Formalizing cover-source mismatch as a robust optimization

Dominik Šepák, Lukáš Adam, and Tomáš Pevný

***Abstract*—Cover-source mismatch (CSM) refers to the use of a steganographic detector on images with a very different probability distribution it has been trained on. This can have a detrimental effect on its accuracy preventing the use of modern steganalytic tools outside laboratories. Despite CSM being introduced almost fifteen years ago, there is no formal definition and no adopted measures for comparing different solutions. This work, therefore, formalizes the cover-source mismatch and proposes and discusses possible error measures. Equipped with these tools, we propose a principled approach to train holistic detectors while minimizing the effects of CSM and experimentally compare them to the prior art, discussing their strength and weaknesses.**

I. Motivation

Steganography is the art of hiding a secret message into an innocuous-looking object, such that the presence of the message is covert. Steganalysis is the opposite problem of detecting objects with a hidden message. In this paper, the objects of interest are digital images, but the theory developed below are equally applicable to other carriers as well.

Modern detectors of hidden messages hidden in digital images (steganography) are usually implemented by machine learning methods. They have a very large number of parameters, which are optimized on a set of images (usually called a training set). When they are applied to images with very different statistical properties, their accuracy can be poor. This phenomenon is called *cover-source mismatch* and it nowadays represents one of the biggest obstacles in moving steganalysis from laboratories to practice.

The first reported evidence of cover source mismatch (CSM) was published in [2]. At the same time, in [13] has been proposed to detect the double compression in JPEG images and use the corresponding detector on them (this approach would later become known as *atomistic*). CSM was also mentioned in [15], where authors naïvely claimed that to mitigate it suffices to optimize the detector on a sufficiently large mix of cover-sources (this approach would later be called *holistic*). CSM has become a discussed topic during the BOSS contest [1], where authors inadvertently introduced it in the test set, but despite it, there has been to date surprisingly little work in this direction. Ref. [8] studied different methods to select detectors in atomistic approach. Alaska contest and its sequel [3, 4] were designed to include CSM, but to our knowledge, most competitors used *holistic* approach and no fundamentally new method were derived. The most serious study [6] was intended to identify sources of CSM, concluding that ISO settings and preprocessing are among the largest contributors.

Despite almost fifteen years have passed since the description of CSM, the problem has never been mathematically formalized, nor a proper error measures discussed. This paper fills these gaps.

Furthermore, the paper enumerates the prior art in CSM and it proposes a variant of a *holistic* approach minimizing the *regret* for not using the detector specialized for a given cover-source. All discussed approaches to CSM are compared on a subset of the Alaska dataset using the same settings as in [6]. The experiments use a relatively simple steganographic algorithm, nsf5 [5] with simulated optimal coding, and relatively simple detectors relying on 22510-dimensional CCJRM feature set [11]. The rationale behind this decision is relatively low computational complexity decreasing the cost of the experiments. The experimental verification with more complex detectors based on convolution neural networks and advanced steganographic algorithms is left to future work.

II. Formalising The cover-source mismatch

A. Definition of terms

We denote the image x, its corresponding label y, the space of all images $\mathcal{X}$, and the space of labels $\mathcal{Y} = \{\mathrm{cover}, \mathrm{stego}\}$.

1) Cover source: The distribution of cover images depends on many factors, such as the image acquisition device (camera, flatbed scanner), settings of the device (ISO, zoom, aperture, shutter time, etc.), captured scenes, post-processing of images (sharpening, white balance, gamma correction, cropping, etc.), and possibly lossy image container (JPEG compression and re-compression, JPEG codec). We denote the concrete combination of the parameters by ω and the set of all their possible combinations by Ω. Therefore ω uniquely defines the probability distributions of cover images $P^{\mathrm{c}}(x|\omega)$ (if it is clear from the context, we drop the arguments for the sake of brewity). The set of all possible distributions of cover images is a convex hull

$$\mathcal{P}^{\mathrm{c}} = \left\{ \sum_{i=1}^{n} \lambda_i P^{\mathrm{c}}(x|\omega_i) \,\middle|\, \sum_{i=1}^{n} \lambda_i = 1), \lambda_i \geq 0, \omega_i \in \Omega, n \in \mathbb{N} \right\}, \tag{1}$$

as images from different $P^{\mathrm{c}}(x|\omega_i)$ will be mixed in a set, where they cannot be distinguished (for example images taken with different ISO, gamma correction, level of sharpening, etc.).

