On Image Segmentation With Noisy Labels: Characterization and Volume Properties of the Optimal Solutions to Accuracy and Dice

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study two of the most popular performance metrics in medical image segmentation, Accuracy and Dice, when the target labels are noisy. For both metrics, several statements related to characterization and volume properties of the set of optimal segmentations are proved, and associated experiments are provided. Our main insights are: (i) the volume of the solutions to both metrics may deviate significantly from the expected volume of the target, (ii) the volume of a solution to Accuracy is always less than or equal to the volume of a solution to Dice and (iii) the optimal solutions to both of these metrics coincide when the set of feasible segmentations is constrained to the set of segmentations with the volume equal to the expected volume of the target.

1 Introduction

One of the most central problems in medical image analysis is to identify the region of an image associated with a certain target structure. This problem, referred to as image segmentation or delineation, is often very time consuming to solve manually. Consequently, there is great interest in the development of methods that can assist in automation of the procedure. During the last years, it has become increasingly popular to address the segmentation problem using machine learning based methods, and in particular, fully convolutions neural networks with U-net architecture. Such methods commonly dominate the winning submissions to segmentation contests and are backed by a large base of supporting literature [25, 1, 6, 7, 11, 17, 28]

Despite the success, performance of these methods will, like any machine learning method, depend on the quality of the available data [33, 4]. Since it is well known that the data commonly used in practice is produced by medical practitioners that delineate structures in an inconsistent manner [22], it is important to understand the impact of label noise. One way to study the influence of label noise is to consider how the noise impacts segmentations that are theoretically optimal with respect to the metric used for measuring performance. Even if theoretically optimal solutions may not be attainable in practice, studying them gives important insights into what the effect label noise has on the particular metric.

Arguably, the most simple and classical choice of metric is Accuracy; the fraction of the image that is correctly delineated. However, since Accuracy may not reflect the desired behaviour when the data is unbalanced, that is, when the target structure is much smaller than the background, alternative metrics are often preferred. The most popular such alternative is the Sørensen-Dice coefficient, or Dice for short and is related to the F_1-metric used in binary classification. Other examples considered in the literature include the Jaccard index and variations of the Haussdorff metric [29].

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

In this work, we conduct a theoretical investigation of the effect label noise has on optimal segmentations with respect to the performance metrics Accuracy and Dice. Because the volume of a proposed segmentation may be used for important properties such as estimating the size of a tumor [8], we pay special attention to the effect noise may have on the volume of the optimal segmentations.

Contributions: A characterization of all of optimal solutions to Accuracy and Dice when the target is noisy is provided. This characterization is used to analyze the volume of the optimal solutions and we prove: (i) sharp upper and lower bounds on the volume of optimal segmentations with respect to Accuracy and Dice, (ii) that the volume of the optimal solutions to Accuracy always is less or equal to the volume of the optimal solutions to Dice and (iii) that the optimal solutions to both metrics coincide when the volume is held fix. We also show the relevance of the problem in a practical setting by including experiments on data from the Gold Atlas project [22].

2 Related work

Deep learning methods is playing an increasingly vital part in the development of medical image analysis. However, deep learning models require large annotated data sets for successful training that rarely are available in the clinics. Even if some data sets are being curated for training deep learning models, it is generally difficult and expensive to accurately annotate large collections of medical images. Moreover, the data sets are often small from the outset, due to privacy issues and institutional policies, which makes it difficult to share data across borders. Moreover, training data may include corrupted or noisy labels. This is particularly the case in image segmentation where different annotators may have different views on the correct delineation of a region of interest, leading to uncertainty about the true label, see e.g. [3, 20]. Noisy labels may also appear due to automated systems or non-expert systems being used to annotate large volumes of data, see [24, 5].

There is a large body of literature on the impact of label noise in image segmentation, see [30] and [12] for recent reviews. The proposed solutions to limit the loss of performance when the labels are noisy include label cleaning and pre-processing, e.g. [9], modification of network architechtures, e.g. [32], robustification of loss functions, e.g. [18], reweighting of training data, see [35, 19], and many others. These approaches are of practical nature and generally address methodology that improve the performance on some chosen noisy data set. On the contrary, the literature that address the effect of label noise from a theoretical point of view in the context of image segmentation is rather limited. It was shown that the loss function soft-Dice, in contrast to the loss function cross-entropy, does not yield optimal predictions that coincide with the pixel-wise marginals and that the associated volume is biased [1, 2]. This motivated methods for post-calibrating uncalibrated marginal estimates [26] and a more general investigation of the relationship between volume and marginal calibration [23].

Another domain of related work can be found in the binary classification literature. From a technical point of view, there is a connection between the study of solutions to Accuracy and Dice in segmentation and the study of Accuracy and F_1 in binary classification [21]. Of importance to us is *optimal plug-in* classifiers to Accuracy and F_1. That is, classifiers that are obtained by processing the posterior class probabilities or estimates thereof using a threshold. For Accuracy, this relates to the classical Bayes classifier which has been studied since the origin of the field [31]. Early work on the existence of such a classifier for F_1, and the fact that this threshold is lower than the threshold for Accuracy can be traced to [34]. Further work showed this threshold to be equal to half of the maximal attainable F_1-score [16]. Lots of extensions of these works have been proposed, but to the best of our knowledge no such extension is in the direction of our work.

