Design of Adaptive Steganographic Schemes for Digital Images

Tomáš Filler and Jessica Fridrich

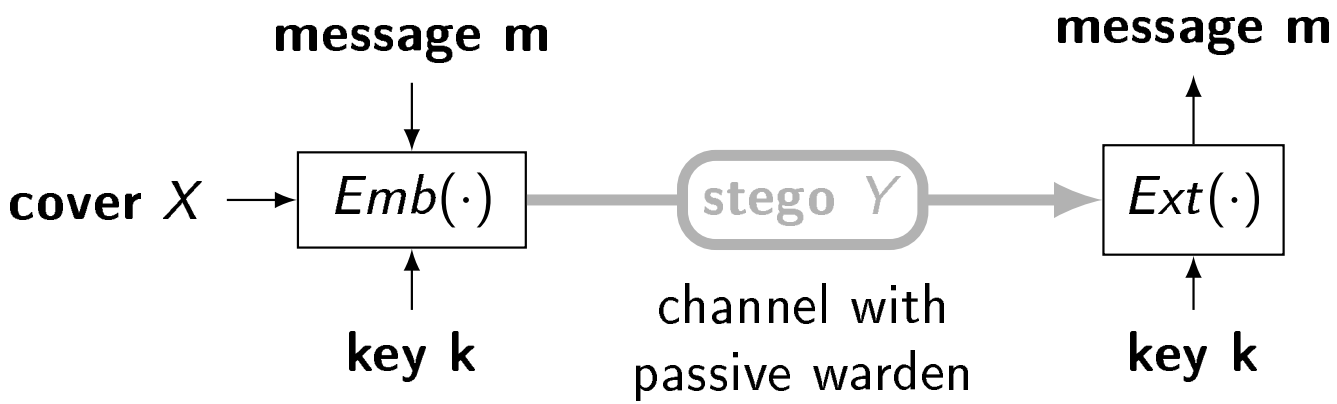
Dept. of Electrical and Computer Engineering
SUNY Binghamton, New York

IS&T / SPIE 2011, San Jose, CA

BINGHAMTON UNIVERSITY
State University of New York

Steganography

Steganography is a mode of covert communication.

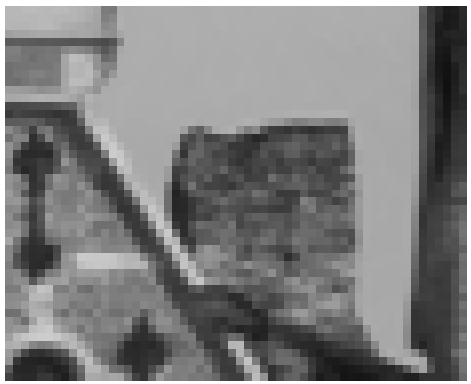


X and Y are r.v. on $\mathcal{X}^n$ — digital images for example
$Emb(\cdot)$, $Ext(\cdot)$ ... embedding, extraction functions

Steganography by cover modification:
Stego object Y is produced by slightly modifying some of the elements (pixels, DCT coefficients, ...) in X.

Which pixels can be changed and how often?

Pixels in hard-to-model content.



Do not change saturated pixels!

Minimal-distortion ± 1 Embedding

Pixels in textured areas can be changed more frequently than those in smooth areas.

± 1 embedding operation: $y_i \in \{x_i - 1, x_i, x_i + 1\}$

Each pixel (DCT element) can be changed by -1, 0, or $+1$.

