Projections onto the canonical simplex with additional linear inequalities

Lukáš Adam*[1,2] and V. Mácha[2]

[1]Southern University of Science and Technology, Shenzhen 518055, China
[2]ÚTIA, The Czech Academy of Sciences, Pod Vodárenskou věží 4, 18208, Prague, Czech Republic

November 6, 2019

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.
Keywords: projection; simplex; distributionally robust optimization; linear obseved 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 $\boldsymbol{q} \in \mathbb{R}^n$, this projection amounts to solving

$$\begin{aligned} \underset{\boldsymbol{p}}{\text{minimize}}\ & \frac{1}{2}\|\boldsymbol{p}-\boldsymbol{q}\|^2 \\ \text{subject to}\ & \sum_{i=1}^{n} p_i = 1, \\ & 0 \leq p_i, \quad \forall i = 1,\dots,n. \end{aligned} \tag{1}$$

The Karush-Kuhn-Tucker optimality conditions imply that if one solves

$$\sum_{i=1}^{n} \max\{q_i - \mu, 0\} = 1 \tag{2}$$

for μ, then one recovers the optimal solution of (1) by thresholding

$$p_i = \max\{q_i - \mu, 0\}. \tag{3}$$

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 $\boldsymbol{q}$, derive an iterative procedure computing the whole function on the left-hand side of

*adam@utia.cas.cz

(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. [24] observed that only those q_i above μ need to be sorted in (2). Using a partially sorted structure called heap, they managed to reduce the complexity to $O(n + k \log n)$, where k is the number of q_i above the optimal μ. [13] realized that many operations in quicksort may be ignored when it is used to solve (2) and reached complexity $O(n)$. [19] proposed a simple method based on the fixed-point theorem with observed complexity $O(n)$. [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. [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. [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. [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]. [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 $\boldsymbol{q}$ and considering the worse outcome when the true distribution $\boldsymbol{p}$ is not far away from $\boldsymbol{q}$. Since $\boldsymbol{p}$ is a distribution and we need to keep close to $\boldsymbol{q}$, 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 ϕ-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 exhibit the observed (almost) linear complexity and outperform 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 ϕ-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 $f(\boldsymbol{x}, \boldsymbol{\xi})$ with respect to a decision variable $\boldsymbol{x}$ while considering the worse possible random outcome of a random variable $\boldsymbol{\xi}$ which is bound to lie in Ξ. This leads to the problem

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \underset{\boldsymbol{\xi} \in \Xi}{\text{maximize}} \, f(\boldsymbol{x}, \boldsymbol{\xi}). \tag{4}$$

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

[1] `https://github.com/VaclavMacha/Projections`

respect to all probability distributions and not to all scenarios. Denoting the probability distribution by P, expectation with respect to P by $\mathbb{E}_P$ and the set of all admissible probability distributions by $\mathscr{P}$, this results in

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \underset{P\in\mathscr{P}}{\text{maximize}}\, \mathbb{E}_P f(\boldsymbol{x},\boldsymbol{\xi}). \tag{5}$$

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

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

$$\underset{\boldsymbol{p}\in\mathscr{P}}{\text{maximize}}\, \mathbb{E}_P f(\boldsymbol{x},\boldsymbol{\xi}) = \underset{\boldsymbol{p}\in\mathscr{P}}{\text{maximize}} \sum_{i=1}^{n} p_i f(\boldsymbol{x},\boldsymbol{\xi}_i) = \underset{\boldsymbol{p}\in\mathscr{P}}{\text{maximize}}\, \boldsymbol{c}^\top \boldsymbol{p}, \tag{6}$$

where we set $c_i := f(\boldsymbol{x},\boldsymbol{\xi}_i)$. However, the probability distribution $\boldsymbol{p}$ is often not known. Then we may assume that $\boldsymbol{p}$ is close to some known estimate $\boldsymbol{q}$ and we may want to hedge against the worst possible small deviation from $\boldsymbol{q}$. Measuring this deviation bu $\hat{d}$, the inner problem in (5) then takes form

$$\begin{aligned} &\underset{\boldsymbol{p}}{\text{maximize}}\ \boldsymbol{c}^\top \boldsymbol{p} \\ &\text{subject to} \sum_{i=1}^{n} p_i = 1, \\ &\qquad 0 \le p_i, \quad \forall i = 1,\dots,n, \\ &\qquad \hat{d}(\boldsymbol{p},\boldsymbol{q}) \le \varepsilon, \end{aligned} \tag{7}$$

The first two constraints prescribe that $\boldsymbol{p}$ is a probability distribution while the last one determines that $\boldsymbol{p}$ is not far from $\boldsymbol{q}$.

2.1 Connection to projection onto the canonical simplex

Since convex programming allows to switch the objective and constraints, for each ε there is some δ such that problem (7) is equivalent to

$$\begin{aligned} &\underset{\boldsymbol{p}}{\text{minimize}}\ \hat{d}(\boldsymbol{p},\boldsymbol{q}) \\ &\text{subject to} \sum_{i=1}^{n} p_i = 1, \\ &\qquad 0 \le p_i, \quad \forall i = 1,\dots,n, \\ &\qquad \boldsymbol{c}^\top \boldsymbol{p} \ge \delta. \end{aligned} \tag{8}$$

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:

$$\begin{aligned} c_{\max} &:= \max_{i=1,\dots,n} c_i, \\ c_{\min} &:= \min_{i=1,\dots,n} c_i, \\ I &:= \{i \mid c_i = c_{\max}\}. \end{aligned}$$

Assumption 2.1. We consider the following assumptions:

(A1) Vector $\boldsymbol{q}$ has positive components which sum to one.

(A2) Vector $\boldsymbol{c}$ is not a constant vector.

Assumption (A1) is natural since we want to measure the distance of $\boldsymbol{p}$ from $\boldsymbol{q}$ which is a distribution. Assumption (A2) is technical only. If $\boldsymbol{c}$ is a constant vector, then the objective $\boldsymbol{c}^\top \boldsymbol{p} = c_{\max}$ is constant and the optimization is trivial.

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 the equation (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 the 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 ϕ-divergences

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

$$\hat{d}(\boldsymbol{p},\boldsymbol{q}) = \sum_{i=1}^{n} d(p_i,q_i) = \sum_{i=1}^{n} q_i \phi\left(\frac{p_i}{q_i}\right), \tag{9}$$

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

$$\begin{aligned} &\underset{\boldsymbol{p}}{\text{maximize}}\ \boldsymbol{c}^\top \boldsymbol{p} \\ &\text{subject to} \sum_{i=1}^{n} p_i = 1, \\ &\qquad 0 \le p_i, \quad \forall i = 1,\dots,n, \\ &\qquad \sum_{i=1}^{n} q_i \phi\left(\frac{p_i}{q_i}\right) \le \varepsilon, \end{aligned} \tag{DRO1}$$

Some examples of ϕ-divergences are listed in Table 1. We also consider the variation distance (l_1 norm). Even though it is a ϕ-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 ε is large enough, then the solution to (DRO1) is trivial.

Table 1: Examples of the ϕ-divergences d and generating functions ϕ. Note that ϕ is convex with $\phi(1) = 0$.

