Analysis of Digital Watermarks Subjected to Optimum Linear Filtering and Additive Noise

Jonathan K. Su[*]
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420-9185
USA

Joachim J. Eggers
Telecommunications Laboratory
University of Erlangen-Nuremberg
Cauerstrasse 7/NT
D-91058 Erlangen
Germany

Bernd Girod
Information Systems Laboratory
Stanford University
350 Serra Mall
Stanford, CA 94305-9510
USA

Keywords

Digital watermarking; Attacks on digital watermarks; Capacity and robustness of digital watermarks; Information theory

Paper Information

- Paper accepted to appear in *Signal Processing*, Special Issue on Information Theoretic Issues in Digital Watermarking

- Manuscript length: 53 pages (includes 3 pages with separate figure captions)

[*] J. K. Su was with the Telecommunications Laboratory at the University of Erlangen-Nuremberg. He is now with MIT Lincoln Laboratory.

- Abstracts: English and German provided
- Tables: None
- Figures: 15
- Galley proofs: Please send galley proofs to J. K. Su at the MIT Lincoln Laboratory address given above. Telephone: +1-781-981-4188, fax: +1-781-981-7271, email: su@ll.mit.edu.

Abstract

Using a theoretical approach based on random processes, signal processing, and information theory, we study the performance of digital watermarks subjected to an attack consisting of linear shift-invariant filtering and additive colored Gaussian noise. Watermarking is viewed as communication over a hostile channel, where the attack takes place. The attacker attempts to minimize the channel capacity under a constraint on the *attack distortion* (distortion of the attacked signal), and the owner attempts to maximize the capacity under a constraint on the *embedding distortion* (distortion of the watermarked signal). The distortion measure is frequency-weighted mean-squared error (MSE). In a conventional additive-noise channel, communication is most difficult when the noise is white and Gaussian, so we first investigate an effective white-noise attack based on this principle. We then consider the problem of resisting this attack and show that capacity is maximized when a *power-spectrum condition* (PSC) is fulfilled. The PSC states that the power spectrum of the watermark should be directly proportional to that of the original signal. However, unlike a conventional channel, the hostile attack channel adapts to the watermark, not vice versa. Hence, the effective white-noise attack is suboptimal. We derive the optimum attack, which minimizes the channel capacity for a given attack distortion. The attack can be roughly characterized by a rule-of-thumb: At low attack distortions, it adds noise, and at high attack distortions, it discards frequency components. Against the optimum attack, the PSC does not maximize capacity at all attack distortions. Also, there is no unique watermark power spectrum that maximizes capacity over the entire range of attack distortions. To find the watermark power spectrum that maximizes capacity against the optimum attack, we apply iterative numerical methods, which alternately adjust the watermark power spectrum and re-optimize the parameters of the optimum attack. Experiments using ordinary MSE distortion lead to a rule-of-thumb: White watermarks perform nearly optimally at low attack distortions, while PSC-compliant watermarks perform nearly optimally at high attack distortions. The effect of interference from the original signal in suboptimal blind watermarking schemes is also considered; experiments examine its influence on the optimized watermark power spectra and the potential increase in capacity when it can be partially suppressed. Additional experiments demonstrate the importance of memory, and compare the optimum attack with suboptimal attack models. Finally, the rule-of-thumb for the defense is extended to the case of frequency-weighted MSE as a distortion measure.

Mit Hilfe eines theoretischen Ansatzes basierend auf stochastischer Signalverarbeitung und Informationstheorie untersuchen wir die Leistungsfähigkeit von digitalen Wasserzeichen im Fall eines Angriffs durch lineare zeitinvariante Filterung und additives, farbiges, gaußverteiltes Rauschen. Die Wasserzeichenproblematik wird als Kommunikation über einen böswilligen Kanal betrachtet, wobei der Kanal den Angriff beinhaltet. Der Angreifer versucht bei begrenzter *Angriffsverzerrung* (Verzerrung des Signals nach dem Angriff) die Kanalkapazität zu minimieren. Der Einbetter versucht bei begrenzter *Einbettungsverzerrung* (Verzerrung des mit einem Wasserzeichen markierten Signals) die Kanalkapazität zu maximieren. Das Verzerrungsmaß ist der frequenzabhängig gewichtete mittlere quadratische Fehler. Kommunikation über einen konventionellen Kanal mit additivem Rauschen ist dann am schwierigsten, wenn das Rauschen weiß und gaußverteilt ist. Hierauf basierend untersuchen wir zunächst einen Angriff mit effektivem weißen Rauschen. Wir betrachten das Problem, wie einem solchen Angriff widerstanden werden kann, und zeigen, dass die Kapazität maximiert wird, wenn eine bestimmte Bedingung (PSC) an das Leistungsdichtespektrum erfüllt ist. Die PSC besagt, dass das Leistungsdichtespektrum des Wasserzeichens direkt proportional zu dem des Originalsignals sein sollte. Allerdings, im Gegensatz zum konventionellen Kanal, passt sich der böswillige Kanal dem Wasserzeichen an und nicht umgekehrt. Daher ist der Angriff mit effektivem weißen Rauschen suboptimal. Wir leiten den optimalen Angriff her, welcher die Kanalkapazität für eine vorgegebene Angriffsverzerrung minimiert. Der Angriff kann grob mit einer Daumenregel charakterisiert werden: Bei geringen Angriffsverzerrungen wird Rauschen addiert und bei starken Angriffsverzerrungen werden komplette Frequenzkomponenten ausgelöscht. Im Fall des optimalen Angriffs maximiert die PSC die Kapazität nicht für alle Stärken der Angriffsverzerrung. Es existiert kein einheitliches Wasserzeichenleistungsdichtespektrum, welches die Kapazität für den gesamten Bereich von Angriffsstärken maximiert. Im Fall eines optimalen Angriffs verwenden wir zur Bestimmung des Wasserzeichenleistungsdichtespektrums mit maximaler Kapazität eine iterative numerische Methode, welche abwechselnd das Wasserzeichenleistungsdichtespektrum anpasst und die Parameter des optimalen Angriffs wieder optimiert. Experimente mit gewöhnlichem MSE-Verzerrungsmaß führten zu der Daumenregel: Weiße Wasserzeichen sind nahezu optimal bei geringen Angriffsverzerrungen, während PSC-angepasste Wasserzeichen nahezu optimal bei starken Angriffsverzerrungen sind. Der Einfluss von Originalsignalinterferenz in suboptimalen blinden Wasserzeichenverfahren wird ebenfalls betrachtet; Experimente zeigen den Originalsignaleinfluss auf das optimierte Wasserzeichenleistungsdichtespektrum und den potentiellen Kapazitätsgewinn, wenn die Originalsignalinterferenz teilweise unterdrückt werden kann. Zusätzliche Experimente zeigen die Bedeutung des Gedächtnisses bei Angriffen und vergleichen den optimalen Angriff mit suboptimalen Angriffsmodellen. Die Daumenregel für die Wahl des Wasserzeichenleistungsdichtespektrums wird abschließend erweitert für den Fall eines frequenzabhängig gewichteten Verzerrungsmaßes.

Nomenclature

ACGN	additive, colored Gaussian noise
AR	autoregressive
AR(p)	p-th order autoregressive process
AWGN	additive, white Gaussian noise
GA	"greedy" annealing
GMA	greedy marginal analysis
IID	independent, identically distributed
LSI	linear, shift-invariant
M-D	M-dimensional
MAP	maximum *a posteriori*
MMSE	minimum mean-squared error
MSE	mean-squared error
RV	random variable
PSC	power-spectrum condition
SA	simulated annealing
WGN	white Gaussian noise
$*$	1-D or M-D convolution
a, $0 \leq a \leq 1$	original-interference suppression factor
a_k, $k \in \{1, 2, \ldots, p\}$	coefficients of 1-D AR(p) process
$A(\vec{\omega})$, $0 \leq A(\vec{\omega}) \leq 1$	freq.-dependent scaling factor in optimum attack
C	channel capacity (or maximum achievable rate for $0 \leq a < 1$)
$C(\lambda)$	channel capacity as a function of λ
$C_{p\%}$	channel capacity or maximum achievable rate when $p\%$ of original-interference power is suppressed
C_t	target channel capacity
$C_{\max}$, $C_{\min}$	maximum and minimum channel capacities after attack
cl$[x]$	clipping function (clip x to interval $[0, 1]$)
D_{embed}	maximum allowable embedding distortion

D_t	target attack distortion
D_{yx}	embedding distortion
$D_{\hat{y}x}$	attack distortion
$D_{\hat{y}x}(\lambda)$	attack distortion as a function of λ
$D_{\hat{y}x,\max}$, $D_{\hat{y}x,\min}$	maximum and minimum attack distortions
$\delta[\vec{n}]$	M-D unit point-sample function
$\mathbf{e}[n] = \hat{\mathbf{y}}[n] - \mathbf{x}[n]$	1-D error or difference signal (random process)
$\Phi_{ee}(\omega)$	1-D power spectrum of $\mathbf{e}[n]$
$F_{ee}(\omega)$	1-D freq.-weighted error power spectrum
$f[n]$, $f[\vec{n}]$	1-D, M-D impulse responses of LSI freq.-weighting filters
$F(\omega)$, $F(\vec{\omega})$	1-D, M-D Fourier transforms of $f[n]$ and $f[\vec{n}]$, resp.
$g[n]$, $g[\vec{n}]$	1-D, M-D impulse response of attack filter
$g^{-1}[\vec{n}]$	M-D impulse response of inverse attack filter
$G(\omega)$, $G(\vec{\omega})$	1-D, M-D transfer functions of attack filter
Ω_G	frequency support of $G(\vec{\omega})$
ω_0	cutoff frequency of $G(\omega)$
$h[\vec{n}]$	M-D impulse response of ideal whitening filter
$H(\vec{\omega})$	M-D transfer function of ideal whitening filter
$\lambda, \lambda_{\min}, \lambda_{\max}$	Lagrange multiplier and its minimum and maximum values
λ^*	solution for λ such that $C(\lambda^*) = C_t$ or $D_{\hat{y}x}(\lambda^*) = D_t$
M	dimensionality
N	number of equal-support subsets used to cover Ω
$n, \vec{n} = (n_1, n_2, \ldots, n_M)$	1-D, M-D time/space indices
$\omega, \vec{\omega} = (\omega_1, \omega_2, \ldots, \omega_M)$	1-D, M-D frequency variables
$\Omega = [-\pi, \pi)^M$	M-D baseband frequency support
$\mathbf{v}[\vec{n}]$	M-D Gaussian noise
$\mathbf{w}[n]$, $\mathbf{w}[\vec{n}]$	watermark (1-D or M-D random process)
σ_w^2	variance of $\mathbf{w}[n]$ or $\mathbf{w}[\vec{n}]$
$\Phi_{ww}(\omega)$, $\Phi_{ww}(\vec{\omega})$	1-D and M-D power spectra of $\mathbf{w}[n]$ and $\mathbf{w}[\vec{n}]$, resp.

$\mathcal{W} = \{\vec{\omega} : \Phi_{ww}(\vec{\omega}) > 0\}$	frequency support of $\Phi_{ww}(\vec{\omega})$
$F_{ww}(\omega)$	1-D freq.-weighted watermark power spectrum
$F'_{ww}(\omega)$	1-D freq.-weighted, PSC-compliant watermark power spectrum
$\mathbf{x}[n]$, $\mathbf{x}[\vec{n}]$	original signal (1-D or M-D random process)
P_x	perceptual power of $\mathbf{x}[n]$ or $\mathbf{x}[\vec{n}]$
$\mathbf{y}[\vec{n}]$	watermarked signal (M-D random process)
$\hat{\mathbf{y}}[\vec{n}]$	attacked signal (M-D random process)
$\mathbf{z}[\vec{n}]$	effective received signal (M-D random process)
$\mathbf{n}_e[\vec{n}]$	effective noise (M-D random process)
$\mathbf{w}_e[\vec{n}]$	effective watermark (M-D random process)
$\mathbf{z}_i[\vec{n}]$	effective inverse-filtered received signal (M-D random process)
$\mathbf{n}_i[\vec{n}]$	effective inverse-filtered noise (M-D random process)
$\mathbf{z}_w[\vec{n}]$	whitened, effective received signal (M-D random process)
$\mathbf{n}_w[\vec{n}]$	whitened, effective noise (M-D random process)
$\mathbf{w}_w[\vec{n}]$	whitened, effective watermark (M-D random process)
F_n, G_n, V_n, W_n, X_n	piecewise-constant approximations of $F(\vec{\omega})$, $G(\vec{\omega})$, $\Phi_{vv}(\vec{\omega})$, $\Phi_{ww}(\vec{\omega})$, and $\Phi_{xx}(\vec{\omega})$, resp., over nth subset of Ω

1 Introduction

Digital watermarking may be described as the secure, imperceptible, robust communication of information by direct embedding in and retrieval from digital data, typically multimedia data such as digital audio, images [15], or video [22, 30]. Potential applications include tracing the distribution path of watermarked data, multimedia annotation, detection of modifications, and copyright protection [49, 23].

Security indicates that only authorized parties should be able to retrieve, and possibly alter, the embedded information. *Imperceptibility* means that the watermarked data should be perceptually equivalent to the original, unwatermarked data (sometimes called “host data” or “cover data”). In some applications this requirement can be relaxed to “unobtrusiveness,” meaning that small perceptible differences between the watermarked and original data can be tolerated. *Robustness* means that it should be possible to retrieve the embedded information reliably even after processing of the watermarked data[1]; any such processing is known as an *attack*. Attacks may be coincidental, such as compression of a legally-obtained, watermarked

[1] A different class of watermarks, known as *fragile watermarks*, are designed to fail in a prescribed manner after mild processing of watermarked data. Data authentication is a primary application area of fragile watermarks, which are not considered here.

audio file or image, or malicious, such as an attempt by a multimedia pirate to destroy the embedded information and prevent tracing of illegal copies of watermarked digital video.

Often, the design of robust watermarking schemes has been motivated by heuristics and intuition. Various authors have argued that watermarks should be embedded in different frequency ranges, e.g., lowpass, bandpass, highpass, or white watermarks, without reaching a consensus. In early spread-spectrum watermarking schemes (e.g., [42, 22]), white-noise watermarks were employed by direct extension from spread spectrum communications and by the idea that robustness would be enhanced by distributing the watermark over all frequencies. In an image watermarking context, Cox *et al.* [15] were among the first to propose embedding in the "perceptually significant frequency components" of the original image. They justified this position by pointing out that these components facilitate perceptual masking [49] and that an attacker cannot alter these components without also severely degrading the watermarked image. Other authors (e.g., [54, 55]) used high-frequency watermarks, which are easier to separate from the typically lowpass original signal. Still others (e.g., [28, 36]) believed that lowpass watermarks would introduce unacceptable embedding distortion and highpass watermarks would be susceptible to attack; as a compromise these authors advocated the used of bandpass watermarks.

Most of the early (and current) work in watermarking has been applied, with robustness and imperceptibility evaluated experimentally [32]. Many attacks consist of additive noise, compression (e.g., MP3 for audio, JPEG for images, and MPEG-2 for video), or geometric transformations such as rotation, shifting, and scaling [35].

Recently, more theoretical approaches have attempted to provide watermarking, and the larger field of information hiding, with a stronger foundation [44, 16, 27, 47, 10, 33, 51, 31, 34, 8, 48]. Of particular note, Moulin and O'Sullivan [34] have introduced a powerful information-theoretic framework for studying watermarking. They cast the problem as a game between the owner and the attacker. The owner's goal is to send and receive as much information as possible, while the attacker's goal is to hinder communication.

This paper focuses on the conflicting requirements of imperceptibility and robustness and takes a theoretical approach based on random processes, signal processing, and information theory. We do not treat the issue of security here; we assume that proper crytographic methods and protocols are used to maintain key security. The intuitive notion of robustness can be stated as follows: "A watermark is robust if communication of the embedded information cannot be impaired without also rendering the attacked data useless." Hence, to evaluate robustness, we must pose two questions simultaneously: "When is communication impaired?" and "When is the attacked data useless?" The first question suggests that we measure the capacity or a related quantity; the latter suggests that we measure the perceptual quality, or distortion, of the attacked data. In addition, to ensure imperceptibility, we should measure the distortion of the watermarked data after watermark embedding.

In the spirit of [34], we consider the conflicting goals of the attack and owner. The attacker wishes to minimize the communication rate while keeping the distortion of the attacked data small enough so that it remains useful, while the owner wishes to maximize the communication rate while keeping the distortion of the watermarked signal acceptably low. We apply Kerckhoff's principle [43] for both the owner and attacker and assume that the attacker knows the owner's methods, and vice versa. We emphasize the use of a well-defined criterion for evaluating robustness, since otherwise it is difficult to compare the utility of different watermarking methods.

Sec. 2 introduces notation, a mathematical model for the attack and defense, and expressions for distortion and capacity. Sec. 3 derives the optimum attack and shows that there may not be a unique defense. Sec. 4 demonstrates the difficulty of finding a defense and describes some numerical methods for computing the defense. Finally, Sec. 5 summarizes the main conclusions and discusses the practical implications of this study.

2 Mathematical Models

We treat the data as a discrete-time/space signal and in turn model signals as ergodic, zero-mean, wide-sense stationary, M-dimensional (M-D) discrete-time/space Gaussian random processes. Indexing of an M-D signal x is denoted by $x[\vec{n}]$, where $\vec{n} = (n_1, n_2, \ldots, n_M)$. Similarly, the M-D Fourier transform is given by $X(\vec{\omega})$ with $\vec{\omega} = (\omega_1, \omega_2, \ldots, \omega_M)$. Throughout this paper, we consider only the baseband frequency support $\Omega = [-\pi, \pi)^M$, with the M-D 2π-periodicity understood. Boldface indicates random quantities, such as $\mathbf{x}[\vec{n}]$. We ignore quantization effects due to digitization of signal values and assume infinite precision.

