

DISCRETE CHOICE PROX-FUNCTIONS ON THE SIMPLEX

David Müller, Yurii Nesterov, Vladimir
Shikhman

REPRINT | 3242

CORE

Voie du Roman Pays 34, L1.03.01

B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email: lidam-library@uclouvain.be

<https://uclouvain.be/en/research-institutes/lidam/core/core-reprints.html>

Discrete choice prox-functions on the simplex

D. Müller *

Yu. Nesterov[†]

V. Shikhman*

September 13, 2019

Abstract

We derive new prox-functions on the simplex from additive random utility models of discrete choice. They are convex conjugates of the corresponding surplus functions. In particular, we explicitly derive the convexity parameter of discrete choice prox-functions associated with generalized extreme value models, and specifically with generalized nested logit models. Incorporated into subgradient schemes, discrete choice prox-functions lead to natural probabilistic interpretations of the iteration steps. As illustration we discuss an economic application of discrete choice prox-functions in consumer theory. The dual averaging scheme from convex programming naturally adjusts demand within a consumption cycle.

Keywords: convex programming, prox-function, discrete choice, additive random utility models, dual averaging, consumption cycle.

1 Introduction

The use of prox-functions in convex programming has become standard in recent decades. Originally, they were introduced in the context of mirror descent methods [12]. In order to explain how prox-functions enter into optimization methods, we recall their definition.

Definition 1 (Prox-function) *We say that $d : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ is a prox-function on a closed convex set $Q \subset \mathbb{R}^n$ if*

- (1) *d is continuous with the domain containing Q , i. e. $Q \subset \text{dom } d$.*
- (2) *d is strongly convex on Q with respect to a norm $\|\cdot\|$, i. e. there exists a constant $\beta > 0$ such that for all $x, y \in Q$ and $\alpha \in [0, 1]$ it holds:*

$$d(\alpha x + (1 - \alpha)y) \leq \alpha d(x) + (1 - \alpha)d(y) - \frac{\beta}{2}\alpha(1 - \alpha)\|x - y\|^2.$$

*Department of Mathematics, Chemnitz University of Technology, Reichenhainer Str. 41, 09126 Chemnitz, Germany; e-mail: david.mueller@mathematik.tu-chemnitz.de, vladimir.shikhman@mathematik.tu-chemnitz.de.

[†]Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium, and National Research University – Higher School of Economics, Russia; e-mail: yurii.nesterov@uclouvain.be.

(3) *The computation of the convex conjugate*

$$d^*(s) = \max_{x \in Q} \langle s, x \rangle - d(x) \quad (\text{A})$$

is simple, i. e. the unique maximizer $x(s)$ can be easily obtained for any $s \in \mathbb{R}^n$.

Auxiliary optimization problem (A) is known to be a key ingredient for minimizing a convex function f on Q by subgradient schemes. Let us take for s subgradients of f or their weighted aggregates. Then, (A) defines a mapping from the dual space with subgradient information s into the primal space with feasible iterates $x(s) \in Q$. This idea leads in particular to primal-dual subgradient methods for convex problems [13].

There are at least two advantages of using prox-functions in (A):

- (a) It is well known that the complexity bounds for optimization methods heavily depend on the size of the feasible set Q . This value has been traditionally defined with respect to Euclidean norm. However, the size of Q , measured with respect to another norm, can be smaller. Thus, by introducing prox-functions, which are strongly convex with respect to an appropriate norm $\|\cdot\|$, it is possible to take into account a particular geometry of the feasible set Q .
- (b) More interestingly, prox-functions often allow natural interpretations of the iteration steps (A) within the convex optimization framework. This feature is important in order to explain agents' behavioral dynamics as being driven by unintentional optimization. Let us illustrate this for the well known entropic prox-function

$$d(p) = \sum_{i=1}^n p^{(i)} \ln p^{(i)}$$

on the $(n-1)$ -dimensional simplex

$$\Delta = \left\{ p \in \mathbb{R}^n \left| \sum_{i=1}^n p^{(i)} = 1, p^{(i)} \geq 0 \right. \right\}.$$

For the feasible set $Q = \Delta$ the auxiliary optimization problem (A) reads:

$$\max_{p \in \Delta} \sum_{i=1}^n s^{(i)} p^{(i)} - \sum_{i=1}^n p^{(i)} \ln p^{(i)}.$$

Its unique solution is given by

$$p^{(i)}(s) = \frac{e^{s^{(i)}}}{\sum_{i=1}^n e^{s^{(i)}}}, \quad i = 1, \dots, n.$$

This formula is in accordance with the logit model of discrete choice: $p^{(i)}(s)$ can be viewed as the choice probability of detecting $s^{(i)}$ to be maximal among $s^{(1)}, \dots, s^{(n)}$.

In this paper we introduce prox-functions on the simplex which allow similar probabilistic interpretation as above. They are derived from the additive random utility models of discrete choice [2]. Note that the logit model is just one prominent example within this class. Our main result in Section 2 states that the convex conjugate of the surplus function, associated with an additive random utility model, is a prox-function on the simplex. For the convex conjugate of the corresponding surplus function we show continuity (Section 2.2.1), strong convexity (Section 2.2.2), and simplicity (Section 2.2.3). In particular, we explicitly derive the convexity parameter for the class of generalized extreme value models [9], and specifically of generalized nested logit models [18]. Section 3 is devoted to an economic application of discrete choice prox-functions in consumer theory. The discrete choice prox-functions are incorporated into the dual averaging scheme from [13] for consumer's utility maximization. This ensures that the update of internal prices for goods' qualities is due to an additive random utility model. The dual averaging scheme corresponds to a natural consumption cycle which successively leads to an optimal consumption of goods (Section 3.3). We mention that the proposed consumption cycle generalizes [14] where the update of internal prices is due to the logit model.

Notation. Our notation is quite standard. We denote by $\mathbb{R}^n$ the space of n -dimensional column vectors $x = (x^{(1)}, \dots, x^{(n)})^T$, by $\mathbb{R}_+^n$ the set of all vectors with nonnegative components. If the components of $x \in \mathbb{R}^n$ are nonnegative (positive), we write $x \geq 0$ ($x > 0$). For $x \in \mathbb{R}^n$ we write $x^{(-i)} \in \mathbb{R}^{n-1}$ meaning that the i -th component of x is missing. Analogously, we write $x^{(-i,j)} \in \mathbb{R}^{n-2}$ meaning that both i -th and j -th components of x are missing. For $x, y \in \mathbb{R}^n$ we introduce the standard scalar product and (if additionally $y > 0$) the vector division:

$$\langle x, y \rangle = \sum_{i=1}^n x^{(i)} y^{(i)}, \quad \frac{x}{y} = \left(\frac{x^{(1)}}{y^{(1)}}, \dots, \frac{x^{(n)}}{y^{(n)}} \right)^T.$$

For $x \in \mathbb{R}^n$ we use the following norms:

$$\|x\|_1 = \sum_{i=1}^n |x^{(i)}|, \quad \|x\|_\infty = \max_{1 \leq i \leq n} |x^{(i)}|.$$

Note that they are dual to each other, i. e.

$$\|x\|_\infty = \sup_{\|y\|_1 \leq 1} \langle y, x \rangle, \quad \|x\|_1 = \sup_{\|y\|_\infty \leq 1} \langle x, y \rangle.$$

We denote by $e_j \in \mathbb{R}^n$ the j -th coordinate vector of $\mathbb{R}^n$. All components of the vector $e \in \mathbb{R}^n$ are equal to one. The space of $(n \times n)$ -matrices with real-valued entries is denoted by $\mathbb{R}^{n \times n}$. We use the induced matrix norm for $A \in \mathbb{R}^{n \times n}$:

$$\|A\|_{\infty,1} = \max_{\|z\|_\infty \leq 1} \|Az\|_1.$$

Given a twice differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, ∇f denotes the gradient and $\nabla^2 f$ stands for the Hessian matrix.

2 Discrete choice prox-functions on the simplex

We derive discrete choice prox-functions on the simplex from additive random utility models.

2.1 Additive random utility models

The additive random utility framework has been first introduced in economic context [9]. It aims to model the discrete choice from a finite number of alternatives $\{1, \dots, n\}$ by a rational decision-maker prone to some random errors. Accordingly, the i -th alternative is endowed with the utility

$$u^{(i)} + \epsilon^{(i)},$$

where $u^{(i)} \in \mathbb{R}$ is its deterministic part and $\epsilon^{(i)}$ is a random error. We denote by

$$u = \left(u^{(1)}, \dots, u^{(n)}\right)^T, \quad \epsilon = \left(\epsilon^{(1)}, \dots, \epsilon^{(n)}\right)^T$$

the vectors of deterministic utilities and of random utility shocks, respectively. The following assumption on the stochastic errors is standard, see e. g. [2].

Assumption 1 *The random vector ϵ follows a joint distribution with finite mean that is absolutely continuous with respect to the Lebesgue measure and fully supported on $\mathbb{R}^n$.*

Since a rational decision-maker chooses alternatives with the maximal utility, the corresponding surplus is given by the expectation

$$E(u) = \mathbb{E}_\epsilon \left(\max_{1 \leq i \leq n} u^{(i)} + \epsilon^{(i)} \right).$$

It is well-known that the surplus function E is convex and differentiable [2]. In particular, its partial derivatives can be expressed as choice probabilities:

$$\frac{\partial E(u)}{\partial u^{(i)}} = \mathbb{P} \left(u^{(i)} + \epsilon^{(i)} = \max_{1 \leq i \leq n} u^{(i)} + \epsilon^{(i)} \right), \quad i = 1, \dots, n. \quad (1)$$

The latter means that the i -th partial derivative of E corresponds to the probability of perceiving the i -th alternative as one with the maximal utility among the others. This result is known as the Williams-Daly-Zachary theorem in the discrete choice literature [9, 10]. The formula (1) is valid due to the fact that, under Assumption 1, the ties between the alternatives occur with zero-probability, i. e.

$$\mathbb{P} \left(\epsilon^{(i)} - \epsilon^{(j)} = c \right) = 0 \quad \text{for all } i \neq j \text{ and } c \in \mathbb{R}.$$

2.2 Convex conjugate of the surplus function

We turn our attention to the convex conjugate $E^* : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ of the surplus function:

$$E^*(p) = \sup_{u \in \mathbb{R}^n} \langle p, u \rangle - E(u),$$

where $p = (p^{(1)}, \dots, p^{(n)})^T \in \mathbb{R}^n$ is the vector of dual variables.