Name	Generating function ϕ	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	$\phi_3(t) = (\sqrt{t}-1)^2$	$d_3(p,q) = \left(\sqrt{p}-\sqrt{q}\right)^2$
χ^2-distance	$\phi_4(t) = \frac{1}{t}(t-1)^2$	$d_4(p,q) = \frac{(p-q)^2}{p}$
Modified χ^2-distance	$\phi_5(t) = (t-1)^2$	$d_5(p,q) = \frac{(p-q)^2}{q}$

Theorem 3.1. *Let Assumption 2.1 hold true and define vector $\hat{\boldsymbol{p}}$ with components*

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

If this solution satisfies

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

then $\hat{\boldsymbol{p}}$ *is the optimal solution of* (DRO1).

Theorem 3.1 provides the optimal solution for the case of large ε. In the opposite case, we provide the solution in the following list of theorems. Each of them handles one ϕ-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* $\mu \in (0, \frac{c_{\max}-c_{\min}}{\varepsilon}]$ *which solves*

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

Moreover, defining the non-normalized weights

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

the optimal solution of (DRO1) *with the Kullback-Leibler divergence* $\phi = \phi_1$ *equals to* $p_i = \frac{\hat{p}_i}{\sum \hat{p}_j}$.

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

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

Moreover, defining the non-normalized weights

$$\hat{p}_i = q_i \frac{1}{\lambda - c_i}, \tag{15}$$

the optimal solution of (DRO1) *with the Burg entropy distance* $\phi = \phi_2$ *equals to* $p_i = \frac{\hat{p}_i}{\sum \hat{p}_j}$.

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* $\lambda \in (c_{\max}, c_{\max} + \frac{(2-\varepsilon)(c_{\max}-c_{\min})}{\varepsilon}]$ *which solves*

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

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* $p_i = \frac{\hat{p}_i}{\sum \hat{p}_j}$.

Theorem 3.5 (χ^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* χ^2*-distance* $\phi = \phi_4$ *equals to* $p_i = \frac{\hat{p}_i}{\sum \hat{p}_j}$.

Theorem 3.6 (Modified χ^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 χ^2-distance $\phi = \phi_5$ equals to $p_i = \frac{\hat{p}_i}{\sum \hat{p}_j}$.*

3.2 Distributionally robust optimization with norms

Anothor possibility to measure the distance between $\boldsymbol{p}$ and $\boldsymbol{q}$ is to use l_p norms. This results in the following problem

$$\begin{aligned} &\underset{\boldsymbol{p}}{\text{maximize}}\ \boldsymbol{c}^\top \boldsymbol{p} \\ &\text{subject to} \sum_{i=1}^{n} p_i = 1, \\ &\qquad 0 \le p_i, \quad \forall i = 1, \dots, n, \\ &\qquad \|\boldsymbol{p} - \boldsymbol{q}\|_p \le \varepsilon, \end{aligned} \tag{DRO2}$$

Note that the l_1 norm also generates a ϕ-divergence but we handle it here. The simple algorithms for solving (DRO2) with the l_1 and l_∞ norms are presented in Appendix A. 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 $\hat{\boldsymbol{p}}$ with components*

$$\hat{p}_i = \begin{cases} q_i + \frac{1}{|I|} - \frac{1}{|I|}\sum_{i \in I} q_j & \text{if } i \in I, \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

If this solution satisfies

$$\|\hat{\boldsymbol{p}} - \boldsymbol{q}\| \le \varepsilon, \tag{23}$$

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

$$\sum_{i=1}^{n} \min^2(\lambda - c_i, \mu q_i) - \varepsilon^2 \mu^2 = 0, \tag{24a}$$

$$\sum_{i=1}^{n} \min(\lambda - c_i, \mu q_i) = 0. \tag{24b}$$

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

$$p_i = \max\left(q_i - \frac{1}{\mu}(\lambda - c_i), 0\right). \tag{25}$$

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

$$g_6(\lambda; \mu) := \sum_{i=1}^{n} \min(\lambda - c_i, \mu q_i). \tag{26}$$

Lemma B.1 states that for each $\mu > 0$, there is unique λ solving $g_6(\lambda; \mu) = 0$. We stress this dependence of λ on μ by writing $\lambda(\mu)$. Moreover, the same lemma states that $\lambda(\mu)$ is a continuous function. Defining the continuous function

$$h_6(\mu) := \sum_{i=1}^{n} \min^2(\lambda(\mu) - c_i, \mu q_i) - \varepsilon^2 \mu^2, \tag{27}$$

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

3.3 Projection onto simplex with additional linear equality

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

$$\begin{aligned} \underset{\boldsymbol{p}}{\text{minimize}}\ & \frac{1}{2}\|\boldsymbol{p}-\boldsymbol{q}\|^2 \\ \text{subject to}\ & \sum_{i=1}^{n} p_i = 1, \\ & l_i \le p_i \le u_i \end{aligned} \tag{SIMPLEX}$$

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*

$$h_7(\lambda) := \sum_{i=1}^{n} \operatorname{clip}_{[l_i,u_i]}(q_i - \lambda) - 1 = 0. \tag{28}$$

Moreover, the optimal solution of (SIMPLEX) *equals to*

$$p_i = \operatorname{clip}_{[l_i,u_i]}(q_i - \lambda). \tag{29}$$

Function h_7 is piecewise linear and non-increasing in λ. 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 λ

For problems (DRO1) including the ϕ-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 λ 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 λ for any fixed μ. For simplicity, we implemented a simpler algorithm which first sorts $\mu\boldsymbol{q}+\boldsymbol{c}$ and then finds λ 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 λ. Since this problem differs from (2) only by the upper bound, we conjecture that there is an algorithm solving it in $O(n)$. Here, we present an algorithm with complexity $O(n \log n)$. We have

$$h_7\big(\min_i(q_i - u_i)\big) = \sum_{i=1}^{n} u_i - 1 \ge 0.$$

The inequality holds due to the assumption that the feasible set of (SIMPLEX) is non-empty.

Denote $\boldsymbol{s}$ the sorted version of $\boldsymbol{q}-\boldsymbol{l}$ and $\boldsymbol{r}$ the sorted version of $\boldsymbol{q}-\boldsymbol{u}$. 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 λ. Once g decreases below 0, we stop the algorithm and linearly interpolate between the last two values. To prevent an overflow, we set $r_{n+1} = \infty$. Concerning the initial values, since $r_1 < s_1$, we set $i = 1$ and $j = 2$.

Algorithm 4.1 For computing λ from (28)

Sort $\boldsymbol{q} - \boldsymbol{l}$ into $\boldsymbol{s}$ and $\boldsymbol{q} - \boldsymbol{u}$ into $\boldsymbol{r}$
$i \leftarrow 1$, $j \leftarrow 2$, $\hat{a} \leftarrow 1$
$\lambda \leftarrow r_1$, $g \leftarrow \sum_{i=1}^{n} u_i - 1$
while $g > 0$ **do**
    if $r_j \leq s_i$ **then**
        $g \leftarrow g - \hat{a}(r_j - \lambda)$
        $\hat{a} \leftarrow \hat{a} + 1$
        $\lambda \leftarrow r_j$, $j \leftarrow j + 1$
    else
        $g \leftarrow g - \hat{a}(s_i - \lambda)$
        $\hat{a} \leftarrow \hat{a} - 1$
        $\lambda \leftarrow s_i$, $i \leftarrow i + 1$
    end if
end while
return linear interpolation of the last two values of λ

4.2 Numerical methods

From the proofs of Theorems 3.2-3.4 we obtain

$$\begin{aligned} \lim_{\mu \downarrow 0} h_1(\mu) &= +\infty, & h_1\left(\frac{c_{\max} - c_{\min}}{\varepsilon}\right) &\leq 0, \\ \lim_{\lambda \downarrow c_{\max}} h_2(\lambda) &= +\infty, & h_2\left(c_{\max} + \frac{c_{\max} - c_{\min}}{\varepsilon}\right) &\leq 0, \\ \lim_{\lambda \downarrow c_{\max}} h_3(\lambda) &= -\infty, & h_3\left(c_{\max} + \frac{(2-\varepsilon)(c_{\max} - c_{\min})}{\varepsilon}\right) &\geq 0. \end{aligned} \tag{31}$$

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).