3 Preliminaries

In our work we find that it is convenient to do the theoretical analysis over a continuous domain. Formally, let $\Omega = [0,1]^n \subset \mathbb{R}^n$ be the unit cube of dimension $n \geq 1$ and λ be the associated standard normalized Lebesgue measure such that $\lambda(\Omega) = 1$. The space of segmentations is denoted by $\mathcal{S}$ and is the space of measurable functions from Ω to the binary numbers $\{0,1\}$. For any segmentation $s \in \mathcal{S}$, $s(w) = 1$ implies that the object of interest occupies the site $\omega \in \Omega$. Note that any segmentation on a discretized domain can be incorporated using this setup by appropriate normalization. Details on this can be found in the end of Section 4.

Beyond the space of segmentations, several other technical constructions are introduced. This includes the space of measurable functions from Ω to $[0,1]$ which we denote by $\mathcal{M}$ and refer to as the *marginal functions*. We also let $|f| \doteq \int_\Omega |f(\omega)|\lambda(d\omega)$ be the volume of f and $\bar{f} \doteq 1 - f$, where f is any measurable function defined on Ω. Throughout we adopt the convention that two λ-measurable functions f, g are equal if they are equal λ-a.e. We will use $I(\cdot)$ to denote the identity function, and when F is a cumulative distribution function, we will denote the left limit $F(t-) = \lim_{s\uparrow t} F(s)$. Finally, for a given volume $v \in [0,1]$, we let $\mathcal{S}_v \doteq \{s \in \mathcal{S}, |s| = v\}$ be the set of segmentations with volume v.

Classically, metrics in medical image segmentation are defined per image as functionals over two deterministic segmentations [29]. When noise is present, the label becomes a random variable and the metrics need to be extended to a functional over one deterministic segmentation and one random label segmentation. In this work, the soft labeling convention for this extension is adopted [13, 10, 27, 14, 15].

Definition 1. *For any $m \in \mathcal{M}$, Accuracy is given by*

$$\mathrm{A}_m(s) \doteq \int_\Omega [s(\omega)m(\omega) + \bar{s}(\omega)\bar{m}(\omega)]\lambda(d\omega), \quad s \in \mathcal{S} \tag{1}$$

Definition 2. *For any $m \in \mathcal{M}$, Dice is given by*

$$\mathrm{D}_m(s) \doteq \frac{2\int_\Omega s(\omega)m(\omega)\lambda(d\omega)}{|s| + |m|}, \quad s \in \mathcal{S}. \tag{2}$$

For a noisy segmentation L, that is, a random variable taking values in $\mathcal{S}$, m can be taken to be the exact marginal success probability $m(\omega) = \mathbb{E}[L(\omega)],\ \omega \in \Omega$. Such marginal functions are important in theory but can seldom be obtained in practice. Alternative choices of marginal functions include finite sample approximations, that is, point-wise averages over finite observations of L, and estimates of m according to a single annotator [13]. These choices of m are important because they are sometimes used for training machine learning models. Finally, also the predictions of machine learning models can be used as marginal functions.

Because of the fact that $|m| = \mathbb{E}[|L|]$ when m is taken to be the exact marginal success probability, $|m|$ plays an important role in the medical segmentation context, either theoretically as the expected volume of the target or as an approximation thereof. Understanding how $|m|$ relates to $|s|$, where s is an optimal segmentation to Accuracy or Dice will be central in our work. To the best of our knowledge, this has not been studied in prior work.

4 Main results

The objective of our analysis is to characterize the optimizers to A_m and D_m and give a detailed description of the volume of the optimal segmentations. That is, for a given $m \in \mathcal{M}$, identify properties (e.g. volume) of the optimal segmentation $s \in \mathcal{S}$ that maximimize Accuracy or Dice. To this end, consider the probability measure on $[0,1]$ given by the push-forward measure $\lambda \circ \bar{m}^{-1}(\cdot)$ and let F_m denote its cumulative distribution function,

$$F_m(t) \doteq \lambda \circ \bar{m}^{-1}([0,t]) = \int_\Omega I\{\bar{m}(\omega) \leq t\}\lambda(d\omega), \quad t \in [0,1]. \tag{3}$$

The function $F_m(t)$ may be interpreted as the volume of the set of sites with non-success probability less than or equal to t. In other words, the volume of the sub-level set of $\bar{m}$ at level t. Since F_m is the cumulative distribution function of a probability distribution on $[0,1]$, it has several well-known properties making it easy to work with, e.g., F_m is non-decreasing and right-continuous with $F_m(0) = 0$ and $F_m(1) = 1$. Of particular interest to us is that it has a generalized inverse given by

$$F_m^{-1}(v) \doteq \inf\{t : F_m(t) \geq v\}, \quad v \in [0,1], \tag{4}$$

which can be interpreted as the the minimum level at which the volume of the corresponding sub-level set of $\bar{m}$, is at least v. This function is often referred to as the quantile function and also has several well known properties; it is non-decreasing and left-continuous. Moreover, it allows us to define the

