SCOD: From Heuristics to Theory
Supplementary material

A Alternative formulations of the SCOD problem

In this paper, we address the SCOD problem (4), further denoted SCOD-tpr, which aims to minimize the SCOD risk $R(h,c) = (1-\alpha)\mathrm{R_S}(h,c) + \alpha\mathrm{fpr}(c)$ subject to a constraint ensuring a minimum TPR of $\mathrm{tpr}_{\min}$. The TPR represents the probability that an ID sample will be accepted for classification by a selector $c(x)$, defined as $\mathrm{tpr}(c) = \mathbb{E}_{x\sim p_I(x)} c(x)$. An alternative formulation, discussed in [27], replaces TPR with total coverage $\rho(c) = \mathrm{tpr}(c)(1-\pi_O) + \mathrm{fpr}(c)\pi_O$ which represents the probability of accepting an input sample, whether generated from ID or OOD. Consequently, we arrive at the SCOD-coverage formulation:

$$\min_{\substack{h\in\mathcal{Y}^{\mathcal{X}} \\ c\in[0,1]^{\mathcal{X}}}} \big[(1-\alpha)\,\mathrm{R_S}(h,c) + \alpha\,\mathrm{fpr}(c)\big] \qquad \text{s.t.} \qquad \rho(c) \geq \rho_{\min}\,, \tag{14}$$

where $\rho_{\min} \in (0,1)$ is a user-defined minimum acceptable coverage.

Both formulations are well-defined, and the choice between them, based on whether one favors accepting a specific portion of ID or a portion of all samples, depends on the specific application. However, the SCOD-tpr formulation analyzed in this paper offers several practical advantages over the SCOD-coverage formulation:

1. The optimal strategy (h^*, c^*) for solving the SCOD-tpr formulation (4) is independent of the portion of OOD data π_O. None of the key quantities, namely $R_S(h,c)$, $\mathrm{tpr}(c)$, and $\mathrm{fpr}(c)$, are influenced by π_O. This property is crucial because, in practice, π_O is not only unknown but often nonstationary. In contrast, the SCOD-coverage formulation relies on the OOD portion π_O through the coverage $\rho(c)$. Consequently, it is not applicable in a nonstationary setup.
2. The optimal selector for both the SCOD-tpr and SCOD-coverage formulations relies on a linear combination of the conditional risk $r(x)$ and the ID/OOD score $g(x) = p_I(x)/p_O(x)$. Specifically, the optimal score is $s(x) = r(x) + \beta\, g(x)$ where $\beta \in \mathbb{R}$ is a multiplier dependent on the problem parameters. In (7), we demonstrated that the optimal multiplier β for the SCOD-tpr can be analytically computed as $\beta = \alpha\mathrm{tpr}_{\min}/(1-\alpha)$. Conversely, for SCOD-coverage, there is no analytical formula for computing β, and it necessitates solving an optimization problem based on quantities reliant on clean OOD data (refer to [27] for further details).

In summary, both SCOD-tpr and SCOD-coverage formulations are well-defined and intuitive. However, SCOD-tpr is independent of π_O, possesses an analytically solvable multiplier β for the optimal combination of $r(x)$ and $g(x)$, and learning the optimal selective classifier does not necessitate clean OOD samples. Conversely, SCOD-coverage formulation inherently depends on π_O, requires optimization of the multiplier β for combining $r(x)$ and $g(x)$, and learning the optimal selective classifier requires clean OOD samples.

B Implementation Details

B.1 Datasets

We adopt datasets from the OpenOOD benchmark [43] and assess SCOD performance across three ID datasets: CIFAR-10, CIFAR-100, and ImageNet-1K.

ID ImageNet-1K In line with [43], we employ 45,000 images from the ImageNet-1K validation set as the *in-distribution* (ID) data for evaluation. The official ImageNet-1K test set could *not* be utilized for this purpose as the ground-truth labels are *not* publicly accessible. For the near-OOD group we employ the SSB_Hard [40] and NINCO [1] datasets. In the far-OOD group we include the iNaturalist [17, 39], Textures [6], and OpenImage_O [21, 41] datasets. A brief overview of the datasets is provided below:

1. **SSB_Hard** [40]: A dataset comprising 49,000 images covering 980 categories selected from ImageNet-21K [32] that are absent in ImageNet-1K. We evaluate on the entire dataset.
2. **NINCO** [1]: A dataset of 5,879 images with manually filtered-out noise. The majority of the dataset was extracted from the species subset of iNaturalist [39]. It is intentionally designed to be challenging to differentiate from ImageNet-1K samples, with examples such as a marbled newt considered OOD, while a common newt is considered ID. We evaluate on the entire dataset.
3. **OpenImage_O** [21, 41]: A dataset consisting of 17,632 images manually selected from the test set of the OpenImages dataset [21]. We use a subset defined by the OpenOOD [43] benchmark, comprising 15,869 images.
4. **iNaturalist** [17,39]: A dataset consisting of 10,000 images randomly sampled from 110 manually selected *plant* classes not present in ImageNet-1K. The samples were obtained by [17] from the full iNaturalist dataset [39]. We evaluate on all 10,000 samples.
5. **Textures** [6]: A dataset consisting of 5,640 images, split into 47 texture classes (e.g., braided, striped, wrinkled). We evaluate on the entire dataset.

ID CIFAR-10 In line with [43], we use 9,000 images from the CIFAR-10 [20] test set as the *in-distribution* (ID) data for evaluation. For the near-OOD group, we employ the CIFAR-100 [20] and Tiny-ImageNet (TIN) [22] datasets. In the far-OOD group, we include the Street View House Numbers (SVHN) [29], Places365 [44], MNIST [23], and Textures [6] datasets. A brief overview of the datasets and a description of how we utilize them are provided below:

1. **CIFAR-100** [20]: Dataset of 60,000 images across 100 classes, selected from the Tiny Images [38] dataset such that there is no semantic overlap with CIFAR-10. The test set comprises 10,000 images, however, we evaluate the methods only on a subset of 9,000 images defined by the OpenOOD [43] benchmark.

2. **TIN** [22]: Dataset of 100,000 images divided into 200 classes. For every class, there are 500 training images, 50 validating images, and 50 test images. Following the OpenOOD [43] benchmark, we evaluate the methods on the validation set and only use 7793 of the 10,000 images as the OOD dataset, removing samples that semantically overlap with CIFAR-10.
3. **SVHN** [29]: Dataset of 99,289 house numbers taken from Google Street View images. We use the pre-processed *cropped* variant of the dataset, and evaluate the methods on the test set comprising 26,032 images.
4. **Places365** [44]: Scene recognition dataset comprising over 10,000,000 images divided into 434 scene classes. Following the OpenOOD benchmark [43], we use the *Places365-Standard* version of the dataset and evaluate on the validation set. We use only 35,195 out of the 36,000 images due to semantic overlap of the samples with CIFAR-10.
5. **Textures** [6]: A dataset consisting of 5,640 images, split into 47 texture classes (e.g., braided, striped, wrinkled). We evaluate on the entire dataset.
6. **MNIST** [23]: Dataset comprising of 70,000 images of handwritten digits. Following the OpenOOD benchmark [43], we evaluate on the entire dataset.

ID CIFAR-100 In line with [43], we use 9,000 images from the CIFAR-100 [20] test set as the *in-distribution* (ID) data for evaluation. For the near-OOD group, we employ the CIFAR-10 [20] and Tiny-ImageNet (TIN) [22] datasets. In the far-OOD group, we include the Street View House Numbers (SVHN) [29], Places365 [44], MNIST [23], and Textures [6] datasets. A brief overview of the datasets and a description of how we utilize them are provided below:

1. **CIFAR-10** [20]: Dataset of 60,000 images across 10 classes, selected from the Tiny Images [38] dataset such that there is no semantic overlap with CIFAR-100. The test set comprises 10,000 images, however, we evaluate the methods only on a subset of 9,000 images defined by the OpenOOD [43] benchmark.
2. **TIN** [22]: Dataset of 100,000 images divided into 200 classes. For every class, there are 500 training images, 50 validating images, and 50 test images. Following the OpenOOD [43] benchmark, we evaluate the methods on the validation set and only use 6,526 of the 10,000 images as the OOD dataset, removing samples that semantically overlap with CIFAR-100.
3. **SVHN** [29]: Dataset of 99,289 house numbers taken from Google Street View images. We use the pre-processed *cropped* variant of the dataset, and evaluate the methods on the test set comprising 26,032 images.
4. **Places365** [44]: Scene recognition dataset comprising over 10,000,000 images divided into 434 scene classes. Following the OpenOOD benchmark [43], we use the *Places365-Standard* version of the dataset and evaluate on the validation set. We use only 33,773 out of the 36,000 images due to semantic overlap of the samples with CIFAR-100.
5. **Textures** [6]: A dataset consisting of 5,640 images, split into 47 texture classes (e.g., braided, striped, wrinkled). We evaluate on the entire dataset.
6. **MNIST** [23]: Dataset comprising of 70,000 images of handwritten digits. Following the OpenOOD benchmark [43], we evaluate on the entire dataset.

B.2 Models

CIFAR-10/100 For the ID classifier $h(x)$ on CIFAR-10/100, we use pre-trained ResNet-18 classifiers from the OpenOOD benchmark by [43]. All of the models were originally trained using standard cross-entropy loss, employing the SGD optimizer with a momentum of 0.9, a learning rate set to 0.1, and a cosine annealing decay schedule. Additionally, a weight decay of 0.0005 was applied. During evaluation, we apply input normalization identical to the one used during the training of the classifier; the normalization is different for CIFAR-10 and CIFAR-100.

ImageNet-1K For the ID classifier $h(x)$ on ImageNet-1K, we use a pre-trained ResNet-50 model from Torchvision [37]. Specifically, we use the `IMAGENET1K_V1` model. During evaluation, we apply input normalization identical to the one used during the training of the classifier.

