Noname manuscript No.
(will be inserted by the editor)

Optimal L_2–norm Empirical Importance Weights for the Change of Probability Measure

Sergio Amaral · Douglas Allaire · Karen Willcox

Received: date / Accepted: date

Abstract This work proposes an optimization formulation to determine a set of empirical importance weights to achieve a change of probability measure. The objective is to estimate statistics from a target distribution using random samples generated from a (different) proposal distribution. This work considers the specific case in which the proposal distribution from which the random samples are generated is unknown; that is, we have available the samples but no explicit description of their underlying distribution. In this setting, the Radon-Nikodym Theorem provides a valid but indeterminable solution to the task, since the distribution from which the random samples are generated is inaccessible. The proposed approach employs the well-defined and determinable empirical distribution function associated with the available samples. The core idea is to compute importance weights associated with the random samples, such that the distance between the weighted proposal empirical distribution function and the desired target distribution function is minimized. The distance metric selected for this work is the L_2–norm and the importance weights are constrained to define a probability measure. The resulting optimization problem is shown to be a single linear equality and box-constrained quadratic program. This problem can be solved efficiently using optimization algorithms that scale well to high dimensions. Under some conditions restricting the class of distribution functions, the solution of the optimization problem is shown to result in a weighted proposal empirical distribution function that converges to the target distribution function in the L_1–norm, as the number of samples tends to infinity. Results on a variety of test cases show that the proposed approach performs well in comparison with other well-known approaches.

Keywords change of measure · empirical measure · Radon-Nikodym Theorem · importance weight · density ratio · quadratic program

1 Introduction

Consider the task of estimating statistics from a distribution of interest, denoted as the *target distribution* (Robert and Casella 2005). In many cases, one may apply standard Monte Carlo simulation to estimate these statistics, using random samples that are generated from the target distribution. However, in some circumstances one may only have available random samples generated from a different distribution, denoted as the *proposal distribution*. The challenge of evaluating statistics from a target distribution given random samples generated from a proposal distribution is acknowledged as the change of measure and arises in a host of domains such as importance sampling, information divergence, and particle filtering (see e.g., Sugiyama et al. (2012) for a fuller discussion of applications). If both proposal and target distributions are known and satisfy additional conditions, then the Radon-Nikodym Theorem provides a solution (Billingsley 2008).

This work considers the case in which the proposal distribution from which the random samples are generated is unknown; that is, we have available the samples

S. Amaral (samaral@mit.edu) · K. Willcox
Department of Aeronautics & Astronautics
Massachusetts Institute of Technology
Cambridge, Massachusetts

D. Allaire
Department of Mechanical Engineering
Texas A&M University
College Station, Texas

but no explicit description of their underlying distribution. Although the Radon-Nikodym Theorem is still valid (if the underlying distribution satisfies the appropriate conditions), it is indeterminable because we cannot compute the Radon-Nikodym derivative (i.e., the ratio of the target probability density function to the proposal probability density function), herein referred to as the probability density ratio. To accomplish the objective, a change of measure, an importance weight proportional to the probability density ratio would be associated with each random sample. However, the importance weights cannot be computed directly in the usual way, since the probability density ratio is indeterminable. This paper presents an approach that overcomes this challenge by formulating and solving a scalable optimization problem to determine a set of empirical importance weights. We first discuss several previously proposed solutions to this problem.

The previous approaches summarized here all assume that the random samples are generated from an unknown distribution (see e.g., Sugiyama et al. (2012) for a detailed discussion of these approaches). As a result, these approaches seek to estimate the probability density ratio using the random samples. A commonly used approach estimates the unknown proposal probability density function from the random samples (Scott 1992). By estimating the probability density function one can then estimate the probability density ratio. The solution to the change of measure problem then follows from the Radon-Nikodym Theorem along with the estimated probability density ratio. However, estimating the unknown probability density function from the random samples is difficult and is particularly challenging in cases of high dimension (Hastie et al. 2009, Scott 1992, Vapnik 1998). In practice, this challenge can be overcome if the random samples are known to be generated from a parametric distribution family, in which case a parametric density estimation method can be employed.

As a result, other approaches have avoided estimating the probability density function and instead estimate directly the probability density ratio using the random samples. The kernel mean matching approach matches the moments using a universal reproducing kernel Hilbert function (Gretton et al. 2009, Huang et al. 2007). The probabilistic classification approach computes the probability density ratio by applying Bayes' Theorem (Qin 1998). The importance estimation filtering approach minimizes the Kullback-Leibler divergence metric between the estimated and actual probability density ratios (Sugiyama et al. 2012). The unconstrained least squares importance filtering approach minimizes the L_2–norm between the estimated and actual probability density ratios (Kanamori et al. 2009). The direct density ratio estimation with dimension reduction solves the previous approach on a lower-dimensional space (Sugiyama et al. 2011). These approaches share in common multiple attributes. They each present a means of computing the probability density ratio using the random samples. They each represent the probability density ratio using a set of basis functions, thereby constraining the solution to exist within a specified basis representation. Finally, these approaches require tuning parameters, which one can choose using a variant of cross-validation.

Our approach avoids estimating or working with the unknown distribution function or its probability density function. Instead, we work with the well-defined and determinable *empirical distribution function* associated with the random samples. Specifically, our approach, illustrated in Figure 1, formulates and solves an optimization problem to determine a set of empirical importance weights that minimize the L_2–norm between the weighted proposal empirical distribution function and the target distribution function. In the example in Figure 1, the target is the uniform distribution function, $\mathcal{U}(0,1)$, and the proposal random samples are generated from the beta distribution function, $\mathcal{B}(0.5, 0.5)$. The core idea of our approach is to compute importance weights associated with the proposal random samples that transform the weighted proposal empirical distribution function to the target distribution function. We also constrain the importance weights to define a probability measure. This requires that these importance weights are non-negative and that the empirical probability measure assigns a unit value to the entire probability space. Our work is differentiated from current practices in that we do not estimate the Radon-Nikodym derivative from the proposal random samples, but rather we find the optimal importance weights for the given set of proposal random samples, where optimality is defined by the closeness of the weighted proposal empirical distribution function to the target distribution function.

The approach proposed in this paper shares resemblance to the recent constructive setting of the density ratio estimate (Vapnik et al. 2014). That work minimizes the regularized L_2–norm between the weighted proposal empirical distribution function and the empirical target distribution function, where the importance weights are defined on a set of basis functions. Those importance weights are shown in Vapnik et al. (2014) to converge in probability to the Radon-Nikodym derivative, as the number of proposal and target random samples tend to infinity. Our approach does not use a basis function representation of the importance weights, since we are only interested in evaluating the importance weights at the random sample locations (i.e., we associate one weight with each random sample). We also do not include regularization, since this modifies the solution and introduces smoothness that may not be desirable. Instead, we rely on the optimization solvers to exploit the structure of the problem. Avoiding regularization allows us to avoid tuning parameters, yet our formulation maintains that the weighted proposal empirical distribution function converges to the target distribution function in the L_1–norm, as the number of random samples tends to infinity. Moreover, our optimization approach can be implemented at large scale (both high-dimensional distribution functions and a large number of random samples). Our approach has an analytic closed-form solution in the case of a unidimensional distribution problem. We show that this closed-form solution for our empirical importance weights results in almost everywhere convergence of the weighted proposal empirical distribution function to the target distribution function, as the number of random samples tends to infinity. Additionally, we demonstrate a relationship between our approach and the trapezoidal integration rule as well as to discrepancy theory.

The organization of this paper is as follows. Section 2 sets nomenclature, formalizes the objective of this work, and presents the proposed optimization formulation. In Section 3, we present the numerical formulation and examine properties of the optimization statement. In Section 4, we prove that the proposed approach achieves convergence in the L_1–norm for multidimensional distributions and weak convergence for unidimensional distributions. In Section 5, we examine the analytic solution to the optimization statement for the case of a unidimensional distribution problem. Section 5 also presents a numerical solution to the optimization statement and discusses techniques that extend our approach to large-scale applications. In Section 6, we demonstrate the properties of the proposed approach on a unidimensional distribution problem. Section 6 also compares our approach to previous approaches on an importance sampling problem over a range of parameters, evaluates the performance of the optimization algorithms, and examines the relationship between discrepancy theory and the proposed approach when the proposal and target are distributed according to the uniform distribution. Finally, Section 7 concludes the paper.

2 Problem Statement

We begin by setting notation for the subsequent developments and establishing the objective of this work.

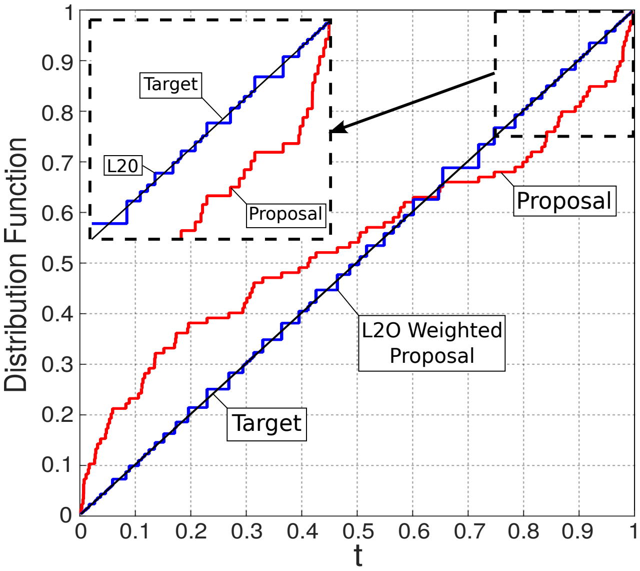


Fig. 1: The proposed approach minimizes, with respect to empirical importance weights associated with the proposal random samples, the L_2–norm between the weighted proposal empirical distribution function and the target distribution function. In this example, we generated $n = 100$ random samples from the proposal beta distribution function, $\mathcal{B}(0.5, 0.5)$. The results show our weighted proposal empirical distribution function, labeled "L2O Weighted Proposal", accurately represents the target uniform distribution function, $\mathcal{U}(0, 1)$.

The section concludes with a description and formulation of our solution to the objective.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, where Ω is a sample space, $\mathcal{F}$ is a σ-field, and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$. Then the random variable $Y : \Omega \to \mathbb{R}^d$ is associated with the *target* measure ν on $\mathbb{R}^d$, such that $\nu(A) = \mathbb{P}(Y^{-1}(A))$ for $A \in \mathbb{R}^d$. Likewise, the random variable $X : \Omega \to \mathbb{R}^d$ is associated with the finite support *proposal* measure μ on $\mathbb{R}^d$, such that $\mu(A) = \mathbb{P}(X^{-1}(A))$ for $A \in \mathbb{R}^d$. In addition, we confine the target measure ν to be absolutely continuous with respect to proposal measure μ. Let $\mathbf{t} \in \mathbb{R}^d$ be a generic point and designate entries of $\mathbf{t}$ by subscript notation as follows $\mathbf{t} = [t_1, t_2, \ldots, t_d]^\top$. Define $F_\nu(\mathbf{t})$ and $f_\nu(\mathbf{t})$ to be the target distribution function and target probability density function of Y evaluated at $\mathbf{t}$, respectively. Similarly, define $F_\mu(\mathbf{t})$ and $f_\mu(\mathbf{t})$ to be the proposal distribution function and proposal probability density function of X evaluated at $\mathbf{t}$, respectively.

In our problem setting, the proposal measure μ is accessible to us only through sampling; that is, we are provided with random samples of the random variable X but we cannot evaluate F_μ or f_μ explicitly. Let $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$ be random samples of X, where n is the number of random samples. The objective of our work is to estimate statistics from the target measure ν given random samples $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$ generated from the proposal measure μ. The challenge with this objective, recognized as a change of measure, is that the proposal measure μ is accessible to us only through sampling.

The typical approach to overcome this challenge is to apply density estimation to the random samples $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$, yielding an estimate of the proposal density f_μ. However, density estimation in high dimensions is notoriously difficult, and state-of-the-art approaches often perform poorly for high-dimensional problems. Therefore, we approach the change of measure challenge in a different way—using instead the well-defined proposal empirical distribution function,

$$F_\mu^n(\mathbf{t}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\mathbf{x}^i \leq \mathbf{t}), \tag{1}$$

where $\mathbb{I}(\mathbf{x}^i \leq \mathbf{t})$ is the maximum convention Heavyside step function defined as

$$\mathbb{I}(\mathbf{x} \leq \mathbf{t}) = \begin{cases} 1, & \text{if } x_i \leq t_i,\ \forall\, i \in \{1, 2, \ldots, d\} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Here we have used the subscript and superscript notation for the empirical distribution function, F_μ^n, to identify the measure of the random samples from which it is built, μ, and the number of random samples, n. The strong law of large numbers (SLLN) states that the estimator F_μ^n converges to the proposal distribution function F_μ defined as

$$F_\mu(\mathbf{t}) = \mu((-\infty, \mathbf{t}]), \tag{3}$$

as n tends to infinity almost everywhere (a.e.) for all continuity points $\mathbf{t}$ of $F_\mu(\mathbf{t})$ (Billingsley 2008).