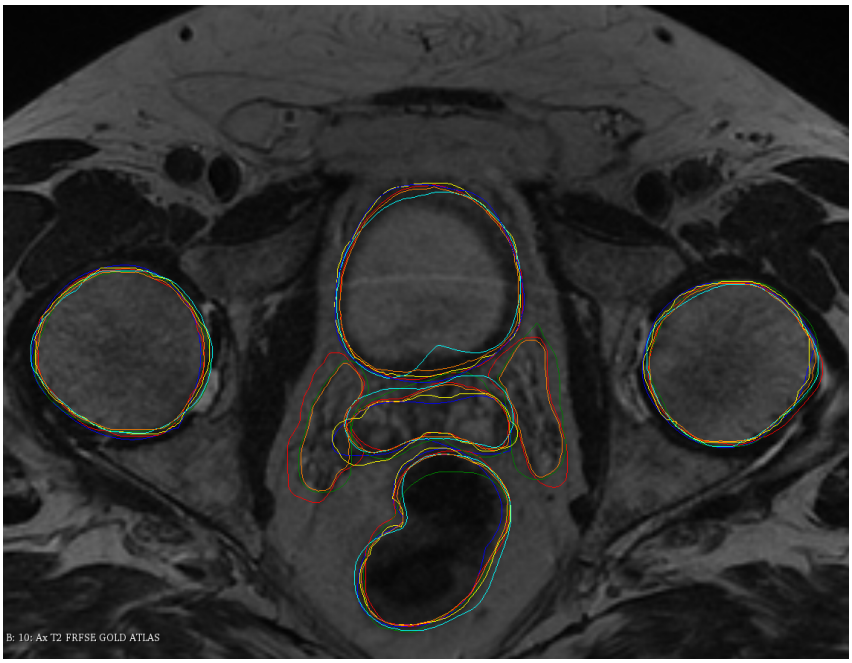

Figure 1: Contours a on a slice for a patient in the Gold Atlas project [22]. Each color is associated with a particular annotator.

following important class of segmentations for a given $m \in \mathcal{M}$.

$$\mathcal{S}_{m,v} \doteq \left\{ s \in \mathcal{S}_v \;\middle|\; \begin{array}{l} \int_\Omega s(\omega) I\{m(\omega) < 1 - F_m^{-1}(v)\}\lambda(d\omega) = 0, \\ \int_\Omega \bar{s}(\omega) I\{m(\omega) > 1 - F_m^{-1}(v)\}\lambda(d\omega) = 0, \end{array} \right\}, \quad v \in [0,1]. \tag{5}$$

The described class $\mathcal{S}_{m,v}$ is informally the set of segmentations with volume v that assigns 1 to sites ω where $m(\omega)$ is large. If $t = F_m^{-1}(v)$ is a continuity point of F_m, i.e. $\lambda(\omega : \bar{m}(\omega) = t) = 0$, then $\mathcal{S}_{m,v}$ only consist of the elements that are λ-a.e. equal to the segmentation $s(\omega) = I\{m(\omega) \geq 1 - F_m^{-1}(v)\}$. A lot of our analysis can be simplified if F_m is assumed to be invertible almost everywhere. However, this would require m to not have any non-neglible constant regions, which for instance excludes any m that is given by an empirical approximations using a finite number of samples. Consequently, in the sequel, we treat general F_m. Our first result contains the essential ingredients for characterizing the optimizers to Accuracy and Dice.

Lemma 1. *For any $m \in \mathcal{M}$ and $v \in [0,1]$*

$$\sup_{s \in \mathcal{S}_v} \int_\Omega s(\omega) m(\omega) \lambda(d\omega) = \int_0^v (1 - F_m^{-1}(u)) du, \tag{6}$$

and the elements where the supremum is attained is given by $\mathcal{S}_{m,v}$.

A complete proof is given in the Supplementary Material and outlined as follows. The first part shows that the class $\mathcal{S}_{m,v}$ is the class of optimal solutions by showing that for any $s^* \in \mathcal{S}_{m,v}$ and $s \in \mathcal{S}_v \setminus \mathcal{S}_{m,v}$, $\int_\Omega (s^*(\omega) - s(\omega)) m(\omega) \lambda(d\omega) > 0$. The second part proves the equality (6) using an application of the quantile transform. That is, if U has uniform distribution on $[0,1]$ then $F_m^{-1}(U)$ has cdf given by F_m and $\int_0^1 t F_m(dt) = \mathbb{E}[F_m^{-1}(U)] = \int_0^1 F_m^{-1}(u) du$.

Lemma 1 allows us to reduce the constrained optimization problem over the rather complicated space of segmentations, to a one-dimensional integral with respect to the quantile function. It is the starting point for our analysis.

In the remaining section our main theoretical results are presented. In Theorem 1, we provide a characterization of all of the optimal solutions to Accuracy based on volume. In addition, sharp upper and lower bounds on the volume of the associated optimal segmentations are provided.

Theorem 1. *For any $m \in \mathcal{M}$, the class of maximizers to A_m is given by $\cup_{v \in \mathcal{V}^{\mathrm{A}_m}} \mathcal{S}_{m,v}$ where*

$$\mathcal{V}^{\mathrm{A}_m} \doteq [F_m(1/2-), F_m(1/2)]. \tag{7}$$

Moreover, $\mathcal{V}^{\mathrm{A}_m}$ satisfies the following bounds

$$\mathcal{V}^{\mathrm{A}_m} \subseteq [\max\{2|m|-1, 0\}, \min\{2|m|, 1\}], \tag{8}$$

where the bounds are sharp in the sense that there for any $v \in [0,1]$ exist $m_0, m_1 \in \mathcal{M}$ such that $|m_0| = |m_1| = v$ and

$$\inf \mathcal{V}^{\mathrm{A}_{m_0}} = \max\{2|m_0|-1, 0\}, \quad \sup \mathcal{V}^{\mathrm{A}_{m_1}} = \min\{2|m_1|, 1\}. \tag{9}$$

The complete proof is given in the Supplementary Material and is outlined as follows. First, the function

$$a_m(v) \doteq v + 1 - |m| - 2\int_0^v F_m^{-1}(u)du, \quad v \in [0,1], \tag{10}$$

is introduced and then Lemma 1 is used to show that $\sup_{s \in \mathcal{S}} \mathrm{A}_m(s) = \sup_{v \in [0,1]} a_m(v)$. Consequently, the class of optimal solutions to A_m is given by $\cup_{v \in \mathcal{V}^{\mathrm{A}_m}} \mathcal{S}_{m,v}$, where $\mathcal{V}^{\mathrm{A}_m}$ is the set of optimizers to a_m. The rest of the proof consists of detailed analysis of a_m and is composed of three parts. The first part is to show (7) by finding one optimal solution and then identifying all volumes that yield the same optimal value. The second part is to provide the lower and upper bounds on the elements of $\mathcal{V}^{\mathrm{A}_m}$ in terms of $|m|$ given by (8). The third part is to provide examples of situations where the extreme cases occur 9. In Figure 2, a case that is extreme both in the lower and in the upper sense is illustrated.