Definition: *Every element $P^{\mathrm{c}} \in \mathcal{P}^{\mathrm{c}}$ is called a cover-source.*

2) The effect of hiding the message: The distribution of images subjected to the detection is influenced by the algorithm used to embed the message, the distribution of embedded messages, the distribution of steganographic keys, and the proportion of cover and stego images. Specific combination, $\gamma \in \Gamma$, together with distribution of cover images P^{c} determines the distribution of stego images $P^{\mathrm{s}}(x, y|P^{\mathrm{c}}, \gamma)$. Unlike established notation, it assumes (i) realizations of P^{s} includes cover images and (ii) the support of P^{s} is a cartesian product $\mathcal{X} \times \mathcal{Y}$ but note that $\mathcal{Y}$ is not observed by the steganalyst. The peculiarities are used to simplify notation. While this work assumes single fixed $\gamma \in \Gamma$, the general notation is introduced since the problem of CSM is similar to that of universal steganalysis when the algorithm used to embed the message or the distribution of messages is not known.

3) Cover source mismatch: Contemporary approach to steganalysis implements the detector by a function $f(x|\theta) : \mathcal{X} \rightarrow \{\mathrm{cover}, \mathrm{stego}\}$ parametrized by $\theta \in \Theta$.[1] Parameters θ are found by minimizing error (or more generally a loss) estimated on a set of training images $\mathcal{X}^{\mathrm{trn}} = \{(x_i, y_i)\}_{i=1}^{n}$, distributed according to $(x_i, y_i) \sim P^{\mathrm{s}}(x, y|P^{\mathrm{trn}}, \gamma)$, $i \in \{1, \ldots, l\}$, where $P^{\mathrm{trn}} \in \mathcal{P}^{\mathrm{c}}$. Due to the large number of parameters influencing the distribution of images, it is almost guaranteed that testing images analyzed by $f(x|\theta)$ will follow a different distribution.

Definition: *Cover-source mismatch occurs iff* $P^{\mathrm{s}}(P^{\mathrm{tst}}, \gamma) \neq P^{\mathrm{s}}(P^{\mathrm{trn}}, \gamma)$.[2]

B. Measuring the error

The main purpose of error measures is to tell steganalyst, how well its detector will perform on testing images (in practice). Thus, the definition should be closely related to the expected use.

The most established method in steganography is $\mathrm{P_E}$ (defined below in Equation (2)), although there were several attempts to replace it with something more related to practice [4, 12, 14, 9]. All the above attempts do not take cover-source mismatch into the account, i.e. they assume $P^{\mathrm{trn}} = P^{\mathrm{tst}}$. These works are not reviewed here, as the object of interest is a cover-source mismatch, how to measure it, and which measures are aligned with objectives.

a) Clairvoyant error: : Most steganalytic literature assumes that the CSM does not happen, therefore $P^{\mathrm{trn}} = P^{\mathrm{tst}}$. Clairvoyant error allows $P^{\mathrm{trn}} \neq P^{\mathrm{tst}}$, but P^{tst} has to be known. Under these conditions, the principled error measure of detector $f(x|\theta)$ reflecting the real costs of its use is

$$\mathbb{E}_{(x,y)\sim P^{\mathrm{s}}(x,y|P^{\mathrm{tst}},\gamma)}[c(y)\mathrm{I}[f(x|\theta) \neq y]], \tag{2}$$

where $c(y)$ is the cost of making error on class y and $\mathrm{I}[\cdot]$ denotes an Iverson bracket which is one if its argument is true and zero otherwise. In practice, the distribution and P^{tst} is unknown and the expectation is replaced by an estimate from a finite number of realizations from P^{tst} (testing set). Furthermore, most steganographic literature assumes that realizations of $P^{\mathrm{s}}(x, y|P, \gamma)$ have equal number of cover and stego images and the cost $c(y) = 1$. Such error is called $\mathrm{P_E}$ and its popularity stems from the fact that neither fraction of cover and stego images nor $c(y)$ are known.[3] The *clairvoyant error* is very unrealistic, as it assumes the perfect knowledge about the future use of the detector.