B.3 Approximating the likelihood ratio

To estimate the OOD/ID likelihood ratio $g(x)$ on real-world data, we train a classifier with BCE of the standard sigmoid (a.k.a logistic regression) and a *corrected sigmoid* model (refer to Sec. 4) using features from the ID classifier. In other words, we replace the last layer of the classifier $h(x)$ with a linear layer with a single output. Additionally, we add a `Dropout` layer with $p_{\text{train}} = 0.2$ before the linear layer. For the experiments on CIFAR-10/100, we allow all weights of $h(x)$ to be modified. For experiments on ImageNet-1K, we only learn the last layer.

Training details During training, we apply several data augmentations, including rotation up to 20 degrees, random resized cropping, and horizontal mirroring. For ImageNet-1K, we extend the augmentation strategy to incorporate variations in brightness, contrast, saturation, and hue. We train the models with binary cross-entropy loss using the AMSGrad variant of the Adam optimizer with a learning rate of 0.003 for a total of 200 epochs. After every 50 epochs, the learning rate is decreased by a factor of 10. The final model is selected based on the validation loss. For further details, please refer to the implementation.

Data splitting When training the likelihood ratio approximation $\hat{g}(x)$, we use the dedicated training sets of ImageNet-1K, CIFAR-10, CIFAR-100, Tiny-ImageNet, and SVHN. Conversely, for MNIST, Places365, SSB_Hard, NINCO, iNaturalist, Textures, and OpenImage_O, we adopt a random split approach, allocating 50% of each dataset for training $\hat{g}(x)$ and the remaining 50% for evaluation. Note that the evaluations are always reported exclusively on samples unseen during training.

Mixture In Sec. 4 we adopt the framework proposed by [18], where in addition to the ID data sample $\mathcal{T}_I = ((x_i^I, y_i^I) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \ldots, m)$ generated from $p_I(x, y)$ we also have an unlabeled sample of a mixture of ID and OOD data

$\mathcal{T}_U = (x_i^U \in \mathcal{X} \mid i = 1, \ldots, n)$ generated from $p(x) = \pi_O p_O^{\text{tr}}(x) + (1 - \pi_O^{\text{tr}}) p_I(x)$. We demonstrate that within this framework, the SCOD problem can be effectively addressed using the proposed POSCOD algorithm. In our experimental setup, we set π_O^{eval} to 50%, indicating an equal distribution of ID and OOD samples in the unlabeled mixture used during evaluation. We use the same evaluation mixture irrespective of the training prior π_O^{tr}. Results for varying π_O^{tr} can be found in Appendix C.

B.4 Hyperparameters of double-score strategies

The current state-of-the-art SCOD selector $s_{\text{SIRC}}(x) = -(S_1^{max} - s_1(x))(1 + \exp(-b(s_2(x) - a))$ employs a nonlinear combination of scores $s_1(x)$ and $s_2(x)$ with two hyperparameters a, b. In contrast, our approach utilizes a linear combination $s_{\text{Linear}}(x) = s_1(x) + \beta s_2(x)$ with a single hyperparameter β, for which we derived an explicit formula, see Eq. (7). When comparing the two stragies in the *tuned* setup, see Sec. 6.2, we search for the best hyperparameters a, b, β, optimizing the SCOD risk on the test set; providing an upper bound on the model's performance.

Hyperparameters of SIRC For practical deployment of s_{SIRC}, a heuristic is provided by [42] to set the parameters a, b using the empirical mean μ_{s_2} and standard deviation σ_{s_2} of the score $s_2(x)$ on ID data. Specifically, they propose setting $a_{\text{plug}} = \mu_{s_2} - 3\sigma_{s_2}$, and $b_{\text{plug}} = {}^1\!/_{\sigma_{s_2}}$. When searching for the optimal parameters a and b, we use these heuristic formulae as a rough estimate. We search over parameters $(a, b) \in [\mu_{s_2} - 6\sigma_{s_2},\ \mu_{s_2}] \times [{}^1\!/_{10\sigma_{s_2}},\ {}^{10}\!/_{\sigma_{s_2}}] = [a_{\text{plug}} - 3\sigma_{s_2},\ a_{\text{plug}} + 3\sigma_{s_2}] \times [\frac{1}{10} \cdot b_{\text{plug}},\ 10 \cdot b_{\text{plug}}]$. We sample both intervals at 40 evenly spaced values and at the heuristic values themselves; resulting in $41^2 = 1681$ possible hyperparameter settings.

Hyperparameter of linear strategy For the linear strategy, we derive an explicit formula $\beta_{\text{plug}} = \frac{\alpha\, \text{tpr}_{\text{min}}}{1-\alpha}$, see Eq. (7). However, when searching for the optimal parameter β on a test set, we parameterize the linear combination by an angle ϑ, $s_{\text{Linear}} = \cos\vartheta s_1(x) + \sin\vartheta s_2(x)$ and search over the interval $\vartheta \in [0, 2\pi]$. This is equivalent to the parameterization $\beta = \tan\vartheta$. The relative scale of the two scores is absorbed in the parameterization by the scale of the decision threshold λ. We sample 1600 evenly spaced values from the interval, as well as $\frac{\pi}{2}, \pi, \frac{3\pi}{2}$; resulting in 1603 possible hyperparameter settings.

B.5 Evaluation Metrics

Definition To summarize the performance of a selective classifier (h, c) where $c(x; \lambda) = [\![s(x) \leq \lambda]\!]$, on the evaluation data $\mathcal{T}$, we employ three metrics:

1. *Area Under Risk - Coverage* (AuRC↓) curve which measures the model's ability to discriminate between correctly classified (✓ ID) and incorrectly classified (✗ ID) data,

2. *Area Under Receiver Operating Characteristic* (AuROC ↑) curve which evaluates how well the model distinguishes between ID and OOD data, and
3. *Area Under SCOD Risk - True positive rate* (AuSRT ↓) curve which is used to evaluate the overall performance of the selective classifier on the SCOD problem, see Def. 1.

AuRC and AuROC are standard metrics for selective classification (SC) and out-of-distribution detection (OODD). AuSRT directly addresses the SCOD objective. Currently, single-score methods tend to address either SC or OODD; but not both at the same time. The SCOD problem addresses both tasks at the same time.

Empirical Estimates As we are only provided with a sample of data, we use the empirical estimates of the metrics. E.g., instead of TPR, we use

$$\widehat{\mathrm{tpr}}(c) = \frac{1}{m} \sum_{(x,y)\in\mathcal{T}_I} c(x).$$

Analogously, for the FPR and the selective risk we use:

$$\widehat{\mathrm{fpr}}(c) = \frac{1}{n} \sum_{(x,\emptyset)\in\mathcal{T}_O} c(x)$$

and

$$\widehat{\mathrm{R}}_{\mathrm{S}}(h,c) = \frac{\frac{1}{m}\sum_{(x,y)\in\mathcal{T}_I} \ell(h(x),y)\, c(x)}{\widehat{\mathrm{tpr}}(c)}.$$

The SCOD risk, Eq. (4), at a given TPR is estimated by

$$\widehat{\mathrm{R}}(\mathrm{tpr}_{\min}; h, c) = \min_{\lambda} (1-\alpha)\, \widehat{\mathrm{R}}_{\mathrm{S}}(h,c) + \alpha\, \widehat{\mathrm{fpr}}(c) \text{ s.t. } \widehat{\mathrm{tpr}}(c) \geq \mathrm{tpr}_{\min},$$

where $c(x;\lambda) = [\![s(x) \leq \lambda]\!]$. The AuSRT is then computed as the area under the curve given by points $\left\{ \left(\widehat{\mathrm{R}}(\mathrm{tpr}_{\min}), \mathrm{tpr}_{\min} \right) \mid \mathrm{tpr}_{\min} \in (0,1] \right\}$.

C Ablation

Relative cost ablation Our analysis on ImageNet, see Tab. 4, reveals that the linear strategy using the plugin score, Eq. (7), consistently performs better than SIRC and single-score strategies, across different values of the relative cost $\alpha \in [0, 1]$. In other words, the linear strategy outperforms alternatives independent of the relative weights assigned to the OODD and SC tasks.

Table 4: AuSRT↓ in % points for the POSCOD approximated likelihood ratio (LR) when varying the relative cost α. Results are shown for practically usable plugin double-score strategies on ID ImageNet.

	Score \ α	0.10	0.25	0.5	0.75	0.9
ssb_hard	SIRC	**8.43**	10.62	14.27	17.93	20.13
	Linear	8.44	**7.13**	**6.37**	**7.22**	**9.97**
ninco	SIRC	**8.10**	9.81	12.67	15.53	17.24
	Linear	8.32	**7.78**	**7.44**	**7.66**	**8.39**
inaturalist	SIRC	7.17	7.49	8.02	8.56	8.88
	Linear	**6.80**	**5.75**	**4.16**	**2.73**	**2.01**
textures	SIRC	7.67	8.73	10.50	12.29	13.35
	Linear	**6.81**	**5.81**	**4.38**	**3.21**	**2.84**
openimage_o	SIRC	7.52	8.35	9.75	11.15	11.99
	Linear	**7.28**	**6.86**	**6.48**	**6.44**	**6.81**

Prior π_O^{tr} ablation In the main paper body, we demonstrate that POSCOD is an effective approach for learning an estimate $\hat{g}(x)$ of the OOD/ID likelihood ratio when the unlabeled mixture $\mathcal{T}_U$ used for training contains an equal ratio of ID and OOD samples, i.e., $\pi_O^{\text{tr}} \approx 1/2$. For other settings of π_O^{tr}, we show results on ID ImageNet in Tab. 5. The linear strategy outperforms other approaches in a wide range of π_O^{tr}. However, with $\pi_O^{\text{tr}} \approx 0$, POSCOD will likely fail, as it would be difficult to estimate $p_O(x)$ using $p(z = M \,|\, x)$.

Note that we use the test time ratio of OOD data $\pi_O^{\text{eval}} \approx 1/2$ irrespective of the prior in the training mixture π_O^{tr}. We can justify this, as the optimal SCOD strategy is independent of π_O^{eval}, see Sec. 2.1. The model can therefore be used in settings where π_O^{eval} is not stationary.