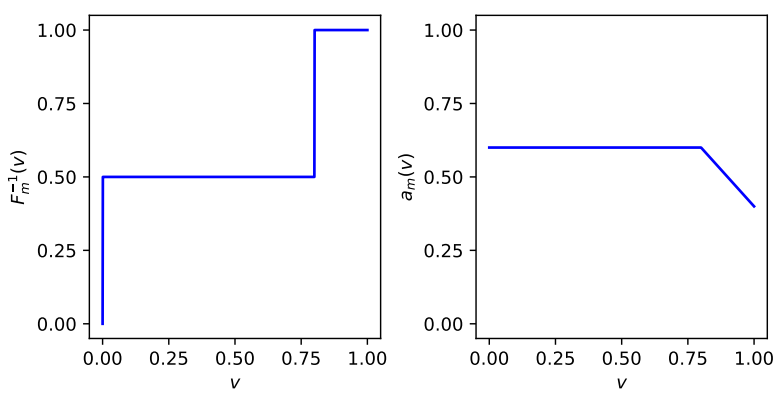

Figure 2: To the left is a particular quantile function F_m^{-1} and to the right is the associated function a_m given by (10). Here $F_m^{-1}(v) = \frac{1}{2} I_{(0,2|m|]}(v) + I_{(2|m|,1]}(v), v \in [0,1]$ with $|m| = 0.4$ that satisfies $\mathcal{V}^{\mathrm{A}_m} = [0, 2|m|] = [\max\{2|m|-1, 0\}, \min\{2|m|, 1\}]$ and consequently is an extreme case to (9) in both the lower and the upper sense.

In Theorem 2, we provide a characterization of all of the optimal solutions to Dice based on volume. In addition, sharp upper and lower bounds on the volume of the associated optimal segmentations are provided.

Theorem 2. *For any $m \in \mathcal{M}$, the class of maximizers to D_m is given by $\cup_{v \in \mathcal{V}^{\mathrm{D}_m}} \mathcal{S}_{m,v}$ where*

$$\mathcal{V}^{\mathrm{D}_m} \doteq [F_m((1 - \sup_{s \in \mathcal{S}} \mathrm{D}_m(s)/2)-), F_m(1 - \sup_{s \in \mathcal{S}} \mathrm{D}_m(s)/2)]. \tag{11}$$

Moreover, $\mathcal{V}^{\mathrm{D}_m}$ satisfies the following bounds

$$\mathcal{V}^{\mathrm{D}_m} \subseteq [|m^2|, 1], \tag{12}$$

where the bounds are sharp in the sense that there for any $v \in (0,1]$ exist $m_0, m_1 \in \mathcal{M}$ such that $|m_0| = |m_1| = v$ and

$$\inf \mathcal{V}^{\mathrm{D}_{m_0}} = |m|^2, \quad \sup \mathcal{V}^{\mathrm{D}_{m_1}} = 1. \tag{13}$$

The complete proof is given in the Supplementary Material and is outlined as follows. First, the function

$$d_m(v) \doteq \frac{2\int_0^v (1 - F^{-1}(u))du}{|m| + v}, \quad v \in [0, 1], \tag{14}$$

is introduced and then Lemma 1 is used to show that $\sup_{s\in\mathcal{S}} \mathrm{D}_m(s) = \sup_{v\in[0,1]} d_m(v)$. Consequently, the class of optimal solutions to D_m is given by $\cup_{v\in\mathcal{V}^{\mathrm{D}_m}} \mathcal{S}_{m,v}$, where $\mathcal{V}^{D_m}$ is the set of optimizers to d_m. The remaining proof consists of detailed analysis of d_m and composed of three parts. The first part is to show (11) which is derived by careful investigation of the properties of the function $\delta(v) = \frac{(|m|+v)^2}{2}\partial_v d_m(v)$, which has the same sign as $\partial_v d_m(v)$ and therefore can be used to identify optimal values of d_m. The second part is to provide the lower and upper bounds on the elements of $\mathcal{V}^{\mathrm{D}_m}$ in terms of $|m|$ given by (12). The third part is to provide examples of situations where the extreme values occur (13). In Figure 3, a case that is extreme both in the lower and in the upper sense is illustrated.

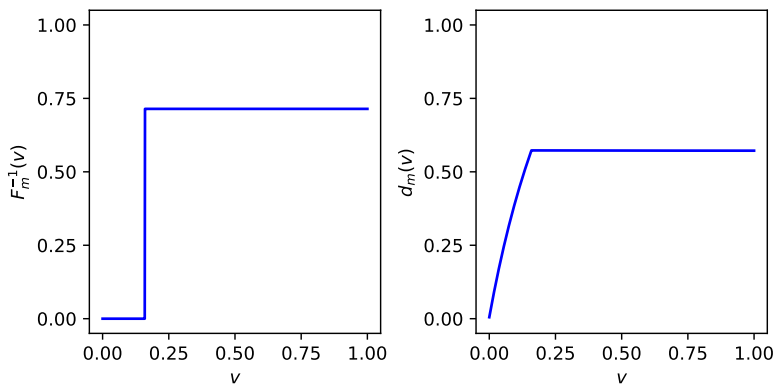

Figure 3: To the left is a particular quantile function F_m^{-1} and to the right is the associated function d_m given by (14). Here $F_m^{-1}(v) = (1 - |m|)(1 - |m|^2)^{-1} I_{(|m|^2,1]}(v)$ with $|m| = 0.4$ that satisfies $\mathcal{V}^{\mathrm{D}_m} = [|m|^2, 1]$ and consequently is an extreme case to (13) in both the lower and the upper sense.