The original signal is modeled by the random process $\mathbf{x}[\vec{n}]$ with variance σ_x^2 and power spectrum $\Phi_{xx}(\vec{\omega})$. Likewise, the embedded watermark is represented by the random process $\mathbf{w}[\vec{n}]$ and has variance σ_w^2 and power spectrum $\Phi_{ww}(\vec{\omega})$. The original $\mathbf{x}[\vec{n}]$ and watermark $\mathbf{w}[\vec{n}]$ are assumed independent. Denote the frequency supports of $\Phi_{xx}(\vec{\omega})$ and $\Phi_{ww}(\vec{\omega})$, respectively, by $\mathcal{X} = \{\vec{\omega} : \Phi_{xx}(\vec{\omega}) > 0\}$ and $\mathcal{W} = \{\vec{\omega} : \Phi_{ww}(\vec{\omega}) > 0\}$.

Although these assumptions are ideal, most watermarking applications deal with multimedia, which can often be modeled as being locally stationary and Gaussian. For example, samples in flat image regions may be treated as realizations of independent, identically distributed (IID) Gaussian random variables (RVs) with a low variance, and samples in textured regions are treated as realizations of IID Gaussian RVs with a high variance.

2.1 Watermark Embedding and Attack

A block diagram of the embedding model and attack appears in Fig. 1. We discuss the components of this diagram in this section. We first model the embedding of watermark $\mathbf{w}[\vec{n}]$ into the original $\mathbf{x}[\vec{n}]$ by simple addition; the watermarked signal is $\mathbf{y}[\vec{n}]$,

$$\mathbf{y}[\vec{n}] = \mathbf{x}[\vec{n}] + \mathbf{w}[\vec{n}], \tag{1}$$

where $\mathbf{x}[\vec{n}]$ and $\mathbf{w}[\vec{n}]$ are assumed independent.

Next, we model the attack. Given $\mathbf{y}[\vec{n}]$, the attacker produces an attacked signal $\hat{\mathbf{y}}[\vec{n}]$. We assume that the attacker employs linear shift-invariant (LSI) filtering and additive colored Gaussian noise (ACGN). Let $g[\vec{n}]$ and $G(\vec{\omega})$, respectively, denote the impulse response and transfer function of the attack filter. Let $\mathbf{v}[\vec{n}]$ denote Gaussian noise that has variance σ_v^2 and power spectrum $\Phi_{vv}(\vec{\omega})$ and is independent of $\mathbf{x}[\vec{n}]$ and $\mathbf{w}[\vec{n}]$. The attacked signal $\hat{\mathbf{y}}[\vec{n}]$ is

$$\hat{\mathbf{y}}[\vec{n}] = g[\vec{n}] * \mathbf{y}[\vec{n}] + \mathbf{v}[\vec{n}] = g[\vec{n}] * (\mathbf{x}[\vec{n}] + \mathbf{w}[\vec{n}]) + \mathbf{v}[\vec{n}]. \tag{2}$$

Applying Kerckhoff's principle, the attacker is assumed to know $\Phi_{xx}(\vec{\omega})$ and $\Phi_{ww}(\vec{\omega})$, and hence, to have complete knowledge of the statistics of $\mathbf{x}[\vec{n}]$ and $\mathbf{w}[\vec{n}]$. In Sec. 3 we explain how the attacker exploits this knowledge and derive the optimum attack.

The attack model (2) is ideal. Gaussian noise is a common channel model and is frequently used to approximate synchronous signal degradations. For example, the noise could model distortions introduced after printing and scanning of a watermarked image that has been re-aligned. Also, many lossy compression schemes operate in the frequency domain; they discard lower-amplitude frequency components and quantize higher-amplitude frequency components. Hence, Eq. (2) can also approximate compression as a combination of frequency-selective filtering and additive (quantization) noise.

In addition, it is not unreasonable to expect that an attacker might use filtering (because of its simple implementation) or Gaussian noise (since in an additive-noise chanel with limited noise variance, communication is most difficult when the noise is Gaussian [40]). Finally, just as a real-world original signal may be modeled as being locally stationary, the attack model can represent locally stationary processing of the watermarked signal.

Recall that the frequency supports of the watermark and original are $\mathcal{W}$ and $\mathcal{X}$, respectively. Clearly $\mathcal{W}$ should be a subset of $\mathcal{X}$, for otherwise the attacker could filter out the portion of the watermark in $(\mathcal{W} - \mathcal{X})$

without introducing any distortion. As will be shown in Sec. 4.3.1, the optimum watermark power spectrum has $\mathcal{W} = \mathcal{X}$ to resist the optimum attack.

2.2 Watermark Reception

Finally we consider retrieval of the information carried by the watermark. Given the attacked signal $\hat{\mathbf{y}}[\vec{n}]$, the receiver attempts to determine the information conveyed by $\mathbf{w}[\vec{n}]$. Applying Kerckhoff's principle, we assume that the receiver has knowledge of $g[\vec{n}]$. This assumption is highly ideal, but it allows us to determine performance limits: The receiver will perform best if it has exact knowledge of $g[\vec{n}]$; less accurate knowledge of $g[\vec{n}]$ can only degrade performance. Hence, the results we find can be interpreted as upper bounds on communication performance.

Depending on the design of the watermarking system, the original $\mathbf{x}[\vec{n}]$ may interfere with reception. Consider two extreme scenarios: *reception-with-original* and *blind reception*. In the first scenario, the receiver has access to $\mathbf{x}[\vec{n}]$; then it can eliminate interference from $\mathbf{x}[\vec{n}]$ by computing $\mathbf{z}_1[\vec{n}] = \hat{\mathbf{y}}[\vec{n}] - g[\vec{n}] * \mathbf{x}[\vec{n}] = g[\vec{n}] * \mathbf{w}[\vec{n}] + \mathbf{v}[\vec{n}]$ and then working with $\mathbf{z}_1[\vec{n}]$. In the second scenario, the receiver has no knowledge of $\mathbf{x}[\vec{n}]$, which acts like an additional source of interference; then $\mathbf{z}_0[\vec{n}] = \hat{\mathbf{y}}[\vec{n}]$.

For the case of a memoryless Gaussian original and the additive white Gaussian noise (AWGN) channel, Chen and Wornell [10] have applied the work of Costa [13] to show that the theoretical capacity of an optimal blind receiver is actually equal to the capacity of the receiver-with-original. This surprising result occurs because the original $\mathbf{x}[\vec{n}]$ is known during watermark embedding [16], so the problem is that of communication with side information ($\mathbf{x}[\vec{n}]$) at the encoder but not at the decoder [41, 21, 13, 26]. Rather than attempting to suppress interference from $\mathbf{x}[\vec{n}]$, the communication system employs channel codes designed with the statistics of $\mathbf{x}[\vec{n}]$ in mind [13, 12, 45].

A few blind watermarking systems based on Costa's result have been proposed [16, 9, 38, 12, 11, 19]. Notably, Chen and Wornell [8] proposed a system that asymptotically approaches capacity. Chou *et al.* [11] recognized the duality between the blind watermarking problem and that of lossy source coding with side information at the decoder but not at the encoder, and they have applied recent results in distributed (lossy) source coding to blind watermarking. A further discussion of this duality appears in [45, 46]. Despite these developments, some interference from the original may be unavoidable in practice; one difficulty is that the codebook can become very large, greatly increasing the complexity of a real system [34, 19].

For these reasons, we introduce an *original-interference suppression factor* a, $0 \leq a \leq 1$, and assume that the *effective received signal* is $\mathbf{z}[\vec{n}]$,

$$\mathbf{z}[\vec{n}] = \hat{\mathbf{y}}[\vec{n}] - ag[\vec{n}] * \mathbf{x}[\vec{n}] = g[\vec{n}] * \mathbf{w}[\vec{n}] + (1-a)g[\vec{n}] * \mathbf{x}[\vec{n}] + \mathbf{v}[\vec{n}]. \tag{3}$$

Note that $\mathbf{x}[\vec{n}]$ may not actually be available to the receiver; Eq. (3) expresses that the watermarking system performs as if it operated on $\mathbf{z}[\vec{n}]$. The case $a = 0$ corresponds to blind reception if the watermarking system does not exploit the knowledge of $\mathbf{x}[\vec{n}]$ available during embedding; we use the term *"conventional blind reception"* to refer to this case.[2] The case $a = 1$ corresponds to reception-with-original and to an optimal blind receiver. If the complexity of a suboptimal blind watermarking scheme can be related to a, it is then possible to evaluate the performance-complexity trade-offs of the scheme.

Finally, in our model, we always assume synchronization between the embedding and retrieval units. As explained in the introduction, some current attacks operate by disrupting synchronization. Such attacks do not actually remove or destroy the watermark, so a more sophisticated receiver should be able to resynchronize [24, 17, 7, 6, 18]. Hence, we assume synchronization throughout this paper.

2.3 Distortion Expressions

In watermarking, the distortion of various signals must also be considered. As a compromise between perceptual relevance and mathematical tractability, we measure distortion using frequency-weighted mean-squared error (MSE). For a signal $\hat{\mathbf{x}}[\vec{n}]$ and a reference signal $\mathbf{x}[\vec{n}]$, define the distortion between $\hat{\mathbf{x}}[\vec{n}]$ and $\mathbf{x}[\vec{n}]$ by $D_{\hat{x}x} = \mathrm{E}\left[\left[f[\vec{n}] * (\hat{\mathbf{x}}[\vec{n}] - \mathbf{x}[\vec{n}])\right]^2\right]$, where $f[\vec{n}]$ is the impulse response of a LSI frequency-weighting filter. Letting $\tilde{\mathbf{x}}[\vec{n}] = \hat{\mathbf{x}}[\vec{n}] - \mathbf{x}[\vec{n}]$, we can write $D_{\hat{x}x} = (2\pi)^{-M} \int_\Omega |F(\vec{\omega})|^2 \Phi_{\tilde{x}\tilde{x}}(\vec{\omega})\, d\vec{\omega}$, where we assume $|F(\vec{\omega})| > 0$, $\forall\vec{\omega}$. Of course, $|F(\vec{\omega})| = 1, \forall\vec{\omega}$, is ordinary MSE distortion. Note that $|F(\vec{\omega})|$ could be made dependent upon $\Phi_{xx}(\vec{\omega})$ to approximate some perceptual masking effects [49, 52].

We are interested in the *embedding distortion* D_{yx} and the *attack distortion* $D_{\hat{y}x}$. From (1), the former is simply

$$D_{yx} = \frac{1}{(2\pi)^M} \int_\Omega |F(\vec{\omega})|^2 \Phi_{ww}(\vec{\omega})\, d\vec{\omega}. \tag{4}$$

For MSE distortion, $D_{yx} = \sigma_w^2$. To find $D_{\hat{y}x}$, we use (2) and find

$$D_{\hat{y}x} = \frac{1}{(2\pi)^M} \int_\Omega |F(\vec{\omega})|^2 \left[|G(\vec{\omega}) - 1|^2 \Phi_{xx}(\vec{\omega}) + |G(\vec{\omega})|^2 \Phi_{ww}(\vec{\omega}) + \Phi_{vv}(\vec{\omega})\right]\, d\vec{\omega}. \tag{5}$$

Finally, since σ_x^2 is the power of the original $\mathbf{x}[\vec{n}]$, we define the *perceptual power* of the original by

$$P_x = \frac{1}{(2\pi)^M} \int_\Omega |F(\vec{\omega})|^2 \Phi_{xx}(\vec{\omega})\, d\vec{\omega}. \tag{6}$$

[2] As a service to the reader, we avoid using the pun "blind-and-dumb reception."

2.4 Capacity Expressions

We also require an expression for the capacity of the watermarking system. For simplicity, let us momentarily assume a 1-D AWGN channel *with no filtering*; i.e., $g[\vec{n}] = \delta[\vec{n}]$. The channel noise is $\mathbf{n}[n] = \mathbf{x}[n]+\mathbf{v}[n]$, where $\mathbf{x}[n]$ and $\mathbf{v}[n]$ are WGN with respective variances σ_x^2 and σ_v^2 and are independent of one another. The state $\mathbf{x}[n]$ is known to the encoder, which transmits a signal $\mathbf{w}[n]$ subject to a power constraint σ_w^2, and the decoder receives $\mathbf{y}[n] = \mathbf{w}[n] + \mathbf{x}[n] + \mathbf{v}[n] = \mathbf{w}[n] + \mathbf{n}[n]$. In watermarking, the state $\mathbf{x}[n]$ is analogous to the original, and the transmitted signal $\mathbf{w}[n]$ is analogous to the watermark signal.

In the reception-with-original scenario, $\mathbf{x}[n]$ is known to both the encoder and decoder. The decoder can just subtract $\mathbf{x}[n]$ from $\mathbf{y}[n]$. The result is just like an AWGN channel with power constraint σ_w^2 and noise power σ_v^2, and the capacity is $C = \frac{1}{2}\log_2(1 + \sigma_w^2/\sigma_v^2)$ [39, 14]. If $\mathbf{w}[\vec{n}]$ and $\mathbf{v}[\vec{n}]$ are M-D with respective power spectra $\Phi_{ww}(\vec{\omega})$ and $\Phi_{vv}(\vec{\omega})$, then the capacity is $C = (2\pi)^{-M}\int_\Omega \frac{1}{2}\log_2\left(1 + \Phi_{ww}(\vec{\omega})/\Phi_{vv}(\vec{\omega})\right)\, d\vec{\omega}$ [40].

In blind watermarking, $\mathbf{x}[n]$ is known to the encoder but not to the decoder. This scenario was considered by Costa [13], who proved the remarkable and surprising result that the capacity is again $C = \frac{1}{2}\log_2(1 + \sigma_w^2/\sigma_v^2)$. Hence, blind watermarking can theoretically perform as well as reception-with-original watermarking[3]! Also, the capacity is independent of the power σ_x^2 of the state/original $\mathbf{x}[n]$. The interested reader is referred to the references in Sec. 2.2 for more details on Costa's solution.

To extend Costa's result to an M-D Gaussian channel with memory and channel state known to the encoder, one can divide the frequency spectrum into parallel, independent Gaussian subchannels, apply the result to each subchannel, and let the number of subchannels go to infinity [40]. Then for fixed power spectra $\Phi_{ww}(\vec{\omega})$, $\Phi_{xx}(\vec{\omega})$, and $\Phi_{vv}(\vec{\omega})$, $C = (2\pi)^{-M}\int_\Omega \frac{1}{2}\log_2\left(1 + \Phi_{ww}(\vec{\omega})/\Phi_{vv}(\vec{\omega})\right)\, d\vec{\omega}$.

Our watermarking model includes filtering and the original-interference suppression factor a. Consequently, we use (3) to write the capacity [29] as

$$C = \frac{1}{(2\pi)^M}\int_\Omega \frac{1}{2}\log_2\left(1 + \frac{|G(\vec{\omega})|^2\Phi_{ww}(\vec{\omega})}{(1-a)^2|G(\vec{\omega})|^2\Phi_{xx}(\vec{\omega}) + \Phi_{vv}(\vec{\omega})}\right) d\vec{\omega}. \tag{7}$$

We may interpret (7) as follows. We say that the *effective watermark* is $\mathbf{w}_e[\vec{n}] = g[\vec{n}] * \mathbf{w}[\vec{n}]$, while the *effective noise* is $\mathbf{n}_e[\vec{n}] = (1-a)g[\vec{n}] * \mathbf{x}[\vec{n}] + \mathbf{v}[\vec{n}]$. Then (7) becomes

$$C = \frac{1}{(2\pi)^M}\int_\Omega \frac{1}{2}\log_2\left(1 + \frac{\Phi_{w_e w_e}(\vec{\omega})}{\Phi_{n_e n_e}(\vec{\omega})}\right) d\vec{\omega}. \tag{8}$$

In the sequel, we will make alternative interpretations where they are useful.

[3] Under these assumptions; for non-Gaussian channels, C_{blind} is likely to be less than $C_{\text{with original}}$.

It is important to appreciate the need for noise $\mathbf{v}[\vec{n}]$ in (2) and (7). Suppose that $g[\vec{n}]$ is invertible, i.e., $G(\vec{\omega}) \neq 0$, $\forall\vec{\omega}$. For an ideal watermarking scheme, $a = 1$, so that the original $\mathbf{x}[\vec{n}]$ does not hinder communication. If the noise $\mathbf{v}[\vec{n}]$ were not present, then the attack would be invertible; the receiver could perfectly undo the effects of the attack, and C would be infinite[4]. The noise $\mathbf{v}[\vec{n}]$ is necessary to make the attack in (2) non-invertible when $g[\vec{n}]$ is invertible. This observation agrees with [34], where the authors pointed out that if an attack is invertible, it does not impair communication at all.

Strictly speaking, "capacity" is the supremum of achievable rates over all possible watermarking systems. When $0 \leq a < 1$, Eq. (7) actually gives the maximum *achievable rate* of a suboptimal watermarking system. However, it is common to speak of "capacity" when describing the best performance of a given, perhaps suboptimal, communications system. For brevity, we use the term "capacity" even when $a \neq 1$.

2.5 Attacks and Defenses

With these expressions for D_{yx}, $D_{\hat{y}x}$, and C, we are ready to look for optimal attacks and defenses. We assume that $\Phi_{xx}(\vec{\omega})$ and $F(\vec{\omega})$ are fixed. We state the attacker's problem formally as:

Problem 1 (Attack) *Let $\Phi_{ww}(\vec{\omega})$ be given. For some target capacity $C_t \geq 0$, choose $G(\vec{\omega})$ and $\Phi_{vv}(\vec{\omega})$ to minimize $D_{\hat{y}x}$ such that $C = C_t$.*

Alternatively, the attacker could attempt to minimize C under the constraint $D_{\hat{y}x} = D_t$.