Table 5: AuSRT↓ in % points for the POSCOD approximated likelihood ratio when varying the portion of OOD samples π_O^{tr} in the mixture. Results are shown for practically usable plugin double-score strategies on ID ImageNet. The evaluation data is identical in all settings. The best single-score and double-score strategies using contemporary scores are also shown; note that they are selected based on test data performance. Settings where the linear combination of $\hat{r}(x)$ and $\hat{g}(x)$ outperforms all other strategies are marked in bold.

	π_O^{tr} / *Score*	0.1	0.2	0.3	0.4	0.5	S. Score	D. Score
ssb_hard	SIRC	17.36	17.34	16.84	16.80	8.63	17.43	16.01
	Linear	17.20	**13.91**	**10.94**	**9.67**	**5.88**		
ninco	SIRC	13.27	13.26	12.79	13.56	11.16	13.50	11.75
	Linear	12.41	12.68	**10.11**	**10.27**	**7.29**		
inaturalist	SIRC	7.66	8.18	8.59	8.02	6.58	6.95	5.53
	Linear	**4.54**	**4.36**	**4.30**	**4.27**	**4.12**		
textures	SIRC	11.61	10.84	11.84	10.85	8.89	7.03	5.91
	Linear	**5.52**	**4.98**	**4.83**	**4.60**	**4.24**		
openimage_o	SIRC	10.42	10.39	9.84	9.60	8.88	8.85	7.41
	Linear	9.11	7.68	7.67	**7.17**	**6.25**		

D Results on ImageNet-1K

Table 6: The metrics defined in Appendix B.5 shown in % points for selective classifiers constructed from an ImageNet ID classifier $h(x)$ and selectors $c(x) = [\![s(x) \leq \lambda]\!]$ for a representative sample of single-scores $s(x)$. Results are shown for in-distribution ImageNet using a relative cost of $\alpha = 0.5$. The best results are marked in bold, with the second best underlined.

	Dataset	ImageNet	ssb_hard		ninco		inaturalist		textures		openimage_o	
	Score	AuRC↓	AuROC↑	AuSRT↓	AuROC↑	AuSRT↓	AuROC↑	AuSRT↓	AuROC↑	AuSRT↓	AuROC↑	AuSRT↓
Single Score	ASH	11.32	<u>72.88</u>	19.03	**83.44**	<u>13.75</u>	**97.06**	**6.94**	**96.90**	**7.02**	**93.25**	**8.85**
	EBO	11.27	72.07	19.59	79.70	15.78	90.63	10.31	88.70	11.28	89.05	11.10
	GradNorm	19.79	71.89	23.94	74.01	22.88	93.89	12.95	92.04	13.87	84.82	17.48
	L_1-norm	25.02	60.47	32.27	52.89	36.06	46.57	39.22	68.27	28.37	59.30	32.85
	MLS	<u>10.73</u>	72.50	19.11	80.40	15.16	91.16	9.78	88.39	11.17	89.16	10.78
	MSP	**6.95**	72.09	**17.43**	79.95	**13.50**	88.40	9.27	82.43	12.26	84.85	11.04
	ODIN	11.32	71.74	19.79	77.76	16.77	91.16	10.07	89.00	11.15	88.23	11.54
	ReAct	10.79	**73.02**	<u>18.88</u>	<u>81.72</u>	14.53	<u>96.34</u>	<u>7.22</u>	<u>92.78</u>	<u>9.00</u>	<u>91.86</u>	<u>9.46</u>
	Residual	24.75	43.31	40.72	52.80	35.97	50.70	37.02	87.57	18.59	60.90	31.92

Table 7: AuSRT↓ in % points for *tuned* selective classifiers constructed from an ID classifier $h(x)$ and selectors $c(x) = [\![s(x) \leq \lambda]\!]$. Results are shown for ID ImageNet and several possible OOD datasets. Rows of the Table correspond to different scores $s(x)$. The relative cost is $\alpha = 0.5$. The best results per dataset are highlighted in green. The best results with contemporary OODD scores are shown in bold.

	Score \ *Dataset*	ssb_hard	ninco	inaturalist	textures	openimage_o
Single Score	ASH [9]	19.04	13.76	6.95	7.03	8.85
	EBO [25]	19.60	15.78	10.32	11.28	11.11
	GradNorm [16]	23.95	22.89	12.95	13.88	17.49
	L_1-norm [16]	32.27	36.06	39.22	28.37	32.86
	MLS [14]	19.12	15.17	9.79	11.17	10.79
	MSP [15]	17.43	13.50	9.27	12.26	11.05
	ODIN [24]	19.79	16.78	10.08	11.16	11.54
	ReAct [35]	18.88	14.53	7.22	9.00	9.46
	Residual [41]	40.72	35.98	37.03	18.59	31.93
	POSCOD LR	14.25	16.67	12.70	13.19	15.65
	Clean LR	14.14	14.04	11.22	11.75	14.00
Linear (MSP, -)	ASH [9]	16.81	11.86	5.57	**5.91**	7.45
	EBO [25]	17.02	12.95	8.26	9.98	9.42
	GradNorm [16]	17.36	13.43	7.74	9.60	10.39
	L_1-norm [16]	17.19	13.50	9.29	11.88	11.04
	MLS [14]	17.03	12.95	8.25	10.08	9.43
	ODIN [24]	16.96	13.06	7.95	9.54	9.36
	ReAct [35]	16.65	12.26	5.92	7.99	8.06
	Residual [41]	17.42	13.22	8.99	6.65	10.16
	POSCOD LR	5.88	7.29	4.12	4.24	6.25
	Clean LR	5.88	6.67	4.04	4.23	5.82
SIRC (MSP, -)	ASH [9]	16.73	**11.75**	**5.53**	6.09	**7.41**
	EBO [25]	17.01	12.93	8.23	10.01	9.41
	GradNorm [16]	**16.01**	12.33	6.42	7.99	8.78
	L_1-norm [16]	16.88	13.43	9.28	11.58	10.87
	MLS [14]	17.02	12.93	8.23	10.14	9.44
	ODIN [24]	16.96	13.03	7.97	9.76	9.43
	ReAct [35]	16.60	12.23	5.91	8.12	8.06
	Residual [41]	17.41	13.13	8.89	6.22	9.90
	POSCOD LR	8.63	11.16	6.58	8.89	8.88
	Clean LR	6.59	6.61	4.78	7.45	6.21

Table 8: AuSRT↓ in % points for *plugin* selective classifiers constructed from an ID classifier $h(x)$ and selectors $c(x) = [\![s(x) \leq \lambda]\!]$. Results are shown for ID ImageNet and several possible OOD datasets. Rows of the Table correspond to different scores $s(x)$. The relative cost is $\alpha = 0.5$. The best results per dataset are highlighted in green. The best results with contemporary OODD scores are shown in bold.

	Score \ *Dataset*	ssb_hard	ninco	inaturalist	textures	openimage_o
Single Score	ASH [9]	19.04	13.76	6.95	7.03	8.85
	EBO [25]	19.60	15.78	10.32	11.28	11.11
	GradNorm [16]	23.95	22.89	12.95	13.88	17.49
	L_1-norm [16]	32.27	36.06	39.22	28.37	32.86
	MLS [14]	19.12	15.17	9.79	11.17	10.79
	MSP [15]	17.43	13.50	9.27	12.26	11.05
	ODIN [24]	19.79	16.78	10.08	11.16	11.54
	ReAct [35]	18.88	14.53	7.22	9.00	9.46
	Residual [41]	40.72	35.98	37.03	18.59	31.93
	POSCOD LR	14.25	16.67	12.70	13.19	15.65
	Clean LR	14.14	14.04	11.22	11.75	14.00
Linear (MSP, -)	ASH [9]	18.74	13.47	6.82	**6.95**	**8.66**
	EBO [25]	18.86	14.97	9.67	10.88	10.56
	GradNorm [16]	23.92	22.86	12.97	13.91	17.48
	L_1-norm [16]	32.11	35.87	39.01	28.20	32.67
	MLS [14]	18.55	14.56	9.36	10.89	10.40
	ODIN [24]	17.40	13.51	9.27	12.21	11.02
	ReAct [35]	17.85	13.37	**6.76**	8.62	8.77
	Residual [41]	27.47	21.46	18.50	9.43	17.02
	POSCOD LR	6.37	7.44	4.16	4.38	6.48
	Clean LR	6.10	8.05	4.35	4.48	7.38
SIRC (MSP, -)	ASH [9]	17.23	13.03	8.01	10.62	9.98
	EBO [25]	17.33	13.37	9.03	11.83	10.71
	GradNorm [16]	**17.00**	**12.96**	7.70	10.24	9.80
	L_1-norm [16]	17.21	13.54	9.50	11.90	10.96
	MLS [14]	17.34	13.38	9.04	11.86	10.72
	ODIN [24]	17.38	13.45	9.16	12.10	10.92
	ReAct [35]	17.23	13.20	8.46	11.54	10.39
	Residual [41]	17.50	13.43	9.20	9.54	10.62
	POSCOD LR	14.28	12.67	8.03	10.51	9.75
	Clean LR	13.24	10.94	6.38	10.49	8.59

E Results on CIFAR-10

Table 9: The mean and standard deviation over 3-folds of metrics defined in Appendix B.5. The results are shown in % points for selective classifiers constructed from a CIFAR-10 ID classifier $h(x)$ and selectors $c(x) = [\![s(x) \leq \lambda]\!]$ for a representative sample of single-scores $s(x)$. Results are shown for in-distribution CIFAR-10 using a relative cost of $\alpha = 0.5$. The best results are marked in bold, with the second best underlined.