In Theorem 3, we relate the volume of the optimal segmentations of Accuracy and the optimal segmentations of Dice for a given marginal probability $m \in \mathcal{M}$.

Theorem 3. *For any $m \in \mathcal{M}$, $\mathcal{V}^{\mathrm{A}_m}$ given by* (7) *and $\mathcal{V}^{\mathrm{D}_m}$ given by* (11) *satisfy*

$$\sup \mathcal{V}^{\mathrm{A}_m} \leq \inf \mathcal{V}^{\mathrm{D}_m}. \tag{15}$$

The complete proof is given in the Supplementary Material and is outlined as follows. First note that $D_m(s) \leq 1$ for any $s \in \mathcal{S}$ and then consider separately the cases when $D_m(s) < 1$ for all $s \in \mathcal{S}$ and when there exist some $s \in \mathcal{S}$ such that $D_m(s) = 1$. For the first case, it is obvious that $\sup_{s\in\mathcal{S}} \mathrm{D}_m(s)/2 < 1/2$ which implies that $F_m(1/2) \leq F_m((1 - \sup_{s\in\mathcal{S}} \mathrm{D}_m(s)/2)-)$. For the second case, we show that the volume of the optimizers are uniquely given by $\mathcal{V}^{\mathrm{A}_m} = \mathcal{V}^{\mathrm{D}_m} = \{F_m(1/2)\}$. In either case, (15) holds.

In Theorem 4, the set of optimal solutions to Accuracy and Dice when constrained to a specific volume is shown to coincide.

Theorem 4. *For any $m \in \mathcal{M}$ and $v \in [0, 1]$ the maximizers to the problems,*

$$\sup_{s\in\mathcal{S}_v} \mathrm{A}_m(s) \quad \textit{and} \quad \sup_{s\in\mathcal{S}_v} \mathrm{D}_m(s), \tag{16}$$

coincide and are given by $\mathcal{S}_{m,v}$.

The complete proof is given in the Supplementary Material and is a straightforward application of Lemma 1 Of particular interest is the case when $v = |m|$, since this correspond to the situation when the metrics are maximized under the constraint that there should be no volume bias.

It follows from Theorem 1 and Theorem 2 that the optimizers to both Accuracy and Dice are of the form $\cup_{v\in\mathcal{V}}\mathcal{S}_{m,v}$, where $\mathcal{V} = [F_m(t-), F_m(t)]$ for some $t \in [0,1]$. This type of charecterization is practical for proving properties on volume, but inconvenient for other tasks. In Theorem 5, we provide an alternative charecterization using threshold segmentations of the form $s(\omega) = I\{m(\omega) > \alpha\}$ or $s(\omega) = I\{m(\omega) \geq \alpha\}$, for some α. Even if there exist optimal segmentations that are not necessarily of threshold type, they can always be bounded, above and below, by optimal segmentations of threshold type.

Theorem 5. *For any $m \in \mathcal{M}$ and $t \in (0,1]$, let $s_1(\omega) = I\{m(\omega) \geq 1-t\}$ and $s_0(\omega) = I\{m(\omega) > 1-t\}$. Then, $|s_0| = F(t-)$, $|s_1| = F(t)$ and for each $s \in \mathcal{S}_{m,v}$, $v \in [F(t-), F(t)]$,*

$$s_0(\omega) \leq s(\omega) \leq s_1(\omega), \quad \lambda - a.e.$$

The complete proof is given in the Supplementary Material and is outlined as follows. First note that $|s_0| = \int_\Omega I\{\bar{m}(\omega) < t\}\lambda(d\omega) = F_m(t-)$ and $|s_1| = \int_\Omega I\{\bar{m}(\omega) \leq t\}\lambda(d\omega) = F_m(t)$. For the upper bound, with $A = \{\omega : s(\omega) > s_1(\omega)\}$, we first observe that $I\{\omega \in A\} = s(\omega)\bar{s}_1(\omega)$ and then, using the definition of $\mathcal{S}_{m,v}$ we prove that

$$\lambda(A) = \int_\Omega s(\omega)\bar{s}_1(\omega)\lambda(d\omega) \leq \int_\Omega s(\omega) I\{\bar{m}(\omega) > F_m^{-1}(v)\}\lambda(d\omega) = 0.$$

The lower bound is similar, but slightly more involved.

In numerical applications, the continuum Ω is usually partitioned into a finite collection of voxels $\{\Omega_i\}_{i\in\mathcal{I}}$. Marginal functions are then constrained to the subset of $\mathcal{M}$ that is compatible with the voxelization in the sense that m is measurable with respect to the σ-field generated by the partition. Note that if $s_1(\omega) = I\{m(\omega) \geq 1-t\}$ and $s_0(\omega) = I\{m(\omega) > 1-t\}$ for some $t \in (0,1]$, then also s_0 and s_1 are compatible with the voxelization. By Theorem 1 (Theorem 2) and Theorem 5, the segmentations with least and greatest volume that are optimal with respect to Accuracy (Dice) are compatible with the voxelization, and, consequently, it is not possible to obtain tighter bounds than (8) by restricting the class of segmentations to the class of segmentations compatible with the voxelization.