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 $(c_{\max}, \infty)$.*
- *If the assumptions of Theorem 3.6 are satisfied, then h_5 is positive on $(-c_{\max}, \lambda_0)$ and decreasing and concave on (λ_0, ∞) for some λ_0.*
- *If the assumptions of Theorem 3.7 are satisfied, then h_6 is positive on $(0, \mu_0)$ and decreasing and concave on (μ_0, ∞) for some μ_0.*

Lemma C.1 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. 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

Table 2: Table showing which problems have an exact algorithm and for which problems, the bisection and Newton's methods are convergent. Note that all problems have at least one convergent algorithm.

		Guaranteed convergence	
	Exact algorithm	Bisection	Newton
(DRO1) with Kullback-Leibler divergence	✗	✓	✗
(DRO1) with Burg entropy	✗	✓	✗
(DRO1) with Hellinger distance	✗	✓	✗
(DRO1) with χ^2-distance	✗	✓	✓
(DRO1) with Modified χ^2-distance	✗	✓	✓
(DRO2) with l_1 norm	✓	·	·
(DRO2) with l_2 norm	✗	✓	✓
(DRO2) with l_∞ norm	✓	·	·
(SIMPLEX)	✓	·	·

Table 3: Properties for the bisection and Newton's method. The table shows which equation needs to be solved for (7), the bounds and whether the function h is convex.

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

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.

Algorithm 4.2 For solving (DRO1) with any ϕ-divergence and (DRO2) with l_2 norm

Compute $\hat{\boldsymbol{p}}$ from (10) or (22), respectively
if $\hat{\boldsymbol{p}}$ satisfies (11) or (23), respectively **then**
    The optimal distribution $\boldsymbol{p}$ equal to $\hat{\boldsymbol{p}}$
else
    if Kullback-Leibler divergence **or** Burg entropy **or** Hellinger distance **then**
        Solve $h = 0$ using the bisection method
    else if χ^2-distance **or** Modified χ^2-distance **or** l_2 norm **then**
        Solve $h = 0$ using the Newton's method
    end if
    Compute the optimal distribution $\boldsymbol{p}$ from the corresponding Theorem 3.2-3.7
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 μ, 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 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_∞ norms and SIMPLEX. We provide a comparison of the theoretical and the observed complexity in Table 5 later.

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 later.

	Evaluation of h	Total
(DRO1)	$O(n)$	$O(n_h n)$
(DRO2) with l_1 norm	–	$O(n \log n)$
(DRO2) with l_2 norm	$O(n)$	$O(n_h n)$
(DRO2) with l_∞ norm	–	$O(n \log n)$
(SIMPLEX)	–	$O(n \log n)$

5 Numerical results

In this section, we present the numerical results. We recall that our codes are available online.[1] 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.

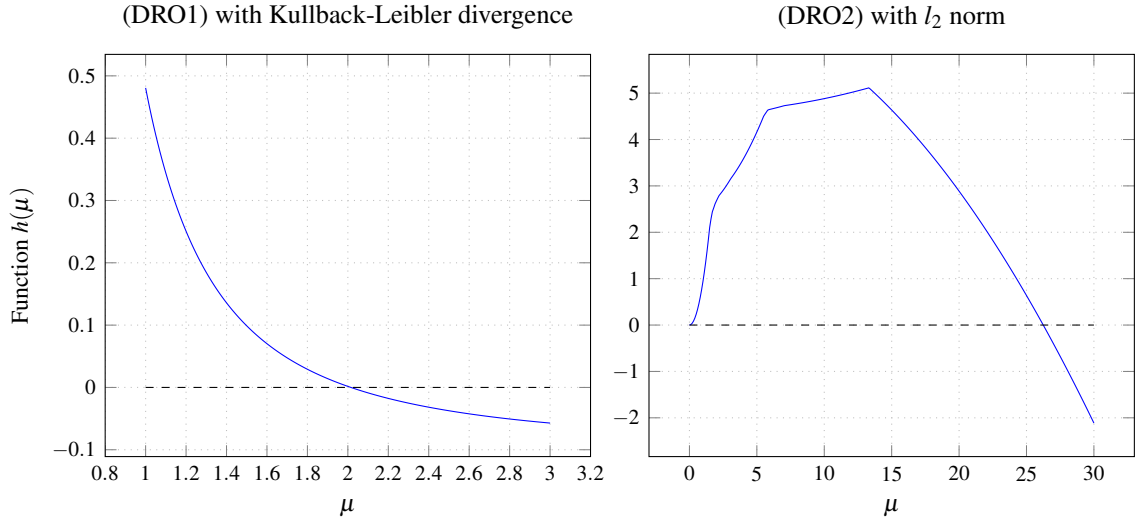

Figure 1: Functions $h_1(\mu)$ and $h_6(\mu)$. Finding zeros of these points is equivalent to solving (DRO1) with Kullback-Leibler divergence and for (DRO2) with l_2 norm.

For numerical comparison, we randomly generated the initial data $\boldsymbol{q}$ and $\boldsymbol{c}$ and solved problems (DRO1), (DRO2) and (SIMPLEX). This was repeated 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

