Projections onto the canonical simplex with additional linear inequalities

L. Adam a,b and V. Mácha b,c

aGuangdong Provincial Key Laboratory of Brain-inspired Intelligent Computation, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, People’s Republic of China; bFaculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic; cFaculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Prague, Czech Republic

ABSTRACT
We consider the distributionally robust optimization and show that computing the distributional worst-case is equivalent to computing the projection onto the canonical simplex with additional linear inequality. We consider several distance functions to measure the distance of distributions. We write the projections as optimization problems and show that they are equivalent to finding a zero of real-valued functions. We prove that these functions possess nice properties such as monotonicity or convexity. We design optimization methods with guaranteed convergence and derive their theoretical complexity. We demonstrate that our methods have (almost) linear observed complexity.

1. Introduction
The projection of a vector onto the unit simplex appears in various fields such as portfolio optimization [18], multi-phase physics [4], mathematical optimization [20], knapsack problem [12] or machine learning applications [5]. Given a vector $q \in \mathbb{R}^n$, this projection amounts to solving

\begin{align}
\text{minimize} \quad & \frac{1}{2} \|p - q\|^2 \\
\text{subject to} \quad & \sum_{i=1}^n p_i = 1, \\
& 0 \leq p_i, \forall i = 1, \ldots, n.
\end{align}

(1)
The Karush–Kuhn–Tucker optimality conditions imply that if one solves
\[
\sum_{i=1}^{n} \max\{q_i - \mu, 0\} = 1
\]
for \(\mu\), then one recovers the optimal solution of (1) by thresholding
\[
p_i = \max\{q_i - \mu, 0\}.
\]
This was discovered for the first time in [10] and then rediscovered many times later [17].

The simplest way to solve (2) is to sort \(q\), derive an iterative procedure computing the whole function on the left-hand side of (2) and then find when its value equals to 1. Since the second part can be done in \(O(n)\), the whole algorithm has complexity \(O(n \log n)\) due to the sorting.

This procedure was improved in numerous papers. Van Den Berg and Friedlander [24] observed that only those \(q_i\) above \(\mu\) need to be sorted in (2). Using a partially sorted structure called a heap, they managed to reduce the complexity to \(O(n + k \log n)\), where \(k\) is the number of \(q_i\) above the optimal \(\mu\). Kiwiel [13] realized that many operations in quicksort may be ignored when it is used to solve (2) and reached complexity \(O(n)\). Michelot [19] proposed a simple method based on the fixed-point theorem with observed complexity \(O(n)\). Condat [6] provided an excellent overview, pointed to some errors in previous papers and designed an improved algorithm.

Besides the standard projection (1), multiple versions appear in the literature. Adam et al. [2] considered an infinite-dimensional optimization problem with partial differential equations in the constraints. To get the independence of the number of iterations on the mesh size, they derived a path-following algorithm. Liu and Ye [16] considered sparse learning problems containing a modified simplex with two vectors of variables whose sum had to be equal. They derived an improved bisection method and a fast-converging subgradient algorithm. A similar simplex appeared in [23] for ranking labels based on feedback and in [15] for a special binary classification problem. Lapin et al. [14] considered SVM with top-k error instead of the standard top-1 error. The resulting modified simplex contained a variable upper bound which was not fixed as in all previous cases. They penalized one constraint and computed an approximate projection. Note that these problems are difficult as observed in [1]. Philpott et al. [21] considered a maximization of a linear function on reduced simplex. Their application came from financial stochastic dual dynamic programming.

In this paper, we also consider projections onto a modified simplex. Our motivation stems from the field of distributionally robust optimization [7] where one hedges against uncertainty by estimating a distribution \(q\) and considering the worse outcome when the true distribution \(p\) is not far away from \(q\). Since \(p\) is a distribution and we need to keep close to \(q\), this may be equivalently written as a projection with additional linear inequality. The projection is taken with respect to the distance between distributions and the additional linear inequality comes from considering the worst case. There are several possibilities of the considered distance function. Probably the most commonly used are the \(\phi\)-divergence [3] and the Wasserstein distance [9]. The authors in [11,21] proposed an algorithm with quadratic complexity for the \(l_2\) norm distance. In [22] the authors provided a closed-form algorithm for \(l_1\) distance.
The resulting problem is convex and depending on the used distance function, it may be even quadratic or linear. This suggests using general-purpose solvers such as CPLEX which is guaranteed to converge. In our paper, we perform a comparison with CPLEX and IPOPT and show that our algorithm exhibits the observed (almost) linear complexity and outperforms the above general-purpose solvers.

The paper is organized as follows. In Section 2 we give a brief introduction into the field of distributionally robust optimization. We derive the problems of interest and focus on measuring the distance by various $\phi$-divergences and $l_p$ norms. In Section 3 we present the main results. Instead of penalizing one constraint as in [14], we write the full KKT system and simplify it into one equation in one variable similar to (2). We derive the thresholding operator similar to (3). For readability, we postpone all proofs to the Appendix. In Section 4 we consider numerical properties and show that the derived equations have nice properties such as monotonicity or convexity. Finally, in Section 5 we focus on numerical results. All codes are available online.1

2. Motivation from distributionally robust optimization

In this section, we provide motivation for the problems from the field of distributionally robust optimization. In the classical robust optimization, one minimizes a random function $f(x, \xi)$ with respect to a decision variable $x$ while considering the worse possible random outcome of a random variable $\xi$ which is bound to lie in $\Xi$. This leads to the problem

$$\min_{x} \max_{\xi \in \Xi} f(x, \xi).$$

(4)

Since (4) considers the worst possible scenario, this approach is usually too conservative. One way to alleviate it, is to consider the distributionally robust optimization, where one takes the worst outcome with respect to all probability distributions and not to all scenarios. Denoting the probability distribution by $P$, expectation with respect to $P$ by $E_P$ and the set of all admissible probability distributions by $\mathcal{P}$, this results in

$$\min_{x} \max_{P \in \mathcal{P}} E_P f(x, \xi).$$

(5)

Note that if $\mathcal{P}$ consists of all Dirac measures concentrated at $\Xi$, then (4) and (5) coincide.

The simplest case appears when we know possible realization $\xi_i$ for the random variable and each may happen with the probability $p_i$. Then the inner maximization problem in (5) reduces to

$$\max_{P \in \mathcal{P}} E_P f(x, \xi) = \max_{P \in \mathcal{P}} \sum_{i=1}^{n} p_i f(x, \xi_i) = \max_{P \in \mathcal{P}} c^T p,$$

(6)

where we set $c_i := f(x, \xi_i)$. However, the probability distribution $p$ is often not known. Then we may assume that $p$ is close to some known estimate $q$ and we may want to hedge against the worst possible small deviation from $q$. Measuring this deviation by $\hat{d}$, the inner
problem in (5) then takes form
\[
\begin{align*}
\text{maximize} & \quad c^\top p \\
\text{subject to} & \quad \sum_{i=1}^n p_i = 1, \\
& \quad 0 \leq p_i, \ \forall \ i = 1, \ldots, n, \\
& \quad \hat{d}(p, q) \leq \varepsilon.
\end{align*}
\] (7)

The first two constraints prescribe that \( p \) is a probability distribution while the last one determines that \( p \) is not far from \( q \).

2.1. Connection to projection onto the canonical simplex

Since convex programming allows to switch the objective and constraints, for each \( \varepsilon \) there is some \( \delta \) such that problem (7) is equivalent to
\[
\begin{align*}
\text{minimize} & \quad \hat{d}(p, q) \\
\text{subject to} & \quad \sum_{i=1}^n p_i = 1, \\
& \quad 0 \leq p_i, \ \forall \ i = 1, \ldots, n, \\
& \quad c^\top p \geq \delta.
\end{align*}
\] (8)

This implies that problem (7) is equivalent to a projection (with respect to distance \( \hat{d} \)) onto the canonical simplex restricted by additional linear constraint. In Section 3.3 we will consider a similar problem where the canonical simplex is restricted by upper bounds.

2.2. Notation and assumptions

In the manuscript, we employ the following notation and assumption:
\[
\begin{align*}
c_{\text{max}} & := \max_{i=1, \ldots, n} c_i, \\
c_{\text{min}} & := \min_{i=1, \ldots, n} c_i, \\
I & := \{ i \mid c_i = c_{\text{max}} \}.
\end{align*}
\]

Assumption 2.1: We consider the following assumptions:

(A1) Vector \( q \) has positive components which sum to one.
(A2) Vector \( c \) is not a constant vector.

Assumption (A1) is natural since we want to measure the distance of \( p \) from \( q \) which is a distribution. Assumption (A2) is technical only. If \( c \) is a constant vector, then the objective \( c^\top p = c_{\text{max}} \) is constant and the optimization is trivial.
Table 1. Examples of the $\phi$-divergences $d$ and generating functions $\phi$.

| Name                  | Generating function $\phi$ | Formula $d$ |
|-----------------------|-----------------------------|-------------|
| Kullback–Leibler divergence | $\phi_1(t) = t \log t$ | $d_1(p, q) = p \log \left( \frac{p}{q} \right)$ |
| Burg entropy         | $\phi_2(t) = -\log t$     | $d_2(p, q) = q \log \left( \frac{q}{p} \right)$ |
| Hellinger distance   | $\phi_3(t) = (\sqrt{t} - 1)^2$ | $d_3(p, q) = \left( \sqrt{p} - \sqrt{q} \right)^2$ |
| $\chi^2$-distance    | $\phi_4(t) = \frac{1}{t} (t - 1)^2$ | $d_4(p, q) = \frac{(p-q)^2}{p}$ |
| Modified $\chi^2$-distance | $\phi_5(t) = (t - 1)^2$ | $d_5(p, q) = \frac{(p-q)^2}{q}$ |

Note: Note that $\phi$ is convex with $\phi(1) = 0$.

3. Reduction of projections onto modified simplex to one equation

In the previous section, we mentioned how a projection onto the unit simplex with an additional constraint arises in the field of distributionally robust optimization. In the introduction, we recalled a way of solving the projection onto the unit simplex (2). Namely, one needs to solve (2) and then apply the thresholding operator (3) to obtain the solution.

In this section, we follow a similar approach to solve problem (7) with two types of distance function $\hat{d}$. Moreover, we use the same technique to derive an algorithm for projection onto the canonical simplex with additional upper bounds. These results allow us to propose numerical methods with linear complexity.