The owner seeks a defense; this problem is defined as:

Problem 2 (Defense) *Let $G(\vec{\omega})$ and $\Phi_{vv}(\vec{\omega})$ be given. For some maximum embedding distortion D_{embed} and some target capacity $C_t \geq 0$, choose $\Phi_{ww}(\vec{\omega})$ to maximize $D_{\hat{y}x}$ such that $C = C_t$ and $D_{yx} \leq D_{\text{embed}}$.*

The owner has an additional constraint on D_{yx}, the embedding distortion. For the Gaussian channel, capacity increases with signal power, so the inequality constraint can be replaced by the equality $D_{yx} = D_{\text{embed}}$, which maximizes the allowable watermark power. Finally, we remark that the owner could instead try to maximize C under the constraints $D_{\hat{y}x} = D_t$ and $D_{yx} = D_{\text{embed}}$.

The attack distortion $D_{\hat{y}x}$ at capacity $C = C_t$ (capacity C at distortion $D_{\hat{y}x} = D_t$) provides a well-defined way of evaluating the robustness of a watermark with a given power spectrum $\Phi_{ww}(\vec{\omega})$ and embedding distortion $D_{yx} = D_{\text{embed}}$. With $C = C_t$ ($D_{\hat{y}x} = D_t$), the greater the attack distortion $D_{\hat{y}x}$ (capacity C), the more robust the watermark.

[4] Of course, in a practical system, signal values are digitized, so the capacity will not actually be infinite but will be limited by the precision of the digital representation.

We have constrained attacks to consist of LSI filtering and ACGN for mathematical tractability. Consequently, there may exist other, more powerful attacks. In this paper, “optimum linear filtering” means the best LSI filtering attack (in conjunction with ACGN), rather than minimum mean-squared error (MMSE) or maximum *a posteriori* (MAP) estimation via filtering. Any claims of optimality in this paper refer to optimality within the class of LSI-filtering/ACGN attacks. However, for $\mathbf{x}[\vec{n}]$ memoryless and Gaussian and MSE distortion, it has been shown [34] that the attack (2) is optimum among *all possible attacks*,[5] and Gaussian-distributed signals $\mathbf{w}[\vec{n}]$ achieve the highest communication rate in the presence of this attack. Hence, for $\mathbf{x}[\vec{n}]$ Gaussian and MSE distortion, our results will describe the ultimate performance limits.

3 Optimum Attack

We present the optimum attack, but first discuss an intuitively appealing, but suboptimal, attack.

3.1 Effective White-Noise Attack and Defense (Power-Spectrum Condition)

Since the receiver knows $g[\vec{n}]$, it can apply the inverse filter with impulse response $g^{-1}[\vec{n}]$ and, without loss of information [29], compute

$$\mathbf{z}_i[\vec{n}] = g^{-1}[\vec{n}] * \mathbf{z}[\vec{n}] = \mathbf{w}[\vec{n}] + (1-a)\mathbf{x}[\vec{n}] + g^{-1}[\vec{n}] * \mathbf{v}[\vec{n}]. \tag{9}$$

We may say that the watermark remains $\mathbf{w}[\vec{n}]$ and define the *effective inverse-filtered noise* by $\mathbf{n}_i[\vec{n}] = (1-a)\mathbf{x}[\vec{n}] + g^{-1}[\vec{n}] * \mathbf{v}[\vec{n}]$. Define the power spectrum of $\mathbf{n}_i[\vec{n}]$ to be

$$\Phi_{n_i n_i}(\vec{\omega}) = \begin{cases} (1-a)^2\Phi_{xx}(\vec{\omega}) + |G(\vec{\omega})|^{-2}\Phi_{vv}(\vec{\omega}), & \text{if } G(\vec{\omega}) \neq 0; \\ \infty, & \text{if } G(\vec{\omega}) = 0. \end{cases} \tag{10}$$

Observe that $\Phi_{n_i n_i}(\vec{\omega})$ is well-defined for all frequencies, even if $G(\vec{\omega}) = 0$ at some frequencies. Then (7) can be written as [40, 29]

$$C = \frac{1}{(2\pi)^M} \int_\Omega \frac{1}{2} \log_2 \left(1 + \frac{\Phi_{ww}(\vec{\omega})}{\Phi_{n_i n_i}(\vec{\omega})}\right) d\vec{\omega}. \tag{11}$$

In a conventional ACGN channel, $\Phi_{n_i n_i}(\vec{\omega})$ remains fixed and $\Phi_{ww}(\vec{\omega})$ is selected to maximize the mutual information between the encoder and decoder. The solution for $\Phi_{ww}(\vec{\omega})$ is a water-filling rule [14, Sec. 10.5], which gives $\mathbf{w}[\vec{n}]$ a power advantage over the noise $\mathbf{n}_i[\vec{n}]$. It is well-known that communication in the presence of additive Gaussian noise is most difficult when the noise is white [40].

[5] In this case, the attack reduces to the Gaussian test channel [34].

An *effective white-noise attack* based on this idea was recently investigated in [48]. The details appear in App. A.[6] Whatever the shape of the watermark power spectrum $\Phi_{ww}(\vec{\omega})$, the attack selects $G(\vec{\omega})$ and $\Phi_{vv}(\vec{\omega})$ to make $\Phi_{n_i n_i}(\vec{\omega})$ directly proportional to $\Phi_{ww}(\vec{\omega})$. Consequently, $\mathbf{n}_i[\vec{n}]$ is white relative to $\mathbf{w}[\vec{n}]$. The owner cannot gain a power advantage by changing the shape of $\Phi_{ww}(\vec{\omega})$ since the attack will re-adjust $G(\vec{\omega})$ and $\Phi_{vv}(\vec{\omega})$ as needed.

As a defense against this attack, it can be shown (see [48] or App. B) that, for any $C = C_t$, $D_{\hat{y}x}$ is maximized when

$$\Phi_{ww}(\vec{\omega}) = \frac{\sigma_w^2}{\sigma_x^2}\,\Phi_{xx}(\vec{\omega}). \tag{12}$$

We refer to Eq. (12) as the *power-spectrum condition* (PSC). In terms of power spectra, the PSC states that *“the watermark should look like the original.”* A watermark that is “spectrally matched” to the original (in the sense that (12) is satisfied) is said to be *PSC-compliant*. The PSC makes sense intuitively: The watermark appears white relative to the original, which makes it hardest to estimate or distinguish from the original. Indeed, the PSC was first derived as a necessary and sufficient condition for resisting MMSE estimation of the watermark from the watermarked signal [47].

For this attack, Eq. (11) simplifies to $C = \frac{1}{2}\log_2\left(1 + \sigma_w^2/\sigma_{n_i}^2\right)$. When the PSC is satisfied, $G(\vec{\omega})$ and $\Phi_{vv}(\vec{\omega})$ become constant for all $\vec{\omega}$, and closed-form expressions for $\sigma_{n_i}^2$ and $D_{\hat{y}x}$ result [48]. Then a direct relationship between C and $D_{\hat{y}x}$ can be obtained (App. C):

$$C = \frac{1}{2}\log_2\left(1 + \frac{(P_x - D_{\hat{y}x})\,D_{\text{embed}}}{P_x^2 - (P_x - D_{\hat{y}x})\,(a(2-a)P_x + D_{\text{embed}})}\right). \tag{13}$$

This relationship was previously derived in [48] for MSE distortion,

$$C = \frac{1}{2}\log_2\left(1 + \frac{\left(\sigma_x^2 - D_{\hat{y}x}\right)\sigma_w^2}{\sigma_x^4 - (\sigma_x^2 - D_{\hat{y}x})\,(a(2-a)\sigma_x^2 + \sigma_w^2)}\right). \tag{14}$$

If $\mathbf{x}[\vec{n}]$ is memoryless and Gaussian, and the distortion measure is MSE, then $\Phi_{ww}(\vec{\omega})$, $G(\vec{\omega})$, and $\Phi_{vv}(\vec{\omega})$ are constant for all frequencies. For this case, the attack model (2) has been shown to be optimum among all possible attacks [34], and Eq. (14) describes the fundamental relationship between C and $D_{\hat{y}x}$.

3.2 Optimum Attack

The preceding attack and defense (PSC) each have an intuitively pleasing motivation. However, the hostile nature of attacks on watermarks means that $\Phi_{n_i n_i}(\vec{\omega})$ can adapt to $\Phi_{ww}(\vec{\omega})$. The attacker has “the last word” on the behavior of the channel, so the attacker, rather than the owner, has a potential power advantage.

[6] The distortion measure in [48] was MSE; the appendix extends the derivation to frequency-weighted MSE.

Consequently, the effective white-noise attack is suboptimal because, by restricting the form of $\Phi_{n_i n_i}(\vec{\omega})$, the attack does not fully exploit its power advantage. This section presents the optimum attack for a given watermark power spectrum $\Phi_{ww}(\vec{\omega})$.

Under the assumptions of IID RVs, a Gaussian original, and MSE distortion, it was shown in [34] that the optimum attack among all possible attacks consists of scaling and additive Gaussian noise. The attack model (2) thus extends the attack in [32] by adding memory. Hence, for $\mathbf{x}[\vec{n}]$ Gaussian and MSE distortion, the attack we derive will be optimum among all attacks.

The attacker's problem is to find $G(\vec{\omega})$ and $\Phi_{vv}(\vec{\omega})$ to minimize $D_{\hat{y}x}$ subject to $C = C_t$ (Problem 1). This problem can be solved by the calculus of variations; the details appear in App. D. The optimum attack filter and noise power spectrum are given by

$$G(\vec{\omega}) = A(\vec{\omega})\,\frac{\Phi_{xx}(\vec{\omega})}{\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})}, \tag{15}$$

$$\Phi_{vv}(\vec{\omega}) = (1 - A(\vec{\omega}))\,G(\vec{\omega})\Phi_{xx}(\vec{\omega}) = (1 - A(\vec{\omega}))\,A(\vec{\omega})\,\frac{\Phi^2_{xx}(\vec{\omega})}{\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})}, \tag{16}$$

where $0 \le A(\vec{\omega}) \le 1$, $\forall\vec{\omega}$. Note that $0 \le G(\vec{\omega}) < 1$, $\forall\vec{\omega}$, so that the filter can only attenuate the watermark; it will never amplify the watermark.

The exact expressions for $A(\vec{\omega})$ are rather complicated, so we provide them in stages. Unfortunately, they do not provide obvious insight into the exact nature of the attack. However, $A(\vec{\omega})$ is parameterized by a Lagrange multiplier λ, which leads to an interpretation of the attack behavior in Sec. 3.2.2.

For any $a \in [0, 1]$,

$$A(\vec{\omega}) = 1, \quad \text{for } \vec{\omega} \text{ such that } \Phi_{xx}(\vec{\omega}) = 0 \text{ or } \Phi_{ww}(\vec{\omega}) = 0. \tag{17}$$

If $\Phi_{xx}(\vec{\omega}) = 0$, this rule results in $G(\vec{\omega}) = \Phi_{vv}(\vec{\omega}) = 0$. There is no power from the original at this frequency, so the only possible power is due to the watermark; the attack can completely eliminate the watermark at this frequency without increasing $D_{\hat{y}x}$. Similarly, if $\Phi_{ww}(\vec{\omega}) = 0$, then $G(\vec{\omega}) = 1$ and $\Phi_{vv}(\vec{\omega}) = 0$. There is no watermark power at this frequency, so only the original signal is present (or else it is zero), and the attack passes this frequency unchanged.

In the equations that follow, we assume that $\Phi_{xx}(\vec{\omega}) > 0$ and $\Phi_{ww}(\vec{\omega}) > 0$ at frequency $\vec{\omega}$. Define

$$\text{cl}[x] = \begin{cases} 1, & \text{if } x > 1; \\ x, & \text{if } 0 \le x \le 1; \\ 0, & \text{if } x < 0. \end{cases} \tag{18}$$

For $a = 0$ (conventional blind reception),

$$A_0(\vec{\omega}) = \text{cl}\left[1 + \frac{\Phi_{xx}(\vec{\omega})}{\Phi_{ww}(\vec{\omega})} - \frac{\lambda}{2\ln 2}\frac{\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})}{\Phi_{xx}^2(\vec{\omega})|F(\vec{\omega})|^2}\right]. \tag{19}$$

For $0 < a \leq 1$, $A(\vec{\omega})$ has a different form than (19) (an explanation appears in App. D),

$$\begin{aligned} &A(\vec{\omega}) = \\ &\text{cl}\left[\left(1 + \frac{\Phi_{ww}}{2a(2-a)\Phi_{xx}} - \frac{\sqrt{\Phi_{xx}^2\Phi_{ww}^2 + (2\lambda/\ln 2)a(2-a)\Phi_{xx}\Phi_{ww}\left(a(2-a)\Phi_{xx} + \Phi_{ww}\right)|F|^{-2}}}{2a(2-a)\Phi_{xx}^2}\right)\right. \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left.\times\left(\frac{\Phi_{xx} + \Phi_{ww}}{a(2-a)\Phi_{xx} + \Phi_{ww}}\right)\right], \end{aligned} \tag{20}$$

where we have omitted the frequency variable $\vec{\omega}$ on the right-hand side. In particular, with $a = 1$ we have reception-with-original/optimal blind reception, and (20) reduces slightly to

$$\begin{aligned} &A_1(\vec{\omega}) = \\ &\text{cl}\left[1 + \frac{\Phi_{ww}(\vec{\omega})}{2\Phi_{xx}(\vec{\omega})} - \frac{\sqrt{\Phi_{xx}^2(\vec{\omega})\Phi_{ww}^2(\vec{\omega}) + (2\lambda/\ln 2)\Phi_{xx}(\vec{\omega})\Phi_{ww}(\vec{\omega})\left(\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})\right)|F(\vec{\omega})|^{-2}}}{2\Phi_{xx}^2(\vec{\omega})}\right]. \end{aligned} \tag{21}$$

Applying (15) and (16), the corresponding distortion is given by

$$\begin{aligned} D_{\hat{y}x} &= P_x - \frac{1}{(2\pi)^M}\int_\Omega |F(\vec{\omega})|^2 A(\vec{\omega})\frac{\Phi_{xx}^2(\vec{\omega})}{\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})}\, d\vec{\omega} &(22)\\ &= P_x - \frac{1}{(2\pi)^M}\int_\Omega |F(\vec{\omega})|^2 G(\vec{\omega})\Phi_{xx}(\vec{\omega})\, d\vec{\omega}. &(23) \end{aligned}$$

Finally, the capacity can be written as

$$C = \frac{1}{(2\pi)^M}\int_\Omega \frac{1}{2}\log_2\left(1 + \frac{A(\vec{\omega})\Phi_{ww}(\vec{\omega})}{\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega}) - A(\vec{\omega})\left(a(2-a)\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})\right)}\right) d\vec{\omega}. \tag{24}$$

3.2.1 Lagrange Multiplier and Relationship between Capacity and Distortion

In the preceding equations, λ is a scalar Lagrange multiplier that determines $A(\vec{\omega})$. For any $a \in [0, 1]$, the limiting values of λ are

$$\lambda_{\min} = 2\ln 2 \min_{\vec{\omega}\in\mathcal{W}} \frac{\Phi_{xx}^2(\vec{\omega})|F(\vec{\omega})|^2}{\Phi_{ww}(\vec{\omega})}\frac{(1-a)^2\Phi_{xx}(\vec{\omega})\left[(1-a)^2\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})\right]}{\left(\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})\right)^2}, \tag{25}$$

and

$$\lambda_{\max} = 2\ln 2 \max_{\vec{\omega}\in\mathcal{W}} \frac{\Phi_{xx}^2(\vec{\omega})|F(\vec{\omega})|^2}{\Phi_{ww}(\vec{\omega})}. \tag{26}$$

When $a = 0$ (conventional blind reception),

$$\lambda_{\min} = 2 \ln 2 \min_{\vec{\omega} \in \mathcal{W}} \frac{\Phi_{xx}^3(\vec{\omega}) |F(\vec{\omega})|^2}{\Phi_{ww}(\vec{\omega}) \left(\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega}) \right)}. \tag{27}$$

When $a = 1$ (reception-with-original/optimal blind reception), $\lambda_{\min} = 0$.

The preceding expressions for $A(\vec{\omega})$ are complicated, but their general behavior depends on λ in a simple way. When $\lambda = \lambda_{\min}$, $A(\vec{\omega}) = 1$, $\forall \vec{\omega}$. As $\lambda \to \lambda_{\max}$, $A(\vec{\omega}) \to 0$, and when $\lambda = \lambda_{\max}$, $A(\vec{\omega}) = 0$, $\vec{\omega} \in \mathcal{W}$. Hence, $A(\vec{\omega})$ is a monotonically decreasing function of λ; the rate of decrease differs from one frequency to another, but the trend holds for all $\vec{\omega} \in \mathcal{W}$.

Thus, the Lagrange multiplier λ parameterizes $A(\vec{\omega})$, which in turn determines $D_{\hat{y}x}$ and C. Hence, $G(\vec{\omega})$ and $\Phi_{vv}(\vec{\omega})$ no longer appear in the expressions for $D_{\hat{y}x}$ and C (Eqs. (22) and (24)), and we can work directly with $D_{\hat{y}x}$, C, and λ. Since $A(\vec{\omega})$ decreases monotonically from unity to zero with λ, $D_{\hat{y}x}$ is a strictly increasing function $D_{\hat{y}x}(\lambda)$ of λ, while C is a strictly decreasing function $C(\lambda)$ of λ. We recognize that, via $A(\vec{\omega})$, λ controls the trade-off between C and $D_{\hat{y}x}$. By sweeping λ from $\lambda_{\min}$ to $\lambda_{\max}$, we can explore the full performance range of a given watermark power spectrum $\Phi_{ww}(\vec{\omega})$. We can thus compute the *distortion-capacity function* $\{(D_{\hat{y}x}(\lambda), C(\lambda)) : \lambda_{\min} \leq \lambda \leq \lambda_{\max}\}$. Because $D_{\hat{y}x}(\lambda)$ and $C(\lambda)$ are invertible functions of λ, we can also define the *capacity-distortion function* $C(D_{\hat{y}x}) = C(\lambda^{-1}(D_{\hat{y}x}))$. $C(D_{\hat{y}x})$ is decreasing since $D_{\hat{y}x}(\lambda)$ and $C(\lambda)$ are strictly increasing and decreasing, respectively.