Additive distortion funct.: $\rho_i(x, y_i) =$ **cost of changing** $x_i \longrightarrow y_i$

cost of changing cover x to stego y $\longrightarrow D(x,y) = \sum_{i=1}^{n} \rho_i(x, y_i)$

Example:

- $\rho_i(x, x_i) = 0$ **and** $\rho_i(x_i - 1, x) = \rho_i(x_i + 1, x) = 1$ **# of changes**
- $\rho_i(x, y_i) \gg 1$ **if** y_i **should almost never be used for pixel** i

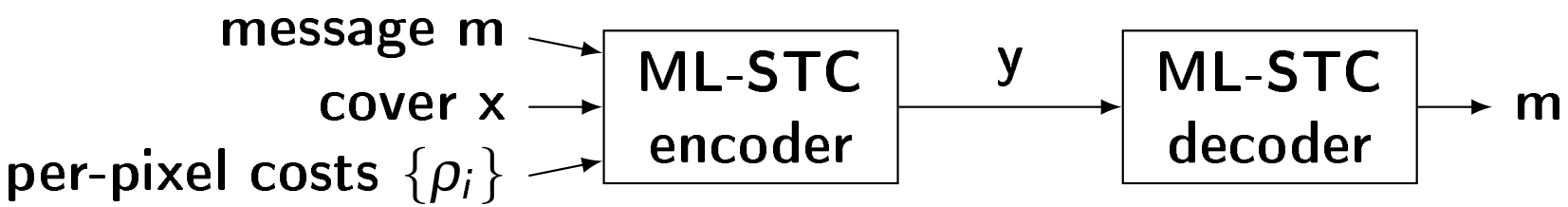
Problem Fomulation for FIXED Cover x

Send m bits in stego y such that $D(\mathbf{x},\mathbf{y})$ is as small as possible.

SOLVED

How to do this in practice for arbitrary distortion?

Use Multi-Layered Syndrome-Trellis Codes [WIFS & SPIE 2010].



ML-STCs utilize $\approx 90\%$ of the theoretical bandwidth.
No need to share cover or costs with the receiver.

How to define $D(\mathbf{x},\mathbf{y}) \Rightarrow$ less detectable steganography?

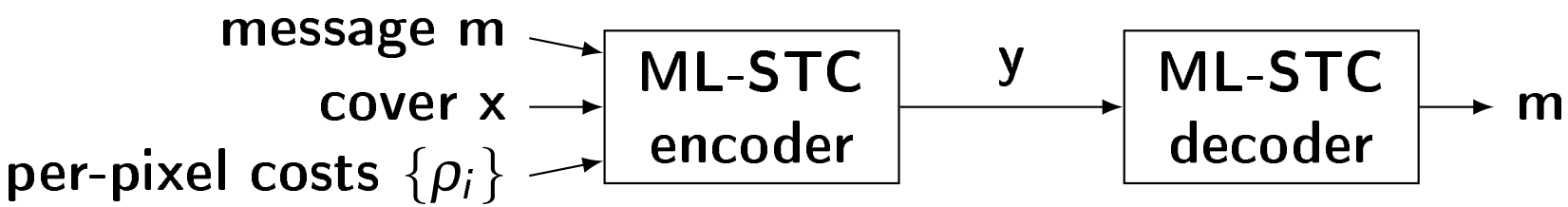
Problem Fomulation for FIXED Cover x

Send m bits in stego y such that $D(\mathbf{x},\mathbf{y})$ is as small as possible.

SOLVED

How to do this in practice for arbitrary distortion?

Use Multi-Layered Syndrome-Trellis Codes [WIFS & SPIE 2010].



ML-STCs utilize $\approx 90\%$ of the theoretical bandwidth.
No need to share cover or costs with the receiver.

MAIN CONTRIBUTION: practical algorithm for designing D w.r.t. a given cover source and steganalytic features.