3.1. Distributionally robust optimization with $\phi$-divergences

In this section, we consider the distributionally robust optimization where the distance function $\hat{d}$ is a $\phi$-divergence. In this case

$$\hat{d}(p, q) = \sum_{i=1}^{n} d(p_i, q_i) = \sum_{i=1}^{n} q_i \phi \left( \frac{p_i}{q_i} \right),$$

where $d$ is a $\phi$-divergence and $\phi$ is the convex generating function with $\phi(1) = 0$. Then the distributionally robust problem (7) takes form

$$\begin{align*}
\text{maximize} & \quad c^\top p \\
\text{subject to} & \quad \sum_{i=1}^{n} p_i = 1, \\
& \quad 0 \leq p_i, \quad \forall \ i = 1, \ldots, n, \\
& \quad \sum_{i=1}^{n} q_i \phi \left( \frac{p_i}{q_i} \right) \leq \varepsilon.
\end{align*}$$

(DRO1)

Some examples of $\phi$-divergences are listed in Table 1. We also consider the variation distance ($l_1$ norm). Even though it is a $\phi$-divergence, we handle it in Section 3.2 as it is a norm as well. We start with the following result which states that if $\varepsilon$ is large enough, then the solution to (DRO1) is trivial.
Theorem 3.1: Let Assumption 2.1 hold true and define vector \( \hat{\mathbf{p}} \) with components

\[
\hat{p}_i = \begin{cases} 
\frac{1}{\sum_{j \in I} q_j} q_i & \text{if } i \in I, \\
0 & \text{otherwise.}
\end{cases}
\] (10)

If this solution satisfies

\[
\sum_{i=1}^n q_i \phi \left( \frac{\hat{p}_i}{q_i} \right) \leq \varepsilon
\] (11)

then \( \hat{\mathbf{p}} \) is the optimal solution of (DRO1).

Theorem 3.1 provides the optimal solution for the case of large \( \varepsilon \). In the opposite case, we provide the solution in the following list of theorems. Each of them handles one \( \phi \)-divergence from Table 1.

**Theorem 3.2 (Kullback–Leibler divergence):** Assume that Assumption 2.1 holds true, that (11) is violated and that \( \varepsilon < -\log(\sum_{i \in I} q_i) \). Then there exists some \( \mu \in (0, (c_{\text{max}} - c_{\text{min}})/\varepsilon] \) which solves

\[
h_1(\mu) := \sum_{i=1}^n q_i \exp \left( \frac{c_i}{\mu} \right) \left( \frac{c_i}{\mu} - \log \left( \sum_{j=1}^n q_j \exp \left( \frac{c_j}{\mu} \right) \right) - \varepsilon \right) = 0.
\] (12)

Moreover, defining the non-normalized weights

\[
\hat{p}_i = q_i \exp \left( \frac{c_i}{\mu} \right),
\] (13)

the optimal solution of (DRO1) with the Kullback-Leibler divergence \( \phi = \phi_1 \) equals to

\[
p_i = \hat{p}_i / \sum \hat{p}_j.
\]

**Theorem 3.3 (Burg entropy):** Assume that Assumption 2.1 holds true and that (11) is violated. Then there exists some \( \lambda \in (c_{\text{max}}, c_{\text{max}} + (c_{\text{max}} - c_{\text{min}})/\varepsilon] \) which solves

\[
h_2(\lambda) := \sum_{i=1}^n q_i \log (\lambda - c_i) + \log \left( \sum_{i=1}^n \frac{q_i}{\lambda - c_i} \right) - \varepsilon = 0.
\] (14)

Moreover, defining the non-normalized weights

\[
\hat{p}_i = q_i \frac{1}{\lambda - c_i},
\] (15)

the optimal solution of (DRO1) with the Burg entropy distance \( \phi = \phi_2 \) equals to

\[
p_i = \hat{p}_i / \sum \hat{p}_j.
\]
Theorem 3.4 (Hellinger distance): Assume that Assumption 2.1 holds true, that (11) is violated and that \( \varepsilon < 2 - 2 \sqrt{\sum_{i \in I} q_i} \). Then there exists some \( \lambda \in (c_{\max}, c_{\max} + (2 - \varepsilon)(c_{\max} - c_{\min})/\varepsilon) \) which solves

\[
h_3(\lambda) := 2 \sum_{i=1}^{n} \frac{q_i}{\lambda - c_i} - (2 - \varepsilon) \sqrt{\frac{1}{\sum_{i=1}^{n} \frac{q_i}{(\lambda - c_i)^2}}} = 0. \tag{16}
\]

Moreover, defining the non-normalized weights

\[
\hat{p}_i = q_i \frac{1}{(\lambda - c_i)^2}, \tag{17}
\]

the optimal solution of (DRO1) with the Hellinger distance \( \phi = \phi_3 \) equals to \( p_i = \hat{p}_i/\sum \hat{p}_j \).

Theorem 3.5 (\( \chi^2 \)-distance): Assume that Assumption 2.1 holds true and that (11) is violated. Then there exists some \( \lambda > c_{\max} \) which solves

\[
h_4(\lambda) := \left( \sum_{i=1}^{n} q_i \sqrt{\lambda - c_i} \right) \left( \sum_{i=1}^{n} q_i \frac{1}{\sqrt{\lambda - c_i}} \right) - 1 - \varepsilon = 0. \tag{18}
\]

Moreover, defining the non-normalized weights

\[
\hat{p}_i = q_i \frac{1}{\sqrt{\lambda - c_i}}, \tag{19}
\]

the optimal solution of (DRO1) with the \( \chi^2 \)-distance \( \phi = \phi_4 \) equals to \( p_i = \hat{p}_i/\sum \hat{p}_j \).

Theorem 3.6 (Modified \( \chi^2 \)-distance): Assume that Assumption 2.1 holds true and that (11) is violated. Then there exists some \( \lambda > -c_{\max} \) which solves

\[
h_5(\lambda) := \sum_{i=1}^{n} q_i \max^2 (c_i + \lambda, 0) - (1 + \varepsilon) \left( \sum_{i=1}^{n} q_i \max (c_i + \lambda, 0) \right)^2 = 0. \tag{20}
\]

Moreover, defining the non-normalized weights

\[
\hat{p}_i = q_i \max (c_i + \lambda, 0), \tag{21}
\]

the optimal solution of (DRO1) with the modified \( \chi^2 \)-distance \( \phi = \phi_5 \) equals to \( p_i = \hat{p}_i/\sum \hat{p}_j \).

In Theorems 3.2 and 3.4 we impose assumptions on \( \varepsilon \). In Appendix 2 we show that if \( q \) has the uniform distribution and the largest component of \( c \) is unique, then this assumption does not actually impose any restriction.
3.2. Distributionally robust optimization with norms

Another possibility to measure the distance between \( p \) and \( q \) is to use \( l_p \) norms. This results in the following problem:

\[
\begin{align*}
\text{maximize} & \quad c^\top p \\
\text{subject to} & \quad \sum_{i=1}^{n} p_i = 1, \\
& \quad 0 \leq p_i, \quad \forall i = 1, \ldots, n, \\
& \quad \|p - q\|_p \leq \varepsilon.
\end{align*}
\]

(DRO2)

Note that the \( l_1 \) norm also generates a \( \phi \)-divergence but we handle it here. The simple algorithms for solving (DRO2) with the \( l_1 \) and \( l_\infty \) norms are presented in Appendix 1. Here, we show the results for the \( l_2 \) norm.

**Theorem 3.7 (l_2 norm):** Let Assumption 2.1 hold true and define vector \( \hat{p} \) with components

\[
\hat{p}_i = \begin{cases} 
q_i + \frac{1}{|I|} - \frac{1}{|I|} \sum_{j \in I} q_j & \text{if } i \in I, \\
0 & \text{otherwise.}
\end{cases}
\]

(22)

If this solution satisfies

\[
\|\hat{p} - q\| \leq \varepsilon,
\]

(23)

then \( \hat{p} \) is the optimal solution of (DRO2). If \( \hat{p} \) violates (23), then there exists some \( \mu > 0 \) and \( \lambda \) which solve

\[
\begin{align*}
\sum_{i=1}^{n} \min^2 \left( \lambda - c_i, \mu q_i \right) - \varepsilon^2 \mu^2 & = 0, \\
\sum_{i=1}^{n} \min \left( \lambda - c_i, \mu q_i \right) & = 0.
\end{align*}
\]

(24a)

(24b)

Moreover, the optimal solution of (DRO2) with the \( l_2 \) norm equals to

\[
p_i = \max \left( q_i - \frac{1}{\mu} (\lambda - c_i), 0 \right).
\]

(25)

Unlike in the previous cases, system (24a) consists of two equations. To reduce them to one, we define first a function \( g_6(\lambda; \mu) \) of \( \lambda \) with fixed parameter \( \mu \) by

\[
g_6(\lambda; \mu) := \sum_{i=1}^{n} \min (\lambda - c_i, \mu q_i).
\]

(26)

Lemma A.2 states that for each \( \mu > 0 \), there is unique \( \lambda \) solving \( g_6(\lambda; \mu) = 0 \). We stress this dependence of \( \lambda \) on \( \mu \) by writing \( \lambda(\mu) \). Moreover, the same lemma states that \( \lambda(\mu) \) is
a continuous function. Defining the continuous function

\[
h_6(\mu) := \sum_{i=1}^{n} \min^2(\lambda(\mu) - c_i, \mu q_i) - \varepsilon^2 \mu^2,
\]  

(27)

we observe that solving system (24a) can be reduced to solving the single equation \( h_6(\mu) = 0 \).

### 3.3. Projection onto simplex with additional bounds

In Section 2.1 we argued that (DRO2) with \( p = 2 \) is equivalent to projecting onto the canonical simplex with additional linear inequality. In this section, we consider one more problem of the projection onto the simplex with additional upper bounds. This problem in a slightly more general form reads

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \| p - q \|^2 \\
\text{subject to} & \quad \sum_{i=1}^{n} p_i = 1, \\
& \quad l_i \leq p_i \leq u_i.
\end{align*}
\]  

(SIMPLEX)

We obtain the reduced optimality conditions as follows, where clip is the projection operator.

**Theorem 3.8 (Simplex with upper bounds):** Assume that the feasible set of (SIMPLEX) is non-empty. Then there exists some \( \lambda \in \mathbb{R} \) which solves

\[
h_7(\lambda) := \sum_{i=1}^{n} \text{clip}_{[l_i, u_i]}(q_i - \lambda) - 1 = 0.
\]

Moreover, the optimal solution of (SIMPLEX) equals to

\[
p_i = \text{clip}_{[l_i, u_i]}(q_i - \lambda).
\]

Function \( h_7 \) is piecewise linear and non-increasing in \( \lambda \). This allows us to find a simple algorithm to find the solution, we present it in Algorithm 4.1.

### 4. Numerical considerations

In the previous section, we derived theoretical results which will be the bases for numerical methods for solving (DRO1), (DRO2) and (SIMPLEX). In this section, we introduce these numerical methods and derive their complexity.