2.2.1 Continuity

We discuss the continuity of the convex conjugate E^* on its domain

$$\text{dom } E^* = \{p \in \mathbb{R}^n \mid E^*(p) < \infty\}.$$

For that, we need some elementary properties of the surplus function E listed below.

Lemma 1 (Elementary properties of E) *For the surplus function E it holds:*

(E1) $E(u + \gamma e) = E(u) + \gamma$ for all $\gamma \in \mathbb{R}, u \in \mathbb{R}^n$.

(E2) $E(u) \geq E(v)$ for all $u, v \in \mathbb{R}^n$ with $u \geq v$.

(E3) $E(u) \geq \max_{1 \leq i \leq n} u^{(i)} + \min_{1 \leq i \leq n} \mathbb{E}_\epsilon \left(\epsilon^{(i)} \right)$ for all $u \in \mathbb{R}^n$.

Proof:

(E1) The linearity of the expectation provides:

$$E(u + \gamma e) = \mathbb{E}_\epsilon \left(\max_{1 \leq i \leq n} \left(u^{(i)} + \gamma + \epsilon^{(i)} \right) \right) = \mathbb{E}_\epsilon \left(\max_{1 \leq i \leq n} \left(u^{(i)} + \epsilon^{(i)} \right) + \gamma \right) = E(u) + \gamma.$$

(E2) The monotonicity of the expectation provides:

$$E(u) = \mathbb{E}_\epsilon \left(\max_{1 \leq i \leq n} u^{(i)} + \epsilon^{(i)} \right) \geq \mathbb{E}_\epsilon \left(\max_{1 \leq i \leq n} v^{(i)} + \epsilon^{(i)} \right) = E(v).$$

(E3) Due to the finite mean condition from Assumption 1, we have for every $i \in \{1, \dots, n\}$:

$$E(u) = \mathbb{E}_\epsilon \left(\max_{1 \leq i \leq n} u^{(i)} + \epsilon^{(i)} \right) \geq \mathbb{E}_\epsilon \left(u^{(i)} + \epsilon^{(i)} \right) \geq u^{(i)} + \min_{1 \leq i \leq n} \mathbb{E}_\epsilon \left(\epsilon^{(i)} \right).$$

□

Theorem 1 (Continuity of E^*) *The convex conjugate E^* is continuous on its domain $\text{dom } E^*$ which coincides with the simplex Δ .*

Proof:

Let us first show that $\text{dom } E^* \subseteq \Delta$. For $p \in \mathbb{R}^n$ with $\langle p, e \rangle \neq 1$ we have:

$$E^*(p) \geq \sup_{\gamma \in \mathbb{R}} \langle p, v + \gamma e \rangle - E(v + \gamma e) \stackrel{\text{(E1)}}{=} \langle p, v \rangle - E(v) + \sup_{\gamma \in \mathbb{R}} \gamma (\langle p, e \rangle - 1) = \infty,$$

where $v \in \mathbb{R}^n$ is fixed. For $p \in \mathbb{R}^n$ with $p^{(i)} < 0$ for an $i \in \{1, \dots, n\}$ we have:

$$E^*(p) \geq \sup_{\gamma \leq 0} \langle p, \gamma e_i \rangle - E(\gamma e_i) \stackrel{\text{(E2)}}{\geq} \sup_{\gamma \leq 0} \gamma p^{(i)} - E(0) = \infty,$$

where e_i denotes the i -th coordinate vector. Secondly, we prove that $\text{dom } E^* \supseteq \Delta$. For that, it is sufficient to show that E^* is bounded from above on Δ . Due to (E3) from Lemma 1, it holds:

$$\begin{aligned} \sup_{p \in \Delta} E^*(p) &= \sup_{p \in \Delta} \left(\sup_{u \in \mathbb{R}^n} \langle p, u \rangle - E(u) \right) = \sup_{u \in \mathbb{R}^n} \left(\sup_{p \in \Delta} \langle p, u \rangle - E(u) \right) \\ &= \sup_{u \in \mathbb{R}^n} \left(\max_{1 \leq i \leq n} u^{(i)} - E(u) \right) \leq - \min_{1 \leq i \leq n} \mathbb{E}_\epsilon \left(\epsilon^{(i)} \right). \end{aligned} \quad (2)$$

Further, we discuss the continuity of E^* on the simplex Δ . Since E^* is convex, it is continuous on the relative interior $\text{rint}(\Delta)$ of its domain. The continuity of E^* on the whole domain Δ can be deduced by an application of the Gale-Klee-Rockafellar theorem. The Gale-Klee-Rockafellar theorem says that a convex function is upper semi-continuous at every point at which its domain is polyhedral [5]. Note that the domain of E^* – the $(n-1)$ -dimensional simplex Δ – is polyhedral. Moreover, the convex conjugate of a function is always lower semi-continuous, so is E^* on Δ . Together, the lower and upper semi-continuity of E^* on Δ provides the claim. $\square$

Theorem 1 says that the convex conjugate E^* is finite on the simplex Δ . The latter can be viewed as the set of probability distributions. Hence, the dual variables p can be interpreted as the probabilities attached to the alternatives $\{1, \dots, n\}$.

Corollary 1 (Upper bound for E^*) *The convex conjugate E^* is bounded from above on its domain Δ , namely it holds:*

$$E^*(p) \leq - \min_{1 \leq i \leq n} \mathbb{E}_\epsilon \left(\epsilon^{(i)} \right) \quad \text{for all } p \in \Delta.$$

Proof:

The assertion follows from the derivation in (2). $\square$

2.2.2 Strong Convexity

We show that the convex conjugate E^* is strongly convex under suitable assumptions, and estimate its convexity parameter.

Definition 2 (Strong convexity of E^*) *The convex conjugate $E^* : \Delta \rightarrow \mathbb{R}$ is β -strongly convex with respect to the $\|\cdot\|_1$ norm if for all $p, q \in \Delta$ and $\alpha \in [0, 1]$ we have:*

$$E^*(\alpha p + (1 - \alpha)q) \leq \alpha E^*(p) + (1 - \alpha)E^*(q) - \frac{\beta}{2} \alpha(1 - \alpha) \|p - q\|_1^2.$$

The positive constant β is called the convexity parameter of E^ .*

The strong convexity of E^* is closely related to the strong smoothness of E .

Definition 3 (Strong smoothness of E) The surplus function $E : \mathbb{R}^n \rightarrow \mathbb{R}$ is L -strongly smooth with respect to the maximum norm $\|\cdot\|_\infty$ if for all $u, v \in \mathbb{R}^n$ we have:

$$E(u + v) \leq E(u) + \langle \nabla E(u), v \rangle + \frac{L}{2} \|v\|_\infty^2.$$

The positive constant L is called the smoothness parameter of E .

The following duality result between the strong convexity of E^* and the strong smoothness of E can be easily deduced from [6, Theorem 6] shown there in the general setting.

Lemma 2 (Strong convex/smooth duality) The convex conjugate E^* is β -strongly convex with respect to the $\|\cdot\|_1$ norm if and only if the surplus function E is $\frac{1}{\beta}$ -strongly smooth with respect to the maximum norm $\|\cdot\|_\infty$.

Proof:

We apply [6, Theorem 6] which says that a closed and convex function is β -strongly convex with respect to a norm if and only if its convex conjugate is $\frac{1}{\beta}$ -strongly smooth with respect to the dual norm. For that, we note that E^* is proper and lower semi-continuous, hence, closed. Moreover, by the Fenchel-Moreau theorem we have

$$E^{**} = E,$$

since E is, in particular, a proper, lower semi-continuous, and convex function. Finally, the dual of the $\|\cdot\|_1$ norm is the maximum norm $\|\cdot\|_\infty$. $\square$

In view of Lemma 2, we may focus on the strong smoothness of E . For the characterization of the latter property, we use the fact that the surplus function E is twice differentiable. Let us compute the second order partial derivatives of E . Recall that its i -th partial derivative can be written as the choice probability

$$\begin{aligned} \frac{\partial E(u)}{\partial u^{(i)}} &= \mathbb{P} \left(\epsilon^{(-i)} - \epsilon^{(i)} \leq u^{(i)} - u^{(-i)} \right) \\ &= \int_{-\infty}^{u^{(i)} - u^{(-i)}} \int_{-\infty}^{\infty} f_\epsilon \left(y^{(-i)} + x^{(i)}, x^{(i)} \right) dx^{(i)} dy^{(-i)}, \end{aligned}$$

where f_ϵ is the probability density function of the random utility shocks ϵ . Here, the inner integral is the probability density function of the $(n-1)$ -dimensional vector of random differences $\epsilon^{(-i)} - \epsilon^{(i)}$ [2]. By differentiating this formula with respect to $u^{(j)}$ for $j \neq i$, we obtain mixed partial derivatives of E :

$$\frac{\partial^2 E(u)}{\partial u^{(i)} \partial u^{(j)}} = - \int_{-\infty}^{u^{(i)} - u^{(-i,j)}} \int_{-\infty}^{\infty} f_\epsilon \left(y^{(-i,j)} + x^{(i)}, u^{(i)} - u^{(j)} + x^{(i)}, x^{(i)} \right) dx^{(i)} dy^{(-i,j)}.$$

This integral can be interpreted as the probability density that $\epsilon^{(j)} - \epsilon^{(i)} = u^{(i)} - u^{(j)}$, and $\epsilon_{-i,j} - \epsilon^{(i)} \leq u^{(i)} - u^{(-i,j)}$, i.e. both alternatives i and j yield the maximal utility. Analogously, we obtain the second order partial derivative of E with respect to $u^{(i)}$:

$$\frac{\partial^2 E(u)}{\partial u^{(i)2}} = \sum_{j \neq i} \int_{-\infty}^{u^{(i)} - u^{(-i,j)}} \int_{-\infty}^{\infty} f_\epsilon \left(y^{(-i,j)} + x^{(i)}, u^{(i)} - u^{(j)} + x^{(i)}, x^{(i)} \right) dx^{(i)} dy^{(-i,j)}.$$