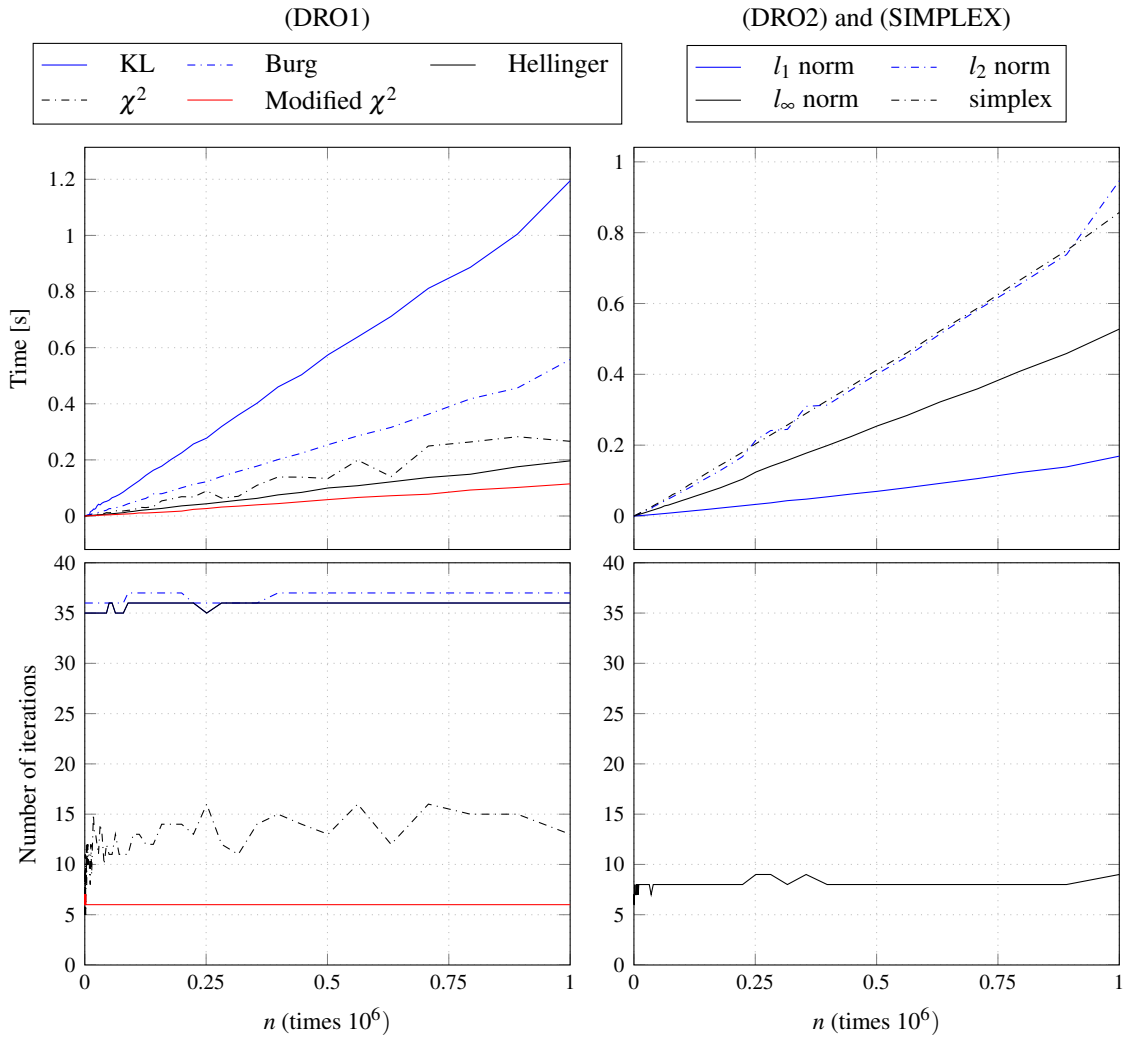

Figure 2: Performance of our methods for (DRO1) (left) and for (DRO2) and (SIMPLEX) (right) for $n \in [10^3, 10^6]$. The first row shows the measured times in seconds while the second row show the number of evaluations of $h(\mu)$ or $h(\lambda)$.

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 $h_1, \ldots, h_6$.

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.

Table 5: Comparison of the observed and theoretical complexity for our methods. In most cases our methods exhibit the complexity of $O(n)$ or $O(n \log n)$ which concurs with Table 4.

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

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 $5.138 \cdot 10^{-7} n^{1.632}$.

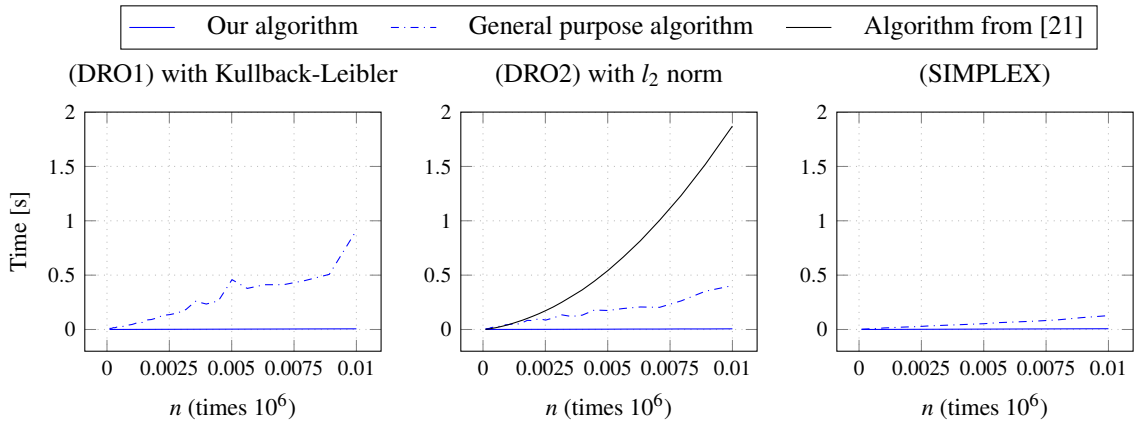

Figure 3: Comparison of our method, general-purpose solvers (IPOPT and CPLEX) and the algorithm from [21].

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant No. 61850410534), the Program for Guangdong Introducing Innovative and Enterpreneurial Teams (Grant No. 2017ZT07X386),

Shenzhen Peacock Plan (Grant No. KQTD2016112514355531) and the Grant Agency of the Czech Republic (18-21409S).

A Problem (DRO2) with l_1 and l_∞ norm

Here we present algorithms for solving (DRO2) for the l_1 and l_∞ norms. Since $\boldsymbol{q}$ is a probability distribution due to Assumption 2.1, if we increase some components of $\boldsymbol{q}$, we have to decrease some components of $\boldsymbol{q}$ by the same margin. The priority is on increasing coordinates of $\boldsymbol{q}$ with the lowest value of $\boldsymbol{c}$ while decreasing those with the largest value. We summarize this procedure in Algorithms A.1 and A.2. Sorting $\boldsymbol{c}$, 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 δ_1 and start decreasing q_j, q_{j-1} and so on until the total reduction δ_2 equals to δ_1. After doing so, we increase i by one and continue until $i = j$. Note that δ_{dec} in Algorithm A.1 measures the decrease of p_j while δ_{tot} in Algorithm A.2 measures the total reduction of $p_j, \dots, p_n$. The first one has to be bounded by ε while the other one by $\frac{\varepsilon}{2}$.

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

Input: Sorted array $\boldsymbol{c}$, probabilities $\boldsymbol{q}$, allowed perturbation level ε

$\boldsymbol{p} \leftarrow \boldsymbol{q}$, $i \leftarrow 1$, $j \leftarrow n$
$\delta_{\text{dec}} \leftarrow 0$
while $i \leq j$ **do**
    $\delta_1 \leftarrow \min\{1 - p_i, \varepsilon\}$, $\delta_2 \leftarrow 0$

    $p_i \leftarrow p_i + \delta_1$
    while $\delta_2 < \delta_1$ **do**
        if $\min\{p_j, \varepsilon - \delta_{\text{dec}}\} \geq \delta_1 - \delta_2$ **then**
            $p_j \leftarrow p_j - \delta_1 + \delta_2$
            $\delta_{\text{dec}} \leftarrow \delta_{\text{dec}} + \delta_1 - \delta_2$
            break (inner while)
        else
            $\delta_2 \leftarrow \delta_2 + \min\{p_j, \varepsilon - \delta_{\text{dec}}\}$
            $p_j \leftarrow p_j - \min\{p_j, \varepsilon - \delta_{\text{dec}}\}$
            $\delta_{\text{dec}} \leftarrow 0$
            $j \leftarrow j - 1$
            if i==j **then**
                $p_i \leftarrow p_i - \delta_1 + \delta_2$
                break (inner while)
            end if
        end if
    end while
end while
return $\boldsymbol{p}$