We can also use the limiting values of λ to find the range of possible values for $D_{\hat{y}x}$ and C. They can be computed by substituting appropriate values of $A(\vec{\omega})$ (given next) into (22) and (24). When $\lambda = \lambda_{\min}$, $A(\vec{\omega}) = 1$, $\forall \vec{\omega}$, In this case, $\Phi_{vv}(\vec{\omega}) = 0$, and $G(\vec{\omega})$ reduces to the Wiener filter, which is the MAP and MMSE estimator for estimating $\mathbf{x}[\vec{n}]$ from $\mathbf{y}[\vec{n}]$. However, the attack is now invertible, so it does not impair communication at all [34] and merely beautifies $\hat{\mathbf{y}}[\vec{n}]$. $D_{\hat{y}x}$ is minimized, and C is maximized. Denote these values by $D_{\hat{y}x,\min}$ and $C_{\max}$. Note that when $a = 1$, $C_{\max} = \infty$ because the attack introduces no noise ($\mathbf{v}[\vec{n}] = 0$, $\forall \vec{n}$) and all interference from the original $\mathbf{x}[\vec{n}]$ can be eliminated.

When $\lambda = \lambda_{\max}$, $A(\vec{\omega}) = 0$, $\vec{\omega} \in \mathcal{W}$, and $A(\vec{\omega}) = 1$, $\vec{\omega} \notin \mathcal{W}$. Then $D_{\hat{y}x}$ is maximized and C is minimized; denote these values by $D_{\hat{y}x,\max}$ and $C_{\min}$. Note that $C_{\min}$ is always zero. If $\mathcal{X} \subseteq \mathcal{W}$, then $D_{\hat{y}x,\max} = P_x$ because the attacked signal becomes $\hat{\mathbf{y}}[\vec{n}] = 0$, $\forall \vec{n}$, so the attack must completely destroy the original signal to stop all communication.

3.2.2 Characterization of Attack Behavior

We can also use the relationship between λ and $A(\vec{\omega})$ to characterize the attack. For envisioned watermarking applications, it is reasonable to assume $\Phi_{xx}(\vec{\omega}) \gg \Phi_{ww}(\vec{\omega})$, $\forall \vec{\omega}$, so that $G(\vec{\omega}) \approx A(\vec{\omega})$ and

$\Phi_{vv}(\vec{\omega}) \approx (1 - A(\vec{\omega}))\, A(\vec{\omega})\Phi_{xx}(\vec{\omega})$. At each frequency $\vec{\omega}$, $G(\vec{\omega})$ decreases from nearly unity to zero with λ, and $\Phi_{vv}(\vec{\omega})$ first increases from zero to $\frac{1}{4}\Phi_{xx}(\vec{\omega})$ before decreasing back to zero. For small λ, $G(\vec{\omega}) \approx 1$ and $\Phi_{vv}(\vec{\omega}) > 0$, so the attack mainly adds noise; for large λ, $G(\vec{\omega}) \to 0$, so the attack chiefly discards frequency components. Thus, we may roughly describe the behavior of the optimum attack with the following **rule-of-thumb**: *At low distortions (high capacities), add noise; at high distortions (low capacities), throw away frequency components.*

3.3 Experimental Results

To examine the theoretical performance of watermarks with different power spectra, we modeled the watermark and original as 1-D autoregressive (AR) processes [25]. AR processes are often used to model naturally occurring signals such as audio, images, and video. Recall that a 1-D AR(p) process $\mathbf{x}[n]$ is generated by the stochastic difference equation $\mathbf{x}[n] = \sum_{k=1}^{p} a_k \mathbf{x}[n-k] + \mathbf{u}[n]$, where $\mathbf{u}[n]$ is WGN. The original signal was modeled as an AR(1) process with $a_1 = 0.95$ and power σ_x^2. The different watermarks, each with power σ_w^2, were modeled as follows: “PSC” (AR(1), $a_1 = 0.95$), “lowpass” (AR(1), $a_1 = 0.90$), “bandpass” (AR(2), $a_1 = 0$, $a_2 = -0.9025$), “highpass” (AR(1), $a_1 = -0.95$), and “white”. Examples of the original, white, and PSC-compliant power spectra appear in Fig. 2.

In decibels, the ratio of original-signal power to attack distortion is $10\log_{10}\frac{P_x}{D_{\hat{y}x}} = 10\log_{10}\frac{P_x}{D_{yx}} - 10\log_{10}\frac{D_{\hat{y}x}}{D_{yx}}$. In most of the experimental results presented in this paper, we employ MSE distortion and use the embedding ratio $10\log_{10}(\sigma_w^2/\sigma_x^2) = -30$ dB. For convenience, we use $D_{yx} = \sigma_w^2 = 1$, so that the distortion $D_{\hat{y}x}$ relates to the above ratio via $10\log_{10}\frac{P_x}{D_{\hat{y}x}} = 10\log_{10}\sigma_x^2 - 10\log_{10}D_{\hat{y}x}$.

The subsequent experiments are briefly described here. Sec. 3.3.1 compares the performances of the effective white-noise attack and the optimum attack and shows that the former attack is indeed suboptimal. Next, Sec. 3.3.2 examines how the behavior of the optimum attack changes depending on the attack distortion $D_{\hat{y}x}$; the results verify the rule-of-thumb in Sec. 3.2.2: “At low distortions, add noise; at high distortions, throw away frequency components.” Finally, Secs. 3.3.3 and 3.3.4 show the effect of interference from the original and suggest that there is not a unique watermark power spectrum that performs best over the entire range of attack distortions.

3.3.1 Comparison with Effective White-Noise Attack

The left-hand graph in Fig. 3 shows the capacity-distortion curves for various watermarks after the effective white-noise attack of Sec. 3.1 when $a = 1$. Clearly, the PSC-compliant watermark is most robust against this attack. However, the right-hand graph shows the performance of PSC-compliant and white watermarks after either the effective white-noise attack or the optimum attack when $a = 1$. It is evident that the PSC-compliant watermark is not most robust over the entire range of $D_{\hat{y}x}$; the white watermark performs much better at low distortions. By fully exploiting the potential power advantage, the optimum attack is clearly more effective than the effective white-noise attack. Against the PSC-compliant watermark, the optimum attack consistently reduces capacity by roughly one order of magnitude over the effective white-noise attack. Against the white watermark, the attacks perform comparably at low distortions, but at high distortions the advantage of the optimum attack becomes obvious.

3.3.2 Examples of Attack Behavior

The preceding results show that the white watermark resists the optimum attack better than the PSC-compliant watermark at lower distortions, while the situation is reversed at higher distortions. Here we provide some explanations for this behavior. We employ the interpretation of the effective watermark $\mathbf{w}_e[n]$ and noise $\mathbf{n}_e[n]$ in Sec. 2.4, as well as another interpretation presented here.

The purpose of this interpretation is to help visualize the relative powers of the watermark and attack. As remarked in Sec. 3.1, an ideal receiver could apply the inverse filter $g^{-1}[\vec{n}]$ to $\mathbf{z}[\vec{n}]$ (Eq. (3)) and then decode from $\mathbf{z}_i[\vec{n}]$ (Eq. (9)). Additionally, the receiver could apply an ideal whitening filter with transfer function $H(\vec{\omega}) = \left(\sigma_w^2/\Phi_{ww}(\vec{\omega})\right)^{1/2}$ and impulse response $h[\vec{n}]$. Thus, define $\mathbf{z}_w[\vec{n}] = h[\vec{n}] * g^{-1}[\vec{n}] * \mathbf{z}[\vec{n}] = \mathbf{w}_w[\vec{n}] + \mathbf{n}_w[\vec{n}]$, where $\mathbf{w}_w[\vec{n}] = h[\vec{n}] * g^{-1}[\vec{n}] * g[\vec{n}] * \mathbf{w}[\vec{n}]$ is the *whitened, effective watermark*, and $\mathbf{n}_w[\vec{n}] = h[\vec{n}] * g^{-1}[\vec{n}] * ((1-a)g[\vec{n}] * \mathbf{x}[\vec{n}] + \mathbf{v}[\vec{n}])$ is the *whitened, effective noise*. Since $G(\vec{\omega})$ may be zero at some frequencies, $\mathbf{w}_w[\vec{n}]$ has power spectrum

$$\Phi_{w_w w_w}(\vec{\omega}) = \begin{cases} \sigma_w^2, & G(\vec{\omega}) \neq 0; \\ 0, & G(\vec{\omega}) = 0. \end{cases} \tag{28}$$

From Eqs. (15) and (16), when $G(\vec{\omega}) = 0$, $\Phi_{vv}(\vec{\omega}) = 0$ as well. Thus, $\mathbf{n}_w[\vec{n}]$ has power spectrum

$$\Phi_{n_w n_w}(\vec{\omega}) = \begin{cases} \dfrac{\sigma_w^2}{\Phi_{ww}(\vec{\omega})} \left[(1-a)^2\Phi_{xx}(\vec{\omega}) + |G(\vec{\omega})|^{-2}\Phi_{vv}(\vec{\omega})\right], & G(\vec{\omega}) \neq 0; \\ 0, & G(\vec{\omega}) = 0. \end{cases} \tag{29}$$

Now (7) can be written as

$$C = \frac{1}{(2\pi)^M} \int_{\Omega_G} \frac{1}{2} \log_2 \left(1 + \frac{\Phi_{w_w w_w}(\vec{\omega})}{\Phi_{n_w n_w}(\vec{\omega})}\right) d\vec{\omega} = \frac{1}{(2\pi)^M} \int_{\Omega_G} \frac{1}{2} \log_2 \left(1 + \frac{\sigma_w^2}{\Phi_{n_w n_w}(\vec{\omega})}\right) d\vec{\omega}, \tag{30}$$

where Ω_G denotes the frequency support of $G(\vec{\omega})$ (within Ω).

Figs. 4 and 5 examine different parts of the attack for low and high distortions. Examples appear for both white and PSC-compliant watermarks. Four graphs appear for each watermark; all graphs in these figures use a decibel scale for the vertical axis. The upper-left graph shows the attack components $A(\omega)$, $G(\omega)$, and $\Phi_{vv}(\omega)$. The upper-right graph shows the power spectrum $\Phi_{ee}(\omega)$ of the error $\mathbf{e}[n] = \hat{\mathbf{y}}[n] - \mathbf{x}[n]$. The lower-left graph contains the power spectra of the effective watermark $\mathbf{w}_e[n]$ and noise $\mathbf{n}_e[n]$, and the lower-right graph shows their whitened versions, $\mathbf{w}_w[n]$ and $\mathbf{n}_w[n]$; in the latter graph, it is easier to see the attacker's power advantage relative to the watermark.

Fig. 4 shows the attack at low distortion, $D_{\hat{y}x} = 6$ dB; the white watermark has $C = 0.189$, and the PSC-compliant watermark has $C = 0.0519$. For both watermarks, $G(\omega) \approx A(\omega) \approx 1$, $\forall \omega$, and $\Phi_{vv}(\omega)$ has approximately the same shape as $\Phi_{ww}(\omega)$. The optimum attack primarily functions by properly shaping $\Phi_{vv}(\omega)$. The shape of $\Phi_{ee}(\omega)$ is similar to that of $\Phi_{vv}(\omega)$, which shows that the distortion is mainly due to the additive noise $\mathbf{v}[n]$. Against the white watermark, the lower set of graphs show that the effective noise power spectrum is nearly flat and about 6 dB greater than the watermark power spectrum at all frequencies; the attack must distribute the noise power evenly over all frequencies and cannot gain a substantial power advantage over any frequency range. Against the PSC-compliant watermark, $\Phi_{n_e n_e}(\omega)$ adapts the watermark power spectrum to make $\Phi_{n_w n_w}(\omega)$ about 10–17 dB greater than $\Phi_{w_w w_w}(\omega)$ at most frequencies; the attack gains a large power advantage except in a small region around the origin, where the watermark power spectrum $\Phi_{ww}(\omega)$ is concentrated. In this way, the white watermark reaches a higher capacity than the PSC-compliant one.

Fig. 5 shows the behavior at high distortion, $D_{\hat{y}x} = 24$ dB; now $C = 2.05 \times 10^{-5}$ for the white watermark, and $C = 2.19 \times 10^{-4}$ for the PSC-compliant watermark. The optimum attack discards frequency components: $G(\omega) = 0$ for $|\omega| > \omega_0$. The graph of $\Phi_{ee}(\omega)$ shows that the distortion is dominated by the portions of the original that have been filtered out. Against the white watermark, the filter has $\omega_0 \approx 0.05\pi$; most of the watermark (and original) is simply discarded. Over the interval $[-\omega_0, \omega_0]$, $\Phi_{n_w n_w}(\omega)$ is 32–38 dB greater than $\Phi_{w_w w_w}(\omega)$. Against the PSC-compliant watermark, $\omega_0 \approx 0.15\pi$; this larger frequency support indicates that more of the watermark passes through the filter $G(\omega)$. In addition, $\Phi_{n_w n_w}(\omega)$ is only 20–30 dB greater than $\Phi_{w_w w_w}(\omega)$ over most of the interval $[-\omega_0, \omega_0]$. Thus, after filtering by $G(\omega)$, more watermark power remains for the PSC-compliant watermark than for the white one, and the former achieves a higher capacity.

These examples agree with the rule-of-thumb in Sec. 3.2.2: "At low distortions, add noise; at high distortions, throw away frequency components." They help explain why the best-performing watermark power spectrum is likely not unique over the entire range of $D_{\hat{y}x}$. The white watermark better resists additive noise, while the PSC-compliant watermark better resists frequency-selective filtering.

3.3.3 Reception-with-Original/Optimal Blind Reception

The left-hand graph in Fig. 6 shows the theoretical performance of various watermarks when subjected to the optimum attack with $a = 1$. The capacity-distortion curves may appear not to be convex, but this is a visual effect due to the logarithmic capacity scale and decibel distortion scale. When drawn with linear scales, the curves are convex but difficult to distinguish.

Again, we immediately see that the PSC-compliant watermark is not optimum for all $D_{\hat{y}x}$. None of the watermarks tested has the best performance over the entire range of $D_{\hat{y}x}$. At low distortions, the white watermark performs best; at high distortions, the PSC-compliant watermark performs best; for a middle range of distortions, the lowpass watermark performs better than both the white and PSC-compliant ones. This behavior suggests that there may be no unique optimum watermark power spectrum that maximizes C over all $D_{\hat{y}x}$.

3.3.4 Interference from the Original

The right-hand graph in Fig. 6 shows performance curves for the watermarks when $a = 0$. Generally, the white watermark performs best at low to medium distortions, while the PSC-compliant watermark performs best at high distortions.

For low to medium distortions, the PSC-compliant watermark has the poorest performance and the white watermark has the best performance. We can explain this behavior as follows. The PSC-compliant watermark power spectrum has the same shape as $\Phi_{xx}(\omega)$, so it suffers most from interference due to the original. The bandpass and highpass watermarks concentrate their power away from the frequencies where $\Phi_{xx}(\omega)$ is largest; these watermarks outperform the PSC-compliant watermark, in contrast to case $a = 1$ (left-hand graph in Fig. 6). Yet because they concentrate power in the middle or high frequencies, these watermarks are easier to attack than the white watermark.

However, at high distortions, most of the power of the bandpass and highpass watermarks is discarded by the attack; the same occurs for the white watermark, although to a lesser extent. The shape of the PSC-compliant watermark power spectrum allows more of its power to survive the attack. In a sense, the original shields the PSC-compliant watermark because the attack cannot discard the frequency components where the watermark power is concentrated without also destroying the original.

4 Optimized Defense

Finding a defense in the presence of the optimum attack is extremely difficult. The owner should select $\Phi_{ww}(\vec{\omega})$ to maximize $D_{\hat{y}x}$ while satisfying constraints $D_{yx} = D_{\text{embed}}$ and $C = C_t$ (Problem 2). However, $D_{\hat{y}x}$ and C depend upon $\Phi_{ww}(\vec{\omega})$ in a complicated manner via $A(\vec{\omega})$. An analytic solution for $\Phi_{ww}(\vec{\omega})$ may be impossible to find. Also, the experimental results of Sec. 3.3 suggest that there may not be a unique $\Phi_{ww}(\vec{\omega})$ that solves the owner's problem over all possible pairs $(D_{\hat{y}x}, C_t)$.

4.1 Piecewise-Constant Approximation

In an attempt to learn more about the possible solution, we make some approximations that may allow us to find $\Phi_{ww}(\vec{\omega})$ numerically. Divide Ω into N non-overlapping, equal-support subsets that cover Ω. Hence, each region has a total support size of $(2\pi)^M/N$. We assume that $|F(\vec{\omega})|$ and all power spectra are constant over each subset. We index the N subsets from 1 to N, so that when $\vec{\omega}$ lies in the nth subset, $\Phi_{xx}(\vec{\omega}) = X_n \geq 0$, $\Phi_{ww}(\vec{\omega}) = W_n \geq 0$, and $|F(\vec{\omega})| = F_n > 0$. We often denote the piecewise-constant quantities as N-vectors. For example, $\Phi_{ww}(\vec{\omega})$ is represented by $W = \begin{bmatrix} W_1 & W_2 & \cdots & W_N \end{bmatrix}^{\mathrm{T}}$.