#### 4.1. Computation of \( \lambda \)

For problems (DRO1) including the \( \phi \)-divergences and for (SIMPLEX), we reduced the optimality conditions into one equation in one variable. For (DRO2) with \( l_2 \) norm, we
reduced it into two equations in two variables from which \( \lambda \) is implicitly computed, further reducing the system into one equation in one variable. It is not difficult to show that this implicit Equation (24b) is equivalent to

\[
\sum_{i=1}^{n} \max (\mu q_i + c_i - \lambda, 0) - \mu = 0. \tag{30}
\]

Since (30) is identical to (2), there are algorithms in \( O(n) \) which compute \( \lambda \) for any fixed \( \mu \). For simplicity, we implemented a simpler algorithm which first sorts \( \mu q + c \) and then finds \( \lambda \) in one pass through the sorted array.

For the analysis of (SIMPLEX) we realize that \( h_7 \) is a piecewise linear function which is non-increasing in \( \lambda \). We have

\[
h_7(\min_i (q_i - u_i)) = \sum_{i=1}^{n} u_i - 1 \geq 0.
\]

The inequality holds due to the assumption that the feasible set of (SIMPLEX) is non-empty. Denote \( s \) the sorted version of \( q - l \) and \( r \) the sorted version of \( q - u \). The main idea of Algorithm 4.1 is to utilize the piecewise linearity of \( h_7 \) with kinks at \( s_i \) and \( r_j \) by tracking the current slope \( \hat{a} \). Algorithm 4.1 is an iterative procedure where at every iteration, we know the values of \( h_7(s_{i-1}) \) and \( h_7(r_{j-1}) \) and we want to evaluate \( h_7 \) at the next point. If \( r_j \leq s_i \), then we consider \( \lambda = r_j \) and increase \( j \) by one. Since a new point enters the active set which contributes to the slope, we increase the slope \( \hat{a} \) by 1. If \( r_j > s_i \), then we consider \( \lambda = s_i \) and increase \( i \) by one. Since one point leaves the active set, we decrease the slope \( \hat{a} \) by 1. In both cases, \( g \) is decreased by \( \hat{a} \) times the difference between the old value and the new values of \( \lambda \). Once \( g \) decreases below 0, we stop the algorithm and linearly interpolate between the last two values. To prevent an overflow, we set \( r_{n+1} = \infty \). Considering the initial values, since \( r_1 < s_1 \), we set \( i = 1 \) and \( j = 2 \). The sorting of \( q - l \) and \( q - u \) can be done in \( O(n \log n) \). The rest of Algorithm 4.1 consists of one pass through an array with complexity \( O(n) \). The total complexity of this algorithm is therefore \( O(n \log n) \). Since finding a zero point of \( h_7 \) differs from (2) only by the upper bound, we conjecture that there is an algorithm solving it in \( O(n) \).

### 4.2. Numerical methods

From the proofs of Theorems 3.2–3.4 we obtain

\[
\lim_{\mu \downarrow 0} h_1(\mu) = +\infty, \quad h_1\left(\frac{c_{\max} - c_{\min}}{\varepsilon}\right) \leq 0,
\]

\[
\lim_{\lambda \downarrow c_{\max}} h_2(\lambda) = +\infty, \quad h_2\left(c_{\max} + \frac{c_{\max} - c_{\min}}{\varepsilon}\right) \leq 0,
\]

\[
\lim_{\lambda \downarrow c_{\max}} h_3(\lambda) = -\infty, \quad h_3\left(c_{\max} + \frac{(2 - \varepsilon)(c_{\max} - c_{\min})}{\varepsilon}\right) \geq 0. \tag{31}
\]

This implies that the bisection method is convergent for solving \( h_1(\mu) = 0, h_2(\lambda) = 0 \) and \( h_3(\lambda) = 0 \) when starting from the bounds suggested by (31).
Algorithm 4.1 For computing $\lambda$ from (22)

1: Sort $q - l$ into $s$ and $q - u$ into $r$
2: $i \leftarrow 1$, $j \leftarrow 2$, $\hat{a} \leftarrow 1$
3: $\lambda \leftarrow r_1$, $g \leftarrow \sum_{i=1}^{n} u_i - 1$
4: while $g > 0$ do
5:   if $r_j \leq s_i$
6:      $g \leftarrow g - \hat{a}(r_j - \lambda)$
7:      $\hat{a} \leftarrow \hat{a} + 1$
8:      $\lambda \leftarrow r_j$, $j \leftarrow j + 1$
9:   else
10:      $g \leftarrow g - \hat{a}(s_i - \lambda)$
11:      $\hat{a} \leftarrow \hat{a} - 1$
12:      $\lambda \leftarrow s_i$, $i \leftarrow i + 1$
13:   end if
14: end while
15: return linear interpolation of the last two values of $\lambda$

To solve $h_4(\lambda) = 0$, $h_5(\lambda) = 0$ and $h_6(\mu) = 0$ we first observe that convexity is present due to the following result.

**Proposition 4.1:** We have the following:

- If the assumptions of Theorem 3.5 are satisfied, then $h_4$ is decreasing and convex on $(c_{\text{max}}, \infty)$.
- If the assumptions of Theorem 3.6 are satisfied, then $h_5$ is positive on $(-c_{\text{max}}, \lambda_0)$ and decreasing and concave on $(\lambda_0, \infty)$ for some $\lambda_0$.
- If the assumptions of Theorem 3.7 are satisfied, then $h_6$ is positive on $(0, \mu_0)$ and decreasing and concave on $(\mu_0, \infty)$ for some $\mu_0$.

Lemma A.3 states that if we start with a point with $h_4(\lambda) > 0$, $h_5(\lambda) < 0$ or $h_6(\mu) < 0$, respectively, the Newton’s method is convergent. Since all these three functions are monotone, such a point is simple to find, for example by starting with $\mu = 1$ and doubling its value until $h_6(\mu) < 0$. Note that general results for convergence of the Newton’s method cannot be used as the zero point needs to lie in a given interval. We summarize this discussion in Tables 2 and 3. The former shows which problem has an exact algorithm and which needs to be solved via an iterative method. Note that convergence is guaranteed for each problem. The latter comments more on the iterative methods and summarizes the equations needed to solve (DRO1) and (DRO2) with the $l_2$ norm. Moreover, it provides the bounds within which the solution lies and shows whether the function in question possesses convexity.

We summarize the whole procedure in Algorithm 4.2. The bisection method may be initialized based on the bounds from Table 3 while the Newton’s method must be initialized based on the paragraph following Proposition 4.1.
Table 2. Table showing which problems have an exact algorithm and for which problems, the bisection and Newton’s methods are convergent.

| Problem                        | Exact algorithm | Bisection | Newton |
|--------------------------------|-----------------|-----------|--------|
| (DRO1) with Kullback-Leibler divergence | ✗               | ✗         | ✗      |
| (DRO1) with Burg entropy       | ✗               | ✗         | ✗      |
| (DRO1) with Hellinger distance | ✗               | ✗         | ✗      |
| (DRO1) with $\chi^2$-distance | ✗               | ✗         | ✗      |
| (DRO1) with Modified $\chi^2$-distance | ✗               | ✗         | ✗      |
| (DRO2) with $l_1$ norm         | ✗               | ✗         | ✗      |
| (DRO2) with $l_2$ norm         | ✗               | ✗         | ✗      |
| (DRO2) with $l_{\infty}$ norm  | ✗               | ✗         | ✗      |
| (SIMPLEX)                      | ✗               | ✗         | ✗      |

Note: Note that all problems have at least one convergent algorithm.

Table 3. Properties for the bisection and Newton’s method.

| Equation                        | Bounds | Convex |
|---------------------------------|--------|--------|
| (DRO1) with Kullback-Leibler divergence | $h_1(\mu) = 0$ | $0 < \mu \leq \frac{c_{\max} - c_{\min}}{\epsilon}$ | ✗ |
| (DRO1) with Burg entropy        | $h_2(\lambda) = 0$ | $c_{\max} < \lambda \leq c_{\max} + \frac{c_{\max} - c_{\min}}{(2-\epsilon)\epsilon}$ | ✗ |
| (DRO1) with Hellinger distance  | $h_3(\lambda) = 0$ | $c_{\max} < \lambda \leq c_{\max} + \frac{c_{\max} - c_{\min}}{(2-\epsilon)\epsilon}$ | ✗ |
| (DRO1) with $\chi^2$-distance  | $h_4(\lambda) = 0$ | $c_{\max} < \lambda$ | ✗ |
| (DRO1) with Modified $\chi^2$-distance | $h_5(\lambda) = 0$ | $-c_{\max} < \lambda$ | ✗ |
| (DRO2) with $l_2$ norm          | $h_6(\mu) = 0$ | $0 < \mu$ | ✗ |

Note: The table shows which equation needs to be solved for (7), the bounds and whether the function $h$ is convex.

Algorithm 4.2 For solving (DRO1) with any $\phi$-divergence and (DRO2) with $l_2$ norm

1: Compute $\hat{p}$ from (10) or (22), respectively
2: if $\hat{p}$ satisfies (11) or (23), respectively
3: The optimal distribution $p$ equal to $\hat{p}$
4: else
5: if Kullback-Leibler divergence or Burg entropy or Hellinger distance
6: Solve $h = 0$ using the bisection method
7: else if $\chi^2$-distance or Modified $\chi^2$-distance or $l_2$ norm
8: Solve $h = 0$ using the Newton’s method
9: end if
10: Compute the optimal distribution $p$ from the corresponding Theorem 3.2–3.7
11: end if

4.3. Complexity

In Table 4 we show the complexity of the evaluation of $h_1, \ldots, h_6$ and the total complexity of the algorithm. For (DRO1), the evaluation of $h_1, \ldots, h_5$ has the complexity of $O(n)$. Similarly for (DRO2) with the $l_2$ norm, for every $\mu$, the computation of $\lambda(\mu)$ can be done in $O(n)$ as shown is Section 4.1. Thus, the evaluation of $h_6$ consumes $O(n)$ as well. We get the total complexity by multiplying this by the number of evaluations $n_h$ of $h$. In order to have a good performance, the number of evaluations $n_h$ needs to stay constant. This
Table 4. Computational complexity evaluating function $h$ and the total complexity for solving problems (DRO1), (DRO2) and (SIMPLEX). Here, $n_h$ refers to the number of evaluation of the function $h$ which is guaranteed to be constant for the bisection method whenever its bracketing interval from Table 3 is uniformly bounded. The observed complexity is shown in Table 5.