Lemma 3 (C^2 -characterization of strong smoothness of E) *The surplus function E is L -strongly smooth with respect to the maximum norm $\|\cdot\|_\infty$ if for all $u \in \mathbb{R}^n$ it holds:*

$$\|\nabla^2 E(u)\|_{\infty,1} \leq L.$$

Proof:

For any $u, v \in \mathbb{R}^n$ we have:

$$\nabla E(u) - \nabla E(v) = \int_0^1 d\nabla E(v + \tau(u - v)) = \int_0^1 \nabla^2 E(v + \tau(u - v)) \cdot (u - v) d\tau.$$

Hence, the gradient of E is Lipschitz continuous:

$$\begin{aligned} \|\nabla E(u) - \nabla E(v)\|_1 &\leq \int_0^1 \|\nabla^2 E(v + \tau(u - v)) \cdot (u - v)\|_1 d\tau \\ &\leq \int_0^1 \|\nabla^2 E(v + \tau(u - v))\|_{\infty,1} \cdot \|u - v\|_\infty d\tau \leq L \cdot \|u - v\|_\infty. \end{aligned}$$

Further, we have:

$$E(u + v) - E(u) = \int_0^1 dE(u + \tau v) = \int_0^1 \langle \nabla E(u + \tau v), v \rangle d\tau.$$

Due to the Lipschitz continuity of ∇E , we obtain:

$$\begin{aligned} E(u + v) - E(u) - \langle \nabla E(u), v \rangle &= \int_0^1 \langle \nabla E(u + \tau v) - \nabla E(u), v \rangle d\tau \\ &\leq \int_0^1 \|\nabla E(u + \tau v) - \nabla E(u)\|_1 \cdot \|v\|_\infty d\tau \\ &\leq \int_0^1 L \cdot \|u + \tau v - u\|_\infty \cdot \|v\|_\infty d\tau = \frac{L}{2} \|v\|_\infty^2. \end{aligned}$$

□

Now, let us consider the set $\mathcal{A}$ of symmetric matrices $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ satisfying:

(A1) $a_{ii} \geq 0$ for all $i = 1, \dots, n$, and $a_{ij} \leq 0$ for all $i \neq j$,

(A2) $a_{ii} + \sum_{j \neq i} a_{ij} = 0$ for all $i = 1, \dots, n$.

Note that the set $\mathcal{A}$ is closed under matrix addition and multiplication by nonnegative scalars, i.e. for any $A, B \in \mathcal{A}$ and $\lambda \geq 0$ it holds:

$$A + B, \lambda A \in \mathcal{A}.$$

Moreover, the Hessian matrix $\nabla^2 E(u)$ of the surplus function is an element of $\mathcal{A}$.

Lemma 4 (Representation of $\|\cdot\|_{\infty,1}$) For $A \in \mathcal{A}$ it holds:

$$\|A\|_{\infty,1} = 4 \max \left\{ \sum_{i,j \in K} a_{ij} \mid K \subset \{1, \dots, n\}, |K| \leq \left\lfloor \frac{n}{2} \right\rfloor \right\}. \quad (3)$$

Proof:

The maximum in the definition of the matrix norm

$$\|A\|_{\infty,1} = \max_{\|z\|_{\infty} \leq 1} \|Az\|_1$$

is attained at some vertex of the feasible set $\{z \in \mathbb{R}^n \mid \|z\|_{\infty} \leq 1\}$. These are the vectors

$$z_K = e_K - e_{K^c},$$

where $e_K \in \mathbb{R}^n$ is the indicator vector of the subset $K \subset \{1, \dots, n\}$, i.e. $e_K^{(i)} = 1$ if $i \in K$, and $e_K^{(i)} = 0$ if $i \notin K$. We may restrict the choice of K by the condition $|K| \leq \left\lfloor \frac{n}{2} \right\rfloor$, since

$$\|Az_{K^c}\|_1 = \|A(-z_K)\|_1 = \|Az_K\|_1.$$

For such a fixed subset K we compute

$$\begin{aligned} \|Az_K\|_1 &= \sum_{i \in K} \left| a_{ii} + \sum_{j \in K \setminus \{i\}} a_{ij} - \sum_{j \in K^c} a_{ij} \right| + \sum_{i \in K^c} \left| \sum_{j \in K} a_{ij} - a_{ii} - \sum_{j \in K^c \setminus \{i\}} a_{ij} \right| \\ &\stackrel{(A2)}{=} 2 \sum_{i \in K} \left| a_{ii} + \sum_{j \in K \setminus \{i\}} a_{ij} \right| + 2 \sum_{i \in K^c} \left| \sum_{j \in K} a_{ij} \right| \\ &\stackrel{(A1)}{=} 2 \sum_{i \in K} \left(a_{ii} + \sum_{j \in K \setminus \{i\}} a_{ij} \right) - 2 \sum_{i \in K^c} \sum_{j \in K} a_{ij} \\ &= 2 \sum_{j \in K} \left(a_{jj} + \sum_{i \in K \setminus \{j\}} a_{ij} - \sum_{i \in K^c} a_{ij} \right) \\ &\stackrel{(A2)}{=} 2 \sum_{j \in K} \left(2a_{jj} + 2 \sum_{i \in K \setminus \{j\}} a_{ij} \right) = 4 \sum_{i,j \in K} a_{ij}. \end{aligned}$$

□

We use Lemma 4 to estimate the $\|\cdot\|_{\infty,1}$ norm on a particular subset of A , which will appear in what follows.

Corollary 2 For $p \in \Delta$ we define the matrix $R = \text{diag}(p) - p \cdot p^T$. It holds for the latter:

$$R \in \mathcal{A}, \quad \|R\|_{\infty,1} \leq 1.$$

Proof:

The symmetric matrix $R = (r_{ij})$ fulfills (A1), since in view of $p \in \Delta$ it holds for all $i = 1, \dots, n$, and $i \neq j$:

$$r_{ii} = p^{(i)} (1 - p^{(i)}) \geq 0, \quad r_{ij} = -p^{(i)} p^{(j)} \leq 0.$$

It also fulfills (A2) due to the following derivation:

$$R \cdot e = \text{diag}(p) \cdot e - p \cdot p^T \cdot e = p - p = 0.$$

Let us fix a subset of indices $K \subset \{1, \dots, n\}$. We estimate the following expression from Lemma 4 uniformly for all $p \in \Delta$:

$$\sum_{i,j \in K} r_{ij} = \sum_{i \in K} p^{(i)} \left(1 - \sum_{j \in K} p^{(j)} \right).$$

For that, let us solve the maximization problem

$$\max \sum_{i \in K} p^{(i)} \left(1 - \sum_{i \in K} p^{(i)} \right) \quad \text{s. t.} \quad \sum_{i \in K} p^{(i)} \leq 1, \quad p^{(i)} \geq 0 \text{ for all } i \in K.$$

Without loss of generality, we may assume that for its solution holds:

$$\sum_{i \in K} p^{(i)} < 1, \quad p^{(i)} > 0 \text{ for all } i \in K.$$

Otherwise, the optimal value vanishes or we pass over to a smaller subset of indices. The first order optimality condition reads:

$$1 - \sum_{i \in K} p^{(i)} - \sum_{i \in K} p^{(i)} = 0.$$

Hence, we get $\sum_{i \in K} p^{(i)} = \frac{1}{2}$, and the optimal value is

$$\sum_{i \in K} p^{(i)} \left(1 - \sum_{i \in K} p^{(i)} \right) = \sum_{i \in K} p^{(i)} \left(1 - \frac{1}{2} \right) = \frac{1}{4}.$$

The application of Lemma 4 provides the assertion. □

Lemma 5 (Estimation of $\|\cdot\|_{\infty,1}$) For $A \in \mathcal{A}$ it holds:

$$\|A\|_{\infty,1} \leq 2 \operatorname{tr}(A),$$

where $\operatorname{tr}(A)$ denotes the trace of the matrix A .

Proof:

For any $z \in \mathbb{R}^n$ it holds:

$$\|Az\|_1 = \sum_{i=1}^n \left| \sum_{j=1}^n a_{ij} z^{(j)} \right| \leq \sum_{i=1}^n \sum_{j=1}^n |a_{ij}| \cdot |z^{(j)}| \leq \sum_{i=1}^n \sum_{j=1}^n |a_{ij}| \cdot \|z\|_{\infty}.$$

Additionally, we have:

$$\sum_{i=1}^n \sum_{j=1}^n |a_{ij}| \stackrel{(A1)}{=} \sum_{i=1}^n \left(a_{ii} - \sum_{j \neq i} a_{ij} \right) \stackrel{(A2)}{=} \sum_{i=1}^n (a_{ii} + a_{ii}) = 2 \operatorname{tr}(A).$$

Overall, we obtain the inequality:

$$\|A\|_{\infty,1} = \max_{\|z\|_{\infty} \leq 1} \|Az\|_1 \leq \max_{\|z\|_{\infty} \leq 1} 2 \operatorname{tr}(A) \|z\|_{\infty} = 2 \operatorname{tr}(A).$$

□

Now, we are ready to state the general result on the strong convexity of E^* . It is given in terms of the differences $\epsilon^{(j)} - \epsilon^{(i)}$, $i \neq j$, of random utility shocks. We recall that the density function of $\epsilon^{(j)} - \epsilon^{(i)}$ can be written as

$$g_{i,j}(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{\epsilon} \left(y^{(-i,j)} + x^{(i)}, z + x^{(i)}, x^{(i)} \right) dx^{(i)} dy^{(-i,j)}.$$

Any point $\bar{z}_{i,j} \in \mathbb{R}$ which maximizes the density function $g_{i,j}$ is called a mode of the random variable $\epsilon^{(j)} - \epsilon^{(i)}$.