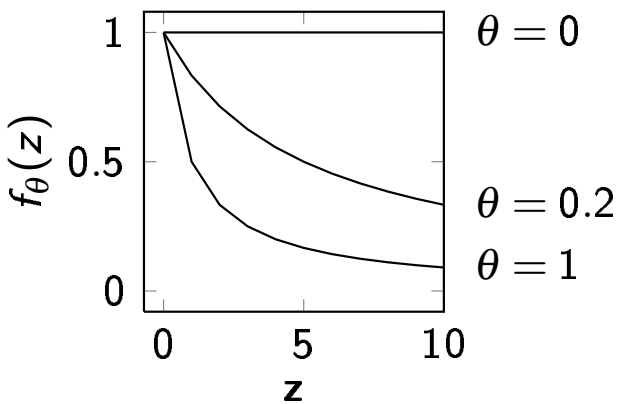
Distortion Function Design

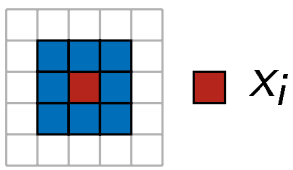
1. parametrize single-pixel distortions $\{\rho_i\}$ with $\theta \in \mathbb{R}^d$
2. find parameter value giving the least detectable method

Example: grayscale spatial-domain imgs, $D(\mathbf{x}, \mathbf{y}) = \sum_i \rho_i(\mathbf{x}, y_i)$

(1) cost of disturbing px. differ. z (2) distortion

$$f_\theta(z) = 1/(1 + \theta|z|)$$





$$\rho_i(\mathbf{x}, x_i) = 0$$

$$\rho_i(\mathbf{x}, y_i) = \sum_{\blacksquare} f_\theta(x_i - \blacksquare) + f_\theta(y_i - \blacksquare)$$

What $\theta \geq 0$ gives us the least detectable method?

Detectability Criterion: Blind Steganalysis

How to evaluate detectability of a given embedding method?

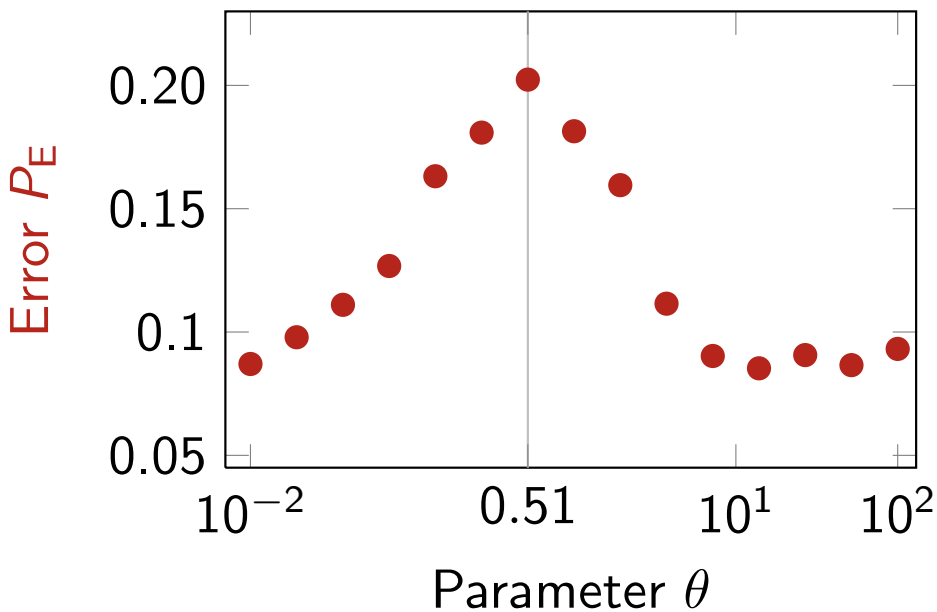
1. embed chosen payload into a large database of images
2. extract features from cover & stego images
3. train machine-learning based classifier (Gaussian SVM)
4. report chosen error metric, such as

$$P_{\mathrm{E}} = \min_{P_{\mathrm{FA}}} \frac{1}{2}(P_{\mathrm{FA}} + P_{\mathrm{MD}})$$