|                      | Evaluation of $h$ | Total          |
|----------------------|-------------------|----------------|
| (DRO1)               | $O(n)$            | $O(n n_h n)$   |
| (DRO2) with $l_1$ norm | $O(n)$            | $O(n n_h n)$   |
| (DRO2) with $l_2$ norm | $O(n)$            | $O(n n_h n)$   |
| (DRO2) with $l_\infty$ norm | $O(n)$         | $O(n n_h n)$   |
| (SIMPLEX)            | $O(n)$            | $O(n n_h n)$   |

Table 5. Comparison of the observed and theoretical complexity for our methods.

| Observed complexity                  | Theoretical complexity |
|--------------------------------------|------------------------|
| (DRO1) with Kullback–Leibler divergence | $3.328 \cdot 10^{-7} n^{1.102}$ | $O(n n_h)$ |
| (DRO2) with $l_1$ norm               | $2.794 \cdot 10^{-8} n^{1.125}$ | $O(n \log n)$ |
| (DRO2) with $l_2$ norm               | $3.759 \cdot 10^{-7} n^{1.056}$ | $O(n n_h)$ |
| (DRO2) with $l_\infty$ norm          | $2.802 \cdot 10^{-7} n^{1.042}$ | $O(n \log n)$ |
| (SIMPLEX)                            | $4.911 \cdot 10^{-7} n^{1.039}$ | $O(n \log n)$ |

Note: In most cases, our methods exhibit the complexity of $O(n)$ or $O(n \log n)$ which concurs with Table 4.

happened in our numerical experiments as the second row of Figure 2 shows. Moreover, $n_h$ is guaranteed to be constant for the bisection method whenever the bracketing interval stays constant. In the table, we also included problem (DRO2) with $l_1$ and $l_\infty$ norms and SIMPLEX. We provide a comparison of the theoretical and the observed complexity in Table 5 later.

5. Numerical results

In this section, we present the numerical results. We recall that our codes are available online. In Section 3, we derived the monotonicity and convexity of functions $h_1, \ldots, h_7$ corresponding to problems (DRO1), (DRO2) and (SIMPLEX) and in Section 4, we argued that finding a zero of these functions should be easy. This is confirmed in Figure 1. We see that $h_1$, which corresponds to (DRO1) with Kullback–Leibler divergence, is decreasing and seems to be convex. Similarly, $h_6$ corresponding to (DRO2) with $l_2$ norm is first increasing and after approximately $\mu = 14$ decreasing and concave. The convexity for $h_1$ was not proven while the concavity for $h_6$ follows from Proposition 4.1.

For numerical comparison, we randomly generated the initial data $q$ and $c$ and solved problems (DRO1), (DRO2) and (SIMPLEX). This was repeated a hundred times and the results were averaged to remove random bias. The main comparison is presented in Figure 2. The left column corresponds to problem (DRO1) while the right column to problems (DRO2) and (SIMPLEX). The x axis always depicts the number of input data $n$ chosen in the range $n \in [10^3, 10^6]$. The first row depicts the computational time in seconds. The second row depicts the number of evaluations of $h_1, \ldots, h_6$. 
Figure 1. Functions $h_1(\mu)$ and $h_6(\mu)$. Finding zeros of these points is equivalent to solving (DRO1) with Kullback–Leibler divergence and for (DRO2) with $l_2$ norm.

Figure 2. Performance of our methods for (DRO1) (left) and for (DRO2) and (SIMPLEX) (right) for $n \in [10^3, 10^6]$. The first row shows the measured times in seconds while the second row shows the number of evaluations of $h(\mu)$ or $h(\lambda)$. 
We observe that the number of evaluations of $h$ in the second row stays relatively constant. Coming back to Table 4, this implies that $n_h$ is constant and the total complexity should be $O(n)$ or $O(n \log n)$. This is confirmed in the first row of Figure 2 where we see the (approximately) linear dependence of time on the data size $n$. To give a more quantitative result, we have interpolated the measured times with function $t(n) = an^b$ for the best possible parameters $a$ and $b$. We show this interpolation in Table 5. We see that the interpolation is close to linear. Note that the larger power of $n$ may hide the logarithm as the domain for $n$ is bounded.

In Figure 3 we compare our results to other solvers. We used the package JuMP for Julia [8]. It employed IPOPT for (DRO1) and CPLEX for (DRO2) and (SIMPLEX). Moreover, for (DRO2) with $l_2$ norm we compared ourselves to the algorithm from [21]. To keep the computation possible, we had to reduce the number of points from $n = 10^6$ to $n = 10^4$. We can see that our algorithms perform significantly better. CPLEX and IPOPT seem to also possess linear complexity unlike the algorithm from [21] with quadratic complexity. Its complexity estimate from Table 5 was $5.138 \cdot 10^{-7} n^{1.632}$.

**Note**

1. https://github.com/VaclavMacha/Projections.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Funding**

This work was supported by the National Natural Science Foundation of China [grant number 61850410534], Guangdong Provincial Key Laboratory [grant number 2020B121201001], the Program for Guangdong Introducing Innovative and Enterpreneurial Teams [grant number 2017ZT07X386], the Grant Agency of the Czech Republic (18-21409S) and by the OP RDE (OPVVV) funded project Research Center for Informatics, reg. No: CZ02.1.01/0.0./0.0./16_019/0000765.
References

[1] L. Adam, M. Červinka, and M. Pištěk, Normally admissible stratifications and calculation of normal cones to a finite union of polyhedral sets, Set-Valued Var. Anal. 24(2) (2016), pp. 207–229.

[2] L. Adam, M. Hintermüller, and T. Surowiec, A semismooth Newton method with analytical path-following for the $H^1$-projection onto the Gibbs simplex, IMA J. Numer. Anal. 39(3) (2018), 1276–1295.

[3] G. Bayraksan and D.K. Love, Data-driven stochastic programming using phi-divergences. In The Operations Research Revolution, INFORMS, 2015, pp. 1–19.

[4] M. Blondsoe and O. Sigmund, Topology Optimization: Theory, Methods, and Applications, Springer, Berlin, 2013.

[5] M. Blondel, A. Fujino, and N. Ueda, Large-scale Multiclass Support Vector Machine Training via Euclidean Projection onto The Simplex, Proceedings of 22nd International Conference on Pattern Recognition, IEEE, 2014, pp. 1289–1294.

[6] L. Condat, Fast projection onto the simplex and the $l_1$ ball, Math. Program. 158(1–2) (2016), pp. 575–585.

[7] E. Delage and Y. Ye, Distributionally robust optimization under moment uncertainty with application to data-driven problems, Oper. Res. 58(3) (2010), pp. 595–612.

[8] I. Dunning, J. Huchette, and M. Lubin, Jump: A modeling language for mathematical optimization, SIAM Rev. 59(2) (2017), pp. 295–320.

[9] R. Gao and A.J. Kleywegt, Distributionally robust stochastic optimization with wasserstein distance, preprint (2016). Available at arXiv, math. 1604.02199.

[10] M. Held, P. Wolfe, and H.P. Crowder, Validation of subgradient optimization, Math. Program. 6(1) (1974), pp. 62–88.

[11] M.G. Kapteyn, K.E. Willcox, and A. Philpott, A Distributionally Robust Approach to Black-box Optimization, 2018 AIAA Non-Deterministic Approaches Conference, 2018, p. 0666.

[12] H. Kellerer, U. Pferschy, and D. Pisinger, Knapsack Problems, Springer, Berlin, 2013.

[13] K.C. Kiwiel, Breakpoint searching algorithms for the continuous quadratic knapsack problem, Math. Program. 112(2) (2008), pp. 473–491.

[14] M. Lapin, M. Hein, and B. Schiele, Top-k multiclass SVM. In Advances in Neural Information Processing Systems, 2015, pp. 325–333.

[15] N. Li, R. Jin, and Z.-H. Zhou, Top Rank Optimization in Linear Time, Proceedings of the 27th International Conference on Neural Information Processing Systems Vol. 1, NIPS’14, MIT Press, Cambridge, MA, USA, 2014, pp. 1502–1510.

[16] J. Liu and J. Ye, Efficient Euclidean Projections in Linear Time, Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 657–664.

[17] N. Maculan and G.G. De Paula Jr, A linear-time median-finding algorithm for projecting a vector on the simplex of $\mathbb{R}^n$, Oper. Res. Lett. 8(4) (1989), pp. 219–222.

[18] H. Markowitz, Portfolio selection, J. Finance 7(1) (1952), pp. 77–91.

[19] C. Michelot, A finite algorithm for finding the projection of a point onto the canonical simplex of $\mathbb{R}^n$, J. Optim. Theory. Appl. 50(1) (1986), pp. 195–200.

[20] J.A. Nelder and R. Mead, A simplex method for function minimization, Comput. J. 7(4) (1965), pp. 308–313.

[21] A. Philpott, V. de Matos, and L. Kapelevich, Distributionally robust SDDP, Comput. Manag. Sci. 15(3–4) (2018), pp. 431–454.

[22] H. Rahimian, G. Bayraksan, and T. Homem-de Mello, Identifying effective scenarios in distributionally robust stochastic programs with total variation distance, Math. Program. 173(1–2) (2019), pp. 393–430.
Appendices

Appendix 1. Problem (DRO2) with $l_1$ and $l_\infty$ norm

Here we present algorithms for solving (DRO2) for the $l_1$ and $l_\infty$ norms. Since $q$ is a probability distribution due to Assumption 2.1, if we increase some components of $q$, we have to decrease some components of $q$ by the same margin. The priority is on increasing coordinates of $q$ with the lowest value of $c$ while decreasing those with the largest value. We summarize this procedure in Algorithms A.1 and A.2. Sorting $c$, the lowest values have the lowest indices and similarly for the largest values. Thus, we start with $i = 1$ and $j = n$. Then we increase $q_i$ by a possible maximal margin $\delta_1$ and start decreasing $q_j$, $q_{j-1}$ and so on until the total reduction $\delta_2$ equals to $\delta_1$. After doing so, we increase $i$ by one and continue until $i = j$. Note that $\delta_{\text{dec}}$ in Algorithm A.1 measures the decrease of $p_i$ while $\delta_{\text{tot}}$ in Algorithm A.2 measures the total reduction of $p_j, \ldots, p_n$. The first one has to be bounded by $\varepsilon$ while the other one by $\varepsilon/2$.