b) Maximum error: : When $P^{\mathrm{trn}} \neq P^{\mathrm{tst}}$ and P^{tst} is not known (a true CSM case), it is difficult for the steganalyst to estimate, how his detector will perform in practice. An approach used in *robust classification* is to assume that distance $\delta(P^{\mathrm{trn}}, P^{\mathrm{tst}})$ between P^{trn} and P^{tst} (measured for example by Total Variation) is bounded by some constant ϵ. This enables to define the error as maximum

$$\max_{\delta(P,P^{\mathrm{trn}})\leq\epsilon} \mathbb{E}_{(x,y)\sim P^{\mathrm{s}}(x,y|P,\gamma)}[c(y)\mathrm{I}[f(x|\theta) \neq y]]. \tag{3}$$

Computing the maximization over a set $\{P|\delta(P, P^{\mathrm{trn}}) \leq \epsilon\}$ is generally difficult. In practice, especially in steganalysis where the accurate and tractable model of images is missing, this is to be replaced by maximization over the finite number of cover-sources $\{P_1, \ldots, P_k\}$. Note that the empirical estimate of maximum error upper-bounds the error over a convex hull of $\{P_1, \ldots, P_k\}$.

The main drawback of *maximum error* is that its value is determined by the performance on the most difficult cover-source and therefore it hides the performance on other cover-sources. As an example, imagine two cover-sources P_1^{c} and P_2^{c} such that the best achievable error of detecting stego images in a cover-source P_1 with embedding γ, $\mathrm{P_E}(P_1, \gamma)$, is much smaller than that on the other cover-source P_2, $\mathrm{P_E}(P_1, \gamma)$, i.e. $\mathrm{P_E}(P_1, \gamma) \ll \mathrm{P_E}(P_2, \gamma)$. In this case, the error of a detector $f(x|\theta)$ measured by *maximum error* will be dominated by the error on the cover-source P_2 and the performance on P_1 might be arbitrarily poor while smaller than that on P_2. This is clearly an undesirable behavior.

c) Maximum regret: : To fix the weakness of maximum error, the steganalyst might measure a difference in errors of the detector $f(x|\theta)$ and the best achievable detector on a given cover-source. The error of the best achievable detector on cover-source P and embedding parameters γ, $\mathrm{P_E}(P, \gamma)$, will therefore serve as a calibration.

$$\max_{\delta(P,P^{\mathrm{trn}})\leq\epsilon} \mathbb{E}_{(x,y)\sim P^{\mathrm{s}}(x,y|P,\gamma)}[c(y)\mathrm{I}[f(x|\theta) \neq y]] - \mathrm{P_E}(P, \gamma). \tag{4}$$

This formulation tells the steganalyst the maximal deviation from optimum over cover-sources. In practice, the maximization over infinite set $\{P|\delta(P, P^{\mathrm{trn}}) \leq \epsilon\}$ will be replaced by a maximization over the finite set of cover-sources $\{P_1, \ldots, P_k\}$. Similarly, $\{\mathrm{P_E}(P_i, \gamma)\}_{i=1}^{k}$ cannot be computed exactly but have to be estimated by training a classifier.

Remark: Replacing $\max$ with mean in Equations (3) and (4) makes both equations equal to Equation (2) with a

[1] θ should be understood here in the wider context including choice and architecture of classifiers.

[2] Since we assume γ to be the same for training and testing, this condition can be translated to $P^{\mathrm{tst}} \neq P^{\mathrm{trn}}$ for all reasonably well behaving steganographic methods.

[3] Few works [4, 12, 9] has used other measures of error than the prevalent $\mathrm{P_E}$. For the sake of simplicity, this work will use $\mathrm{P_E}$.

suitably selected P^{tst} (Equation (4) would contain additional term $\mathrm{mean}_{\delta(P,P^{\mathrm{trn}})}\mathrm{P_E}(P,\gamma)$ constant with respect to $f(x|\theta)$). Setting $\mathrm{P_E}(P,\gamma)$ to zero in Equation (4) makes it equivalent to Equation (3).