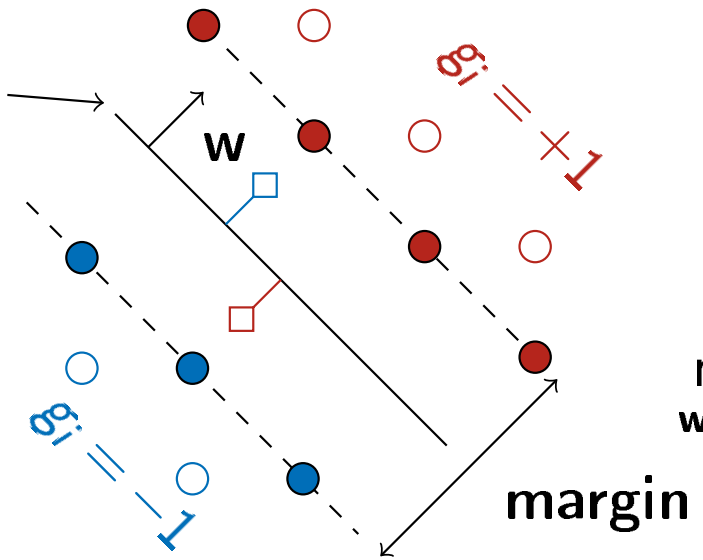
✓ trustworthy measure

✗ very slow
- requires many images
- slow to train classifier

0.5 bpp, CDF set, 512×512 images



Detectability Criterion: Margin of Linear SVM



support vectors

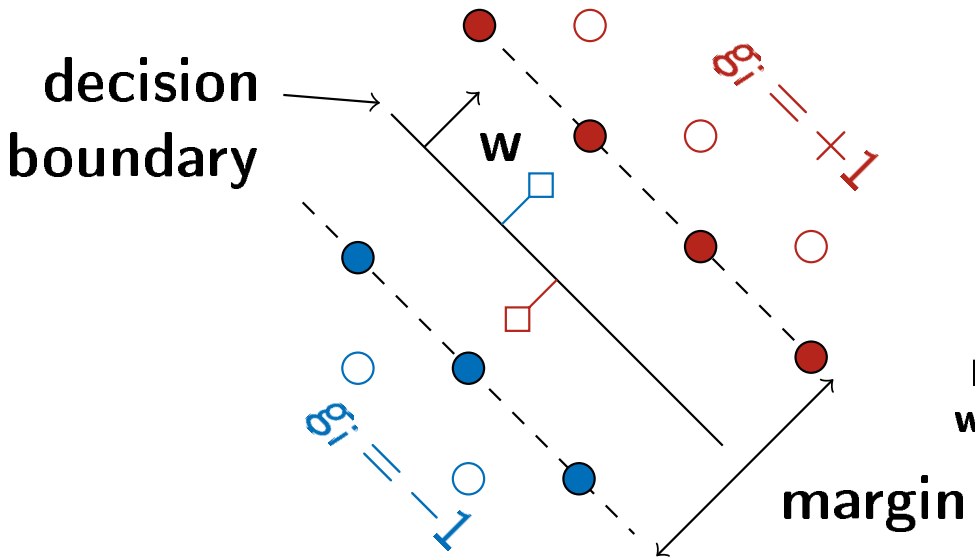
misclassified samples

SVM training problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi(\mathbf{w};\mathbf{f}_i,g_i)$$

regularization term

Detectability Criterion: Margin of Linear SVM



support vectors

misclassified samples

SVM training problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi(\mathbf{w};\mathbf{f}_i,g_i)$$