Consequently, $A(\vec{\omega})$, $G(\vec{\omega})$, and $\Phi_{vv}(\vec{\omega})$ are also constant over each subset; the corresponding values are respectively denoted by A_n, G_n, and V_n. We may therefore view each subset as an independent subchannel, so that we have N parallel subchannels. The integrals involving D_{yx}, $D_{\hat{y}x}$, and C can be replaced by summations. Eqs. (5), (22) and (24) become

$$D_{yx} = \frac{1}{N} \sum_{n=1}^{N} F_n^2 W_n, \tag{31}$$

$$D_{\hat{y}x} = P_x - \frac{1}{N} \sum_{n=1}^{N} F_n^2 A_n \frac{X_n^2}{X_n + W_n}, \tag{32}$$

$$C = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{2} \log_2 \left(1 + \frac{A_n W_n}{X_n + W_n - A_n \left(a(2-a) X_n + W_n \right)} \right). \tag{33}$$

4.2 Description of Optimization Algorithms

We have implemented three iterative numerical methods to optimize the watermark vector W. The first uses a *greedy marginal-analysis* (GMA) algorithm, the second employs *simulated annealing* (SA), and the third uses *"greedy" annealing* (GA) [50]. A brief description of the algorithms follows; more details are given in App. E,

Let D_{embed} and C_t be given, so $D_{\hat{y}x}$ should be maximized. An initial vector W that satisfies D_{embed} is selected; we choose W to distribute the embedding distortion evenly over all N subchannels. During each iteration, W is perturbed slightly, and the attack is re-optimized. Attack re-optimization can be performed efficiently because $C(\lambda)$ is a decreasing function of λ (Sec. 3.2.1). Hence, a bisection search can be used to find λ^* such that $|C(\lambda^*) - C_t|/C_t < \varepsilon$. Once λ^* has been found, $D_{\hat{y}x}(\lambda^*)$ can be computed. When the perturbations no longer produce increases in $D_{\hat{y}x}$, the algorithms stop. The algorithms all operate in this way but differ in the way W is perturbed and the stopping criterion.

4.3 Experimental Results

In these experiments, we used the AR models described in Sec. 3.3 and approximated the power spectra using $N = 64$ subchannels. In our initial experiments, we found that the results produced by the SA and GA methods were almost identical, but the SA algorithms took much longer to converge. For this reason, we only present results for the GMA, GA/normal, and GA/scaled methods here.

4.3.1 Reception-with-Original/Optimal Blind Reception

The left-hand graph of Fig. 7 shows the results for optimized watermark power spectra with $a = 1$. The optimization algorithms gave nearly identical capacity-distortion curves. The greedy nature of the algorithms sometimes caused them to become trapped in local maxima, but overall they performed well. It is evident that a white watermark performs nearly as well as the optimized watermarks at low distortions, while a PSC-compliant watermark performs almost as well at high distortions. These results confirm that there is not a unique optimum defense for all attack distortions.

Fig. 8 shows examples of the optimized watermark power spectra produced by the optimization methods at three different attack distortions ($D_{\hat{y}x} = 6$, 15, and 24 dB); the corresponding capacities appear in the table in the figure. The white and PSC-compliant power spectra appear for comparison. At 6 dB, the upper plot shows that all three optimized watermarks are nearly white. At 15 dB, the middle plot demonstrates that the optimized power spectra reach a compromise between the white and PSC-compliant power spectra. The optimized power spectra are similar to the power spectrum of the lowpass AR model, which explains its good performance at medium distortions in Sec. 3.3.3 and Fig. 6. Lastly, at 24 dB, the optimized watermark power spectra are very similar to the PSC-compliant one. These results agree with the discussion in Sec. 3.3.2, and we have a **rule-of-thumb** for the defense: *At low distortions, white watermarks perform well; at high distortions, PSC-compliant watermarks perform well.*

Lastly, observe that at high distortion (24 dB), the optimized watermark power spectra are not zero at high frequencies, even though the attack discards these frequency components. A small amount of watermark power remains in these frequency components, which corresponds to the requirement $W_n' \geq \varepsilon$, $\forall n$, in the optimization algorithms (see App. E). It would appear that performance could be improved by putting all of the watermark power in the frequency components that are not discarded by the attack. However, the attack would re-optimize itself and no longer discard frequency components where the watermark power was zero. The attack would no longer need to incur large distortions while reducing C; in particular, $D_{\hat{y}x,\max}$ could become less than P_x. By leaving a small amount of its power at high frequencies, the optimized watermark has $\mathcal{W} = \mathcal{X}$ and forces the attack to discard these frequency components and incur additional distortion.

4.3.2 Interference from the Original

Experiments were also conducted for the case $a = 0$, corresponding to conventional blind reception when no knowledge of the original is exploited by the watermarking system. Now the original $\mathbf{x}[n]$ acts like high-power, low-frequency noise, so the original signal's power spectrum $\Phi_{xx}(\omega)$ forms a large portion of the effective noise power spectrum $\Phi_{n_e n_e}(\omega)$ (Sec. 2.4). The resulting capacity-distortion curves appear in the right-hand graph of Fig. 7. The GMA algorithm occasionally became trapped, but the GA methods gave consistently good performance. At low distortions, the white watermark again performs close to the optimized watermarks, while the PSC-compliant watermark does so at high distortions. However, the curves show that, unlike the case $a = 1$, optimizing the watermark power spectrum results in additional improvement.

Fig. 9 shows the optimized watermark power spectra for $D_{\hat{y}x} = 6$, 15, and 24 dB and the corresponding capacities. The power spectra reflect a compromise between resisting the attack and water-filling to avoid interference from $\Phi_{xx}(\omega)$. At low distortion, $\Phi_{xx}(\omega)$ dominates $\Phi_{n_e n_e}(\omega)$. The optimized watermark power spectra have roughly the same shape as $\Phi_{xx}(\omega)$ (so they appear white relative to $\Phi_{n_e n_e}(\omega)$), but they place very little power at the frequencies where $\Phi_{xx}(\omega)$ is largest (so they avoid most of the original-signal interference). As the attack distortion becomes larger, the optimized watermarks become more like a PSC-compliant watermark because the attack begins to discard the frequency components where $\Phi_{xx}(\omega)$ is not concentrated. Although $\Phi_{xx}(\omega)$ interferes with watermark reception, it also acts like a shield that prevents the low-frequency portions of the watermarks from being filtered out, as suggested in Sec. 3.3.4.

4.3.3 Comparison of Blind Schemes

This section looks at the potential capacity improvement in blind watermarking when knowledge of the original signal is exploited during watermark embedding, rather than using a conventional blind watermarking scheme that treats the original as noise. The left-hand graph in Fig. 10 displays the capacity-distortion curves for the optimized watermarks produced by the GA/scaled algorithm for four different values of the original-interference suppression factor a: 0 (conventional blind reception), 0.2929 (50% original-interference power suppression), 0.6838 (90% power suppression), and 1 (optimal blind reception). It is clear that a suboptimal blind watermarking scheme that suppresses some of the original-interference power can achieve a substantial increase in capacity. Promising methods for practical schemes may be found in [16, 10, 38, 8, 11].

The right-hand graph in Fig. 10 shows the ratio $C_{p\%}/C_{0\%}$, where $C_{p\%}$ denotes the capacity when p percent of the original-interference power is suppressed. For $D_{\hat{y}x} < 10$ dB, $C_{50\%}$ is about 1.8 times $C_{0\%}$, $C_{90\%}$ about 3–7 times, and $C_{100\%}$ is 7–148 times (not shown). For 10 dB $\leq D_{\hat{y}x} \leq 25$ dB, the increases are more modest; the capacities have the approximate ratio $C_{100\%} : C_{90\%} : C_{50\%} : C_{0\%} \approx 6 : 3 : 1.5 : 1$. At higher distortions, there is still some room for improvement over conventional blind reception, but $C_{100\%}$ is itself very small (less than 10^{-4}). In applications where communication even at very high distortions is necessary, a conventional blind watermarking scheme may remain a practical choice. However, when the number of samples is limited (e.g., image watermarking), even an optimal scheme may not be able to communicate sufficient information at high distortions.

4.3.4 Comparison with Memoryless Case

It is also worthwhile to compare performance for a correlated original and for a white or memoryless original, for which Eq. (13) gives a closed-form relationship between C and $D_{\hat{y}x}$. Of course, in the memoryless case, there are no frequencies where the optimum attack or optimized defense can gain a power advantage. We now show that memory can significantly affect watermark capacity.

Fig. 11 shows the capacity-distortion curves for lowpass (AR(1), $a_1 = 0.95$) and memoryless original signals. Consider low distortions. When $a = 1$ (left-hand graph), the curves are almost identical. This behavior can be explained by recalling that, for the colored original in this distortion range, the best watermarks are almost white, and the optimum attack mainly operates by adding nearly-white noise. When $a = 0$ (right-hand graph), watermark capacity with a colored original is much greater than with a white original. Recall that the optimized watermarks do not place much power at frequencies where $\Phi_{xx}(\omega)$ is large, but they cannot do this in the memoryless case, where $\Phi_{xx}(\omega)$ is flat.

For either $a = 0$ or $a = 1$ at higher distortions, watermark capacity with a memoryless original is significantly greater than that with a colored original. When the original is colored, the attack can exploit a power advantage at frequencies where $\Phi_{xx}(\omega)$ is small and eventually discard these frequency components when the distortion becomes large. This is not possible in the memoryless case.

4.3.5 Comparison with Suboptimal Attacks

This section compares the performance of three different attacks and their respective defenses. For simplicity, we employ MSE distortion and assume $a = 1$ (reception-with-original/optimal blind reception). First, we consider an *additive-noise attack*, consisting only of ACGN $\mathbf{v}[\vec{n}]$, so $\hat{\mathbf{y}}[\vec{n}] = \mathbf{x}[\vec{n}]+\mathbf{w}[\vec{n}]+\mathbf{v}[\vec{n}]$. It is clear that such an attack is suboptimal, but this model has been used frequently in the watermarking literature. For this attack, $D_{\hat{y}x} = \sigma_w^2 + \sigma_v^2$. Then the best watermark has a white power spectrum; the attack cannot gain a power advantage at any frequency and must use white noise. Hence, $C = \frac{1}{2}\log_2\left(1 + \sigma_w^2/(D_{\hat{y}x} - \sigma_w^2)\right)$. Second, we consider the effective white-noise attack of Sec. 3.1. The best defense against this attack requires that $\Phi_{ww}(\vec{\omega})$ satisfy the PSC (12). The capacity-distortion curve is given by (14). Third, we present the results for the optimum attack of Sec. 3.2 and the optimized defense (watermark power spectra) generated by the GA/scaled algorithm described in Sec. 4.2.

Fig. 12 shows the capacity-distortion curves for the three attacks. At low distortions, the additive-noise attack provides a fairly accurate approximation to the optimum attack, which functions mainly as additive noise in this distortion range. As the distortion increases, the capacities predicted by the additive-noise and effective white-noise attack models become erroneously optimistic. At high distortions, the latter anticipates a capacity 10–50 times greater than that actually produced by the optimum attack; for the former, the capacity may be overestimated by factors as large as 50, 100, or more. These results demonstrate that neither of these attack models is adequate when communication must be maintained even at high distortions after hostile attacks. However, the additive-noise attack model may suffice for applications in which only a modest amount of distortion must be tolerated or where resistance to hostile attacks is unnecessary.

4.3.6 Frequency-Weighted Distortion

All of the results presented to this point employ MSE distortion. To observe how the frequency-weighted MSE affects the attack and watermark, we present a few experimental results here. We set $|F(\omega)|^2 = 2.0009/\left(1 + 0.5e^{j2\omega}\right)\left(1 + 0.5e^{-j2\omega}\right)$. The scale factor of 2.0009 is chosen so that $P_x = 10^3 = \sigma_x^2$, and hence the frequency-weighted distortion covers the same range as MSE distortion. A plot of the piecewise-constant curves that correspond to $\Phi_{xx}(\omega)$ and $|F(\omega)|^2$ appear in Fig. 13. This choice of $|F(\omega)|^2$ means that distortions at the middle frequencies are more perceptible than those at low or high frequencies.

Examples of the GA/scaled-optimized watermark and the attack behavior appear in Fig. 14 for $D_{\hat{y}x} = 6$ and 24 dB. The original-interference suppression factor is $a = 1$, and the perceptual embedding distortion is maintained at $D_{yx}/P_x = -30$ dB. In each set of four graphs, the upper-left graph includes both $\Phi_{ww}(\omega)$ and the frequency-weighted watermark $F_{ww}(\omega) = |F(\omega)|^2\Phi_{ww}(\omega)$; likewise, the lower-right graph shows the error power spectrum $\Phi_{ee}(\omega)$ and its frequency-weighted counterpart $F_{ee}(\omega) = |F(\omega)|^2\Phi_{ee}(\omega)$.

At low distortion (6 dB), the optimized watermark is not white—unlike the case of MSE distortion—but the frequency-weighted watermark $F_{ww}(\omega)$ is. We say that the latter is *perceptually white*. The attack filter $G(\omega)$ remains almost flat, but the noise power spectrum $\Phi_{vv}(\omega)$ is now shaped to avoid introducing excessive power at frequencies where $|F(\omega)|^2$ is large. The power spectra $\Phi_{w_e w_e}(\omega)$ and $\Phi_{n_e n_e}(\omega)$ of the effective watermark and noise have almost the same shape. As a result, $\Phi_{n_e n_e}(\omega)$ is approximately white relative to $\Phi_{w_e w_e}(\omega)$; this offers a direct analogy to the lower-left graphs in Fig. 4 for MSE distortion. Finally, the error power spectrum $\Phi_{ee}(\omega)$ is also shaped such that its weighted counterpart $F_{ee}(\omega)$ has a flat power spectrum.

At high distortion (24 dB), the optimized watermark power spectrum is roughly PSC-compliant. The attack discards frequency components, but it cannot discard as many middle frequency components as with MSE distortion. Next, let $F'_{ww}(\omega) = |F(\omega)|^2(D_{\text{embed}}/P_x)\Phi_{xx}(\omega)$, which corresponds to a *frequency-weighted* PSC-compliant watermark power spectrum. $F'_{ww}(\omega)$ is drawn as a dotted curve in the upper-left graph. We also say that $F'_{ww}(\omega)$ is *perceptually PSC-compliant*. Clearly, the frequency-weighted, optimized power spectrum $F_{ww}(\omega)$ closely matches $F'_{ww}(\omega)$.

Fig. 15 shows the distortion-capacity curves with this frequency-weighted MSE distortion measure and an original-interference suppression factor $a = 1$. Three optimized curves are shown, as well as curves for perceptually white and perceptually PSC-compliant watermarks. The curves show that the perceptually white watermark performs nearly optimally at low distortions, and the perceptually PSC-complaitn watermark performs nearly optimally at high distortions.

Based on these results, we can extend the observations for MSE distortion (e.g., Sec. 4.3.1) to frequency-weighted MSE in a simple way to obtain the following **rule-of-thumb**: *At low perceptual distortions, a perceptually white watermark performs nearly optimally, while at high perceptual distortions, a perceptually PSC-compliant watermark performs nearly optimally.*

5 Conclusions and Remarks

5.1 Summary and Conclusions

We have analyzed the theoretical performance of watermarks and employed a well-defined robustness criterion that measures watermark capacity and attack distortion. Watermarking was viewed as communication over a hostile channel, where attacks take place. Our attack channel model consisted of LSI filtering and additive colored Gaussian noise. In a conventional additive-noise channel, communication is most difficult when the noise is white and Gaussian. This observation inspired the investigation of an effective white-noise attack. It was shown that the best defense (watermark power spectrum) against this attack results when a *power-spectrum condition* (PSC) is fulfilled. The PSC states that the watermark power spectrum should be directly proportional to the power spectrum of the original signal; in other words, *"the watermark should look like the original"* in a statistical sense.

However, unlike conventional channels, the hostile attack channel is not fixed but adapts to the watermark; the attacker, not the owner, has "the last word." The optimum attack was derived and shown to be superior to the effective white-noise attack. The optimum attack is difficult to describe exactly, but its behavior may be roughly described by a rule-of-thumb: "*At low attack distortions (high capacities), add noise; at high attack distortions (low capacities), throw away frequency components.*" Experiments demonstrated that the PSC is not always the best defense against this attack; they also showed that there is no unique optimum watermark power spectrum over the entire range of attack distortions.

Next, an optimized defense (watermark power spectrum) against the optimum attack was investigated. Because of difficulties in finding an analytical solution, and because the defense is intimately tied to the attack, numerical optimization methods were applied. Like the optimum attack, the optimized defense is difficult to describe precisely. However, experimental results with the MSE distortion measure produced a rule-of-thumb for the defense: "*White watermarks perform nearly optimally at low distortion, and PSC-compliant watermarks perform nearly optimally at high distortions.*" These results agree with the description of the attack behavior because a white watermark resists additive noise well, while a PSC-compliant watermark resists frequency-selective filtering well. For applications where only mild attack distortions must be tolerated, a white watermark is preferable because it should provide a higher capacity than a PSC-compliant watermark. For applications where communication must be possible even at high attack distortions, a PSC-compliant watermark is more suitable because it should offer a greater capacity than a white watermark.

Also, the optimized watermarks distribute their power over the entire frequency support of the original signal's power spectrum. They leave a small amount of power at frequency components where the original power is small. Doing so forces the attack to spread its effort over all frequencies.

When the receiver fails to suppress all of the interference from the original, the original acts like additional channel noise. The optimized watermark power spectrum strikes a balance between resisting the attack, which is hostile and adaptive, and water-filling to resist original-signal interference, which is coincidental and passive. Experiments with partial original-interference suppression, likely in practical blind watermarking schemes, indicate that significant capacity gains over conventional blind reception are possible.

Additional experiments indicate that there can be a significant performance difference between the cases of a memoryless original and an original with memory. When the original signal is highly correlated, modeling it as memoryless may overestimate the capacity of the watermarking system. Likewise, modeling the attack as additive noise or effective white-noise can lead to a large discrepancy between predicted and actual capacity-distortion performance. The optimum attack is more powerful than either of these attacks.