For Accuracy and Dice respectively, we denote the segmentations with the greatest volume by:

$$s^{\mathrm{A}_m}(\omega) \doteq I\{m(\omega) \geq 1/2\}, \quad \omega \in \Omega, \tag{17}$$

$$s^{\mathrm{D}_m}(\omega) \doteq I\{m(\omega) \geq \sup_{s\in\mathcal{S}} \mathrm{D}_m(s)/2\}, \quad \omega \in \Omega. \tag{18}$$

Note that s^{A_m} is analogous to the Bayes classifier and s^{D_m} is analogous to the threshold classifier described in [16]. Both of these are trivial to compute from a given marginal function m compatible to some voxelization.

5 Experiments

The sharp bounds on volume in Thereom 1 and Theorem 2 implies that there exist marginal functions for which the volume of the optimal segmentations to Accuracy and Dice deviate significantly from the expected volume of the target. In this section we conduct experiments on marginal functions formed from real world data to compare the volume of optimal segmentations to the expected volume in practice.

The data set we investigate contains segmentations in the pelvic area and is part of the Gold Atlas project [22]. It is freely availible for reserch and educational purposes. More specifically, the data is in 3D with a resolution of 512×512 pixels per slice and consist of 18 patients with 9 different ROI's (region of interest), each of which have been delineated by several medical practitioners (see Figure 1 for an illustration of the segmentations associated with one patient and ROI). For each ROI in every patient, a marginal function m is formed by taking the pixel-wise average of the different segmentations, that is, a finite sample approximation is considered. The resulting marginal functions are then used compute the segmentations s^{A_m} (17) and s^{D_m} (18). Details on the implementation and how to obtain the data is available in the Supplementary Material.

Results of the experiments are illustrated in Table 1 and Figure 4. The quantities of interest are $|s^{\mathrm{A}_m}|/|m|$ and $|s^{\mathrm{D}_m}|/|m|$, which in a relative sense describe how much the volume of the computed

Table 1: Results of our experiments on the pelvic data in the Gold Atlas project [22]. For each ROI and patient, the associated m is formed (finite sample approximation of the segmentations obtained from the contributing annotators) and used to compute s^{A_m} as defined by (17) and s^{D_m} as defined by (18). This has then been used to generate statistical quantities related to $|s^{\mathrm{A}_m}|/|m|$ and $|s^{\mathrm{D}_m}|/|m|$ aggregated over the different patients.

	$\lvert s^{\mathrm{A}_m}\rvert/\lvert m\rvert$				$\lvert s^{\mathrm{D}_m}\rvert/\lvert m\rvert$			
ROI	Mean	Std	Min	Max	Mean	Std	Min	Max
Urinary bladder	1.004	0.010	0.991	1.035	1.004	0.010	0.991	1.035
Rectum	0.993	0.048	0.912	**1.094**	0.993	0.048	0.912	1.094
Anal canal	0.915	0.070	0.753	1.067	1.082	0.185	0.876	**1.560**
Penile bulb	0.924	0.070	0.696	1.004	1.068	0.161	0.863	1.365
Neurovascular bundles	0.771	0.109	**0.461**	0.928	1.267	0.118	1.070	1.481
Femoral head R	0.989	0.009	0.968	1.006	0.989	0.009	0.968	1.006
Femoral head L	0.996	0.022	0.958	1.070	0.996	0.022	0.958	1.070
Prostate	0.977	0.027	0.894	1.011	0.977	0.027	0.894	1.011
Seminal vesicles	0.904	0.087	0.669	1.028	1.017	0.135	**0.855**	1.325

optimal segmentations with respect to Accuracy and Dice deviate from the expected volume of the target. In Table 1, aggregated statistics of these quantities with respect to all patients are shown and in Figure 4, these quantities are illustrated for each patient and ROI in scatter plots. We make three observations from the results. Firstly, there exist a marginal function m for which $|s^{\mathrm{A}_m}|/|m| < 1/2$ and a marginal function m for which $|s^{\mathrm{D}_m}|/|m| > 1.5$. This confirms that the volume of optimal segmentations to Accuracy and Dice in practice may indeed deviate significantly from the expected volume of the target. Secondly, on average there seems to be a tendency that $|s^{\mathrm{A}_m}|/|m| \leq 1$ and $|s^{\mathrm{D}_m}|/|m| \geq 1$. An interesting question is if this holds exactly for some important class of marginals. Thirdly, there seems to be a relationship between the relative volume bias and the shape of the ROI, where large structures that are more *spherical* like the Urinary bladder are less affected than small structures that are less *spherical* like the Neurovascular bundles. This could be connected to the fact that larger more *spherical* structures have greater volume in relation to the boundary than smaller less *spherical* structures, and that noise primarily exist on the boundary.

6 Conclusion

In this work, we have theoretically investigated the optimal segmentations with respect to the performance metrics Accuracy and Dice. We have given a detailed rigorous characterization of the optimizers and upper and lower bounds on the volume of optimal segmentations. Finally, we have shown the relevance of our theoretical observations in practice by comparing the volume of optimal segmentations with respect to the performance metrics to the expected volume, on a real world data set. We conclude that the noise has a significant effect on the volume of optimal segmentations, and that the effect may be related to the shape of the target structure.

Broader impacts: Formalizing the evaluation process of automated segmentation methods can be done in many ways, each with its pros and cons. Even if this work can be interpreted as describing the problems with using Dice for this formalization, it can still paradoxically contribute to an unhealthy fixation of Dice as the gold standard for segmentation evaluation in medical image analysis. This in turn can lead to that medical practitioners put two much faith in models that have been shown to perform well with respect to the metric on some test data. One solution to this is to make sure that clinical practitioners using such models are educated in the problems associated with the metric.