	Dataset	ID cifar10	cifar100		mnist		places365	
	Score	AuRC↓	AuROC↑	AuSRT↓	AuROC↑	AuSRT↓	AuROC↑	AuSRT↓
Single Score	ASH	1.84 ± 0.27	74.11 ± 1.55	13.87 ± 0.89	83.16 ± 4.66	9.34 ± 2.35	79.89 ± 3.69	10.97 ± 1.71
	EBO	0.90 ± 0.10	86.36 ± 0.58	7.27 ± 0.33	94.32 ± 2.53	3.29 ± 1.22	89.25 ± 0.78	5.82 ± 0.35
	GradNorm	3.89 ± 0.38	54.43 ± 1.59	24.73 ± 0.92	63.72 ± 7.37	20.08 ± 3.54	60.50 ± 5.33	21.69 ± 2.47
	KNN	$\mathbf{0.56 \pm 0.04}$	$\mathbf{89.73 \pm 0.14}$	$\mathbf{5.41 \pm 0.09}$	94.26 ± 0.38	3.15 ± 0.17	$\mathbf{91.77 \pm 0.23}$	$\mathbf{4.40 \pm 0.11}$
	MLS	0.89 ± 0.10	86.31 ± 0.59	7.29 ± 0.33	94.15 ± 2.48	3.37 ± 1.19	89.14 ± 0.76	5.87 ± 0.34
	MSP	$\underline{0.60 \pm 0.03}$	87.19 ± 0.33	6.70 ± 0.18	92.63 ± 1.57	3.98 ± 0.77	88.92 ± 0.47	5.84 ± 0.22
	ODIN	1.23 ± 0.22	82.18 ± 1.87	9.53 ± 1.04	$\mathbf{95.24 \pm 1.96}$	$\mathbf{3.00 \pm 0.91}$	85.07 ± 1.24	8.08 ± 0.69
	ReAct	0.86 ± 0.11	85.93 ± 0.83	7.46 ± 0.46	92.81 ± 3.03	4.02 ± 1.49	$\underline{90.35 \pm 0.78}$	$\underline{5.26 \pm 0.34}$
	VIM	0.79 ± 0.05	$\underline{87.75 \pm 0.28}$	$\underline{6.52 \pm 0.16}$	$\underline{94.76 \pm 0.38}$	$\underline{3.02 \pm 0.19}$	89.49 ± 0.39	5.65 ± 0.22

	Dataset	ID cifar10	svhn		textures		tin	
	Score	AuRC↓	AuROC↑	AuSRT↓	AuROC↑	AuSRT↓	AuROC↑	AuSRT↓
Single Score	ASH	1.84 ± 0.27	73.46 ± 6.41	14.19 ± 3.28	77.45 ± 2.39	12.20 ± 1.33	76.44 ± 0.61	12.70 ± 0.44
	EBO	0.90 ± 0.10	91.79 ± 0.98	4.56 ± 0.45	89.47 ± 0.70	5.72 ± 0.40	88.80 ± 0.36	6.05 ± 0.22
	GradNorm	3.89 ± 0.38	53.91 ± 6.36	24.99 ± 3.23	52.07 ± 4.09	25.91 ± 2.23	55.37 ± 0.41	24.26 ± 0.39
	KNN	$\mathbf{0.56 \pm 0.04}$	$\underline{92.67 \pm 0.30}$	$\underline{3.94 \pm 0.13}$	$\underline{93.16 \pm 0.24}$	$\underline{3.70 \pm 0.14}$	$\mathbf{91.56 \pm 0.26}$	$\mathbf{4.50 \pm 0.15}$
	MLS	0.89 ± 0.10	91.69 ± 0.94	4.60 ± 0.43	89.41 ± 0.71	5.74 ± 0.41	88.72 ± 0.36	6.08 ± 0.23
	MSP	$\underline{0.60 \pm 0.03}$	91.46 ± 0.40	4.57 ± 0.19	89.89 ± 0.71	5.36 ± 0.37	88.87 ± 0.19	5.86 ± 0.11
	ODIN	1.23 ± 0.22	84.58 ± 0.77	8.33 ± 0.43	86.94 ± 2.26	7.15 ± 1.24	83.55 ± 1.84	8.84 ± 1.03
	ReAct	0.86 ± 0.11	89.12 ± 3.19	5.87 ± 1.59	89.38 ± 1.49	5.74 ± 0.80	88.29 ± 0.44	6.29 ± 0.28
	VIM	0.79 ± 0.05	$\mathbf{94.51 \pm 0.48}$	$\mathbf{3.14 \pm 0.22}$	$\mathbf{95.16 \pm 0.34}$	$\mathbf{2.82 \pm 0.19}$	$\underline{89.62 \pm 0.33}$	$\underline{5.59 \pm 0.19}$

Table 10: The mean and standard deviation over 3-folds of AuSRT↓ in % points for *tuned* selective classifiers constructed from an ID classifier $h(x)$ and selectors $c(x) = [\![s(x) \leq \lambda]\!]$. Results are shown for ID CIFAR-10 and several possible OOD datasets. Rows of the Table correspond to different scores $s(x)$. The relative cost is $\alpha = 0.5$. The best results per dataset are highlighted in green. The best results with contemporary OODD scores are shown in bold.

	Score \ *Dataset*	cifar100	mnist	places365	svhn	textures	tin
Single Score	ASH [9]	13.87 ± 0.89	9.34 ± 2.35	10.97 ± 1.71	14.19 ± 3.28	12.20 ± 1.33	12.70 ± 0.44
	EBO [25]	7.27 ± 0.33	3.29 ± 1.22	5.82 ± 0.35	4.56 ± 0.45	5.72 ± 0.40	6.05 ± 0.22
	GradNorm [16]	24.73 ± 0.92	20.08 ± 3.54	21.69 ± 2.47	24.99 ± 3.23	25.91 ± 2.23	24.26 ± 0.39
	KNN [36]	5.41 ± 0.09	3.15 ± 0.17	$\mathbf{4.40 \pm 0.11}$	3.94 ± 0.13	3.70 ± 0.14	$\mathbf{4.50 \pm 0.15}$
	MLS [14]	7.29 ± 0.33	3.37 ± 1.19	5.87 ± 0.34	4.60 ± 0.43	5.74 ± 0.41	6.08 ± 0.23
	MSP [15]	6.70 ± 0.18	3.98 ± 0.77	5.84 ± 0.22	4.57 ± 0.19	5.36 ± 0.37	5.86 ± 0.11
	ODIN [24]	9.53 ± 1.04	3.00 ± 0.91	8.08 ± 0.69	8.33 ± 0.43	7.15 ± 1.24	8.84 ± 1.03
	ReAct [35]	7.46 ± 0.46	4.02 ± 1.49	5.26 ± 0.34	5.87 ± 1.59	5.74 ± 0.80	6.29 ± 0.28
	VIM [41]	6.52 ± 0.16	3.02 ± 0.19	5.65 ± 0.22	3.14 ± 0.22	2.82 ± 0.19	5.59 ± 0.19
	POSCOD LR	10.80 ± 0.11	2.17 ± 0.17	7.89 ± 0.95	2.25 ± 0.12	3.97 ± 0.30	2.73 ± 0.07
	Clean LR	4.24 ± 0.18	2.17 ± 0.16	5.99 ± 0.13	2.14 ± 0.13	3.01 ± 0.16	2.31 ± 0.18
Linear (MSP, -)	ASH [9]	6.60 ± 0.20	3.79 ± 0.79	5.63 ± 0.28	4.51 ± 0.26	5.25 ± 0.38	5.75 ± 0.12
	EBO [25]	6.28 ± 0.18	3.03 ± 1.03	5.20 ± 0.27	3.96 ± 0.30	4.84 ± 0.34	5.31 ± 0.12
	GradNorm [16]	6.69 ± 0.18	3.93 ± 0.81	5.80 ± 0.24	4.55 ± 0.22	5.35 ± 0.37	5.85 ± 0.11
	KNN [36]	$\mathbf{5.40 \pm 0.08}$	3.12 ± 0.19	$\mathbf{4.40 \pm 0.11}$	3.91 ± 0.13	3.70 ± 0.14	$\mathbf{4.50 \pm 0.15}$
	MLS [14]	6.31 ± 0.18	3.11 ± 1.01	5.25 ± 0.27	4.01 ± 0.28	4.88 ± 0.34	5.36 ± 0.12
	ODIN [24]	6.52 ± 0.21	$\mathbf{2.67 \pm 0.81}$	5.63 ± 0.22	4.48 ± 0.16	4.80 ± 0.39	5.73 ± 0.14
	ReAct [35]	6.33 ± 0.23	3.31 ± 1.00	5.00 ± 0.22	4.24 ± 0.43	4.86 ± 0.45	5.38 ± 0.17
	VIM [41]	6.15 ± 0.18	2.88 ± 0.11	5.29 ± 0.09	$\mathbf{3.10 \pm 0.21}$	$\mathbf{2.81 \pm 0.19}$	5.28 ± 0.17
	POSCOD LR	5.07 ± 0.10	0.29 ± 0.02	3.75 ± 0.46	0.34 ± 0.02	1.55 ± 0.12	0.98 ± 0.05
	Clean LR	3.02 ± 0.06	0.29 ± 0.02	3.20 ± 0.05	0.30 ± 0.02	1.15 ± 0.11	0.42 ± 0.00
SIRC (MSP, -)	ASH [9]	6.57 ± 0.19	3.77 ± 0.80	5.51 ± 0.39	4.51 ± 0.26	5.24 ± 0.38	5.73 ± 0.11
	EBO [25]	6.30 ± 0.18	3.10 ± 1.01	5.23 ± 0.27	3.99 ± 0.29	4.87 ± 0.34	5.34 ± 0.12
	GradNorm [16]	6.66 ± 0.18	3.89 ± 0.82	5.72 ± 0.29	4.53 ± 0.22	5.33 ± 0.38	5.83 ± 0.11
	KNN [36]	5.94 ± 0.13	3.35 ± 0.40	4.91 ± 0.14	4.16 ± 0.19	4.17 ± 0.24	5.03 ± 0.12
	MLS [14]	6.33 ± 0.19	3.18 ± 0.98	5.29 ± 0.26	4.04 ± 0.27	4.91 ± 0.35	5.38 ± 0.12
	ODIN [24]	6.51 ± 0.21	2.86 ± 0.82	5.63 ± 0.22	4.47 ± 0.16	4.81 ± 0.38	5.73 ± 0.14
	ReAct [35]	6.33 ± 0.23	3.35 ± 0.97	5.08 ± 0.24	4.25 ± 0.41	4.88 ± 0.45	5.40 ± 0.16
	VIM [41]	6.14 ± 0.16	2.92 ± 0.20	5.27 ± 0.02	3.17 ± 0.23	2.93 ± 0.23	5.30 ± 0.14
	POSCOD LR	5.09 ± 0.23	0.29 ± 0.02	5.04 ± 0.18	0.36 ± 0.03	1.81 ± 0.30	1.72 ± 0.28
	Clean LR	3.89 ± 0.61	0.29 ± 0.02	5.40 ± 0.18	0.32 ± 0.02	1.78 ± 0.12	0.66 ± 0.02

Table 11: The mean and standard deviation over 3-folds AuSRT↓ in % points for *plugin* selective classifiers constructed from an ID classifier $h(x)$ and selectors $c(x) = [\![s(x) \leq \lambda]\!]$. Results are shown for ID CIFAR-10 and several possible OOD datasets. Rows of the Table correspond to different scores $s(x)$. The relative cost is $\alpha = 0.5$. The best results per dataset are highlighted in green. The best results with contemporary OODD scores are shown in bold.