To accomplish the change of measure objective, we propose to compute a set of importance weights, defined here as *empirical importance weights*, to transform the proposal empirical distribution function into the target distribution function. We introduce n empirical importance weights, denoted by the vector $\mathbf{w} = [w_1, w_2, \ldots, w_n]^\top$. Each empirical importance weight w_i is associated with a random sample $\mathbf{x}^i$. We use the notation

$$F_{\mu;\mathbf{w}}^n(\mathbf{t}) = \frac{1}{n}\sum_{i=1}^{n} w_i \mathbb{I}(\mathbf{x}^i \leq \mathbf{t}) \tag{4}$$

to represent a weighted empirical distribution function that is composed of n random samples generated from the measure μ and weighted by $\mathbf{w}$. The empirical importance weights are dependent on the random samples, $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$; however, for simplicity we do not show the dependency in the notation.

We now cast the change of measure objective as an optimization statement. The objective is to minimize, with respect to the empirical importance weights, the distance between $F_{\mu;\mathbf{w}}^n$, defined in Equation (4), and the target distribution function, F_ν. The criterion selected is the L_2–norm distance metric. Thus, the L_2–norm objective function is defined as

$$\omega^2(\mathbf{w}) = \frac{1}{2}\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(F_{\mu;\mathbf{w}}^n(\mathbf{t}) - F_\nu(\mathbf{t})\right)^2 \, \mathrm{d}\mathbf{t}, \tag{5}$$

conditioned on the scaled empirical importance weights being a probability measure. That is, $\mathbf{w}$ satisfies the non-negativity box-constraint, $w_i \geq 0,\ \forall\, i \in \{1, 2, \ldots, n\}$, and the single equality constraint, $\mathbf{1}_n^\top \mathbf{w} = n$, where $\mathbf{1}_n \in \mathbb{R}^n$ is a vector with all entries equal to 1. The optimization statement that determines the empirical importance weights associated with the proposal random samples for the change of measure is thus stated as follows:

$$\begin{aligned} \arg\min_{\mathbf{w}} \ & \omega^2(\mathbf{w}) \\ s.t. \ & w_i \geq 0,\ \forall\, i \in \{1, 2, \ldots, n\} \\ & \mathbf{1}_n^\top \mathbf{w} = n. \end{aligned} \tag{6}$$

In the above optimization statement, we have assumed that the target distribution F_ν is known explicitly. However, our approach can be applied to the case where the the target measure is represented only through random samples of the random variable Y. In that case, we replace F_ν in Equation (5) with the target empirical distribution function F_ν^m, where m is the number of random samples of the random variable Y. In the following development, we work mostly with the formulation defined in Equations (5) and (6); when applicable we introduce the target empirical distribution function into the optimization statement.

3 Numerical Formulation

This section describes how the optimization statement (6) can be formulated as a single linear equality and box-constrained quadratic program (Section 3.1). Section 3.2 examines the properties of the optimization statement using the Karush Kuhn Tucker (KKT) conditions.

3.1 Single Linear Equality and Box-Constrained Quadratic Program

Upon substituting Equation (4) into Equation (5) and, without loss of generality, confining the support of μ to the unit hypercube, we obtain

$$\omega^2(\mathbf{w}) = \\ \frac{1}{2}\int_0^1 \cdots \int_0^1 \left(\frac{1}{n}\sum_{i=1}^{n} w_i \mathbb{I}(\mathbf{x}^i \leq \mathbf{t}) - F_\nu(\mathbf{t}) \right)^2 \mathrm{d}\mathbf{t}. \tag{7}$$

This expression can be expanded as follows:

$$\begin{aligned}\omega^2(\mathbf{w}) =& \frac{1}{2}\int_0^1 \cdots \int_0^1 \left[\left(\frac{1}{n}\sum_{i=1}^{n} w_i \mathbb{I}(\mathbf{x}^i \leq \mathbf{t}) \right)^2 - \right. \\ & \left. \frac{2F_\nu(\mathbf{t})}{n}\sum_{i=1}^{n} w_i \mathbb{I}(\mathbf{x}^i \leq \mathbf{t}) + (F_\nu(\mathbf{t}))^2 \right] \mathrm{d}\mathbf{t}.\end{aligned} \tag{8}$$

The third term in the integrand of Equation (8) is independent of the optimization parameter and thus can be discarded from the optimization statement without affecting the optimal solution $\mathbf{w}$. We now examine the first term and second term individually and formulate their respective numerical representations.

The first term of the integrand in Equation (8) can be represented as

$$\begin{aligned}&\int_0^1 \cdots \int_0^1 \left(\frac{1}{n}\sum_{i=1}^{n} w_i \mathbb{I}(\mathbf{x}^i \leq \mathbf{t}) \right)^2 \mathrm{d}\mathbf{t} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \int_0^1 \cdots \int_0^1 \mathbb{I}(\mathbf{x}^i \leq \mathbf{t})\mathbb{I}(\mathbf{x}^j \leq \mathbf{t}) \ \mathrm{d}\mathbf{t} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \prod_{k=1}^{d} \int_0^1 \mathbb{I}(x_k^i \leq t)\mathbb{I}(x_k^j \leq t) \ \mathrm{d}t_k \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \prod_{k=1}^{d} \int_{z_k^{i,j}}^1 \mathrm{d}t_k, \\ &= \mathbf{w}^\top H \mathbf{w},\end{aligned} \tag{9}$$

where $z_k^{i,j} = \max(x_k^i, x_k^j)$ and x_k^i is the k^{th} entry of random sample $\mathbf{x}^i$. Note that $H \in \mathbb{R}^{n \times n}$ is a reproducing kernel and by definition a positive definite matrix (see e.g., Novak and Wozniakowski (2009) for a review of this analysis). Additionally, the H matrix is the Hadamard product of d individual matrices. To obtain the Hadamard construction of H, we define the matrix corresponding to the single dimension k, H^k, where the $(i,j)^{th}$ entry of H^k is

$$H_{i,j}^k = \int_{z_k^{i,j}}^1 \mathrm{d}t_k, \tag{10}$$

and $k \in \{1, 2, \ldots, d\}$. Then the $(i,j)^{th}$ entry of H can be defined as

$$H_{i,j} = \frac{1}{n^2}\prod_{k=1}^{d} H_{i,j}^k, \tag{11}$$

which allows us to construct matrix H as

$$H = \frac{1}{n^2}\big(H^1 \circ H^2 \circ \cdots \circ H^d\big), \tag{12}$$

where "$\circ$" represents the Hadamard product.

The second term of the integrand in Equation (8) can be represented as

$$\begin{aligned}&\int_0^1 \cdots \int_0^1 \frac{2F_\nu(\mathbf{t})}{n}\sum_{i=1}^{n} w_i \mathbb{I}(\mathbf{x}^i \leq \mathbf{t}) \ \mathrm{d}\mathbf{t} \\ &= \frac{1}{n}\sum_{i=1}^{n} w_i \int_0^1 \cdots \int_0^1 F_\nu(\mathbf{t})\mathbb{I}(\mathbf{x}^i \leq \mathbf{t}) \ \mathrm{d}\mathbf{t} \\ &= \frac{1}{n}\sum_{i=1}^{n} w_i \int_{x_1^i}^1 \cdots \int_{x_d^i}^1 F_\nu(\mathbf{t}) \ \mathrm{d}\mathbf{t}, \\ &= \mathbf{w}^\top \mathbf{b},\end{aligned} \tag{13}$$

where the i^{th} entry of $\mathbf{b} \in \mathbb{R}^n$ is

$$b_i = \frac{1}{n}\int_{x_1^i}^1 \cdots \int_{x_d^i}^1 F_\nu(\mathbf{t}) \ \mathrm{d}\mathbf{t}. \tag{14}$$

If the target distribution function, F_ν, is unknown and instead we have m random samples of the random variable Y, $\{\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^m\}$, then the i^{th} entry of $\mathbf{b}$ is

$$b_i = \frac{1}{n}\int_{x_1^i}^1 \cdots \int_{x_d^i}^1 \frac{1}{m}\sum_{j=1}^{m} \mathbb{I}(\mathbf{y}^j \leq \mathbf{t}) \ \mathrm{d}\mathbf{t}. \tag{15}$$

Our modified optimization statement is now

$$\begin{aligned}\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \ & \hat{\omega}^2(\mathbf{w}) \\ s.t. \ & w_i \geq 0, \ \forall \, i \in \{1, \ldots, n\} \\ & \mathbf{1}_n^\top \mathbf{w} = n,\end{aligned} \tag{16}$$

where

$$\hat{\omega}^2(\mathbf{w}) = \frac{1}{2}\big(\mathbf{w}^\top H \mathbf{w} - 2\mathbf{w}^\top \mathbf{b}\big). \tag{17}$$

Solving (16) yields the optimal empirical importance weights $\hat{\mathbf{w}}$ that minimize our original L_2–norm distance metric while satisfying the requirement of $\hat{\mathbf{w}}/n$ forming a probability measure.

3.2 Karush Kuhn Tucker Conditions

The Lagrangian of the optimization statement (16) is

$$\mathcal{L}(\mathbf{w}, \delta, \boldsymbol{\lambda}) = \\ \frac{1}{2}\big(\mathbf{w}^\top H \mathbf{w} - 2\mathbf{w}^\top \mathbf{b}\big) + \delta(\mathbf{1}_n^\top \mathbf{w} - n) - \boldsymbol{\lambda}^\top \mathbf{w}, \tag{18}$$

where $\delta \in \mathbb{R}$ and $\boldsymbol{\lambda} \in \mathbb{R}^n$ are the equality and inequality constraint Lagrange multipliers, respectively. The optimal solution to (16) satisfies the following KKT conditions:

$$\begin{aligned} \frac{\partial \mathcal{L}(\hat{\mathbf{w}}, \delta, \boldsymbol{\lambda})}{\partial \mathbf{w}} &= \mathbf{0}_n = H\hat{\mathbf{w}} - \mathbf{b} + \delta \mathbf{1}_n - \boldsymbol{\lambda} \\ \hat{w}_i &\geq 0, \ \forall\, i \in \{1, 2, \ldots, n\} \\ \lambda_i &\geq 0, \ \forall\, i \in \{1, 2, \ldots, n\} \\ \mathbf{1}_n^\top \mathbf{w} &= n \\ \delta &\text{ is sign unrestricted} \\ \lambda_i \hat{w}_i &= 0 \ \ \forall\, i \in \{1, 2, \ldots, n\} \end{aligned} \tag{19}$$

where $\mathbf{1}_n \in \mathbb{R}^n$ and $\mathbf{0}_n \in \mathbb{R}^n$ are vectors with all entries equal to 1 and 0 respectively.

Since the optimization statement is a strictly convex quadratic program with linear constraints, one may show that the solution $\hat{\mathbf{w}}$ of (19) is the global solution to (16) (Boyd and Vandenberghe 2004). This implies that for all n the following inequality holds,

$$\begin{aligned} &\int_A (F^n_{\mu,\hat{\mathbf{w}}}(\mathbf{t}) - F_\nu(\mathbf{t}))^2 \ \mathrm{d}\mathbf{t} \\ &\qquad \leq \int_A (F^n_{\mu,\bar{\mathbf{w}}}(\mathbf{t}) - F_\nu(\mathbf{t}))^2 \ \mathrm{d}\mathbf{t}, \end{aligned} \tag{20}$$

where $\bar{\mathbf{w}} = [\bar{w}_1, \bar{w}_2, \ldots, \bar{w}_n]^\top$ is any set of importance weights that satisfies the constraints of the optimization statement (16).

The active set method is one numerical method that solves (16), and has been shown to converge and terminate in a finite number of steps (Lawson and Hanson 1974). This method employs an iterative approach that splits the solution space into an active set, $\mathcal{A} = \{i : w_i = 0\}$, and a passive set, $\mathcal{P} = \{i : w_i > 0\}$. The active and passive sets are updated iteratively until the KKT conditions are satisfied. At each iteration, the method solves an optimization problem for the passive set importance weights that has a closed-form solution. We use this closed-form solution to derive an analytic solution for the special case $d = 1$ (Section 5.1); however, our general numerical results employ optimization methods that are more amenable to large-scale problems, as described in Section 5.2. Before discussing the optimization solution strategies in detail, we first analyze the convergence properties of our approach.

4 Convergence

The following section demonstrates that our approach, based on (16), converges to the target distribution function in the L_1–norm, as the number of random samples tends to infinity. To demonstrate convergence in the L_1–norm we require the Radon-Nikodym derivative, which we recall in this section. The section concludes with the convergence theorem and proof.