regularization term

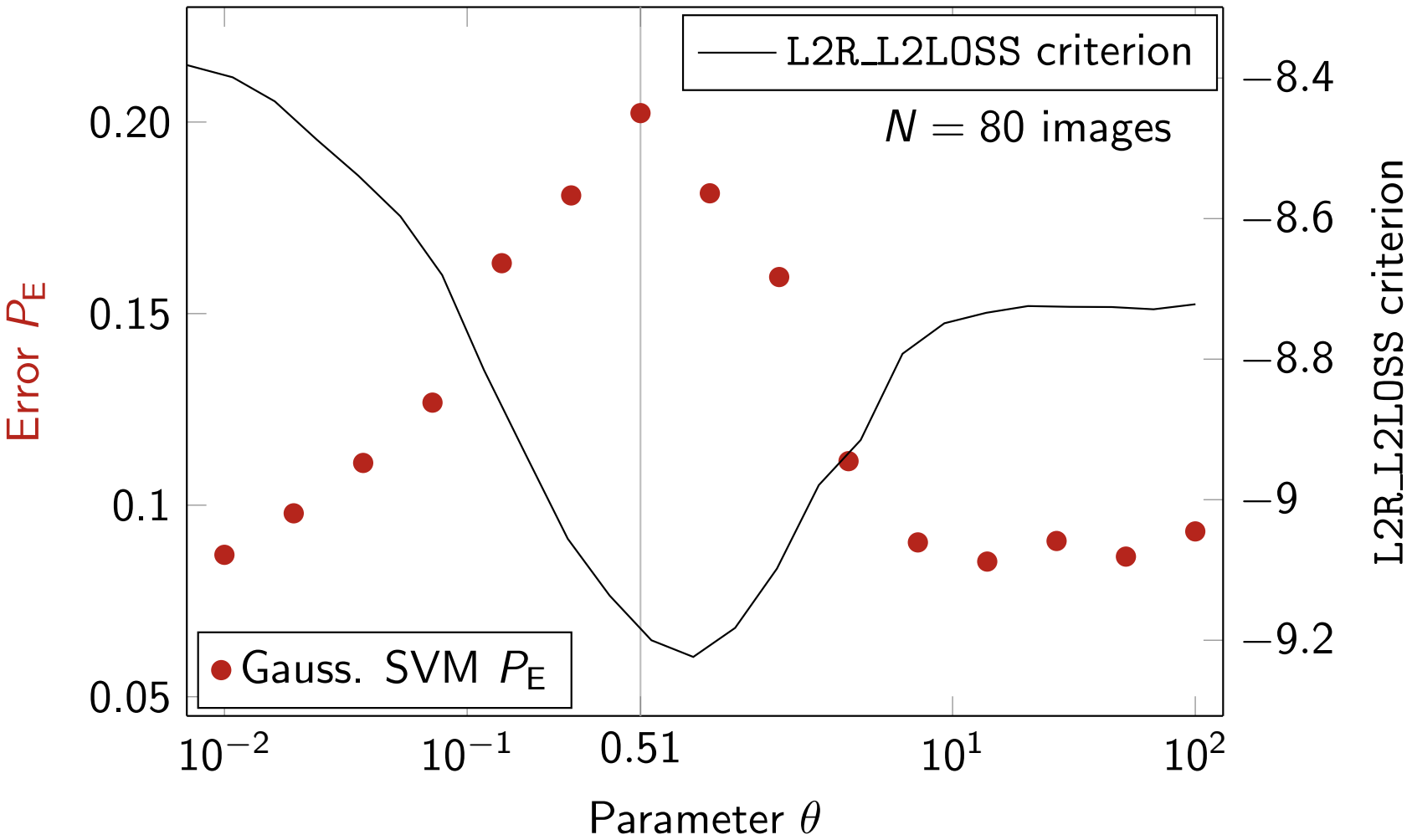
`L2R_L2LOSS` **detectability criterion:**

1. embed chosen payload in N images and extract features
2. `L2R_L2LOSS` = size of the margin from soft-margin L_2-regularized L_2-loss linear SVM

We used $N = 80$ images to evaluate `L2R_L2LOSS`.
Takes $1-2$ seconds on 40CPU cluster for 512×512 images.