Theorem 2 (Strong convexity of E^*) Let the differences $\epsilon^{(j)} - \epsilon^{(i)}$ of random utility shocks have modes $\bar{z}_{i,j} \in \mathbb{R}$, $i \neq j$. Then, the corresponding convex conjugate E^* is β -strongly convex with respect to the $\|\cdot\|_1$ norm, where the convexity parameter is given by

$$\beta = \frac{1}{2 \sum_{i=1}^n \sum_{j \neq i} g_{i,j}(\bar{z}_{i,j})}.$$

Proof:

We estimate the second order derivative of the surplus function E with respect to $u^{(i)}$:

$$\begin{aligned} \frac{\partial^2 E(u)}{\partial u^{(i)2}} &= \sum_{j \neq i} \int_{-\infty}^{u^{(i)} - u^{(-i,j)}} \int_{-\infty}^{\infty} f_{\epsilon} \left(y^{(-i,j)} + x^{(i)}, u^{(i)} - u^{(j)} + x^{(i)}, x^{(i)} \right) dx^{(i)} dy^{(-i,j)} \\ &\leq \sum_{j \neq i} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{\epsilon} \left(y^{(-i,j)} + x^{(i)}, u^{(i)} - u^{(j)} + x^{(i)}, x^{(i)} \right) dx^{(i)} dy^{(-i,j)} \\ &= \sum_{j \neq i} g_{i,j} \left(u^{(i)} - u^{(j)} \right) \leq \sum_{j \neq i} g_{i,j}(\bar{z}_{i,j}). \end{aligned}$$

Hence, we obtain for the trace of $\nabla^2 E$ the following inequality:

$$\text{tr}(\nabla^2 E(u)) = \sum_{i=1}^n \frac{\partial^2 E(u)}{\partial u^{(i)2}} \leq \sum_{i=1}^n \sum_{j \neq i} g_{i,j}(\bar{z}_{i,j}).$$

In view of Lemma 5, whose application is justified by the fact that the matrix $\nabla^2 E(u)$ fulfills (A1)–(A2), it holds:

$$\|\nabla^2 E(u)\|_{\infty,1} \leq 2 \text{tr}(\nabla^2 E(u)) \leq 2 \sum_{i=1}^n \sum_{j \neq i} g_{i,j}(\bar{z}_{i,j}) = \frac{1}{\beta}.$$

The latter provides, due to Lemma 3, that the surplus function E is $\frac{1}{\beta}$ -strongly smooth with respect to the maximum norm $\|\cdot\|_{\infty}$. Finally, we apply Lemma 2 to conclude that the convex conjugate E^* is β -strongly convex with respect to the $\|\cdot\|_1$ norm. $\square$

Remark 1 (Existence of modes) We note that the condition on the existence of modes in Theorem 2 cannot be weakened. This can be seen already in case of two alternatives, i.e. $n = 2$. Then, the second order derivative of the surplus function E is

$$\nabla^2 E(u) = \begin{pmatrix} g_{1,2}(u^{(1)} - u^{(2)}) & -g_{1,2}(u^{(1)} - u^{(2)}) \\ -g_{2,1}(u^{(2)} - u^{(1)}) & g_{2,1}(u^{(2)} - u^{(1)}) \end{pmatrix}.$$

After a moment of reflection we realize that

$$g_{1,2}(u^{(1)} - u^{(2)}) = g_{2,1}(u^{(2)} - u^{(1)}).$$

Due to Lemma 4, it holds:

$$\|\nabla^2 E(u)\|_{\infty,1} = 4g_{1,2}(u^{(1)} - u^{(2)}).$$

From now on we assume that $g_{1,2}$ is continuous. Hence, the reverse implication in Lemma 3 becomes valid, and E is strongly smooth – or, equivalently, E^* is strongly convex – if and only if the density function $g_{1,2}$ is bounded on $\mathbb{R}$. The latter property can be characterized by the existence of a mode $\bar{z}_{1,2}$ of $\epsilon^{(2)} - \epsilon^{(1)}$. Additionally, the convexity parameter of E^* from Theorem 2 can be expressed as

$$\beta = \frac{1}{2 \sum_{i=1}^n \sum_{j \neq i} g_{i,j}(\bar{z}_{i,j})} = \frac{1}{4g_{1,2}(\bar{z}_{1,2})} = \frac{1}{\|\nabla^2 E(\bar{u})\|_{\infty,1}},$$

where the utilities $\bar{u} = (\bar{u}^{(1)}, \bar{u}^{(2)})^T$ are chosen to satisfy $\bar{u}^{(1)} - \bar{u}^{(2)} = \bar{z}_{1,2}$. This formula provides that the convexity parameter of E^* cannot be larger than β . $\square$

The estimation of the convexity parameter of E^* in Theorem 2 is rather pessimistic when the number of alternatives increases. The dependence of β on n becomes explicit in case of independent and identically distributed random utility shocks.

Corollary 3 (Strong convexity of E^* for IID utility shocks) *Let the random utility shocks $\epsilon^{(i)}$, $i = 1, \dots, n$, be independent and identically distributed with the common probability density function f having a mode $\bar{z} \in \mathbb{R}$. Then, the corresponding convex conjugate E^* is β -strongly convex with respect to the $\|\cdot\|_1$ norm, where the convexity parameter is given by*

$$\beta = \frac{1}{2n(n-1)f(\bar{z})}.$$

Proof:

We estimate the probability distribution function of $\epsilon^{(j)} - \epsilon^{(i)}$:

$$\begin{aligned} g_{i,j}(z) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{\epsilon} \left(y^{(-i,j)} + x^{(i)}, z + x^{(i)}, x^{(i)} \right) dx^{(i)} dy^{(-i,j)} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \prod_{k \neq i,j} f \left(y_k + x^{(i)} \right) f \left(z + x^{(i)} \right) f \left(x^{(i)} \right) dx^{(i)} dy^{(-i,j)} \\ &= \int_{-\infty}^{\infty} f \left(z + x^{(i)} \right) f \left(x^{(i)} \right) \prod_{k \neq i,j} \int_{-\infty}^{\infty} f \left(y_k + x^{(i)} \right) dy_k dx^{(i)}. \end{aligned}$$

For all $k \neq i, j$, and $x^{(i)} \in \mathbb{R}$ we have:

$$\int_{-\infty}^{\infty} f \left(y_k + x^{(i)} \right) dy_k = 1.$$

Moreover, $f \left(z + x^{(i)} \right) \leq f(\bar{z})$ for all $x^{(i)} \in \mathbb{R}$. Altogether, we get

$$g_{i,j}(z) \leq f(\bar{z}).$$

The application of Theorem 2 yields the assertion. $\square$

In what follows we concentrate on some special distributions of random utility shocks widely used in the discrete choice literature. For them we obtain better estimations of the convexity parameter of E^* , in particular, not dependent on the number of alternatives. First, we consider the class of generalized extreme value models [9]. The vector $\epsilon = (\epsilon^{(1)}, \dots, \epsilon^{(n)})^T$ of random utility shocks defines a generalized extreme value model (GEV) if it follows the joint distribution given by the probability density function

$$f_{\epsilon} \left(y^{(1)}, \dots, y^{(n)} \right) = \frac{\partial^n \exp \left(-G \left(e^{-y^{(1)}}, \dots, e^{-y^{(n)}} \right) \right)}{\partial y^{(1)} \dots \partial y^{(n)}},$$

where the generating function $G : \mathbb{R}_+^n \rightarrow \mathbb{R}_+$ has the following properties:

- (G1) G is homogeneous of degree $1/\mu > 0$.
- (G1) $G \left(x^{(1)}, \dots, x^{(i)}, \dots, x^{(n)} \right) \rightarrow \infty$ as $x^{(i)} \rightarrow \infty$, $i = 1, \dots, n$.

(G3) For the partial derivatives of G with respect to k distinct variables it holds:

$$\frac{\partial^k G(x^{(1)}, \dots, x^{(n)})}{\partial x^{(i_1)} \dots \partial x^{(i_k)}} \geq 0 \text{ if } k \text{ is odd, } \quad \frac{\partial^k G(x^{(1)}, \dots, x^{(n)})}{\partial x^{(i_1)} \dots \partial x^{(i_k)}} \leq 0 \text{ if } k \text{ is even.}$$

It is well known from [9] that the surplus function for GEV is

$$E(u) = \mu \ln G(e^u) + \mu \gamma,$$

where γ is Euler's constant and we set $e^u = (e^{u^{(1)}}, \dots, e^{u^{(n)}})^T$ for the sake of brevity. The choice probability of the i -th alternative is given by the i -th partial derivative of the GEV surplus function E :

$$\mathbb{P}\left(u^{(i)} + \epsilon^{(i)} = \max_{1 \leq i \leq n} u^{(i)} + \epsilon^{(i)}\right) = \frac{\partial E(u)}{\partial u^{(i)}} = \mu \frac{\partial G(e^u)}{\partial x^{(i)}} \cdot \frac{e^{u^{(i)}}}{G(e^u)}.$$

We state a sufficient condition for the strong convexity of E^* in terms of the generating function G .