The Radon-Nikodym Theorem states that

$$\nu(A) = \int_A h \ \mathrm{d}\mu \tag{21}$$

for any measurable subset $A \in \mathcal{F}$, where the measurable function $h : \mathbb{R}^d \to \mathbb{R}$ is called the Radon-Nikodym derivative and is defined by the probability density ratio, $h = f_\nu / f_\mu$ (Billingsley 2008). In our problem setting, the Radon-Nikodym derivative exists but is unknown. Let $\{h(\mathbf{x}^1), h(\mathbf{x}^2), \ldots \ , h(\mathbf{x}^n)\}$ be the Radon-Nikodym derivatives corresponding to proposal random samples $\{\mathbf{x}^1, \mathbf{x}^2, \ldots \ , \mathbf{x}^n\}$. To construct a probability measure, define the Radon-Nikodym importance weights as $\hat{h}(\mathbf{x}^i) = h(\mathbf{x}^i)/\bar{h}$ where $\bar{h} = \frac{1}{n}\sum_{i=1}^n h(\mathbf{x}^i)$. If weighted by $\hat{\mathbf{h}} = [\hat{h}(\mathbf{x}^1), \hat{h}(\mathbf{x}^2), \ldots, \hat{h}(\mathbf{x}^n)]^\top$, the Radon-Nikodym importance weighted empirical distribution function,

$$F^n_{\mu;\hat{\mathbf{h}}}(\mathbf{t}) = \frac{1}{n}\sum_{i=1}^n \hat{h}(\mathbf{x}^i)\mathbb{I}(\mathbf{x}^i \leq \mathbf{t}), \tag{22}$$

converges almost everywhere to the distribution function F_ν by the SLLN as n tends to infinity for all continuity points $\mathbf{t}$ of $F_\nu(\mathbf{t})$ (Tokdar et al. 2011).

We now present the convergence proof using our empirical importance weight vector $\hat{\mathbf{w}}$. We emphasize that Theorem 1 given below does not imply that the empirical importance weights converge pointwise to the Radon-Nikodym importance weights as the number of random samples tends to infinity. The proof establishes that the sequence of functions $\{F^1_{\mu;\hat{\mathbf{w}}},\ F^2_{\mu;\hat{\mathbf{w}}},\ \ldots\}$, defined by Equation (4), converges to the target distribution function in the L_1–norm, as the number of random samples tends to infinity.

Theorem 1 *Let F_ν be the distribution function of ν and $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$ be random samples generated from the finite support probability measure μ where ν is absolutely continuous with respect to μ. Then there exists a set of empirical importance weights $\hat{\mathbf{w}} = [\hat{w}_1, \hat{w}_2, \cdots, \hat{w}_n]^\top$ satisfying (16) such that*

$$\lim_{n\to\infty} \int_A \left| F^n_{\mu,\hat{\mathbf{w}}}(\mathbf{t}) - F_\nu(\mathbf{t}) \right| d\mathbf{t} = 0, \tag{23}$$

where $A = \{t \in \mathbb{R} \mid f_\mu(t) > 0\}$ is a bounded set.

Proof We begin with the Radon-Nikodym importance weights $\hat{\mathbf{h}}$, which satisfy the constraints in the optimization statement (16). As stated previously, by the SLLN we have

$$\lim_{n\to\infty} F^n_{\mu,\hat{\mathbf{h}}}(\mathbf{t}) \overset{a.e.}{=} F_\nu(\mathbf{t}), \tag{24}$$

for every continuity point $\mathbf{t}$ of $F_\nu(\mathbf{t})$. Since there exists an integrable function dominating $F^n_{\mu,\hat{\mathbf{h}}}(\mathbf{t}) \leq 1$ for all $\mathbf{t} \in A$ and n, we can apply the dominated convergence theorem to obtain convergence in the L_1–norm:

$$\lim_{n\to\infty} \int_A \left| F^n_{\mu,\hat{\mathbf{h}}}(\mathbf{t}) - F_\nu(\mathbf{t}) \right| \, \mathrm{d}\mathbf{t} = 0. \tag{25}$$

Using the inequality $\left| F^n_{\mu,\hat{\mathbf{h}}}(\mathbf{t}) - F_\nu(\mathbf{t}) \right| \leq 1$, for all $\mathbf{t} \in A$ and all n, we obtain a bound on the L_2–norm,

$$\begin{aligned} \int_A (F^n_{\mu,\hat{\mathbf{h}}}(\mathbf{t}) - F_\nu(\mathbf{t}))^2 \, \mathrm{d}\mathbf{t} \leq \\ \int_A 1 \cdot \left| F^n_{\mu,\hat{\mathbf{h}}}(\mathbf{t}) - F_\nu(\mathbf{t}) \right| \, \mathrm{d}\mathbf{t}. \end{aligned} \tag{26}$$

Combining Equation (26) with Equation (25) we show convergence in the L_2–norm:

$$\lim_{n\to\infty} \int_A (F^n_{\mu,\hat{\mathbf{h}}}(\mathbf{t}) - F_\nu(\mathbf{t}))^2 \, \mathrm{d}\mathbf{t} = 0. \tag{27}$$

Since $\hat{\mathbf{h}}$ satisfies the constraints of the optimization statement (16), we use Equation (20) to show that

$$\begin{aligned} &\int_A (F^n_{\mu,\hat{\mathbf{w}}}(\mathbf{t}) - F_\nu(\mathbf{t}))^2 \, \mathrm{d}\mathbf{t} \\ &\qquad \leq \int_A (F^n_{\mu,\hat{\mathbf{h}}}(\mathbf{t}) - F_\nu(\mathbf{t}))^2 \, \mathrm{d}\mathbf{t}. \end{aligned} \tag{28}$$

This result coupled with Equation (27) states that convergence of $F^n_{\mu,\hat{\mathbf{h}}}$ to F_ν in the L_2–norm implies convergence of $F^n_{\mu,\hat{\mathbf{w}}}$ to F_ν in the L_2–norm,

$$\lim_{n\to\infty} \int_A (F^n_{\mu,\hat{\mathbf{w}}}(\mathbf{t}) - F_\nu(\mathbf{t}))^2 \, \mathrm{d}\mathbf{t} = 0. \tag{29}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned} &\int_A \left| F^n_{\mu,\hat{\mathbf{w}}}(\mathbf{t}) - F_\nu(\mathbf{t}) \right| \, \mathrm{d}\mathbf{t} \\ &\leq \left(\int_A (F^n_{\mu,\hat{\mathbf{w}}}(\mathbf{t}) - F_\nu(\mathbf{t}))^2 \mathrm{d}\mathbf{t} \right)^{1/2} \cdot \left(\int_A (1)^2 \, \mathrm{d}\mathbf{t} \right)^{1/2} \\ &\leq M \cdot \left(\int_A (F^n_{\mu,\hat{\mathbf{w}}}(\mathbf{t}) - F_\nu(\mathbf{t}))^2 \, \mathrm{d}\mathbf{t} \right)^{1/2}, \end{aligned} \tag{30}$$

where $M < \infty$. Coupling this with (29), we show convergence in the L_1–norm,

$$\lim_{n\to\infty} \int_A \left| F^n_{\mu,\hat{\mathbf{w}}}(\mathbf{t}) - F_\nu(\mathbf{t}) \right| \, \mathrm{d}\mathbf{t} = 0. \tag{31}$$

□

For the unidimensional case (i.e., $d = 1$), Equation 31, is the Kantorovich or L_1-Wasserstein distance metric (Gibbs 2002). Convergence in the L_1-Wasserstein distance metric under our stated assumption, that μ is finitely supported, establishes weak convergence.

Corollary 1 *Let $\{\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^m\}$ be m random samples generated from the probability measure ν and $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$ be n random samples generated from the finite support probability measure μ where ν is absolutely continuous with respect to μ. Then there exists a set of empirical importance weights $\hat{\mathbf{w}} = [\hat{w}_1, \hat{w}_2, \cdots, \hat{w}_n]^\top$ satisfying (16) with vector $\mathbf{b}$ defined by Equation (15) such that*

$$\lim_{\min n,m\to\infty} \int_A \left| F^n_{\mu,\hat{\mathbf{w}}}(\mathbf{t}) - F^m_\nu(\mathbf{t}) \right| \, d\mathbf{t} = 0, \tag{32}$$

where $A = \{t \in \mathbb{R} \mid f_\mu(t) > 0\}$ is a bounded set.

Proof By combining the SLLN and the dominated convergence theorem we establish that the estimator F^m_ν converges to the target distribution function F_ν in the L_1–norm. That is, we have

$$\lim_{\min(m)\to\infty} \int_A \left| F_\nu(\mathbf{t}) - F^m_\nu(\mathbf{t}) \right| \, \mathrm{d}\mathbf{t} = 0. \tag{33}$$

By Theorem 1 in combination with Equation (33) and the triangle inequality, we define a bound on the quantity of interest and conclude the proof (Rudin 1987),

$$\begin{aligned} &\int_A \left| F^n_{\mu,\hat{\mathbf{w}}}(\mathbf{t}) - F^m_\nu(\mathbf{t}) \right| \, \mathrm{d}\mathbf{t} \\ &\leq \int_A \left| F^n_{\mu,\hat{\mathbf{w}}}(\mathbf{t}) - F_\nu(\mathbf{t}) \right| \, \mathrm{d}\mathbf{t} + \int_A \left| F_\nu(\mathbf{t}) - F^m_\nu(\mathbf{t}) \right| \, \mathrm{d}\mathbf{t}. \end{aligned} \tag{34}$$

□

5 Solving the Optimization Statement

In this section we examine the solution to the optimization statement (16). We begin by presenting the analytical solution to the optimization statement for $d = 1$ as this solution provides a better understanding of the optimization statement. The section concludes with the general solution to the optimization statement by numerical methods. Here we introduce methods that extend our approach to large-scale applications and demonstrate how to incorporate a target empirical distribution function.

5.1 Analytic Solution for $\mathbb{R}$

For the case when $d = 1$, we present the analytic solution to (16) and demonstrate that this solution satisfies the KKT conditions (19). Note that for this case the random variable is unidimensional, but the dimension of the optimization problem is still n, the number of proposal random samples. Without loss of generality, let the random samples of $X : \Omega \to \mathbb{R}$, $\{x^1, x^2, \dots, x^n\}$, be ordered such that $x^i < x^{i+1}$, $\forall\, i \in \{1, 2, \dots, n-1\}$. Using Equation (12), the matrix H is

$$H = \frac{1}{n^2}\begin{bmatrix} (1-x^1) & (1-x^2) & (1-x^3) & \dots & (1-x^n) \\ (1-x^2) & (1-x^2) & (1-x^3) & \dots & (1-x^n) \\ (1-x^3) & (1-x^3) & (1-x^3) & \dots & (1-x^n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (1-x^n) & (1-x^n) & (1-x^n) & \dots & (1-x^n) \end{bmatrix}. \tag{35}$$

Similarly, using Equation (14), the vector $\mathbf{b}$ is

$$\mathbf{b} = \frac{1}{n}\begin{bmatrix} \int_{x^1}^1 F_\nu(t)\ \mathrm{d}t \\ \int_{x^2}^1 F_\nu(t)\ \mathrm{d}t \\ \dots \\ \int_{x^n}^1 F_\nu(t)\ \mathrm{d}t \end{bmatrix}. \tag{36}$$

Then the solution to (16) is

$$\boldsymbol{\lambda} = \mathbf{0}_n, \tag{37}$$

$$\delta = \frac{1}{n}\int_{x^n}^1 F_\nu(t)\ \mathrm{d}t + \frac{x^n - 1}{n}, \tag{38}$$

and

$$\hat{\mathbf{w}} = \begin{bmatrix} \frac{n}{(x^2-x^1)}\int_{x^1}^{x^2} F_\nu(t)\ \mathrm{d}t & \\ \frac{n}{(x^3-x^2)}\int_{x^2}^{x^3} F_\nu(t)\ \mathrm{d}t & -\sum_{i=1}^{1} w_i \\ \vdots & \\ \frac{n}{(x^n-x^{n-1})}\int_{x^{n-1}}^{x^n} F_\nu(t)\ \mathrm{d}t & -\sum_{i=1}^{n-2} w_i \\ n & -\sum_{i=1}^{n-1} w_i \end{bmatrix}. \tag{39}$$

This solution can be derived using the active set method (Lawson and Hanson 1974); we omit the details of the derivation here for brevity, but show that this solution satisfies the KKT conditions (19).

First, it can be seen that the empirical importance weights (39) are by construction non-negative, since F_ν is a monotonically non-decreasing function. Thus, in this $d = 1$ case, the non-negativity constraints on the importance weights do not play a role in constraining the optimal solution and all the corresponding Lagrange multipliers are zero, $\lambda_i = 0$, $\forall\ i \in \{1, 2, \dots, n\}$. This result means that the complementarity conditions are satisfied. Second, summing the terms in (39), it is easy to show that the equality constraint $\mathbf{1}_n^\top \hat{\mathbf{w}} = n$ is satisfied. Lastly, we show that $H\hat{\mathbf{w}} = \mathbf{b} - \delta\mathbf{1}_n$ holds for each row entry $j \in \{1, 2, \dots, n\}$. That is, we show

$$\begin{aligned} &\frac{1-x^j}{n^2}\sum_{i=1}^{j}\hat{w}^i + \frac{1}{n^2}\sum_{i=j+1}^{n-1}\hat{w}^i(1-x^i) + \frac{\hat{w}^n(1-x^n)}{n^2} \\ &\quad = \frac{1}{n}\int_{x^j}^1 F_\nu(t)\ \mathrm{d}t - \left(\frac{1}{n}\int_{x^n}^1 F_\nu(t)\ \mathrm{d}t + \frac{x^n-1}{n}\right). \end{aligned} \tag{40}$$

By substituting the empirical importance weights (39) into the left-hand side of Equation (40) and simplifying, we obtain,

$$\begin{aligned} &\frac{1-x^j}{n^2}\sum_{i=1}^{j}\hat{w}^i + \frac{1}{n^2}\sum_{i=j+1}^{n-1}\hat{w}^i(1-x^i) + \frac{\hat{w}^n(1-x^n)}{n^2} \\ &\quad = \frac{1}{n}\int_{x^j}^{x^n} F_\nu(t)\ \mathrm{d}t + \frac{1-x^n}{n}. \end{aligned} \tag{41}$$

We obtain Equation (40) upon adding and subtracting b_n in Equation (41),

$$\begin{aligned} &\frac{1}{n}\int_{x^j}^{x^n} F_\nu(t)\ \mathrm{d}t + \frac{1}{n}(1-x^n) + b_n - b_n \\ &\quad = \frac{1}{n}\int_{x^j}^1 F_\nu(t)\ \mathrm{d}t - \left(\frac{1}{n}\int_{x^n}^1 F_\nu(t)\ \mathrm{d}t + \frac{x^n-1}{n}\right). \end{aligned} \tag{42}$$

Since the KKT conditions are satisfied, (37)–(39) represent the solution to the optimization problem (16) for $d = 1$.