III. Compared methods robust to cover-source mismatch

This section describes methods compared in the experimental Section. While some of them have been already published, we discuss the optimization of criterion (4). The exposition is divided into two parts: the first describes different approaches to train detectors robust to CSM and the second discusses how linear and non-linear classifiers were optimized. All classifiers assume images to be described by features, which is done to save the computational resources.

A. Holisitic

Refs. [15, 6] suggests a *holistic* approach to CSM consisting of training a single detector on all available cover-sources. This corresponds to minimizing error in Equation (2) with respect to $f(x|\theta)$ with an Iverson bracket replaced by differentiable surrogate and with a fixed P^{trn}, which usually follows the number of images from each cover-source available during optimization of the detector. If P^{trn} is close to P^{tst}, this is from a theoretical point of view a very good approach. A linear holistic detector [6] (denoted as *holistic linear*) was implemented by Ridge Regression (see Subsection IV-A). Ref. [15] has used a non-linear classifier, specifically Support Vector Machine with Gaussian Kernel. Below, it was replaced by a multi-layer perceptron due to better scaling with respect to the number of training samples (see Subsection IV-A). This approach is denoted later as *holistic MLP.* The non-linear classifier is used, since the cover and stego images in the mixture of cover-sources might be difficult to separate.

B. Atomistic

In atomistic approach [6, 13, 8], the steganographer creates one detector for each cover-source available during training. During steganalysis, it outputs the decision of the most suitable detector for a given image. The first part, creating a detector for each cover-source, is straightforward. Below, linear classifiers were used as in [6], as they should be sufficient here due to the high number of features describing each object and relatively low number of training samples from each cover-source (see below).

The main difficulty of the atomistic approach stems from finding the best classifier from the pool for a given image. Sometimes, this selection can be done on basis of available information (for example a quality factor (QF) in JPEG images used in [13]), but more frequently it has to be estimated. In this work, the cover-source is estimated by a trained classifier (multilayer perceptron constructed as described in Subsection IV-A). The experimental section compares two variants: predicting just the camera model (the final detector is called *atomistic QF predicted*) and predicting camera model and quality factor (the final detector is called *atomistic QF known*). While the QF can be read from a JPEG container, the authors were curious to see the difference inaccuracy.

C. Maximum regret

The Equation 4 can be used as a loss function in the optimization of a detector as long as the Iverson bracket is replaced by a differential surrogate (the choice usually depends on the used classifier). Here, $f(x|\theta)$ were implemented as a multi-layer perceptron (see Section IV-A).

As already mentioned, the maximization over the set $\{P|\delta(P,P^{\mathrm{trn}})\leq\epsilon\}$ in Equation (4) has been replaced by maximization over finite number of cover-sources $\{p_1,\ldots,p_k\}$. The best achievable error per cover-source, $\mathrm{P_E}(p_i,\gamma), i\in\{1,\ldots,k\}$, has been estimated by the error of the linear classifier on the validation set. During optimization, cover-source with largest regret *on the validation set* was selected in each training step. But the gradients for the update of model parameters were computed on the *training set*. This dichotomy stems from the fact that due to the relatively small number of samples, it is trivial to achieve zero error on all cover-sources in the training set. The detector was trained for 750 000 training steps.

IV. Experimental details

A. Implementation details of classifiers

1) Linear classifier: All linear classifiers $f(x|w,b)=w^{\top}x+b$ with parameters w and b were trained using Ridge Regression [7], which replaces Iverson bracket in Equation (2) by a differentiable $\mathrm{L}_2(y,\hat{y})=\frac{1}{2}(y-\hat{y})^2$. This has the advantage that w has a closed-form solution as $w=(\mathbf{X}^{\top}\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^{\top}y$, where $\mathbf{X}\in\mathbb{R}^{n,d}$ is a matrix containing features extracted from training images in their rows, $y\in-1,+1^n$ is vector with corresponding labels, and $\lambda\geq 0$ is a regularization parameter. Its value was selected from $\lambda\in\{10^i|i\in\{-7,-6,\ldots,3,4\}\}$ to minimize the $\mathrm{P_E}$ estimated on the validation dataset.

