mathcal{B}}(\delta) = -1/10$, and $\lim \hat{q}_{\mathcal{B}}(\delta) = (54, 44, 1)/99 \in S_{\mathcal{P}}$.

The next example provides a numeric illustration of the pointwise convergence of a sequence of perturbed primals to $\mathcal{P}$. The perturbed primal solutions are obtained through their BD-duals. It is worth stressing that $\mathcal{B}$ to $\mathcal{P}_\delta$ is, for a linear C, an unconstrained optimization problem; cf. Section 4.1.1.

Example 23 (Qin and Lawless [38], Ex. 1). Consider a discrete-case analogue of Example 1 from Qin and Lawless [38]. Let $\mathcal{X} = \{-2, -1, 0, 1, 2\}$ and

$$C_\theta = \{q \in \Delta_{\mathcal{X}} : \mathrm{E}_q(X - \theta) = 0, \mathrm{E}_q(X^2 - 2\theta^2 - 1) = 0\},$$

where $\theta \in \Theta = [-2, 2]$. Then $C_\Theta = \bigcup_{\theta\in\Theta} C_\theta$ is the estimating equations model; cf. (3.1). Clearly, $u_1(\theta) = (-2-\theta, -1-\theta, -\theta, 1-\theta, 2-\theta)$ and $u_2(\theta) = (3 - 2\theta^2, -2\theta^2, -1-2\theta^2, -2\theta^2, 3-2\theta^2)$. Let $\nu = (0, 0, 7, 3, 0)/10$.

For a fixed $\theta \in \Theta$ and a perturbed type $\nu(\delta)$, the BD-dual to the perturbed primal $\mathcal{P}_\delta$ is (equivalent to)

$$\hat{\alpha}_{\mathcal{B}}(\delta) = \operatorname*{argmin}_{\alpha\in\mathbb{R}^2} I_1(\nu^a \,\|\, \langle \alpha, u(\theta)\rangle)$$

and the corresponding $\hat{q}_{\mathcal{B}}(\delta) = \frac{\nu(\delta)}{1+\langle\hat{\alpha}(\delta), u(\theta)\rangle}$. For $\theta = 0$ and $\nu(\delta)$ given by (6.4) with $\delta = 10^{-j}$ ($j = 3, 5, 7, 9$), Table 1 illustrates the pointwise convergence of $\hat{q}_{\mathcal{B}}(\delta)$ to $\hat{q}_{\mathcal{P}}$, $-\hat{\ell}_{\mathcal{B}}(\delta)$ to $\hat{\ell}_{\mathcal{P}}$.

The optimal $\hat{\alpha}_{\mathcal{B}}(\delta)$'s were computed by `optim` of `R` [39]. The rightmost three columns in Table 1 state orders of the precision $10^{-\gamma}$ of satisfaction of the constraints $\mathrm{E}_{\hat{q}(\delta)} X = 0$, $\mathrm{E}_{\hat{q}(\delta)}(X^2 - 1) = 0$, and $\sum \hat{q}(\delta) - 1 = 0$. The solution of $\mathcal{P}$, obtained by `solnp` from the `R` library `Rsolnp` (cf. Ghalanos and Theussl [17]; based on Ye [48]), is $\hat{q}_{\mathcal{P}} = (0.1625, 0, 0.525, 0.3, 0.0125)$ and $\hat{\ell}_{\mathcal{P}} = 0.812242$.

TABLE 1
The pointwise convergence of $\mathcal{P}_\delta$ to $\mathcal{P}$.

j	$\hat{q}_{\mathcal{B}}(\delta)$					$-\hat{\ell}_{\mathcal{B}}(\delta)$	γ_1	γ_2	γ_3
3	0.161439	0.001013	0.528326	0.294553	0.014669	0.823788	7	7	8
5	0.162488	0.000010	0.525041	0.299936	0.012525	0.812404	7	7	6
7	0.162501	1e-7	0.525000	0.299999	0.012500	0.812242	6	6	6
9	0.162501	1e-8	0.525000	0.300000	0.012502	0.812242	5	5	6

For $j > 9$ the numerical effects become noticeable. For instance, for $j = 20$, the precision of the constraints satisfaction is of the order 10^{-1}.

As an aside, note that for this type ν the BD-dual to the original, unperturbed primal $\mathcal{P}$ breaks down, since C_θ is an H-set. In fact, it is an H-set with respect to this ν for any $\theta \in \Theta$; cf. the empty set problem in Grendár and Judge [19].

The convergence theorems suggest that the practice of replacing the zero counts by an 'ad hoc' value can be superseded by the *PP algorithm*; i.e., by a sequence of the perturbed primal problems, for $\nu(\delta) > 0$ such that $\lim_{\delta\searrow 0} \nu(\delta) = \nu$. Since each $\nu(\delta) > 0$, the PP algorithm can be implemented through the BD-dual to $\mathcal{P}_\delta$, by the Fisher scoring algorithm, or by the Gokhale algorithm [18], among other methods.

7. Implications for empirical likelihood

Annotation. Consequences of the presented results for the empirical likelihood (EL) method (cf. Owen [34]) are pointed out. It is noted that the distribution that maximizes empirical likelihood differs, in general, from the distribution that maximizes multinomial likelihood. The multinomial likelihood ratio may lead to different inferential and evidential conclusions than the empirical likelihood ratio. The case of continuous data is discussed in Section 7.1, where Fisher's original notion of likelihood is recalled.

In most settings, including the discrete one, empirical likelihood is 'a multinomial likelihood on the sample', Owen [35, p. 15]. It is usually applied to an *empirical estimating equations* model, which is in the discrete setting defined as $C_{\Theta,\nu^a} \triangleq \bigcup_{\theta\in\Theta} C_{\theta,\nu^a}$, where

$$C_{\theta,\nu^a} \triangleq \{p \in \Delta_{\mathcal{X}^a} : \langle p, u_h^a(\theta)\rangle = 0 \text{ for } h = 1, 2, \dots, r\}$$

and $u_h^a : \Theta \to \mathbb{R}^{m_a}$ are the *empirical estimating functions*. The empirical likelihood estimator is defined through

$$\inf_{\theta\in\Theta} \inf_{p\in C_{\theta,\nu^a}} -\langle \nu^a, \log p\rangle\,. \tag{7.1}$$

For a fixed $\theta \in \Theta$, the data-supported feasible set C_{θ,ν^a} is a convex set and the inner optimization in (7.1) is the *empirical likelihood inner problem*

$$\hat{\ell}_{\mathcal{E}} \triangleq \inf_{p\in C_{\theta,\nu^a}} \ell(p), \qquad S_{\mathcal{E}} \triangleq \{\hat{p}_{\mathcal{E}} \in C_{\theta,\nu^a} : \ell(\hat{p}_{\mathcal{E}}) = \hat{\ell}_{\mathcal{E}}\}. \tag{$\mathcal{E}$}$$

Since C_{θ,ν^a} is just the 0^p-slice $C_\theta^a(0^p)$ of C_θ (given by (3.1)), the EL inner problem $\mathcal{E}$ can equivalently be expressed as

$$\hat{\ell}_{\mathcal{E}} = \inf_{q^a\in C_\theta^a(0^p)} \ell(q^a).$$

Its dual is

$$\inf_{\alpha\in\mathbb{R}^r} I_1\left(\nu^a \,\Big\|\, \sum_h \alpha_h u_h^a(\theta)\right). \tag{7.2}$$

Note that (7.2) is just Smith's simplified Lagrangean (4.1), that is, the BD-dual $\mathcal{B}$ to the multinomial likelihood primal problem $\mathcal{P}$, for the linear set C_θ. This connection implies, through Theorem 16, that the maximum of empirical likelihood does not exist if C_θ is either an H-set or a Z-set with respect to ν. The two possibilities are recognized in the literature on EL, where an H-set is referred to as the *convex hull problem* (cf. Owen [35, Sect. 10.4]), and a Z-set is known as the *zero likelihood problem* (cf. Bergsma et al. [8]). Theorem 16 also implies that these are the only ways the EL inner problem may fail to have a solution. Note that $\mathcal{E}$ may fail to have a solution for any $\theta \in \Theta$; cf. the *empty set problem*, Grendár and Judge [19].

In addition, Theorem 16 implies that, besides failing to exist, the EL inner problem $\mathcal{E}$ may have different solution than the multinomial likelihood primal problem $\mathcal{P}$. If C_θ is neither an H-set nor a Z-set then, by Theorem 16(b), it is possible that

1. either $\hat{\ell}_{\mathcal{P}} = -\hat{\ell}_{\mathcal{B}}$ (no BD-duality gap), or
2. $\hat{\ell}_{\mathcal{P}} < -\hat{\ell}_{\mathcal{B}}$ (BD-duality gap).

Since $\hat{\ell}_{\mathcal{B}} = -\hat{\ell}_{\mathcal{E}}$, in the latter case $\hat{\ell}_{\mathcal{P}} < \hat{\ell}_{\mathcal{E}}$ and $S_{\mathcal{P}} \neq S_{\mathcal{E}}$. This happens when any of the conditions (i)–(iv) from Theorem 16(b) is not satisfied. Then the distribution that maximizes empirical likelihood differs from the distribution that maximizes multinomial likelihood. Moreover, the multinomial likelihood ratio may lead to different inferential and evidential conclusions than the empirical likelihood ratio. The points are illustrated in the next example.

Example 24 (LR vs. ELR). Let $\mathcal{X} = \{-2, -1, 0, 1, 2\}$, $\Theta = \{\theta_1, \theta_2\}$ with $\theta_1 = 1.01$ and $\theta_2 = 1.05$. Let $C_{\theta_j} = \{q \in \Delta_{\mathcal{X}} : \mathrm{E}_q(X^2) = \theta_j\}$. Clearly, $u(\theta_j) = (4-\theta_j, 1-\theta_j, -\theta_j, 1-\theta_j, 4-\theta_j)$ for $j = 1, 2$. Let $\nu = (6, 3, 0, 0, 1)/10$.

The solution of $\mathcal{P}$ is

- $\hat{q}_{\mathcal{P}}(\theta_1) = (0.1515, 0.3030, 0.52025, 0, 0.02525)$, for θ_1,
- $\hat{q}_{\mathcal{P}}(\theta_2) = (0.1575, 0.3150, 0.50125, 0, 0.02625)$, for θ_2.

Note that each of the solutions assigns a positive weight to the unobserved outcome 0. Such a solution cannot be obtained by the BD dual, due to the presence of the BD-duality gap. Recall that there is no BD-gap if $\hat{y}^p_{\mathcal{B}} \geq -1$; cf. Theorem 16 and Sect. 4.3.1. For θ_1, $\hat{\alpha}_{\mathcal{B}}(\theta_1) = 69.8997$, hence $\hat{y}^p_{\mathcal{B}}(\theta_1) = (-70.5987, -0.6990)$, thus the BD-gap is present. Similarly for θ_2, where $\hat{\alpha}_{\mathcal{B}}(\theta_2) = 13.8983$, hence $\hat{y}^p_{\mathcal{B}}(\theta_2) = (-14.5932, -0.6949)$.

As the two solutions are very close, the multinomial likelihood ratio is

- $\mathrm{LR}_{21} = \exp\left(n[\ell(\hat{q}_{\mathcal{P}}(\theta_1)) - \ell(\hat{q}_{\mathcal{P}}(\theta_2))]\right) = 1.4746$,

which indicates inconclusive evidence.

However, the empirical likelihood ratio leads to a very different conclusion. Note that the active letters are $\mathcal{X}^a = \{-2, -1, 2\}$, and C_θ is neither an H-set nor a Z-set with respect to the observed type ν, for the considered θ_j $(j = 1, 2)$. Hence for both θ's the solution of $\mathcal{E}$ exists and it is

- $\hat{q}_{\mathcal{E}}(\theta_1) = (0.00286, 0.99\bar{6}, 0.00048)$, for θ_1,
- $\hat{q}_{\mathcal{E}}(\theta_2) = (0.01429, 0.98\bar{3}, 0.00238)$, for θ_2.

The weights given by EL to -2 are very different in the two models; the same holds for 2. The empirical likelihood ratio is

- $\mathrm{ELR}_{21} = \exp\left(n[\ell(\hat{q}_{\mathcal{E}}(\theta_1)) - \ell(\hat{q}_{\mathcal{E}}(\theta_2))]\right) = 75031.31$,

which indicates decisive support for θ_2; cf. Zhang [50].

The BD-duality gap thus implies that in the discrete *iid* setting, when C is given by linear equality constraints, EL-based inferences from finite samples may be grossly misleading.

7.1. Continuous case and Fisher likelihood

As far as the continuous random variables are concerned, due to the finite precision of any measurement 'all actual sample spaces are discrete, and all observable random variables have discrete distributions', Pitman [37, p. 1]. Already Fisher's original notion of the likelihood [15] (see also Lindsey [31, p. 75]) reflects the finiteness of the sample space. For an *iid* sample $X_1^n \triangleq (X_1, X_2, \dots, X_n)$ and a finite partition $\mathcal{A} = \{A_l\}_1^m$ of a sample space $\mathcal{X}$, the *Fisher likelihood* $L_{\mathcal{A}}(q; X_1^n)$ which the data X_1^n provide to a pmf $q \in \Delta_{\mathcal{X}}$ is

$$L_{\mathcal{A}}(q; X_1^n) \triangleq \prod_{A_l \in \mathcal{A}} e^{n(A_l) \log q(A_l)},$$

where $n(A_l)$ is the number of observations in X_1^n that belong to A_l. Thus, this view carries the discordances between the multinomial and empirical likelihoods also to the continuous *iid* setting.

Example 25 (FL with estimating equations). To give an illustration of the Fisher likelihood as well as yet another example that $\hat{q}_{\mathcal{E}}$ may be different than $\hat{q}_{\mathcal{P}}$, consider the setting of Example 23 with $\mathcal{X} = \{-4, -3.9, \dots, 3.9, 4\}$ and $\theta \in \Theta = [-4, 4]$. The letters of the alphabet are taken to be the representative points of the partition $\mathcal{A} \triangleq \{(-\infty, -3.95), [-3.95, -3.85), \dots, [3.85, 3.95), [3.95, \infty)\}$ of $\mathbb{R}$. This way the alphabet captures the finite precision of measurements of a continuous random variable. The type ν exhibited at the panel c) of Figure 1 is induced by a random sample of size $n = 100$ from the $\mathcal{A}$-quantized standard normal distribution. The EL estimate of θ is -0.052472 and the associated EL-maximizing distribution $\hat{q}_{\mathcal{E}}$ is different than the multinomial likelihood maximizing distribution $\hat{q}_{\mathcal{P}}$, which is associated with the estimated value 0.000015 and assigns a positive weight also to the passive letters -4 and 4.

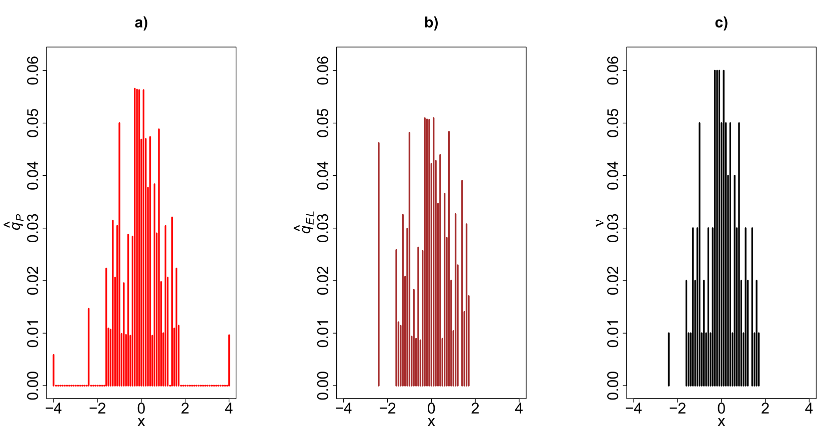

FIG 1. *Panel a) the multinomial likelihood maximizing distribution $\hat{q}_{\mathcal{P}}$; panel b) the empirical likelihood maximizing distribution $\hat{q}_{\mathcal{E}}$; and panel c) the observed type ν.*

8. Implications for other methods

Annotation. Besides the empirical likelihood, the minimum discrimination information, Lindsay geometry, compositional data analysis, bootstrap in the presence of auxiliary information, and Lagrange multiplier test ignore information about the alphabet, and are restricted to the observed data. Thus, they are affected by the above findings.

8.1. Contingency tables with given marginals and MDI

In the analysis of contingency tables with given marginals, the minimum discrimination information (MDI) method by Ireland and Kullback [25] is more popular than the maximum multinomial likelihood method. This is because the former is more computationally tractable, thanks to the generalized iterative scaling algorithm (cf. Ireland et al. [24]). MDI minimizes $I_0(q \,\|\, \nu)$ over q, so that a solution of the MDI problem must assign a zero mass to a passive, unobserved letter. Thus, MDI is effectively an empirical method. This implies that the MDI-minimizing distribution restricted to a convex closed set C may not exist; however, the multinomial likelihood-maximizing distribution always exists (cf. Theorem 2), and may assign a positive mass to an unobserved outcome(s).

Example 26 (Contingency table with given marginals). Consider a 3×3 contingency table with given marginals. Let $\mathcal{X} = \{1, 2, 3\} \times \{1, 2, 3\}$, and let the observed bi-variate type ν have all the mass concentrated to $(1, 1)$; the remaining eight possibilities have got zero counts. Let the column and raw marginals be $f_c = (1, 2, 7)/10$, $f_r = (5, 4, 1)/10$, respectively. One of the multinomial likelihood maximizing distributions $\hat{q}_{\mathcal{P}}$ is displayed in Table 2. In the active letter $\hat{q}^a_{\mathcal{P}} = 0.1$ is unique, in the passive letters $\hat{q}^p_{\mathcal{P}} \in C^p(q^a_{\mathcal{P}})$. The table exhibits the $\hat{q}^p_{\mathcal{P}}$ which can also be obtained by the PP algorithm with the uniform activation (6.4). Note that $C^a(0^p) = \emptyset$, so that the MDI-minimizing distribution does not exist.

TABLE 2
A solution of $\mathcal{P}$.

$\hat{q}_{\mathcal{P}}$		
0.1000	0.0000	0.0000
0.0945	0.0800	0.0255
0.3055	0.3200	0.0745

It is worth stressing that the PP algorithm makes the multinomial likelihood primal problem $\mathcal{P}$ computationally feasible. In particular, the BD-dual implementation of the PP algorithm can be used to reduce dimensionality of the optimization problem from card $\mathcal{X}$ to the number of the given marginals.

Example 26 (cont'd). A reviewer suggested to consider the above example with a type having $1/3$ assigned to ν_{11}, ν_{12}, and ν_{32}. One of the solutions of $\mathcal{P}$, obtained by `solnp` from the `R` library `Rsolnp` with the uniform initialization, is presented in Table 3. The same solution can be obtained by the PP algorithm with the uniform activation.

TABLE 3
A solution of $\mathcal{P}$.

$\hat{q}_{\mathcal{P}}$		
0.0535	0.0465	0.0000
0.1710	0.0000	0.0290
0.2755	0.3535	0.0710

The same active components of the solution are obtained by the Fenchel dual. As C is linear, a solution of $\mathcal{F}$ takes the form $\sum_h \hat{\alpha}_h u_h$, where $\hat{\alpha} = (6.2291, 0.0001, -0.0004, 0.9417)$, and u_h's are stacked row-wise into

$$U = \begin{pmatrix} 9 & 9 & 9 & -1 & -1 & -1 & -1 & -1 & -1 \\ -2 & -2 & -2 & 8 & 8 & 8 & -2 & -2 & -2 \\ 5 & -5 & -5 & 5 & -5 & -5 & 5 & -5 & -5 \\ -4 & 6 & -4 & -4 & 6 & -4 & -4 & 6 & -4 \end{pmatrix} \frac{1}{10}.$$

By (3.5), the corresponding $\hat{q}^a_{\mathcal{P}} = (0.0535, 0.0465, 0.3537)$. In the passive components, a solution belongs to $C^p(q^a_{\mathcal{P}})$. Furthermore, since $\hat{y}^p_i \neq 1$ implies that $\hat{q}_i = 0$, the letters $(1,3)$ and $(2,2)$ are assigned the zero weight.

8.2. Marginal homogeneity in contingency tables

As a real-life example, requested by a reviewer, consider the marginal homogeneity model for the data on patient histology for two imaging modalities, studied by Sharma et al. [43]. The data form a two-way, 5×5 contingency table, displayed in Table 4.

TABLE 4
Type, corresponding to the data from Table 1 in [43], concerning an agreement of examinations of patients with Barrett's oesophagus by the High-definition White Light Endoscopy and by the Narrow Band Imaging.

	HD-WLE				
NBI	No IM	IM	LGD	HGD	OAC
No IM	0.0813	0.0488	0.0244	0.0000	0.0000
IM	0.0650	0.3577	0.0894	0.0081	0.0000
LGD	0.0000	0.1463	0.0732	0.0081	0.0000
HGD	0.0081	0.0081	0.0244	0.0325	0.0000
OAC	0.0000	0.0000	0.0000	0.0163	0.0081

The marginal homogeneity model (cf. Bishop, Fienberg and Holland [11]) assumes that the marginals of the bi-variate q are equal. The model can be represented as a system of linear constraints on the vectorized q. Thus, the problem of finding the maximum multinomial likelihood estimate of q under the marginal homogeneity constraints is an instance of the $\mathcal{P}$ problem.

There are nine cells with zero counts in Table 4. In the presence of zero-counts Simth's dual (cf. Sect. 4.1.1), considered in Madansky [33] or Bishop, Fienberg and Holland [11, p. 294], may not lead to the solution of $\mathcal{P}$. A solution of $\mathcal{P}$ obtained by the PP algorithm is exhibited in Table 5. Note that a positive weight is assigned to one of the nine passive letters. The solution cannot be obtained by Smith's simplified Lagrange dual. The marginal distribution, induced by the solution, is (0.1547, 0.5451, 0.2176, 0.0648, 0.0178).