Finally, experiments with frequency-weighted MSE distortion generalize the results for MSE distortion in a simple way: "At low perceptual distortions, perceptually white watermarks have nearly optimal performance, and at high perceptual distortions, perceptually PSC-compliant watermarks have nearly optimal performance." For applications where only mild attacks must be resisted, perceptually white watermarks are desirable. For robustness at high perceptual distortions, these observations strongly encourage the heuristic rule of Cox *et al.* [15] and others (e.g., [49, 24, 47]) that the watermark should be embedded in the "perceptually significant frequency components."

5.2 Remarks and Practical Implications

Because of the theoretical nature of this paper, some remarks on its significance for practical watermarking schemes are in order. Many of the assumptions in the analysis are ideal; nevertheless, the results can also provide helpful insights and useful tools for constructing and evaluating practical watermarking systems. These remarks are of a more speculative nature than the rest of this paper.

First, we have applied Kerckhoff's principle from both the owner's and attacker's viewpoints. Of course, the former viewpoint should be used in the responsible design of a watermarking scheme: The designer should be pessimistic and assume that the attacker has complete knowledge of the statistics of the original and the watermark. However, the latter viewpoint is optimistic: The owner is assumed to have complete knowledge of the attack filter and noise statistics. Using this knowledge, the watermark receiver compensates for the attack. In practice, it is unlikely that the watermark receiver will be fortunate enough to have such accurate knowledge. Consequently, the results in this paper represent upper bounds on performance.

Second, the results (e.g., see Figs. 7 and 15) indicate very low capacities when the attack distortion becomes large. At such distortions, thousands of samples may be required to communicate a single information bit. One should be careful not to draw any sweeping conclusions from our theoretical analysis, but it does suggest that it could be difficult or even impossible to communicate a significant amount of information if the attack distortion is high and the number of signal samples is severely limited. Such a conclusion could have important consequences for practical image and audio watermarking schemes. In contrast, the number of available samples is virtually unlimited in video watermarking; however, synchronization and production costs could be problematic for such long signals.

Third, the analysis shows that there is not a unique watermark power spectrum that provides the best performance over the entire range of attack distortons. This result implies an unavoidable "you can't have it all" trade-off: A single watermarking strategy (e.g., white or PSC-compliant) cannot achieve the highest communication rate or information payload at both low and high distortions. For example, if a system uses a PSC-compliant watermark to maintain communication at high distortion, but the attack is less severe than anticipated, then the payload will be lower than if the system had employed a white watermark.

This trade-off does *not* mean that PSC-compliant watermarks are superior to white watermarks, or vice versa. Rather, it means that the choice of watermark power spectrum is highly application-dependent. For example, consider applications like embedding meta-information or broadcast monitoring. It may not be important if the meta-information cannot be retrieved after mild signal degradations. Likewise, a broadcasting site is unlikely to introduce much distortion intentionally, and it does not matter if consumers later process the watermarked data after broadcast. For such applications, a white watermark may be appropriate. However, in applications such as access control and the protection of intellectual property rights, the hidden information should also be decodable even at extremely high distortions. PSC-compliant watermarks are more suitable for these applications.

Fourth, although ordinary MSE and frequency-weighted MSE are imperfect distortion measures for real data, the analysis may still provide a useful guideline for practical watermarking schemes that employ more accurate perceptual models. A watermark whose embedding distortion is spread fairly uniformly over the original signal would be analogous to a perceptually white watermark; it would likely provide a large payload but low robustness. On the other hand, a watermark whose embedding distortion is concentrated in the perceptually significant portions of the original signal would be analogous to a perceptually PSC-compliant watermark; it would probably yield a small payload but high robustness.

Finally, we remark that the presented optimum attack is not a purely theoretical entity; it could actually be implemented by an attacker. Even in practice, an attacker may be able to acquire reasonably accurate knowledge of the statistics of the original and watermark signals, and then the attack could easily be applied. The optimum attack may thus be a useful tool for evaluating practical watermarking schemes; it has recently been applied to an image watermarking scheme in [20].

Acknowledgment

The authors thank the anonymous reviewers for their helpful comments, which have improved the quality of this paper.

A Effective White-Noise Attack

For this attack, the attacker should set

$$\Phi_{n_i n_i}(\vec{\omega}) = \frac{\sigma_{n_i}^2}{\sigma_w^2}\Phi_{ww}(\vec{\omega}). \tag{34}$$

Then $C = \frac{1}{2}\log_2\left(1 + \sigma_w^2/\sigma_{n_i}^2\right)$. Since σ_w^2 is fixed, the attacker can ensure that $C = C_t$ by selecting $\sigma_{n_i}^2$ appropriately. From (10), it follows that

$$\Phi_{vv}(\vec{\omega}) = |G(\vec{\omega})|^2\left[\frac{\sigma_{n_i}^2}{\sigma_w^2}\Phi_{ww}(\vec{\omega}) - (1-a)^2\Phi_{xx}(\vec{\omega})\right], \tag{35}$$

so that once $G(\vec{\omega})$ is known, $\Phi_{vv}(\vec{\omega})$ is also specified. Regardless of $G(\vec{\omega})$, to ensure that $\Phi_{vv}(\vec{\omega})$ remains non-negative at all frequencies, $\sigma_{n_i}^2$ is restricted by

$$\sigma_{n_i}^2 \geq (1-a)^2\sigma_w^2 \max_{\vec{\omega}\in\mathcal{W}} \frac{\Phi_{xx}(\vec{\omega})}{\Phi_{ww}(\vec{\omega})}. \tag{36}$$

The distortion expression (5) becomes

$$D_{\hat{y}x} = \frac{1}{(2\pi)^M}\int_\Omega |F(\vec{\omega})|^2\left[\left(|G(\vec{\omega})-1|^2 - (1-a)^2|G(\vec{\omega})|^2\right)\Phi_{xx}(\vec{\omega}) + |G(\vec{\omega})|^2 K\Phi_{ww}(\vec{\omega})\right]d\vec{\omega}, \tag{37}$$

where $K = 1 + \sigma_{n_i}^2/\sigma_w^2$. Write $G(\vec{\omega})$ in magnitude-phase form, $G(\vec{\omega}) = |G(\vec{\omega})|e^{j\theta(\vec{\omega})}$, and substitute this form into (37) to obtain

$$\begin{aligned} D_{\hat{y}x} = \frac{1}{(2\pi)^M}\int_\Omega |F(\vec{\omega})|^2\Big[&\left(|G(\vec{\omega})|^2 - 2|G(\vec{\omega})|\cos\theta(\vec{\omega}) + 1 - (1-a)^2|G(\vec{\omega})|^2\right)\Phi_{xx}(\vec{\omega}) \\ &+ |G(\vec{\omega})|^2 K\Phi_{ww}(\vec{\omega})\Big]d\vec{\omega}. \end{aligned} \tag{38}$$

Let $D'_{\hat{y}x}$ denote the integrand of (38). To minimize $D_{\hat{y}x}$, compute the partial derivatives of $D'_{\hat{y}x}$ with respect to $\theta(\vec{\omega})$ and $|G(\vec{\omega})|$, and set them equal to zero. First,

$$\frac{\partial D'_{\hat{y}x}}{\partial\theta(\vec{\omega})} = |F(\vec{\omega})|^2 \times 2|G(\vec{\omega})|\sin\theta(\vec{\omega})\Phi_{xx}(\vec{\omega}) = 0.$$

Thus, $\theta(\vec{\omega}) = k\pi$, and $\cos\theta(\vec{\omega}) = \pm 1$. Second,

$$\frac{\partial D'_{\hat{y}x}}{\partial|G(\vec{\omega})|} = |F(\vec{\omega})|^2 \times \left[\left(2|G(\vec{\omega})| - 2\cos\theta(\vec{\omega}) - 2(1-a)^2|G(\vec{\omega})|\right)\Phi_{xx}(\vec{\omega}) + 2|G(\vec{\omega})|K\Phi_{ww}(\vec{\omega})\right] = 0. \tag{39}$$

Solving (39) for $|G(\vec{\omega})|$ gives

$$|G(\vec{\omega})| = \frac{\Phi_{xx}(\vec{\omega})}{a(2-a)\Phi_{xx}(\vec{\omega}) + K\Phi_{ww}(\vec{\omega})} \cos\theta(\vec{\omega}).$$

Since $|G(\vec{\omega})| \geq 0$, $\cos\theta(\vec{\omega}) = 1$; hence, $G(\vec{\omega}) = |G(\vec{\omega})|$, and choose $\theta(\vec{\omega}) = 0$, $\forall\vec{\omega}$. Thus, the solution for $G(\vec{\omega})$ is

$$G(\vec{\omega}) = \frac{\Phi_{xx}(\vec{\omega})}{a(2-a)\Phi_{xx}(\vec{\omega}) + K\Phi_{ww}(\vec{\omega})}. \tag{40}$$

$\Phi_{vv}(\vec{\omega})$ can then be computed from (35).

The attack distortion (37) becomes

$$D_{\hat{y}x} = P_x - \frac{1}{(2\pi)^M} \int_\Omega |F(\vec{\omega})|^2 \frac{\Phi_{xx}^2(\vec{\omega})}{a(2-a)\Phi_{xx}(\vec{\omega}) + K\Phi_{ww}(\vec{\omega})} \, d\vec{\omega}. \tag{41}$$

B Power-Spectrum Condition

Let $G(\vec{\omega})$ be given by (40). Then apply the calculus of variations with the Lagrangian

$$J = |F(\vec{\omega})|^2 \left[\left((G(\vec{\omega}) - 1)^2 - (1-a)^2 G^2(\vec{\omega}) \right) \Phi_{xx}(\vec{\omega}) + G^2(\vec{\omega}) K \Phi_{ww}(\vec{\omega}) \right] + \lambda |F(\vec{\omega})|^2 \Phi_{ww}(\vec{\omega}).$$

Next,

$$\frac{dJ}{d\Phi_{ww}(\vec{\omega})} = \frac{|F(\vec{\omega})|^2 K \Phi_{xx}^2(\vec{\omega})}{\left(a(2-a)\Phi_{xx}(\vec{\omega}) + K\Phi_{ww}(\vec{\omega})\right)^2} + \lambda |F(\vec{\omega})|^2 = 0.$$

Solving this equation for $\Phi_{ww}(\vec{\omega})$ yields

$$\Phi_{ww}(\vec{\omega}) = \frac{1}{\sqrt{K}} \left(\pm \frac{1}{\sqrt{-\lambda}} - \frac{a(2-a)}{\sqrt{K}} \right) \Phi_{xx}(\vec{\omega}). \tag{42}$$

It is already evident that $\Phi_{ww}(\vec{\omega})$ is directly proportional to $\Phi_{xx}(\vec{\omega})$.

For completeness, continue analysis to verify that $\Phi_{ww}(\vec{\omega})$ is always a valid power spectrum. Since power spectra are real, select $\lambda < 0$. Since power spectra are non-negative, choose the plus-case and then set $\frac{1}{\sqrt{-\lambda}} - \frac{a(2-a)}{\sqrt{K}} \geq 0$. Solving for λ gives $\lambda \geq -K/a^2(2-a)^2$. Hence, $0 < -\lambda \leq K/a^2(2-a)^2$, and let $-\lambda = \mu^2 K/a^2(2-a)^2$, where μ must satisfy $0 < \mu \leq 1$ for $\Phi_{ww}(\vec{\omega})$ to be valid. Then $\Phi_{ww}(\vec{\omega})$ becomes

$$\Phi_{ww}(\vec{\omega}) = \frac{a(2-a)}{K} \left(\frac{1-\mu}{\mu} \right) \Phi_{xx}(\vec{\omega}) = \frac{\sigma_w^2}{\sigma_x^2} \Phi_{xx}(\vec{\omega}).$$

Solving for μ gives

$$\mu = \frac{a(2-a)\sigma_x^2}{a(2-a)\sigma_x^2 + K\sigma_w^2}.$$

Recall $0 \leq a \leq 1$ and $K = 1 + \sigma_{n_i}^2/\sigma_w^2$, so μ always satisfies $0 < \mu \leq 1$. Consequently, $\Phi_{ww}(\vec{\omega})$ in (42) is always a valid power spectrum and is directly proportional to $\Phi_{xx}(\vec{\omega})$.

C Capacity Expression for Effective White-Noise Attack and PSC

Since $\Phi_{ww}(\vec{\omega})$ is PSC-compliant, substitute $\Phi_{ww}(\vec{\omega}) = (\sigma_w^2/\sigma_x^2)\Phi_{xx}(\vec{\omega})$ into the distortion expression (41), which yields

$$\begin{aligned}
D_{\hat{y}x} &= P_x - \frac{1}{(2\pi)^M}\int_\Omega |F(\vec{\omega})|^2 \frac{\Phi_{xx}^2(\vec{\omega})}{a(2-a)\Phi_{xx}(\vec{\omega}) + K\frac{\sigma_w^2}{\sigma_x^2}\Phi_{xx}(\vec{\omega})}\,d\vec{\omega} \\
&= P_x - \frac{\sigma_x^2}{a(2-a)\sigma_x^2 + K\sigma_w^2} \times \frac{1}{(2\pi)^M}\int_\Omega |F(\vec{\omega})|^2\Phi_{xx}(\vec{\omega})\,d\vec{\omega} \\
&= P_x\left[1 - \frac{\sigma_x^2}{a(2-a)\sigma_x^2 + \sigma_w^2 + \sigma_{n_i}^2}\right], \qquad (43)
\end{aligned}$$

where the last line follows because $K = 1 + \sigma_{n_i}^2/\sigma_w^2$. Next, solve (43) for $\sigma_{n_i}^2$, which produces

$$\sigma_{n_i}^2 = \frac{P_x\sigma_x^2 - (P_x - D_{\hat{y}x})\left(a(2-a)\sigma_x^2 + \sigma_w^2\right)}{P_x - D_{\hat{y}x}}.$$

The capacity is $C = \frac{1}{2}\log_2\left(1 + \sigma_w^2/\sigma_{n_i}^2\right)$, so write

$$\begin{aligned}
\frac{\sigma_w^2}{\sigma_{n_i}^2} &= \frac{(P_x - D_{\hat{y}x})\sigma_w^2}{P_x\sigma_x^2 - (P_x - D_{\hat{y}x})\left(a(2-a)\sigma_x^2 + \sigma_w^2\right)} \times \frac{P_x/\sigma_x^2}{P_x/\sigma_x^2} \\
&= \frac{(P_x - D_{\hat{y}x})\frac{\sigma_w^2}{\sigma_x^2}P_x}{P_x^2 - (P_x - D_{\hat{y}x})\left(a(2-a)P_x + \frac{\sigma_w^2}{\sigma_x^2}P_x\right)}.
\end{aligned}$$

Because $\Phi_{ww}(\vec{\omega})$ is PSC-compliant, $D_{\text{embed}} = (\sigma_w^2/\sigma_x^2)P_x$, and (13) follows.

D Optimum Attack

Let $G(\vec{\omega}) = |G(\vec{\omega})|e^{j\theta(\vec{\omega})}$. Write the integrand of $D_{\hat{y}x}$ in (5) as

$$D'_{\hat{y}x} = |F(\vec{\omega})|^2\left[\left(|G(\vec{\omega})|^2 - 2|G(\vec{\omega})|\cos\theta(\vec{\omega}) + 1\right)\Phi_{xx}(\vec{\omega}) + |G(\vec{\omega})|^2\Phi_{ww}(\vec{\omega}) + \Phi_{vv}(\vec{\omega})\right],$$

and the integrand of C in (7) as $C'/\ln 2$, where

$$C' = \frac{1}{2}\ln\left(1 + \frac{|G(\vec{\omega})|^2\Phi_{ww}(\vec{\omega})}{(1-a)^2|G(\vec{\omega})|^2\Phi_{xx}(\vec{\omega}) + \Phi_{vv}(\vec{\omega})}\right).$$

Then apply the calculus of variations with the Lagrangian $J = D'_{\hat{y}x} + (\lambda/\ln 2)C'$.

First, we have $\partial J/\partial\theta(\vec{\omega}) = 2|G(\vec{\omega})|\sin\theta(\vec{\omega})\Phi_{xx}(\vec{\omega}) = 0$, so $\theta(\vec{\omega}) = k\pi$, and $\cos\theta(\vec{\omega}) = \pm 1$.