Algorithm A.2 for solving (DRO2) with $p = 1$

Input: Sorted array $\boldsymbol{c}$, probabilities $\boldsymbol{q}$, allowed perturbation level ε

$\boldsymbol{p} \leftarrow \boldsymbol{q}$, $i \leftarrow 1$, $j \leftarrow n$
$\delta_{\text{tot}} \leftarrow 0$
while $i \leq j$ **and** $\delta_{\text{tot}} \leq \frac{\varepsilon}{2}$ **do**
    $\delta_1 \leftarrow \min\{1 - p_i, \frac{\varepsilon}{2} - \delta_{\text{tot}}\}$, $\delta_2 \leftarrow 0$
    $\delta_{\text{tot}} \leftarrow \delta_{\text{tot}} + \delta_1$
    $p_i \leftarrow p_i + \delta_1$
    while $\delta_2 < \delta_1$ **do**
        if $p_j \geq \delta_1 - \delta_2$ **then**
            $p_j \leftarrow p_j - \delta_1 + \delta_2$

            break (inner while)
        else
            $\delta_2 \leftarrow \delta_2 + p_j$
            $p_j \leftarrow 0$

            $j \leftarrow j - 1$
            if i==j **then**
                $p_i \leftarrow p_i - \delta_1 + \delta_2$
                break (inner while)
            end if
        end if
    end while
end while
return $\boldsymbol{p}$

B 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.

B.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 ϕ is convex in ϕ, since $\varepsilon > 0$, since $d(q_i, q_i) = 0$ and since $\boldsymbol{q}$ defines a -probability distribution, the Slater constraint qualification is satisfied at $\boldsymbol{q}$. Thus, problem (DRO1) is equivalent to its KKT optimality conditions. The Lagrangian for (DRO1) reads

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

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

$$\frac{\partial L(\cdot)}{\partial p_i} = -c_i - \alpha_i + \lambda + \mu \nabla_p d(p_i, q_i) = 0. \tag{32a}$$

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

$$\alpha_i p_i = 0, \quad \forall i = 1, \ldots, n, \tag{32b}$$

$$\mu \left(\sum_{i=1}^{n} d(p_i, q_i) - \varepsilon \right) = 0. \tag{32c}$$

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

Case 1 of $\mu = 0$: If $\mu = 0$, then from (32a) 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 $\boldsymbol{p}$ 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

$$\begin{aligned} \underset{\boldsymbol{p}}{\text{minimize}} \ & \sum_{i \in I} d(p_i, q_i) \\ \text{subject to} \ & \sum_{i \in I} p_i = 1, \\ & 0 \leq p_i, \quad \forall i \in I, \end{aligned} \tag{33}$$

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

Since for ϕ-divergences, we have $d(p_i, q_i) = q_i \phi(\frac{p_i}{q_i})$, it is not difficult to verify that the ratio $\frac{p_i}{q_i}$ is constant. This implies that $\hat{\boldsymbol{p}}$ defined in (10) is the solution to (33). This finishes the proof of Theorem 3.1.

Case 2 of $\mu > 0$: In the opposite case we have $\mu > 0$, which due to (32c) implies that the feasibility conditions change into

$$\begin{aligned} \sum_{i=1}^{n} p_i &= 1, \\ \sum_{i=1}^{n} d(p_i, q_i) &= \varepsilon. \end{aligned} \tag{34}$$

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 (32a) reads

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

from which we deduce

$$p_i = q_i \exp\left(\frac{c_i - \lambda}{\mu}\right). \tag{35}$$

Plugging (35) into the feasibility conditions (34) yields

$$\sum_{i=1}^{n} q_i \exp\left(\frac{c_i - \lambda}{\mu}\right) = 1, \tag{36}$$

$$\sum_{i=1}^{n} q_i \exp\left(\frac{c_i - \lambda}{\mu}\right) \frac{c_i - \lambda}{\mu} = \varepsilon. \tag{37}$$

We can express μ from (36) via

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

which together with (37) gives the final equation (12). The optimal probabilities (13) then follow from (35).

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

$$\frac{c_i}{\mu} - \log\left(\sum_{j=1}^{n} q_j \exp\left(\frac{c_j}{\mu}\right)\right) \leq \frac{c_i}{\mu} - \log\left(\sum_{j=1}^{n} q_j \exp\left(\frac{c_{\min}}{\mu}\right)\right) = \frac{c_i}{\mu} - \frac{c_{\min}}{\mu} \leq \frac{c_{\max} - c_{\min}}{\mu},$$

which implies that

$$h_1(\mu) \leq \sum_{i=1}^{n} q_i \exp\left(\frac{c_i}{\mu}\right) \left(\frac{c_{\max} - c_{\min}}{\mu} - \varepsilon\right) \leq 0 \qquad \text{whenever} \qquad \mu \geq \frac{c_{\max} - c_{\min}}{\varepsilon}. \tag{38}$$

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 μ_0 such that for all $\mu \in (0, \mu_0)$ we have

$$-\log\left(\sum_{j=1}^{n} q_j \exp\left(\frac{c_j}{\mu}\right)\right) \geq -\log\left((1+\alpha)\sum_{j\in I} q_j \exp\left(\frac{c_j}{\mu}\right)\right) = -\log(1+\alpha) - \log\left(\sum_{j\in I} q_j\right) - \frac{c_{\max}}{\mu}.$$

This implies

$$h_1(\mu) \geq \sum_{i=1}^{n} q_i \exp\left(\frac{c_i}{\mu}\right) \left(\frac{c_i}{\mu} - \log(1+\alpha) - \log\left(\sum_{j\in I} q_j\right) - \frac{c_{\max}}{\mu} - \varepsilon\right).$$

The right-most term is positive and independent of μ whenever $i \in I$ and α is sufficiently small due to the assumptions of Theorem 3.2. Moreover as

$$\exp\left(\frac{c_{\max}}{\mu}\right) \gg \exp\left(\frac{c_i}{\mu}\right) \frac{1}{\mu}$$

for all $i \notin I$, we deduce that $h_1(\mu) \to \infty$ as $\mu \downarrow 0$. This combined with (38) and the continuity of h_1 implies that the equation $h_1(\mu) = 0$ has a solution on $(0, \frac{c_{\max} - c_{\min}}{\varepsilon}]$. 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 3.3. The form of ϕ_2 implies $p_i > 0$ for all i, which further means $\alpha_i = 0$. Then the optimality condition (32a) reads

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

from which we deduce

$$p_i = q_i \frac{\mu}{\lambda - c_i}. \tag{39}$$

Plugging (39) into the feasibility conditions (34) yields

$$\sum_{i=1}^{n} q_i \frac{\mu}{\lambda - c_i} = 1, \tag{40}$$

$$\sum_{i=1}^{n} q_i \log(\lambda - c_i) = \varepsilon + \log \mu. \tag{41}$$

We can express μ from (40) via

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

which together with (41) gives the final equation (14). The optimal probabilities (15) then follow from (39). The constraint $\mu > 0$ transfers to $\lambda > c_{\max}$ due to (39).