Limitations: In order for the volume bounds to be sharp in Theorem 1 and Theorem 2, we construct extreme cases of m. These extreme cases might only be representable approximately with step functions for a particular choice of voxelization. Consequently, the most extreme cases we can construct in a numerical setting may not be as extreme as those we have constructed in the continuous setting. However, in medical image analysis it is common to deal voxelizations of the order of $512 \times 512 \times 100$ voxels which means that the approximation error would be negligable. Our

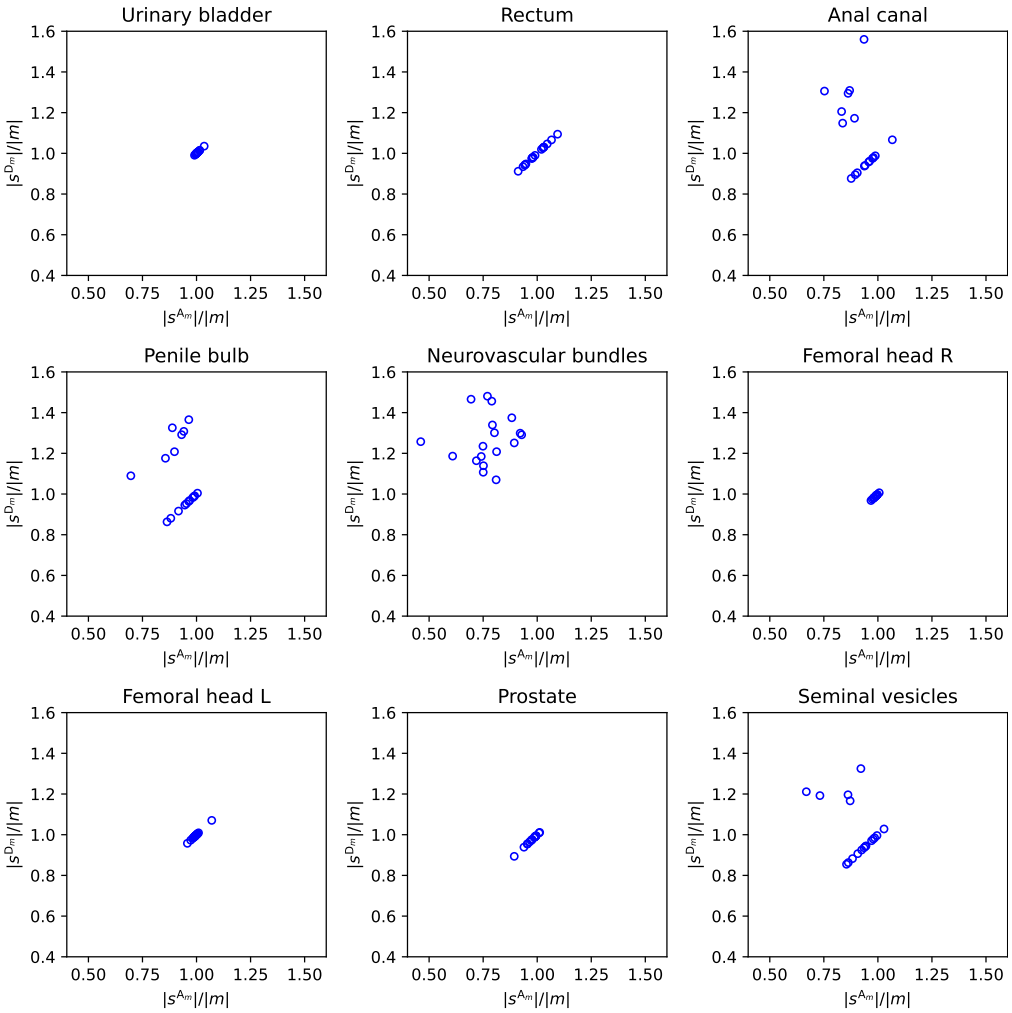

Figure 4: Results of our experiments on the pelvic data in the Gold Atlas project [22]. For each ROI and patient, the associated m is formed (finite sample approximation of the segmentations obtained from the contributing annotators) and used to compute s^{A_m} as defined by (17) and s^{D_m} as defined by (18). These are then plotted with $|s^{\mathrm{A}_m}|/|m|$ on the horizontal axis and $|s^{\mathrm{D}_m}|/|m|$ over on the vertical axis with one subplot per ROI.

work is also limited by the amount of experiments included. Additional numerical experiments on a wider range of data sets would give a more comprehensive picture on the impact of different noise distributions.

References

[1] Jeroen Bertels, Tom Eelbode, Maxim Berman, Dirk Vandermeulen, Frederik Maes, Raf Bisschops, and Matthew B Blaschko. Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 92–100. Springer, 2019.

[2] Jeroen Bertels, David Robben, Dirk Vandermeulen, and Paul Suetens. Theoretical analysis and experimental validation of volume bias of soft dice optimized segmentation maps in the context of inherent uncertainty. *Medical Image Analysis*, 67:101833, 2021.

[3] Pete Bridge, Andrew Fielding, Pamela Rowntree, and Andrew Pullar. Intraobserver variability: should we worry? *Journal of medical imaging and radiation sciences*, 47(3):217–220, 2016.

[4] Weicheng Chi, Lin Ma, Junjie Wu, Mingli Chen, Weiguo Lu, and Xuejun Gu. Deep learning-based medical image segmentation with limited labels. *Physics in Medicine & Biology*, 65(23): 235001, 2020.

[5] Florent Chiaroni, Mohamed-Cherif Rahal, Nicolas Hueber, and Frédéric Dufaux. Hallucinating a cleanly labeled augmented dataset from a noisy labeled dataset using gan. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3616–3620. IEEE, 2019.