Second, it is useful to compute the partial derivatives of C' before continuing. We find $\partial C'/\partial\Phi_{vv}(\vec{\omega}) = -|G(\vec{\omega})|^2\Phi_{ww}(\vec{\omega})/2\Psi(\vec{\omega})$, and $\partial C'/\partial|G(\vec{\omega})| = |G(\vec{\omega})|\Phi_{ww}(\vec{\omega})\Phi_{vv}(\vec{\omega})/\Psi(\vec{\omega})$, where

$$\Psi(\vec{\omega}) = \left((1-a)^2|G(\vec{\omega})|^2\Phi_{xx}(\vec{\omega}) + \Phi_{vv}(\vec{\omega})\right)\left((1-a)^2|G(\vec{\omega})|^2\Phi_{xx}(\vec{\omega}) + \Phi_{vv}(\vec{\omega}) + |G(\vec{\omega})|^2\Phi_{ww}(\vec{\omega})\right).$$

Then

$$\frac{\partial J}{\partial \Phi_{vv}(\vec{\omega})} = |F(\vec{\omega})|^2 + \left(\frac{\lambda}{\ln 2}\right) \frac{\partial C'}{\partial \Phi_{vv}(\vec{\omega})} = |F(\vec{\omega})|^2 - \left(\frac{\lambda}{\ln 2}\right) \frac{|G(\vec{\omega})|^2 \Phi_{ww}(\vec{\omega})}{2\Psi(\vec{\omega})} = 0,$$

which we write as

$$2\frac{|F(\vec{\omega})|^2}{|G(\vec{\omega})|} = \left(\frac{\lambda}{\ln 2}\right) \frac{|G(\vec{\omega})| \Phi_{ww}(\vec{\omega})}{\Psi(\vec{\omega})}. \tag{44}$$

We multiply both sides of (44) by $\Phi_{vv}(\vec{\omega})$ to obtain

$$2\,\frac{|F(\vec{\omega})|^2 \Phi_{vv}(\vec{\omega})}{|G(\vec{\omega})|} = \left(\frac{\lambda}{\ln 2}\right) \frac{|G(\vec{\omega})| \Phi_{ww}(\vec{\omega}) \Phi_{vv}(\vec{\omega})}{\Psi(\vec{\omega})} = \left(\frac{\lambda}{\ln 2}\right) \frac{\partial C'}{\partial |G(\vec{\omega})|}. \tag{45}$$

Third, we compute

$$\frac{\partial J}{\partial |G(\vec{\omega})|} = |F(\vec{\omega})|^2 \left[\left(2|G(\vec{\omega})| - 2\cos\theta(\vec{\omega})\right) \Phi_{xx}(\vec{\omega}) + 2|G(\vec{\omega})| \Phi_{ww}(\vec{\omega}) \right] + \left(\frac{\lambda}{\ln 2}\right) \frac{\partial C'}{\partial |G(\vec{\omega})|} = 0.$$

From (45),

$$2|F(\vec{\omega})|^2 \left[\left(|G(\vec{\omega})| - \cos\theta(\vec{\omega})\right) \Phi_{xx}(\vec{\omega}) + |G(\vec{\omega})| \Phi_{ww}(\vec{\omega}) \right] + 2\frac{|F(\vec{\omega})|^2 \Phi_{vv}(\vec{\omega})}{|G(\vec{\omega})|} = 0.$$

Solving for $\Phi_{vv}(\vec{\omega})$ produces

$$\Phi_{vv}(\vec{\omega}) = |G(\vec{\omega})| \cos\theta(\vec{\omega}) \Phi_{xx}(\vec{\omega}) - |G(\vec{\omega})|^2 \left(\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})\right) \geq 0, \tag{46}$$

where the inequality has been added to ensure that $\Phi_{vv}(\vec{\omega})$ is a valid power spectrum.

To satisfy this inequality, we must have $\cos\theta(\vec{\omega}) = +1, \forall \vec{\omega}$, so $G(\vec{\omega}) = |G(\vec{\omega})| \geq 0, \forall \vec{\omega}$. We choose $\theta(\vec{\omega}) = 0, \forall \vec{\omega}$. Eq. (46) yields $G(\vec{\omega}) \leq \Phi_{xx}(\vec{\omega}) / \left(\Phi_{xx}(\vec{\omega}) + \Phi_{ww}(\vec{\omega})\right)$. We can then write $G(\vec{\omega})$ as in (15). Substituting (15) into (46) gives $\Phi_{vv}(\vec{\omega})$ in (16).

It remains to find an expression for $A(\vec{\omega})$. We substitute (15) and (16) into (44). After some algebra, we find

$$\begin{aligned} &|F|^2 \Phi_{xx}^2 \left[(1-a)^2 A \Phi_{xx} + (1-A)\left(\Phi_{xx} + \Phi_{ww}\right) \right] \\ &\qquad \times \left[(1-a)^2 A \Phi_{xx} + (1-A)\left(\Phi_{xx} + \Phi_{ww}\right) + A\Phi_{ww} \right] = \frac{\lambda}{2\ln 2} \Phi_{ww} \left(\Phi_{xx} + \Phi_{ww}\right)^2, \end{aligned} \tag{47}$$

where we have omitted the frequency variable $\vec{\omega}$.

For $0 < a \leq 1$, Eq. (47) can be written as a quadratic expression in $A(\vec{\omega})$. It has roots

$$\begin{aligned} &A(\vec{\omega}) = \\ &\left(1 + \frac{\Phi_{ww}}{2a(2-a)\Phi_{xx}} \pm \frac{\sqrt{\Phi_{xx}^2 \Phi_{ww}^2 + (2\lambda/\ln 2) a(2-a) \Phi_{xx} \Phi_{ww} \left(a(2-a)\Phi_{xx} + \Phi_{ww}\right) |F|^{-2}}}{2a(2-a)\Phi_{xx}^2} \right) \\ &\qquad \times \left(\frac{\Phi_{xx} + \Phi_{ww}}{a(2-a)\Phi_{xx} + \Phi_{ww}} \right). \end{aligned} \tag{48}$$

We select the minus-case since otherwise $A(\vec{\omega}) > 1, \forall\vec{\omega}$. Since $0 \leq A(\vec{\omega}) \leq 1, \forall\vec{\omega}$, as well, to satisfy the Kuhn-Tucker conditions [53], we impose the $\text{cl}[\cdot]$ operator (18) and arrive at (20).

In the case $a = 0$, (48) is not well-defined. Substitute $a = 0$ into (47), which produces a linear equation in $A(\vec{\omega})$; imposing the constraint $A(\vec{\omega}) \geq 0, \forall\vec{\omega}$, yields (19).

E Optimization Algorithms

We outline the optimization algorithms for the case where D_{embed} and C_t are given and $D_{\hat{y}x}$ should be maximized. They can easily be modified to replace the capacity constraint with an attack-distortion constraint $D_{\hat{y}x} = D_t$ and maximize C (see Sec. 2.5).

Initialization The algorithms require the N-vectors X and F, and the scalars D_{embed} and C_t. The initial watermark N-vector W should be selected such that $W_n \geq 0, \forall n$, and $(1/N)\sum_{n=1}^{N} F_n^2 W_n = D_{\text{embed}}$; the latter condition enforces the embedding distortion constraint. Typically, we set $W_n = D_{\text{embed}}/F_n^2$, $\forall n$, so that the initial embedding distortion is distributed evenly over all subchannels.

Bisection Search for Optimum Attack Given W, it is necessary to find the optimum attack that minimizes $D_{\hat{y}x}$ such that the capacity $C = C_t$. With all other parameters fixed, the attack is parameterized by λ. Because $C(\lambda)$ is a decreasing function of λ (Sec. 3.2.1), a simple bisection search can be employed to find λ^* such that $|C(\lambda^*) - C_t|/C_t < \varepsilon_C$ for some small tolerance $\varepsilon_C > 0$. Once λ^* has been determined, it is a simple matter to compute the corresponding attack distortion $D_{\hat{y}x}(\lambda^*)$.

If the alternate approach to the attack is desired (minimize C such that $D_{\hat{y}x} = D_t$), a bisection search can also be used since $D_{\hat{y}x}(\lambda)$ is an increasing function of λ. After λ^* has been found such that $|D_{\hat{y}x}(\lambda^*) - D_t|/D_t < \varepsilon_D$, for some $\varepsilon_D > 0$, the corresponding capacity $C(\lambda^*)$ can be computed.

Perturbation When perturbing the watermark vector W, it is useful to work with the embedding distortion vector D, whose elements are $D_n = F_n^2 W_n$, $\forall n$. The perturbed watermark distortion vector D' is given by $D' = D + \Delta D$, where ΔD is the perturbation vector with elements ΔD_n. Then the perturbed watermark W' has elements $W_n' = D_n'/F_n^2$, $\forall n$. To fulfill the constraint on D_{embed}, ΔD must satisfy $(1/N)\sum_{n=1}^{N} \Delta D_n = 0$. To ensure that $W_n' \geq 0$, $\forall n$, the elements ΔD_n must also satisfy $\Delta D_n \geq -D_n, \forall n$. In actuality, W_n' is restricted such that $W_n' \geq \varepsilon$, where ε is a small, positive number (e.g., $\varepsilon = 10^{-6}$). An explanation for this restriction appears in Sec. 4.3.1.

In the following, T corresponds to temperature in the annealing schedule. For the GMA algorithm, one element ΔD_n is set equal to $\pm T$, and the other $(N-1)$ elements of ΔD are set to

$\mp T/(N-1)$. For the annealing methods, there are two different ways for generating perturbations. In the first method, referred to as *SA/normal* or *GA/normal*, ΔD is produced by concatenating N random deviates from a normally distributed random-number generator with standard deviation T. In the second method, *SA/scaled* or *GA/scaled*, the elements of ΔD are $\Delta D_n = \varepsilon_n D_n$, where ε_n is the output of a uniform-$[-T, T]$ random-number generator and $0 < T < 1$; hence $D'_n = (1+\varepsilon_n)D_n$.

Decision After a perturbation has been made, the bisection search is used to re-optimize the attack and find λ' such that $|C'(\lambda') - C_t|/C_t < \varepsilon_C$; then $D'_{\hat{y}x}(\lambda')$ is computed. Each algorithm must decide whether or not to accept the perturbed watermark W' or keep the previous watermark W.

The GMA algorithm tries all single-subchannel perturbations and selects the one that produced the greatest value of $D'_{\hat{y}x} > D_{\hat{y}x}$; then $W \leftarrow W'$ and $D_{\hat{y}x} \leftarrow D'_{\hat{y}x}$. If all single-subchannel perturbations result in $D'_{\hat{y}x} < D_{\hat{y}x}$, the algorithm stops.

The SA algorithms use the Metropolis decision rule [37]. If $D'_{\hat{y}x} > D_{\hat{y}x}$, the perturbation is accepted, and $W \leftarrow W'$, $D_{\hat{y}x} \leftarrow D'_{\hat{y}x}$. Otherwise, the algorithm computes $p = \exp\big[\big(D'_{\hat{y}x} - D_{\hat{y}x}\big)/KT\big]$, where $K > 0$ is the temperature constant, and picks a random number r uniformly distributed over $[0, 1]$. If $r < p$, the perturbation is accepted and W and $D_{\hat{y}x}$ are updated; if $r \geq p$, the perturbation is rejected.

The GA algorithms make a random perturbation, but only accept the perturbation if $D'_{\hat{y}x} > D_{\hat{y}x}$. This is analogous to setting $K = 0$ in the corresponding SA algorithms.

Annealing Schedule All the algorithms, including GMA, employ an annealing schedule that gradually reduces the temperature T according to $T \leftarrow cT$, where $c \in (0, 1)$ is the cooling factor. For a given temperature T, many perturbations (typically $100N$) are performed before reducing the temperature by c. If none of the perturbations are accepted at a single temperature, the algorithms stop.

In the annealing algorithms, if a large number (e.g., $10N$) of the perturbations are accepted, T is immediately reduced. The logic is that the system is "too hot" and simply jumping randomly from one state (watermark vector W) to another; hence, it is reasonable to cool the system prematurely.

References

[1] *Proc. IEEE Intl. Conf. Image Proc.*, Lausanne, Switzerland, Sep. 1996.

[2] *Proc. IEEE Intl. Conf. Image Proc.*, Santa Barbara, CA, USA, Oct. 1997.

[3] *Prelim. Proc. Third Intl. Information Hiding Workshop*, Dresden, Germany, Sep.–Oct. 1999.

[4] *Proc. SPIE Security & Watermarking Multimedia Contents*, Vol. 3657, San Jose, CA, USA, Jan. 1999.

[5] *Proc. SPIE Security & Watermarking Multimedia Contents II*, Vol. 3971, San Jose, CA, USA, Jan. 2000.

[6] M. Alghoniemy and A. H. Tewfik, "Geometric distortion correction in image watermarking," in [5], pp. 82–89.

[7] G. W. Braudaway and F. C. Mintzer, "Automatic recovery of invisible image watermarks from geometrically distorted images," in [5], pp. 74–81.

[8] B. Chen and G. Wornell, "Preprocessed and postprocessed quantization index modulation methods for digital watermarking," in [5], pp. 48–59.

[9] B. Chen and G. W. Wornell, "Dither modulation: A new approach to digital watermarking and information embedding," in [4], pp. 342–353.

[10] B. Chen and G. W. Wornell, "Provably robust digital watermarking," in *Proc. SPIE Multimedia Systems and Applications II*, Vol. 3845, pp. 43–54, 1999.

[11] J. Chou, S. Pradhan, L. El Ghaoui, and K. Ramchandran, "A robust optimization solution to the data hiding problem using distributed source coding principles," in *Proc. SPIE Image and Video Communications and Processing 2000*, Vol. 3974, pp. 270–279, San Jose, CA, USA, Jan. 2000.

[12] J. Chou, S. S. Pradhan, and K. Ramchandran, "On the duality between distributed source coding and data hiding," in *Proc. 33rd Asilomar Conf. Signals, Systems, and Computers*, Vol. 2, pp. 1503–1507, Pacific Grove, CA, USA, Oct. 1999.

[13] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, Vol. IT-29, No. 3, pp. 439–441, May 1983.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, 1991.

[15] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for images, audio, and video," in [1].

[16] I. J. Cox, M. L. Miller, and A. L. McKellips, "Watermarking as communications with side information," *Proc. IEEE*, Vol. 87, No. 7, pp. 1127–1141, Jul. 1999.

[17] G. Csurka, F. Deguillaume, J. J. K. Ó Ruandaidh, and T. Pun, "A Bayesian approach to affine transformation resistant image and video watermarking," in [3].

[18] J. Dugelay and F. A. P. Petitcolas, "Image watermarking: Possible counter-attacks against random geometric distortions," in [5], pp. 338–345.

[19] J. J. Eggers, J. K. Su, and B. Girod, "A blind watermarking scheme based on structured codebooks," in *Secure Images and Image Authentication, IEE Colloquium*, pp. 4/1–4/6, London, UK, April 2000.

[20] J. J. Eggers, J. K. Su, and B. Girod, "Robustness of a blind image watermarking scheme," in *Proc. IEEE Intl. Conf. Image Proc.*, Vancouver, Canada, Sep. 2000.

[21] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Control and Inform. Theory*, Vol. 9, No. 1, pp. 19–31, 1980.

[22] F. Hartung and B. Girod, "Watermarking of uncompressed and compressed video," *Signal Processing*, Vol. 66, pp. 283–301, 1998.

[23] F. Hartung and M. Kutter, "Multimedia watermarking techniques," *Proc. IEEE*, Vol. 87, No. 7, pp. 1079–1107, Jul. 1999.

[24] F. Hartung, J. K. Su, and B. Girod, "Spread spectrum watermarking: Malicious attacks and counterattacks," in [4], pp. 147–158.

[25] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*, John Wiley and Sons, New York, NY, USA, 1996.

[26] C. Heegard and A. A. El Gamal, "On the capacity of computer memory with defects," *IEEE Trans. Inform. Theory*, Vol. IT-29, No. 5, pp. 731–739, Sep. 1983.

[27] J. R. Hernández and F. Pérez-González, "Statistical analysis of watermarking schemes for copyright protection of images," *Proc. IEEE*, Vol. 87, No. 7, pp. 1142–1166, Jul. 1999.

[28] C.-T. Hsu and J.-L. Wu, "Hidden signatures in images," in [1], pp. 223–226.

[29] J. Huber, *Trelliscodierung*, Springer-Verlag, Berlin, Germany, 1992. In German.

[30] T. Kalker, G. Depovere, J. Haitsma, and M. J. Maes, "A video watermarking system for broadcast monitoring," in [4], pp. 103–112.

[31] T. Kalker and A. J. E. M. Janssen, “Analysis of watermark detection using SPOMF,” in *Proc. IEEE Intl. Conf. Image Proc.*, Kobe, Japan, Oct. 1999.

[32] M. Kutter and F. A. P. Petitcolas, “A fair benchmark for image watermarking systems,” in [4], pp. 226–239.

[33] T. Mittelholzer, “An information-theoretic approach to steganography and watermarking,” in [3].

[34] P. Moulin and J. A. O’Sullivan, “Information-theoretic analysis of information hiding,” Preprint, Sep. 1999.

[35] F. Peticolas and M. G. Kuhn, StirMark 2.3 watermark robustness testing software, available at URL http:// www.cl.cam.ac.uk/˜fapp2/watermarking/image_watermarking/stirmark/, Oct. 1998.

[36] A. Piva, M. Barni, F. Bartolini, and V. Cappellini, “DCT-based watermark recovering without resorting to the uncorrupted original image,” in [2], pp. 520–523.

[37] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge Univ. Press, Cambridge, United Kingdom, 1990.

[38] M. Ramkumar, *Data Hiding in Multimedia: Theory and Applications*, PhD thesis, New Jersey Institute of Technology, Kearny, NJ, USA, Nov. 1999.

[39] C. E. Shannon, “A mathematical theory of communication,” *Bell Sys. Tech. Journal*, Vol. 27, pp. 379–423, Jul. 1948.

[40] C. E. Shannon, “Communication in the presence of noise,” *Proc. IRE*, Vol. 37, pp. 10–21, 1949.

[41] C. E. Shannon, “Channels with side information at the transmitter,” *IBM J. Res. and Dev.*, Vol. 2, pp. 289–293, 1958.

[42] J. R. Smith and B. O. Comiskey, “Modulation and information hiding in images,” in *Proc. First Information Hiding Workshop*, Vol. 1174, May 1996.

[43] D. R. Stinson, *Cryptography: Theory and Practice*, CRC Press, New York, NY, USA, 1995.

[44] H. S. Stone, “Analysis of attacks of image watermarks with randomized coefficients,” tech. report, NEC Research Inst., May 1996.

[45] J. K. Su, J. J. Eggers, and B. Girod, "Channel coding and rate distortion with side information: Geometric interpretation and illustration of duality," Submitted to *IEEE Trans. Inform. Theory*, May 2000.

[46] J. K. Su, J. J. Eggers, and B. Girod, "Illustration of the duality between channel coding and rate distortion with side information," in *Proc. 34th Asilomar Conf. Signals, Systems, & Computers*, Pacific Grove, CA, USA, Oct. 29-Nov. 1, 2000.

[47] J. K. Su and B. Girod, "On the imperceptibility and robustness of digital fingerprints," In *Proc. IEEE Intl. Conf. Multimedia Computing & Systems*, Vol. 2, pp. 530–535, Florence, Italy, Jun. 1999.