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

$$\begin{aligned} h_2(\lambda) &\le \sum_{i=1}^{n} q_i \log(\lambda - c_{\min}) + \log\left(\sum_{i=1}^{n} \frac{q_i}{\lambda - c_{\max}}\right) - \varepsilon = \log(\lambda - c_{\min}) - \log(\lambda - c_{\max}) - \varepsilon \\ &= \log\left(\frac{\lambda - c_{\min}}{\lambda - c_{\max}}\right) - \varepsilon = \log\left(1 + \frac{c_{\max} - c_{\min}}{\lambda - c_{\max}}\right) - \varepsilon \le \frac{c_{\max} - c_{\min}}{\lambda - c_{\max}} - \varepsilon \end{aligned}$$

Thus

$$h_2(\lambda) \le 0 \qquad \text{whenever} \qquad \lambda \ge c_{\max} + \frac{c_{\max} - c_{\min}}{\varepsilon}. \tag{42}$$

We consider now the limit of $h_2(\lambda)$ as $\lambda \downarrow c_{\max}$. Denoting $c_{\max 2}$ the second largest distinct component value of $\boldsymbol{c}$, we have

$$\begin{aligned} h_2(\lambda) &\ge \sum_{i \in I} q_i \log(\lambda - c_i) + \sum_{i \notin I} q_i \log(\lambda - c_i) + \log\left(\sum_{i \in I} \frac{q_i}{\lambda - c_i}\right) - \varepsilon \\ &\ge \sum_{i \in I} q_i \log(\lambda - c_i) + \sum_{i \notin I} q_i \log(c_{\max} - c_{\max 2}) + \log\left(\sum_{i \in I} \frac{q_i}{\lambda - c_i}\right) - \varepsilon \\ &= \left(\sum_{i \in I} q_i - 1\right) \log(\lambda - c_{\max}) + \sum_{i \notin I} q_i \log(c_{\max} - c_{\max 2}) + \log\left(\sum_{i \in I} q_i\right) - \varepsilon \end{aligned}$$

Due to the assumptions of Theorem 3.6, we obtain $h_2(\lambda) \to \infty$ as $\lambda \downarrow c_{\max}$. This combined with (42) and the continuity of h_2 implies that the equation $h_2(\lambda) = 0$ has a solution on $(c_{\max}, c_{\max} + \frac{c_{\max} - c_{\min}}{\varepsilon}]$. □

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 (32a) reads

$$-c_i + \lambda - \mu \sqrt{\frac{q_i}{p_i}} = 0. \tag{43}$$

from which we deduce

$$p_i = q_i \frac{\mu^2}{(\lambda - c_i)^2}. \tag{44}$$

Plugging (44) into the feasibility conditions (34) yields

$$\sum_{i=1}^{n} q_i \frac{\mu^2}{(\lambda - c_i)^2} = 1, \tag{45}$$

$$2\sum_{i=1}^{n} q_i \frac{\mu}{|\lambda - c_i|} = 2 - \varepsilon. \tag{46}$$

We can express μ from (45) via

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

which together with (46) gives the final equation (16). The optimal probabilities (17) then follow from (44). The constraint $\mu > 0$ transfers to $\lambda > c_{\max}$ due to (43), which also allows us to remove the absolute value from (46).

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

$$\begin{aligned} h_3(\lambda) &\geq 2\sum_{i=1}^{n} \frac{q_i}{\lambda - c_{\min}} - (2-\varepsilon)\sqrt{\sum_{i=1}^{n} \frac{q_i}{(\lambda - c_{\max})^2}} = \frac{2}{\lambda - c_{\min}} - \frac{2-\varepsilon}{\lambda - c_{\max}} \\ &= \frac{2(\lambda - c_{\max}) - (2-\varepsilon)(\lambda - c_{\min})}{(\lambda - c_{\min})(\lambda - c_{\max})} = \frac{\varepsilon\lambda - 2(c_{\max} - c_{\min}) - \varepsilon c_{\min}}{(\lambda - c_{\min})(\lambda - c_{\max})} \end{aligned}$$

Thus

$$h_3(\lambda) \geq 0 \qquad \text{whenever} \qquad \lambda \geq c_{\min} + \frac{2(c_{\max} - c_{\min})}{\varepsilon} = c_{\max} + \frac{(2-\varepsilon)(c_{\max} - c_{\min})}{\varepsilon}. \tag{47}$$

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

$$\begin{aligned} h_3(\lambda) &\leq 2\sum_{i\in I} \frac{q_i}{\lambda - c_i} + 2\sum_{i\notin I} \frac{q_i}{\lambda - c_i} - (2-\varepsilon)\sqrt{\sum_{i\in I} \frac{q_i}{(\lambda - c_i)^2}} \\ &\leq 2\frac{\sum_{i\in I} q_i}{\lambda - c_{\max}} + \frac{2}{c_{\max} - c_{\max 2}} - (2-\varepsilon)\frac{\sqrt{\sum_{i\in I} q_i}}{\lambda - c_{\max}} \\ &\leq \frac{2\sum_{i\in I} q_i - (2-\varepsilon)\sqrt{\sum_{i\in I} q_i}}{\lambda - c_{\max}} + \frac{2}{c_{\max} - c_{\max 2}} \end{aligned}$$

Due to the assumptions of Theorem 3.4, we obtain $h_3(\lambda) \to -\infty$ as $\lambda \downarrow c_{\max}$. This combined with (47) and the continuity of h_3 implies that the equation $h_3(\lambda) = 0$ has a solution on $(c_{\max}, c_{\max} + \frac{(2-\varepsilon)(c_{\max} - c_{\min})}{\varepsilon}]$. 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 ϕ_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 (32a) reads

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

from which we deduce

$$p_i = q_i \sqrt{\frac{\mu}{\lambda - c_i}}. \tag{48}$$

Plugging (48) into the feasibility conditions (34) yields

$$\sum_{i=1}^{n} q_i \sqrt{\frac{\mu}{\lambda - c_i}} = 1, \tag{49}$$

$$\sum_{i=1}^{n} q_i \sqrt{\frac{\lambda - c_i}{\mu}} = 1 + \varepsilon. \tag{50}$$

We can express μ from (50) via

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

which together with (49) gives the final equation (18). The optimal probabilities (19) then follow from (48). The constraint $\mu > 0$ transfers to $\lambda > c_{\max}$ due to (48). □

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 (32a) reads (for the uniformity of results, we flipped the sign of λ)

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

from which we due to the complementarity conditions (32b) deduce

$$p_i = \frac{1}{2\mu} q_i \max(c_i + \lambda, 0). \tag{51}$$

Plugging (51) into the feasibility conditions (34) yields

$$\frac{1}{2\mu} \sum_{i=1}^{n} q_i \max(c_i + \lambda, 0) = 1, \tag{52}$$

$$\frac{1}{4\mu^2} \sum_{i=1}^{n} q_i \max^2(c_i + \lambda, 0) = 1 + \varepsilon. \tag{53}$$

We can express μ from (52) via

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

which together with (53) gives the final equation (20). The optimal probabilities (21) then follow from (51). The constraint $\mu > 0$ transfers to $\lambda > -c_{\max}$ due to (51). □