Theorem 3 (Strong convexity of E^* for GEV) *Let a generating function G for GEV satisfy the following inequality for all $x = (x^{(1)}, \dots, x^{(n)})^T \in \mathbb{R}_+^n$:*

$$\sum_{i=1}^n \frac{\partial^2 G(x)}{\partial x^{(i)2}} \cdot x^{(i)2} \leq M \cdot G(x)$$

with some constant $M \in \mathbb{R}$. Then, the corresponding convex conjugate E^ is β -strongly convex with respect to the $\|\cdot\|_1$ norm, where the convexity parameter is given by*

$$\beta = \frac{1}{2(\mu M + 1) - 1/\mu}.$$

Proof:

For the second order partial derivative of E with respect to $u^{(i)}$ it holds:

$$\begin{aligned} \frac{\partial^2 E(u)}{\partial u^{(i)2}} &= \mu \left(\frac{\partial G(e^u)}{\partial x^{(i)}} \cdot \frac{e^{u^{(i)}}}{G(e^u)} \left(1 - \frac{\partial G(e^u)}{\partial x^{(i)}} \cdot \frac{e^{u^{(i)}}}{G(e^u)} \right) + \frac{\partial G^2(e^u)}{\partial x^{(i)2}} \cdot \frac{(e^{u^{(i)}})^2}{G(e^u)} \right) \\ &= \frac{1}{\mu} \frac{\partial E(u)}{\partial u^{(i)}} \left(1 - \frac{\partial E(u)}{\partial u^{(i)}} \right) + \left(1 - \frac{1}{\mu} \right) \frac{\partial E(u)}{\partial u^{(i)}} + \mu \frac{\partial G^2(e^u)}{\partial x^{(i)2}} \cdot \frac{(e^{u^{(i)}})^2}{G(e^u)}. \end{aligned}$$

Analogously, we obtain the mixed partial derivative of E for $j \neq i$:

$$\begin{aligned} \frac{\partial^2 E(u)}{\partial u^{(i)} \partial u^{(j)}} &= \mu \left(-\frac{\partial G(e^u)}{\partial x^{(i)}} \frac{e^{u^{(i)}}}{G(e^u)} \cdot \frac{\partial G(e^u)}{\partial x^{(j)}} \frac{e^{u^{(j)}}}{G(e^u)} + \frac{\partial G^2(e^u)}{\partial x^{(i)} \partial x^{(j)}} \cdot \frac{e^{u^{(i)}} e^{u^{(j)}}}{G(e^u)} \right) \\ &= -\frac{1}{\mu} \frac{\partial E(u)}{\partial u^{(i)}} \cdot \frac{\partial E(u)}{\partial u^{(j)}} + \mu \frac{\partial G^2(e^u)}{\partial x^{(i)} \partial x^{(j)}} \cdot \frac{e^{u^{(i)}} e^{u^{(j)}}}{G(e^u)}. \end{aligned}$$

Equivalently, we have in matrix form:

$$\nabla^2 E(u) = \frac{1}{\mu} R(u) + S(u),$$

where

$$\begin{aligned} R(u) &= \text{diag}(\nabla E(u)) - \nabla E(u) \cdot \nabla^T E(u), \\ S(u) &= \left(1 - \frac{1}{\mu}\right) \text{diag}(\nabla E(u)) + \mu \frac{\text{diag}(e^u) \cdot \nabla^2 G(e^u) \cdot \text{diag}(e^u)}{G(e^u)}. \end{aligned}$$

Since $\nabla E(u) \in \Delta$, we may apply Corollary 2 to derive that

$$R(u) \in \mathcal{A}, \quad \|R(u)\|_{\infty,1} \leq 1.$$

In view of $\nabla^2 E(u)$, $R(u) \in \mathcal{A}$, the matrix $S(u)$ fulfills (A2):

$$S(u) \cdot e = \underbrace{\nabla^2 E(u) \cdot e}_{=0} - \frac{1}{\mu} \underbrace{R(u) \cdot e}_{=0} = 0.$$

The off-diagonal entries of $S(u)$ are nonpositive due to (G3), hence, $S(u)$ fulfills also (A1). We apply Lemma 5 to estimate the $\|\cdot\|_{\infty,1}$ norm of $S(u) \in \mathcal{A}$:

$$\begin{aligned} \|S(u)\|_{\infty,1} &\leq 2 \text{tr}(S(u)) \\ &= 2 \left(\left(1 - \frac{1}{\mu}\right) \sum_{i=1}^n \frac{\partial E(u)}{\partial u^{(i)}} + \mu \sum_{i=1}^n \frac{\partial G^2(e^u)}{\partial x^{(i)2}} \cdot \frac{(e^{u^{(i)}})^2}{G(e^u)} \right) \\ &\leq 2 \left(\left(1 - \frac{1}{\mu}\right) + \mu M \right). \end{aligned}$$

Overall, we have:

$$\|\nabla^2 E(u)\|_{\infty,1} \leq \frac{1}{\mu} \|R(u)\|_{\infty,1} + \|S(u)\|_{\infty,1} \leq \frac{1}{\mu} + 2 \left(\left(1 - \frac{1}{\mu}\right) + \mu M \right) = \frac{1}{\beta}.$$

The latter provides, due to Lemma 3, that the surplus function E is $\frac{1}{\beta}$ -strongly smooth with respect to the maximum norm $\|\cdot\|_{\infty}$. Finally, we apply Lemma 2 to conclude that the convex conjugate E^* is β -strongly convex with respect to the $\|\cdot\|_1$ norm. $\square$

Now, we consider generalized nested logit models (GNL) introduced in [18]. GNL is a particular class of GEV models defined by the generating function

$$G(x) = \sum_{\ell \in L} \left(\sum_{i=1}^n (\sigma_{i\ell} \cdot x^{(i)})^{1/\mu_\ell} \right)^{\mu_\ell/\mu}.$$

Here, L is a generic set of nests. The parameters $\sigma_{i\ell} \geq 0$ denote the shares of the i -th alternative with which it is attached to the ℓ -th nest. For any fixed $i \in \{1, \dots, n\}$ they sum up to one:

$$\sum_{\ell \in L} \sigma_{i\ell} = 1,$$

and $\sigma_{i\ell} = 0$ means that the ℓ -th nest does not contain the i -th alternative. Hence, the set of alternatives within the ℓ -th nest is

$$N_\ell = \{i \mid \sigma_{i\ell} > 0\}.$$

The nest parameters $\mu_\ell > 0$ describe the variance of the random errors while choosing alternatives within the ℓ -th nest. Analogously, $\mu > 0$ describes the variance of the random errors while choosing among the nests. For the function G to fulfill (G1)-(G3) we require:

$$\mu_\ell \leq \mu \quad \text{for all } \ell \in L.$$

The underlying choice process can be viewed to comprise two stages:

- (1) the probability of choosing the ℓ -th nest is

$$q_\ell = \frac{e^{v_\ell/\mu}}{\sum_{\ell \in L} e^{v_\ell/\mu}},$$

where

$$v_\ell = \mu_\ell \ln \left(\sum_{i=1}^n \left(\sigma_{i\ell} \cdot e^{u^{(i)}} \right)^{1/\mu_\ell} \right)$$

stands for the utility attached to the ℓ -th nest;

- (2) the probability of choosing the i -th alternative within the ℓ -th nest is

$$p_{i\ell} = \frac{\left(\sigma_{i\ell} \cdot e^{u^{(i)}} \right)^{1/\mu_\ell}}{\sum_{i=1}^n \left(\sigma_{i\ell} \cdot e^{u^{(i)}} \right)^{1/\mu_\ell}}.$$

Overall, the choice probability of the i -th alternative according to GNL amounts to

$$\mathbb{P} \left(u^{(i)} + \epsilon^{(i)} = \max_{1 \leq i \leq n} u^{(i)} + \epsilon^{(i)} \right) = \mu \frac{\partial G(e^u)}{\partial x^{(i)}} \cdot \frac{e^{u^{(i)}}}{G(e^u)} = \sum_{\ell \in L} q_\ell \cdot p_{i\ell}.$$

Corollary 4 (Strong convexity of E^* for GNL) *For GNL the corresponding convex conjugate E^* is β -strongly convex with respect to the $\|\cdot\|_1$ norm, where the convexity parameter is given by*

$$\beta = \frac{1}{\frac{2}{\min_{\ell \in L} \mu_\ell} - 1/\mu}.$$

Proof:

Let us estimate the constant M from Theorem 3. We have:

$$\frac{\partial G(x)}{\partial x^{(i)}} = \frac{1}{\mu} \sum_{\ell \in L} \left(\sum_{i=1}^n (\sigma_{i\ell} \cdot x^{(i)})^{1/\mu_\ell} \right)^{\mu_\ell/\mu-1} (\sigma_{i\ell} \cdot x^{(i)})^{1/\mu_\ell-1} \cdot \sigma_{i\ell},$$

and, further:

$$\begin{aligned} \frac{\partial^2 G(x)}{\partial x^{(i)2}} &= \frac{1}{\mu} \sum_{\ell \in L} \frac{1}{\mu_\ell} \left(\frac{\mu_\ell}{\mu} - 1 \right) \left(\sum_{i=1}^n (\sigma_{i\ell} \cdot x^{(i)})^{1/\mu_\ell} \right)^{\mu_\ell/\mu-2} \left((\sigma_{i\ell} \cdot x^{(i)})^{1/\mu_\ell-1} \cdot \sigma_{i\ell} \right)^2 \\ &+ \frac{1}{\mu} \sum_{\ell \in L} \left(\frac{1}{\mu_\ell} - 1 \right) \left(\sum_{i=1}^n (\sigma_{i\ell} \cdot x^{(i)})^{1/\mu_\ell} \right)^{\mu_\ell/\mu-1} (\sigma_{i\ell} \cdot x^{(i)})^{1/\mu_\ell-2} \cdot \sigma_{i\ell}^2. \end{aligned}$$

Due to $\mu_\ell \leq \mu$, $\ell \in L$, we get:

$$\frac{\partial^2 G(x)}{\partial x^{(i)2}} \leq \frac{1}{\mu} \sum_{\ell \in L} \left(\frac{1}{\mu_\ell} - 1 \right) \left(\sum_{i=1}^n (\sigma_{i\ell} \cdot x^{(i)})^{1/\mu_\ell} \right)^{\mu_\ell/\mu-1} (\sigma_{i\ell} \cdot x^{(i)})^{1/\mu_\ell-2} \cdot \sigma_{i\ell}^2.$$

We multiply these terms by $x^{(i)2}$ and sum up over $i = 1, \dots, n$:

$$\begin{aligned} \sum_{i=1}^n \frac{\partial^2 G(x)}{\partial x^{(i)2}} \cdot x^{(i)2} &\leq \frac{1}{\mu} \sum_{\ell \in L} \left(\frac{1}{\mu_\ell} - 1 \right) \left(\sum_{i=1}^n (\sigma_{i\ell} \cdot x^{(i)})^{1/\mu_\ell} \right)^{\mu_\ell/\mu-1} \\ &\quad \cdot \sum_{i=1}^n (\sigma_{i\ell} \cdot x^{(i)})^{1/\mu_\ell-2} \cdot \sigma_{i\ell}^2 \cdot x^{(i)2} \\ &= \frac{1}{\mu} \sum_{\ell \in L} \left(\frac{1}{\mu_\ell} - 1 \right) \left(\sum_{i=1}^n (\sigma_{i\ell} \cdot x^{(i)})^{1/\mu_\ell} \right)^{\mu_\ell/\mu} \\ &\leq \frac{1}{\mu} \max_{\ell \in L} \left(\frac{1}{\mu_\ell} - 1 \right) \cdot G(x) = \frac{1}{\mu} \left(\frac{1}{\min_{\ell \in L} \mu_\ell} - 1 \right) \cdot G(x). \end{aligned}$$