	Score \ *Dataset*	cifar100	mnist	places365	svhn	textures	tin
Single Score	ASH [9]	13.87 ± 0.89	9.34 ± 2.35	10.97 ± 1.71	14.19 ± 3.28	12.20 ± 1.33	12.70 ± 0.44
	EBO [25]	7.27 ± 0.33	3.29 ± 1.22	5.82 ± 0.35	4.56 ± 0.45	5.72 ± 0.40	6.05 ± 0.22
	GradNorm [16]	24.73 ± 0.92	20.08 ± 3.54	21.69 ± 2.47	24.99 ± 3.23	25.91 ± 2.23	24.26 ± 0.39
	KNN [36]	$\mathbf{5.41 \pm 0.09}$	3.15 ± 0.17	$\mathbf{4.40 \pm 0.11}$	3.94 ± 0.13	3.70 ± 0.14	$\mathbf{4.50 \pm 0.15}$
	MLS [14]	7.29 ± 0.33	3.37 ± 1.19	5.87 ± 0.34	4.60 ± 0.43	5.74 ± 0.41	6.08 ± 0.23
	MSP [15]	6.70 ± 0.18	3.98 ± 0.77	5.84 ± 0.22	4.57 ± 0.19	5.36 ± 0.37	5.86 ± 0.11
	ODIN [24]	9.53 ± 1.04	$\mathbf{3.00 \pm 0.91}$	8.08 ± 0.69	8.33 ± 0.43	7.15 ± 1.24	8.84 ± 1.03
	ReAct [35]	7.46 ± 0.46	4.02 ± 1.49	5.26 ± 0.34	5.87 ± 1.59	5.74 ± 0.80	6.29 ± 0.28
	VIM [41]	6.52 ± 0.16	3.02 ± 0.19	5.65 ± 0.22	$\mathbf{3.14 \pm 0.22}$	$\mathbf{2.82 \pm 0.19}$	5.59 ± 0.19
	POSCOD LR	10.80 ± 0.11	2.17 ± 0.17	7.89 ± 0.95	2.25 ± 0.12	3.97 ± 0.30	2.73 ± 0.07
	Clean LR	4.24 ± 0.18	2.17 ± 0.16	5.99 ± 0.13	2.14 ± 0.13	3.01 ± 0.16	2.31 ± 0.18
Linear (MSP, -)	ASH [9]	13.60 ± 0.84	9.13 ± 2.30	10.78 ± 1.68	13.69 ± 3.07	11.91 ± 1.30	12.43 ± 0.41
	EBO [25]	7.28 ± 0.33	3.37 ± 1.20	5.87 ± 0.35	4.59 ± 0.44	5.74 ± 0.40	6.08 ± 0.22
	GradNorm [16]	24.71 ± 0.91	20.07 ± 3.53	21.68 ± 2.47	24.97 ± 3.22	25.89 ± 2.22	24.24 ± 0.39
	KNN [36]	5.69 ± 0.09	3.45 ± 0.30	4.79 ± 0.16	4.06 ± 0.10	4.11 ± 0.15	4.84 ± 0.16
	MLS [14]	7.31 ± 0.34	3.45 ± 1.17	5.92 ± 0.34	4.64 ± 0.42	5.77 ± 0.41	6.12 ± 0.23
	ODIN [24]	6.79 ± 0.21	3.87 ± 0.80	5.93 ± 0.23	4.64 ± 0.20	5.32 ± 0.39	5.96 ± 0.12
	ReAct [35]	7.49 ± 0.44	4.11 ± 1.45	5.34 ± 0.34	5.87 ± 1.51	5.79 ± 0.77	6.33 ± 0.26
	VIM [41]	6.49 ± 0.17	$\mathbf{3.00 \pm 0.17}$	5.61 ± 0.21	3.16 ± 0.21	2.86 ± 0.19	5.55 ± 0.18
	POSCOD LR	8.93 ± 0.34	0.30 ± 0.02	4.30 ± 0.38	0.37 ± 0.03	1.73 ± 0.12	1.20 ± 0.08
	Clean LR	3.66 ± 0.12	0.30 ± 0.02	4.16 ± 0.07	0.32 ± 0.02	1.25 ± 0.12	0.48 ± 0.0
SIRC (MSP, -)	ASH [9]	6.70 ± 0.18	3.94 ± 0.80	5.79 ± 0.26	4.61 ± 0.24	5.36 ± 0.39	5.86 ± 0.11
	EBO [25]	6.61 ± 0.18	3.76 ± 0.85	5.70 ± 0.23	4.43 ± 0.20	5.24 ± 0.37	5.74 ± 0.11
	GradNorm [16]	6.76 ± 0.18	3.99 ± 0.81	5.86 ± 0.25	4.62 ± 0.21	5.42 ± 0.40	5.91 ± 0.12
	KNN [36]	6.53 ± 0.18	3.81 ± 0.71	5.61 ± 0.21	4.48 ± 0.20	5.12 ± 0.36	5.67 ± 0.11
	MLS [14]	6.62 ± 0.18	3.77 ± 0.85	5.71 ± 0.23	4.44 ± 0.20	5.24 ± 0.37	5.75 ± 0.11
	ODIN [24]	6.62 ± 0.19	3.49 ± 0.82	5.77 ± 0.22	4.52 ± 0.18	5.08 ± 0.37	5.81 ± 0.12
	ReAct [35]	6.62 ± 0.19	3.81 ± 0.85	5.68 ± 0.22	4.49 ± 0.26	5.24 ± 0.39	5.75 ± 0.12
	VIM [41]	6.57 ± 0.18	3.78 ± 0.67	5.71 ± 0.18	4.29 ± 0.25	4.59 ± 0.39	5.73 ± 0.12
	POSCOD LR	6.02 ± 0.25	0.30 ± 0.02	5.72 ± 0.22	0.44 ± 0.06	2.76 ± 0.37	3.40 ± 0.49
	Clean LR	5.28 ± 0.57	0.30 ± 0.02	5.73 ± 0.2	0.36 ± 0.03	2.83 ± 0.17	1.26 ± 0.07

F Results on CIFAR-100

Table 12: The mean and standard deviation over 3-folds of metrics defined in Appendix B.5. Results are shown in % points for selective classifiers constructed from a CIFAR-100 ID classifier $h(x)$ and selectors $c(x) = [\![s(x) \leq \lambda]\!]$ for a representative sample of single-scores $s(x)$. Results are shown for in-distribution CIFAR-100 using a relative cost of $\alpha = 0.5$. The best results are marked in bold, with the second best underlined.

	Dataset	ID cifar100	cifar10		mnist		places365	
	Score	AuRC↓	AuROC↑	AuSRT↓	AuROC↑	AuSRT↓	AuROC↑	AuSRT↓
Single Score	ASH	8.34 ± 0.20	76.48 ± 0.30	15.93 ± 0.24	77.23 ± 0.46	15.56 ± 0.13	78.76 ± 0.16	14.79 ± 0.05
	EBO	7.83 ± 0.06	$\underline{79.05 \pm 0.11}$	14.39 ± 0.07	79.18 ± 1.37	14.33 ± 0.67	79.52 ± 0.23	14.16 ± 0.13
	GradNorm	14.17 ± 0.18	70.32 ± 0.20	21.92 ± 0.16	65.35 ± 1.12	24.41 ± 0.48	69.69 ± 0.17	22.24 ± 0.14
	KNN	$\underline{7.28 \pm 0.12}$	77.02 ± 0.25	15.13 ± 0.08	$\underline{82.36 \pm 1.52}$	$\underline{12.46 \pm 0.79}$	79.43 ± 0.47	13.93 ± 0.18
	MLS	7.59 ± 0.07	$\mathbf{79.21 \pm 0.10}$	$\underline{14.19 \pm 0.08}$	78.91 ± 1.47	14.34 ± 0.72	$\underline{79.75 \pm 0.24}$	13.92 ± 0.14
	MSP	$\mathbf{6.19 \pm 0.12}$	78.47 ± 0.07	$\mathbf{13.86 \pm 0.09}$	76.08 ± 1.86	15.06 ± 0.93	79.22 ± 0.29	$\mathbf{13.49 \pm 0.16}$
	ODIN	8.13 ± 0.02	78.18 ± 0.14	14.98 ± 0.07	$\mathbf{83.79 \pm 1.31}$	$\mathbf{12.17 \pm 0.64}$	79.45 ± 0.26	14.34 ± 0.14
	ReAct	7.66 ± 0.02	78.65 ± 0.05	14.50 ± 0.02	78.37 ± 1.59	14.65 ± 0.79	$\mathbf{80.03 \pm 0.11}$	$\underline{13.82 \pm 0.06}$
	VIM	8.79 ± 0.18	72.21 ± 0.41	18.29 ± 0.21	81.89 ± 1.02	13.45 ± 0.59	75.85 ± 0.37	16.47 ± 0.10