Algorithm A.1 for solving (DRO2) with $p = \infty$

**Input:** Sorted array $c$, probabilities $q$, allowed perturbation level $\varepsilon$

1. $p \leftarrow q$, $i \leftarrow 1$, $j \leftarrow n$
2. $\delta_{\text{dec}} \leftarrow 0$
3. while $i \leq j$
   4. $\delta_1 \leftarrow \min\{1 - p_i, \varepsilon\}$, $\delta_2 \leftarrow 0$
   5. $p_i \leftarrow p_i + \delta_1$
   6. while $\delta_2 < \delta_1$
      7. if $\min\{p_j, \varepsilon - \delta_{\text{dec}}\} \geq \delta_1 - \delta_2$
         8. $p_j \leftarrow p_j - \delta_1 + \delta_2$
         9. $\delta_{\text{dec}} \leftarrow \delta_{\text{dec}} + \delta_1 - \delta_2$
         10. break (inner while)
      11. else
         12. $\delta_2 \leftarrow \delta_2 + \min\{p_j, \varepsilon - \delta_{\text{dec}}\}$
         13. $p_j \leftarrow p_j - \min\{p_j, \varepsilon - \delta_{\text{dec}}\}$
         14. $\delta_{\text{dec}} \leftarrow 0$
         15. $j \leftarrow j - 1$
      16. if $i = j$
         17. $p_i \leftarrow p_i - \delta_1 + \delta_2$
         18. break (inner while)
   19. end if
   20. end if
21. end while
22. return $p$
Algorithm A.2 for solving (DRO2) with \( p = 1 \)

**Input:** Sorted array \( c \), probabilities \( q \), allowed perturbation level \( \epsilon \)

1. \( p \leftarrow q \), \( i \leftarrow 1 \), \( j \leftarrow n \)
2. \( \delta_{\text{tot}} \leftarrow 0 \)
3. \( \textbf{while } i \leq j \) and \( \delta_{\text{tot}} \leq \frac{\epsilon}{2} \)
4. \( \delta_1 \leftarrow \min\{1 - p_i, \frac{\epsilon}{2} - \delta_{\text{tot}}\} \), \( \delta_2 \leftarrow 0 \)
5. \( \delta_{\text{tot}} \leftarrow \delta_{\text{tot}} + \delta_1 \)
6. \( p_i \leftarrow p_i + \delta_1 \)
7. \( \textbf{while } \delta_2 < \delta_1 \)
8. \( \textbf{if } p_j \geq \delta_1 - \delta_2 \)
9. \( p_j \leftarrow p_j - \delta_1 + \delta_2 \)
10. \( \textbf{break (inner while)} \)
11. \( \textbf{else} \)
12. \( \delta_2 \leftarrow \delta_2 + p_j \)
13. \( p_j \leftarrow 0 \)
14. \( j \leftarrow j - 1 \)
15. \( \textbf{if } i = j \)
16. \( p_i \leftarrow p_i - \delta_1 + \delta_2 \)
17. \( \textbf{break (inner while)} \)
18. \( \textbf{end if} \)
19. \( \textbf{end if} \)
20. \( \textbf{end while} \)
21. \( \textbf{end while} \)
22. \( \textbf{return } p \)

Appendix 2. Bounds on \( \epsilon \) in Theorems 3.2 and 3.4

In Theorems 3.2 and 3.4 we imposed assumptions on \( \epsilon \). In this section, we show that these assumptions do not impose any restrictions in an important case.

**Lemma A.1:** Consider the uniform discrete distribution \( q = (1/n, \ldots, 1/n) \) and an arbitrary discrete distribution \( p \). Then if \( \hat{d}(p, q) = \sum_{i=1}^{n} d(p_i, q_i) \) with \( d \) convex function, then we have

\[
\hat{d}(p, q) \leq \hat{d}(e_1, q),
\]

where \( e_1 = (1, 0, \ldots, 0) \).

**Proof:** Since \( p \) is a distribution, it can be expressed as a convex combination \( p = \sum_{i=1}^{n} p_i e_i \), where \( e_i \) has 1 position \( i \) and 0 elsewhere. The convexity of \( d \) implies

\[
\hat{d}(p, q) = \hat{d}\left(\sum_{i=1}^{n} p_i e_i, q\right) \leq \sum_{i=1}^{n} p_i \hat{d}(e_i, q) = \sum_{i=1}^{n} p_i \hat{d}(e_1, q) = \hat{d}(e_1, q),
\]

where we used that \( q \) has the uniform distribution and therefore \( \hat{d}(p, e_i) \) has the same value for all \( i \). \( \blacksquare \)
We fix again $q$ to be the uniform distribution and compute $\hat{d}(e_1, q)$ for the Kullback–Leibler divergence and the Hellinger distance. For the former, we have

$$\hat{d}_1(e_1, q) = 1 \log n + 0 \log 0 = \log n,$$

while in the latter, we get

$$\hat{d}_3(e_1, q) = \left(1 - \frac{1}{\sqrt{n}}\right)^2 + \sum_{i=2}^{n} \left(0 - \frac{1}{\sqrt{n}}\right)^2 = \cdots = 2 - 2 \frac{1}{\sqrt{n}}.$$ 

Since all $\phi$-divergences are convex, Lemma A.1 implies

$$\hat{d}_1(p, q) \leq \log n,$$

$$\hat{d}_3(p, q) \leq 2 - 2 \frac{1}{\sqrt{n}}.$$ 

In Theorem 3.2 and 3.4 the bounds on $\varepsilon$ were

$$\varepsilon_1 < \log n - \log |I|,$$

$$\varepsilon_3 < 2 - 2 \frac{\sqrt{|I|}}{\sqrt{n}}.$$ 

In other words, when $q$ follows the uniform distribution and when $|I| = 1$ (the largest component of $c$ is unique), then there is no point to choose $\varepsilon$ greater or equal to the bounds from these theorems because then any perturbation would be allowed.

Appendix 3. Proofs

In this section, we present all proofs. The proofs are divided into subsections as in the manuscript body. We omit the proof of Theorem 3.8 as it is similar to other proofs.

A.1 Optimality conditions for Theorems 3.1–3.6

We start with a general part which is common to all Theorems 3.1–3.6. Since $\phi$ is convex in $\phi$, since $\varepsilon > 0$, since $d(q_i, q_i) = 0$ and since $q$ defines a probability distribution, the Slater constraint qualification is satisfied at $q$. Thus, problem (DRO1) is equivalent to its KKT optimality conditions. The Lagrangian for (DRO1) reads

$$L(p; \alpha, \lambda, \mu) = -\sum_{i=1}^{n} c_i p_i - \sum_{i=1}^{n} \alpha_i p_i + \lambda \left(\sum_{i=1}^{n} p_i - 1\right) + \mu \left(\sum_{i=1}^{n} d(p_i, q_i) - \varepsilon\right).$$

The minus in front of the first term needs to be present since (DRO1) is a maximization problem. The KKT conditions then amount to the optimality conditions

$$\frac{\partial L(\cdot)}{\partial p_i} = -c_i - \alpha_i + \lambda + \mu \nabla_p d(p_i, q_i) = 0. \quad (A1a)$$

the primal feasibility conditions (DRO1), the dual feasibility conditions $\alpha_i \geq 0, \lambda \in \mathbb{R}, \mu \geq 0$ and finally the complementarity conditions

$$\alpha_i p_i = 0, \quad \forall i = 1, \ldots, n, \quad (A1b)$$

$$\mu \left(\sum_{i=1}^{n} d(p_i, q_i) - \varepsilon\right) = 0. \quad (A1c)$$

Since $\mu \geq 0$, there are two possibilities. 

Case 1 of $\mu = 0$: If $\mu = 0$, then from (A1a) we get $\lambda = \alpha_i + c_i$. This, together with $\alpha_i \geq 0$ and $\alpha_i p_i = 0$ implies that $\alpha_i = 0$ for $i \in I$ and that $p_i = 0$ for $i \notin I$. Then the KKT system is satisfied
if there exists a feasible \( \mathbf{p} \) with \( p_i = 0 \) for \( i \notin I \) such that \( \sum_{i=1}^{n} d(p_i, q_i) \leq \varepsilon \). This problem can be verified by solving the convex problem

\[
\begin{align*}
\text{minimize} & \quad \sum_{i \in I} d(p_i, q_i) \\
\text{subject to} & \quad \sum_{i \in I} p_i = 1, \\
& \quad 0 \leq p_i, \quad \forall i \in I,
\end{align*}
\]

and checking whether its optimal value is smaller or equal than \( \varepsilon - \sum_{i \notin I} d(0, q_i) \).

Since for \( \phi \)-divergences, we have \( d(p_i, q_i) = q_i \phi(p_i/q_i) \), it is not difficult to verify that the ratio \( p_i/q_i \) is constant. This implies that \( \hat{\mathbf{p}} \) defined in (10) is the solution to (A2). This finishes the proof of Theorem 3.1.

Case 2 of \( \mu > 0 \): In the opposite case we have \( \mu > 0 \), which due to (A1c) implies that the feasibility conditions change into

\[
\begin{align*}
\sum_{i=1}^{n} p_i &= 1, \\
\sum_{i=1}^{n} d(p_i, q_i) &= \varepsilon.
\end{align*}
\]

We now split the proof into five parts for Theorem 3.2–3.6.

**Proof of Theorem 3.2:** The feasibility constraint may be due to Assumption 2.1 written as

\[
\sum_{i=1}^{n} \left( p_i \log \left( \frac{p_i}{q_i} \right) - p_i \right) = \varepsilon - 1.
\]

For now we assume that \( p_i > 0 \) for all \( i \) and remove this assumption later. This implies \( \alpha_i = 0 \). Then the optimality condition (A1a) reads

\[- c_i + \lambda + \mu \log p_i - \mu \log q_i = 0,
\]

from which we deduce

\[
p_i = q_i \exp \left( \frac{c_i - \lambda}{\mu} \right).
\]

Plugging (A4) into the feasibility conditions (A3) yields

\[
\begin{align*}
\sum_{i=1}^{n} q_i \exp \left( \frac{c_i - \lambda}{\mu} \right) &= 1, \\
\sum_{i=1}^{n} q_i \exp \left( \frac{c_i - \lambda}{\mu} \right) \frac{c_i - \lambda}{\mu} &= \varepsilon.
\end{align*}
\]

We can express \( \mu \) from (A5) via

\[
\sum_{i=1}^{n} q_i \exp \left( \frac{c_i}{\mu} \right) = \exp \left( \frac{\lambda}{\mu} \right),
\]

which together with (A6) gives the final Equation (12). The optimal probabilities (13) then follow from (A4).
Now we need to remove the assumption of \( p_i > 0 \). Recall that
\[
h_1(\mu) = \sum_{i=1}^{n} q_i \exp\left( \frac{c_i}{\mu} \right) \left( \frac{c_i}{\mu} - \log \left( \sum_{j=1}^{n} q_j \exp\left( \frac{c_j}{\mu} \right) \right) - \varepsilon \right).
\]