2) Non-linear classifier: All non-linear classifiers were implemented by multi-layer perceptron. Due to its large flexibility in terms of architecture and training parameters, a random search selecting the best configuration with minimal $\mathrm{P_E}$ error on the validation dataset was used. Architectures were sampled from the following Cartesian product of: (1, 2 or 3) hidden layers, (8, 16, 32, 64, or 128) number of neurons, use of batch normalization, and (ReLU, $\tanh$) activation function of hidden layers. The optimization was done with an Adam [10] variant of gradient descend with full batch, with learning rate sampled from the set $\{ab|a\in\{2,4,\ldots,10\},b\in\{1\times10^{-4},1\times10^{-5},1\times10^{-6}\}\}$ and weight decay parameter from the set $\{0\}\bigcup\{1\times10^{-2},1\times10^{-3},\ldots,1\times10^{-6}\}$. The network was trained for 200 000 iterations with full batch minimizing binary cross-entropy with early stopping, which has selected the best model with lowest $\mathrm{P_E}$ on the validation set.

B. Image database

A cover-source is simulated by camera model and by the quality factor, keeping all other parameters of the processing pipeline constant. Raw images were sourced from the

Camera name	ISO	no. of images	cover source labels
Panasonic FZ28	100	1164	A
Nikon D610	100	1060	B
iPad Pro (12.9” gen. 2)	20	1571	C
Canon EOS 500D	1600	1665	D
Sony ILCE-7R	800	926	E
Pentax K10D	400	835	F
Panasonic GM1	3200	417	G

TABLE I
LABELS OF COVER SOURCES, THEIR CAMERA MODELS AND ISO VALUES OF ITS IMAGES, AND THEIR NUMBER OF IMAGES AVAILABLE FOR EXPERIMENTS.

ALASKA dataset, from which seven camera models with a large number of images and with different ISO sensitivity were selected. For brevity, cover-sources are denoted by letters A–G, their details can be found in Table I. Images were converted from RAW format to TIFF using RawTherapee version 5.4 with the default neutral profile. The TIFF images were cropped to 512×512 pixels. These were then converted to JPEG images with quality factors 75 and 100 using and 4:4:4 subsampling. The conversion was performed in Python 3.6.9 using Pillow 8.1.2 library.

The creation of stego images mimicked [6]. They were created by simulating embedding of a random message with payload 0.04 bits per AC coefficient (bpAC) by nsF5 algorithm with optimal embedding. All images were represented by cc-JRM [11] feature vectors with dimension 22510.

Data from each cover-source were split into training, validation, and testing sets according to the 6:2:2 ratio. This split was three times repeated, therefore all results reported below are averages from metrics computed on the testing set (unless said otherwise). Furthermore, the seven cover-source identified by letters A–G (see Table I for their detail) were divided such that images from cover-sources F and G were *never* part of the training set nor the *validation* set. They were only used for computing metrics on the testing set, which allowed to study accuracy on cover-sources *unseen* during the development of the detector. The validation set was used to optimize all hyperparameters described in Section IV-A.

V. EXPERIMENTAL RESULTS

Each row in Table (II) contains quantities computed on the testing set from a given cover-source (for camera models corresponding to letter see Table IV). The quantities in a column captioned *Clairvoyant* contains $\mathrm{P_E}$ of a linear detector trained and tested on the same cover-source indentified by the row. This is an optimistic (”optimal”) result, to which other detectors are compared to. The remaining quantities contain $\mathrm{P_E}$ of a corresponding detector (in a column) on a given cover-source (in a row) minus $\mathrm{P_E}$ of the clairvoyant detector. This means that if there is a negative number, the detector is on that cover-source better than clairvoyant. Authors confess that they have not anticipated seeing negative numbers, which are likely