TABLE 5
The maximum multinomial likelihood estimate of the sampling distribution under the marginal homogeneity model, based on the data in Table 4.

	HD-WLE				
NBI	No IM	IM	LGD	HGD	OAC
No IM	0.0813	0.0473	0.0260	0.0000	0.0000
IM	0.0671	0.3577	0.0986	0.0120	0.0097
LGD	0.0000	0.1339	0.0732	0.0106	0.0000
HGD	0.0063	0.0061	0.0198	0.0325	0.0000
OAC	0.0000	0.0000	0.0000	0.0097	0.0081

8.3. Lindsay geometry

Lindsay [30, Sect. 7.2] discusses multinomial mixtures under linear constraints on the mixture components, and assumes that it is sufficient to consider the distributions supported in the data (i.e., in the active alphabet). Though the objective function $\ell(\cdot)$ in $\mathcal{P}$ is a 'single-component' multinomial likelihood, the present results for the H-set, Z-set, and BD-gap suggest that it would be more appropriate to work with the complete alphabet; see also Anaya-Izquierdo et al. [5, Sect. 5.1].

8.4. Compositional data analysis

Multinomial likelihood maximization has the same solution regardless of whether the proportions ν or the counts $(n_i)_{i\in\mathcal{X}}$ are used. Note that the vector ν of proportions is an instance of the compositional data. In the analysis of compositional data, it is assumed that the compositional data $(x_1,\dots,x_m)$ belong to $\{(x_1,\dots,x_m) : x_1 > 0,\dots,x_m > 0, \sum x_i = 1\}$; cf. Aitchison [3, Sect. 2.2]. This assumption transforms $\nu \in \Delta_{\mathcal{X}}$ into $\nu^a \in \Delta_{\mathcal{X}^a}$. Consequently, the multinomial likelihood problem $\mathcal{P}$ is replaced by the empirical likelihood problem $\mathcal{E}$. However, this replacement is not without consequences, as the solution of the empirical likelihood problem $\mathcal{E}$ (if it exists) may differ from the solution of $\mathcal{P}$; cf. Section 7.

8.5. Bootstrap with auxiliary information

Bootstrap in the presence of auxiliary information (cf. Zhang [49], Hall and Presnell [22]) in the form of a convex closed set, resamples from the EL-maximizing distribution $\hat{q}_{\mathcal{E}}$. Hence, this method intentionally discards information about the alphabet. Resampling from $\hat{q}_{\mathcal{P}}$ seems to be a better option.

8.6. Score test

The Lagrange multiplier (score) test (cf. Silvey [44]) of the linear restrictions on q (cf. C given by (2.3)) fails if C is an H-set or a Z-set with respect to ν, because the Lagrangean first-order conditions do not lead to a finite solution of $\mathcal{P}$. However, the multinomial likelihood ratio exists.

9. Concluding comments

Computational aspects. There are several methods for obtaining a solution of $\mathcal{P}$, numerically.

- The primal problem $\mathcal{P}$ can be solved by the augmented lagrangean methods for constrained optimization. This may become burdensome for a large alphabet.
- A solution of $\mathcal{P}$ can also be obtained from the Fenchel dual $\mathcal{F}$. For a linear set C the optimization is unconstrained over $\mathbb{R}^r$, where r is the number of linear constraints. This way the Fenchel dual may greatly decrease dimensionality of the optimization problem. Evaluation of the convex conjugate is not computationally demanding, as it involves nothing more complex than a univariate root finding; cf. (3.4). Once a solution of $\mathcal{F}$ is found, the corresponding active component of the solution of $\mathcal{P}$ can be obtained by (3.5). Concerning the passive component, if $\hat{y}_i^p \neq 1$, then the passive letter is assigned the zero weight. Furthermore, the passive component belongs to $C^p\left(\frac{\nu^a}{1+\hat{y}^a_{\mathcal{F}}}\right)$.
- For a linear or polyhedral C a solution of $\mathcal{F}$ may also be obtained by a minimax (saddle point) convex optimization; cf. Sect. 3.1. For a polyhedral C it involves maximization over a positive half space and minimization over a simplex.
- A solution of $\mathcal{P}$ can also be obtained by the sequential active-passive dualization; cf. Sect. 5.
- Perhaps the most convenient way of obtaining a solution of $\mathcal{P}$ is by the PP algorithm; cf. Sect. 6. The PP algorithm can be implemented in various ways. For instance, by the Fisher scoring algorithm, or by the Gokhale algorithm. It can as well be implemented by the BD dual. The BD dual implementation of the PP algorithm is particularly convenient in the case of a linear or polyhedral C, where it may reduce dimensionality of the optimization. The dimensionality reduction is achieved also by the Fenchel dual, but unlike the Fenchel dual, the PP algorithm seamlessly leads both active and passive components of a solution. In the Fenchel dual approach, the passive components of the solution have to be determined separately.

Applications. The presented results can be used to obtain the *correct* Maximum multinomial Likelihood (MmL) estimates and inferences in the many models, where the feasible set C is convex. Marginal homogeneity models, isotonic cone models, mean response models, multinomial-Poisson homogeneous models, contingency tables with given marginals, constrained 'density' estimation, estimating equations, and many others are among such a models. The correct MmL cannot, in general, be obtained by the simplified Lagrange and Fenchel duals, which are considered in the Statistics literature for decades. Besides the flawed duality theory of MmL, the current practice of obtaining MmL estimates and inferences is equally unsatisfactory. Indeed, the replacement of zero count by an *ad hoc*, surrogate value leads a 'solution' of $\mathcal{P}$ that depends on the surrogate value. The PP algorithm avoids this inconvenience, by letting the perturbations to zero, in an appropriate way. Taken together, the presented results can be instrumental in reviving interest in the likelihood estimation and inferences in the many convex-constrained models, where they are overshadowed by the methods such as the MDI, minimum χ^2, or the least squares.

MmL vs. EL. Empirical Likelihood is the multinomial likelihood which assumes the support in a data. Owen [34, p. 238] motivates such an 'empiricism' by noting, that restricting to empirical measures is "... convenient because the statistician might not be willing to specify a bound M ..." on the support $[-M, M]$. Though EL was developed primarily with the continuous data in mind, it should lead meaningful results also in the discrete case, where the support is usually known. Once it is recognized that the MmL may exploit also the unobserved outcomes, a limitation of restricting to observed outcomes becomes visible. Namely, the EL-maximizing distribution (if it exists at all) may be different than the MmL distribution (which always exists). Consequently, the multinomial likelihood ratio may lead to different inferential and evidential conclusions than the EL ratio.

As a yet another illustration of the extent of the difference between ignoring support and taking it into account, consider the unimodal probability mass function estimation (cf. Balabdaoui and Jankowski [7]), where a sample of size 50 is obtained from the negative binomial distribution with parameters $(0.1, 1)$. Figure 2 exhibits the observed empirical probability mass function, a unimodal EL estimate, and the MmL estimate under unimodality.

Also in the continuous case MmL could be preferred to EL, if a statistician is willing to follow Fisher, and take the finite precision of a data into account.

10. Proofs

10.1. Notation and preliminaries

In this section we introduce notation and recall notions and results which will be used later; it is based mainly on Bertsekas [10] and Rockafellar [41, 40]. We will not repeat the definitions introduced in the previous part of the paper.

We assume that the extended real line $\bar{\mathbb{R}} = [-\infty, \infty]$ is equipped with the order topology; so it is a compact metrizable space homeomorphic to the unit

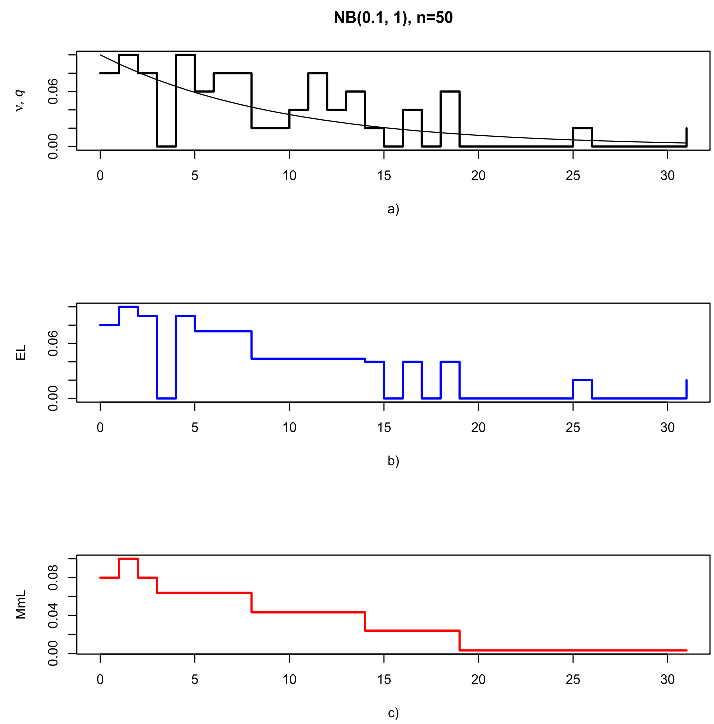

FIG 2. *Panel a) the type ν and sampling distribution q; panel b) a unimodal EL distribution; and panel c) the unimodal MmL.*

interval. The arithmetic operations on $\bar{\mathbb{R}}$ are defined in a usual way; further we put $0 \cdot (\pm\infty) \triangleq 0$. For $\alpha \le 0$ we define $\log(\alpha) \triangleq -\infty$; then $\log : \mathbb{R} \to \bar{\mathbb{R}}$ is continuous.

For $m \ge 0$ put $\mathbb{R}^m_+ \triangleq \{x \in \mathbb{R}^m : x \ge 0\}$ and $\mathbb{R}^m_- \triangleq \{x \in \mathbb{R}^m : x \le 0\}$ (recall that, for $m = 0$, $\mathbb{R}^m = \{0\}$). In the matrix operations, the members of $\mathbb{R}^m$ are considered to be column matrices. If no confusion can arise, a vector with constant values is denoted by a scalar.

Let C be a nonempty subset of $\mathbb{R}^m$. The *convex hull* of C is denoted by $\text{conv}(C)$. The *polar cone* of C is the set $C^* \triangleq \{y \in \mathbb{R}^m : \langle y, q\rangle \le 0 \text{ for every } q \in C\}$. This is a nonempty closed convex cone [10, p. 166]. Assume that C is convex. The *relative interior* $\text{ri}(C)$ of C is the interior of C relative to the affine hull $\text{aff}(C)$ of C [10, p. 40]; it is nonempty and convex [10, Prop. 1.4.1].

The *recession cone* of a convex set C is the convex cone

$$R_C \triangleq \{z \in \mathbb{R}^m : \; x + \alpha z \in C \text{ for every } x \in C, \alpha > 0\} \tag{10.1}$$

[10, p. 50]. Every $z \in R_C$ is called a *direction of recession* of C. Clearly, $0 \in R_C$; if $R_C = \{0\}$ it is said to be *trivial*. The *lineality space* L_C of C is defined by $L_C \triangleq R_C \cap (-R_C)$ [10, p. 54]; it is a linear subspace of $\mathbb{R}^m$. Note that if C is a cone then $R_C = C$ and $L_C = C \cap (-C)$.

Let X be a subset of $\mathbb{R}^m$ and $f : X \to \bar{\mathbb{R}}$ be a function. By $f'(x; y)$ we denote the *directional derivative* of f at x in the direction y [10, p. 17]. By $\nabla f(x)$ and $\nabla^2 f(x)$ we denote the *gradient* and the *Hessian* of f at x. For a nonempty set $C \subseteq X$, $\operatorname{argmin}_C f$ and $\operatorname{argmax}_C f$ denote the sets of all minimizing and maximizing points of f over C, respectively; that is,

$$\operatorname*{argmin}_C f = \left\{ \bar{x} \in C : \; f(\bar{x}) = \inf_{x \in C} f(x) \right\},$$
$$\operatorname*{argmax}_C f = \left\{ \bar{x} \in C : \; f(\bar{x}) = \sup_{x \in C} f(x) \right\}.$$

The *(effective) domain* and the *epigraph* of f are the sets [10, p. 25]

$$\operatorname{dom}(f) \triangleq \{x \in X : \; f(x) < \infty\} \quad \text{and} \quad \operatorname{epi}(f) \triangleq \{(x, w) \in X \times \mathbb{R} : f(x) \leq w\}.$$

A function $f : X \to \bar{\mathbb{R}}$ is

- *proper* if $f > -\infty$ and there is $x \in \mathbb{R}^m$ with $f(x) < \infty$ [10, p. 25] (this should not be confused with the properness associated with compactness of point preimages);
- *closed* if $\operatorname{epi}(f)$ is closed in $\mathbb{R}^{m+1}$ [10, p. 28];
- *lower semicontinuous (lsc)* if $f(x) \leq \liminf_k f(x_k)$ for every $x \in X$ and every sequence $(x_k)_k$ in X converging to x [10, p. 27]; analogously for the *upper semicontinuity (usc)*;
- *convex* if both X and $\operatorname{epi}(f)$ are convex [10, Def. 1.2.4];
- *concave* if $(-f)$ is convex.

When dealing with closedness of f, we will often use the following simple lemma [10, Prop. 1.2.2 and p. 28].

Lemma 27. *Let $f : X \to \bar{\mathbb{R}}$ be a map defined on a set $X \subseteq \mathbb{R}^m$. Define*

$$\tilde{f} : \mathbb{R}^m \to \bar{\mathbb{R}}, \qquad \tilde{f}(x) \triangleq \begin{cases} f(x) & \text{if } x \in X; \\ \infty & \text{if } x \notin X. \end{cases}$$

Then the following are equivalent:

(a) f is closed;
(b) $\tilde{f}$ is closed;
(c) $\tilde{f}$ is lower semicontinuous;
(d) the level sets $V_\gamma \triangleq \{x \in \mathbb{R}^m : \tilde{f}(x) \leq \gamma\} = \{x \in X : f(x) \leq \gamma\}$ are closed for every $\gamma \in \mathbb{R}$.

The *recession cone* R_f of a proper convex closed function $f : \mathbb{R}^m \to \bar{\mathbb{R}}$ is the recession cone of any of its nonempty level sets V_γ [10, p. 93]. The lineality space L_f of R_f is, due to the convexity of f, the set of directions y in which f is constant (that is, $f(x + \alpha y) = f(x)$ for every $x \in \operatorname{dom}(f)$ and $\alpha \in \mathbb{R}$); thus L_f is also called the *constancy space* of f [10, p. 97]. If $g : \mathbb{R}^m \to \bar{\mathbb{R}}$ is *concave*, the corresponding notions for g are defined via the convex function $(-g)$.

The fundamental results underlying the importance of recession cones, are the following theorems ([10, Props. 2.3.2 and 2.3.4] or [41, Thm. 27.3]).

Theorem 28. *Let C be a nonempty convex closed subset of $\mathbb{R}^m$ and $f:\mathbb{R}^m\to\bar{\mathbb{R}}$ be a proper convex closed function such that $C\cap\operatorname{dom}(f)\neq\emptyset$. Then the following are equivalent:*

(a) the set $\operatorname{argmin}_C f$ *of minimizing points of f over C is nonempty and compact;*
(b) C and f have no common nonzero direction of recession, that is,

$$R_C\cap R_f=\{0\}.$$

Both conditions are satisfied, in particular, if $C\cap\operatorname{dom}(f)$ is bounded.

Theorem 29. *Let C be a nonempty convex closed subset of $\mathbb{R}^m$ and $f:\mathbb{R}^m\to\bar{\mathbb{R}}$ be a convex closed function such that $C\cap\operatorname{dom}(f)\neq\emptyset$. If*

$$R_C\cap R_f=L_C\cap L_f, \tag{10.2}$$

or if

$$C \text{ is polyhedral} \qquad \text{and} \qquad R_C\cap R_f\subseteq L_f,$$

then the set $\operatorname{argmin}_C f$ *of minimizing points of f over C is nonempty. Under condition (10.2),* $\operatorname{argmin}_C f$ *can be written as $\tilde{C}+(L_C\cap L_f)$, where $\tilde{C}$ is compact.*

Standing Assumption. If not stated otherwise, in the sequel it is assumed that a nonempty convex closed subset C of $\Delta_{\mathcal{X}}$ having support $\mathcal{X}$, and a type $\nu\in\Delta_{\mathcal{X}}$ are given.

Since no confusion can arise, by ℓ we denote also an extension of the original function ℓ (defined in (2.1)) to $\mathbb{R}^m$:

$$\ell:\mathbb{R}^m\to\bar{\mathbb{R}},\qquad \ell(x)\triangleq-\langle\nu,\log(x)\rangle=-\sum_{i\in\mathcal{X}^a}\nu_i^a\log(x_i^a); \tag{10.3}$$

the conventions $0\cdot(-\infty)=0$ and $\log\beta=-\infty$ for every $\beta\le 0$ apply.

10.2. Proof of Theorem 2 (Primal problem)

In the next lemma we prove that ℓ is a proper convex closed function. Since C is compact, C has no nonzero direction of recession. From this and Theorem 28, Theorem 2 will follow.

Lemma 30. *If $\nu\in\Delta_{\mathcal{X}}$ and ℓ is given by (10.3), then*

(a) $\ell(x)>-\infty$ for every $x\in\mathbb{R}^m$, and $\ell(x)<\infty$ if and only if $x^a>0$; so

$$\operatorname{dom}(\ell)=\{x\in\mathbb{R}^m:\ x^a>0\};$$

(b) ℓ is a proper continuous (hence closed) convex function;
(c) the restriction of ℓ to its domain $\operatorname{dom}(\ell)$ *is strictly convex if and only if $\nu>0$;*

(d) *ℓ is differentiable on* $\mathrm{dom}(\ell)$ *with the gradient given by*

$$\nabla \ell(x) = -(\nu^a/x^a, 0^p);$$

the one-sided directional derivative of ℓ at $x \in \mathrm{dom}(\ell)$ in the direction y is

$$\ell'(x; y) = \langle \nabla \ell(x), y \rangle = -\sum_{i \in \mathcal{X}^a} \frac{\nu_i^a y_i^a}{x_i^a};$$

(e) *the recession cone R_ℓ and the constancy space L_ℓ of ℓ are*

$$R_\ell = \{z \in \mathbb{R}^m : \ z^a \geq 0\} \quad \textit{and} \quad L_\ell = \{z \in \mathbb{R}^m : \ z^a = 0\}.$$

Proof. The properties (a) and (d) are trivial and the property (b) follows from the continuity and concavity of the logarithm (extended to the whole real line); closedness of ℓ follows from continuity by Lemma 27. For the property (c), use that the Hessian $\nabla^2 \ell(x) = (\partial^2 \ell(x)/\partial x_i \partial x_j)_{ij}$ is a diagonal matrix $\mathrm{diag}(\nu^a/(x^a)^2, 0^p)$. Hence it is positive definite and ℓ is strictly convex if and only if $\nu > 0$ [10, Prop. 1.2.6].

It remains to prove (e). Fix any $\gamma \in \mathbb{R}$ such that the level set $V = \{x : \ \ell(x) \leq \gamma\}$ is nonempty. If z is such that $z^a \ngeq 0$ then there is an active letter i with $z_i < 0$. In such a case, for any $x \in V$ there is $\alpha > 0$ with $y_i + \alpha z_i \leq 0$ and so $y + \alpha z \notin \mathrm{dom}(\ell) \supseteq V$. Hence, by (10.1), $z \notin R_V = R_\ell$.

Now take any z with $z^a \geq 0$. Then, for every $x \in V$ and $\alpha > 0$, $\ell(x + \alpha z) \leq \ell(x)$ by the monotonicity of the logarithm. That is, $x + \alpha z \in V$ and so $z \in R_V = R_\ell$. The property (e) is proved. □

Proposition 31. *If the support of C is $\mathcal{X}$, then the primal $\mathcal{P}$ is finite and its solution set $S_\mathcal{P}$ is nonempty and compact. Moreover, if $\nu > 0$ then $S_\mathcal{P}$ is a singleton.*

Proof. Since C is compact, its recession cone is trivial. Thus the first assertion follows from Theorem 28. The fact that $S_\mathcal{P}$ is a singleton provided $\nu > 0$, follows from the strict convexness of f. □

Proposition 32. *Let $\nu \in \Delta_\mathcal{X}$ and C be a convex closed set having support $\mathcal{X}$. Then the π^a-projection of $S_\mathcal{P}$ onto active letters is always a singleton $\{\hat{q}^a_\mathcal{P}\}$ and*

$$S_\mathcal{P} = \{q \in C : \ q^a = \hat{q}^a_\mathcal{P}\} = \{\hat{q}^a_\mathcal{P}\} \times C^p(\hat{q}^a_\mathcal{P}).$$

Consequently, $S_\mathcal{P}$ is a singleton if and only if $C^p(\hat{q}^a_\mathcal{P})$ is a singleton.