Minimizing L2R_L2LOSS ⇒ better undetectability

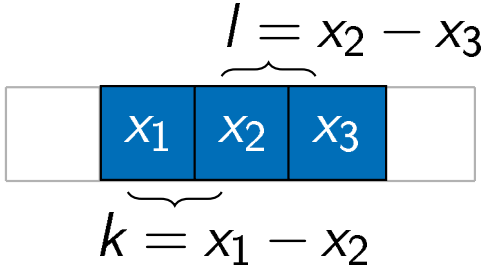
0.5 bpp, CDF set, 512×512 images from BOWS2 database



Further Application to Spatial-Domain Images

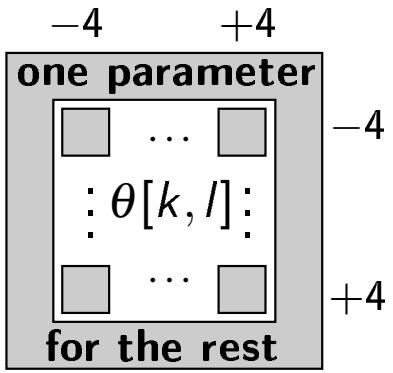
2-difference distortion model: $(2 \cdot 4 + 1)^2 + 1 = 82$ **parameters**

$\rho_i(\mathbf{x}, y_i)$ **is derived from model parameters** $\theta[k, l]$ **for all 3-pixel lines containing** x_i**.**