	QF	**Clairvoyant**	**holistic linear**	**holistic MLP**	**maximum re-gret**	**atomistic QF predicted**	**atomistic QF known**
A	75	7.9	0.4	-1.6	1.1	-0.1	-0.2
B	75	5.7	-0.1	-0.3	0.6	0.3	0.1
C	75	9.0	-0.3	-0.1	0.6	0.5	0.5
D	75	2.0	0.7	0.4	1.0	0.0	0.0
E	75	8.3	-2.1	-2.6	-1.3	0.2	0.0
F	75	7.2	-2.3	-1.0	-1.0	-0.2	-0.2
G	75	11.5	-1.0	2.2	3.4	11.1	9.9
A	100	40.7	-1.3	-3.4	-0.3	-0.7	0.0
B	100	34.0	-2.5	-0.9	0.9	0.1	0.0
C	100	41.4	-2.0	-2.5	-1.0	0.2	0.2
D	100	22.4	11.3	7.0	-0.3	0.0	0.0
E	100	38.6	-0.4	-1.6	1.4	-0.4	-0.2
F	100	39.8	-0.8	-2.6	1.3	0.3	0.9
G	100	44.4	-0.6	-2.8	-1.2	0.8	0.6

TABLE II
EACH ROW CONTAINS QUANTITIES ESTIMATED FROM THE TESTING SET ON ONE COVER-SOURCE. COLUMN DENOTED BY *Clairvoyant* $\mathrm{P_E}$, OTHER COLUMNS CONTAINS $\mathrm{P_E}$ MINUS $\mathrm{P_E}$ OF *Clairvoyant* ON THE SAME ROW. NEGATIVE NUMBERS THEREFORE INDICATES THAT THE DETECTOR IS BETTER THAN *Clairvoyant*.

	QF	**Clairvoyant**	**holistic linear**	**holistic MLP**	**maximum re-gret**	**atomistic QF predicted**	**atomistic QF known**
A–E	mean $\mathrm{P_E}$		21.4	**20.4**	21.3	21.0	21.0
	max. $\mathrm{P_E}$		39.5	**38.9**	40.4	41.6	41.6
	max. regret		11.3	7.0	1.4	**0.5**	**0.5**
A–G	mean $\mathrm{P_E}$		22.3	**21.6**	22.7	23.2	23.2
	max. $\mathrm{P_E}$		43.8	**41.7**	43.3	45.2	45.0
	max. regret		11.3	7.0	**3.4**	11.1	9.9

TABLE III
ERRORS (2), (3), AND (4) AGGREGATED OVER COVER SOURCES A–E AND A–G FROM VALUES IN TABLE II.

	QF	**A QF75**	**B QF75**	**C QF75**	**D QF75**	**E QF75**	**A QF100**	**B QF100**	**C QF100**	**D QF100**	**E QF100**
A	75	77.6	1.3	11.7	0.1	6.4	0.0	0.0	0.0	0.0	0.0
B	75	0.9	96.7	1.9	0.0	0.9	0.0	0.0	0.0	0.0	0.0
C	75	15.1	1.3	81.9	0.0	6.7	0.0	0.0	0.1	0.0	0.0
D	75	0.1	0.0	0.0	99.9	0.0	0.0	0.0	0.0	0.0	0.0
E	75	6.4	0.6	4.5	0.0	86.0	0.0	0.0	0.0	0.0	0.0
A	100	0.0	0.0	0.0	0.1	0.0	82.1	0.0	6.8	0.0	7.6
B	100	0.0	0.0	0.0	0.0	0.0	0.2	100.0	0.3	0.0	0.0
C	100	0.0	0.0	0.0	0.0	0.0	10.0	0.0	90.3	0.0	2.3
D	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.9	0.0
E	100	0.0	0.2	0.0	0.0	0.0	7.7	0.0	2.6	0.1	90.1

TABLE IV
CONFUSION MATRIX OF THE DETECTOR OF COVER-SOURCE PREDICTING CAMERA MODEL AND THE QUALITY FACTOR.

caused by the fact that holistic and maximal regret classifiers are trained on all samples.

Table (III) contains error measures introduced Section II-B, specifically average $\mathrm{P_E}$, where each cover-source has the same probability of occurrence (Equation (2)), maximum error over cover-source (Equation (3)), and finally maximum regret (Equation (4)). Comparing the approaches using these quantities, suggests that if all testing cover-sources have been present in the training set, the *atomistic approach* is the best on average $\mathrm{P_E}$ and surprisingly also on maximum regret. Its great performance is likely caused by the high accuracy of the cover-source detector which is generally above 95% (see Table IV). On maximum $\mathrm{P_E}$, the best results have been achieved by the holistic approach. When the testing set contains *additional* cover-sources, maximum regret seems to be the best on maximal regret and maximal error. It is also very good on mean $\mathrm{P_E}$, where it is just by 0.4 worse than the best linear holistic approach. These results, together with good results on cover-sources A–E make it a very good option.