Proof. Note that $C^a = \pi^a(C)$ is a nonempty convex closed subset of $\mathbb{R}^{m_a}$. Define $\ell^a : \mathbb{R}^{m_a} \to \bar{\mathbb{R}}$ by $\ell^a(x^a) \triangleq -\langle \nu^a, \log x^a \rangle$ for $x^a \in C^a$ and $\ell^a(x^a) \triangleq \infty$ otherwise. Since

$$\ell(q) = \ell^a(q^a) \qquad \text{for every} \quad q \in C, \tag{10.4}$$

it holds that

$$\inf_{x^a \in C^a} \ell^a(x^a) = \hat{\ell}_\mathcal{P}.$$

The map ℓ^a is proper, convex and closed (use Lemma 27 and the fact that the restriction of ℓ^a to the closed set C^a is continuous, hence closed). Since $\mathrm{dom}(\ell^a) \subseteq C^a$ is bounded, Theorem 28 gives that $\operatorname{argmin} \ell^a$ is a nonempty compact set. This set is a subset of $\mathrm{dom}(\ell^a)$ and the restriction of ℓ^a to $\mathrm{dom}(\ell^a)$ is strictly convex (Lemma 30(c)), so $\operatorname{argmin} \ell^a$ is a singleton. Hence there is a unique point $\hat{q}^a_{\mathcal{P}} \in C^a$ such that $\ell^a(\hat{q}^a_{\mathcal{P}}) = \hat{\ell}_{\mathcal{P}}$. Now, (10.4) gives that $q \in S_{\mathcal{P}}$ if and only if $q^a = \hat{q}^a_{\mathcal{P}}$; so $S_{\mathcal{P}} = \{\hat{q}^a_{\mathcal{P}}\} \times C^p(\hat{q}^a_{\mathcal{P}})$. □

Theorem 2 immediately follows from Propositions 31 and 32.

10.3. Proof of Theorem 5 (Convex conjugate by Lagrange duality)

In this section we prove Theorem 5 on the convex conjugate ℓ^*, defined by the convex conjugate primal problem (cc-primal, for short)

$$\ell^*(z) = \sup_{q \in \Delta_{\mathcal{X}}} (\langle q, z\rangle - \ell(q)).$$

The proof is based on the following reformulation of the cc-primal

$$\ell^*(z) = \sup_{x \geq 0} \inf_{\mu \in \mathbb{R}} K_z(x, \mu) = \inf_{\mu \in \mathbb{R}} \sup_{x \geq 0} K_z(x, \mu),$$

where

$$K_z(x, \mu) = \langle x, z\rangle + \langle \nu, \log x\rangle - \mu \left(\sum x - 1\right)$$

is the Lagrangian function; cf. Lemma 34. Then we will show that the map $\mu \mapsto \sup_{x \geq 0} K_z(x, \mu)$ is minimized at $\hat{\mu}(z) = \max\{\bar{\mu}(z^a), \max(z^p)\}$; cf. Section 10.3.2. Structure of the solution set $S_{cc}(z)$ of the cc-primal is described in Section 10.3.3. Additional properties of the convex conjugate, which will be utilized in the proof of Theorem 6, are stated in Section 10.3.4.

For every $z^a \in \mathbb{R}^{m_a}$ and $\mu > \max(z^a)$ put

$$\xi(\mu) = \xi_{z^a}(\mu) \triangleq \sum \frac{\nu^a}{\mu - z^a} \tag{10.5}$$

and recall that $\hat{\mu}(z) \triangleq \max\{\bar{\mu}(z^a), \max(z^p)\}$, where $\bar{\mu}(z^a) > \max(z^a)$ solves $\xi(\mu) = 1$. Since

$$\xi \text{ is strictly decreasing,} \qquad \lim_{\mu \searrow \max(z^a)} \xi(\mu) = \infty \quad \text{and} \quad \lim_{\mu \to \infty} \xi(\mu) = 0, \tag{10.6}$$

$\bar{\mu}(z^a)$ is well-defined. We start with a simple lemma.

Lemma 33. *For every $z \in \mathbb{R}^m$ and $c \in \mathbb{R}$,*

$$\bar{\mu}(z^a + c) = \bar{\mu}(z^a) + c, \quad \hat{\mu}(z + c) = \hat{\mu}(z) + c, \quad \textit{and} \quad \ell^*(z + c) = \ell^*(z) + c.$$

Proof. The first two equalities follow from the facts that $\xi_{z^a+c}(\cdot + c) = \xi_{z^a}(\cdot)$ and that $\max(z^p + c) = \max(z^p) + c$. The final one is a trivial consequence of the definition of ℓ^*; indeed, since $\langle q, c\rangle = c$ for every $q \in \Delta_{\mathcal{X}}$, $\ell^*(z + c) = \sup_q(\langle q, (z + c)\rangle - \ell(q)) = \ell^*(z) + c$. □

10.3.1. Lagrange duality for the convex conjugate

Assume that $\nu \in \Delta_{\mathcal{X}}$ and $z \in \mathbb{R}^m$ are given. For $x \in \mathbb{R}^m_+$ put

$$h_z(x) \triangleq \langle x, z\rangle - \ell(x) = \langle x, z\rangle + \langle \nu^a, \log x^a\rangle$$

and define extended-real-valued functions

$$\begin{aligned} K_z : \mathbb{R}^m \times \mathbb{R} \to \bar{\mathbb{R}}, \quad & K_z(x,\mu) && \triangleq \begin{cases} h_z(x) - \mu\left(\sum x - 1\right) & \text{if } x \in \mathbb{R}^m_+, \mu \in \mathbb{R}, \\ \infty & \text{otherwise;} \end{cases} \\ k_z : \mathbb{R} \to \bar{\mathbb{R}}, \quad & k_z(\mu) && \triangleq \sup_{x \in \mathbb{R}^m_+} K_z(x,\mu). \end{aligned}$$

Lemma 34 (Lagrange duality for the convex conjugate)**.** *For every $\nu \in \Delta_{\mathcal{X}}$ and $z \in \mathbb{R}^m$,*

$$\ell^*(z) = \inf_{\mu\in\mathbb{R}} k_z(\mu) = \inf_{\mu\in\mathbb{R}} \sup_{x\geq 0} K_z(x,\mu).$$

Proof. We follow [40, Sect. 4]. Denote by $\mathbb{R}^m_\oplus$ the subset $\{x \in \mathbb{R}^m : x^a > 0, x^p \geq 0\}$ of $\mathbb{R}^m_+$. Define $F_z : \mathbb{R}^m \times \mathbb{R}^2 \to \bar{\mathbb{R}}$ and $f_z : \mathbb{R}^m \to \bar{\mathbb{R}}$ by

$$\begin{aligned} F_z(x,u) &\triangleq \begin{cases} -h_z(x) & \text{if } x \in \mathbb{R}^m_\oplus, 1-u_2 \leq \sum x \leq 1+u_1; \\ \infty & \text{otherwise;} \end{cases} \\ f_z(x) &\triangleq F_z(x,0). \end{aligned}$$

Note that, in the definition of F_z, $x \in \mathbb{R}^m_\oplus$ can be replaced by $x \in \mathbb{R}^m_+$; indeed, if $x \in \mathbb{R}^m_+ \setminus \mathbb{R}^m_\oplus$ then $x^a_i = 0$ for some $i \in \mathcal{X}^a$ and hence $-h_z(x) = \infty$. The set $D \triangleq \{(x,u) \in \mathbb{R}^m \times \mathbb{R}^2 : x \geq 0, 1-u_2 \leq \sum x \leq 1+u_1\}$ is closed convex (in fact, polyhedral) and the map $\tilde{F}_z : D \to \bar{\mathbb{R}}$, $(x,u) \mapsto -h_z(x)$ is convex and continuous, hence closed. Since the epigraphs of $\tilde{F}_z$ and F_z coincide,

$$F_z \quad \text{is convex and closed jointly in } x \text{ and } u. \tag{10.7}$$

The corresponding optimal value function $\varphi_z : \mathbb{R}^2 \to \bar{\mathbb{R}}$ is defined by (cf. [40, (4.7)])

$$\varphi_z(u) \triangleq \inf_{x\in\mathbb{R}^m} F_z(x,u).$$

We are going to show that

$$\varphi_z(0) = \liminf_{u\to 0} \varphi_z(u). \tag{10.8}$$

To this end, take any $\varepsilon \in (0,1)$ and $u \in \mathbb{R}^2$ with $|u| < 1$. Assume that $u_1+u_2 \geq 0$. Then

$$\varphi_z(u) = \inf_{\substack{x\geq 0 \\ \sum x \in [1-u_2, 1+u_1]}} (-h_z(x)) = \inf_{\theta\in[1-u_2,1+u_1]} \inf_{x\in\Delta_{\mathcal{X}}} (-h_z(\theta x)).$$

For any $\theta > 0$,

$$\sup_{x\in\Delta_{\mathcal{X}}} |h_z(\theta x) - h_z(x)| \le \sup_{x\in\Delta_{\mathcal{X}}} (|\log\theta| + |1-\theta|\cdot|\langle x, z\rangle|)$$
$$= |\log\theta| + |1-\theta|\cdot\max|z| \triangleq \psi_z(\theta).$$

Since ψ_z is continuous at $\theta = 1$ and $\psi_z(1) = 0$, there is $\delta > 0$ such that $\psi_z(\theta) < \varepsilon$ for every $\theta \in [1-\delta, 1+\delta]$. Thus $|\varphi_z(u) - \varphi_z(0)| < \varepsilon$ whenever $|u| < \delta$ and $u_1 + u_2 \ge 0$. This gives

$$\varphi_z(0) = \lim_{\substack{u\to 0\\ u_1+u_2\ge 0}} \varphi_z(u).$$

Since $\varphi_z(u) = \infty$ if $u_1 + u_2 < 0$ (indeed, for such u, $F_z(x,u) = \infty$ for every x), (10.8) is proved.

The Lagrangian function $L_z : \mathbb{R}^m \times \mathbb{R}^2 \to \bar{\mathbb{R}}$ associated with F_z is defined by (cf. [40, (4.2)])

$$L_z(x,y) \triangleq \inf_{u\in\mathbb{R}^2} (F_z(x,u) + \langle u, y\rangle).$$

A simple computation yields (cf. [40, (4.4)] with $f_0(x) = h_z(x)$, $f_1(x) = \sum x - 1$, and $f_2(x) = 1 - \sum x$, all restricted to $C = \mathbb{R}^m_\oplus$)

$$L_z(x,y) = \begin{cases} -h_z(x) + (y_1 - y_2)\left(\sum x - 1\right) & \text{if } x \in \mathbb{R}^m_\oplus, y \in \mathbb{R}^2_+; \\ -\infty & \text{if } x \in \mathbb{R}^m_\oplus, y \notin \mathbb{R}^2_+; \\ \infty & \text{if } x \notin \mathbb{R}^m_\oplus. \end{cases}$$

(Indeed, fix any $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^2$. If $x \notin \mathbb{R}^m_\oplus$ then $F_z(x,u) = \infty$ for every u, hence $L_z(x,y) = \infty$. If $x \in \mathbb{R}^m_\oplus$ and $y \not\ge 0$, then $F_z(x,u) = -h_z(x)$ whenever both $u_1, u_2 > 0$ are sufficiently large; for such u, $\langle u, y\rangle$ is not bounded from below (use that $y_1 < 0$ or $y_2 < 0$), hence $L_z(x,y) = -\infty$. Finally, assume that $x \in \mathbb{R}^m_\oplus$ and $y \ge 0$. If $u_1 + u_2 < 0$ then $F_z(x,u) = \infty$ by the definition of F_z. If $u_1 + u_2 \ge 0$ then $\langle u, y\rangle \ge y_1(\sum x - 1) + y_2(1 - \sum x)$, with the equality if $u_1 = \sum x - 1$, $u_2 = 1 - \sum x$. Thus $L_z(x,y) = -h_z(x) + (y_1 - y_2)\left(\sum x - 1\right)$.)

If we define (cf. [40, (4.6)])

$$g_z : \mathbb{R}^2 \to \bar{\mathbb{R}}, \qquad g_z(y) \triangleq \inf_{x\in\mathbb{R}^m} L_z(x,y),$$

then [40, Thm. 7], (10.7), and (10.8) imply

$$\inf_{x\in\mathbb{R}^m} f_z(x) = \varphi_z(0) = \liminf_{u\to 0} \varphi_z(u) = \sup_{y\in\mathbb{R}^2} g_z(y).$$

By the definition of f_z, $\inf_x f_z(x) = -\ell^*(z)$; thus to finish the proof of the lemma it suffices to show that

$$\inf_{\mu\in\mathbb{R}} k_z(\mu) = -\sup_{y\in\mathbb{R}^2} g_z(y). \tag{10.9}$$

If $y \notin \mathbb{R}^2_+$ then $g_z(y) = -\infty$; to see this, take arbitrary $x \in \mathbb{R}^m_\oplus$ and realize that $g_z(y) \leq L_z(x, y) = -\infty$. Fix any $y \in \mathbb{R}^2_+$. Then, for every $x \in \mathbb{R}^m_\oplus$, $K_z(x, y_1 - y_2) = -L_z(x, y)$; hence

$$k_z(y_1 - y_2) = -g_z(y).$$

Since $\{y_1 - y_2 : y \in \mathbb{R}^2_+\} = \mathbb{R}$, we have

$$\inf_{\mu \in \mathbb{R}} k_z(\mu) = \inf_{y \in \mathbb{R}^2_+} (-g_z(y)) = -\sup_{y \in \mathbb{R}^2} g_z(y).$$

Thus (10.9) is established and the proof of the lemma is finished. □

10.3.2. Proof of Theorem 5

With Lemma 34, Theorem 5 can be proved. Recall that the theorem states that, for every $z \in \mathbb{R}^m$,

$$\ell^*(z) = -1 + \hat{\mu}(z) + I_{\hat{\mu}(z)}(\nu^a \,\|\, -z^a), \qquad \text{where} \quad \hat{\mu}(z) \triangleq \max\{\bar{\mu}(z^a), \max(z^p)\}.$$

Proof of Theorem 5. Keep the notation from Section 10.3.1. By Lemma 34, $\ell^*(z) = \inf_{\mu \in \mathbb{R}} k_z(\mu)$, and, using partial maximization,

$$k_z(\mu) = \sup_{x^a \geq 0} \tilde{k}_z(x^a, \mu), \quad \text{where} \quad \tilde{k}_z(x^a, \mu) \triangleq \sup_{x^p \geq 0} K_z((x^a, x^p), \mu).$$

Note that $K_z(x, \mu) = c + \sum_{i \in \mathcal{X}^p} x^p_i (z^p_i - \mu)$, where c does not depend on x^p. Thus

$$\tilde{k}_z(x^a, \mu) = \begin{cases} \langle x^a, z^a \rangle + \langle \nu^a, \log x^a \rangle - \mu \left(\sum x^a - 1\right) & \text{if } \mu \geq \max(z^p); \\ \infty & \text{otherwise.} \end{cases} \tag{10.10}$$

The second case immediately gives

$$k_z(\mu) = \infty \qquad \text{if} \quad \mu < \max(z^p). \tag{10.11}$$

If $\mu \geq \max(z^p)$ and $z^a_i - \mu \geq 0$ for some $i \in \mathcal{X}^a$, then $k_z(\mu) = \infty$ (use (10.10) and the fact that, for any $a \geq 0$ and $b > 0$, the map $f(x) \triangleq ax + b \log x$ is strictly increasing and $\lim_{x \to \infty} f(x) = \infty$). That is

$$k_z(\mu) = \infty \qquad \text{if} \quad \mu \leq \max(z^a). \tag{10.12}$$

Assume now that $\mu \geq \max(z^p)$ and $\mu > \max(z^a)$. Note that

$$k_z(\mu) = \max_{x^a > 0} \tilde{k}_z(x^a, \mu).$$

By (10.10), $\tilde{k}_z(\cdot, \mu)$ is differentiable and strictly concave on the open set $\{x^a \in \mathbb{R}^{m_a} : x^a > 0\}$ (use that the Hessian, which is equal to $\operatorname{diag}(-\nu^a_i/(x^a_i)^2)_{i=1}^{m_a}$, is negative definite). Thus the basic necessary and sufficient condition for unconstrained optimization [10, p. 258] gives that

$$k_z(\mu) = \tilde{k}_z(\hat{x}^a, \mu),$$

where $\hat{x}^a = \hat{x}^a_\mu > 0$ is the unique solution of $\nabla_{x^a} \tilde{k}_z(\hat{x}^a, \mu) = 0$; that is,

$$z_i^a + \frac{\nu_i^a}{\hat{x}_i^a} - \mu = 0 \qquad \text{for every} \quad i \in \mathcal{X}^a.$$

The above equation immediately yields

$$\hat{x}^a = \frac{\nu^a}{\mu - z^a}$$

and

$$k_z(\mu) = -1 + \mu + I_\mu(\nu^a \,\|\, -z^a) \qquad \text{if} \quad \mu \geq \max(z^p), \mu > \max(z^a). \tag{10.13}$$

Put $J \triangleq \{\mu \in \mathbb{R} : \mu > \max(z^a), \mu \geq \max(z^p)\}$ and $\underline{\mu} \triangleq \max(z)$; then either $J = (\underline{\mu}, \infty)$ or $J = [\underline{\mu}, \infty)$. Equations (10.11), (10.12), and (10.13) yield

$$k_z(\mu) = \begin{cases} -1 + \mu + I_\mu(\nu^a \,\|\, -z^a) & \text{if } \mu \in J; \\ \infty & \text{otherwise.} \end{cases}$$

For $\mu > \underline{\mu}$ we have

$$k_z'(\mu) = 1 - \xi(\mu).$$

Since ξ is decreasing on J and $\xi(\bar{\mu}(z^a)) = 1$, $k_z(\mu)$ is decreasing on $[\underline{\mu}, \bar{\mu}(z^a)]$ and increasing on $[\bar{\mu}(z^a), \infty)$, provided $\underline{\mu} \leq \bar{\mu}(z^a)$; otherwise $k_z(\mu)$ is increasing on $[\underline{\mu}, \infty]$. Thus,

$$\ell^*(z) = \inf_{\mu \in \mathbb{R}} k_z(\mu) = \begin{cases} k_z(\bar{\mu}(z^a)) & \text{if } \bar{\mu}(z^a) \geq \underline{\mu}; \\ k_z(\underline{\mu}) & \text{otherwise.} \end{cases}$$

So $\operatorname{argmin} k_z = \max\{\bar{\mu}(z^a), \max(z^p)\} = \hat{\mu}(z)$ and Theorem 5 is proved. □

Corollary 35. *If $\nu > 0$ then, for every $z \in \mathbb{R}^m$,*

$$\begin{aligned} \ell^*(z) &= -1 + \bar{\mu}(z) + I_{\bar{\mu}(z)}(\nu \,\|\, -z) \\ &= -1 + \bar{\mu}(z) + \langle \nu, \log \nu \rangle - \langle \nu, \log(\bar{\mu}(z) - z) \rangle \,. \end{aligned}$$

Corollary 35 was first proved by El Barmi and Dykstra [12, Lemma 2.1].

10.3.3. The structure of the solution set $S_{cc}(z)$

The structure of the cc-primal solution set $S_{cc}(z)$, defined by (3.2), is described. First, recall the definition (10.5) of ξ.

Proposition 36. *For every $z \in \mathbb{R}^m$, $S_{cc}(z)$ is a nonempty compact set and*

$$S_{cc}(z) = \{\hat{q}^a_{cc}(z)\} \times \{q^p \geq 0 : \sum q^p = 1 - \xi(\hat{\mu}(z)),$$
$$q^p_i = 0 \text{ whenever } z^p_i \neq \hat{\mu}(z)\},$$

where

$$\hat{q}^a_{cc}(z) \triangleq \frac{\nu^a}{\hat{\mu}(z) - z^a}.$$

In particular, if $\bar{\mu}(z^a) \geq \max(z^p)$ then $S_{cc}(z) = \{(\hat{q}^a_{cc}(z), 0^p)\}$ is a singleton.

Proof. Fix $z \in \mathbb{R}^m$ and put $\mu \triangleq \hat{\mu}(z)$, $\hat{q}^a_{cc} \triangleq \hat{q}^a_{cc}(z)$; then, by Theorem 5, $\ell^*(z) = -1 + \mu + I_\mu(\nu^a \,\|\, - z^a)$. The fact that $S_{cc}(z)$ is nonempty and compact follows from Theorem 28.

Take any $\bar{q} \in S_{cc}(z)$. Then $\bar{q}^a > 0$ and, for every $q \in \Delta_{\mathcal{X}}$,

$$0 \geq f'_z(\bar{q}; q - \bar{q}) = \langle z, (q - \bar{q})\rangle + \sum \frac{\nu^a q^a}{\bar{q}^a} - 1.$$

If also $q \in S_{cc}(z)$ then, analogously,

$$0 \geq \langle z, (\bar{q} - q)\rangle + \sum \frac{\nu^a \bar{q}^a}{q^a} - 1.$$

Thus, by combining these two inequalities,

$$\sum \nu^a \left(\frac{q^a}{\bar{q}^a} + \frac{\bar{q}^a}{q^a}\right) \leq 2.$$

For positive x and y, $(x/y + y/x) \geq 2$, with equality if and only if $x = y$. Thus, $q^a = \bar{q}^a$ for every $\bar{q}, q \in S_{cc}(z)$; hence,

$$\pi^a(S^a_{cc}(z)) \quad \text{is a singleton.} \tag{10.14}$$

Distinguish two cases. First assume that $\mu = \bar{\mu}(z^a) \geq \max(z^p)$ and put $\bar{q} \triangleq (\hat{q}^a_{cc}, 0^p)$. Then $\bar{q} \in \Delta_{\mathcal{X}}$ (use that $\sum \bar{q} = \xi(\mu) = 1$) and

$$\begin{aligned} f_z(\bar{q}) &= \langle \hat{q}^a_{cc}, z^a\rangle + \langle \nu^a, \log(\hat{q}^a_{cc})\rangle = \sum \frac{(z^a - \mu + \mu)\nu^a}{\mu - z^a} + I_\mu(\nu^a \,\|\, - z^a) \\ &= -1 + \mu + I_\mu(\nu^a \,\|\, - z^a) = \ell^*(z). \end{aligned}$$

Hence $\bar{q} \in S_{cc}(z)$ and, since $\sum \bar{q}^a = 1$, $S_{cc}(z) = \{\bar{q}\}$ by (10.14).