If instead we are given a target empirical distribution function represented by m random samples $\{y^1, y^2, \dots, y^m\}$ generated from ν, then the optimal solution remains the same, with F_ν in (38)–(39) replaced by F_ν^m.

We conclude this subsection by demonstrating that the empirical importance weights defined in (39) result in weak convergence of the weighted proposal empirical distribution function to the target distribution function. That is, we show that

$$\lim_{n\to\infty} F^n_{\mu;\hat{\mathbf{w}}}(t) \overset{a.e.}{=} F_\nu(t), \tag{43}$$

for every continuity point $t \in A$ of $F_\nu(t)$ where $A = \{t \in \mathbb{R} \mid f_\mu(t) > 0\}$ is a bounded set. Given ν is absolutely continuous with respect to μ, let $i = \{j \in \{1, 2, \dots, n-1\} \mid \hat{t} \in [x^j, x^{j+1})\}$ where $\hat{t}$ is a continuity point of $F_\nu(\hat{t})$. We expand $F^n_{\mu;\hat{\mathbf{w}}}(\hat{t})$ using the empirical importance weights from (39):

$$\begin{aligned} F^n_{\mu;\hat{\mathbf{w}}}(\hat{t}) &= \frac{1}{n}\sum_{i=1}^{n} \hat{w}^i \mathbb{I}(x^i \le \hat{t}) \\ &= \frac{1}{x^{i+1}-x^i}\int_{x^i}^{x^{i+1}} F_\nu(t)\ \mathrm{d}t. \end{aligned} \tag{44}$$

Given that F^n_ν is monotonically non-decreasing and using Equation (44), we obtain the following inequality:

$$F_\nu(x^i) \le F_\nu(\hat{t}), F^n_{\mu;\hat{\mathbf{w}}}(\hat{t}) < F_\nu(x^{i+1}). \tag{45}$$

Since the target distribution is continuous at $\hat{t}$, this ensures for every $\epsilon > 0$ there exists a $\delta > 0$ such that $|F_\nu(x) - F_\nu(\hat{t})| \le \epsilon$ for all points $x \in A$ for which $|x - \hat{t}| \le \delta$. Now, since ν is absolutely continuous with respect to μ, there exists a finite n which is sufficiently large that we can find an $i = \{j \in \{1, 2, \ldots, n-1\} \,|\, \hat{t} \in [x^j, x^{j+1}]\}$ that yields $|x^i - \hat{t}| \le \delta$ and $|x^{i+1} - \hat{t}| \le \delta$. This implies $|F_\nu(x^i) - F_\nu(\hat{t})| \le \epsilon$ and $|F_\nu(x^{i+1}) - F_\nu(\hat{t})| \le \epsilon$. Lastly, by Equation (45) and application of the triangle inequality, we obtain

$$\begin{aligned} &|F^n_{\mu;\hat{\mathbf{w}}}(\hat{t}) - F_\nu(\hat{t})| \\ &\quad < |F_\nu(x^i) - F_\nu(x^{i+1})| \\ &\quad \le |F_\nu(x^i) - F_\nu(\hat{t})| + |F_\nu(\hat{t}) - F_\nu(x^{i+1})| \\ &\quad \le 2\epsilon, \end{aligned} \tag{46}$$

which yields the desired result for every continuity point $\hat{t} \in A$ of $F_\nu(\hat{t})$ as n tends to infinity.

5.2 Optimization Algorithm

Here we focus on the optimization statement for the general case when $d > 1$ and examine algorithms that extend our approach to large-scale applications (i.e., a large number proposal random samples, n). The challenge with solving the optimization statement (16) when $d > 1$ is that the matrix H is not analytically invertible as was the case for $d = 1$. As a result, we rely on a numerical optimization routine to solve (16).

The optimization statement in (16) is classified as a single linear equality and box-constrained quadratic program. A popular application which falls into this class of problems is the dual form of the nonlinear support vector machine optimization statement (Vapnik 1998). That application resulted in algorithms to extend single linear equality and box-constrained quadratic programs to large-scale applications (Dai and Fletcher 2006, Lin et al. 2009, Platt 1999, Zanni 2006). For this work we have selected two large-scale optimization algorithms that exploit our problem structure: the Frank-Wolfe algorithm (Frank and Wolfe 1956) and the Dai-Fletcher algorithm (Dai and Fletcher 2006).

The Frank-Wolfe algorithm is well-suited for solving (16) since the objective is a differentiable convex function and the constraints are a bounded convex set. The core idea behind the Frank-Wolfe algorithm is to approximate the objective with a linear function and then take a step in the descent direction. The Frank-Wolfe algorithm is particularly attractive because it has well established convergence rates, low computational complexity, and can generate sparse solutions. The pseudo algorithm describing the Frank-Wolfe algorithm tailored to the optimization statement (16) is given in Algorithm 1. Note that the step length α can be chosen to be the deterministic value $2/(2+k)$, where k is the iteration number, or alternatively α can be chosen such that it minimizes the objective function of (16) at that particular iteration. The computational complexity of the Frank-Wolfe algorithm per iteration is low since it requires only a rank-one update to the gradient vector at each iteration. With the structure of our problem, this update can be computed very efficiently.

Algorithm 1: Frank-Wolfe Algorithm for solving (16).

```
Data: Random samples $\mathbf{x}$, vector $\mathbf{b}$, initial
      feasible solution $\mathbf{w}_0$, and termination
      criteria.
Result: Empirical importance weight vector $\hat{\mathbf{w}}$.
Initialization: $\hat{\mathbf{w}} = \mathbf{w}_0$
$\mathbf{a} = H\hat{\mathbf{w}}$,
$\mathbf{g} = \mathbf{a} - \mathbf{b}$,
for $k = 1, 2, \ldots$ do
    · Steepest descent direction:
      $\ell = \arg\min_{i \in \{1,2,\ldots,n\}}(g_i)$,
    · $\bar{w}_i = \begin{cases} 1, & \text{if } i = \ell \\ 0, & \text{otherwise} \end{cases}$
    · Set $\hat{\mathbf{w}} = \hat{\mathbf{w}} + \alpha(\bar{\mathbf{w}} - \hat{\mathbf{w}})$, where $\alpha \in [0, 1]$,
    · $\mathbf{a} = (1-\alpha)\mathbf{a} + \alpha H_{(\cdot,\ell)}$,
    · $\mathbf{g} = \mathbf{a} - \mathbf{b}$,
    if (termination criteria satisfied) then
        Exit
    end
end
```

As a second option, we examine the Dai-Fletcher optimization algorithm. The general idea of the Dai-Fletcher algorithm is to construct the Lagrangian penalty function

$$L(\mathbf{w}; \delta) = \frac{1}{2}\big(\mathbf{w}^\top H\mathbf{w} - 2\mathbf{w}^\top\mathbf{b}\big) - \delta(\mathbf{1}_n^\top\hat{\mathbf{w}} - n), \tag{47}$$

where δ is the equality constraint Lagrangian multiplier. Then for any fixed δ, the box-constrained quadratic program (Dai and Fletcher 2005),

$$\begin{aligned} \hat{\mathbf{w}}(\delta) = \arg\min_{\mathbf{w}} \;\; & L(\mathbf{w};\delta) \\ s.t. \;\; & w_i \geq 0, \; \forall\, i \in \{1,2,\dots,n\}, \end{aligned} \tag{48}$$

is solved. Next, δ is adjusted in an outer secant-like method to solve the single nonlinear equation,

$$r(\delta) = \mathbf{1}_n^\top \hat{\mathbf{w}}(\delta) - n = 0. \tag{49}$$

That is, Equation (49) enforces that the solution of (48) satisfies the equality constraint (i.e., exists in the feasible region). In summary, for each iteration of the Dai-Fletcher algorithm, a box-constrained projected gradient-based algorithm is used to compute a new solution for (48). This solution is projected into a feasible region using a secant projection approximation method, thereby satisfying Equation (49). A summary of the Dai-Fletcher algorithm is given in Algorithm 2.

Algorithm 2: Dai-Fletcher Algorithm for solving (16).

```
Data: Random samples $\mathbf{x}$, vector $\mathbf{b}$, initial
      solution $\mathbf{w}_0$, and termination criteria.
Result: Empirical importance weight vector $\hat{\mathbf{w}}$.
Initialization: $\hat{\mathbf{w}} = \mathbf{w}_0$
for $k = 1, 2, \dots$ do
    · Compute gradient of (47),
    · Take a steepest descent step,
    · Project into feasible region by (49),
    · Possibly carry out a line search,
    · Calculate a Barzilai-Borwein step length,
    · Update the line search control parameters,
    if (termination criteria satisfied) then
        Exit
    end
end
```

The termination criteria in Algorithm 1 or Algorithm 2 may incorporate a maximum number of iterations and a minimum tolerance associated with the gradient of the objective function in (16) or the Lagrangian penalty function, Equation (47), respectively. Although Algorithm 1 and Algorithm 2 may in some cases terminate prior to locating the global optimal solution, by construction they generate a sequence of feasible iterates. In Section 6, we evaluate the performance of these two algorithms over a range of parameters. The remainder of this section discusses numerical techniques to extend our approach to large-scale applications and to incorporate the target empirical distribution function.

The largest computational expense in Algorithm 2 is in the calculation the matrix-vector product, $H\mathbf{w}$. The matrix-vector product, $H\mathbf{w}$, is also required in Algorithm 1, but since it only needs to be evaluated once, it has less impact on the computational performance of Algorithm 1. In the circumstance where the matrix H is small, the matrix can be assembled and stored for computations; however, large-scale applications (many samples) may prohibit assembly of the matrix H. In these cases, one option is to use the Frank-Wolfe algorithm and avoid repeated matrix-vector products altogether. Since in some cases the Dai-Fletcher algorithm may yield improved convergence rates, another option is to exploit the structure in the problem to reduce the numerical complexity of the matrix-vector product calculations. In particular, we recognize that since active set empirical importance weights are zero, they do not contribute to the matrix-vector product. As a result, only the columns of matrix H associated with passive set empirical importance weights are required for the matrix-vector product calculation. Thus, the numerical complexity of the gradient evaluation is $\mathcal{O}(n|\mathcal{P}|d^2 + n|\mathcal{P}|)$, where the first term captures the construction of matrix H, the second term captures the matrix-vector product, and $|\mathcal{P}|$ denotes the cardinality of the passive set. In addition, efficient algorithms which rely on the divide-and-conquer technique have been developed and applied successfully to Equation (9) (Bentley 1980, Heinrich 1996). Lastly, one may take advantage of parallel routines to divide and conquer the matrix-vector product (Nickolls et al. 2008, Zanni et al. 2006).