	Dataset	ID cifar100	svhn		textures		tin	
	Score	AuRC↓	AuROC↑	AuSRT↓	AuROC↑	AuSRT↓	AuROC↑	AuSRT↓
Single Score	ASH	8.34 ± 0.20	$\mathbf{85.60 \pm 1.40}$	$\mathbf{11.37 \pm 0.60}$	80.72 ± 0.70	13.81 ± 0.25	79.92 ± 0.20	14.21 ± 0.17
	EBO	7.83 ± 0.06	82.03 ± 1.74	12.90 ± 0.90	78.35 ± 0.83	14.74 ± 0.45	82.76 ± 0.08	12.54 ± 0.04
	GradNorm	14.17 ± 0.18	76.95 ± 4.73	18.61 ± 2.40	64.58 ± 0.13	24.79 ± 0.15	69.95 ± 0.79	22.11 ± 0.38
	KNN	$\underline{7.28 \pm 0.12}$	$\underline{84.15 \pm 1.09}$	$\underline{11.57 \pm 0.48}$	$\underline{83.66 \pm 0.83}$	$\underline{11.81 \pm 0.44}$	$\mathbf{83.34 \pm 0.16}$	$\mathbf{11.97 \pm 0.07}$
	MLS	7.59 ± 0.07	81.65 ± 1.49	12.97 ± 0.78	78.39 ± 0.84	14.60 ± 0.46	$\underline{82.90 \pm 0.05}$	12.35 ± 0.05
	MSP	$\mathbf{6.19 \pm 0.12}$	78.42 ± 0.89	13.89 ± 0.51	77.32 ± 0.71	14.44 ± 0.41	82.07 ± 0.17	$\underline{12.06 \pm 0.13}$
	ODIN	8.13 ± 0.02	74.54 ± 0.76	16.80 ± 0.39	79.33 ± 1.08	14.40 ± 0.55	81.63 ± 0.08	13.25 ± 0.04
	ReAct	7.66 ± 0.02	83.01 ± 0.97	12.33 ± 0.48	80.15 ± 0.46	13.76 ± 0.23	82.88 ± 0.08	12.39 ± 0.04
	VIM	8.79 ± 0.18	83.14 ± 3.71	12.83 ± 1.77	$\mathbf{85.91 \pm 0.78}$	$\mathbf{11.44 \pm 0.46}$	77.76 ± 0.16	15.52 ± 0.13

Table 13: The mean and standard deviation over 3-folds AuSRT↓ in % points for *tuned* selective classifiers constructed from an ID classifier $h(x)$ and selectors $c(x) = [\![s(x) \leq \lambda]\!]$. Results are shown for ID CIFAR-100 and several possible OOD datasets. Rows of the Table correspond to different scores $s(x)$. The relative cost is $\alpha = 0.5$. The best results per dataset are highlighted in green. The best results with contemporary OODD scores are shown in bold.

	Score \ *Dataset*	cifar10	mnist	places365	svhn	textures	tin
Single Score	ASH [9]	15.93 ± 0.24	15.56 ± 0.13	14.79 ± 0.05	11.37 ± 0.60	13.81 ± 0.25	14.21 ± 0.17
	EBO [25]	14.39 ± 0.07	14.33 ± 0.67	14.16 ± 0.13	12.90 ± 0.90	14.74 ± 0.45	12.54 ± 0.04
	GradNorm [16]	21.92 ± 0.16	24.41 ± 0.48	22.24 ± 0.14	18.61 ± 2.40	24.79 ± 0.15	22.11 ± 0.38
	KNN [36]	15.13 ± 0.08	12.46 ± 0.79	13.93 ± 0.18	11.57 ± 0.48	11.81 ± 0.44	11.97 ± 0.07
	MLS [14]	14.19 ± 0.08	14.34 ± 0.72	13.92 ± 0.14	12.97 ± 0.78	14.60 ± 0.46	12.35 ± 0.05
	MSP [15]	13.86 ± 0.09	15.06 ± 0.93	13.49 ± 0.16	13.89 ± 0.51	14.44 ± 0.41	12.06 ± 0.13
	ODIN [24]	14.98 ± 0.07	12.17 ± 0.64	14.34 ± 0.14	16.80 ± 0.39	14.40 ± 0.55	13.25 ± 0.04
	ReAct [35]	14.50 ± 0.02	14.65 ± 0.79	13.82 ± 0.06	12.33 ± 0.48	13.76 ± 0.23	12.39 ± 0.04
	VIM [41]	18.29 ± 0.21	13.45 ± 0.59	16.47 ± 0.10	12.83 ± 1.77	11.44 ± 0.46	15.52 ± 0.13
	POSCOD LR	19.18 ± 0.18	11.29 ± 0.09	16.66 ± 0.60	11.53 ± 0.20	14.54 ± 0.20	12.28 ± 0.22
	Clean LR	12.10 ± 0.24	11.33 ± 0.10	15.60 ± 0.29	10.98 ± 0.11	12.93 ± 0.16	11.78 ± 0.12
Linear (MSP ,-)	ASH [9]	13.70 ± 0.10	13.96 ± 0.51	13.17 ± 0.10	$\mathbf{10.98 \pm 0.59}$	12.96 ± 0.08	11.90 ± 0.12
	EBO [25]	13.56 ± 0.09	13.99 ± 0.79	13.23 ± 0.14	12.58 ± 0.86	14.03 ± 0.41	11.67 ± 0.06
	GradNorm [16]	13.86 ± 0.09	14.91 ± 0.88	13.48 ± 0.16	13.31 ± 0.96	14.40 ± 0.38	12.06 ± 0.13
	KNN [36]	13.58 ± 0.04	12.31 ± 0.83	$\mathbf{13.03 \pm 0.18}$	11.39 ± 0.44	11.59 ± 0.47	$\mathbf{11.35 \pm 0.11}$
	MLS [14]	13.56 ± 0.09	14.06 ± 0.82	13.23 ± 0.14	12.72 ± 0.77	14.04 ± 0.41	11.68 ± 0.07
	ODIN [24]	13.66 ± 0.09	$\mathbf{11.83 \pm 0.69}$	13.17 ± 0.15	13.85 ± 0.51	13.48 ± 0.48	11.81 ± 0.09
	ReAct [35]	13.60 ± 0.08	14.24 ± 0.86	13.17 ± 0.12	12.09 ± 0.46	13.49 ± 0.24	11.65 ± 0.06
	VIM [41]	13.74 ± 0.05	12.62 ± 0.71	13.12 ± 0.18	11.41 ± 1.11	$\mathbf{10.30 \pm 0.42}$	11.70 ± 0.14
	POSCOD LR	9.62 ± 0.07	3.09 ± 0.06	8.57 ± 0.27	3.20 ± 0.05	6.78 ± 0.23	4.22 ± 0.12
	Clean LR	6.63 ± 0.17	3.08 ± 0.06	7.71 ± 0.22	3.10 ± 0.06	5.73 ± 0.08	3.60 ± 0.16
SIRC (MSP ,-)	ASH [9]	13.68 ± 0.10	13.94 ± 0.50	13.14 ± 0.10	11.15 ± 0.43	12.95 ± 0.05	11.87 ± 0.12
	EBO [25]	$\mathbf{13.54 \pm 0.09}$	14.03 ± 0.84	13.21 ± 0.14	12.66 ± 0.81	14.01 ± 0.42	11.66 ± 0.06
	GradNorm [16]	13.69 ± 0.09	14.74 ± 0.85	13.36 ± 0.15	12.74 ± 1.09	14.24 ± 0.37	11.98 ± 0.11
	KNN [36]	13.55 ± 0.04	12.59 ± 0.91	$\mathbf{13.03 \pm 0.17}$	11.60 ± 0.32	11.88 ± 0.47	$\mathbf{11.35 \pm 0.10}$
	MLS [14]	13.55 ± 0.09	14.12 ± 0.86	13.21 ± 0.14	12.81 ± 0.73	14.03 ± 0.42	11.67 ± 0.07
	ODIN [24]	13.64 ± 0.09	12.31 ± 0.76	13.13 ± 0.15	13.85 ± 0.51	13.57 ± 0.47	11.80 ± 0.09
	ReAct [35]	13.58 ± 0.08	14.24 ± 0.87	13.15 ± 0.12	12.31 ± 0.49	13.59 ± 0.29	11.63 ± 0.06
	VIM [41]	13.70 ± 0.05	12.48 ± 0.73	13.08 ± 0.19	11.38 ± 1.04	10.34 ± 0.41	11.64 ± 0.15
	POSCOD LR	9.41 ± 0.31	3.09 ± 0.06	12.81 ± 0.28	3.21 ± 0.07	7.49 ± 0.81	6.55 ± 0.46
	Clean LR	7.67 ± 0.57	3.09 ± 0.06	13.16 ± 0.26	3.12 ± 0.06	7.20 ± 1.51	4.28 ± 0.50

Table 14: The mean and standard deviation over 3-folds AuSRT↓ in % points for *plugin* selective classifiers constructed from an ID classifier $h(x)$ and selectors $c(x) = [\![s(x) \leq \lambda]\!]$. Results are shown for ID CIFAR-100 and several possible OOD datasets. Rows of the Table correspond to different scores $s(x)$. The relative cost is $\alpha = 0.5$. The best results per dataset are highlighted in green. The best results with contemporary OODD scores are shown in bold.