Assume now that $\mu = \max(z^p) > \bar{\mu}(z^a)$; then $\gamma \triangleq \xi(\mu) < \xi(\bar{\mu}(z^a)) = 1$. Take any $\bar{q} \in \Delta_{\mathcal{X}}$ with $\bar{q}^a = \hat{q}^a_{cc}$ and note that $\sum \bar{q}^p = 1 - \gamma$. An argument analogous to the first case gives

$$f_z(\bar{q}) = \langle \bar{q}^p, z^p\rangle + \langle \hat{q}^a, z^a\rangle + \langle \nu^a, \log(\hat{q}^a_{cc})\rangle = \langle \bar{q}^p, z^p\rangle - 1 + \mu\gamma + I_\mu(\nu^a \,\|\, - z^a).$$

Hence $\bar{q} \in S_{cc}(z)$ if and only if $\langle \bar{q}^p, z^p\rangle = \mu(1 - \gamma)$. Since $\sum \bar{q}^p = 1 - \gamma$, the last condition is equivalent to the fact that $\bar{q}^p_i = 0$ for every $i \in \mathcal{X}^p$ with $z^p_i < \mu$. In view of (10.14) this proves the proposition. $\square$

10.3.4. Additional properties of the convex conjugate

Some additional properties of the convex conjugate ℓ^*, concerning its monotonicity and differentiability, are summarized in the next lemmas.

Lemma 37. *The following is true for every* $z, \tilde{z} \in \mathbb{R}^m$*:*

(a) If $z^a \geq \tilde{z}^a$ *then* $\bar{\mu}(z^a) \geq \bar{\mu}(\tilde{z}^a)$*, with the equality if and only if* $z^a = \tilde{z}^a$*.*
(b) If $z \geq \tilde{z}$ *then* $\ell^*(z) \geq \ell^*(\tilde{z})$ *(that is,* ℓ^* *is nondecreasing), with the equality if and only if* $z^a = \tilde{z}^a$ *and* $\hat{\mu}(z) = \hat{\mu}(\tilde{z})$*.*

Proof. (a) If $z^a \geq \tilde{z}^a$ then $\xi_{z^a}(\mu) \geq \xi_{\tilde{z}^a}(\mu)$ for every $\mu > \max(z^a)$; moreover, $\xi_{z^a}(\mu) = \xi_{\tilde{z}^a}(\mu)$ for some $\mu > \max(z^a)$ if and only if $z^a = \tilde{z}^a$. Using (10.6), (a) follows.

(b) Assume that $z \geq \tilde{z}$. Then $\langle z, q\rangle \geq \langle \tilde{z}, q\rangle$ for every $q \in \Delta_{\mathcal{X}}$, hence $\ell^*(z) \geq \ell^*(\tilde{z})$. If $z^a = \tilde{z}^a$ and $\hat{\mu}(z) = \hat{\mu}(\tilde{z})$ then $\ell^*(z) = \ell^*(\tilde{z})$ by Theorem 5. It suffices to prove that if $z^a \neq \tilde{z}^a$ or $\hat{\mu}(z) \neq \hat{\mu}(\tilde{z})$ then $\ell^*(z) > \ell^*(\tilde{z})$.

To this end, fix some $\tilde{q} \in S_{cc}(\tilde{z})$. If $z^a \neq \tilde{z}^a$ then $\ell^*(z) \geq \langle \tilde{q}, z\rangle - \ell(\tilde{q}) \geq \langle \tilde{q}^a, (z^a - \tilde{z}^a)\rangle + \ell^*(\tilde{z})$. Since $\tilde{q}^a > 0$ and $(z^a - \tilde{z}^a) \geq 0$ is not zero, we have $\ell^*(z) > \ell^*(\tilde{z})$. Finally, assume that $z^a = \tilde{z}^a$ and $\hat{\mu}(z) > \hat{\mu}(\tilde{z})$. The map

$$g : [\bar{\mu}(z^a), \infty) \to \mathbb{R}, \qquad g(\mu) \triangleq \mu + I_\mu(\nu^a \,\|\, -z^a)$$

is strictly increasing (indeed, $g'(\mu) = 1 - \xi(\mu) > 0$ for $\mu > \bar{\mu}(z^a)$ by (10.6)). Thus, by Theorem 5 and the fact that $\hat{\mu}(\tilde{z}) \geq \bar{\mu}(\tilde{z}^a) = \bar{\mu}(z^a)$, it follows that $\ell^*(z) - \ell^*(\tilde{z}) = g(\hat{\mu}(z)) - g(\hat{\mu}(\tilde{z})) > 0$. ◻

Since the convex conjugate ℓ^* is finite-valued and convex, by [41, Thm. 10.4] it is locally Lipschitz. The following lemma claims that ℓ^* is even globally Lipschitz with the Lipschitz constant equal to 1. (Here we assume that $\mathbb{R}^m$ is equipped with the sup-norm $\|x\|_\infty = \max|x_i|$.)

Lemma 38. *The convex conjugate* $\ell^* : \mathbb{R}^m \to \mathbb{R}$ *is a (finite-valued) convex function which is Lipschitz with* $\mathrm{Lip}(\ell^*) = 1$*.*

Proof. The fact that ℓ^* is always finite is obvious. To prove that ℓ^* is Lipschitz-1, fix any $z, z' \in \mathbb{R}^m$. Then, for any $q \in S_{cc}(z)$,

$$\begin{aligned} \ell^*(z) - \ell^*(z') &\leq (\langle z, q\rangle - \ell(q)) - (\langle z', q\rangle - \ell(q)) = \langle (z - z'), q\rangle \\ &\leq \|z - z'\|_\infty. \end{aligned}$$

Analogously $\ell^*(z') - \ell^*(z) \leq \|z - z'\|_\infty$. Thus $\mathrm{Lip}(\ell^*) \leq 1$. Since $\mathrm{Lip}(\ell^*) \geq 1$ by Lemma 33 (use that $\|(c, c, \ldots, c)\|_\infty = |c|$ for $c \in \mathbb{R}$), we have $\mathrm{Lip}(\ell^*) = 1$. ◻

Lemma 39. *The map* $\bar{\mu} : \mathbb{R}^{m_a} \to \mathbb{R}$ *is differentiable (even* C^∞*) and*

$$\nabla \bar{\mu}(z^a) = \frac{1}{\sum_i \alpha_i}(\alpha_1, \ldots, \alpha_{m_a}), \qquad \alpha_k \triangleq \frac{\nu_k}{(\bar{\mu}(z^a) - z^a_k)^2} \quad (1 \leq k \leq m_a).$$

The map $\hat{\mu} : \mathbb{R}^m \to \mathbb{R}$ *is continuous.*

Proof. Since $\hat{\mu}(z) = \max(\bar{\mu}(z^a), \max(z^p))$, the continuity of $\hat{\mu}$ follows from the continuity of $\bar{\mu}$; thus it suffices to prove the first part of the lemma.

To this end, we may assume that $\mathcal{X}^p = \emptyset$, that is, $\nu > 0$ and $m_a = m$. Put $\Omega \triangleq \{(z, \mu) : z \in \mathbb{R}^m, \mu > \max(z)\}$ and define $F : \Omega \to (0, \infty)$ by

$$F(z, \mu) \triangleq \sum_{i=1}^{m} \frac{\nu_i}{\mu - z_i} .$$

Note that, for every z, $\bar{\mu}(z)$ is the unique solution of $F(z, \mu) = 1$ such that $(z, \mu) \in \Omega$. The set Ω is open and connected and F is C^∞ on Ω. Moreover,

$$\nabla_z F(z, \mu) = \frac{\nu}{(\mu - z)^2} \qquad \text{and} \qquad \nabla_\mu F(z, \mu) = -\sum_{i=1}^{m} \frac{\nu_i}{(\mu - z_i)^2} \neq 0.$$

By the local implicit function theorem [10, Prop. 1.1.14], $\bar{\mu}$ is C^∞ on $\mathbb{R}^m$ and $\nabla\bar{\mu}(z)$ is given by $-\nabla_z F(z, \bar{\mu}(z)) \cdot [\nabla_\mu F(z, \bar{\mu}(z))]^{-1}$. From this the lemma follows. □

Lemma 40. *For every $z, v \in \mathbb{R}^m$, the subgradient and the directional derivative of ℓ^* are given by*

$$\partial\ell^*(z) = S_{cc}(z) \qquad \text{and} \qquad (\ell^*)'(z; v) = \max_{q \in S_{cc}(z)} \langle q, v\rangle .$$

In particular, if $\bar{\mu}(z^a) \geq \max(z^p)$ then ℓ^* is differentiable at z and

$$\nabla\ell^*(z) = (\hat{q}^a_{cc}(z), 0^p).$$

Proof. This is a consequence of Danskin's theorem [10, Prop. 4.5.1]. To see this, fix $\bar{z} \in \mathbb{R}^m$ and take $\varepsilon > 0$ such that $\hat{q}^a_{cc}(\bar{z}) > \varepsilon$. Put $\Delta^\varepsilon_{\mathcal{X}} \triangleq \{q \in \Delta_{\mathcal{X}} : q^a \geq \varepsilon\}$ and define

$$\begin{aligned} &\varphi : \mathbb{R}^m \times \Delta^\varepsilon_{\mathcal{X}} \to \mathbb{R}, && \varphi(z, q) \triangleq \langle z, q\rangle - \ell(q), \\ &f : \mathbb{R}^m \to \mathbb{R}, && f(z) \quad \triangleq \max_{q \in \Delta^\varepsilon_{\mathcal{X}}} \varphi(z, q). \end{aligned}$$

The map φ is continuous and the partial functions $\varphi(\cdot, q)$ are (trivially) convex and differentiable for every $q \in \Delta^\varepsilon_{\mathcal{X}}$ with (continuous) $\nabla_z\varphi(z, q) = q$. Thus, by Danskin's theorem,

$$f'(\bar{z}; v) = \max_{q \in S_{cc}(\bar{z})} \langle q, v\rangle \qquad (v \in \mathbb{R}^m)$$

and

$$\partial f(\bar{z}) = \operatorname{conv}\{\nabla_z\varphi(\bar{z}, q) : q \in S_{cc}(\bar{z})\} = S_{cc}(\bar{z}).$$

Note that $\hat{\mu}(\cdot)$ is continuous by Lemma 39 and hence also $\hat{q}^a_{cc}(\cdot)$ is continuous. Thus $\hat{q}^a_{cc}(z) > \varepsilon$ on a neighborhood U of $\bar{z}$, and so $\ell^* = f$ on U. This proves the first assertion.

If $\bar{\mu}(\bar{z}^a) \geq \max(\bar{z}^p)$ then $S_{cc}(\bar{z})$ is a singleton $\{\bar{q}\}$, where $\bar{q} \triangleq (\hat{q}^a_{cc}(\bar{z}), 0^p)$. In such a case Danskin's theorem gives that ℓ^* is differentiable at $\bar{z}$ with $\nabla\ell^*(z) = \nabla_z\varphi(z, \bar{q}) = \bar{q}$. So also the second assertion of the lemma is proved. □

The convex conjugate ℓ^* is not strictly convex, even if $\nu > 0$. This is so because $\ell^*(z+c) = \ell^*(z) + c$ for every constant c, cf. Lemma 33. However, the following holds.

Lemma 41. *Let $\nu > 0$. Then ℓ^* is C^∞ and, for every $z \in \mathbb{R}^m$,*

$$\nabla \ell^*(z) = \frac{\nu}{\bar{\mu}(z) - z}, \qquad \nabla^2 \ell^*(z) = \operatorname{diag}(\alpha) - \frac{1}{\sum \alpha}\alpha\alpha',$$

where $\alpha \triangleq (\alpha_1, \ldots, \alpha_m)'$, $\alpha_i \triangleq \nu_i/(\bar{\mu}(z) - z_i)^2$. Consequently, for every $x \in \mathbb{R}^m$, $x'(\nabla^2 \ell^*(z))x \geq 0$ and

$$x'(\nabla^2 \ell^*(z))x = 0 \qquad \text{if and only if} \qquad x = (c, \ldots, c)' \text{ for some } c \in \mathbb{R}.$$

Proof. We use the notation from the proof of Lemma 39. By Corollary 35 and Lemma 39,

$$\frac{\partial \ell^*(z)}{\partial z_k} = \frac{\partial \bar{\mu}(z)}{\partial z_k}\left(1 - \sum_i \frac{\nu_i}{\bar{\mu}(z) - z_i}\right) + \frac{\nu_k}{\bar{\mu}(z) - z_k} = \frac{\nu_k}{\bar{\mu}(z) - z_k}$$

since $\sum_i \nu_i/(\bar{\mu}(z) - z_i) = F(z, \bar{\mu}(z)) = 1$. Further,

$$\frac{\partial^2 \ell^*(z)}{\partial z_k \partial z_l} = -\frac{\nu_k}{(\bar{\mu}(z) - z_k)^2}\frac{\partial \bar{\mu}(z)}{\partial z_l} = -\frac{\alpha_k \alpha_l}{\sum \alpha}$$

for $k \neq l$, and

$$\frac{\partial^2 \ell^*(z)}{\partial z_k^2} = -\frac{\nu_k}{(\bar{\mu}(z) - z_k)^2}\left(\frac{\partial \bar{\mu}(z)}{\partial z_k} - 1\right) = \alpha_k - \frac{\alpha_k^2}{\sum \alpha}.$$

Thus the first assertion of the lemma is proved.

To prove the second part of the lemma, fix any $z, x \in \mathbb{R}^m$ and put $a \triangleq \sum \alpha$, $A \triangleq a\nabla^2 \ell^*(z)$. Then

$$x'Ax = a\sum_i \alpha_i x_i^2 - \left(\sum_i \alpha_i x_i\right)^2 = a^2\left[\sum_i w_i x_i^2 - \left(\sum_i w_i x_i\right)^2\right]$$

($w_i \triangleq \alpha_i/a$). Since $a \neq 0$, Jensen's inequality gives that $x'Ax \geq 0$, and $x'Ax = 0$ if and only if $x_1 = \cdots = x_m$. The lemma is proved. □

10.4. Proof of Theorem 6 (Relation between $\mathcal{F}$ and $\mathcal{P}$)

The proof of Theorem 6 goes through several lemmas. First, in Lemma 43, the extremality relation between $\mathcal{P}$ and $\mathcal{F}$ is established using the Primal Fenchel duality theorem. Then the minimax equality for L (cf. (10.15)) is proved, see Corollary 45. Proposition 47 gives a relation between the solution set $S_{\mathcal{P}}$ of the primal and the solution set $S_{cc}(-\hat{y})$ of the convex conjugate ℓ^* for $\hat{y} \in S_{\mathcal{F}}$. Lemma 48 provides a key for establishing the second part of Theorem 6. The structure of the solution set $S_{\mathcal{F}}$ is described in Lemma 50.

10.4.1. Extremality relation

In the following denote by $\Delta_{\mathcal{X}}^{+}$ the set $\{q \in \Delta_{\mathcal{X}} : q^a > 0\}$; note that neither the primal nor the convex conjugate are affected by restricting to $\Delta_{\mathcal{X}}^{+}$.

Lemma 42. *Let $C \subseteq \Delta_{\mathcal{X}}$. If $y \in C^*$ then $y + \mathbb{R}_{-}^{m} \subseteq C^*$.*

Proof. Take any $y \in C^*$ and $z \leq 0$. Then, for every $q \in C$,

$$\langle q, (y+z)\rangle = \langle q, y\rangle + \langle q, z\rangle \leq \langle q, y\rangle \leq 0.$$

Thus $y + z \in C^*$. □

Lemma 43. *For every convex closed set $C \subseteq \Delta_{\mathcal{X}}$ having support $\mathcal{X}$ and for every $\nu \in \Delta_{\mathcal{X}}$ it holds that*

$$\hat{\ell}_{\mathcal{F}} = -\hat{\ell}_{\mathcal{P}} \qquad \textit{and} \qquad S_{\mathcal{F}} \neq \emptyset.$$

Proof. Write the primal problem $\mathcal{P}$ in the form

$$\hat{\ell}_{\mathcal{P}} = \inf_{x \in \Delta_{\mathcal{X}}^{+} \cap K_C} \ell(x), \qquad \text{where} \quad K_C \triangleq \{\alpha q : q \in C, \alpha \geq 0\}$$

is the convex cone generated by C. Since C is convex, there is $q \in \mathrm{ri}(C)$ [10, Prop. 1.4.1(b)]. Further, $\mathcal{X}$ is the support of C, so there is $q' \in C$ satisfying $q' > 0$. By [10, Prop. 1.4.1(a)], $\tilde{q} \triangleq \frac{1}{2}(q + q') > 0$ belongs to $\mathrm{ri}(C)$. Further, by [41, p. 50],

$$\mathrm{ri}(K_C) = \{\alpha q : q \in \mathrm{ri}(C), \alpha > 0\};$$

thus $\tilde{q} \in \mathrm{ri}(K_C)$. Since trivially $\mathrm{ri}(\Delta_{\mathcal{X}}^{+}) = \{q \in \Delta_{\mathcal{X}} : q > 0\}$, it holds that

$$\mathrm{ri}(\Delta_{\mathcal{X}}^{+}) \cap \mathrm{ri}(K_C) \neq \emptyset.$$

Now the Primal Fenchel duality theorem [10, Prop. 7.2.1, pp. 439–440], applied to the convex set $\Delta_{\mathcal{X}}^{+}$, the convex cone K_C and the real-valued convex function $\ell|_{\Delta_{\mathcal{X}}^{+}}$, gives that

$$\inf_{x \in \Delta_{\mathcal{X}}^{+} \cap K_C} \ell(x) = \sup_{y \in K_C^*} (-\ell^*(-y)) = -\inf_{y \in C^*} \ell^*(-y)$$

and that the supremum in the right-hand side is attained. That is, $\hat{\ell}_{\mathcal{P}} = -\hat{\ell}_{\mathcal{F}}$ and $S_{\mathcal{F}} \neq \emptyset$. □

10.4.2. Minimax equality for L

Let $L : C^* \times \Delta_{\mathcal{X}}^{+} \to \mathbb{R}$ be given by

$$L(y, q) \triangleq -\langle y, q\rangle - \ell(q) = -\langle y, q\rangle + \langle \nu^a, \log q^a\rangle. \tag{10.15}$$

Lemma 44. *For every $y \in C^*$ and $q \in \Delta_{\mathcal{X}}^+$,*

$$\sup_{q\in\Delta_{\mathcal{X}}^+} L(y,q) = \ell^*(-y) \qquad \text{and} \qquad \inf_{y\in C^*} L(y,q) = \begin{cases} -\ell(q) & \text{if } q \in C; \\ -\infty & \text{if } q \notin C. \end{cases}$$

Proof. The first equality is immediate since if $q \in \Delta_{\mathcal{X}} \setminus \Delta_{\mathcal{X}}^+$ then $\ell(q) = \infty$. To show the second one, realize that $\inf_{y\in C^*} L(y,q) = -\ell(q) - \sup_{y\in C^*} \langle y, q\rangle$. If $q \in C$ then $\langle y, q\rangle \le 0$ for every $y \in C^*$, and $\langle y, q\rangle = 0$ for $y = 0 \in C^*$; hence $\inf_{y\in C^*} L(y,q) = -\ell(q)$. If $q \in \Delta_{\mathcal{X}}^+ \setminus C$, there is $\bar{y} \in C^*$ with $\langle \bar{y}, q\rangle > 0$ (use that C is convex closed, thus $C^{**} = K_C^{**} = K_C$ [10, Prop. 3.1.1] and $K_C \cap \Delta_{\mathcal{X}} = C$). Since $\lambda\bar{y} \in C^*$ for every $\lambda \ge 0$,

$$\sup_{y\in C^*} \langle y, q\rangle \ge \sup_{\lambda\ge 0} \lambda \langle \bar{y}, q\rangle = \infty.$$

Thus $\inf_{y\in C^*} L(y,q) = -\infty$ provided $q \in \Delta_{\mathcal{X}}^+ \setminus C$. □

Lemmas 43 and 44 immediately yield the minimax equality for L.

Corollary 45. *We have*

$$\hat{\ell}_{\mathcal{F}} = \inf_{y\in C^*} \sup_{q\in\Delta_{\mathcal{X}}^+} L(y,q) = \sup_{q\in\Delta_{\mathcal{X}}^+} \inf_{y\in C^*} L(y,q) = -\hat{\ell}_{\mathcal{P}}.$$

Lemma 46. *For every $\bar{y} \in \mathbb{R}^m$ and $\bar{q} \in \Delta_{\mathcal{X}}^+$,*

$$(\bar{y}, \bar{q}) \text{ is a saddle point of } L \qquad \iff \qquad \bar{y} \in S_{\mathcal{F}} \text{ and } \bar{q} \in S_{\mathcal{P}}.$$

Recall that $(\bar{y}, \bar{q})$ is a *saddle point* of L [10, Def. 2.6.1] if, for every $y \in C^*$ and $q \in \Delta_{\mathcal{X}}^+$,

$$L(\bar{y}, q) \le L(\bar{y}, \bar{q}) \le L(y, \bar{q}). \tag{10.16}$$

Proof. This result is an immediate consequence of Corollary 45, Lemma 44, and [10, Prop. 2.6.1]. □

10.4.3. Relation between $S_{\mathcal{P}}$ and $S_{cc}(-\hat{y})$

Proposition 47. *Let $\hat{y} \in S_{\mathcal{F}}$. Then*

$$S_{\mathcal{P}} \subseteq S_{cc}(-\hat{y}).$$

Proof. Take any $\hat{y} \in S_{\mathcal{F}}$ and any $\hat{q} \in S_{\mathcal{P}}$. By Lemma 46, $(\hat{y}, \hat{q})$ is a saddle point of L. Hence, by (10.16),

$$-\langle \hat{y}, \hat{q}\rangle - \ell(\hat{q}) = L(\hat{y}, \hat{q}) \ge L(\hat{y}, q) = -\langle \hat{y}, q\rangle - \ell(q)$$

for any $q \in \Delta_{\mathcal{X}}^+$. Thus $\ell^*(-\hat{y}) = L(\hat{y}, \hat{q})$ and so $\hat{q} \in S_{cc}(-\hat{y})$. □

10.4.4. From $\hat{q}_{\mathcal{P}}$ to $\hat{y}_{\mathcal{F}}$

The proof of the following lemma is inspired by that of [12, Thm. 2.1].