[6] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[7] Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. The importance of skip connections in biomedical image segmentation. In *Deep learning and data labeling for medical applications*, pages 179–187. Springer, 2016.

[8] Hans-Hermann Dubben, Howard D Thames, and Hans-Peter Beck-Bornholdt. Tumor volume: a basic and specific response predictor in radiotherapy. *Radiotherapy and oncology*, 47(2): 167–174, 1998.

[9] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017.

[10] Charley Gros, Andreanne Lemay, and Julien Cohen-Adad. Softseg: Advantages of soft versus binary training for image segmentation. *Medical image analysis*, 71:102038, 2021.

[11] Zeyu Jiang, Changxing Ding, Minfeng Liu, and Dacheng Tao. Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In *International MICCAI brainlesion workshop*, pages 231–241. Springer, 2019.

[12] Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020.

[13] Eytan Kats, Jacob Goldberger, and Hayit Greenspan. Soft labeling by distilling anatomical knowledge for improved ms lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1563–1566. IEEE, 2019.

[14] Andreanne Lemay, Charley Gros, and Julien Cohen-Adad. Label fusion and training methods for reliable representation of inter-rater uncertainty. *arXiv preprint arXiv:2202.07550*, 2022.

[15] Hang Li, Dong Wei, Shilei Cao, Kai Ma, Liansheng Wang, and Yefeng Zheng. Superpixel-guided label softening for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 227–237. Springer, 2020.

[16] Zachary C Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. Optimal thresholding of classifiers to maximize f1 measure. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 225–239. Springer, 2014.

[17] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88, 2017.

[18] Damian J Matuszewski and Ida-Maria Sintorn. Minimal annotation training for segmentation of microscopy images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 387–390. IEEE, 2018.

[19] Zahra Mirikharaji, Yiqi Yan, and Ghassan Hamarneh. Learning to segment skin lesions from noisy annotations. In *Domain adaptation and representation transfer and medical image learning with less labels and imperfect data*, pages 207–215. Springer, 2019.

[20] Guy Nir, Soheil Hor, Davood Karimi, Ladan Fazli, Brian F Skinnider, Peyman Tavassoli, Dmitry Turbin, Carlos F Villamil, Gang Wang, R Storey Wilson, et al. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Medical image analysis*, 50:167–180, 2018.

[21] Marcus Nordström, Han Bao, Fredrik Löfman, Henrik Hult, Atsuto Maki, and Masashi Sugiyama. Calibrated surrogate maximization of dice. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 269–278. Springer, 2020.

[22] Tufve Nyholm, Stina Svensson, Sebastian Andersson, Joakim Jonsson, Maja Sohlin, Christian Gustafsson, Elisabeth Kjellén, Karin Söderström, Per Albertsson, Lennart Blomqvist, et al. Mr and ct data with multiobserver delineations of organs in the pelvic area—part of the gold atlas project. *Medical physics*, 45(3):1295–1300, 2018.

[23] Teodora Popordanoska, Jeroen Bertels, Dirk Vandermeulen, Frederik Maes, and Matthew B Blaschko. On the relationship between calibrated predictors and unbiased volume estimation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 678–688. Springer, 2021.

[24] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29, 2016.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[26] Axel-Jan Rousseau, Thijs Becker, Jeroen Bertels, Matthew B Blaschko, and Dirk Valkenborg. Post training uncertainty calibration of deep networks for medical image segmentation. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1052–1056. IEEE, 2021.

[27] Joao Lourenço Silva and Arlindo L Oliveira. Using soft labels to model uncertainty in medical image segmentation. *arXiv preprint arXiv:2109.12622*, 2021.

[28] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017.

[29] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):1–28, 2015.

[30] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020.

[31] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[32] Jiangchao Yao, Jiajie Wang, Ivor W Tsang, Ya Zhang, Jun Sun, Chengqi Zhang, and Rui Zhang. Deep learning from noisy image labels with quality embedding. *IEEE Transactions on Image Processing*, 28(4):1909–1922, 2018.

[33] Shaode Yu, Mingli Chen, Erlei Zhang, Junjie Wu, Hang Yu, Zi Yang, Lin Ma, Xuejun Gu, and Weiguo Lu. Robustness study of noisy annotation in deep learning based medical image segmentation. *Physics in Medicine & Biology*, 65(17):175007, 2020.

[34] Ming-Jie Zhao, Narayanan Edakunni, Adam Pocock, and Gavin Brown. Beyond fano's inequality: Bounds on the optimal f-score, ber, and cost-sensitive risk and their implications. *The Journal of Machine Learning Research*, 14(1):1033–1090, 2013.

[35] Haidong Zhu, Jialin Shi, and Ji Wu. Pick-and-learn: automatic quality evaluation for noisy-labeled image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 576–584. Springer, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Our main claims are informal descriptions of the statements we prove in the Main results section.
   (b) Did you describe the limitations of your work? [Yes] See Limitations in the end of the Conclusions section.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Broader impacts in the Discussion section.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Generalities are described in the preliminaries and the rest in the Main results section.
   (b) Did you include complete proofs of all theoretical results? [Yes] The, proofs are listed in the Supplementary and a short scetch is provided after each stated theorem.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All these details are found in the Supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] .
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] No random experiments of this kind was conducted.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Our experiments does not require any heavy computation. It can be done on a regular laptop. We mention this in the Supplementary.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Several times with [22]
   (b) Did you mention the license of the assets? [Yes] Yes, in the Supplementary material.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Yes, code is included in the Supplementary Material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Yes, see the Supplementary Material.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Yes, see the Supplementary Material.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]