For its middle part we have
\[
\frac{c_i}{\mu} - \log \left( \sum_{j=1}^{n} q_j \exp\left( \frac{c_j}{\mu} \right) \right) \leq \frac{c_i}{\mu} - \log \left( \sum_{j \in I} q_j \exp\left( \frac{c_j}{\mu} \right) \right) \leq \frac{c_i}{\mu} - \frac{c_{\text{min}}}{\mu} \leq \frac{c_{\text{max}} - c_{\text{min}}}{\mu},
\]
which implies that
\[
h_1(\mu) \leq \sum_{i=1}^{n} q_i \exp\left( \frac{c_i}{\mu} \right) \left( \frac{c_{\text{max}} - c_{\text{min}}}{\mu} - \varepsilon \right) \leq 0 \quad \text{whenever } \mu \geq \frac{c_{\text{max}} - c_{\text{min}}}{\varepsilon}. \tag{A7}
\]

We consider now the limit of \( h_1(\mu) \) as \( \mu \downarrow 0 \). Due to the properties of the exponential function, for all \( \alpha > 0 \), there is some \( \mu_0 \) such that for all \( \mu \in (0, \mu_0) \) we have
\[
-\log \left( \sum_{j=1}^{n} q_j \exp\left( \frac{c_j}{\mu} \right) \right) \geq -\log \left( \frac{1+\alpha}{\mu} \right),
\]
which implies
\[
h_1(\mu) \geq \sum_{i=1}^{n} q_i \exp\left( \frac{c_i}{\mu} \right) \left( \frac{c_i}{\mu} - \log(1+\alpha) - \log \left( \sum_{j \in I} q_j \right) - \frac{c_{\text{max}}}{\mu} \right).
\]

This implies
\[
h_1(\mu) \rightarrow \infty \quad \text{as } \mu \downarrow 0.
\]

Proof of Theorem 3.3: The form of \( \phi_2 \) implies \( p_i > 0 \) for all \( i \), which further means \( \alpha_i = 0 \). Then the optimality condition \((A1a)\) reads
\[
-c_i + \lambda - \mu \frac{q_i}{p_i} = 0,
\]
from which we deduce
\[
p_i = q_i \frac{\mu}{\lambda - c_i}. \tag{A8}
\]
Plugging \((A8)\) into the feasibility conditions \((A3)\) yields
\[
\sum_{i=1}^{n} q_i \frac{\mu}{\lambda - c_i} = 1, \tag{A9}
\]
\[
\sum_{i=1}^{n} q_i \log (\lambda - c_i) = \varepsilon + \log \mu. \tag{A10}
\]

We can express \(\mu\) from (A9) via
\[
\sum_{i=1}^{n} \frac{q_i}{\lambda - c_i} = \frac{1}{\mu},
\]
which together with (A10) gives the final Equation (14). The optimal probabilities (15) then follow from (A8). The constraint \(\mu > 0\) transfers to \(\lambda > c_{\text{max}}\) due to (A8).

Now we are interested in the limits. Recall that
\[
h_2(\lambda) = \sum_{i=1}^{n} q_i \log (\lambda - c_i) + \log \left( \sum_{i=1}^{n} \frac{q_i}{\lambda - c_i} \right) - \varepsilon.
\]

Then we have
\[
h_2(\lambda) \leq \sum_{i=1}^{n} q_i \log (\lambda - c_{\text{min}}) + \log \left( \sum_{i=1}^{n} \frac{q_i}{\lambda - c_{\text{max}}} \right) - \varepsilon = \log (\lambda - c_{\text{min}}) - \log (\lambda - c_{\text{max}}) - \varepsilon
\]
\[
= \log \left( \frac{\lambda - c_{\text{min}}}{\lambda - c_{\text{max}}} \right) - \varepsilon = \log \left( 1 + \frac{c_{\text{max}} - c_{\text{min}}}{\lambda - c_{\text{max}}} \right) - \varepsilon \leq \frac{c_{\text{max}} - c_{\text{min}}}{\lambda - c_{\text{max}}} - \varepsilon
\]

Thus
\[
h_2(\lambda) \leq 0 \quad \text{whenever} \quad \lambda \geq c_{\text{max}} + \frac{c_{\text{max}} - c_{\text{min}}}{\varepsilon}. \tag{A11}
\]

We consider now the limit of \(h_2(\lambda)\) as \(\lambda \downarrow c_{\text{max}}\). Denoting \(c_{\text{max}2}\) the second largest distinct component value of \(c\), we have
\[
h_2(\lambda) \geq \sum_{i \in I} q_i \log (\lambda - c_i) + \sum_{i \notin I} q_i \log (c_{\text{max}} - c_{\text{max}2}) + \log \left( \sum_{i \in I} \frac{q_i}{\lambda - c_i} \right) - \varepsilon
\]
\[
\geq \sum_{i \in I} q_i \log (\lambda - c_i) + \sum_{i \notin I} q_i \log (c_{\text{max}} - c_{\text{max}2}) + \log \left( \sum_{i \in I} \frac{q_i}{\lambda - c_i} \right) - \varepsilon
\]
\[
= \left( \sum_{i \in I} q_i - 1 \right) \log (\lambda - c_{\text{max}}) + \sum_{i \notin I} q_i \log (c_{\text{max}} - c_{\text{max}2}) + \log \left( \sum_{i \in I} q_i \right) - \varepsilon
\]

Due to the assumptions of Theorem 3.6, we obtain \(h_2(\lambda) \to \infty\) as \(\lambda \downarrow c_{\text{max}}\). This combined with (A11) and the continuity of \(h_2\) implies that the equation \(h_2(\lambda) = 0\) has a solution on \((c_{\text{max}}, c_{\text{max}} + (c_{\text{max}} - c_{\text{min}})/\varepsilon)]\). \(\blacksquare\)

**Proof of Theorem 3.4:** The feasibility constraint may due to Assumption 2.1 be written as
\[
-2 \sum_{i=1}^{n} \sqrt{p_i q_i} = \varepsilon - 2.
\]

For now we assume that \(p_i > 0\) for all \(i\) and remove this assumption later. This implies \(\alpha_i = 0\). Then the optimality condition (A1a) reads
\[
- c_i + \lambda - \mu \sqrt{\frac{q_i}{p_i}} = 0, \tag{A12}
\]
from which we deduce
\[
p_i = q_i \left( \frac{\mu^2}{(\lambda - c_i)^2} \right). \tag{A13}
\]
Plugging (A13) into the feasibility conditions (A3) yields

\[ \sum_{i=1}^{n} q_i \frac{\mu^2}{(\lambda - c_i)^2} = 1, \quad (A14) \]

\[ 2 \sum_{i=1}^{n} q_i \frac{\mu}{|\lambda - c_i|} = 2 - \varepsilon. \quad (A15) \]

We can express \( \mu \) from (A14) via

\[ \sum_{i=1}^{n} q_i \frac{\mu^2}{(\lambda - c_i)^2} = 1 \]

which together with (A15) gives the final Equation (16). The optimal probabilities (17) then follow from (A13). The constraint \( \mu > 0 \) transfers to \( \lambda > c_{\text{max}} \) due to (A12), which also allows us to remove the absolute value from (A15).

Now we need to remove the assumption of \( p_i > 0 \). Recall that

\[ h_3(\lambda) = 2 \sum_{i=1}^{n} \frac{q_i}{\lambda - c_i} - (2 - \varepsilon) \sqrt{\sum_{i=1}^{n} \frac{q_i}{(\lambda - c_i)^2}}. \]

Then we have

\[ h_3(\lambda) \geq 2 \sum_{i=1}^{n} \frac{q_i}{\lambda - c_{\text{min}}} - (2 - \varepsilon) \sqrt{\sum_{i=1}^{n} \frac{q_i}{(\lambda - c_{\text{max}})^2}} = \frac{2}{\lambda - c_{\text{min}}} - \frac{2 - \varepsilon}{\lambda - c_{\text{max}}} \]

\[ = \frac{2(\lambda - c_{\text{max}}) - (2 - \varepsilon)(\lambda - c_{\text{min}})}{(\lambda - c_{\text{min}})(\lambda - c_{\text{max}})} = \frac{\varepsilon\lambda - 2(c_{\text{max}} - c_{\text{min}}) - \varepsilon c_{\text{min}}}{(\lambda - c_{\text{min}})(\lambda - c_{\text{max}})}. \]

Thus

\[ h_3(\lambda) \geq 0 \quad \text{whenever} \quad \lambda \geq c_{\text{min}} + \frac{2(c_{\text{max}} - c_{\text{min}})}{\varepsilon} = c_{\text{max}} + \frac{(2 - \varepsilon)(c_{\text{max}} - c_{\text{min}})}{\varepsilon}. \quad (A16) \]

We consider now the limit of \( h_3(\lambda) \) as \( \lambda \downarrow c_{\text{max}} \). Denoting \( c_{\text{max}}^2 \) the second largest distinct component value of \( c \), we have

\[ h_3(\lambda) \leq 2 \sum_{i \in I} \frac{q_i}{\lambda - c_i} + 2 \sum_{i \notin I} \frac{q_i}{\lambda - c_i} - (2 - \varepsilon) \sqrt{\sum_{i \in I} \frac{q_i}{(\lambda - c_i)^2}} \]

\[ \leq \frac{2}{\lambda - c_{\text{max}}} + \frac{2}{c_{\text{max}} - c_{\text{max}}^2} - (2 - \varepsilon) \frac{2}{\lambda - c_{\text{max}}} \]

\[ \leq \frac{2}{\lambda - c_{\text{max}}} - (2 - \varepsilon) \frac{\sum_{i \in I} q_i}{\varepsilon} + \frac{2}{c_{\text{max}} - c_{\text{max}}^2}. \]

Due to the assumptions of Theorem 3.4, we obtain \( h_3(\lambda) \to -\infty \) as \( \lambda \downarrow c_{\text{max}} \). This combined with (A16) and the continuity of \( h_3 \) implies that the equation \( h_3(\lambda) = 0 \) has a solution on \( (c_{\text{max}}, c_{\text{max}} + (2 - \varepsilon)(c_{\text{max}} - c_{\text{min}})/\varepsilon) \]. Since the optimality conditions are equivalent to problem (DRO1) due to convexity, the existence of solution also implies that the assumption of \( p_i > 0 \) may be alleviated. ■
Proof of Theorem 3.5: The form of $\phi_4$ implies $p_i > 0$ for all $i$, which further means $\alpha_i = 0$. Due to Assumption 2.1, the feasibility constraints on distance amounts to

$$\sum_{i=1}^{n} \frac{q_i^2}{p_i} = 1 + \varepsilon.$$ 

Then the optimality condition (A1a) reads

$$-c_i + \lambda - \mu \frac{q_i^2}{p_i^2} = 0,$$

from which we deduce

$$p_i = q_i \sqrt{\frac{\mu}{\lambda - c_i}}.$$ (A17)

Plugging (A17) into the feasibility conditions (A3) yields

$$\sum_{i=1}^{n} q_i \sqrt{\frac{\mu}{\lambda - c_i}} = 1,$$ (A18)

$$\sum_{i=1}^{n} q_i \sqrt{\frac{\lambda - c_i}{\mu}} = 1 + \varepsilon.$$ (A19)

We can express $\mu$ from (A19) via

$$\frac{1}{1 + \varepsilon} \sum_{i=1}^{n} q_i \sqrt{\lambda - c_i} = \sqrt{\mu},$$

which together with (A18) gives the final Equation (18). The optimal probabilities (19) then follow from (A17). The constraint $\mu > 0$ transfers to $\lambda > c_{\text{max}}$ due to (A17).