Lemma 48. *Let $\nu \in \Delta_{\mathcal{X}}$ and $\hat{q}^a_{\mathcal{P}}$ be as in Theorem 2. Then*

$$\hat{y}_{\mathcal{F}} \triangleq (\hat{y}^a_{\mathcal{F}}, -1^p) \in S_{\mathcal{F}}, \qquad \textit{where} \quad \hat{y}^a_{\mathcal{F}} \triangleq \frac{\nu^a}{\hat{q}^a_{\mathcal{P}}} - 1^a\,.$$

Further, $\hat{\mu}(-\hat{y}_{\mathcal{F}}) = 1$, $\hat{y}_{\mathcal{F}} \perp S_{\mathcal{P}}$, and

(a) if $\sum \hat{q}^a_{\mathcal{P}} = 1$ then $\bar{\mu}(-\hat{y}^a_{\mathcal{F}}) = 1$;
(b) if $\sum \hat{q}^a_{\mathcal{P}} < 1$ then $\bar{\mu}(-\hat{y}^a_{\mathcal{F}}) < 1$.

Proof. Take any $\hat{q} \in S_{\mathcal{P}}$. Then $\hat{q} \in \mathrm{dom}(\ell)$ and, by Lemma 30(d),

$$\begin{aligned} 0 \le \ell'(\hat{q}; q - \hat{q}) &= -\sum_{i \in \mathcal{X}^a} \frac{\nu^a_i (q^a_i - \hat{q}^a_i)}{\hat{q}^a_i} = 1 - \sum_{i \in \mathcal{X}^a} \frac{\nu^a_i q^a_i}{\hat{q}^a_i} \\ &= \sum_{i \in \mathcal{X}^a} q^a_i \left(1 - \frac{\nu^a_i}{\hat{q}^a_i}\right) + \sum_{i \in \mathcal{X}^p} q^p_i = -\langle q, \hat{y}_{\mathcal{F}} \rangle \end{aligned}$$

for every $q \in C$. Hence $\hat{y}_{\mathcal{F}} \in C^*$. Further,

$$\langle \hat{q}, \hat{y}_{\mathcal{F}} \rangle = \sum_{i \in \mathcal{X}^a} \hat{q}^a_i \left(\frac{\nu^a_i}{\hat{q}^a_i} - 1\right) + \sum_{i \in \mathcal{X}^p} \hat{q}^p_i \cdot (-1) = \sum \nu^a - \sum \hat{q} = 0$$

and so $\hat{y}_{\mathcal{F}} \perp \hat{q}$.

We claim that

$$\bar{\mu}(-\hat{y}^a_{\mathcal{F}}) \begin{cases} = 1 & \text{if } \sum \hat{q}^a = 1; \\ < 1 & \text{if } \sum \hat{q}^a < 1. \end{cases} \tag{10.17}$$

Since $\hat{y}^a_{\mathcal{F}} = (\nu^a / \hat{q}^a_{\mathcal{P}}) - 1^a > -1^a$, it holds that $1 > \max(-\hat{y}^a_{\mathcal{F}})$. Moreover,

$$\xi_{-y^a_{\mathcal{F}}}(1) = \sum \frac{\nu^a}{1 + y^a_{\mathcal{F}}} = \sum \hat{q}^a_{\mathcal{P}}.$$

Thus (10.17) follows from (10.6).

Assume that $\sum \hat{q}^a = 1$. Then $\bar{\mu}(-\hat{y}^a_{\mathcal{F}}) = 1$. Since $-\hat{y}^p_{\mathcal{F}} = 1^p$, $\hat{\mu}(-\hat{y}_{\mathcal{F}}) = 1$. Theorem 5 and Lemma 43 yield

$$\ell^*(-\hat{y}_{\mathcal{F}}) = I_1(\nu^a \,\|\, \hat{y}^a_{\mathcal{F}}) = \langle \nu^a, \log \hat{q}^a \rangle = -\ell(\hat{q}) = -\hat{\ell}_{\mathcal{P}} = \hat{\ell}_{\mathcal{F}}.$$

Since $\hat{y}_{\mathcal{F}} \in C^*$, we have $\hat{y}_{\mathcal{F}} \in S_{\mathcal{F}}$.

If $\sum \hat{q}^a < 1$ then $\bar{\mu}(-\hat{y}^a_{\mathcal{F}}) < 1$ by (10.17). Since $-\hat{y}^p_{\mathcal{F}} = 1^p$, again $\hat{\mu}(-\hat{y}_{\mathcal{F}}) = 1$. As before we obtain $\ell^*(-\hat{y}_{\mathcal{F}}) = -\ell(\hat{q}) = \hat{\ell}_{\mathcal{F}}$ and $\hat{y}_{\mathcal{F}} \in S_{\mathcal{F}}$. The lemma is proved. □

10.4.5. Structure of $S_{\mathcal{F}}$

Lemma 49. *Let $\hat{q}^a_{\mathcal{P}}$ be as in Theorem 2. Then*

$$S_{\mathcal{F}} \subseteq \left\{ \hat{y} \in C^* : \hat{y}^a = \frac{\nu^a}{\hat{q}^a_{\mathcal{P}}} - \hat{\mu}(-\hat{y}) \right\}.$$

Proof. Take any $\hat{y} \in S_{\mathcal{F}}$. By Propositions 47 and 36,

$$\hat{q}^a_{\mathcal{P}} = \hat{q}^a_{cc}(-\hat{y}) = \frac{\nu^a}{\hat{\mu}(-\hat{y}) + \hat{y}^a}.$$

From this the result follows immediately. □

Lemma 50. *Let $\hat{y}^a_{\mathcal{F}}$ be as in Lemma 48. Then*

$$S_{\mathcal{F}} = \{\hat{y} \in C^* : \ \hat{y}^a = \hat{y}^a_{\mathcal{F}}, \ \hat{\mu}(-\hat{y}) = 1\}.$$

Proof. One inclusion follows from Lemmas 48 and 37(b). To prove the other one, take any $\hat{y} \in S_{\mathcal{F}}$ and put $\mu = \hat{\mu}(-\hat{y})$. Let $\hat{y}_{\mathcal{F}}$ be as in Lemma 48. Then $\hat{y}^a = \hat{y}^a_{\mathcal{F}} - (\mu - 1)$ by Lemma 49. Theorem 5 yields

$$0 = \ell^*(\hat{y}) - \ell^*(\hat{y}_{\mathcal{F}}) = (\mu + I_\mu(\nu^a \,\|\, \hat{y}^a)) - (1 + I_1(\nu^a \,\|\, \hat{y}^a_{\mathcal{F}})) = \mu - 1.$$

Thus $\mu = 1$ and $\hat{y}^a = \hat{y}^a_{\mathcal{F}}$. □

10.4.6. Proof of Theorem 6

Now we are ready to prove Theorem 6.

Proof of Theorem 6. The facts that $\hat{\ell}_{\mathcal{F}} = -\hat{\ell}_{\mathcal{P}}$ and $S_{\mathcal{F}} \neq \emptyset$ were proved in Lemma 43. Further, the set $S_{\mathcal{F}}$ is convex and closed due to the fact that the conjugate ℓ^* is convex and closed (even continuous, see Lemma 38). To show compactness of $S_{\mathcal{F}}$ it suffices to prove that it is bounded.

We already know from Lemmas 48 and 50 that $\hat{y}^a > -1$ and $\hat{y}^p \geq -\hat{\mu}(-\hat{y}) = -1$ for every $\hat{y} \in S_{\mathcal{F}}$; thus $S_{\mathcal{F}}$ is bounded from below by -1.

Since $\operatorname{supp} C = \mathcal{X}$, there is $q \in C$ with $q > 0$; put $\beta \triangleq \min q_i > 0$. Take any $\hat{y} \in S_{\mathcal{F}}$ and put $I \triangleq \{i \in \mathcal{X} : \hat{y}_i > 0\}$. Then, since $\hat{y} \in C^*$ and $\hat{y} \geq -1$,

$$0 \geq \langle \hat{y}, q \rangle \geq \beta \sum_{i \in I} \hat{y}_i - \sum_{j \in \mathcal{X} \setminus I} q_j \geq \beta \hat{y}_i - 1$$

for every $i \in I$. Thus $\hat{y} \leq (1/\beta)$ and so $S_{\mathcal{F}}$ is bounded, hence compact.

The rest of Theorem 6 follows from Lemmas 48 and 50. □

10.5. Proof of Theorem 16 (Relation between $\mathcal{B}$ and $\mathcal{P}$)

First, basic properties of $\ell^*_{\mathcal{B}}$ are proven. Then Lemma 52 provides a preparation for the recession cone considerations of $\mathcal{B}$. This leads to Proposition 53 giving the conditions of finiteness of $\hat{\ell}_{\mathcal{B}}$. There $\mathcal{B}$ is seen as a primal problem and

Theorem 29 is applied to it. The solution set $S_{\mathcal{B}}$ is described in Lemma 54. Lemma 55 provides properties of $\hat{q}_{\mathcal{B}} \in C$ defined via $\hat{y}_{\mathcal{B}}$, noting that $\hat{q}_{\mathcal{B}}$ need not belong to $S_{\mathcal{P}}$. Conditions equivalent to $\hat{q}_{\mathcal{B}} \in S_{\mathcal{P}}$ are stated in Lemma 58. Its proof utilizes also Lemmas 56 and 57.

Let $\nu \in \Delta_{\mathcal{X}}$ be a type. Recall from Section 4.1 that the map $\ell^*_{\mathcal{B}} : \mathbb{R}^m \to \bar{\mathbb{R}}$ is for $y^a > -1$ defined by

$$\ell^*_{\mathcal{B}}(y) = I_1(\nu^a \,\|\, y^a) = \left\langle \nu^a, \log \frac{\nu^a}{1+y^a} \right\rangle .$$

Since

$$\ell^*_{\mathcal{B}}(y) = \ell(1+y) + \langle \nu^a, \log \nu^a \rangle$$

for every $y \in \mathbb{R}^m$ (where ℓ is defined on $\mathbb{R}^m$ by (10.3)), Lemma 30 yields the following result.

Lemma 51. *If $\nu \in \Delta_{\mathcal{X}}$, then*

(a) $\ell^*_{\mathcal{B}}(y) > -\infty$ *for every* $y \in \mathbb{R}^m$, *and* $\ell^*_{\mathcal{B}}(y) < \infty$ *if and only if* $y^a > -1$; *so*

$$\mathrm{dom}(\ell^*_{\mathcal{B}}) = \{y \in \mathbb{R}^m :\ y^a > -1\};$$

(b) $\ell^*_{\mathcal{B}}$ *is a proper continuous (hence closed) convex function;*
(c) the restriction of $\ell^*_{\mathcal{B}}$ *to its domain* $\mathrm{dom}(\ell^*_{\mathcal{B}})$ *is strictly convex if and only if* $\nu > 0$;
(d) $\ell^*_{\mathcal{B}}$ *is differentiable on* $\mathrm{dom}(\ell^*_{\mathcal{B}})$ *with the gradient given by*

$$\nabla \ell^*_{\mathcal{B}}(y) = -(\nu^a/(1+y^a), 0^p);$$

(e) the recession cone $R_{\ell^*_{\mathcal{B}}}$ *and the constancy space* $L_{\ell^*_{\mathcal{B}}}$ *of* $\ell^*_{\mathcal{B}}$ *are*

$$R_{\ell^*_{\mathcal{B}}} = \{z \in \mathbb{R}^m :\ z^a \geq 0\}, \qquad L_{\ell^*_{\mathcal{B}}} = \{z \in \mathbb{R}^m :\ z^a = 0\}.$$

10.5.1. Finiteness of the BD-dual $\mathcal{B}$

Lemma 52. *Let $\nu \in \Delta_{\mathcal{X}}$ and C be a convex closed subset of $\Delta_{\mathcal{X}}$ having support $\mathcal{X}$. Assume that C is either an H-set or a Z-set. Then there is an active letter i with the following property: For any $\gamma > 0$ and $v > 0$*

$$\text{there is} \quad y \in C^* \quad \text{such that} \quad y^a \geq -\gamma \quad \text{and} \quad y^a_i = v. \tag{10.18}$$

Proof. Assume first that C is a Z-set, that is, $C^a(0^p) \neq \emptyset$ and there is an active letter i such that $q^a_i = 0$ whenever $q \in C$ satisfies $q^p = 0$. Fix any $\gamma > 0, v > 0$, and choose $\varepsilon \in (0,1)$ such that $\varepsilon < \gamma/(v + 2\gamma)$. By compactness of C we can find $0 < \delta < \varepsilon$ such that, for every $q \in C$, $\sum q^p \leq \delta$ implies $q^a_i < \varepsilon$. (For if not, there is a sequence $(q^{(n)})_n$ in C with $\sum (q^{(n)})^p \leq 1/n$ and $(q^{(n)})^a_i \geq \varepsilon$ for every n; by compactness, there is a limit point $q \in C$ of this sequence and any such q satisfies $\sum q^p = 0$ and $q^a_i \geq \varepsilon$, a contradiction.) Finally, take any $w \geq v/\delta$.

Define $y \in \mathbb{R}^m$ by

$$y_i^a = v, \quad y_j^a = -\gamma \text{ for } j \in \mathcal{X}^a \setminus \{i\}, \quad \text{and} \quad y_j^p = -w \text{ for } j \in \mathcal{X}^p. \tag{10.19}$$

Take arbitrary $q \in C$; we are going to show that $\langle y, q\rangle \le 0$. If $\sum q^p \le \delta$ then $q_i^a \le \varepsilon$ and $\sum q^a \ge 1-\delta > 1-\varepsilon$, so

$$\langle y, q\rangle = (v+\gamma)q_i^a - \gamma \sum q^a - w \sum q^p < (v+\gamma)\varepsilon - \gamma(1-\varepsilon) < 0$$

by the choice of ε. On the other hand, if $\sum q^p > \delta$ then, using $q_i^a \le \min\{1, \sum q^a\}$,

$$\langle y, q\rangle = vq_i^a - \gamma \left(\sum q^a - q_i^a\right) - w \sum q^p < v - w\delta \le 0$$

by the choice of w. Thus $y \in C^*$.

Now assume that C is an H-set; that is, $C^a(0^p) = \emptyset$. By compactness, there exists $\delta > 0$ such that $\sum q^p > \delta$ for every $q \in C$. Continuing as above we obtain that, for any $\gamma > 0, v > 0$, and $w \ge v/\delta$, the vector y given by (10.19) belongs to C^*. □

Proposition 53. *Let $\nu \in \Delta_{\mathcal{X}}$ and C be a convex closed subset of $\Delta_{\mathcal{X}}$ having support $\mathcal{X}$. Then the following are equivalent:*

(a) the dual $\mathcal{B}$ is finite;
(b) the set C is neither an H-set nor a Z-set (with respect to ν).

Proof. Assume that C is either an H-set or a Z-set. Fix any $\gamma \in (0,1)$. Then, by Lemma 52, for arbitrary $v > 0$ there is $y \in C^*$ satisfying (10.18). For such y,

$$\begin{aligned} -\ell_{\mathcal{B}}^*(y) + \langle \nu^a, \log \nu^a\rangle &\ge \sum_{j \in \mathcal{X}^a \setminus \{i\}} \nu_j^a \log(1-\gamma) + \nu_i^a \log(1+v) \\ &= (1-\nu_i^a)\log(1-\gamma) + \nu_i^a \log(1+v). \end{aligned}$$

Since γ is fixed and $v > 0$ is arbitrary, $\hat{\ell}_{\mathcal{B}} = -\infty$. This proves that (a) implies (b).

Now we show that (b) implies (a). Assume that C is neither an H-set nor a Z-set; that is, there is $q \in C$ with $q^a > 0$, $q^p = 0$. Put $D = \{q\}^*$. Since $R_D = D = \{y : \langle y^a, q^a\rangle \le 0\}$, $L_D = D \cap (-D) = \{y : \langle y^a, q^a\rangle = 0\}$ and, by Lemma 51, $R_{\ell_{\mathcal{B}}^*} = \{y : y^a \ge 0\}$ and $L_{\ell_{\mathcal{B}}^*} = \{y : y^a = 0\}$, it follows that

$$R_{\ell_{\mathcal{B}}^*} \cap R_D = L_{\ell_{\mathcal{B}}^*} \cap L_D = \{y : y^a = 0\}.$$

Now Theorem 29 (applied to $\ell_{\mathcal{B}}^*$ and D) implies that $\operatorname{argmin}_D \ell_{\mathcal{B}}^*$ is nonempty, hence $\ell_{\mathcal{B}}^*$ is bounded from below on D. Since $C^* \subseteq D$, (a) is established. □

10.5.2. The solution set $S_{\mathcal{B}}$ of the BD-dual

Lemma 54. *Let $\nu \in \Delta_{\mathcal{X}}$. Let C be a convex closed subset of $\Delta_{\mathcal{X}}$ having support $\mathcal{X}$, which is neither an H-set nor a Z-set (with respect to ν). Then the solution set $S_{\mathcal{B}}$ is nonempty, and there is $\hat{y}_{\mathcal{B}}^a \in C^{*a}$ such that $\hat{y}_{\mathcal{B}}^a > -1$ and*

$$S_{\mathcal{B}} = \{\hat{y} \in C^* : \hat{y}^a = \hat{y}_{\mathcal{B}}^a\} = \{\hat{y}_{\mathcal{B}}^a\} \times C^{*p}(\hat{y}_{\mathcal{B}}^a).$$

Moreover, $S_{\mathcal{B}}$ is a singleton if and only if $\nu > 0$.

Proof. The proof is analogous to that of Proposition 32. Define $\ell^{*a}_{\mathcal{B}} : C^{*a} \to \bar{\mathbb{R}}$ by $\ell^{*a}_{\mathcal{B}}(y^a) \triangleq \langle \nu^a, \log(\nu^a/(1+y^a))\rangle$ if $y^a > -1$, $\ell^{*a}_{\mathcal{B}}(y^a) \triangleq \infty$ otherwise. Then $\ell^*_{\mathcal{B}}(y) = \ell^{*a}_{\mathcal{B}}(y^a)$ for any $y \in C^*$, hence

$$\hat{\ell}_{\mathcal{B}} = \inf_{y^a \in C^{*a}} \ell^{*a}_{\mathcal{B}}(y^a) \quad \text{and} \quad S_{\mathcal{B}} = \left\{ y \in C^* : y^a \in \operatorname*{argmin}_{C^{*a}} \ell^{*a}_{\mathcal{B}} \right\}. \tag{10.20}$$

By Lemma 51, $\ell^{*a}_{\mathcal{B}}$ is strictly convex, so $\operatorname{argmin}_{C^{*a}} \ell^{*a}_{\mathcal{B}}$ contains at most one point. The set C^* is neither an H-set nor a Z-set, so there is $q \in C$ with $q^a > 0$ and $q^p = 0$. Thus, as in the proof of Proposition 53, $R_{\ell^{*a}_{\mathcal{B}}} \cap R_{C^{*a}} \subseteq R_{\ell^{*a}_{\mathcal{B}}} \cap R_{\{q^a\}^*} = \{0^a\}$. Since $C^{*a} \cap \operatorname{dom}(\ell^{*a}_{\mathcal{B}}) \neq \emptyset$, Theorem 28 gives that $\operatorname{argmin}_{C^{*a}} \ell^{*a}_{\mathcal{B}}$ is nonempty, hence a singleton; denote its unique point by $\hat{y}^a_{\mathcal{B}}$. Since trivially $\hat{y}^a_{\mathcal{B}} > -1$, the first assertion of the lemma follows from (10.20). The second assertion follows from the first one and Lemma 42. ◻

10.5.3. No BD-duality gap; proof of Theorem 16

The proof of the following lemma is inspired by that of [12, Thm. 2.1].

Lemma 55 (Relation between $\hat{y}_{\mathcal{B}}$ and $\hat{q}_{\mathcal{B}}$)**.** *Let $\nu \in \Delta_{\mathcal{X}}$. Let C be a convex closed subset of $\Delta_{\mathcal{X}}$ having support $\mathcal{X}$, which is neither an H-set nor a Z-set (with respect to ν). Let $\hat{y}^a_{\mathcal{B}}$ be as in Lemma 54 and put*

$$\hat{q}_{\mathcal{B}} \triangleq \left(\frac{\nu^a}{1+\hat{y}^a_{\mathcal{B}}}, 0^p \right). \tag{10.21}$$

Then

$$\bar{\mu}(-\hat{y}^a_{\mathcal{B}}) = 1, \qquad \hat{q}_{\mathcal{B}} \in C, \qquad \hat{q}_{\mathcal{B}} \perp S_{\mathcal{B}}, \qquad \textit{and} \qquad \ell(\hat{q}_{\mathcal{B}}) = -\hat{\ell}_{\mathcal{B}}.$$

Note that $\hat{q}_{\mathcal{B}}$ need not belong to $S_{\mathcal{P}}$; for conditions equivalent to $\hat{q}_{\mathcal{B}} \in S_{\mathcal{P}}$, see Lemma 58.