Solving the optimization problem also requires evaluating the vector $\mathbf{b}$. Here we will describe two special circumstances for which the vector $\mathbf{b}$ can be directly evaluated: an independently distributed target distribution function and a target empirical distribution function. For an independently distributed target distribution function we can define the target measure ν as the product of d individual measures, $\nu = \nu_1 \otimes \cdots \otimes \nu_d$, where ν_i is the i^{th} target measure on $\mathbb{R}$. Then the resulting target distribution function can be expanded using a product series as $F_\nu(\mathbf{t}) = \prod_{k=1}^d F_{\nu_k}(t_k)$. The vector $\mathbf{b}$, Equation (14), can then be evaluated as a Hadamard product over each dimension:

$$\mathbf{b} = \begin{bmatrix} \int_{x_1^1}^1 F_{\nu_1}(t)\,\mathrm{d}t \\ \int_{x_1^2}^1 F_{\nu_1}(t)\,\mathrm{d}t \\ \cdots \\ \int_{x_1^n}^1 F_{\nu_1}(t)\,\mathrm{d}t \end{bmatrix} \circ \cdots \circ \begin{bmatrix} \int_{x_d^1}^1 F_{\nu_d}(t)\,\mathrm{d}t \\ \int_{x_d^2}^1 F_{\nu_d}(t)\,\mathrm{d}t \\ \cdots \\ \int_{x_d^n}^1 F_{\nu_d}(t)\,\mathrm{d}t \end{bmatrix}. \tag{50}$$

If the target distribution function is unknown and is instead estimated by the target empirical distribution function given m random samples $\{\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^m\}$ generated from ν, then there also exists an approach to directly construct the vector $\mathbf{b}$. The approach requires expanding Equation (15) as follows:

$$\begin{aligned} \mathbf{b} &= \frac{1}{nm} \sum_{j=1}^{m} \int_0^1 \cdots \int_0^1 \mathbb{I}(\mathbf{x}^i \leq \mathbf{t}) \mathbb{I}(\mathbf{y}^j \leq \mathbf{t}) \; \mathrm{d}\mathbf{t}, \\ &= \hat{H}\mathbf{v}, \end{aligned} \tag{51}$$

and noting the similarities with the matrix-vector product $H\mathbf{w}$. Here we define $\mathbf{v} \in \mathbb{R}^m$ as the importance weights of the target random samples (i.e., $v_i = 1$, $\forall\, i \in \{1, 2, \ldots, m\}$). Additionally, we define an entry of matrix $\hat{H} \in \mathbb{R}^{n \times m}$ as

$$\hat{H}_{i,j} = \frac{1}{nm} \prod_{k=1}^{d} \int_{\hat{z}_k^{i,j}}^{1} \; \mathrm{d}t_k, \tag{52}$$

where $\hat{z}_k^{i,j} = max(x_k^i, y_k^j)$. The vector $\mathbf{b}$ is then computed by the matrix-vector product (51).

6 Applications

In this section we apply the proposed approach to a number of numerical experiments. In Section 6.1, we demonstrate the properties of the proposed approach on a unidimensional distribution problem. Section 6.2 compares the proposed approach to previous approaches on an importance sampling problem over a range of parameters. Lastly, in Section 6.3, we examine the relationship between discrepancy theory and the proposed approach when the proposal and target are distributed according to the uniform distribution. We also use this opportunity to evaluate the performance of the Frank-Wolfe algorithm and Dai-Fletcher algorithm.

6.1 Unidimensional Numerical Example

This analysis revisits the problem presented in Figure 1. However, instead of using the analytic empirical importance weights (39), as was done in Figure 1, this example uses the Frank-Wolfe algorithm with a step length $\alpha = 2/(2 + k)$ and premature termination to obtain sparse solutions (recall that the Frank-Wolfe algorithm updates only one weight at each iteration). To initialize the Frank-Wolfe algorithm (i.e., $\mathbf{w}_0$), we choose an empirical importance weight vector with entries equal to

$$w_{0,i} = \begin{cases} n, \text{ if } i = \ell \\ 0, \text{ otherwise} \end{cases}, \tag{53}$$

where $\ell \in \{1, 2, \ldots, n\}$ is selected uniformly at random. The results of this numerical experiment using $n = 100$ proposal random samples are presented in Figure 2. The top and center plots show the results after 25 and 100 iterations, respectively, of the Frank-Wolfe algorithm.

These results illustrate that the proposed approach produces accurate representations of the target distribution function. Since the support of the proposal distribution function is finite, we can guarantee weak convergence by Theorem 1 (i.e., L_1-Wasserstein distance metric); permitting the Frank-Wolfe algorithm to run for more iterations would recover the analytic empirical importance weights (39). However, the sparse empirical importance weights, shown on the top plot of Figure 2, are already a good approximation and may be advantageous if one wishes to evaluate computationally expensive statistics with respect to a complex or unknown target distribution function. That is, with the proposed approach, we have demonstrated one can approximate a target distribution function using a small set of optimally weighted proposal random samples. These results also illustrate that the proposed approach naturally accounts for clustering of the proposal random samples and other deviations from the original proposal distribution function. If we were to use the Radon-Nikodym importance weights the clustering of the proposal random samples would not have been accounted for as demonstrated on the bottom plot of Figure 2. In the next section we compare our approach to previous approaches over a range of multiple-dimensional distributions with the application of the target empirical distribution function.

6.2 Importance Sampling

Importance sampling is a commonly used technique for estimating statistics of a target distribution given random samples generated from a proposal distribution. As an example, we consider the setting where we have a model $g : \mathbb{R}^d \to \mathbb{R}$ that maps a d-dimensional input $\mathbf{x}$ to a scalar output $g(\mathbf{x})$. g could, for example, be a computational model that estimates a system performance metric (output of interest) as a function of system geometric parameters (model inputs). For many applications of interest, g embodies partial differential equations and is expensive to evaluate. We have available to us the proposal random samples of the input $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$, drawn from the (unknown) input proposal probability density function f_μ. We also have available the corresponding model evaluations for each proposal sample (i.e., $\{g(\mathbf{x}^1), g(\mathbf{x}^2), \ldots, g(\mathbf{x}^n)\}$).

We consider the case where the goal is to evaluate statistics of g, but where the inputs $\mathbf{x}$ are now distributed according to a target distribution function. This situation occurs if we gain additional knowledge of the input distribution (e.g., refined estimates from experts or from upstream models), or if we want to study the system under a variety of different input scenarios (e.g., during a design process). If the model g is expensive to evaluate, as is the case for many applications in science and engineering, then it becomes intractable to re-evaluate g over samples from many different input target distributions; instead we use importance sampling to reweight the available proposal samples. The decomposition-based uncertainty analysis approach proposed in Amaral et al. (2014) is one concrete example of this setting.

In the numerical example presented here, the proposal random samples are distributed according to $X \sim \mathcal{N}(\mathbf{1}/\sqrt{d}, \mathbf{I})$, where $\mathbf{I} \in \mathbb{R}^{d\times d}$ is the identity matrix, and the target random samples are distributed according to $Y \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Since these measures have infinite support, although our approach is still applicable, we cannot guarantee convergence in the L_1–norm. In this illustration, we assume that we do not know f_μ or f_ν, but are provided with random samples from each: $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$ drawn from f_μ and $\{\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^m\}$ drawn from f_ν. We have the proposal model evaluations $\{g(\mathbf{x}^1), g(\mathbf{x}^2), \ldots, g(\mathbf{x}^n)\}$, but *not* the target model evaluations $\{g(\mathbf{y}^1), g(\mathbf{y}^2), \ldots, g(\mathbf{y}^m)\}$. Instead of evaluating the computational model over the target random samples, we use our proposal model evaluations and perform a change of measure to approximate the statistics of interest with regards to g where the inputs to g are distributed according to the target distribution function.

The statistic of interest for this numerical study is

$$E \equiv \mathbb{E}_\nu[g(\mathbf{t})] = \int_\Omega g(\mathbf{t}) f_\nu(\mathbf{t})\ \mathrm{d}\mathbf{t}. \tag{54}$$

We will estimate E as defined in Equation (54) using the following weighted Monte Carlo integration rule:

$$E_n = \sum_{i=1}^{n} \hat{w}^i g(\mathbf{x}^i), \tag{55}$$

where $\hat{\mathbf{w}}$ is obtained using one of the approaches described below. For comparison purposes we perform an exhaustive Monte Carlo simulation using the target distribution function to approximate Equation (54). The result is used as the “truth value” for E when comparing to the approximate estimates computed using Equation (55).

Our numerical experiment compares the following approaches:

1. **Radon-Nikodym Derivative (RND)**: The Radon-Nikodym importance weights are obtained by computing the ratio of the target and proposal probability density functions. These results are provided

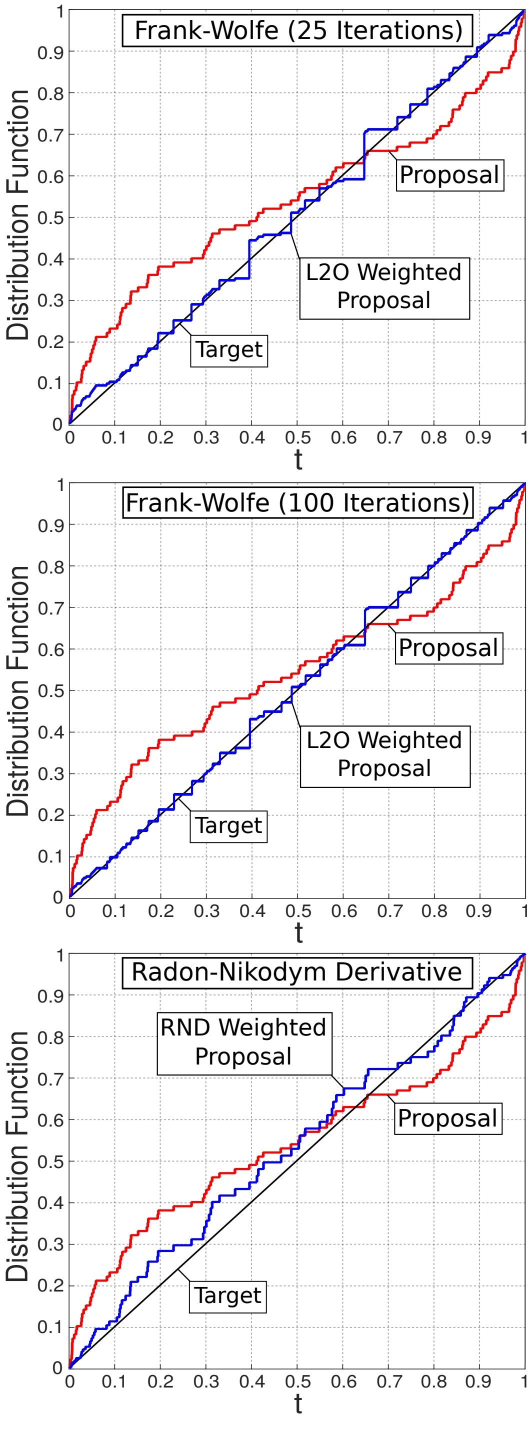


Fig. 2: Our empirical importance weights are determined using the Frank-Wolfe algorithm with step length $\alpha = 2/(2+k)$ and premature termination. We use $n = 100$ proposal random samples generated from a beta distribution function (i.e., $\mathcal{B}(0.5, 0.5)$) and the target is the uniform distribution function (i.e., $\mathcal{U}(0,1)$). Terminating the Frank-Wolfe algorithm after 25 iterations (top) results in a sparse empirical importance weight vector. Terminating the Frank-Wolfe algorithm after 100 iterations (center) results in a dense solution and a more accurate representation of the target distribution function. For comparison, the Radon-Nikodym importance weighted empirical proposal distribution function is provided in the (bottom) plot.

here for comparison purposes only, since this approach uses additional information that is not available to the other approaches.

2. L_2**–norm Optimal Weight (L2O)**: Our optimal empirical importance weights are obtained by solving the optimization statement developed in this paper (16). For $d = 1$, we use analytic empirical importance weights (39) where F_ν is replaced by the target empirical distribution function F_ν^m. For $d > 1$, we use the Dai-Fletcher algorithm and terminate after $2\max(n, m)$ iterations where n and m are the number of proposal and target random samples, respectively. For the implementation of the Dai-Fletcher algorithm, we compute the matrix H once for each case considered and store it for use at each optimization iteration.

3. **Kernel Density Estimation (KDE)**: The kernel density estimation (Scott 1992) approach is applied to approximate f_ν and f_μ, denoted by $\tilde{f}_\nu$ and $\tilde{f}_\mu$, from their respective random samples. We compute the Radon-Nikodym importance weights by approximating the Radon-Nikodym derivative with $\tilde{f}_\nu/\tilde{f}_\mu$. The KDE uses Gaussian kernels where the kernel bandwidth is selected using the minimal mean squared error.

4. **Kernel Mean Matching (KMM)**: The kernel mean matching method (Huang et al. 2007) aims to match the moments between the proposal and target distribution using a Gaussian reproducing kernel (i.e., $K(\mathbf{t}, \mathbf{t}')$). The KMM empirical optimization statement using proposal samples $\mathbf{x}^i$, $i = 1, \ldots, n$, and target samples $\mathbf{y}^j$, $j = 1, \ldots, m$, is formulated as

$$\begin{aligned}\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \quad & \frac{1}{2}\sum_{i,j=1}^{n} \beta_i\beta_j K(\mathbf{x}^i, \mathbf{x}^j) - \\ & \frac{n}{m}\sum_{i=1}^{n} \beta_i \sum_{j=1}^{m} K(\mathbf{x}^i, \mathbf{y}^j) \\ s.t. \quad & 0 \leq \beta_i, \leq B, \ \forall\, i \in \{1, \ldots, n\} \\ & \left| \frac{1}{n}\sum_{i=1}^{n} \beta_i - 1 \right| \leq \epsilon,\end{aligned}$$

where the optimization variables $\boldsymbol{\beta}$ are the density ratio estimates, B is an upper limit on the density ratio, and ϵ is a user specified tolerance, recommended to be set as $\frac{B}{\sqrt{m}}$. The optimization problem solution directly provides the density ratio estimates at their respective proposal samples,

$$\tilde{h}(\mathbf{x}^i) = \beta_i, \ i = 1, \ldots, n. \tag{56}$$

The Gaussian kernel variance parameter is selected based on a five-fold cross validation.