Hence, we may apply Theorem 3 with

$$M = \frac{1}{\mu} \left(\frac{1}{\min_{\ell \in L} \mu_\ell} - 1 \right).$$

□

Example 1 (Multinomial logit) Let in GNL there be just one nest, i. e. $L = \{1\}$, and $\mu_1 = \mu$. Then, the generating function

$$G(x) = \sum_{i=1}^n \left(x^{(i)} \right)^{1/\mu}$$

leads to the multinomial logit (ML). The corresponding surplus function is

$$E(u) = \mu \ln \sum_{i=1}^n e^{u^{(i)}/\mu} + \mu\gamma,$$

and the choice probabilities are

$$\mathbb{P}\left(u^{(i)} + \epsilon^{(i)} = \max_{1 \leq i \leq n} u^{(i)} + \epsilon^{(i)}\right) = \frac{e^{u^{(i)}/\mu}}{\sum_{i=1}^n e^{u^{(i)}/\mu}}, \quad i = 1, \dots, n.$$

Note that ML can be deduced from the IID random utility shocks $\epsilon^{(i)}$, $i = 1, \dots, n$, each of them following the Gumbel distribution with zero mode and variance $\mu\pi/\sqrt{6}$ [2]. For ML the convex conjugate of the surplus function can be explicitly given:

$$E^*(p) = \mu \sum_{i=1}^n p^{(i)} \ln p^{(i)} - \mu\gamma = \mu H(p) - \mu\gamma,$$

where H is the (negative) entropy. It is well known that H is 1-strongly convex with respect to the $\|\cdot\|_1$ norm due to the Pinsker inequality. Hence, E^* is μ -strongly convex with respect to the $\|\cdot\|_1$ norm. The same result also follows from Corollary 4 with the convexity parameter

$$\beta = \frac{1}{\frac{2}{\min_{\ell \in L} \mu_\ell} - 1/\mu} = \frac{1}{2/\mu - 1/\mu} = \mu.$$

□

Example 2 (Nested logit) Let in GNL for every alternative $i \in \{1, \dots, n\}$ there be a unique nest $\ell_i \in L$ with $\sigma_{i\ell_i} = 1$, and $\mu = 1$. Then, the nests $N_\ell = \{i \mid \ell_i = \ell\}$ are mutually exclusive, and the generating function

$$G(x) = \sum_{\ell \in L} \left(\sum_{i \in N_\ell} x^{(i)1/\mu_\ell} \right)^{\mu_\ell}$$

leads to the nested logit (NL). The corresponding surplus function is

$$E(u) = \mu \ln \sum_{\ell \in L} \left(\sum_{i \in N_\ell} e^{u^{(i)}/\mu_\ell} \right)^{\mu_\ell} + \mu\gamma,$$

and the choice probabilities for $i \in N_\ell$, $\ell \in L$ are

$$\mathbb{P}\left(u^{(i)} + \epsilon^{(i)} = \max_{1 \leq i \leq n} u^{(i)} + \epsilon^{(i)}\right) = \frac{e^{\mu_\ell \ln \sum_{i \in N_\ell} e^{u^{(i)}/\mu_\ell}}}{\sum_{\ell \in L} e^{\mu_\ell \ln \sum_{i \in N_\ell} e^{u^{(i)}/\mu_\ell}}} \cdot \frac{e^{u^{(i)}/\mu_\ell}}{\sum_{i \in N_\ell} e^{u^{(i)}/\mu_\ell}}.$$

For NL the convex conjugate of the surplus function is explicitly given in [4]:

$$E^*(p) = \sum_{\ell \in L} \mu_\ell \sum_{i \in N_\ell} p^{(i)} \ln p^{(i)} + \sum_{\ell \in L} (1 - \mu_\ell) \left(\sum_{i \in N_\ell} p^{(i)} \right) \ln \left(\sum_{i \in N_\ell} p^{(i)} \right) - \mu\gamma.$$

By examining the first part of this formula, it can be shown that E^* is $\left(\min_{\ell \in L} \mu_\ell\right)$ -strongly convex with respect to the $\|\cdot\|_1$ norm. Corollary 4 provides the convexity parameter, which is of the same order:

$$\beta = \frac{1}{\frac{2}{\min_{\ell \in L} \mu_\ell} - 1/\mu} = \frac{1}{\frac{2}{\min_{\ell \in L} \mu_\ell} - 1} > \frac{1}{2} \min_{\ell \in L} \mu_\ell.$$

□

Example 3 (GNL's from literature) For other specifications of GNL – except of ML and NL – the explicit form of the convex conjugate E^* is not known yet. In this case Corollary 4 can be applied in order to estimate the convexity parameter of E^* .

(i) **Ordered GEV** [17] is a GNL model with

$$L = \{1, \dots, n+m\}, \quad \mu = 1,$$

$$\sigma_{i\ell} > 0 \text{ for all } \ell \in \{i, \dots, i+m\}, \quad \sigma_{i\ell} = 0 \text{ for all } \ell \in L \setminus \{i, \dots, i+m\}.$$

There are $n+m$ overlapping nests $N_\ell = \{i \mid \ell - m \leq i \leq \ell\}$, and every alternative lies exactly in $m+1$ of them, namely $i \in N_\ell$ for $\ell = i, \dots, i+m$. Then, the generating function is

$$G(x) = \sum_{\ell=1}^{n+m} \left(\sum_{i \in N_\ell} \left(\sigma_{i\ell} x^{(i)} \right)^{1/\mu_\ell} \right)^{\mu_\ell}.$$

(ii) **Paired combinatorial logit** [7] is a GNL model with

$$L = \{(i, j) \in \{1, \dots, n\} \mid i \neq j\}, \quad \mu = 1,$$

$$\sigma_{i\ell} = \begin{cases} \frac{1}{2(n-1)} & \text{if } \ell = (i, j), (j, i) \text{ with } j \neq i, \\ 0 & \text{else.} \end{cases}$$

There are $n^2 - n$ nests corresponding to the pairs of alternatives, and every alternative lies in $2(n-1)$ of them. Then, the generating function is

$$G(x) = \sum_{\ell=(i,j), i \neq j} \left(\left(\sigma_{i\ell} x^{(i)} \right)^{1/\mu_\ell} + \left(\sigma_{j\ell} x^{(j)} \right)^{1/\mu_\ell} \right)^{\mu_\ell}.$$

(iii) **Principles of differentiation GEV** [3] is a GNL model with

$$L = \bigcup_{d \in D} L_d, \quad \mu = 1, \quad \mu_\ell = \mu_d \text{ for all } \ell \in L_d,$$

$$\sigma_{i\ell} = \begin{cases} \sigma_d & \text{if } i \in N_{\ell d} \text{ and } \ell \in L_d, \\ 0 & \text{else,} \end{cases}$$

where

$$\{1, \dots, n\} = \bigcup_{\ell \in L_d} N_{\ell d}.$$

The set D represents the dimensions of alternatives. Along the d -th dimension alternatives can be clustered into the disjoint nests $N_{\ell d}$, $\ell \in L_d$. The shares $\sigma_{i\ell}$ of all alternatives within these nests depend only on the dimension d . The parameters μ_ℓ also coincide for all $\ell \in L_d$. Then, the generating function is

$$G(x) = \sum_{d \in D} \sigma_d \sum_{\ell \in L_d} \left(\sum_{i \in N_{\ell d}} (x^{(i)})^{1/\mu_d} \right)^{\mu_d}.$$

Note that the convexity parameter of E^* for GNL specifications (i)-(iii) depends only on the smallest of the nest parameters μ_ℓ , $\ell \in L$:

$$\beta = \frac{1}{\frac{2}{\min_{\ell \in L} \mu_\ell} - 1}.$$

□

Let us relate our results on the strong convexity of E^* to those existing in the literature.

Remark 2 (Cross moment model) In CMM [11] the random vector $\epsilon \sim (0, \Sigma)$ follows a joint distribution with zero mean and a given covariance matrix Σ . Additionally, it maximizes the surplus function:

$$Z(u) = \max_{\epsilon \sim (0, \Sigma)} \mathbb{E}_\epsilon \left(\max_{1 \leq i \leq n} u^{(i)} + \epsilon^{(i)} \right).$$

Assuming that the covariance matrix Σ is positive definite, the following dual representation of Z has been derived in [1]:

$$Z(u) = \max_{p \in \Delta} \langle p, u \rangle + \text{tr} \left(\left(\Sigma^{1/2} (\text{diag}(p) - pp^T) \Sigma^{1/2} \right)^{1/2} \right).$$

Moreover, the solution $p \in \Delta$ of the latter optimization problem provides the choice probabilities corresponding to the random error $\epsilon(u) \sim (0, \Sigma)$ that maximizes the surplus function above, i. e.

$$p^{(i)} = \mathbb{P} \left(u^{(i)} + \epsilon^{(i)}(u) = \max_{1 \leq i \leq n} u^{(i)} + \epsilon^{(i)}(u) \right), \quad i = 1, \dots, n.$$

Hence, the convex conjugate of Z is

$$Z^*(p) = -\text{tr} \left(\left(\Sigma^{1/2} (\text{diag}(p) - pp^T) \Sigma^{1/2} \right)^{1/2} \right).$$

In [1], Z^* is shown to be strongly convex on the simplex w.r.t. Euclidean norm (however, its convexity parameter is not given explicitly). We point out that Z^* can be therefore used as a discrete choice prox-function on the simplex as well. $\square$

2.2.3 Simplicity

In view of the discussion in Section 2.2.2, we assume that the convex conjugate E^* is strongly convex.