Distortion of keeping the cover pixel $\rho_i(\mathbf{x}, x_i) = 0$

lookup table

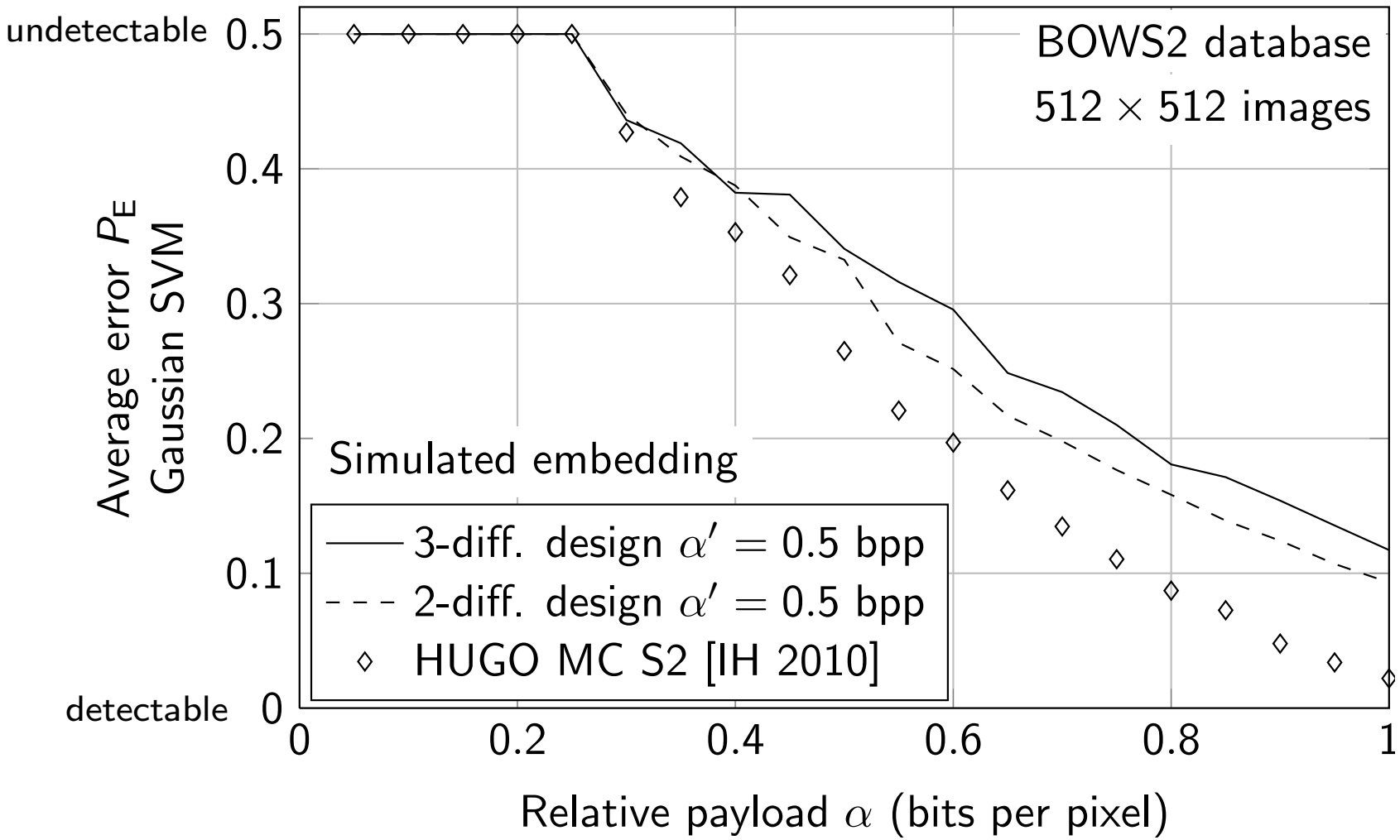


3-difference distortion model: $(2 \cdot 4 + 1)^3 + 1 = 730$ **parameters**

Similar, uses differences between pixels in all 4-pixel lines.

`L2R_L2LOSS` **was minimized using derivative-free Nelder-Mead simplex-reflection algorithm.**

Detectability of Optimized Distortion Models



Optimized w.r.t. SPAM features and tested using CDF set.

Application to DCT-Domain Images

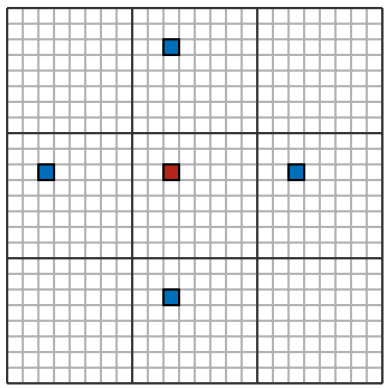
- ± 1 **embedding operation**
- **modify all (even zero) AC DCT coefficients**

Inter-block distortion model: $(2 \cdot 6 + 1) + 1 = 14$ **parameters**

lookup table $\boxed{\theta[\bullet]}\ \boxed{\theta[-6]}\ \boxed{\theta[-5]} \cdots \boxed{\theta[+5]}\ \boxed{\theta[+6]}\ \boxed{\theta[\bullet]}$