	Score \ *Dataset*	cifar10	mnist	places365	svhn	textures	tin
Single Score	ASH [9]	15.93 ± 0.24	15.56 ± 0.13	14.79 ± 0.05	$\mathbf{11.37 \pm 0.60}$	13.81 ± 0.25	14.21 ± 0.17
	EBO [25]	14.39 ± 0.07	14.33 ± 0.67	14.16 ± 0.13	12.90 ± 0.90	14.74 ± 0.45	12.54 ± 0.04
	GradNorm [16]	21.92 ± 0.16	24.41 ± 0.48	22.24 ± 0.14	18.61 ± 2.40	24.79 ± 0.15	22.11 ± 0.38
	KNN [36]	15.13 ± 0.08	12.46 ± 0.79	13.93 ± 0.18	11.57 ± 0.48	11.81 ± 0.44	11.97 ± 0.07
	MLS [14]	14.19 ± 0.08	14.34 ± 0.72	13.92 ± 0.14	12.97 ± 0.78	14.60 ± 0.46	12.35 ± 0.05
	MSP [15]	13.86 ± 0.09	15.06 ± 0.93	13.49 ± 0.16	13.89 ± 0.51	14.44 ± 0.41	12.06 ± 0.13
	ODIN [24]	14.98 ± 0.07	12.17 ± 0.64	14.34 ± 0.14	16.80 ± 0.39	14.40 ± 0.55	13.25 ± 0.04
	ReAct [35]	14.50 ± 0.02	14.65 ± 0.79	13.82 ± 0.06	12.33 ± 0.48	13.76 ± 0.23	12.39 ± 0.04
	VIM [41]	18.29 ± 0.21	13.45 ± 0.59	16.47 ± 0.10	12.83 ± 1.77	11.44 ± 0.46	15.52 ± 0.13
	POSCOD LR	19.18 ± 0.18	11.29 ± 0.09	16.66 ± 0.60	11.53 ± 0.20	14.54 ± 0.20	12.28 ± 0.22
	Clean LR	12.10 ± 0.24	11.33 ± 0.10	15.60 ± 0.29	10.98 ± 0.11	12.93 ± 0.16	11.78 ± 0.12
Linear (MSP ,-)	ASH [9]	15.66 ± 0.20	15.41 ± 0.19	14.56 ± 0.05	11.44 ± 0.52	13.74 ± 0.21	13.91 ± 0.15
	EBO [25]	14.18 ± 0.08	14.30 ± 0.71	13.92 ± 0.14	12.91 ± 0.82	14.57 ± 0.45	12.32 ± 0.05
	GradNorm [16]	21.91 ± 0.16	24.40 ± 0.48	22.22 ± 0.13	18.60 ± 2.39	24.78 ± 0.15	22.09 ± 0.37
	KNN [36]	14.20 ± 0.04	$\mathbf{14.12 \pm 0.85}$	13.46 ± 0.17	13.00 ± 0.21	13.36 ± 0.42	$\mathbf{11.81 \pm 0.12}$
	MLS [14]	14.08 ± 0.08	14.35 ± 0.75	13.79 ± 0.14	13.01 ± 0.74	14.51 ± 0.45	12.24 ± 0.06
	ODIN [24]	13.94 ± 0.10	14.95 ± 0.87	13.58 ± 0.18	14.09 ± 0.51	14.43 ± 0.44	12.17 ± 0.14
	ReAct [35]	14.28 ± 0.02	14.60 ± 0.84	13.61 ± 0.07	12.42 ± 0.42	13.69 ± 0.23	12.19 ± 0.02
	VIM [41]	17.32 ± 0.16	13.13 ± 0.61	15.64 ± 0.13	12.32 ± 1.63	$\mathbf{10.99 \pm 0.45}$	14.54 ± 0.13
	POSCOD LR	13.13 ± 0.13	3.10 ± 0.06	11.60 ± 0.37	3.25 ± 0.06	7.12 ± 0.28	4.49 ± 0.21
	Clean LR	7.56 ± 0.20	3.10 ± 0.06	11.27 ± 0.14	3.15 ± 0.06	6.07 ± 0.10	3.75 ± 0.24
SIRC (MSP ,-)	ASH [9]	13.82 ± 0.09	14.85 ± 0.85	13.41 ± 0.15	13.50 ± 0.44	14.21 ± 0.37	12.02 ± 0.12
	EBO [25]	13.82 ± 0.09	14.97 ± 0.93	13.44 ± 0.16	13.77 ± 0.53	14.38 ± 0.41	12.00 ± 0.12
	GradNorm [16]	13.82 ± 0.09	14.95 ± 0.88	13.46 ± 0.16	13.54 ± 0.64	14.39 ± 0.38	12.09 ± 0.11
	KNN [36]	$\mathbf{13.80 \pm 0.08}$	14.75 ± 0.93	13.37 ± 0.17	13.48 ± 0.46	14.01 ± 0.42	11.91 ± 0.13
	MLS [14]	13.82 ± 0.09	14.97 ± 0.93	13.45 ± 0.16	13.78 ± 0.53	14.38 ± 0.41	12.00 ± 0.12
	ODIN [24]	13.84 ± 0.09	14.89 ± 0.93	13.44 ± 0.16	13.93 ± 0.51	14.37 ± 0.41	12.03 ± 0.12
	ReAct [35]	13.81 ± 0.09	14.96 ± 0.92	13.43 ± 0.16	13.73 ± 0.52	14.34 ± 0.40	11.98 ± 0.12
	VIM [41]	13.84 ± 0.08	14.71 ± 0.96	$\mathbf{13.34 \pm 0.17}$	13.48 ± 0.37	13.45 ± 0.42	11.95 ± 0.14
	POSCOD LR	12.36 ± 0.23	3.10 ± 0.06	13.46 ± 0.14	3.39 ± 0.20	11.16 ± 1.14	10.23 ± 0.36
	Clean LR	10.70 ± 0.72	3.10 ± 0.06	13.45 ± 0.16	3.15 ± 0.06	10.44 ± 1.51	6.26 ± 0.92

G Proof of Theorem 1

The proof is essentially identical to that of [11, Theorem 1]. We include it here for the sake of completeness.

For any ID classifier h and a stochastic selector c, the definition of h_B allows to derive $\mathrm{R_S}(h_B, c) \leq \mathrm{R_S}(h, c)$ as follows:

$$\begin{aligned}
\mathrm{R_S}(h_B, c) &= \frac{1}{\mathrm{tpr}(c)} \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y)\, \ell(y, h_B(x))\, c(x)\, dx \\
&= \frac{1}{\mathrm{tpr}(c)} \int_{\mathcal{X}} p(x) c(x) \left(\sum_{y \in \mathcal{Y}} p(y \mid x)\, \ell(y, h_B(x)) \right) dx \\
&\leq \frac{1}{\mathrm{tpr}(c)} \int_{\mathcal{X}} p(x) c(x) \left(\sum_{y \in \mathcal{Y}} p(y \mid x)\, \ell(y, h(x)) \right) dx \\
&= \frac{1}{\mathrm{tpr}(c)} \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y)\, \ell(y, h(x))\, c(x)\, dx \\
&= \mathrm{R_S}(h, c)\,.
\end{aligned}$$

H Proof of Theorem 2

If c^* is an optimal solution to Problem 1 fulfilling $\mathrm{tpr}(c^*) = a \cdot \mathrm{tpr}_{\min}$, where $a > 1$, then $c = c^*/a$ is also an optimal solution such that $\mathrm{tpr}(c) = \mathrm{tpr}_{\min}$. Hence, to find an optimal solution to Problem 1, it suffices to minimize the objective function

$$\begin{aligned}
&\frac{1-\alpha}{\mathrm{tpr}_{\min}} \int_{\mathcal{X}} R(h_B, x) c(x) dx + \alpha \int_{\mathcal{X}} p_O(x) c(x) dx \\
&= \int_{\mathcal{X}} \left[\frac{1-\alpha}{\mathrm{tpr}_{\min}} R(h_B, x) + \alpha p_O(x) \right] c(x) dx
\end{aligned}$$

subject to

$$\int_{\mathcal{X}} p_I(x) c(x) \geq \mathrm{tpr}_{\min}\,.$$

By [11, Theorem 3], for a bounded risk function $f : \mathcal{X} \to \mathbb{R}_+$, a probability distribution $p : \mathcal{X} \to \mathbb{R}_+$, and $b \in \mathbb{R}_+$, whenever the problem

$$\min_{c \in [0,1]^{\mathcal{X}}} \int_{\mathcal{X}} f(x) c(x) dx \qquad \text{s.t.} \qquad \int_{\mathcal{X}} p(x) c(x) \geq b$$

is feasible, the set of its optimal solutions contains

$$c^*(x) = \begin{cases} 0 \text{ if } \frac{f(x)}{p(x)} > \lambda \\ \tau \text{ if } \frac{f(x)}{p(x)} = \lambda \\ 1 \text{ if } \frac{f(x)}{p(x)} < \lambda \end{cases}$$

for suitable $\tau \in [0,1]$ and $\lambda \in \mathbb{R}$. This implies that Problem 1 has an optimal solution c^* induced by the score function $s(x) = \frac{1-\alpha}{\mathrm{tpr}_{\min}} r(x) + \alpha g(x)$ and the threshold value λ. It is further easy to see that the score $s(x)$ can be replaced by $s'(x) = r(x) + \frac{\alpha \mathrm{tpr}_{\min}}{1-\alpha} g(x)$ if the threshold is set to $\frac{\mathrm{tpr}_{\min}}{1-\alpha}\lambda$.

I Proof of Theorem 3

It was proved in [10] that the problem

$$\min_{h \in \mathcal{Y}^{\mathcal{X}}, c \in [0,1]^{\mathcal{X}}} R_C(h,c) \tag{15}$$

where

$$R_C(h,c) = \int_{\mathcal{X}} \Big[(1-\pi_O) R(h,x) c(x) + L_{FN}(1-\pi_O) p_I(x)(1-c(x)) \\ + L_{FP} \pi_O p_O(x) c(x) \Big] dx$$

is not PAC learnable for any triple of constants $\pi_O \in (0,1)$, $L_{FN} > 0$, $L_{FP} > 0$.

Observe that

$$R_C(h,c) = L_{FN}(1-\pi_O) + \int_{\mathcal{X}} G(h,x) c(x) dx$$

where

$$G(h,x) = (1-\pi_O) R(h,x) - L_{FN}(1-\pi_O) p_I(x) + L_{FP} \pi_O p_O(x)\,.$$

An optimal solution (h_B, c^*) to Problem 15 can thus be established by prescribing $c^*(x) = 1$ whenever $G(h_B, x) < 0$, and $c^*(x) = 0$ whenever $G(h_B, x) \geq 0$. This means that the considered optimal solution is equivalently determined by a score function $s_1(x) = r(x) + \frac{\pi_O L_{FP}}{1-\pi_O} g(x)$ and a threshold $\lambda_1 = L_{FN}$.