Proof of Theorem 3.6: The feasibility constraint may due to Assumption 2.1 be written as

$$\sum_{i=1}^{n} \frac{p_i^2}{q_i} = 1 + \varepsilon.$$ 

Then the optimality condition (A1a) reads (for the uniformity of results, we flipped the sign of $\lambda$)

$$-c_i - \alpha_i - \lambda + 2\mu \frac{p_i}{q_i} = 0,$$

from which we due to the complementarity conditions (A1b) deduce

$$p_i = \frac{1}{2\mu} q_i \max (c_i + \lambda, 0).$$ (A20)

Plugging (A20) into the feasibility conditions (A3) yields

$$\frac{1}{2\mu} \sum_{i=1}^{n} q_i \max (c_i + \lambda, 0) = 1,$$ (A21)

$$\frac{1}{4\mu^2} \sum_{i=1}^{n} q_i \max^2 (c_i + \lambda, 0) = 1 + \varepsilon.$$ (A22)
We can express $\mu$ from (A21) via
\[
\frac{1}{2} \sum_{i=1}^{n} q_i \max (c_i + \lambda, 0) = \mu,
\]
which together with (A22) gives the final Equation (20). The optimal probabilities (21) then follow from (A20). The constraint $\mu > 0$ transfers to $\lambda > -c_{\text{max}}$ due to (A20).

### A.2 Optimality conditions for Theorem 3.7

To prove Theorem 3.7, we first realize that the first part of Section A.1 including the first paragraph of 'Case 1 of $\mu = 0'$ holds true with $d(p_i, q_i) = \frac{1}{2}(p_i - q_i)^2$. Moreover, problem (A2) takes the form

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \sum_{i \in I} (p_i - q_i)^2 \\
\text{subject to} & \quad \sum_{i \in I} p_i = 1, \\
& \quad 0 \leq p_i, \forall i \in I.
\end{align*}
\]

(A23)

Since $q_i > 0$ and $\sum_{i \in I} q_i < 1$ due to Assumption 2.1, it is not difficult to verify that $\hat{p}$ defined in (22) is the optimal solution of (A23). This proves the first part of Theorem 3.7.

For the second part, we realize that the feasibility constraint may be written as
\[
\frac{1}{2} \sum_{i=1}^{n} (p_i - q_i)^2 = \frac{1}{2} \varepsilon^2.
\]

Then the optimality condition (A1a) reads
\[
-c_i - \alpha_i + \lambda + \mu (p_i - q_i) = 0,
\]
from which we due to the primal feasibility condition $p_i \geq 0$, the dual feasibility conditions $\alpha_i \geq 0$ and the complementarity condition (A1b) deduce
\[
p_i = \max \left( q_i - \frac{1}{\mu} (\lambda - c_i), 0 \right). \tag{A24}
\]

Plugging (A24) into the feasibility conditions (A3) yields
\[
\sum_{i=1}^{n} \max \left( q_i - \frac{1}{\mu} (\lambda - c_i), 0 \right) = 1, \tag{A25}
\]
\[
\sum_{i=1}^{n} \left( \max \left( q_i - \frac{1}{\mu} (\lambda - c_i), 0 \right) - q_i \right)^2 = \varepsilon^2. \tag{A26}
\]

The final Equations (24a) are obtained from (A25) and (A26) by simple calculus, the formula $\max(-x, -y) = -\min(x, y)$ and Assumption 2.1.

### A.3 Convexity for Proposition 4.1

We first show an auxiliary result which states the continuity on $h_6$.

**Lemma A.2:** For each $\mu > 0$ there is a unique $\lambda(\mu)$ which solves $g_6(\lambda; \mu) = 0$. Moreover, function $h_6(\mu)$ is continuous on $(0, \infty)$. 
Proof: Fix any \( \mu > 0 \) and recall that
\[
g_{6}(\lambda; \mu) = \sum_{i=1}^{n} \min(\lambda - c_i, \mu q_i).
\]
Since \( \mu > 0 \) for at least one \( i \) we have \( \lambda(\mu) - c_i < \mu q_i \) which implies that \( g_{6} \) is strictly increasing in \( \lambda \) around \( \lambda(\mu) \). Thus, the solution \( \lambda(\mu) \) is unique.

Consider now any \( \mu_k \to \mu > 0 \). Since the corresponding \( \lambda(\mu_k) \) are bounded, we may select a converging subsequence, say \( \lambda(\mu_k) \to \lambda^{*} \). Then
\[
0 = \sum_{i=1}^{n} \min(\lambda(\mu_k) - c_i, \mu_k q_i) \to \sum_{i=1}^{n} \min(\lambda^{*} - c_i, \mu q_i),
\]
which implies that \( \lambda^{*} = \lambda(\mu) \). Thus, \( \lambda \) is a continuous function of \( \mu \). This further implies that \( h_{6} \) is a continuous function of \( \mu \). \( \blacksquare \)

Concerning the proof of Proposition 4.1, we divide it into three parts.
Proof for \( h_{4} \): After rearranging of terms, we get
\[
h_{4}(\lambda) = \sum_{i=1}^{n} \sum_{j=1}^{n} q_{ij} g_{ij}(\lambda) - 1 - \varepsilon,
\]
where for \( i, j = 1, \ldots, n \) we define
\[
g_{ij}(\lambda) := (\lambda - c_i)^{\frac{1}{2}} (\lambda - c_j)^{-\frac{1}{2}}.
\]
Computing the first derivative
\[
g_{ij}'(\lambda) = \frac{1}{2}(\lambda - c_i)^{-\frac{1}{2}} (\lambda - c_j)^{-\frac{1}{2}} - \frac{1}{2}(\lambda - c_i)^{\frac{1}{2}} (\lambda - c_j)^{-\frac{1}{2}}
= \frac{1}{2}(\lambda - c_i)^{-\frac{1}{2}} (\lambda - c_j)^{-\frac{1}{2}} (\lambda - c_j - (\lambda - c_i))
= \frac{1}{2}(\lambda - c_i)^{-\frac{1}{2}} (\lambda - c_j)^{-\frac{1}{2}} (c_i - c_j)
\]
and the second derivative
\[
g_{ij}''(\lambda) = \frac{1}{2} (c_i - c_j) \left( -\frac{1}{2}(\lambda - c_i)^{-\frac{3}{2}} (\lambda - c_j)^{-\frac{1}{2}} - \frac{3}{2}(\lambda - c_i)^{-\frac{1}{2}} (\lambda - c_j)^{-\frac{3}{2}} \right)
= -\frac{1}{4} (c_i - c_j) (\lambda - c_i)^{-\frac{3}{2}} (\lambda - c_j)^{-\frac{1}{2}} + 3(\lambda - c_i)^{-\frac{1}{2}} (\lambda - c_j)^{-\frac{5}{2}} (\lambda - c_j + 3(\lambda - c_i))
= -\frac{1}{4} (c_i - c_j)(\lambda - c_i)^{-\frac{3}{2}} (\lambda - c_j)^{-\frac{5}{2}} (4\lambda - 3c_i - c_j),
\]
we realize that
\[
g_{ij}'(\lambda) + g_{ij}''(\lambda)
= -\frac{1}{4} (c_i - c_j)(\lambda - c_i)^{-\frac{3}{2}} (\lambda - c_j)^{-\frac{5}{2}} (4\lambda - 3c_i - c_j)(\lambda - c_i) - (4\lambda - c_i - 3c_j)(\lambda - c_j)
= -\frac{1}{4} (c_i - c_j)(\lambda - c_i)^{-\frac{3}{2}} (\lambda - c_j)^{-\frac{5}{2}} (3(c_i - c_j)(2\lambda - c_i - c_j))
= \frac{3}{4} (c_i - c_j)^{2}(\lambda - c_i)^{-\frac{3}{2}} (\lambda - c_j)^{-\frac{5}{2}} (2\lambda - c_i - c_j) \geq 0.
\]
This together with (A27) implies that $h_4$ is convex.

**Proof for $h_5$:** Recall that

$$h_5(\lambda) = \sum_{i=1}^{n} q_i \max^2 (c_i + \lambda, 0) - (1 + \varepsilon) \left( \sum_{i=1}^{n} q_i \max (c_i + \lambda, 0) \right)^2. \tag{A28}$$

For simplicity assume that $-c_1 < \cdots < -c_n$. Then on $\lambda \in (-c_j, -c_{j+1})$ we have

$$h_5(\lambda) = \sum_{i=1}^{j} q_i (c_i + \lambda)^2 - (1 + \varepsilon) \left( \sum_{i=1}^{j} q_i (c_i + \lambda) \right)^2.$$

The derivative at this interval equals to

$$h_5' = 2 \sum_{i=1}^{j} q_i (c_i + \lambda) - 2(1 + \varepsilon) \left( \sum_{i=1}^{j} q_i (c_i + \lambda) \right) \sum_{i=1}^{j} q_i$$

$$= 2 \sum_{i=1}^{j} q_i (c_i + \lambda) \left( 1 - (1 + \varepsilon) \sum_{i=1}^{j} q_i \right)$$

$$= 2 \left( \sum_{i=1}^{n} q_i (c_i + \lambda, 0) \right) \left( 1 - (1 + \varepsilon) \sum_{i=1}^{j} q_i \right). \tag{A29}$$

Due to the assumption of $-c_1 < \cdots < -c_n$, the violation of (11) implies $(1 + \varepsilon)q_1 < 1$. This due to (A29) means that $h_5' > 0$ on $\lambda \in (-c_1, -c_2)$. Since $h_5(-c_1) = 0$ due to (A28), this implies that $h_5$ is positive on $(-c_1, -c_2)$. Moreover, $h_5'$ equals to a product of two terms, the first one is always positive while the second one is piecewise constant with decreasing values on individual pieces. This implies that there is some $\lambda_0$ such that $h_5$ is non-decreasing on $(-c_{\text{max}}, \lambda_0)$ and decreasing on $(\lambda_0, \infty)$. By the same arguments we have that $h_5'$ is decreasing on $(\lambda_0, \infty)$, which implies that $h_5$ is concave on this interval. This finishes the proof of the second part.