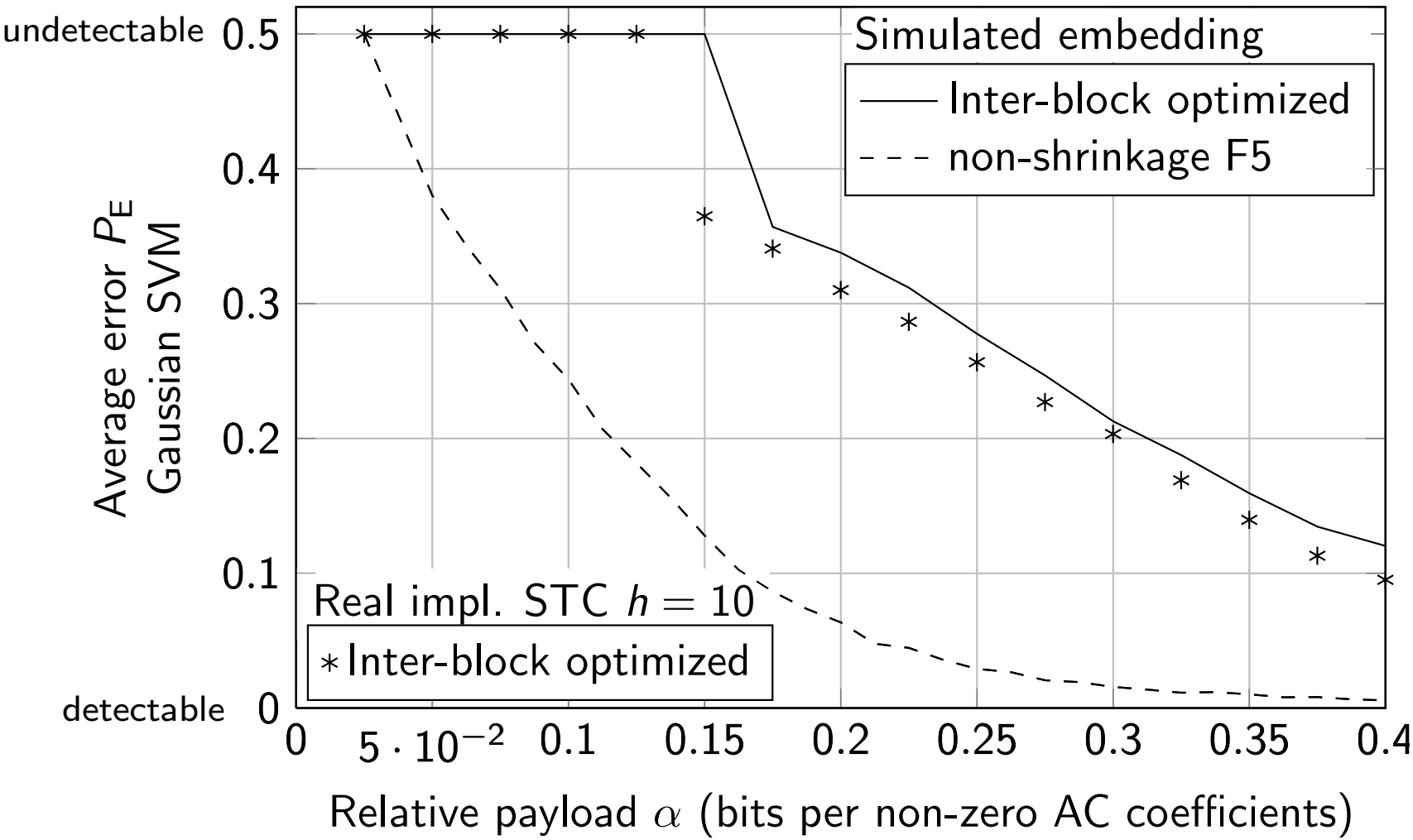
$$\rho_i(\mathbf{x}, y_i) = \begin{cases} 0 & y_i = x_i \\ \sum_{\blacksquare} \theta[\blacksquare - \blacksquare]^2 & \end{cases}$$

$\theta[\blacksquare - \blacksquare] = \theta[\bullet]$ **if** $|\blacksquare - \blacksquare| \geq 7$.



Minimize `L2R_L2LOSS` **using the Nelder-Mead algorithm with Cartesian-Calibrated Pevný features.**

Detectability of Optimized Distortion Models



Optimized w.r.t. CC-PEV features and tested with CDF set.

Conclusion

Minimum-distortion steganography - embed m bits by minimizing distortion between cover and stego.

ML-STCs allow implementing such algorithms in practice.

Proposed technique allows optimizing the distortion funct. by minimizing the margin between feature vectors.

- **strong connection to feature-based steganalysis**
- **fast evaluation allows numerical optimization (80 imgs)**
- **embedding algorithms are NOT over-trained to specific features.**

Simple inter-block-optimized distortion function doubles the payload w.r.t. nsF5 in JPEGs.