5. **Ratio Fitting (uLS)**: The unconstrained least squares importance fitting (Kanamori et al. 2009) approach is applied to approximate $h = f_\nu/f_\mu$. Here h is represented by the linear model,

$$\tilde{h}(\mathbf{t}) = \sum_{i=1}^{b} \hat{\beta}_i \phi_i(\mathbf{t}), \tag{57}$$

where b is the number of basis functions, $\{\phi_i\}_{i=1}^b$ are the basis functions, and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_b)^\top$ are the parameters to be learned. The parameters are obtained by solving the following optimization statement,

$$\begin{aligned}\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \quad & \frac{1}{2}\int \left(\tilde{h}(\mathbf{t}) - h(\mathbf{t})\right)^2 f_\nu(\mathbf{t})\ \mathrm{d}\mathbf{t} + \gamma\boldsymbol{\beta}^\top\mathbf{1} \\ s.t. \quad & \beta_i \geq 0, \ \forall\, i \in \{1, \ldots, b\},\end{aligned}$$

where γ is the regularization parameter. The basis functions are Gaussian kernel models centered at the target random samples. The Gaussian kernel variance and regularization parameter are selected based on a 5-fold cross validation. Note that although the unknown Radon-Nikodym derivative appears in the objective, it is not explicitly evaluated.

6. **Divergence Fitting (KLD)**: The Kullback-Liebler (divergence) importance estimation (Sugiyama et al. 2012) approach applies the linear model in Equation (57). The parameters are obtained by solving the following optimization statement,

$$\begin{aligned}\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \quad & \int f_\nu(\mathbf{t}) log\left(\frac{h(\mathbf{t})}{\tilde{h}(\mathbf{t})}\right)\ \mathrm{d}\mathbf{t} + \lambda\boldsymbol{\beta}^\top\mathbf{1} \\ s.t. \quad & \sum_{i=1}^{n}\sum_{j=1}^{b} \beta_j \phi(\mathbf{x}^i) = n \\ s.t. \quad & \beta_i \geq 0, \ \forall\, i \in \{1, \ldots, b\},\end{aligned}$$

where the equality constraint ensures that $\tilde{h}$ defines a probability density function. The basis functions are Gaussian kernel models centered at the target random samples. The Gaussian kernel variance and regularization parameter are selected based on a 5-fold cross validation. Note that although the unknown Radon-Nikodym derivative appears in the objective, it is not explicitly evaluated.

The five approaches presented above are tested over the following four scenarios:

1. $n = 2^{10}, m = 2^{10}$ and $d = \{1, 2, 5, 10\}$,
2. $n = 2^{10}, m = 2^{12}$ and $d = \{1, 2, 5, 10\}$,
3. $n = 2^{12}, m = 2^{10}$ and $d = \{1, 2, 5, 10\}$,
4. $n = 2^{12}, m = 2^{12}$ and $d = \{1, 2, 5, 10\}$.

For all scenarios, the results are the average over 100 independent trials and the quality of the results is quantified by

$$r_n = \frac{|E_n - E|}{E}. \tag{58}$$

Table 1 presents the results for each scenario, where g is Ackley's function (Galletly 1998),

$$\begin{aligned} g(\mathbf{t}) = &-20 \exp\left(-0.2\sqrt{\frac{1}{d}\sum_{i=1}^{d} t_i^2}\right) - \\ &\exp\left(\frac{1}{d}\sum_{i=1}^{d} cos(2\pi t_i)\right) + 20 + \exp(1). \end{aligned} \tag{59}$$

The results demonstrate that our approach accurately estimates E using only the proposal and target random samples and the proposal model evaluations. Our approach is also shown to result in estimates that are comparable in accuracy to using the Radon-Nikodym importance weights, although we did not require that our empirical importance weights converge to the Radon-Nikodym importance weights. This outcome can be attributed to the results presented in Figure 2; that is, we only required that the weighted proposal empirical distribution function matches the target distribution function. By increasing the number of proposal and target random samples we enrich the proposal and target empirical distribution functions and as a result improve the accuracy of our estimate. Overall, the results in Table 1 demonstrate that the proposed approach evaluates the statistic of interest more accurately than current practices for all scenarios considered. The normalized computational time required to compute the importance weights for scenario 4 is shown in Table 2. The computational time required by our approach is comparable to the computational times required by the other approaches.

Additionally, we repeat the scenario $\{n = 2^{12}, m = 2^{12}\}$ and $d = 5$ using four different models for g:

1. G-Function (Saltelli 1995),

$$g_1(\mathbf{t}) = \prod_{i=1}^{d} \frac{|4t_i - 2| + (i-2)/2}{1 + (i-2)/2}$$

2. Morokoff & Caflisch (Morokoff 1995),

$$g_2(\mathbf{t}) = (1 + 1/d)^d \prod_{i=1}^{d} (|t_i|)^{1/d}$$

Table 1: The error metric r_n, Equation (58), measured as a percentage, for the six methods and four scenarios. Results are averaged over 100 independent trials and the term in parentheses is the corresponding standard deviation. Bold text indicates the best estimate for each scenario (not considering the Radon-Nikodym derivative, which is shown only for illustration). The "truth values" used in these computations are computed using an exhaustive Monte Carlo simulation: $E_\nu[g(\mathbf{t})|d = 1] = 4.2830$, $E_\nu[g(\mathbf{t})|d = 2] = 4.7650$, $E_\nu[g(\mathbf{t})|d = 5] = 5.1018$, $E_\nu[g(\mathbf{t})|d = 10] = 5.2212$.

	Low: $n = 2^{10}$, Low: $m = 2^{10}$			
	$d = 1$	$d = 2$	$d = 5$	$d = 10$
RND	2.79(2.07)	1.65(1.27)	0.78(0.59)	0.57(0.51)
L2O	**1.32(0.95)**	**0.85(0.67)**	**0.52(0.39)**	**0.47(0.34)**
KDE	2.65(1.80)	5.83(1.04)	12.4(1.02)	10.3(0.73)
KMM	2.03(1.46)	2.59(1.08)	2.84(0.75)	1.72(0.48)
uLS	4.07(2.65)	7.13(1.37)	5.36(5.85)	2.34(0.44)
KLD	11.1(2.15)	6.38(1.61)	7.31(0.79)	4.56(0.52)

	Low: $n = 2^{10}$, High: $m = 2^{12}$			
	$d = 1$	$d = 2$	$d = 5$	$d = 10$
RND	2.79(2.07)	1.65(1.27)	0.78(0.59)	0.57(0.51)
L2O	**0.69(0.57)**	**0.52(0.42)**	**0.32(0.24)**	**0.35(0.27)**
KDE	2.54(0.96)	5.86(0.68)	12.8(0.87)	11.5(0.66)
KMM	1.68(0.92)	2.46(0.65)	2.98(0.59)	1.80(0.37)
uLS	3.59(1.72)	6.90(1.34)	1.19(2.39)	2.34(0.44)
KLD	11.0(1.89)	5.99(1.22)	7.95(0.76)	4.48(0.45)

	High: $n = 2^{12}$, Low: $m = 2^{10}$			
	$d = 1$	$d = 2$	$d = 5$	$d = 10$
RND	1.48(1.05)	0.91(0.68)	0.48(0.35)	0.31(0.28)
L2O	**1.24(0.85)**	**0.78(0.60)**	**0.54(0.39)**	**0.36(0.27)**
KDE	1.68(1.25)	4.22(0.89)	11.8(0.68)	10.2(0.48)
KMM	1.46(1.08)	1.27(0.83)	1.25(0.66)	0.78(0.38)
uLS	2.94(2.02)	5.70(1.35)	9.98(0.62)	2.36(0.24)
KLD	11.4(1.42)	6.28(1.49)	7.33(0.69)	4.44(0.42)

	High: $n = 2^{12}$, High: $m = 2^{12}$			
	$d = 1$	$d = 2$	$d = 5$	$d = 10$
RND	1.48(1.05)	0.91(0.68)	0.48(0.35)	0.31(0.28)
L2O	**0.64(0.42)**	**0.47(0.36)**	**0.29(0.21)**	**0.26(0.16)**
KDE	1.43(0.76)	4.28(0.53)	12.3(0.51)	11.4(0.39)
KMM	0.93(0.67)	1.06(0.56)	1.43(0.45)	0.81(0.26)
uLS	2.37(1.28)	5.34(0.95)	11.0(1.15)	2.36(0.24)
KLD	11.5(1.08)	5.86(0.96)	7.77(0.69)	4.35(0.28)

Table 2: Normalized computational times required for computing the last scenario in Table 1. The results are normalized by the fastest method in each case and are averaged over 100 independent trials.

	High: $n = 2^{12}$, High: $m = 2^{12}$		
	$d = 2$	$d = 5$	$d = 10$
L2O	7.17	5.27	3.51
KDE	1	1	3.95
KMM	17.9	5.18	1.37
uLS	8.42	5.22	2.46
KLD	14.2	7.03	1

3. Oscillatory Integrand Family (Genz 1984),

$$g_3(\mathbf{t}) = cos(\pi + \sum_{i=1}^{d} t_i)$$

4. Product Peak Integrand Family (Genz 1984),

$$g_4(\mathbf{t}) = \prod_{i=1}^{d} \frac{1}{2^{-2} + (t_i - 0.5)^2}.$$

The results from this numerical study are provided in Table 3 and show that our approach performs the change of measure more accurately than previous approaches on three of the four test functions. For the model g_3 (the Oscillatory Integrand Family), the accuracy of our approach is slightly worse on average but comparable to the other approaches.

Lastly, we demonstrate the applicability of our approach for an example where $g(\mathbf{t})$ is computationally expensive. For this application problem, $g(\mathbf{t})$ represents a computational tool that evaluates an aircraft's performance using low-order physical models. The output of $g(\mathbf{t})$ is the fuel energy consumption per payload-range (PFEI) of an aircraft. The proposal and target input distributions to $g(\mathbf{t})$ are provided in Table 4. In this illustration, we assume that we do not know f_μ or f_ν, but are provided with random samples from each (i.e., random samples may have been generated from upstream models or experimentations). We thus have $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$ drawn from the unknown f_μ and $\{\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^m\}$ drawn from the unknown f_ν, where $n = m = 50,000$. We also have available the proposal model evaluations $\{g(\mathbf{x}^1), g(\mathbf{x}^2), \ldots, g(\mathbf{x}^n)\}$, which took approximately 68 minutes to generate on a desktop computer[1], but we do *not* have available the target model evaluations $\{g(\mathbf{y}^1),\, g(\mathbf{y}^2), \ldots, g(\mathbf{y}^m)\}$.

Table 3: The error metric r_n, Equation (58), measured as a percentage, for the six methods and all four functions under the last scenario in Table 1. Results are averaged over 100 independent trials and the term in parentheses is the corresponding standard deviation. Bold text indicates the best estimate for each scenario not considering the Radon-Nikodym derivative. The "truth values" used in these computations are computed using an exhaustive Monte Carlo simulation: $E_\nu[g_1(\mathbf{t})] = 21.7970$, $E_\nu[g_2(\mathbf{t})] = 1.4733$, $E_\nu[g_3(\mathbf{t})] = -0.0820$, $E_\nu[g_4(\mathbf{t})] = 3.5483 \times 10^{-4}$.

	$d = 5$; High: $n = 2^{12}$, High: $m = 2^{12}$			
	$g_1(\mathbf{t})$	$g_2(\mathbf{t})$	$g_3(\mathbf{t})$	$g_4(\mathbf{t})$
RND	5.41(4.55)	1.01(0.80)	20.9(14.7)	1.16(0.93)
L2O	**2.42(1.53)**	**0.62(0.49)**	58.3(41.0)	**0.69(0.55)**
KDE	37.8(1.78)	20.1(0.94)	76.0(21.7)	42.34(1.32)
KMM	6.96(1.99)	2.60(0.96)	**11.7(8.91)**	2.57(0.79)
uLS	23.9(3.61)	18.1(1.94)	119.9(20.3)	29.9(1.49)
KLD	34.6(2.72)	12.8(1.24)	59.7(29.7)	28.6(1.61)

Table 4: The proposal and target distributions for $g(\mathbf{t})$ with $\mathbf{t} = \{t_1, t_2, t_3, t_4, t_5\}$. The physical representation of the inputs are; t_1 =turbine metal temperature [K], t_2 = turbine inlet total temperature for cruise [K], t_3 = operating pressure ratio [-], t_4 = max allowable wing spar cap stress [psi], and t_5 = start-of-cruise altitude [ft]. We use $\mathcal{T}(a, b, c)$ to represent a triangular distribution with lower limit a, mode b, and upper limit c. We use $\mathcal{U}(a, b)$ to represent a uniform distribution with lower limit a and upper limit b. In this study, the target distributions are absolutely continuous with respect to the proposal distributions.