Proof. Take any $\hat{y} = (\hat{y}^a_{\mathcal{B}}, \hat{y}^p) \in S_{\mathcal{B}}$. Then

$$0 \le (\ell^*_{\mathcal{B}})'(\hat{y}; y - \hat{y}) = -\sum \frac{\nu^a(y^a - \hat{y}^a_{\mathcal{B}})}{1+\hat{y}^a_{\mathcal{B}}} \qquad \text{for every } y \in C^*. \tag{10.22}$$

Applying (10.22) to $y = 2\hat{y}$ and then to $y = (1/2)\hat{y}$ (both belonging to C^* since C^* is a cone) gives

$$\langle \hat{q}_{\mathcal{B}}, \hat{y} \rangle = \sum \frac{\nu^a \hat{y}^a_{\mathcal{B}}}{1+\hat{y}^a_{\mathcal{B}}} = 0. \tag{10.23}$$

Now (10.22) and (10.23) yield $\langle \hat{q}_{\mathcal{B}}, y \rangle \le 0$ for every $y \in C^*$, that is, $\hat{q}_{\mathcal{B}} \in C^{**} = K_C$ (for the last equality use that C is convex closed). Further (recall the definition (10.5) of ξ),

$$1 = \sum \frac{\nu^a(1+\hat{y}^a_{\mathcal{B}})}{1+\hat{y}^a_{\mathcal{B}}} = \sum \frac{\nu^a}{1+\hat{y}^a_{\mathcal{B}}} = \xi_{-\hat{y}^a_{\mathcal{B}}}(1)$$

by (10.23) and (10.21). Since $\max(-\hat{y}^a_{\mathcal{B}}) < 1$, $\bar{\mu}(-\hat{y}^a_{\mathcal{B}}) = 1$. Moreover, $\hat{q}_{\mathcal{B}} \in \Delta_{\mathcal{X}}$ (use that $\hat{q}_{\mathcal{B}} \geq 0$), and so $\hat{q}_{\mathcal{B}} \in (K_C \cap \Delta_{\mathcal{X}}) = C$. Finally,

$$\ell(\hat{q}_{\mathcal{B}}) = -\left\langle \nu^a, \log \hat{q}^a_{\mathcal{B}} \right\rangle = -I_1(\nu^a \,\|\, \hat{y}^a_{\mathcal{B}}) = -\ell^*_{\mathcal{B}}(\hat{y}) = -\hat{\ell}_{\mathcal{B}}.$$

The lemma is proved. □

Lemma 56. *Let $\nu \in \Delta_{\mathcal{X}}$ and C be a convex closed subset of $\Delta_{\mathcal{X}}$ having support $\mathcal{X}$. Then*

$$\hat{\ell}_{\mathcal{B}} \leq \hat{\ell}_{\mathcal{F}}.$$

Moreover, if $\nu > 0$ then $\hat{\ell}_{\mathcal{B}} = \hat{\ell}_{\mathcal{F}}$.

Proof. It follows from Theorems 6 and 5 that

$$\hat{\ell}_{\mathcal{F}} = \inf_{\substack{y \in C^* \\ \hat{\mu}(-y)=1}} \ell^*(-y) = \inf_{\substack{y \in C^* \\ \hat{\mu}(-y)=1}} I_1(\nu^a \,\|\, y^a) \geq \inf_{y \in C^*} I_1(\nu^a \,\|\, y^a) = \hat{\ell}_{\mathcal{B}}.$$

Hence the inequality is proved. If $\nu > 0$ then $\hat{\mu}(-y) = \bar{\mu}(-y^a)$ for every y; so, by Lemma 55,

$$\hat{\ell}_{\mathcal{B}} = \inf_{\substack{y \in C^* \\ \bar{\mu}(-y^a)=1}} I_1(\nu^a \,\|\, y^a) = \hat{\ell}_{\mathcal{F}}.$$ □

Lemma 57. *Let $\nu \in \Delta_{\mathcal{X}}$, C be a convex closed subset of $\Delta_{\mathcal{X}}$ having support $\mathcal{X}$, and $\hat{q}^a_{\mathcal{P}}$ be as in Theorem 2. Assume that C is neither an H-set nor a Z-set (with respect to ν), and that $\sum \hat{q}^a_{\mathcal{P}} = 1$. Then $\hat{\ell}_{\mathcal{B}} = \hat{\ell}_{\mathcal{F}}$ and $\hat{y} \triangleq (\nu^a/\hat{q}^a_{\mathcal{P}} - 1^a, -1^p)$ belongs to $S_{\mathcal{B}} \cap S_{\mathcal{F}}$.*

Proof. By Theorems 6(a) and 5, $\hat{y} \in S_{\mathcal{F}}$ and $\hat{\ell}_{\mathcal{F}} = \ell^*(-\hat{y}) = I_1(\nu^a \,\|\, \hat{y}) = \ell^*_{\mathcal{B}}(\hat{y})$. Moreover, $\xi_{-\hat{y}^a}(1) = \sum \hat{q}^a_{\mathcal{P}} = 1$, so $\bar{\mu}(-\hat{y}^a) = 1$. Thus

$$\ell^*_{\mathcal{B}}(\hat{y}) = \hat{\ell}_{\mathcal{F}} \qquad \text{and} \qquad \hat{\mu}(-\hat{y}) = \bar{\mu}(-\hat{y}^a) = 1. \tag{10.24}$$

Now we prove that $\hat{\ell}_{\mathcal{B}} = \hat{\ell}_{\mathcal{F}}$. For $\nu > 0$ this follows from Lemma 56. In the other case put $C' \triangleq C^a(0^p) = C^a \cap \Delta_{\mathcal{X}^a}$. This is a nonempty convex closed set with $\operatorname{supp}(C') = \mathcal{X}^a$ (use that C is neither an H-set nor a Z-set). Put $\nu' \triangleq \nu^a > 0$ and

$$\hat{\ell}'_{\mathcal{B}} \triangleq \inf_{y' \in C'^*} I_1(\nu' \,\|\, y'), \qquad \hat{\ell}'_{\mathcal{F}} \triangleq \inf_{y' \in C'^*} \ell'^*(-y'), \qquad \hat{\ell}'_{\mathcal{P}} \triangleq \inf_{q' \in C'} \ell'(q'),$$

where $\ell'(q') \triangleq -\langle \nu', \log q' \rangle$ is defined on $\Delta_{\mathcal{X}^a}$. Since $\hat{q}^a_{\mathcal{P}} \in C'$ we trivially have $\hat{\ell}'_{\mathcal{P}} = \hat{\ell}_{\mathcal{P}}$. Further, $C'^* \supseteq C^{*a}$. (To see this, take any $y^a \in C^{*a}$; then there is y^p such that $\langle y^a, q^a \rangle + \langle y^p, q^p \rangle \leq 0$ for every $q \in C$. Since $(q', 0^p) \in C$ for any $q' \in C'$, for every such q' we have $\langle y^a, q' \rangle \leq 0$; that is, $y^a \in C'^*$.) Hence $\hat{\ell}'_{\mathcal{B}} \leq \inf_{y^a \in C^{*a}} I_1(\nu^a \,\|\, y^a) = \hat{\ell}_{\mathcal{B}}$ (for the equality see the definition ($\mathcal{B}$) of $\hat{\ell}_{\mathcal{B}}$). By Lemma 56 and Theorem 6, $\hat{\ell}'_{\mathcal{B}} = \hat{\ell}'_{\mathcal{F}} = -\hat{\ell}'_{\mathcal{P}}$, thus,

$$\hat{\ell}_{\mathcal{B}} \geq \hat{\ell}'_{\mathcal{B}} = -\hat{\ell}'_{\mathcal{P}} = -\hat{\ell}_{\mathcal{P}} = \hat{\ell}_{\mathcal{F}}.$$

Now $\hat{\ell}_{\mathcal{B}} = \hat{\ell}_{\mathcal{F}}$ by the inequality from Lemma 56. To finish the proof it suffices to realize that this fact and (10.24) yield $\ell^*_{\mathcal{B}}(\hat{y}) = \hat{\ell}_{\mathcal{B}}$, and so $\hat{y} \in S_{\mathcal{B}}$. □

Lemma 58. *Keep the assumptions and notation from Lemma 55. Then there is no BD-duality gap, that is,*

$$\hat{\ell}_{\mathcal{B}} = -\hat{\ell}_{\mathcal{P}},$$

if and only if any of the following (equivalent) conditions hold:

(a) $\hat{q}^a_{\mathcal{B}} = \hat{q}^a_{\mathcal{P}}$*;*
(b) $\hat{q}_{\mathcal{B}} \in S_{\mathcal{P}}$*;*
(c) $S_{\mathcal{P}} = \{\hat{q}_{\mathcal{B}}\}$*;*
(d) $\sum \hat{q}^a_{\mathcal{P}} = 1$*;*
(e) $\hat{y}^a_{\mathcal{B}} = \hat{y}^a_{\mathcal{F}}$*;*
(f) $S_{\mathcal{F}} \cap S_{\mathcal{B}} \neq \emptyset$*;*
(g) $S_{\mathcal{F}} = S_{\mathcal{B}} \cap \{\hat{y} \in C^* : \hat{\mu}(-\hat{y}) = 1\}$*;*
(h) $\hat{\mu}(-\hat{y}_{\mathcal{B}}) = 1$ *for some* $\hat{y}_{\mathcal{B}} \in S_{\mathcal{B}}$*;*
(i) $\hat{y}^p_{\mathcal{B}} \geq -1$ *for some* $\hat{y}_{\mathcal{B}} \in S_{\mathcal{B}}$*;*
(j) $\ell(\hat{q}_{\mathcal{B}}) + \ell^*(-\hat{y}_{\mathcal{B}}) = 0$ *for some* $\hat{y}_{\mathcal{B}} \in S_{\mathcal{B}}$ *(*extremality relation*);*
(k) $(\hat{y}^a_{\mathcal{B}}, -1^p) \in C^*$.

Proof. We first prove that the conditions (a)–(k) are equivalent. First, Theorem 2 and Lemma 55 yield that the conditions (a)–(c) are equivalent and that any of them implies (d). By the definitions of $\hat{y}^a_{\mathcal{F}}$ and $\hat{q}^a_{\mathcal{B}}$ from Theorem 6 and Lemma 55, (a) is equivalent to (e). Theorem 6 and Lemma 54 yield that the conditions (e)–(g) are equivalent. Since (d) implies (f) by Lemma 57, it follows that (a)–(g) are equivalent.

Since $\bar{\mu}(-\hat{y}^a_{\mathcal{B}}) = 1$ for any $\hat{y}_{\mathcal{B}} \in S_{\mathcal{B}}$ by Lemmas 54 and 55, (h) is equivalent to (i). Since $S_{\mathcal{F}}$ is nonempty, (g) implies (h). If (h) is true then, for some $\hat{y}^{\mathcal{B}} \in S_{\mathcal{B}}$,

$$\hat{\ell}_{\mathcal{B}} = \ell^*_{\mathcal{B}}(\hat{y}_{\mathcal{B}}) = I_1(\nu^a \,\|\, \hat{y}^a_{\mathcal{B}}) = \ell^*(-\hat{y}_{\mathcal{B}}) \geq \hat{\ell}_{\mathcal{F}}$$

by Theorem 5; now Lemma 56 gives that $\hat{y}_{\mathcal{B}} \in S_{\mathcal{F}}$. Hence (h) implies (f). Further, (b), (e), and (h) imply (j) by Theorem 5 and Lemma 55. On the other hand, the extremality relation (j) implies that $\hat{q}_{\mathcal{B}} \in S_{\mathcal{P}}$ and that $\hat{y}_{\mathcal{B}} \in S_{\mathcal{F}}$; that is, (j) implies both (b) and (f). So (a)–(j) are equivalent

By Theorem 6, (e) implies (k). By Lemmas 55 and 54, (k) implies that $\hat{y} \triangleq (\hat{y}^a_{\mathcal{B}}, -1^p)$ belongs to $S_{\mathcal{B}}$ and that $\hat{\mu}(-\hat{y}) = 1$; thus, (k) implies (h).

We have proved that the conditions (a)–(k) are equivalent. To finish we show that no BD-duality gap is equivalent to the condition (b). If (b) is true (that is, $\hat{q}_{\mathcal{B}} \in S_{\mathcal{P}}$) then $\hat{\ell}_{\mathcal{P}} = \ell(\hat{q}_{\mathcal{B}}) = -\hat{\ell}_{\mathcal{B}}$ by Lemma 55; hence (b) implies no BD-duality gap. Assume now that $\hat{\ell}_{\mathcal{B}} = -\hat{\ell}_{\mathcal{P}}$. Then Lemma 55 yields $\ell(\hat{q}_{\mathcal{B}}) = -\hat{\ell}_{\mathcal{B}} = \hat{\ell}_{\mathcal{P}}$. So (b) is satisfied and the proof is finished. □

Proof of Theorem 16. The theorem immediately follows from Proposition 53 and Lemmas 54, 55, 56 and 58. □

10.6. Proof of Theorem 19 (Active-passive dualization $\mathcal{A}_\kappa$)

Proof of Theorem 19. Fix $q^p \in C^p$ with $\mathrm{supp}(C^a(q^p)) = \mathcal{X}^a$. Then $\sum q^p < 1$. Recall that $\kappa = \kappa(q^p) = 1/(1 - \sum q^p)$ and note that $C^a(q^p)$ is a convex closed subset of $\{q^a \in \mathbb{R}_+^{m_a} : \sum q^a = 1/\kappa\}$. Thus

$$\tilde{C} \triangleq \kappa \cdot C^a(q^p) = \{\kappa q^a : q^a \in C^a(q^p)\}$$

is a convex closed subset of $\Delta_{\mathcal{X}^a}$ with $\text{supp}(\tilde{C}) = \mathcal{X}^a$. Put $\tilde{\nu} \triangleq \nu^a$, and denote by $\tilde{\mathcal{P}}$ and $S_{\tilde{\mathcal{P}}}$ the primal problem and its solution set for minimizing $\tilde{\ell}(\tilde{q}) \triangleq -\langle \tilde{\nu}, \log \tilde{q}\rangle = \ell(\tilde{q}/\kappa, q^p) - \log\kappa$ over $\tilde{C}$. Then $S_{\tilde{\mathcal{P}}}$ is a singleton $\{\hat{q}_{\tilde{\mathcal{P}}}\}$ by Theorem 2, and $S_{\tilde{\mathcal{P}}} = \kappa S_{\mathcal{P}}(q^p)$; hence

$$\hat{q}^a_{\mathcal{P}}(q^p) = \kappa^{-1}\hat{q}_{\tilde{\mathcal{P}}}. \tag{10.25}$$

Consider now the BD-dual $\tilde{\mathcal{B}}$ to $\tilde{\mathcal{P}}$. First, $\ell^*_{\tilde{\mathcal{B}}}(\tilde{y}) = I_1(\tilde{\nu} \,\|\, \tilde{y})$ and, by Corollary 17,

$$S_{\tilde{\mathcal{B}}} = \{\hat{y}_{\tilde{\mathcal{B}}}\}, \qquad \hat{y}_{\tilde{\mathcal{B}}} = \frac{\tilde{\nu}}{\hat{q}_{\tilde{\mathcal{B}}}} - 1. \tag{10.26}$$

Further, $\tilde{C} = [C^a(q^p)]^*$ by the definition of the polar cone. The fact that $I_\kappa(\nu^a \,\|\, y^a) = I_1(\nu^a \,\|\, \kappa^{-1} y^a) - \langle \nu^a, \log\kappa\rangle = \ell^*_{\tilde{\mathcal{B}}}(\kappa^{-1}y^a) - \log\kappa$ now implies that $S_{\mathcal{B}}(q^p) = \kappa S_{\tilde{\mathcal{B}}}$. That is,

$$S_{\mathcal{B}}(q^p) = \{\hat{y}^a_{\mathcal{B}}\}, \qquad \text{where} \quad \hat{y}^a_{\mathcal{B}}(q^p) = \kappa \hat{y}_{\tilde{\mathcal{B}}}. \tag{10.27}$$

Finally, (10.25), (10.26), and (10.27) yield (5.1). The rest follows from (5.1) and Corollary 17. □

10.7. *Proof of Theorem 20 (Perturbed primal $\mathcal{P}_\delta$ – the general case, epi-convergence)*

We start with some notation, which will be used also in Section 10.8. Then we embark on proving the epi-convergence of perturbed primal problems.

10.7.1. *Perturbed primal $\mathcal{P}_\delta$ – notation*

Fix any $\nu \in \Delta_{\mathcal{X}}$. Recall that m_a, m_p denote the cardinalities of $\mathcal{X}^a, \mathcal{X}^p$. For every $\delta > 0$ take $\nu(\delta) \in \Delta_{\mathcal{X}}$ such that (6.1) is true; that is,

$$\nu(\delta) > 0 \qquad \text{and} \qquad \lim_{\delta \searrow 0} \nu(\delta) = \nu.$$

Recall also the definitions of $\ell_\delta, \hat{\ell}_{\mathcal{P}}(\delta), S_{\mathcal{P}}(\delta)$ from ($\mathcal{P}_\delta$) and the fact that (cf. (6.2))

$$S_{\mathcal{P}}(\delta) = \{\hat{q}_{\mathcal{P}}(\delta)\} \qquad \text{for } \delta > 0.$$

The maps $\ell^C, \ell^C_\delta : \mathbb{R}^m \to \bar{\mathbb{R}}$ are defined by (see Section 6)

$$\ell^C(x) \triangleq \begin{cases} \ell(x) & \text{if } x \in C, \\ \infty & \text{if } x \notin C, \end{cases} \qquad \text{and} \qquad \ell^C_\delta(x) \triangleq \begin{cases} \ell_\delta(x) & \text{if } x \in C, \\ \infty & \text{if } x \notin C. \end{cases}$$

10.7.2. Epi–convergence

For the definition of epi-convergence, see [42, Chap. 7]. We will use [42, Ex. 7.3, p. 242] stating that, for every sequence $(g_n)_n$ of maps $g_n : \mathbb{R}^m \to \bar{\mathbb{R}}$ and for every $x \in \mathbb{R}^m$,

$$\begin{aligned}(\text{e-}\liminf_n g_n)(x) &= \lim_{\varepsilon\to 0+} \liminf_{n\to\infty} \inf_{x'\in B(x,\varepsilon)} g_n(x'),\\ (\text{e-}\limsup_n g_n)(x) &= \lim_{\varepsilon\to 0+} \limsup_{n\to\infty} \inf_{x'\in B(x,\varepsilon)} g_n(x'),\end{aligned} \tag{10.28}$$

where $B(x,\varepsilon)$ is the closed ball with the center x and radius ε. If $\text{e-}\liminf_n g_n = \text{e-}\limsup_n g_n \triangleq g$, we say that the sequence $(g_n)_n$ *epi-converges* to g and we write $\text{e-}\lim_n g_n = g$. For a system $(g_\delta)_{\delta>0}$ of maps indexed by real numbers $\delta > 0$, we say that $(g_\delta)_\delta$ *epi-converges* to g for $\delta \searrow 0$ and we write $\text{e-}\lim_\delta g_\delta = g$, provided $\text{e-}\lim_n g_{\delta_n} = g$ for every sequence $(\delta_n)_n$ decreasing to zero.

Before proving the epi-convergence of ℓ^C_δ to ℓ^C, two simple lemmas are presented.

Lemma 59. *Let C be a convex closed subset of $\Delta_{\mathcal{X}}$ with $\operatorname{supp}(C) = \mathcal{X}$. Then there exists a positive real β such that*

$$\hat{q}_{\mathcal{P}}(\nu) \geq \beta\nu \qquad \textit{for every} \quad \nu \in \Delta_{\mathcal{X}}, \nu > 0,$$

where, for $\nu > 0$, $\hat{q}_{\mathcal{P}}(\nu)$ denotes the unique point of $S_{\mathcal{P}}(\nu)$.

Proof. Fix some $q > 0$ from C and put $\beta = \min_i q_i$. For any $\nu > 0$ and $\hat{q} = \hat{q}_{\mathcal{P}}(\nu)$ we have $\hat{q} > 0$; so, by Lemma 30(d), ℓ_ν is differentiable at $\hat{q}$ and the directional derivative $\ell'_\nu(\hat{q}; q - \hat{q}) = -\sum_{i\in\mathcal{X}}[\nu_i(q_i - \hat{q}_i)/\hat{q}_i]$ is nonnegative. Thus, by the choice of β,

$$0 \geq \sum_{i\in\mathcal{X}} \frac{\nu_i q_i}{\hat{q}_i} - 1 \geq \frac{\nu_j \beta}{\hat{q}_j} - 1$$

for every j. From this the lemma follows. □

Lemma 60. *Fix $\beta > 0$ and take the map*

$$F : D \to \bar{\mathbb{R}}, \qquad F(\nu, q) \triangleq \ell_\nu(q) = -\langle \nu, \log q\rangle$$

defined on

$$D \triangleq \{(\nu, q) \in \Delta_{\mathcal{X}} \times C : q \geq \beta\nu\}.$$

Then F is continuous on D.

Proof. Take any sequence $(\nu^n, q^n)_n$ from D such that $\nu^n \to \nu$, $q^n \to q$. Then $q \geq \beta\nu$. The sum $\sum_{i:\nu_i>0} \nu_i^n \log q_i^n$ converges to $\sum_{i:\nu_i>0} \nu_i \log q_i$. Further, for every i such that $\nu_i = 0$,

$$0 \geq \nu_i^n \log q_i^n \geq \nu_i^n \log(\beta\nu_i^n)$$

(which is true also in the case when $\nu_i^n = 0$) and the right-hand side converges to zero. Hence $\lim_n F(\nu^n, q^n) = F(\nu, q)$ and the lemma is proved. □

Note that F need not be continuous on $\Delta_{\mathcal{X}}\times C$. For example, take $\mathcal{X}=\{0,1\}$, $C=\{(0,1)\}$, $\nu^n=(1/n,1-(1/n))$ converging to $\nu=(0,1)$, and $q^n=q=(0,1)$. Then $(\nu^n,q^n)\to(\nu,q)$, but $F(\nu^n,q^n)=\infty$ does not converge to $F(\nu,q)=0$.