VI. Discussion

This paper has formalized the cover-source mismatch in terms of probability distributions of images, which allowed to define different concepts of errors aligned with different objectives of the steganalyst: while *clairvoyant* error assumes the knowledge of the distribution of images on the testing set, *maximum error* and *maximum regret* enable to principally integrate uncertainty about it.

Using these types of errors, prior art approaches to CSM and the proposed minimization of maximal regret were experimentally compared on the subset of the Alaska 2 dataset. According to the results, *atomistic* approach is very good if cover-sources from the testing set are present in the training set, *maximum regret* is very good even when they were not known during training. This seems that *maximum regret* delivers more robust *holistic* classifier. Unfortunately, the training complexity of both maximum regret and atomistic approaches is very high, as in both cases one needs to train classifier per cover-source.

VII. Acknowledgement

The research leading to this paper has received funding from by OP VVV project CZ.02.1.01/0.0/0.0/16_019/0000765 ”Research Center for Informatics”, by Czech Ministry of Education 19-29680L, and the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 101021687. This material reflects only the authors views and the Commission is not liable for any use that may be made from the information contained therein.

References

[1] P. Bas, T. Filler, and T. Pevný. ” break our steganographic system”: the ins and outs of organizing boss. In *International workshop on information hiding*, pages 59–70. Springer, 2011.

[2] G. Cancelli, G. Doërr, M. Barni, and I. J. Cox. A comparative study of±steganalyzers. In *2008 IEEE 10th Workshop on Multimedia Signal Processing*, pages 791–796. IEEE, 2008.

[3] R. Cogranne, Q. Giboulot, and P. Bas. The alaska steganalysis challenge: A first step towards steganalysis. In *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, pages 125–137, 2019.

[4] R. Cogranne, Q. Giboulot, and P. Bas. Alaska# 2: Challenging academic research on steganalysis with realistic images. In *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–5. IEEE, 2020.

[5] J. Fridrich, T. Pevný, and J. Kodovský. Statistically undetectable jpeg steganography: dead ends challenges, and opportunities. In *Proceedings of the 9th workshop on Multimedia & security*, pages 3–14, 2007.

[6] Q. Giboulot, R. Cogranne, D. Borghys, and P. Bas. Effects and solutions of cover-source mismatch in image steganalysis. *Signal Processing: Image Communication*, 86:115888, 2020.

[7] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[8] A. D. Ker and T. Pevný. A mishmash of methods for mitigating the model mismatch mess. In *Media Watermarking, Security, and Forensics 2014*, volume 9028, pages 189–203. SPIE, 2014.

[9] A. D. Ker and T. Pevný. The steganographer is the outlier: Realistic large-scale steganalysis. *IEEE Transactions on information forensics and security*, 9(9):1424–1435, 2014.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] J. Kodovský and J. Fridrich. Steganalysis of jpeg images using rich models. In *Media Watermarking, Security, and Forensics 2012*, volume 8303, page 83030A. International Society for Optics and Photonics, 2012.

[12] T. Pevný. Detecting messages of unknown length. In *Media Watermarking, Security, and Forensics III*, volume 7880, pages 300–311. SPIE, 2011.

[13] T. Pevný and J. Fridrich. Multiclass detector of current steganographic methods for jpeg format. *IEEE Transactions on Information Forensics and Security*, 3(4):635–650, 2008.

[14] T. Pevný and A. D. Ker. Towards dependable steganalysis. In *Media Watermarking, Security, and Forensics 2015*, volume 9409, page 94090I. International Society for Optics and Photonics, 2015.

[15] T. Pevný, P. Bas, and J. Fridrich. Steganalysis by subtractive pixel adjacency matrix. *IEEE Transactions on information Forensics and Security*, 5(2):215–224, 2010.