	Proposal	Target
t_1	$\mathcal{T}(1122, 1222, 1322)$	$\mathcal{U}(1172, 1272)$
t_2	$\mathcal{U}(1541, 1692)$	$\mathcal{U}(1541, 1641)$
t_3	$\mathcal{T}(22.2, 27.2, 28.6)$	$\mathcal{U}(24.2, 28.2)$
t_4	$\mathcal{T}(27500, 30000, 32500)$	$\mathcal{U}(28500, 31500)$
t_5	$\mathcal{U}(33000, 38000)$	$\mathcal{U}(34000, 36000)$

The objective of this numerical study is to evaluate statistics of interest with regards to the target distribution. This is beneficial if one has already performed all the proposal evaluations in an "offline" phase and would like to evaluate the target statistics of interest in an "online" phase. The results from this numerical study are provided in Figure 3. These results indicate that

[1] All computations were performed on a six-core Intel Core i7-5930K CPU desktop computer.

our approach accurately quantifies the output of interest distribution function from the proposal model evaluations. The target distribution function, shown here for comparison, required approximately 68 minutes to compute on a desktop computer. In comparison, our approach required 85 seconds to evaluate the empirical importance weighted proposal distribution function.

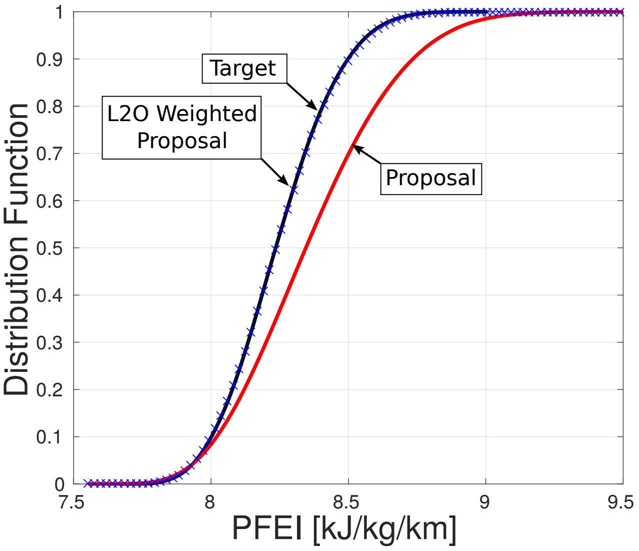


Fig. 3: The proposal distribution function for the output of interest, PFEI, in an aircraft performance example. Note that we show the target distribution function here for comparison purposes. Our L_2–norm optimal empirical importance weights, denoted by the crosses, are determined using the Frank-Wolfe algorithm with step length $\alpha = 2/(2+k)$.

6.3 Uniform Distribution and the L_2–norm Discrepancy

In this example we present the relationship between our proposed approach and discrepancy theory (Dick and Pillichshammer 2010). To illustrate this relationship, the proposal and target distributions are the uniform distribution on the unit hypercube. We also take this opportunity to evaluate the performance of the Frank-Wolfe algorithm and Dai-Fletcher algorithm over a range of parameters.

Substituting the uniform distribution function, $F_\nu(\mathbf{t}) = \prod_{i=1}^{d} t_i$, for the target distribution function in Equation (8), we obtain

$$\tilde{\omega}^2(\hat{\mathbf{w}}) = \frac{1}{2}\Bigg(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{w}_i\hat{w}_j\prod_{k=1}^{d}\left(1-\max(x_k^i,x_k^j)\right) - \frac{2}{n}\sum_{i=1}^{n}\hat{w}_i\prod_{k=1}^{d}\frac{1-(x_k^i)^2}{2} + \frac{1}{3^d}\Bigg), \tag{60}$$

where we use $\tilde{\omega}$ to denote our L_2–norm distance metric in the special case of a uniform target distribution. If the proposal random samples are uniformly weighted (i.e., $\hat{w}_i = 1, \forall i \in \{1,2,\ldots,n\}$), then Equation (60) relates directly to the L_2–norm discrepancy. The L_2–norm discrepancy is defined as

$$D_2 = \sqrt{2}\tilde{\omega}(\mathbf{1}_n), \tag{61}$$

and is sometimes referred to as Warnock's formula (Matoušek 1998, Warnock 1972).

In the following numerical study, we compare the ratio between the weighted L_2–norm discrepancy that results from using (60) with our optimal empirical importance weights and Warnock's formula (61),

$$r = \frac{\sqrt{2}\tilde{\omega}(\hat{\mathbf{w}})}{D_2} = \frac{\tilde{\omega}(\hat{\mathbf{w}})}{\tilde{\omega}(\mathbf{1}_n)}. \tag{62}$$

We investigate two scenarios: proposal random samples drawn from a pseudo-random (PR) sequence and from a randomized Sobol' low discrepancy (i.e., quasi-random, QR) sequence (Niederreiter 1978). A pseudo-random number generator combines randomness from various low-entropy input streams to generate a sequence of outputs that are in practice statistically indistinguishable from a truly random sequence, whereas a quasi-random number generator constructs a sequence of outputs deterministically such that the output obtains a small discrepancy (Caflisch 1998, Niederreiter 1978).

For the case $d = 1$, the analytic empirical importance weights (39) are

$$\hat{\mathbf{w}} = \frac{1}{2}\begin{bmatrix} x^2 + x^1 \\ x^3 - x^1 \\ \cdots \\ x^n - x^{n-2} \\ 2 - x^n - x^{n-1} \end{bmatrix}. \tag{63}$$

Table 5 presents the results for the $d = 1$ case. Shown are the ratios r (in percentages), averaged over 100 independent trials. The results illustrate that the optimal empirical importance weights consistently reduce the L_2–norm discrepancy with respect to the uniformly weighted proposal random samples (i.e., $r < 1$). The reduction is more pronounced for the pseudo-random samples than the quasi-random samples. This is expected because quasi-random samples are constructed to reduce the discrepancy among the samples.

Table 5: The ratio of discrepancy computed using our optimal empirical importance weights and uniform importance weights, Equation (62) measured as a percentage. Shown are results for the $d = 1$ case, averaged over 100 independent trials. The term in parentheses is the corresponding standard deviation. n is the number of proposal random samples.

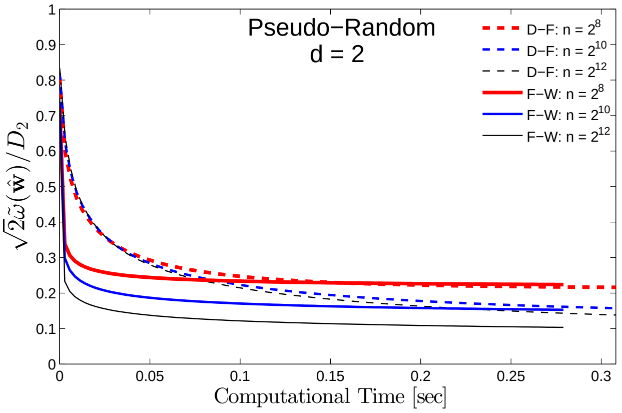
	$n = 2^8$	$n = 2^{10}$	$n = 2^{12}$
PR	12.2(4.80)	6.96(2.45)	3.38(1.17)
QR	86.4(6.48)	86.7(6.10)	85.9(6.86)

Since we have available the analytic representation of the empirical importance weights (63), we can also see that the resulting weighted Monte Carlo integration rule for an integrable function g is

$$\int_0^1 g(t)\ \mathrm{d}t = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\hat{w}_i g(x^i) = \lim_{n\to\infty}\frac{1}{2}\Bigg((x^2+x^1)g(x^1)+\sum_{i=1}^{n}(x^{i+1}-x^{i-1})g(x^i) + (2-x^n-x^{n-1})g(x^n)\Bigg), \qquad (64)$$

which was previously shown to be the trapezoidal integration rule (Yakowitz et al. 1978).

For the general case $d > 1$, the empirical importance weights are computed using the Frank-Wolfe algorithm with an optimal step length α, and the Dai-Fletcher algorithm. For all simulations presented the Dai-Fletcher algorithm computes the matrix H once and stores it. The Frank-Wolfe algorithm using a deterministic step length α halves the computational time compared to using an optimal step length, but leads to poor results early in the optimization process. We selected a maximum number of iterations as the termination criterion for both algorithms. The maximum number of iterations were selected such that both algorithms have similar computational run times.[2] The purpose of this study is to evaluate our proposed approach and to compare the computational performance of the Frank-Wolfe algorithm to the Dai-Fletcher algorithm over a range of parameters. These parameters include the number of proposal random samples n, the initial solution $\mathbf{w}_0$, and dimension d. The initial solution for all simulations is uniform importance weights (i.e., $\mathbf{w}_0 = \mathbf{1}_n$). Figures 4, 5, and 6 show the results averaged over 100 independent trials for $d = 2$, $d = 5$, and $d = 10$, respectively.



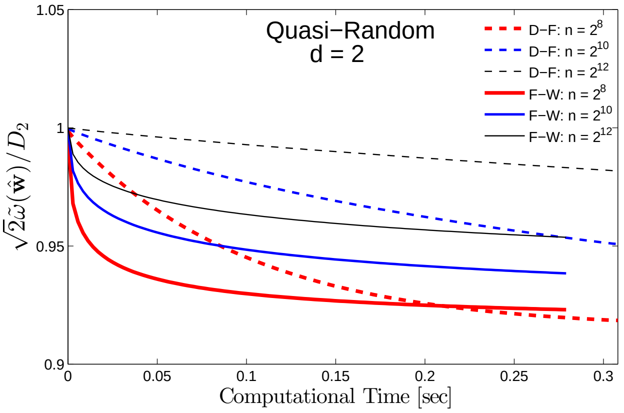


Fig. 4: Discrepancy reduction for $d = 2$. Both algorithms reduce the L_2–norm discrepancy (i.e., $r < 1$) in both scenarios. The Frank-Wolfe algorithm converges more quickly than the Dai-Fletcher algorithm.

As was the case for $d = 1$, these results illustrate that the optimal empirical importance weights consistently reduce the L_2–norm discrepancy with respect to uniformly weighted proposal random samples. Again, the reduction is more pronounced for the pseudo-random samples than the quasi-random samples. In general, if the proposal random samples are drawn from a pseudo-random sequence, then increasing n leads to further decrease in the discrepancy (r decreases further); however, if the proposal random samples are drawn from a quasi-random sequence, then increasing n leads to less discrepancy reduction (r shows less decrease). This can be explained since the pseudo-random proposal samples have poor (high) initial discrepancy and including more

[2] All computations were performed on a dual Intel Core Xeon E5410 CPU desktop computer.

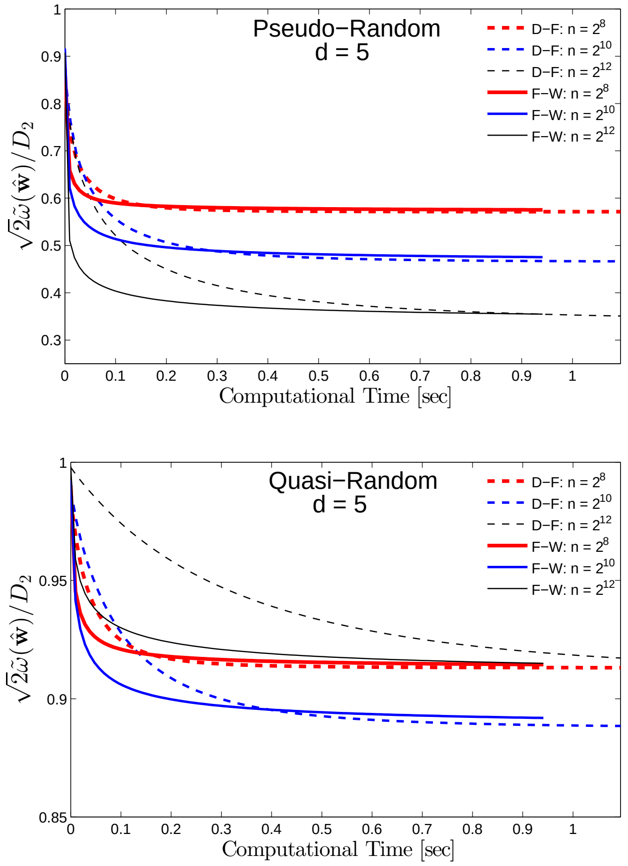


Fig. 5: Discrepancy reduction for $d = 5$. Both algorithms reduce the L_2–norm discrepancy (i.e., $r < 1$) in both scenarios. The Frank-Wolfe algorithm converges more quickly than the Dai-Fletcher algorithm, although the final results are similar.