Lemma 61. *Let C be a convex closed subset of $\Delta_{\mathcal{X}}$ with $\operatorname{supp}(C)=\mathcal{X}$. Then the maps ℓ_δ^C epi-converge to ℓ^C, that is,*

$$\operatorname*{e-lim}_{\delta\searrow 0}\ell_\delta^C=\ell^C.$$

Proof. Take any sequence $(\delta_n)_n$ decreasing to 0. To prove that $\text{e-lim}_n\,\ell_{\delta_n}^C=\ell^C$ we use (10.28). Take $x\in\mathbb{R}^m$, $\varepsilon>0$, and put

$$\psi_n(x,\varepsilon)\triangleq\inf_{y\in B(x,\varepsilon)}\ell_{\delta_n}^C(y)=\inf_{y\in C\cap B(x,\varepsilon)}\ell_{\delta_n}(y).$$

If $x\notin C$ then, since C is closed, $\psi_n(x,\varepsilon)=\infty$ provided ε is small enough; hence $\lim_\varepsilon\lim_n\psi_n(x,\varepsilon)=\infty=\ell^C(x)$.

Assume now that $x\in C$. Then, by the monotonicity of the logarithm,

$$\psi_n(x,\varepsilon)\in[\ell_{\delta_n}(x+\varepsilon),\ell_{\delta_n}(x)].$$

If $x>0$ and $0<\varepsilon<\min_i x_i$, then $\lim_n\ell_{\delta_n}(x)=\ell^C(x)$ and $\lim_n\ell_{\delta_n}(x+\varepsilon)=\ell^C(x+\varepsilon)$, so continuity of ℓ at x easily gives that $\lim_\varepsilon\lim_n\psi_n(x,\varepsilon)=\ell^C(x)$. If there is $i\in\mathcal{X}^a$ with $x_i^a=0$, then $\ell^C(x)=\infty$ and $\lim_\varepsilon\lim_n(-\nu(\delta_n)_i^a\log(x_i^a+\varepsilon))=-\nu_i^a\lim_\varepsilon\log\varepsilon=\infty$, so again $\lim_\varepsilon\lim_n\psi_n(x,\varepsilon)=\ell^C(x)$.

Finally, assume that $x\in C$ is such that $x^a>0$ and there is $i\in\mathcal{X}^p$ with $x_i^p=0$. By Theorem 2 (applied to $\nu(\delta_n)>0$ and to the nonempty convex closed subset $C\cap B(x,\varepsilon)$ of $\Delta_{\mathcal{X}}$), for every n there is unique $\hat{x}_{n,\varepsilon}>0$ from $C\cap B(x,\varepsilon)$ such that

$$\psi_n(x,\varepsilon)=\ell_{\delta_n}(\hat{x}_{n,\varepsilon}).$$

Moreover, by Lemma 59 there is $\beta>0$ such that

$$\hat{x}_{n,\varepsilon}\geq\beta\nu(\delta_n)\qquad\text{for every } n.\tag{10.29}$$

Take any convergent subsequence of $(\hat{x}_{n,\varepsilon})_n$, denoted again by $(\hat{x}_{n,\varepsilon})_n$, and let $\bar{x}_\varepsilon\in C\cap B(x,\varepsilon)$ denote its limit. Then, by (10.29) and Lemma 60, $\ell_{\delta_n}(\hat{x}_{n,\varepsilon})$ converges to $\ell(\bar{x}_\varepsilon)$. Hence all cluster points of $(\psi_n(x,\varepsilon))_n$ belong to $\ell(C\cap B(x,\varepsilon))$. Since ℓ is continuous, $\lim_\varepsilon\ell(C\cap B(x,\varepsilon))=\{\ell(x)\}$. This implies that $\text{e-lim}_n\,\ell_{\delta_n}^C(x)=\ell^C(x)$ and the lemma is proved. □

Lemma 62. *Let $\operatorname{supp}(C)=\mathcal{X}$. Then*

$$\lim_{\delta\searrow 0}\hat{\ell}_{\mathcal{P}}(\delta)=\hat{\ell}_{\mathcal{P}}\qquad\textit{and}\qquad\lim_{\delta\searrow 0}\inf_{\hat{q}_{\mathcal{P}}\in S_{\mathcal{P}}}d(\hat{q}_{\mathcal{P}}(\delta),\hat{q}_{\mathcal{P}})=0.$$

Proof. Take any $(\delta_n)_n$ decreasing to 0. Since $(\ell_{\delta_n}^C)_n$ epi-converges to ℓ^C and $\hat{\ell}_{\mathcal{P}}$ is finite, the result follows from [42, Thm. 7.1, p. 264] (in part (a) take $B\triangleq C$; use that $\hat{\ell}_{\mathcal{P}}(\delta_n)=\min\ell_{\delta_n}^C$ and that $\hat{\ell}_{\mathcal{P}}=\min\ell^C$). □

The lemma states that every cluster point of a sequence $(\hat{q}_{\mathcal{P}}(\delta_n))_n$ of solutions of the perturbed primal problems $\mathcal{P}_{\delta_n}$ ($\delta_n \searrow 0$) is a solution of the (unperturbed) primal $\mathcal{P}$. Of course, not every solution of $\mathcal{P}$ can be obtained as a cluster point of a sequence of perturbed solutions. For example, it is shown in the following section that if $(\nu(\delta))_\delta$ satisfies the regularity condition (6.3), then the perturbed solutions converge to a unique solution of $\mathcal{P}$, regardless of whether $S_{\mathcal{P}}$ is a singleton.

Proof of Theorem 20. The first part of Theorem 20 was shown in Lemmas 61 and 62. The second part on the convergence of active coordinates then follows from Theorem 2. □

10.8. Proof of Theorem 21 (Perturbed primal $\mathcal{P}_\delta$ – the linear case, pointwise convergence)

Let a model C be given by finitely many linear constraints (2.3), that is,

$$C = \{q \in \Delta_{\mathcal{X}} : \langle q, u_h \rangle = 0 \text{ for } h = 1, \dots, r\},$$

where u_h ($h = 1, \dots, r$) are fixed vectors from $\mathbb{R}^m$. Let a type $\bar{\nu} \in \Delta_{\mathcal{X}}$ be given and let $\mathcal{X}^a$ and $\mathcal{X}^p$ be the sets of active and passive coordinates with respect to $\bar{\nu}$.

Assume that perturbed types $\nu(\delta)$ ($\delta \in (0,1)$) are such that the conditions (6.1) and (6.3) are true (with ν replaced by $\bar{\nu}$); that is, $\nu : (0,1) \to \Delta_{\mathcal{X}}$ is continuously differentiable,

$$\nu(\delta) > 0, \qquad \lim_{\delta \searrow 0} \nu(\delta) = \bar{\nu},$$

and there is a constant $c > 0$ such that, for every $i \in \mathcal{X}^p$,

$$|\vartheta_i'(\delta)| \le c\, \vartheta_i(\delta), \qquad \text{where} \quad \vartheta_i(\delta) \triangleq \frac{\nu_i(\delta)}{\sum_{j \in \mathcal{X}^p} \nu_j(\delta)}\,.$$

The aim of this section is to prove that, under these conditions, solutions $\hat{q}_{\mathcal{P}}(\delta)$ of the perturbed primal problems $\mathcal{P}_\delta$ *converge* to a solution of the unperturbed primal $\mathcal{P}$; that is,

$$\lim_{\delta \searrow 0} \hat{q}_{\mathcal{P}}(\delta) \quad \text{exists and belongs to } S_{\mathcal{P}}.$$

10.8.1. Outline of the proof

The proof is based on the following 'passive-active' reformulation of the perturbed primal problem

$$\min_{q \in C} \ell_\delta(q) = \min_{q^a \in C^a} \min_{q^p \in C^p(q^a)} \ell_\delta(q^a, q^p);$$

hence, the passive projection $\hat{q}^p_{\mathcal{P}}(\delta)$ of the optimal solution $\hat{q}_{\mathcal{P}}(\delta)$ is

$$\hat{q}^p_{\mathcal{P}}(\delta) = \operatorname*{argmin}_{q^p \in C^p(\hat{q}^a_{\mathcal{P}}(\delta))} \ell_\delta(\hat{q}^a_{\mathcal{P}}(\delta), q^p)$$

(recall that $C^a = \pi^a(C)$ is the projection of C onto the active coordinates and, for $q^a \in C^a$, that $C^p(q^a) = \{q^p \geq 0 : (q^a, q^p) \in C\}$ is the q^a-slice of C). Employing the implicit function theorem, it is then shown that the passive projections $\hat{q}^p_{\mathcal{P}}(\delta)$ can be obtained from the active ones $\hat{q}^a_{\mathcal{P}}(\delta)$ via a uniformly continuous map φ. Since $\hat{q}^a_{\mathcal{P}}(\delta)$ converges by Theorem 20, the uniform continuity of φ ensures that also $\hat{q}^p_{\mathcal{P}}(\delta)$ converges.

The above argument is implemented in the following steps:

1. Assume that $S_{\mathcal{P}}$ is not a singleton, the other case being trivial. Then the linear constraints (2.3) defining C can be rewritten into a parametric form

$$q^p = A\lambda + Bq^a + c \qquad (q^a \in C^a, \lambda \in \Lambda(q^a))$$

(see Lemma 63). There, A is an $m_p \times s$ matrix of rank s, B is an $m_p \times m_a$ matrix, $c \in \mathbb{R}^{m_p}$, none of A, B, c depends on q^a, and the (polyhedral) closed subset $\Lambda(q^a)$ of $\mathbb{R}^s$ is defined by

$$\Lambda(q^a) \triangleq \{\lambda \in \mathbb{R}^s : \ A\lambda + Bq^a + c \geq 0\}.$$

2. Define an open bounded polyhedral set $G \subset \mathbb{R}^{1+m_a+s}$ by

$$G \triangleq \{(\delta, x, y) \in (0,1) \times \mathbb{R}^{m_a} \times \mathbb{R}^s : x > 0, Ay + Bx + c > 0\}; \qquad (10.30)$$

there x stands for q^a and y stands for λ, for brevity. Let Z denote the projection of G onto the first $(1+m_a)$ coordinates. For $z = (\delta, x) \in Z$ put

$$\begin{aligned} G(z) &\triangleq \{y \in \mathbb{R}^s : Ay + Bx + c > 0\}, \\ \bar{G}(z) &\triangleq \{y \in \mathbb{R}^s : Ay + Bx + c \geq 0\}. \end{aligned} \qquad (10.31)$$

In Lemma 65, it is proven that, for every $z = (\delta, x) \in Z$, the map

$$\psi_z : \bar{G}(z) \to \bar{\mathbb{R}}, \qquad \psi_z(y) \triangleq -\sum_{i \in \mathcal{X}^p} \vartheta_i(\delta) \log(\langle \alpha_i, y\rangle + \langle \beta_i, x\rangle + c_i)$$

(where α_i and β_i denote the i-th rows of A and B, respectively) has a unique minimizer $\varphi(z) \in G(z)$.

3. Since $\varphi(z)$ is also a local minimum of ψ_z, it satisfies $F(z, \varphi(z)) = 0$, where $F : G \to \mathbb{R}^s$ is given by

$$F(z, y) \triangleq -\nabla \psi_z(y). \qquad (10.32)$$

Using the implicit function theorem and an algebraic result (Proposition 68), it is shown that $\varphi : Z \to \mathbb{R}^s$ is uniformly continuous (see Lemma 70).

4. Finally, in Lemma 71, it is demonstrated that, for every $\delta > 0$,

$$(\delta, \hat{q}^a_{\mathcal{P}}(\delta)) \in Z \qquad \text{and} \qquad \hat{q}^p_{\mathcal{P}}(\delta) = A\varphi(\delta, \hat{q}^a_{\mathcal{P}}(\delta)) + B\hat{q}^a_{\mathcal{P}}(\delta) + c.$$

This fact, together with the convergence of $\hat{q}^a_{\mathcal{P}}(\delta)$ and the uniform continuity of φ, implies the convergence of $\hat{q}^p_{\mathcal{P}}(\delta)$, and thus proves Theorem 21.

10.8.2. Parametric expression for $q^p \in C^p(q^a)$

Let C be given by (2.3). Denote by U^p the $(r+1) \times m_p$ matrix whose first r rows are equal to u_h^p (the passive projections of u_h, $h = 1, \dots, r$), and the last row is a vector of 1's; analogously define U^a using the active projections u_h^a of u_h. Let $b \in \mathbb{R}^{r+1}$ be such that $b_i = 0$ for $i \le r$ and $b_{r+1} = 1$. Then

$$C = \{(q^a, q^p) \ge 0 : \ U^p q^p = -U^a q^a + b\}. \tag{10.33}$$

Let V^p (of dimension $m_p \times (r+1)$) be the Moore-Penrose inverse of U^p. By [23, p. 5–12], (10.33) can be written in the following form

$$C = \{(q^a, q^p) \ge 0 : \ q^p = (I - V^p U^p)\gamma + Bq^a + c \ \text{ for some } \ \gamma \in \mathbb{R}^{m_p}\},$$

where I denotes the $m_p \times m_p$ identity matrix,

$$B \triangleq -V^p U^a \qquad \text{and} \qquad c \triangleq V^p b.$$

Put $s \triangleq \operatorname{rank}(I - V^p U^p)$.

If $s = 0$ then $(I - V^p U^p) = 0$ and q^p uniquely depends on q^a. That is, for every $q^a \in C^a$, the set $C^p(q^a)$ is a singleton. By Theorems 2 and 20, also $S_{\mathcal{P}}$ is a singleton and $\hat{q}_{\mathcal{P}}(\delta)$ converges to the unique member of $S_{\mathcal{P}}$; so in this case Theorem 21 is proved.

Hereafter, we assume that $s \ge 1$ and, without loss of generality, that the first s columns of $(I - V^p U^p)$ are linearly independent. Put

$$A \triangleq (I - V^p U^p)[\{1, \dots, m_p\}, \{1, \dots, s\}]$$

(the submatrix of entries that lie in the first m_p rows and the first s columns). Since $\{(I - V^p U^p)\gamma : \gamma \in \mathbb{R}^{m_p}\}$ equals $\{A\lambda : \lambda \in \mathbb{R}^s\}$, the next lemma follows.

Lemma 63. *Let C be given by (2.3) and U^p, V^p be as above. Assume that $V^p U^p \neq I$. Then there are $s \ge 1$, an $m_p \times s$ matrix A of rank s, an $m_p \times m_a$ matrix B, and a vector $c \in \mathbb{R}^{m_p}$ such that*

$$C = \{(q^a, L(q^a, \lambda)) : \ q^a \in C^a, \lambda \in \Lambda(q^a)\}.$$

There,

$$L(q^a, \lambda) \triangleq A\lambda + Bq^a + c$$

and

$$\Lambda(q^a) \triangleq \{\lambda \in \mathbb{R}^s : L(q^a, \lambda) \ge 0\}$$

is a nonempty closed polyhedral subset of $\mathbb{R}^s$.

Note that, for $\lambda \neq \lambda'$, $L(q^a, \lambda) \neq L(q^a, \lambda')$ since A has full column rank.

Keep the notation from Lemma 63. Write x for q^a, y for λ, and define $G \subseteq \mathbb{R}^{1+m_a+s}$ by (10.30). Let $\pi_1 : G \to (0,1) \times \mathbb{R}^{m_a}$ be the natural projection mapping (δ, x, y) onto (δ, x). Put

$$Z \triangleq \pi_1(G)$$

and, for $z = (\delta, x) \in Z$, define $G(z), \bar{G}(z)$ by (10.31). Since $G(z)$ is the z-slice of G,

$$G = \{(z, y) : z \in Z, y \in G(z)\}.$$

Note also that, by Lemma 63,

$$\begin{aligned} \Lambda(q^a) &= \bar{G}(\delta, q^a), \\ C^p(q^a) &= A\bar{G}(\delta, q^a) + Bq^a + c, \\ \{q^p \in C^p(q^a) : q^p > 0\} &= AG(\delta, q^a) + Bq^a + c, \end{aligned} \tag{10.34}$$

for every $q^a \in C^a$, $q^a > 0$, and every $\delta \in (0, 1)$. Moreover,

$$\{q \in C : q > 0\} = \{(x, Ay + Bx + c) : (\delta, x, y) \in G \text{ for } \delta \in (0, 1)\}. \tag{10.35}$$

Lemma 64. *The following are true:*

(a) G is a nonempty open bounded polyhedral set;
(b) Z is a nonempty open bounded set;
(c) $G(z)$ ($\bar{G}(z)$) is a nonempty open (closed) bounded polyhedral set for every $z \in Z$.

Proof. (a) The fact that G is open and polyhedral is immediate from (10.30). It is nonempty since $(\delta, q) \in G$ for every $\delta \in (0, 1)$ and every $q \in C$ such that $q > 0$ (such q exists since C has support $\mathcal{X}$). To finish the proof of (a) it remains to show that G is bounded. Since A has full column rank, (10.34) and [23, p. 5–12] yield

$$\begin{aligned} G(\delta, q^a) &= \{A^+(q^p - Bq^a - c) : q^p \in C^p(q^a), q^p > 0\} \\ &\subseteq A^+[0, 1]^{m_p} - A^+(Bq^a + c), \end{aligned}$$

where A^+ is the Moore-Penrose inverse of A. Thus G is bounded.

(b) Since Z is the natural projection of G, it is open, bounded and nonempty.

(c) Boundedness follows from (a); the rest is trivial. □

10.8.3. Global minima of the maps ψ_z ($z \in Z$)

Fix $z = (\delta, x) \in Z$ and define a map $\psi_z : \bar{G}(z) \to \bar{\mathbb{R}}$ by

$$\psi_z(y) \triangleq -\sum_{i \in \mathcal{X}^p} \vartheta_i(\delta) \log d_i(x, y), \qquad d_i(x, y) \triangleq \langle \alpha_i, y \rangle + \langle \beta_i, x \rangle + c_i \tag{10.36}$$

(the convention $\log 0 = -\infty$ applies), where α_i and β_i denote the i-th rows of A and B, respectively. Since ψ_z is differentiable on $G(z)$, we can define a map $F : G \to \mathbb{R}^s$ by (10.32); that is,

$$F_h(\delta, x, y) \triangleq \sum_{i \in \mathcal{X}^p} \frac{a_{ih} \vartheta_i(\delta)}{d_i(x, y)} \qquad \text{for } h = 1, \ldots, s \text{ and } (\delta, x, y) \in G.$$

Easy computation gives (recall that $\nu(\cdot)$, hence also $\vartheta(\cdot)$, is differentiable)

$$\frac{\partial F}{\partial \delta} = A'E1^p, \qquad \frac{\partial F}{\partial x} = -A'DB, \qquad \frac{\partial F}{\partial y} = -A'DA, \tag{10.37}$$

where A' is the transpose of A, and $D = D(\delta, x, y), E = E(\delta, x, y)$ are given by

$$D \triangleq \operatorname{diag}\left(\frac{\vartheta_i(\delta)}{d_i^2(x,y)}\right)_{i=1}^{m_p} \qquad \text{and} \qquad E \triangleq \operatorname{diag}\left(\frac{\vartheta_i'(\delta)}{d_i(x,y)}\right)_{i=1}^{m_p}. \tag{10.38}$$

Lemma 65. *For every $z \in Z$ there is a unique minimizer $\varphi(z) \in G(z)$ of ψ_z:*

$$\operatorname*{argmin}_{\bar{G}(z)} \psi_z = \{\varphi(z)\}.$$

Proof. Let $z = (\delta, x) \in Z$. Since $\bar{G}(z)$ is compact (Lemma 64) and $\psi_z : \bar{G}(z) \to \bar{\mathbb{R}}$ is continuous, argmin ψ_z is nonempty. If $y \in \bar{G}(z) \setminus G(z)$ then there is $1 \leq i \leq m_p$ such that $d_i(x, y) = 0$ and so $\psi_z(y) = \infty$. Hence argmin $\psi_z \subseteq G(z)$. The map ψ_z is strictly convex on $G(z)$ (use [10, Prop. 1.2.6(b)] and the fact that, by (10.37), $\nabla^2\psi_z = A'DA$ is positive definite). Hence the minimum is unique. □

Since $G(z)$ is open for every z, the necessary condition for optimality gives the following result.

Lemma 66. *Let $\varphi : Z \to \mathbb{R}^s$ be as in Lemma 65. Then, for every $(z, y) \in G$,*

$$F(z, y) = 0 \qquad \textit{if and only if} \qquad y = \varphi(z).$$

Since F is continuously differentiable and $\partial F/\partial y = -A'DA$ is always regular, the local implicit function theorem [10, Prop. 1.1.14] and (10.37) immediately imply the next lemma.

Lemma 67. *The map $\varphi : Z \to \mathbb{R}^s$ is continuously differentiable and, for every $z = (\delta, x) \in Z$,*

$$\frac{\partial \varphi}{\partial \delta} = (A'DA)^{-1}A'E1^p \qquad \textit{and} \qquad \frac{\partial \varphi}{\partial x} = -(A'DA)^{-1}A'DB,$$

where $D = D(z, \varphi(z)), E = E(z, \varphi(z))$ are given by (10.38).

In the next section an algebraic result (Proposition 68) implying that φ is Lipschitz on Z (Lemma 70) is proven.

10.8.4. Boundedness of $(A'DA)^{-1}A'D$

Let $\|\cdot\|$ denote the spectral matrix norm [23, p. 37–4], that is, the matrix norm induced by the Euclidean vector norm, which is also denoted by $\|\cdot\|$. If A is a matrix, A' denotes its transpose. The following result must be known, but the authors are not able to give a reference for it.

Proposition 68. *Let $1 \le s \le m$ and A be an $m \times s$ matrix with full column rank* $\operatorname{rank}(A) = s$. *Then there is $\sigma > 0$ such that*

$$\|(A'DA)^{-1}A'D\| \le \sigma$$

for every $m \times m$-diagonal matrix D with positive entries on the diagonal.