If $-c_1 \leq \cdots \leq -c_n$ instead of the assumed $-c_1 < \cdots < -c_n$, then the proof can be performed in exactly the same way but $(1 + \varepsilon)q_1 < 1$ changes to $(1 + \varepsilon) \sum_{i=1}^{j} q_i < 1$, where $j$ is the cardinal of $I$. Then we can get the same estimates of (A29) as in the previous paragraphs.

**Proof for $h_6$:** Define

$$I(\mu) = \{ i \mid \lambda(\mu) - c_i < \mu q_i \} \tag{A30}$$

and consider any $0 < \mu_1 < \mu_2$. Since $\lambda(\mu_1)$ solves (24b) for $\mu = \mu_1$ and $\lambda(\mu_2)$ solves the same equation for $\mu = \mu_2$, we obtain $\lambda(\mu_1) \geq \lambda(\mu_2)$. This due to (A30) implies $I(\mu_1) \subseteq I(\mu_2)$. Thus, $I(\mu)$ is a non-decreasing function (with respect to set inclusion) of $\mu$.

From (24b) we have

$$0 = \sum_{i=1}^{n} \min \left( \lambda(\mu) - c_i, \mu q_i \right) = \sum_{i \in I(\mu)} (\lambda(\mu) - c_i) + \sum_{i \notin I(\mu)} \mu q_i$$

$$= |I(\mu)| \lambda(\mu) - \sum_{i \in I(\mu)} c_i + \mu \sum_{i \notin I(\mu)} q_i,$$

from which we deduce

$$\lambda(\mu) = \frac{1}{|I(\mu)|} \left( \sum_{i \in I(\mu)} c_i - \mu \sum_{i \notin I(\mu)} q_i \right). \tag{A31}$$
Since $I(\mu)$ is a non-decreasing function of $\mu$, this implies that $\lambda(\mu)$ is a piecewise linear function with a finite number of pieces. On each of these pieces, we have

$$\lambda'(\mu) = -\frac{1}{|I(\mu)|} \sum_{i \in I(\mu)} q_i$$

and consequently

$$h_6(\mu) = \sum_{i \in I(\mu)} (\lambda(\mu) - c_i)^2 + \sum_{i \notin I(\mu)} \mu^2 q_i^2 - \varepsilon^2 \mu^2.$$  

Differentiating this relation yields

$$\frac{1}{2} h_6'(\mu) = \lambda'(\mu) \sum_{i \in I(\mu)} (\lambda(\mu) - c_i) + \mu \sum_{i \notin I(\mu)} q_i^2 - \mu \varepsilon^2$$

$$= -\lambda'(\mu) \sum_{i \in I(\mu)} c_i + \lambda'(\mu) |I(\mu)| \lambda(\mu) + \mu \sum_{i \notin I(\mu)} q_i^2 - \mu \varepsilon^2$$

$$= -\lambda'(\mu) \sum_{i \in I(\mu)} c_i + \lambda'(\mu) \left( \sum_{i \in I(\mu)} c_i - \mu \sum_{i \notin I(\mu)} q_i \right) + \mu \sum_{i \notin I(\mu)} q_i^2 - \mu \varepsilon^2$$

$$= -\mu \lambda'(\mu) \sum_{i \notin I(\mu)} q_i + \mu \sum_{i \notin I(\mu)} q_i^2 - \mu \varepsilon^2$$

$$= \mu \left( \frac{1}{|I(\mu)|} \sum_{i \notin I(\mu)} q_i \sum_{i \notin I(\mu)} q_i + \sum_{i \notin I(\mu)} q_i^2 - \varepsilon^2 \right), \quad \text{(A33)}$$

where in the third equality we used (A31) and in the last one (A32).

Denote $c_{\max 2}$ the second largest distinct component value of $\varepsilon$ and define $\hat{\lambda} = \frac{1}{2} (c_{\max} + c_{\max 2})$. Fix any $\mu$ sufficiently small but positive. Then we have

$$g_6(\hat{\lambda}, \mu) = \sum_{i=1}^n \min \left( \hat{\lambda} - c_i, \mu q_i \right) = \sum_{i \in I} \min \left( \hat{\lambda} - c_i, \mu q_i \right) + \sum_{i \notin I} \min \left( \hat{\lambda} - c_i, \mu q_i \right)$$

$$= \frac{1}{2} |I| (c_{\max 2} - c_{\max}) + \sum_{i \notin I} \min \left( \hat{\lambda} - c_i, \mu q_i \right) \leq \frac{1}{2} |I| (c_{\max 2} - c_{\max}) + \mu \sum_{i \notin I} q_i < 0,$n

whenever $\mu$ is sufficiently small. Since $g_6(\lambda(\mu); \mu) = 0$ due to definition and since $g_6$ is non-decreasing in $\lambda$, this means that $\lambda(\mu) \geq \hat{\lambda} = \frac{1}{2} (c_{\max} + c_{\max 2})$. This due to (A30) implies $I(\mu) = I$ whenever $\mu$ is sufficiently small. It is possible to show that the violation of (23) is equivalent to

$$\frac{1}{|I|} \sum_{i \notin I} q_i \sum_{i \notin I} q_i + \sum_{i \notin I} q_i^2 - \varepsilon^2 > 0.$$

Combining this with (A33) and $I(\mu) = I$, this implies that $h_6$ is strictly increasing on some interval $(0, \mu_1)$. Moreover, since $\lambda(0) = c_{\max}$, we have $I(0) = \emptyset$ and thus $h_6(0) = 0$. This implies that $h_6$ is positive on $(0, \mu_1)$.

Moreover, $h_6'$ equals to a product of two terms, the first one is always positive while the second one is piecewise constant with decreasing values on individual pieces. This implies that there is some $\mu_0$ such that $h_6$ is non-decreasing on $(0, \mu_0)$ and decreasing on $(\mu_0, \infty)$. By the same arguments, we have that $h_6''$ is decreasing on $(\mu_0, \infty)$, which implies that $h_6$ is concave on this interval. This finishes the proof of the second part.

**Appendix 4. Convergence of Newton’s method**

In this section, we state a convergence result of the Newton’s method.
Lemma A.3: Consider a continuous concave function \( h : [a, b] \to \mathbb{R} \) with \( h(a) > 0 \) and \( h(b) < 0 \). Consider the Newton’s updates

\[
\lambda^{k+1} = \lambda^k - \frac{h(\lambda^k)}{\mu^k},
\]

where \( \mu^k \) is any element of the convex subdifferential \(-\partial(-h)(\lambda^k)\). If this update is started from any point \( \lambda^0 \in (a, b) \) with \( h(\lambda^0) < 0 \), then \( \lambda^k \) is a non-increasing sequence which converges to some \( \bar{\lambda} \in (a, b) \) with \( h(\bar{\lambda}) = 0 \).

The proof immediately follows from the following simpler formulation for convex function by considering \(-h\) instead of \( h \).

Lemma A.4: Consider a continuous convex function \( h : [a, b] \to \mathbb{R} \) with \( h(a) < 0 \) and \( h(b) > 0 \). Consider the Newton’s updates

\[
\lambda^{k+1} = \lambda^k - \frac{h(\lambda^k)}{\mu^k},
\]

where \( \mu^k \) is any element of the convex subdifferential \( \partial h(\lambda^k) \). If this update is started from any point \( \lambda^0 \in (a, b) \) with \( h(\lambda^0) > 0 \), then \( \lambda^k \) is a non-increasing sequence which converges to some \( \lambda \in (a, b) \) with \( h(\lambda) = 0 \).

Proof: Due to convexity of \( h \) and \( h(a) < 0 \) with \( h(b) > 0 \), there exists unique \( \lambda^* \in (a, b) \) with \( h(\lambda^*) = 0 \). Since \( h(a) < 0 \) and \( h(\lambda^*) = 0 \), convexity implies that for all \( \lambda \geq \lambda^* \) all elements of \( \partial h(\lambda) \) are positive.

We will show that for all \( k \) we have \( h(\lambda^k) \geq 0 \) and \( \lambda^k \geq \lambda^* \). We will use the proof by induction. Since due to the assumptions we have \( h(\lambda^0) > 0 \) and \( \lambda^0 > \lambda^* \), we need only to make the induction step \( k \to k + 1 \). Assume therefore that \( h(\lambda^k) \geq 0 \) and \( \lambda^k \geq \lambda^* \). Due to the convexity and \( \mu^k \in \partial h(\lambda^k) \) we have

\[
h(\lambda^{k+1}) \geq h(\lambda^k) + \mu^k(\lambda^{k+1} - \lambda^k).
\]

Since the right-hand side equals to 0 due to Newton’s step (A34), we get \( h(\lambda^{k+1}) \geq 0 \). The previous paragraph and the assumed \( \lambda^k \geq \lambda^* \) implies \( \mu^k > 0 \). Then the Newton’s step results in \( \lambda^{k+1} \leq \lambda^k \). Now we will show that \( \lambda^{k+1} \geq a \). If this is not true then \( \lambda^{k+1} \leq a < \lambda^* \leq \lambda^k \) and the convexity implies \( h(a) \geq \min\{h(\lambda^{k+1}), h(\lambda^k)\} \). Since the left-hand side is negative due to the lemma assumption and the right-hand side is positive, we arrived at a contradiction and therefore \( a \leq \lambda^{k+1} \leq \lambda^k \). But since \( h(\lambda) < 0 \) for all \( \lambda \in (a, \lambda^*) \) and since \( h(\lambda^{k+1}) \geq 0 \), we may strengthen the previous into \( \lambda^* \leq \lambda^{k+1} \leq \lambda^k \). This shows the induction step and therefore for all \( k \) we have \( h(\lambda^k) \geq 0 \) and \( \lambda^* \leq \lambda^k \). Moreover, we also showed \( \lambda^* \leq \lambda^{k+1} \leq \lambda^* \).

Since \( (\lambda^k) \) is a non-increasing sequence which is bounded below by \( \lambda^* \), it is convergent. At the same time, all elements of \( \partial h(\lambda^k) \) are uniformly bounded below by zero as the subdifferential is a monotone operator and all elements of \( \partial h(\lambda^*) \) are positive. This implies

\[
\frac{h(\lambda^k)}{g^k} = \lambda^k - \lambda^{k+1} \to 0.
\]

The uniform boundedness of \( g^k \) from zero implies \( h(\lambda^k) \to 0 \). But this due to the continuity of \( h \) means that \( \lambda^k \to \bar{\lambda} \) with \( h(\bar{\lambda}) = 0 \).