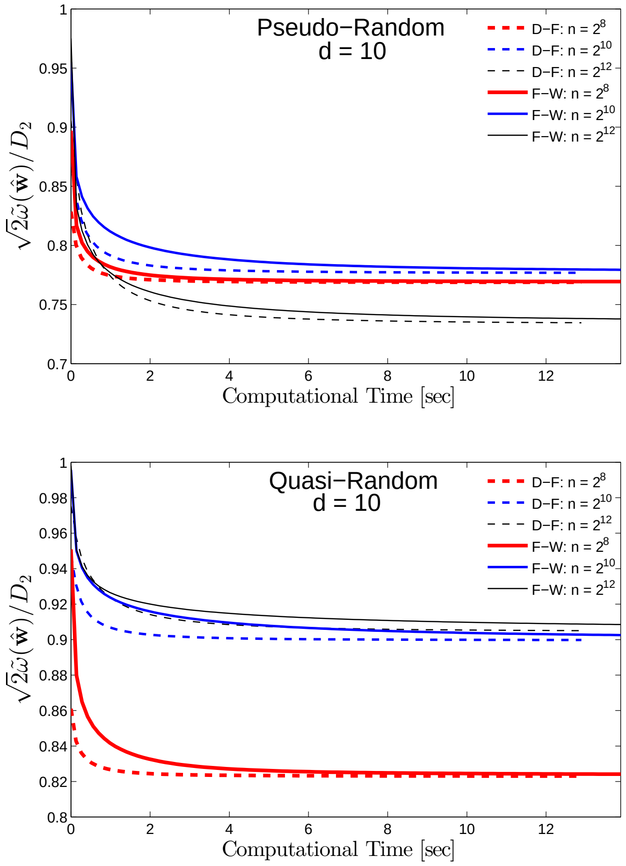


Fig. 6: Discrepancy reduction for $d = 10$. Both algorithms reduce the L_2–norm discrepancy (i.e., $r < 1$) in both scenarios. The Dai-Fletcher algorithm converges more quickly the Frank-Wolfe algorithm, although the final results are similar.

proposal samples gives our approach more degrees of freedom over which to optimize. Conversely, the quasi-random proposal samples already have low discrepancy; including more samples in this case makes it more difficult for the optimization to find a lower-discrepancy solution.

The results generally show that the Frank-Wolfe algorithm converges more quickly for cases using pseudo-random samples, while the Dai-Fletcher exhibits better performance for quasi-random samples. This suggests that the Frank-Wolfe algorithm may be preferred when the initial proposal empirical distribution function is far from the target distribution function, while the Dai-Fletcher algorithm is a better choice when the initial empirical importance weights are already close to optimal. Examining the results with increasing dimension d (i.e., increasing condition number of matrix H (Visick 2000)), illustrates that both algorithms require more computational time to converge. This is expected since both algorithms implement gradient descent techniques whose convergence rates are expected to depend on the condition number of H.

The results presented in Figure 7 demonstrate our approach on a large-scale application problem. In this example we extended the results presented in Figure 5 using the Frank-Wolfe algorithm to proposal sample sizes $n = [8192, 32768, 131072]$. The computational times presented do not include the time required to evaluate the initial gradient (i.e., initial matrix-vector product; $\mathbf{a} = H\hat{\mathbf{w}}$). The results suggest our approach scales well with large number of samples. Numerical strategies such as divide-and-conquering methods and parallelization can be implemented to further improve the computational run times.

From these results, we recommend using the Frank-Wolfe algorithm when the dimension d is small or when the initial proposal empirical distribution function is far from the target distribution function. Otherwise, we recommend the Dai-Fletcher algorithm if the dimension d is large or if the initial proposal empirical distribution

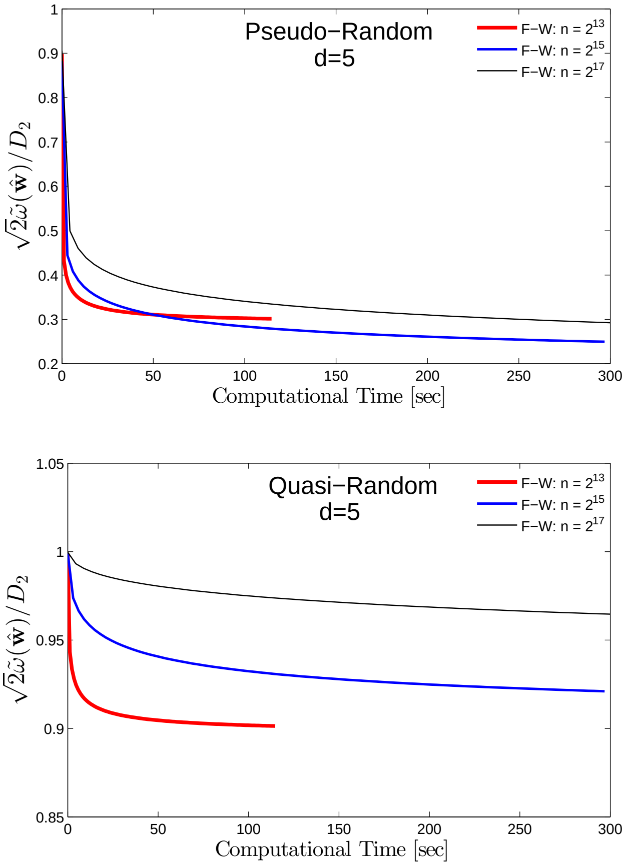


Fig. 7: Discrepancy reduction for $d = 5$ and a large number of samples. The Frank-Wolfe algorithm reduces the L_2–norm discrepancy (i.e., $r < 1$) in both scenarios for a large-scale application problem (i.e., large n). The results presented are the average over 100 simulations.

function is close to the target distribution function. If the number of proposal samples n is so large such that the matrix H cannot be stored, then we recommend using the Frank-Wolfe algorithm since the Dai-Fletcher algorithm will require constructing the matrix H on the fly at each iteration, which will drastically increase computational time.

7 Conclusion

This paper presents a new approach that defines and computes empirical importance weights, and shows its connections to other discrepancy metrics and studies. A key attribute of the approach is its scalability: it lends itself well to handling a large number of samples through a scalable optimization algorithm. The approach also scales to problems with high-dimensional distributions, although numerical challenges will arise due to ill-conditioning of the matrix H. These challenges can be addressed, as they have in other fields such as optimization of systems governed by partial differential equations (Biros and Ghattas 2005), through a combination of preconditioning techniques and use of optimization solvers that are tolerant to ill-conditioned matrices. Future efforts are required to extend the convergence results in the L_1–norm presented here, to demonstrate almost everywhere convergence. Other future directions of interest include exploitation of the optimization solution process to generate sparse solutions, which may yield a way to derive efficient Monte Carlo integration rules that rely on a condensed set of samples (Girolami and He 2003), and exploring different objective function metrics (in particular replacing the L_2–norm metric with an L_1–norm metric).

8 Acknowledgements

This work was supported in part by the International Design Center at the Singapore University of Technology and Design, by the DARPA META program through AFRL Contract FA8650-10-C-7083 and Vanderbilt University Contract VU-DSR#21807-S7, and by the US Federal Aviation Administration Office of Environment and Energy under FAA Award No. 09-C-NE-MIT, Amendment Nos. 028, 033, and 038, managed by Rhett Jefferies and James Skalecky, FAA. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the FAA. The authors thank A. Philpott, R. Freund, and S. Wright for helpful advice regarding the optimization solvers.

References

Adams, M.R., Guillemin, V.: Measure theory and probability. Springer, Boston, Massachusetts (1996)

Amaral, S., Allaire, D., Willcox, K.: A decomposition-based approach to uncertainty analysis of feed-forward multicomponent systems. International Journal for Numerical Methods in Engineering **100**(13), 982–1005 (2014)

Bentley, J.L.: Multidimensional divide-and-conquer. Communications of the ACM **23**(4), 214–229 (1980)

Billingsley, P.: Probability and measure. John Wiley & Sons (2008)

Biros, G., Ghattas, O.: Parallel Lagrange–Newton–Krylov–Schur methods for PDE-constrained optimization. Part I: The Krylov–Schur solver. SIAM Journal on Scientific Computing **27**(2), 687–713 (2005)

Boyd, S.P., Vandenberghe, L.: Convex optimization. Cambridge University Press, Cambridge, UK (2004)

Caflisch, R.: Monte Carlo and quasi-Monte Carlo methods. Acta Numerica **7**, 1–49 (1998)

Dai, Y.H., Fletcher, R.: Projected Barzilai-Borwein methods for large-scale box-constrained quadratic programming. Numerische Mathematik **100**(1), 21–47 (2005)

Dai, Y.H., Fletcher, R.: New algorithms for singly linearly constrained quadratic programs subject to lower and upper bounds. Mathematical Programming **106**(3), 403–421 (2006)

Dick, J., Pillichshammer, F.: Digital nets and sequences: discrepancy theory and quasi-Monte Carlo integration. Cambridge University Press, New York, NY, USA (2010)

Frank, M., Wolfe, P.: An algorithm for quadratic programming. Naval Research Logistics Quarterly **3**(1-2), 95–110 (1956)

Galletly, J.: Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms. Kybernetes **27**(8), 979–980 (1998)

Genz, A.: Testing multidimensional integration routines. Proceedings of International Conference on tools, methods and languages for scientific and engineering computation, 81–94 (1984)

Gibbs, A., Su, F.: On choosing and bounding probability metrics. International statistical review **70**(3), 419–435 (2002)

Girolami, M., He, C.: Probability density estimation from optimally condensed data samples. Pattern Analysis and Machine Intelligence, IEEE Transactions on **25**(10), 1253–1264 (2003)

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., Schölkopf, B.: Covariate shift by kernel mean matching. Dataset Shift in Machine Learning **3**(4), 1–38 (2009)

Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning, vol. 2. Springer (2009)

Heinrich, S.: Efficient algorithms for computing the $_2$-discrepancy. Mathematics of Computation of the American Mathematical Society **65**(216), 1621–1633 (1996)

Huang, J., Smola, A.J., Gretton, A., Borgwardt, K.M., Schölkopf, B.: Correcting sample selection bias by unlabeled data. Advances in Neural Information Processing Systems **19**, 601–608 (2007)

Kanamori, T., Hido, S., Sugiyama, M.: A least-squares approach to direct importance estimation. The Journal of Machine Learning Research **10**, 1391–1445 (2009)

Lawson, C.L., Hanson, R.J.: Solving least squares problems, vol. 161. SIAM (1974)

Lin, C.J., Lucidi, S., Palagi, L., Risi, A., Sciandrone, M.: Decomposition algorithm model for singly linearly-constrained problems subject to lower and upper bounds. Journal of Optimization Theory and Applications **141**(1), 107–126 (2009)

Morokoff, W., Caflisch, R.: quasi-Monte Carlo integration. Journal of Computational Physics **122**(2), 218–230 (1995)

Matoušek, J.: On the L2–discrepancy for anchored boxes. Journal of Complexity **14**(4), 527–556 (1998)

Nickolls, J., Buck, I., Garland, M., Skadron, K.: Scalable parallel programming with CUDA. Queue **6**(2), 40–53 (2008)

Niederreiter, H.: quasi-Monte Carlo methods and pseudo-random numbers. Bulletin of the American Mathematical Society **84**(6), 957–1041 (1978)

Novak, E., Wozniakowski, H.: L_2-Discrepancy and multivariate integration. Analytic Number Theory: Essays in Honour of Klaus Roth pp. 359–388 (2009)

Platt, J.: Fast training of support vector machines using sequential minimal optimization. MIT Press, Cambridge, Massachusetts (1999)

Qin, J.: Inferences for case-control and semiparametric two-sample density ratio models. Biometrika **85**(3), 619–630 (1998)

Robert, C.P., Casella, G.: Monte Carlo Statistical Methods (Springer Texts in Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA (2005)

Rudin, W.: Principles of mathematical analysis, vol. 3. McGraw-Hill New York, New York, NY, USA (1976)

Rudin, W.: Real and complex analysis, 3rd Ed. McGraw-Hill, Inc., New York, NY, USA (1987)

Saltelli, A., Sobol', I. M.: Sensitivity analysis for nonlinear mathematical models numerical experience. Matematicheskoe Modelirovanie **7**(1), 16–28 (1992)

Scott, D.: Multivariate density estimation: theory, practice, and visualization. A Wiley-Interscience publication. Wiley (1992)

Sugiyama, M., Suzuki, T., Kanamori, T.: Density ratio estimation in machine learning. Cambridge University Press, Cambridge, UK (2012)

Sugiyama, M., Yamada, M., Von Buenau, P., Suzuki, T., Kanamori, T., Kawanabe, M.: Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. Neural Networks **24**(2), 183–198 (2011)

Tokdar, S., Kass, R.: Importance sampling: a review. Wiley Interdisciplinary Reviews: Computational Statistics **2**(1), 54–60 (2010)

Vapnik, V., Braga, I., Izmailov, R.: A constructive setting for the problem of density ratio estimation. In:

Proceedings of the 2014 SIAM International Conference on Data Mining, pp. 434–442 (2014)

Vapnik, V.: Statistical learning theory, vol. 2. Wiley New York, New York, NY, USA (1998)

Visick, G.: A quantitative version of the observation that the hadamard product is a principal submatrix of the kronecker product. Linear Algebra and Its Applications **304**(1), 45–68 (2000)

Warnock, T.: Computational investigations of low-discrepancy point sets, applications of number theory to numerical analysis (S. K. Zaremba, ed.). Academic Press (1972)

Yakowitz, S., Krimmel, J., Szidarovszky, F.: Weighted Monte Carlo integration. SIAM Journal on Numerical Analysis **15**(6), 1289–1300 (1978)

Zanni, L.: An improved gradient projection-based decomposition technique for support vector machines. Computational Management Science **3**(2), 131–145 (2006)

Zanni, L., Serafini, T., Zanghirati, G.: Parallel software for training large scale support vector machines on multiprocessor systems. The Journal of Machine Learning Research **7**, 1467–1492 (2006)