B.2 Optimality conditions for Theorem 3.7

To prove Theorem 3.7, we first realize that the first part of Section B.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 (33) takes the form

$$\begin{aligned} \underset{\boldsymbol{p}}{\text{minimize}}\ & \frac{1}{2}\sum_{i\in I}(p_i - q_i)^2 \\ \text{subject to}\ & \sum_{i\in I} p_i = 1, \\ & 0 \le p_i, \quad \forall i \in I. \end{aligned} \tag{54}$$

Since $q_i > 0$ and $\sum_{i\in I} q_i < 1$ due to Assumption 2.1, it is not difficult to verify that $\hat{\boldsymbol{p}}$ defined in (22) is the optimal solution of (54). 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 (32a) reads

$$-c_i - \alpha_i + \lambda + \mu(p_i - q_i) = 0,$$

from which we due to the primal feasibility condition $p_i \ge 0$, the dual feasibility conditions $\alpha_i \ge 0$ and the complementarity condition (32b) deduce

$$p_i = \max\left(q_i - \frac{1}{\mu}(\lambda - c_i), 0\right). \tag{55}$$

Plugging (55) into the feasibility conditions (34) yields

$$\sum_{i=1}^{n}\max\left(q_i - \frac{1}{\mu}(\lambda - c_i), 0\right) = 1, \tag{56}$$

$$\sum_{i=1}^{n}\left(\max\left(q_i - \frac{1}{\mu}(\lambda - c_i), 0\right) - q_i\right)^2 = \varepsilon^2. \tag{57}$$

The final equations (24) are obtained from (56) and (57) by simple calculus, the formula $\max(-x,-y) = -\min(x,y)$ and Assumption 2.1.

B.3 Convexity for Proposition 4.1

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

Lemma B.1. *For each $\mu > 0$ there is a unique $\lambda(\mu)$ which solves $g_6(\lambda;\mu)$. 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 λ 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, λ is a continuous function of μ. This further implies that h_6 is a continuous function of μ. □

Concerning the proof of Proposition 4.1, we divide it into three parts.

Proof for h_4: After rearranging of terms, we get

$$h_4(\lambda) = \sum_{i=1}^{n}\sum_{j=1}^{n} q_i q_j g_{ij}(\lambda) - 1 - \varepsilon, \tag{58}$$

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

$$\begin{aligned} g'_{ij}(\lambda) &= \frac{1}{2}(\lambda - c_i)^{-\frac{1}{2}}(\lambda - c_j)^{-\frac{1}{2}} - \frac{1}{2}(\lambda - c_i)^{\frac{1}{2}}(\lambda - c_j)^{-\frac{3}{2}} \\ &= \frac{1}{2}(\lambda - c_i)^{-\frac{1}{2}}(\lambda - c_j)^{-\frac{3}{2}}(\lambda - c_j - (\lambda - c_i)) \\ &= \frac{1}{2}(\lambda - c_i)^{-\frac{1}{2}}(\lambda - c_j)^{-\frac{3}{2}}(c_i - c_j) \end{aligned}$$

and the second derivative

$$\begin{aligned} g''_{ij}(\lambda) &= \frac{1}{2}(c_i - c_j)\left(-\frac{1}{2}(\lambda - c_i)^{-\frac{3}{2}}(\lambda - c_j)^{-\frac{3}{2}} - \frac{3}{2}(\lambda - c_i)^{-\frac{1}{2}}(\lambda - c_j)^{-\frac{5}{2}}\right) \\ &= -\frac{1}{4}(c_i - c_j)\left((\lambda - c_i)^{-\frac{3}{2}}(\lambda - c_j)^{-\frac{3}{2}} + 3(\lambda - c_i)^{-\frac{1}{2}}(\lambda - c_j)^{-\frac{5}{2}}\right) \\ &= -\frac{1}{4}(c_i - c_j)(\lambda - c_i)^{-\frac{3}{2}}(\lambda - c_j)^{-\frac{5}{2}}(\lambda - c_j + 3(\lambda - c_i)) \\ &= -\frac{1}{4}(c_i - c_j)(\lambda - c_i)^{-\frac{3}{2}}(\lambda - c_j)^{-\frac{5}{2}}(4\lambda - 3c_i - c_j). \end{aligned}$$

we realize that

$$\begin{aligned} g''_{ij}(\lambda) + g''_{ji}(\lambda) &= -\frac{1}{4}(c_i - c_j)(\lambda - c_i)^{-\frac{5}{2}}(\lambda - c_j)^{-\frac{5}{2}}\big(4\lambda - 3c_i - c_j)(\lambda - c_i) - (4\lambda - c_i - 3c_j)(\lambda - c_j)\big) \\ &= -\frac{1}{4}(c_i - c_j)(\lambda - c_i)^{-\frac{5}{2}}(\lambda - c_j)^{-\frac{5}{2}}\big(3(c_j - c_i)(2\lambda - c_i - c_j)\big) \\ &= \frac{3}{4}(c_i - c_j)^2(\lambda - c_i)^{-\frac{5}{2}}(\lambda - c_j)^{-\frac{5}{2}}(2\lambda - c_i - c_j) \geq 0. \end{aligned}$$

This together with (58) implies that h_4 is convex.

Proof for h_5: Recall that

$$h_5(\lambda) = \sum_{i=1}^{n} q_i \max^2(c_i + \lambda, 0) - (1+\varepsilon)\left(\sum_{i=1}^{n} q_i \max(c_i + \lambda, 0)\right)^2 \tag{59}$$

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

$$\begin{aligned} h'_5(\lambda) &= 2\sum_{i=1}^{j} q_i (c_i + \lambda) - 2(1+\varepsilon)\left(\sum_{i=1}^{j} q_i (c_i + \lambda)\right)\sum_{i=1}^{j} q_i \\ &= 2\sum_{i=1}^{j} q_i (c_i + \lambda)\left(1 - (1+\varepsilon)\sum_{i=1}^{j} q_i\right) \\ &= 2\left(\sum_{i=1}^{n} q_i \max(c_i + \lambda, 0)\right)\left(1 - (1+\varepsilon)\sum_{i=1}^{j} q_i\right). \end{aligned} \tag{60}$$

Due to the assumption of $-c_1 < \cdots < -c_n$, the violation of (11) implies $(1+\varepsilon)q_1 < 1$. This due to (60) means that $h_4'(\lambda) > 0$ on $\lambda \in (-c_1, -c_2)$. Since $h_5(-c_1) = 0$ due to (59), 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 λ_0 such that h_5 is non-decreasing on $(-c_{\max}, \lambda_0)$ and decreasing on (λ_0, ∞). By the same arguments we have that h_5' is decreasing on (λ_0, ∞), 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 cardinaly of I. Then we can get the same estimates of (60) as in the previous paragraphs.

Proof for h_6: Define

$$I(\mu) = \{i \mid \lambda(\mu) - c_i < \mu q_i\} \tag{61}$$

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 (61) implies $I(\mu_1) \subset I(\mu_2)$. Thus, $I(\mu)$ is a non-decreasing function (with respect to set inclusion) of μ.