Before proving this proposition we give a formula for the inverse of $A'DA$, which is a simple consequence of the Cauchy-Binet formula; cf. [23, p. 4–4]. To this end some notation is needed. If C is an $m \times k$ matrix and $H \subseteq \{1, \dots, m\}$, $K \subseteq \{1, \dots, k\}$ are nonempty, by $C[H, K]$ we denote the submatrix of C of entries that lie in the rows of C indexed by H and the columns indexed by K. C_H, $C^{(j)}$, $C^{(ij)}$, and $C_H^{(j)}$ are shorthands for $C[H, \{1, \dots, k\}]$, $C[\{1, \dots, m\}, \{j\}^c]$, $C[\{i\}^c, \{j\}^c]$, and $C[H, \{j\}^c]$, respectively (there, K^c means the complement of K). For $l \le m$, $\mathcal{H}_{m,l}$ denotes the system of all subsets of $\{1, \dots, m\}$ of cardinality l. Finally, if $(x_i)_{i \in \mathcal{I}}$ is an indexed system of numbers and $I \subseteq \mathcal{I}$ is finite, put

$$x_I \triangleq \prod_{i \in \mathcal{I}} x_i.$$

Lemma 69. *Let $1 \le s \le m$, A be an $m \times s$ matrix with full rank* $\operatorname{rank}(A) = s$, *and $D = \operatorname{diag}(d_1, \dots, d_m)$ be a diagonal matrix with every $d_i > 0$. Put*

$$M \triangleq A'DA.$$

Then M is regular and the following are true:

(a) $\det(M) = \sum_{H \in \mathcal{H}_{m,s}} d_H \aleph_H^2 > 0$,
(b) $\det(M^{(hk)}) = \sum_{I \in \mathcal{H}_{m,s-1}} d_I \aleph_{I,h} \aleph_{I,k}$ *for every $1 \le h, k \le s$,*
(c) $M^{-1} = (c_{hk})_{hk=1}^s$,

where $\aleph_H \triangleq \det(A_H)$, $\aleph_{I,h} \triangleq \det\left(A_I^{(h)}\right)$, *and* $c_{hk} \triangleq (-1)^{h+k} \det(M^{(hk)})/\det(M)$ *for every $H \in \mathcal{H}_{m,s}$, $I \in \mathcal{H}_{m,s-1}$, and $h, k \in \{1, \dots, s\}$.*

Proof. Put $\bar{A} \triangleq DA = (d_i a_{ij})_{ij}$. By the Cauchy-Binet formula,

$$\det(A'\bar{A}) = \sum_{H \in \mathcal{H}_{m,s}} \det(A_H) \det(\bar{A}_H) = \sum_{H \in \mathcal{H}_{m,s}} d_H \aleph_H^2,$$

thus (a) and the regularity of M are proved. Also the property (b) immediately follows from the Cauchy-Binet formula, since $M^{(hk)} = (A^{(h)})'\bar{A}^k$. Finally, (c) follows from the formula $M^{-1} = (1/\det(M)) \operatorname{adj} M$, where $\operatorname{adj} M$ is the adjugate of M, and the fact that M is symmetric. □

Proof of Proposition 68. The proof is by induction on m. If $m = s$ then A is a regular square matrix and $(A'DA)^{-1}A'D = A^{-1}$, so it suffices to put $\sigma \triangleq \|A^{-1}\|$. Assume that $m \ge s$ and that the assertion is true for every matrix A of type $m \times s$. Take any $(m+1) \times s$ matrix $\tilde{A}$ of rank s and any diagonal matrix $\tilde{D} = \operatorname{diag}(d_1, \dots, d_m, \delta)$ with $d_i > 0, \delta > 0$. Since $\|(A'(\theta D)A)^{-1}A'(\theta D)\| =$

$\|(A'DA)^{-1}A'D\|$ for any $\theta > 0$, we may assume that $\delta > 1$ and $d_i > 1$ for every i. Put $D \triangleq \mathrm{diag}(d_1, \dots, d_m)$ and write $\tilde{A}$ in the form

$$\tilde{A} = \begin{pmatrix} A \\ \alpha' \end{pmatrix}$$

with A being an $m \times s$ matrix and $\alpha \in \mathbb{R}^s$. Without loss of generality assume that $\mathrm{rank}(A) = s$. By the induction hypothesis, there is a constant $\sigma > 0$ not depending on D such that $\|(A'DA)^{-1}A'D\| \le \sigma$.

An easy computation gives

$$\tilde{A}'\tilde{D}\tilde{A} = M + \delta\alpha\alpha', \qquad \text{where } M \triangleq A'DA.$$

By the Sherman-Morrison formula [23, p. 14–15], for any $u, v \in \mathbb{R}^s$, it holds that

$$(M + uv')^{-1} = M^{-1} - \frac{1}{1 + v'M^{-1}u} M^{-1}uv'M^{-1}.$$

Hence

$$(\tilde{A}'\tilde{D}\tilde{A})^{-1} = M^{-1} - \varsigma M^{-1}\alpha\alpha'M^{-1} \qquad \left(\varsigma \triangleq \frac{\delta}{1 + \delta\alpha'M^{-1}\alpha}\right)$$

and

$$(\tilde{A}'\tilde{D}\tilde{A})^{-1}\tilde{A}'\tilde{D} = (M^{-1}A'D - \varsigma M^{-1}\alpha\alpha'M^{-1}A'D,\ \varsigma M^{-1}\alpha).$$

To finish the proof it suffices to show that $\|\varsigma M^{-1}\alpha\|$ is bounded by a constant σ' not depending on D, δ. (Indeed, if this is true then the norm of $(\tilde{A}'\tilde{D}\tilde{A})^{-1}\tilde{A}'\tilde{D}$ is bounded from above by $(\sigma + \sigma'\|\alpha\|\sigma) + \sigma'$; it follows from the matrix norm triangle inequality applied to $(E, F) = (E, 0_{s\times 1}) + (0_{s\times m}, F)$, and from the matrix norm consistency property; cf. [23, p. 37–4].)

If $\alpha = 0$, then $\varsigma M^{-1}\alpha = 0$, so one can take $\sigma' = 0$. Assume now that $\alpha = (\alpha_h)_h \neq 0$. Using the notation from Lemma 69,

$$\begin{aligned} \alpha'M^{-1}\alpha &= \sum_{h,k=1}^{s} c_{hk}\alpha_h\alpha_k \\ &= \frac{1}{\det(M)} \sum_{h,k=1}^{s} \sum_{I \in \mathcal{H}_{m,s-1}} (-1)^{h+k}\alpha_h\alpha_k d_I \aleph_{I,h}\aleph_{I,k} \\ &= \frac{1}{\det(M)} \sum_{I \in \mathcal{H}_{m,s-1}} d_I\varepsilon_I^2, \qquad \text{where} \quad \varepsilon_I \triangleq \sum_{k=1}^{s} (-1)^k\alpha_k\aleph_{I,k}. \end{aligned}$$

On the other hand,

$$\begin{aligned} (M^{-1}\alpha)_h &= \sum_{k=1}^{s} c_{hk}\alpha_k \\ &= \frac{1}{\det(M)} \sum_{k=1}^{s} \sum_{I \in \mathcal{H}_{m,s-1}} (-1)^{h+k}\alpha_k d_I\aleph_{I,h}\aleph_{I,k} \\ &= \frac{(-1)^h}{\det(M)} \sum_{I \in \mathcal{H}_{m,s-1}} d_I\varepsilon_I\aleph_{I,h}. \end{aligned}$$

Thus, the h-th coordinate of $v \triangleq \varsigma M^{-1}\alpha$ satisfies

$$
\begin{aligned}
|v_h| &= \left| \sum_{I\in\mathcal{H}_{m,s-1}} \frac{d_I \varepsilon_I \aleph_{I,h}}{\frac{1}{\delta}\det(M) + \sum_{J\in\mathcal{H}_{m,s-1}} d_J \varepsilon_J^2} \right| \\
&\le \sum_{I\in\mathcal{H}_{m,s-1}} \frac{d_I |\varepsilon_I \aleph_{I,h}|}{\frac{1}{\delta}\det(M) + \sum_{J\in\mathcal{H}_{m,s-1}} d_J \varepsilon_J^2} \\
&\le \sum_{I\in\mathcal{H}_{m,s-1}} \frac{d_I |\varepsilon_I \aleph_{I,h}|}{\frac{1}{\delta}\det(M) + d_I \varepsilon_I^2} \\
&\le \sum_{I\in\mathcal{H}_{m,s-1},\, \varepsilon_I\neq 0} \frac{|\aleph_{I,h}|}{|\varepsilon_I|}\,.
\end{aligned}
$$

This proves that the absolute value of every coordinate of $v = \varsigma M^{-1}\alpha$ is bounded from above by a constant not depending on D, δ. It finishes the proof of Proposition 68. □

10.8.5. Lipschitz property of φ

The result of the previous section and Lemma 67 yield that φ is Lipschitz, hence uniformly continuous on Z.

Lemma 70. *The map $\varphi : Z \to \mathbb{R}^s$ is Lipschitz on Z. Consequently, there exists a continuous extension $\bar{\varphi} : \bar{Z} \to \mathbb{R}^s$ of φ to the closure $\bar{Z}$ of Z.*

Proof. By (6.3), the norm of $D^{-1}E = \operatorname{diag}(d_i(x,y)\vartheta_i'(\delta)/\vartheta_i(\delta))_{i=1}^{m_p}$ is bounded from above by a constant not depending on $z = (\delta, x)$ and y (use that $0 < d_i \le 1$ by (10.35)). Thus, by Proposition 68 and Lemma 67, φ has bounded derivative on Z. Now the Lipschitz property of φ follows from the mean value theorem. □

10.8.6. Proof of Theorem 21

The following lemma demonstrates that, for $\delta > 0$, the passive projection $\hat{q}^p_{\mathcal{P}}(\delta)$ of the solution of $\mathcal{P}_\delta$ can be expressed via φ and the active projection $\hat{q}^a_{\mathcal{P}}(\delta)$.

Lemma 71. *Let C be given by (2.3) and A, B, c be given by Lemma 63. Let $(\nu(\delta))_{\delta\in(0,1)}$ satisfy (6.1) and (6.3). Then, for every $\delta > 0$,*

$$(\delta, \hat{q}^a_{\mathcal{P}}(\delta)) \in Z \qquad \textit{and} \qquad \hat{q}^p_{\mathcal{P}}(\delta) = A\varphi(\delta, \hat{q}^a_{\mathcal{P}}(\delta)) + B\hat{q}^a_{\mathcal{P}}(\delta) + c.$$

Proof. Fix any $\delta > 0$ and put $x \triangleq \hat{q}^a_{\mathcal{P}}(\delta)$, $z \triangleq (\delta, x)$. Let $\hat{y} \in \Lambda(x)$ be such that $\hat{q}^p_{\mathcal{P}}(\delta) = A\hat{y} + Bx + c$ (note that $\hat{y}$ is unique since A has full column rank; see Lemma 63). Then $z \in Z$ since $(\delta, x, \hat{y})$ belongs to G (use that $\hat{q}^p_{\mathcal{P}}(\delta) > 0$). Further, $G(z)$ is a subset of $\Lambda(x)$ by (10.34), and $\ell_\delta(x, Ay + Bx + c) = \infty$ for every $y \in \Lambda(x) \setminus G(z)$ (indeed, for such y some coordinate of $Ay + Bx + c$ is zero). Thus $\hat{y} \in G(z)$ and, by (10.36) and (6.3),

$$\begin{aligned}\hat{y} &= \operatorname*{argmin}_{y \in G(z)} \ell_\delta(x, Ay + Bx + c) \\ &= \operatorname*{argmin}_{y \in G(z)} \left(-\sum_{i \in \mathcal{X}^p} \nu_i(\delta) \log d_i(x, y) \right) \\ &= \operatorname*{argmin}_{y \in G(z)} \psi_z(y).\end{aligned}$$

Hence, by Lemma 65, $\hat{y} = \varphi(z)$ and so $\hat{q}^p_{\mathcal{P}}(\delta) = A\varphi(z) + Bx + c$. □

Proof of Theorem 21. Let $\bar{\varphi} : \bar{Z} \to \mathbb{R}^s$ denote the continuous extension of φ to the closure $\bar{Z}$ of Z (see Lemma 70). By Theorem 20, $\lim_{\delta \searrow 0} \hat{q}^a_{\mathcal{P}}(\delta)$ exists and equals the unique member $\hat{q}^a_{\mathcal{P}}$ of $S_{\mathcal{P}}$. Since $(\delta, \hat{q}^a_{\mathcal{P}}(\delta)) \in Z$ for every $\delta > 0$ (see Lemma 71), $(0, \hat{q}^a_{\mathcal{P}})$ belongs to $\bar{Z}$. Thus, by Lemma 71,

$$\begin{aligned}\lim_{\delta \searrow 0} \hat{q}^p_{\mathcal{P}}(\delta) &= \lim_{\delta \searrow 0} \left(A\varphi(\delta, \hat{q}^a_{\mathcal{P}}(\delta)) + B\hat{q}^a_{\mathcal{P}}(\delta) + c\right) \\ &= A\bar{\varphi}(0, \hat{q}^a_{\mathcal{P}}) + B\hat{q}^a_{\mathcal{P}} + c.\end{aligned}$$

This proves convergence of $\hat{q}^p_{\mathcal{P}}(\delta)$. Now, by Theorem 20, Theorem 21 follows. □

Acknowledgements

Substantive feedback from Sanjay Chaudhuri, Bing-yi Jing, Jana Majerová, Long Nguyen, and Danny Pfeffermann is gratefully acknowledged. The thanks extend to a two anonymous reviewers and an Associate Editor. This research is an outgrowth of the project "SPAMIA", MŠ SR-3709/2010-11, supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic, under the heading of the state budget support for research and development. The first author acknowledges support from VEGA 2/0038/12, APVV-0096-10, VEGA 2/0047/15, CZ.1.07/2.3.00/20.0170 grants and by "Biomedical Center Martin", ITMS code 262202220187, co-financed from EU sources. The second author acknowledges support from VEGA 1/0786/15 and APVV-15-0439 grants. Research was partially completed while Marian Grendár was visiting the Institute for Mathematical Sciences, National University of Singapore in 2016.

Supplementary Material

R code and data to reproduce the numerical examples
(doi: 10.1214/17-EJS1294SUPP; .zip).

References

[1] Agresti, A. (2002). *Categorical data analysis*, Second ed. Wiley-Interscience [John Wiley & Sons], New York. MR1914507

[2] Agresti, A. and Coull, B. A. (2002). The analysis of contingency tables under inequality constraints. *J. Statist. Plann. Inference* **107** 45–73. MR1927754

[3] Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman & Hall, London. MR865647

[4] Aitchison, J. and Silvey, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.* **29** 813–828. MR0094873

[5] Anaya-Izquierdo, K., Critchley, F., Marriott, P. and Vos, P. W. Computational information geometry: theory and practice. *arxiv:1209.1988*.

[6] Baker, R., Clarke, M. and Lane, P. (1985). Zero entries in contingency tables. *Comput. Statist. Data Anal.* **3** 33–45.

[7] Balabdaoui, F. and Jankowski, H. (2016). Maximum likelihood estimation of a unimodal probability mass function. *Stat. Sin.* **26** 1061–1086. MR3559943

[8] Bergsma, W., Croon, M. and van der Ark, L. A. (2012). The empty set and zero likelihood problems in maximum empirical likelihood estimation. *Electron. J. Stat.* **6** 2356–2361. MR3020267

[9] Bergsma, W., Croon, M. A. and Hagenaars, J. A. (2009). *Marginal models*. Springer.

[10] Bertsekas, D. P. (2003). *Convex analysis and optimization*. Athena Scientific, Belmont, MA With Angelia Nedić and Asuman E. Ozdaglar. MR2184037

[11] Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete multivariate analysis: theory and practice*. The MIT Press, Cambridge, Mass.-London With the collaboration of Richard J. Light and Frederick Mosteller. MR0381130

[12] El Barmi, H. and Dykstra, R. L. (1994). Restricted multinomial maximum likelihood estimation based upon Fenchel duality. *Statist. Probab. Lett.* **21** 121–130. MR1309885

[13] El Barmi, H. and Dykstra, R. L. (1998). Maximum likelihood estimates via duality for log-convex models when cell probabilities are subject to convex constraints. *Ann. Statist.* **26** 1878–1893. MR1673282

[14] Fienberg, S. E. and Rinaldo, A. (2012). Maximum likelihood estimation in log-linear models. *Ann. Statist.* **40** 996–1023. MR2985941

[15] Fisher, R. A. (1925). Theory of statistical estimation. *Math. Proc. Cambridge Philos. Soc.* **22** 700–725.

[16] Geyer, C. J. (2009). Likelihood inference in exponential families and directions of recession. *Electron. J. Stat.* **3** 259–289. MR2495839

[17] Ghalanos, A. and Theussl, S. (2012). *Rsolnp: general non-linear optimization using augmented Lagrange multiplier method*. R package version 1.14.

[18] Gokhale, D. V. (1973). Iterative maximum likelihood estimation for discrete distributions. *Sankhyā Ser. B* **35** 293–298. MR0415851

[19] Grendár, M. and Judge, G. (2009). Empty set problem of maximum empirical likelihood methods. *Electron. J. Stat.* **3** 1542–1555. MR2578837
[20] Grendár, M. and Špitalský, V. (2014). Supplement to "Multinomial and empirical likelihood under convex constraints: directions of recession, Fenchel duality, the PP algorithm". DOI: 10.1214/17-EJS1294SUPP
[21] Haber, M. (1985). Maximum likelihood methods for linear and log-linear models in categorical data. *Comput. Statist. Data Anal.* **3** 1–10. MR812132
[22] Hall, P. and Presnell, B. (1999). Intentionally biased bootstrap methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 143–158. MR1664116
[23] Hogben, L., ed. (2007). *Handbook of linear algebra*. *Discrete Mathematics and Its Applications (Boca Raton)*. Chapman & Hall/CRC, Boca Raton, FL. MR2279160
[24] Ireland, C. T., Ku, H. H. and Kullback, S. (1969). Symmetry and marginal homogeneity of an $r \times r$ contingency table. *J. Amer. Statist. Assoc.* **64** 1323–1341. MR0251877
[25] Ireland, C. T. and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika* **55** 179–188. MR0229329
[26] Kall, P. (1986). Approximation to optimization problems: an elementary review. *Math. Oper. Res.* **11** 9–18. MR830103
[27] Kerridge, D. F. (1961). Inaccuracy and inference. *J. Roy. Statist. Soc. Ser. B* **23** 184–194. MR0123375
[28] Klotz, J. H. (1978). Testing a linear constraint for multinomial cell frequencies and disease screening. *Ann. Statist.* **6** 904–909. MR0494636
[29] Lang, J. B. (2004). Multinomial-Poisson homogeneous models for contingency tables. *Ann. Statist.* **32** 340–383. MR2051011
[30] Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics* i–163.
[31] Lindsey, J. K. (1996). *Parametric statistical inference*. The Clarendon Press, Oxford University Press, New York. MR1420412
[32] Little, R. J. and Wu, M.-M. (1991). Models for contingency tables with known margins when target and sampled populations differ. *J. Amer. Statist. Assoc.* **86** 87–95.
[33] Madansky, A. (1963). Tests of homogeneity for correlated samples. *J. Amer. Statist. Assoc.* **58** 97–119. MR0149613
[34] Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249. MR946049
[35] Owen, A. B. (2001). *Empirical likelihood*. Chapman & Hall/CRC.
[36] Pelz, W. and Good, I. J. (1986). Estimating probabilities from contingency tables when the marginal probabilities are known, by using additive objective functions. *Statistician* **35** 45–50.
[37] Pitman, E. J. G. (1979). *Some basic theory for statistical inference*. Chapman and Hall, London; A Halsted Press Book, John Wiley & Sons, New York. MR549771
[38] Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22** 300–325. MR1272085

[39] R-Core-Team (2012). R: a language and environment for statistical computing R Foundation for Statistical Computing, Vienna, Austria.

[40] Rockafellar, R. T. (1974). *Conjugate duality and optimization*. Society for Industrial and Applied Mathematics, Philadelphia, Pa. MR0373611

[41] Rockafellar, R. T. (1997). *Convex analysis*. Princeton University Press, Princeton, NJ. MR1451876

[42] Rockafellar, R. T. and Wets, R. J. B. (1998). *Variational analysis* **317**. Springer-Verlag, Berlin. MR1491362

[43] Sharma, P., Hawes, R. H., Bansal, A., Gupta, N., Curvers, W., Rastogi, A., Singh, M., Hall, M., Mathur, S. C., Wani, S. B. et al. (2013). Standard endoscopy with random biopsies versus narrow band imaging targeted biopsies in Barrett's oesophagus: a prospective, international, randomised controlled trial. *Gut* **62** 15–21.

[44] Silvey, S. D. (1959). The Lagrangian multiplier test. *Ann. Math. Statist.* **30** 389–407. MR0104307

[45] Smith, J. H. (1947). Estimation of linear functions of cell proportions. *Ann. Math. Statist.* **18** 231–254. MR0020760

[46] Stirling, W. D. (1986). Testing linear hypotheses in contingency tables with zero cell counts. *Comput. Statist. Data Anal.* **4** 1–13. MR2240855

[47] Wets, R. J. B. (1999). Statistical estimation from an optimization viewpoint. *Ann. Oper. Res.* **85** 79–101. MR1685588

[48] Ye, Y. (1987). Interior algorithms for linear, quadratic, and linearly constrained non-linear programming PhD thesis, Department of ESS, Stanford University.

[49] Zhang, B. (1999). Bootstrapping with auxiliary information. *Canad. J. Statist.* **27** 237–249. MR1704411

[50] Zhang, Z. (2009). Interpreting statistical evidence with empirical likelihood functions. *Biom. J.* **51** 710–720. MR2744088