Theorem 4 (Simplicity) *The unique maximizer of the optimization problem*

$$E(u) = \sup_{p \in \Delta} \langle p, u \rangle - E^*(p) \quad (4)$$

is given by the choice probabilities

$$p^{(i)} = \mathbb{P} \left(u^{(i)} + \epsilon^{(i)} = \max_{1 \leq i \leq n} u^{(i)} + \epsilon^{(i)} \right), \quad i = 1, \dots, n.$$

Proof:

Let a vector $u \in \mathbb{R}^n$ of deterministic utilities be given. The vector p of corresponding probabilities lies in the relative interior of the simplex Δ , i.e. $p \in \text{rint}(\Delta)$. This is due to the fact that the distribution of the random vector ϵ is fully supported on $\mathbb{R}^n$, see [15]. Hence, we have:

$$p = \nabla E(u),$$

or, equivalently, by the convex duality [16]:

$$u \in \partial E^*(p).$$

The latter is the first order optimality condition for the optimization problem (4). Additionally, note that the dual representation of the surplus function is valid due to the Fenchel-Moreau theorem. $\square$

Corollary 5 (Lower bound for E^*) *The unique minimizer p_0 of the convex conjugate E^* consists of the choice probabilities with respect to the zero-utility, i.e.*

$$p_0^{(i)} = \mathbb{P} \left(\epsilon^{(i)} = \max_{1 \leq i \leq n} \epsilon^{(i)} \right), \quad i = 1, \dots, n.$$

Moreover, it holds:

$$E^*(p) \geq E^*(p_0) = -\mathbb{E}_\epsilon \left(\max_{1 \leq i \leq n} \epsilon^{(i)} \right) \quad \text{for all } p \in \Delta.$$

Proof:

Setting $u = 0$ in Theorem 4 yields the assertion. $\square$

3 Economic application

We apply discrete choice prox-functions for the natural adjustment of consumer's demand.

3.1 Lancaster's approach to consumer theory

We briefly describe the Lancaster's approach to consumer theory as presented in [8]. For that, let $x \in \mathbb{R}_+^m$ denote the consumer's demand vector of m goods. Every such demand vector x generates the vector $z \in \mathbb{R}^n$ of n qualities (sometimes called characteristics):

$$z = Qx,$$

where $Q = (q_{ij}) \in \mathbb{R}^{n \times m}$ is a fixed quality matrix. Its entries q_{ij} denote the amounts of the i -th quality while consuming one unit of the j -th good. Further, the consumer assigns utility $u(z)$ to the goods' qualities z , and tries to maximize it by adjusting the demand x . Hereby, the budget constraint need to be satisfied, i. e.

$$\langle \pi, x \rangle \leq w,$$

where $\pi \in \mathbb{R}_+^m$ is a fixed vector of positive goods' prices, i.e. $\pi > 0$, and $w > 0$ is a fixed available budget. Overall, the Lancaster's approach to consumer theory consists in solving the following maximization problem:

$$\max_{x \geq 0} u(z) \quad \text{s.t.} \quad z = Qx, \quad \langle \pi, x \rangle \leq w.$$

In what follows we focus on the Leontieff utility function by setting

$$u(z) = \min \left\{ \frac{z^{(1)}}{\sigma^{(1)}}, \dots, \frac{z^{(n)}}{\sigma^{(n)}} \right\},$$

where $\sigma = (\sigma^{(1)}, \dots, \sigma^{(n)})^T \in \mathbb{R}_+^n$ is a fixed vector of positive quality standards, i. e. $\sigma > 0$. Additionally, we assume that there exists a feasible demand vector which delivers positive Leontieff utility. Otherwise, the Lancaster's consumption problem is trivially solved by driving the demand to zero. On the contrary, under the proposed assumption the whole budget w will be spent at any optimal demand. Hence, the budget constraint can be taken tight without loss of generality. The primal Lancaster's consumption problem becomes

$$\text{P:} \quad \max_{\substack{x \geq 0 \\ \langle \pi, x \rangle = w}} U(x),$$

with the concave objective function

$$U(x) = u(Qx) = \min_{1 \leq i \leq n} \frac{(Qx)^{(i)}}{\sigma^{(i)}}.$$

The optimization problem (P) consists in adjusting demand x by spending the budget w in order to maximize the worst ratio of consumed qualities $(Qx)^{(i)}$ in relation to their desired standards $\sigma^{(i)}$ taking over $i = 1, \dots, n$.

Let us derive a dual optimization problem for (P). For that, we introduce dual variables

$$\lambda = \left(\lambda^{(1)}, \dots, \lambda^{(n)} \right)^T \in \mathbb{R}^n,$$

which can be interpreted as internal prices of qualities. Due to the duality of linear programming, we have:

$$\begin{aligned} \max_{\substack{x \geq 0 \\ \langle \pi, x \rangle = w}} U(x) &= \max_{\substack{x \geq 0 \\ \langle \pi, x \rangle = w}} \min_{1 \leq i \leq n} \frac{(Qx)^{(i)}}{\sigma^{(i)}} \\ &= \max_{\substack{x \geq 0 \\ \langle \pi, x \rangle = w}} \min_{\substack{\lambda \geq 0 \\ \langle \sigma, \lambda \rangle = 1}} \langle Qx, \lambda \rangle \\ &= \min_{\substack{\lambda \geq 0 \\ \langle \sigma, \lambda \rangle = 1}} \max_{\substack{x \geq 0 \\ \langle \pi, x \rangle = w}} \langle x, Q^T \lambda \rangle \\ &= \min_{\substack{\lambda \geq 0 \\ \langle \sigma, \lambda \rangle = 1}} w \max_{1 \leq j \leq m} \frac{(Q^T \lambda)^{(j)}}{\pi^{(j)}} = \min_{\substack{\lambda \geq 0 \\ \langle \sigma, \lambda \rangle = 1}} \Phi(\lambda). \end{aligned}$$

The dual Lancaster's consumption problem becomes

$$\text{D: } \min_{\substack{\lambda \geq 0 \\ \langle \sigma, \lambda \rangle = 1}} \Phi(\lambda),$$

with the convex objective function

$$\Phi(\lambda) = w \max_{1 \leq j \leq m} \frac{(Q^T \lambda)^{(j)}}{\pi^{(j)}}.$$

The optimization problem (D) consists in adjusting internal prices of qualities λ in order to minimize the best ratio of quality estimates $(Q^T \lambda)^{(j)}$ of goods in relation to their market prices $\pi^{(j)}$ taking over $j = 1, \dots, m$. Note that usually the number n of goods is considerably bigger than the number m of their qualities. From a consumer it will be thus more plausible to expect that (D) is successively solved rather than (P).

3.2 Dual Averaging Scheme

We rewrite the dual optimization problem (D) by introducing new variables

$$p^{(i)} = \sigma^{(i)} \lambda^{(i)}, \quad i = 1, \dots, n.$$

In terms of the variables $p = (p^{(1)}, \dots, p^{(n)})^T$ we equivalently obtain the following auxiliary optimization problem on the simplex:

$$\text{A: } \min_{p \in \Delta} \Psi(p),$$

where the objective function is

$$\Psi(p) = \Phi\left(\frac{p}{\sigma}\right).$$

For solving (A) we apply the dual averaging scheme from [13]. For that, we use the family of prox-functions on the simplex discussed in Section 2:

$$d(p) = E^*(p) - E^*(p_0).$$

Note that d is continuous on the simplex Δ (Theorem 1), strongly convex with convexity parameter $\beta > 0$ (Theorems 2 and 3, Corollary 4). The computation of its convex conjugate $d^*(u) = E(u) - E(0)$ is simple (Theorem 4). Additionally, Corollary 5 provides us with the prox-center of Δ :

$$p_0 = \arg \min_{p \in \Delta} d(p).$$

In view of $d(p_0) = 0$, the dual averaging scheme can be initialized by p_0 .

Dual Averaging Scheme for (A)
1. Compute $\nabla \Psi(p_k)$. 2. Set $s_{k+1} = \frac{1}{k+1} \sum_{\ell=0}^k \nabla \Psi(p_\ell)$. 3. Update $p_{k+1} = \arg \min_{p \in \Delta} \left\{ \langle s_{k+1}, p \rangle + \frac{d(p)}{\sqrt{k+1}} \right\}$.

The convergence properties of the dual averaging scheme follow from [13, Theorem 1]. For $k \geq 0$ it holds:

$$\delta_k \leq \frac{D}{\sqrt{k+1}} + \frac{M^2}{2\beta} \cdot \frac{1}{k+1} \left(1 + \sum_{\ell=1}^k \frac{1}{\sqrt{\ell}} \right), \quad (5)$$

where

$$\delta_k = \max_{p \in \Delta} \frac{1}{k+1} \sum_{\ell=0}^k \langle \nabla \Psi(p_\ell), p_\ell - p \rangle, \quad M = \max_{p \in \Delta} \|\nabla \Psi(p)\|_\infty, \quad D \geq \max_{p \in \Delta} d(p).$$

3.3 Consumption cycle

Let us state the dual averaging scheme from Section 3.2 in terms of the primal and dual Lancaster consumption problem (P) and (D), respectively.

1. We compute

$$\nabla \Psi(p_k) = \nabla \Phi\left(\frac{p_k}{\sigma}\right) = \frac{\nabla \Phi(\lambda_k)}{\sigma},$$

where λ_k denotes the internal prices of qualities at the k -th iteration. It holds:

$$\nabla \Phi(\lambda_k) = \nabla \left(w \max_{1 \leq j \leq m} \frac{(Q^T \lambda_k)^{(j)}}{\pi^{(j)}} \right) = w \frac{Q y_k}{\pi},$$

where the sharing vector $y_k \in \Delta$ fulfills

$$y_k^{(j)} = 0 \quad \text{for } j \notin J(\lambda_k)$$

with the active index set

$$J(\lambda_k) = \left\{ j \in \{1, \dots, m\} \mid \frac{(Q^T \lambda_k)^{(j)}}{\pi^{(j)}} = \max_{1 \leq j \leq m} \frac{(Q^T \lambda_k)^{(j)}}{\pi^{(j)}} \right\}.$$

In other words, $J(\lambda_k)$ contains goods with the best quality/price ratio estimated by means of internal prices λ_k . We set the demand at the k -th iteration as

$$x_k = w \frac{y_k}{\pi}.$$

Note that x_k is feasible for (P), since

$$\langle \pi, x_k \rangle = \left\langle \pi, w \frac{y_k}{\pi} \right\rangle = w \langle e, y_k \rangle = w.$$

Moreover, the demand is concentrated on the goods with the best quality/price ratio, i. e. $x_k^{(j)} \neq 0$ if and only if $j \in J(\lambda_k)$. Overall, we obtain:

$$\nabla \Psi(p_k) = \frac{Q x_k}{\sigma},$$

which are the ratios of the consumed qualities $Q x_k$ in relation to their standards σ .