From (24b) we have

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \min(\lambda(\mu) - c_i, \mu q_i) = \sum_{i \in I(\mu)} (\lambda(\mu) - c_i) + \sum_{i \notin I(\mu)} \mu q_i \\ &= |I(\mu)|\lambda(\mu) - \sum_{i \in I(\mu)} c_i + \mu \sum_{i \notin I(\mu)} q_i, \end{aligned}$$

from which we deduce

$$\lambda(\mu) = \frac{1}{|I(\mu)|}\left(\sum_{i \in I(\mu)} c_i - \mu \sum_{i \notin I(\mu)} q_i\right). \tag{62}$$

Since $I(\mu)$ is a non-decreasing function of μ, this implies that $\lambda(\mu)$ is a piecewise linear function with a finite number of pieces. On each of these pieces, we have

$$\lambda'(\mu) = -\frac{1}{|I(\mu)|} \sum_{i \notin I(\mu)} q_i \tag{63}$$

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

$$\begin{aligned} \frac{1}{2} h_6'(\mu) &= \lambda'(\mu) \sum_{i \in I(\mu)} (\lambda(\mu) - c_i) + \mu \sum_{i \notin I(\mu)} q_i^2 - \mu\varepsilon^2 \\ &= -\lambda'(\mu) \sum_{i \in I(\mu)} c_i + \lambda'(\mu)|I(\mu)|\lambda(\mu) + \mu \sum_{i \notin I(\mu)} q_i^2 - \mu\varepsilon^2 \\ &= -\lambda'(\mu) \sum_{i \in I(\mu)} c_i + \lambda'(\mu)\left(\sum_{i \in I(\mu)} c_i - \mu \sum_{i \notin I(\mu)} q_i\right) + \mu \sum_{i \notin I(\mu)} q_i^2 - \mu\varepsilon^2 \\ &= -\mu\lambda'(\mu) \sum_{i \notin I(\mu)} q_i + \mu \sum_{i \notin I(\mu)} q_i^2 - \mu\varepsilon^2 \\ &= \mu\left(\frac{1}{|I(\mu)|} \sum_{i \notin I(\mu)} q_i \sum_{i \notin I(\mu)} q_i + \sum_{i \notin I(\mu)} q_i^2 - \varepsilon^2\right), \end{aligned} \tag{64}$$

where in the third equality we used (62) and in the last one (63).

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

$$\begin{aligned} g_6(\hat{\lambda}, \mu) &= \sum_{i=1}^{n} \min\Big(\hat{\lambda} - c_i, \mu q_i\Big) = \sum_{i \in I} \min\Big(\hat{\lambda} - c_i, \mu q_i\Big) + \sum_{i \notin I} \min\Big(\hat{\lambda} - c_i, \mu q_i\Big) \\ &= \frac{1}{2}|I|(c_{\max 2} - c_{\max}) + \sum_{i \notin I} \min\Big(\hat{\lambda} - c_i, \mu q_i\Big) \leq \frac{1}{2}|I|(c_{\max 2} - c_{\max}) + \mu \sum_{i \notin I} q_i < 0 \end{aligned}$$

whenever μ is sufficiently small. Since $g_6(\lambda(\mu); \mu) = 0$ due to definition and since g_6 is non-decreasing in λ, this means that $\lambda(\mu) \geq \hat{\lambda} = \frac{1}{2}(c_{\max} + c_{\max 2})$. This due to (61) implies $I(\mu) = I$ whenever μ 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 (64) 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 μ_0 such that h_6 is non-decreasing on $(0, \mu_0)$ and decreasing on (μ_0, ∞). By the same arguments, we have that h_6' is decreasing on (μ_0, ∞), which implies that h_6 is concave on this interval. This finishes the proof of the second part.

C Convergence of Newton's method

We show the following theorem only for differentiable functions. However, it holds for any concave function by replacing the derivative by its concave superdifferential.

Lemma C.1. *Consider a continuous concave function $h : [a, b] \to \mathbb{R}$ with $h(a) > 0$ and $h(b) < 0$. Then the Newton's method*

$$\lambda^{k+1} = \lambda^k - \frac{h(\lambda^k)}{h'(\lambda^k)}$$

started from any point λ^0 with $h(\lambda^0) < 0$ gives a decreasing sequence which converges to some $\bar{\lambda} \in (a, b)$ with $h(\bar{\lambda}) = 0$.

Proof. Due to concavity of h and $h(a) > 0$ with $h(b) < 0$, there exists unique $\lambda^* \in (a, b)$ with $h(\lambda^*) = 0$. Moreover, due to concavity again we have $h'(\lambda^*) < 0$. Since h is concave and since $h(a) > 0$ and $h(\lambda^0) < 0$, we obtain $\lambda^0 > \lambda^*$. Then the Newton's method forms a decreasing sequence $\{\lambda^k\}$ bounded below by λ^*, which is therefore convergent. At the same time, $h'(\lambda^k)$ is uniformly bounded above by zero as $h'(\lambda^k) < h'(\lambda^*) < 0$. This implies

$$\frac{h(\lambda^k)}{h'(\lambda^k)} = \lambda^k - \lambda^{k+1} \to 0,$$

which due to the uniform boundedness of $h'(\lambda^k)$ from zero implies $h(\lambda^k) \to 0$. But this due to the continuity of h means that $\lambda^k \to \bar{\lambda}$. □

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 and Variational Analysis*, 24(2):207–229, 2016.

[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 Journal of Numerical Analysis*, 2018.

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

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

[5] M. Blondel, A. Fujino, and N. Ueda. Large-scale multiclass support vector machine training via euclidean projection onto the simplex. In *2014 22nd International Conference on Pattern Recognition*, pages 1289–1294. IEEE, 2014.

[6] L. Condat. Fast projection onto the simplex and the l_1 ball. *Mathematical Programming*, 158(1-2):575–585, 2016.

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

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

[9] R. Gao and A. J. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[10] M. Held, P. Wolfe, and H. P. Crowder. Validation of subgradient optimization. *Mathematical programming*, 6(1):62–88, 1974.

[11] M. G. Kapteyn, K. E. Willcox, and A. Philpott. A distributionally robust approach to black-box optimization. In *2018 AIAA Non-Deterministic Approaches Conference*, page 0666, 2018.

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

[13] K. C. Kiwiel. Breakpoint searching algorithms for the continuous quadratic knapsack problem. *Mathematical Programming*, 112(2):473–491, 2008.

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

[15] N. Li, R. Jin, and Z.-H. Zhou. Top rank optimization in linear time. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pages 1502–1510, Cambridge, MA, USA, 2014. MIT Press.

[16] J. Liu and J. Ye. Efficient Euclidean projections in linear time. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 657–664. ACM, 2009.

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

[18] H. Markowitz. Portfolio selection. *The journal of finance*, 7(1):77–91, 1952.

[19] C. Michelot. A finite algorithm for finding the projection of a point onto the canonical simplex of $\mathbb{R}^n$. *Journal of Optimization Theory and Applications*, 50(1):195–200, 1986.

[20] J. A. Nelder and R. Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[21] A. Philpott, V. de Matos, and L. Kapelevich. Distributionally robust sddp. *Computational Management Science*, 15(3-4):431–454, 2018.

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

[23] S. Shalev-Shwartz and Y. Singer. Efficient learning of label ranking by soft projections onto polyhedra. *Journal of Machine Learning Research*, 7(Jul):1567–1599, 2006.

[24] E. Van Den Berg and M. P. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912, 2008.