By Theorem 2, there are optimal solutions to Problem 1 determined by the score function $s_2(x) = r(x) + \frac{\alpha \mathrm{tpr}_{\min}}{1-\alpha} g(x)$ and a threshold $\lambda_2 > 0$. Hence, for a given instance of Problem 1 with an optimal solution (h_B, c^*) consistent with the score function s_2, we obtain an instance of Problem 15 for which (h_B, c^*) is also an optimal solution just by setting $\pi_O := \alpha$, $L_{FP} := \mathrm{tpr}_{\min}$, and $L_{FN} := \lambda_2$. In accordance with the proof of Theorem 2, we assume that $\mathrm{tpr}(c^*) = \mathrm{tpr}_{\min}$.

Let us now consider we have a solution (h,c) to Problem 1 such that

$$R_S(h,c) - R_S(h_B, c^*) \leq \varepsilon_1\,, \tag{16}$$

$$\mathrm{tpr}(c) \geq \mathrm{tpr}_{\min} - \varepsilon_2 \tag{17}$$

for some $\varepsilon_1, \varepsilon_2 \in \mathbb{R}_+$.

Our goal is to show that

$$R_C(h,c) - R_C(h_B, c^*) \leq \varepsilon_1 \mathrm{tpr}(c) + \alpha \varepsilon_2 \mathrm{fpr}(c)\,. \tag{18}$$

As the right-hand side expression can be made arbitrarily close to zero by choosing sufficiently small values of ε_1 and ε_2, inequality (18) enforces that Problem 1 is not PAC learnable, otherwise Problem 15 would be PAC learnable as well, which is impossible due to [10, Theorem 4].

Inequality (16) implies

$$\frac{1-\alpha}{\mathrm{tpr}(c)}\int_{\mathcal{X}} R(h,x)c(x)dx - \frac{1-\alpha}{\mathrm{tpr}_{\min}}\int_{\mathcal{X}} R(h_B,c^*)c^*(x)dx + \alpha\,\mathrm{fpr}(c) - \alpha\,\mathrm{fpr}(c^*) \leq \varepsilon_1$$

which can further be rewritten to

$$\begin{aligned}
&(1-\alpha)\int_{\mathcal{X}} R(h,x)c(x)dx - (1-\alpha)\int_{\mathcal{X}} R(h_B,c^*)c^*(x)dx && (19)\\
&\leq \alpha\,\mathrm{tpr}(c)\,(\mathrm{fpr}(c^*) - \mathrm{fpr}(c)) && (20)\\
&\qquad + (1-\alpha)\frac{\mathrm{tpr}(c) - \mathrm{tpr}_{\min}}{\mathrm{tpr}_{\min}}\int_{\mathcal{X}} R(h_B,x)c^*(x)dx + \varepsilon_1\mathrm{tpr}(c)\,.
\end{aligned}$$

The score function s_2 ensures that

$$R(h_B,x)c^*(x) + \frac{\alpha}{1-\alpha}\mathrm{tpr}_{\min}p_O(x)c^*(x) \leq \lambda_2 p_I(x)c^*(x)$$

for all $x \in \mathcal{X}$, yielding

$$\int_{\mathcal{X}} R(h_B,x)c^*(x)dx + \frac{\alpha}{1-\alpha}\mathrm{tpr}_{\min}\mathrm{fpr}(c^*) \leq \lambda_2\mathrm{tpr}(c^*) = \lambda_2\mathrm{tpr}_{\min}\,. \tag{21}$$

Now, using (17), (19) and (21), we derive

$$\begin{aligned}
R_C(h,c) - R_C(h_B,c^*) &= (1-\alpha)\int_{\mathcal{X}} R(h,x)c(x)dx\\
&\quad - (1-\alpha)\int_{\mathcal{X}} R(h_B,c^*)c^*(x)dx - \lambda_2(1-\alpha)\mathrm{tpr}(c) + \lambda_2(1-\alpha)\mathrm{tpr}_{\min}\\
&\quad + \alpha\,\mathrm{tpr}_{\min}\mathrm{fpr}(c) - \alpha\,\mathrm{tpr}_{\min}\mathrm{fpr}(c^*)\\
&\leq \varepsilon_1\mathrm{tpr}(c) + (1-\alpha)\frac{\mathrm{tpr}(c) - \mathrm{tpr}_{\min}}{\mathrm{tpr}_{\min}}\left(\lambda_2\mathrm{tpr}_{\min} - \frac{\alpha}{1-\alpha}\mathrm{tpr}_{\min}\mathrm{fpr}(c^*)\right)\\
&\quad + \alpha\,\mathrm{tpr}(c)(\mathrm{fpr}(c^*) - \mathrm{fpr}(c)) - \lambda_2(1-\alpha)(\mathrm{tpr}(c) - \mathrm{tpr}_{\min})\\
&\quad - \alpha\,\mathrm{tpr}_{\min}(\mathrm{fpr}(c^*) - \mathrm{fpr}(c))\\
&= \varepsilon_1\mathrm{tpr}(c) + \lambda_2(1-\alpha)(\mathrm{tpr}(c) - \mathrm{tpr}_{\min}) + \alpha\,\mathrm{fpr}(c^*)(\mathrm{tpr}_{\min} - \mathrm{tpr}(c))\\
&\quad - \lambda_2(1-\alpha)(\mathrm{tpr}(c) - \mathrm{tpr}_{\min}) - \alpha(\mathrm{fpr}(c^*) - \mathrm{fpr}(c))(\mathrm{tpr}_{\min} - \mathrm{tpr}(c))\\
&= \varepsilon_1\mathrm{tpr}(c) + \alpha\,\mathrm{fpr}(c)(\mathrm{tpr}(c) - \mathrm{tpr}_{\min}) \leq \varepsilon_1\mathrm{tpr}(c) + \alpha\varepsilon_2\mathrm{fpr}(c)\,.
\end{aligned}$$

J Proof of Theorem 4

Assume there exists a map $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^d$ that renders the ID and ODD data normal distributed, i.e.,

$$\begin{aligned}
p_I(x) &= (2\pi)^{-\frac{d}{2}}\det(\boldsymbol{C})^{-\frac{1}{2}}\exp\big(-\tfrac{1}{2}(\boldsymbol{\phi}(x) - \boldsymbol{\mu}_I)^T\boldsymbol{C}^{-1}(\boldsymbol{\phi}(x) - \boldsymbol{\mu}_I)\big)\\
p_O(x) &= (2\pi)^{-\frac{d}{2}}\det(\boldsymbol{C})^{-\frac{1}{2}}\exp\big(-\tfrac{1}{2}(\boldsymbol{\phi}(x) - \boldsymbol{\mu}_O)^T\boldsymbol{C}^{-1}(\boldsymbol{\phi}(x) - \boldsymbol{\mu}_O)\big)
\end{aligned} \tag{22}$$

where $(\boldsymbol{\mu}_I, \boldsymbol{C})$ and $(\boldsymbol{\mu}_O, \boldsymbol{C})$ are mean and covariance matrix of the ID and OOD, respectively. Under assumption Eq. (22), the OOD/ID likelihood ratio reads

$$\frac{p_O(x)}{p_I(x)} = \exp\big(\frac{1}{2}(\boldsymbol{\mu}_O^T \boldsymbol{C}^{-1} \boldsymbol{\mu}_O - \boldsymbol{\mu}_I^T \boldsymbol{C}^{-1} \boldsymbol{\mu}_I) + \boldsymbol{\phi}(x)^T \boldsymbol{C}^{-1}\big(\boldsymbol{\mu}_I - \boldsymbol{\mu}_O)\big)\,. \tag{23}$$

From (10), which defines the mixture of ID and unlabeled mixture of ID and OOD, we can derive that

$$\begin{aligned} p(z = I \mid x) &= \frac{p(x, z = I)}{p(x, z = I) + p(x, z = U)} \\ &= \frac{(1 - \pi_U) p_I(x)}{(1 - \pi_U) p_I(x) + \pi_U \pi_O p_O(x) + (1 - \pi_O) \pi_U p_I(x)} \\ &= \frac{1}{1 + \dfrac{\pi_U \pi_O}{(1 - \pi_U)} \dfrac{p_O(x)}{p_I(x)} + \pi_U \dfrac{(1 - \pi_O)}{(1 - \pi_U)}}\,. \end{aligned} \tag{24}$$

After substituting Eq. (23) to Eq. (24), we obtain

$$\begin{aligned} (z = I \mid x; \boldsymbol{\theta}, a) &= \frac{1}{1 + \pi_U \dfrac{(1 - \pi_O)}{(1 - \pi_U)} + \exp\Bigg(\ln\Big(\dfrac{\pi_U \pi_O}{(1 - \pi_U)}\Big) \\ &\qquad + \tfrac{1}{2}(\boldsymbol{\mu}_O^T \boldsymbol{C}^{-1} \boldsymbol{\mu}_O - \boldsymbol{\mu}_I^T \boldsymbol{C}^{-1} \boldsymbol{\mu}_I) + \boldsymbol{\phi}(x)^T \boldsymbol{C}^{-1}\big(\boldsymbol{\mu}_I - \boldsymbol{\mu}_O\big)\Bigg)} \\ &= \frac{1}{1 + |a| + \exp\big(\boldsymbol{\theta}^T [\boldsymbol{\phi}(x); 1]\big)} \end{aligned} \tag{25}$$

where

$$\boldsymbol{\theta} = \left[\boldsymbol{C}^{-1}\big(\boldsymbol{\mu}_I - \boldsymbol{\mu}_O\big); \ln\Big(\frac{\pi_U \pi_O}{(1 - \pi_U)}\Big) + \frac{1}{2}(\boldsymbol{\mu}_O^T \boldsymbol{C}^{-1} \boldsymbol{\mu}_O - \boldsymbol{\mu}_I^T \boldsymbol{C}^{-1} \boldsymbol{\mu}_I)\right]$$

and

$$a = \pi_U \frac{(1 - \pi_O)}{(1 - \pi_U)}\,.$$

Note that we use $\boldsymbol{u} = [\boldsymbol{v}; b]$ to denote a column vector obtained by extending the vector $\boldsymbol{v}$ by a new coordinate b. This ends the proof.