2. We set

$$s_{k+1} = \frac{1}{k+1} \sum_{\ell=0}^k \nabla \Psi(p_\ell) = \frac{1}{k+1} \sum_{\ell=0}^k \frac{Q x_\ell}{\sigma} = \frac{Q \bar{x}_k}{\sigma}$$

with the average demand

$$\bar{x}_k = \frac{1}{k+1} \sum_{\ell=0}^k x_\ell.$$

Again, s_{k+1} relates the average consumption $Q \bar{x}_k$ to the standards σ .

3. We update

$$p_{k+1} = \arg \min_{p \in \Delta} \left\{ \langle s_{k+1}, p \rangle + \frac{d(p)}{\sqrt{k+1}} \right\}.$$

Due to Theorem 4, we equivalently obtain for $i = 1, \dots, n$:

$$p_{k+1}^{(i)} = \mathbb{P} \left(s_{k+1}^{(i)} - \frac{\epsilon^{(i)}}{\sqrt{k+1}} = \min_{1 \leq i \leq n} s_{k+1}^{(i)} - \frac{\epsilon^{(i)}}{\sqrt{k+1}} \right).$$

For the internal prices we have:

$$\lambda_{k+1}^{(i)} = \frac{1}{\sigma^{(i)}} \mathbb{P} \left(s_{k+1}^{(i)} - \frac{\epsilon^{(i)}}{\sqrt{k+1}} = \min_{1 \leq i \leq n} s_{k+1}^{(i)} - \frac{\epsilon^{(i)}}{\sqrt{k+1}} \right).$$

Thus, the internal price $\lambda_{k+1}^{(i)}$ of the i -th quality is proportional to the probability of detecting its average consumption $(Q\bar{x}_k)^{(i)}$ as the lowest one in comparison to the standard $\sigma^{(i)}$. Moreover, this detecting process is in accordance with additive random utility models allowing behavioral interpretations. E.g., the additive random errors $\frac{\epsilon^{(i)}}{\sqrt{k+1}}$ are diminishing with respect to the iteration number k , which accounts for the learning effect. We further mention that λ_{k+1} is feasible for (D), since

$$\langle \sigma, \lambda_{k+1} \rangle = \langle e, p_{k+1} \rangle = 1.$$

So are the average internal prices

$$\bar{\lambda}_k = \frac{1}{k+1} \sum_{\ell=0}^k \lambda_\ell.$$

Let us examine the convergence properties of the proposed consumption cycle.

Theorem 5 (Consumption cycle) *The duality gap between (P) and (D) evaluated at the average demand and the average internal prices is closing at the optimal rate $O\left(\frac{1}{\sqrt{k+1}}\right)$. Namely, it holds for $k \geq 0$:*

$$0 \leq \Phi(\bar{\lambda}_k) - U(\bar{x}_k) \leq \left(D + \frac{M^2}{\beta}\right) \frac{1}{\sqrt{k+1}},$$

where

$$M = w \max_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}} \frac{|q_{i,j}|}{\sigma^{(i)} \cdot \pi^{(j)}}, \quad D = \mathbb{E}_\epsilon \left(\max_{1 \leq i \leq n} \epsilon^{(i)} \right) - \min_{1 \leq i \leq n} \mathbb{E}_\epsilon \left(\epsilon^{(i)} \right).$$

Proof:

We compute the gap bound

$$\begin{aligned} \delta_k &= \max_{p \in \Delta} \frac{1}{k+1} \sum_{\ell=0}^k \langle \nabla \Psi(p_\ell), p_\ell - p \rangle = \frac{1}{k+1} \sum_{\ell=0}^k w \left\langle \frac{Q y_\ell}{\pi}, \lambda_\ell \right\rangle - \min_{p \in \Delta} \langle s_{k+1}, p \rangle \\ &= \frac{1}{k+1} \sum_{\ell=0}^k w \sum_{j \in J(\lambda_\ell)} y_\ell^{(j)} \cdot \frac{(Q^T \lambda_\ell)^{(j)}}{\pi^{(j)}} - \min_{1 \leq i \leq n} s_{k+1}^{(i)} \\ &= \frac{1}{k+1} \sum_{\ell=0}^k w \max_{1 \leq j \leq m} \frac{(Q^T \lambda_\ell)^{(j)}}{\pi^{(j)}} - \min_{1 \leq i \leq n} \frac{(Q \bar{x}_k)^{(i)}}{\sigma^{(i)}} = \frac{1}{k+1} \sum_{\ell=0}^k \Phi(\lambda_\ell) - U(\bar{x}_k). \end{aligned}$$

The uniform bound for subgradients of Ψ is given by

$$M = \max_{p \in \Delta} \|\nabla \Psi(p)\|_\infty = \max_{\substack{x \geq 0 \\ \langle \pi, x \rangle = w}} \left\| \frac{Qx}{\sigma} \right\|_\infty = w \max_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}} \frac{|q_{i,j}|}{\sigma^{(i)} \cdot \pi^{(j)}}.$$

By applying Corollaries 1 and 5, we estimate the constant D :

$$\max_{p \in \Delta} d(p) = \max_{p \in \Delta} E^*(p) - E^*(p_0) \leq - \min_{1 \leq i \leq n} \mathbb{E}_\epsilon \left(\epsilon^{(i)} \right) + \mathbb{E}_\epsilon \left(\max_{1 \leq i \leq n} \epsilon^{(i)} \right) = D.$$

It is straightforward to see that

$$\frac{1}{k+1} \left(1 + \sum_{\ell=1}^k \frac{1}{\sqrt{\ell}} \right) \leq \frac{2}{\sqrt{k+1}}.$$

Finally, by using (5) we get:

$$\begin{aligned} \Phi(\bar{\lambda}_k) - U(\bar{x}_k) &\leq \frac{1}{k+1} \sum_{\ell=0}^k \Phi(\lambda_\ell) - U(\bar{x}_k) = \delta_k \\ &\leq \frac{D}{\sqrt{k+1}} + \frac{M^2}{2\beta} \cdot \frac{1}{k+1} \left(1 + \sum_{\ell=1}^k \frac{1}{\sqrt{\ell}} \right) = \left(D + \frac{M^2}{\beta} \right) \frac{1}{\sqrt{k+1}}. \end{aligned}$$

□

References

- [1] Ahipasaoglu, S. D., Li, X., Natarajan K.: A Convex Optimization Approach for Computing Correlated Choice Probabilities With Many Alternatives. *IEEE Transactions on Automatic Control*, **64**, 190–205 (2019).
- [2] Anderson, S. P., de Palma, A., Thisse, J.-F.: *Discrete choice theory of product differentiation*. MIT Press, Cambridge (1992).
- [3] Bresnahan, T. F., Stern, S., Trajtenberg M.: Market segmentation and the sources of rents from innovation: Personal Computers in the Late 1980s. *The RAND Journal of Economics*, **28**, 17–44 (1997).
- [4] Fosgerau, M., Melo E., de Palma, A., Shum, M.: Discrete choice and rational inattention: a general equivalence result. *Discussion Papers* 17-26, p. 31 (2017).
- [5] Gale, D., Klee, V., Rockafellar, R. T.: Convex functions on convex polytopes. *Proc. Am. Math. Soc.*, **19**, 867–873 (1968).
- [6] Kakade, S., Shalev-Shwartz, S., Tewari, A.: On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. Technical report, TTI, (2009).

- [7] Koppelman, F. S., Wen, C.-H.: The paired combinatorial logit model: Properties, estimation and application. *Transportation Research*, **34**, 75–89 (2000).
- [8] Lancaster, K. J.: A new approach to consumer theory. *Journal of Political Economy*, **74**, 132–157 (1966).
- [9] McFadden, D.: Modelling the choice of residential location. In: Karlquist A. et. al. (eds.): *Spatial Interaction Theory and Residential Location*. Transportation Research Board, North-Holland, Amsterdam, 72–77 (1978).
- [10] McFadden, D.: Econometric models of probabilistic choice. In: Manski, C., McFadden D. (eds.): *Structural Analysis of Discrete Data with Economic Applications*. MIT Press, Cambridge, 198–272 (1981).
- [11] Mishra, V., Natarajan, K., Tao, H., Teo, C.-P.: Choice prediction with semidefinite optimization when utilities are correlated. *IEEE Transactions on Automatic Control*, **57**, 2450–2463 (2012).
- [12] Nemirovsky, A. S., Yudin, D. B.: *Problem complexity and method efficiency in optimization*. Wiley, New York (1983).
- [13] Nesterov, Yu.: Primal-dual subgradient methods for convex problems. *Mathematical Programming*, **120**, 221–259 (2013).
- [14] Nesterov, Yu., Shikhman, V.: Excessive revenue model of competitive markets. In: Mordukhovich, B. S., Reich, S., Zaslavski, A. J. (eds.): *Nonlinear Analysis and Optimization, Contemporary Mathematics* **659**. AMS, Providence, 189–221, (2016).
- [15] Norets, A., Takahashi, S.: On the surjectivity of the mapping between utilities and choice probabilities. *Quantitative Economics*, **4**, 149–155 (2013).
- [16] Rockafellar, R. T.: *Convex Analysis*. Princeton University Press, Princeton, (1970).
- [17] Small, K. A.: A discrete choice model for ordered alternatives. *Econometrica*, **55**, 409–424. (1987).
- [18] Wen, C.-H., Koppelman, F. S.: The generalized nested logit model. *Transportation Research*, **35**, 627–641. (2001).