[48] J. K. Su and B. Girod, "Fundamental performance limits of power-spectrum condition-compliant watermarks," in [5], pp. 314–325.

[49] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data-embedding and watermarking techniques," *Proc. IEEE*, Vol. 86, No. 6, pp. 1064–1087, Jun. 1998.

[50] P. J. M. van Laarhoven and E. H. L. Aarts, *Simulated Annealing: Theory and Applications*, Kluwer Academic, Norwell, Massachusetts, 1987.

[51] S. Voloshynovskiy, A. Herrigel, N. Baumgaertner, and T. Pun, "A stochastic approach to content adaptive digital image watermarking," in [3].

[52] R. B. Wolfgang, C. I. Podilchuk, and E. J. Delp, "Perceptual watermarks for digital images and video," *Proc. IEEE*, Vol. 87, No. 7, pp. 1108–1126, Jul. 1999.

[53] W. Woodside, "Lagrange multipliers for engineers—some applications," *Intl. J. Applied Engineering Education*, Vol. 1, No. 1, pp. 61–64, 1985.

[54] X.-G. Xia, C. G. Boncelet, and G. R. Arce, "A multiresolution watermark for digital images," in [2], pp. 548–551.

[55] W. Zhu, Z. Xiong, and Y.-Q. Zhang, "Multiresolution watermarking for images and video," *IEEE Trans. Circ. Sys. Video Tech.*, Vol. 9, No. 4, pp. 545–550, Jun. 1999.

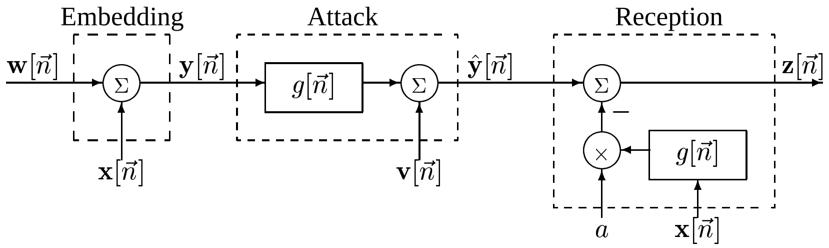


Figure 1: Block diagram of watermark embedding, LSI filtering-additive noise attack, and watermark reception. The original $\mathbf{x}[\vec{n}]$ may not be physically available to the receiver, but the watermarking system may exploit knowledge of $\mathbf{x}[\vec{n}]$ during embedding so that it behaves as if it eliminates some or all of the interference from $\mathbf{x}[\vec{n}]$. Thus, the receiver can *effectively* remove $ag[\vec{n}] * \mathbf{x}[\vec{n}]$ from $\hat{\mathbf{y}}[\vec{n}]$.

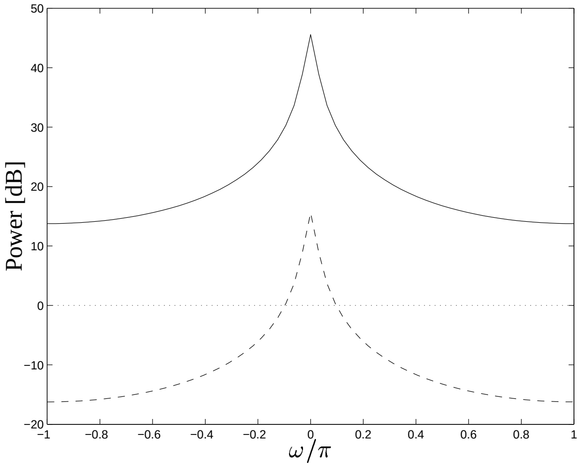


Figure 2: Example power spectra for original and watermarks used in experiments. All power spectra are shown with a decibel scale.

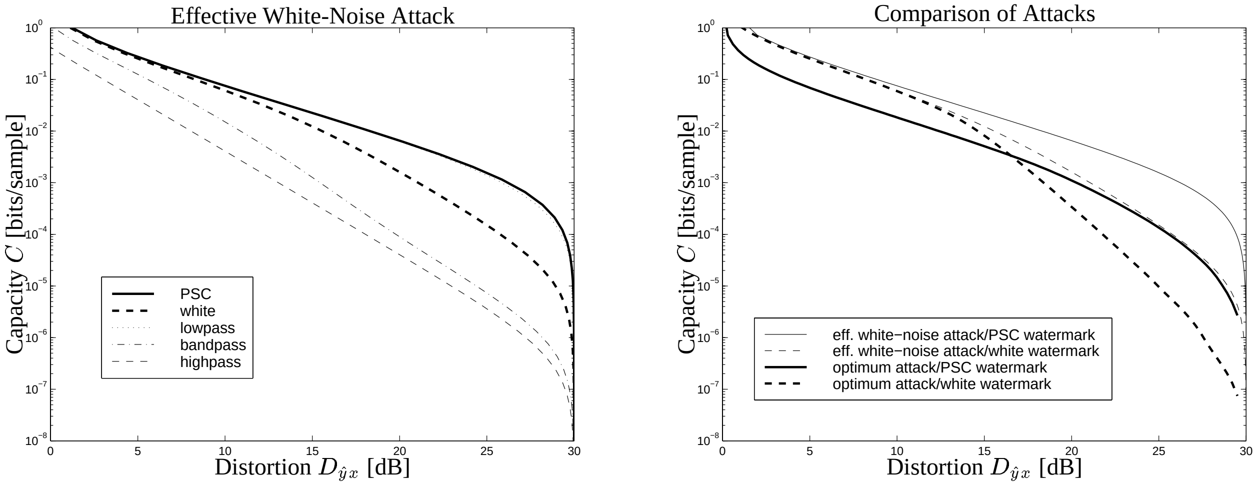
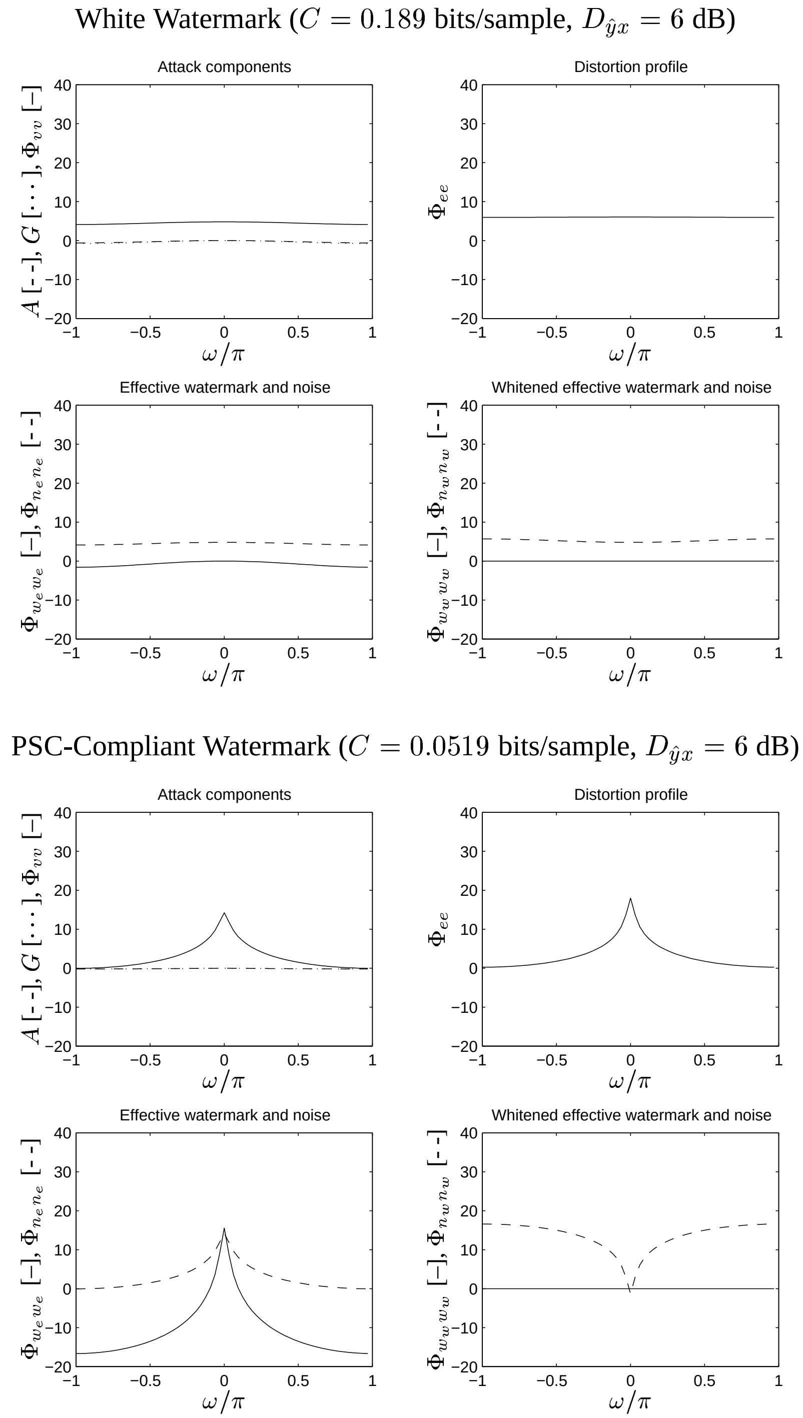


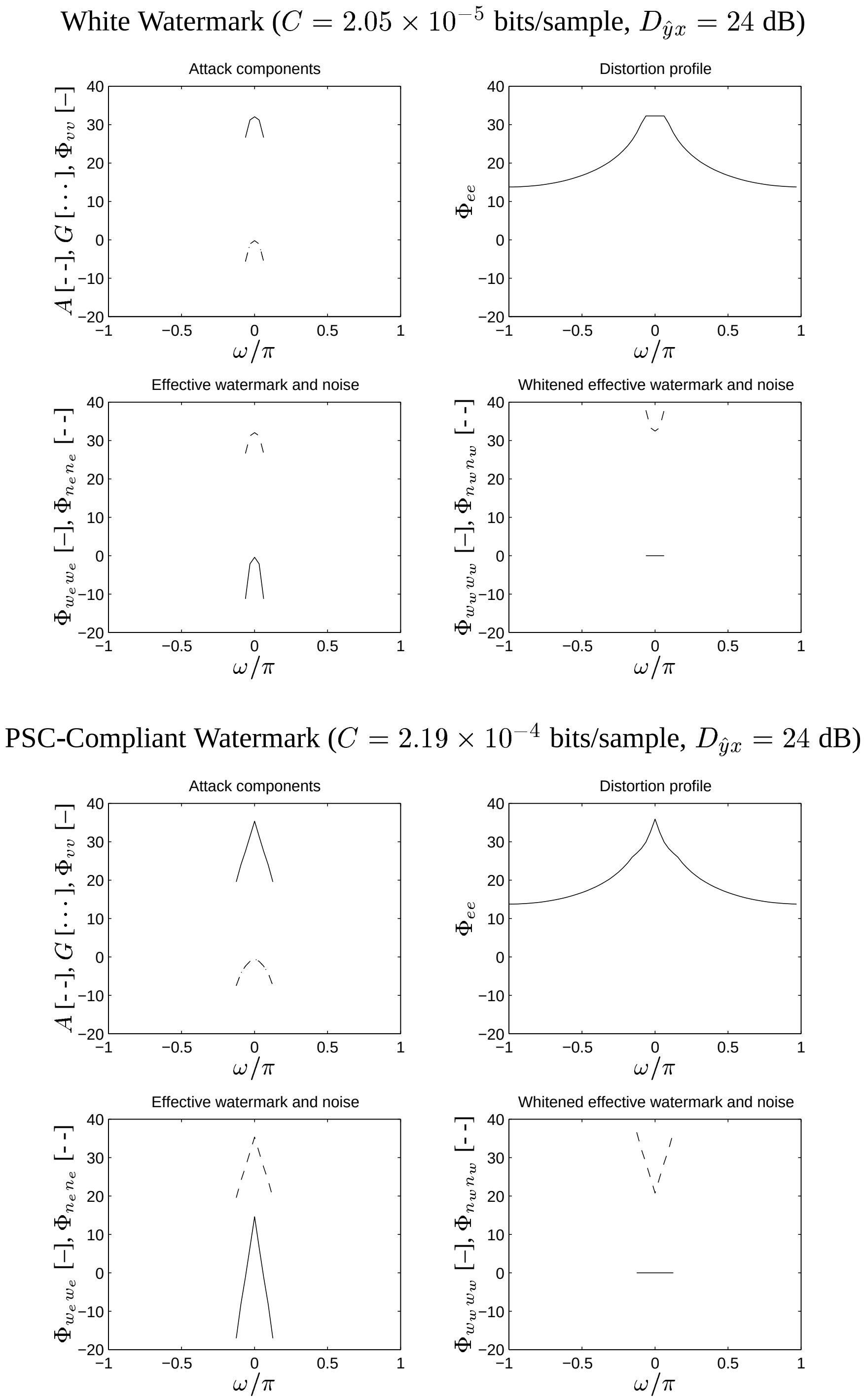
Figure 3: Comparison of effective white-noise attack and optimum attack. Left: performance of various watermarks subject to the effective white-noise attack. Right: performance of PSC-compliant and white watermarks subject to either attack. For both graphs, the original-interference suppression factor is $a = 1$.



White Watermark ($C = 0.189$ bits/sample, $D_{\hat{y}x} = 6$ dB)

PSC-Compliant Watermark ($C = 0.0519$ bits/sample, $D_{\hat{y}x} = 6$ dB)

Figure 4: Example of attack behavior for white and PSC-compliant watermarks after optimum attack at low distortion. The original-interference suppression factor is $a = 1$. All power spectra are shown with a decibel scale.



White Watermark ($C = 2.05 \times 10^{-5}$ bits/sample, $D_{\hat{y}x} = 24$ dB)

PSC-Compliant Watermark ($C = 2.19 \times 10^{-4}$ bits/sample, $D_{\hat{y}x} = 24$ dB)

Figure 5: Example of attack behavior for white and PSC-compliant watermarks after optimum attack at high distortion. The original-interference suppression factor is $a = 1$.

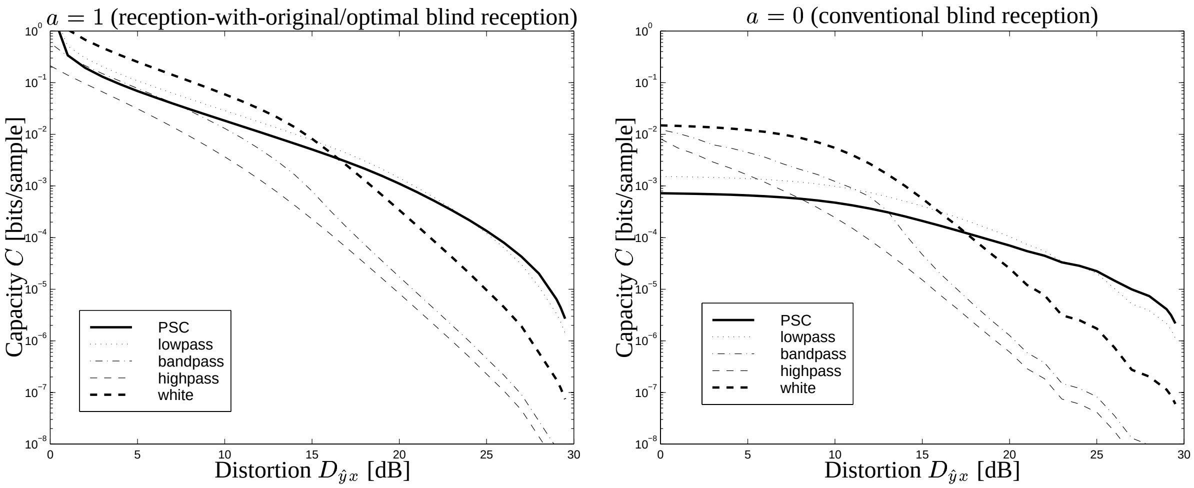


Figure 6: Performance of various fixed watermark power spectra subject to the optimum attack.

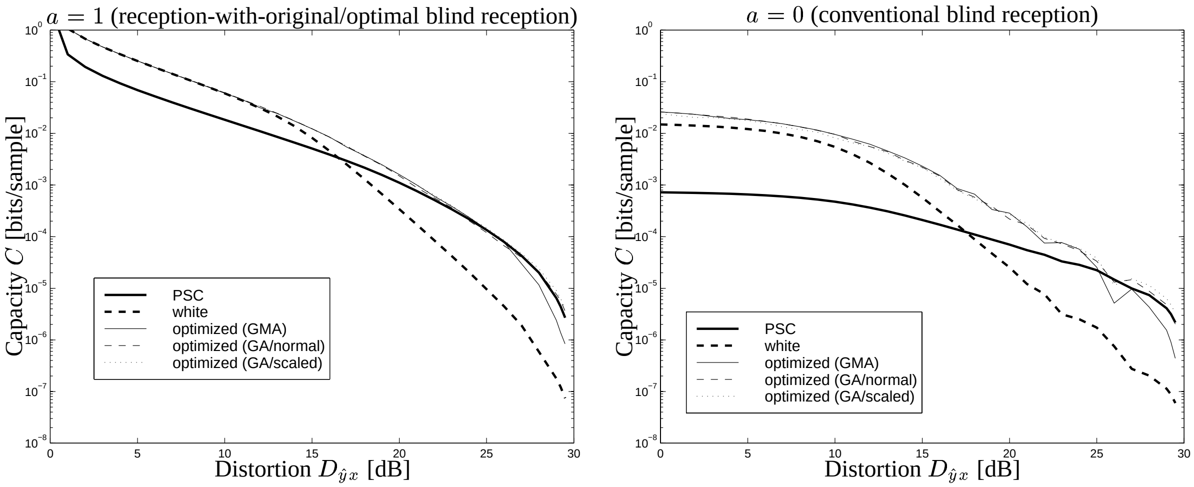
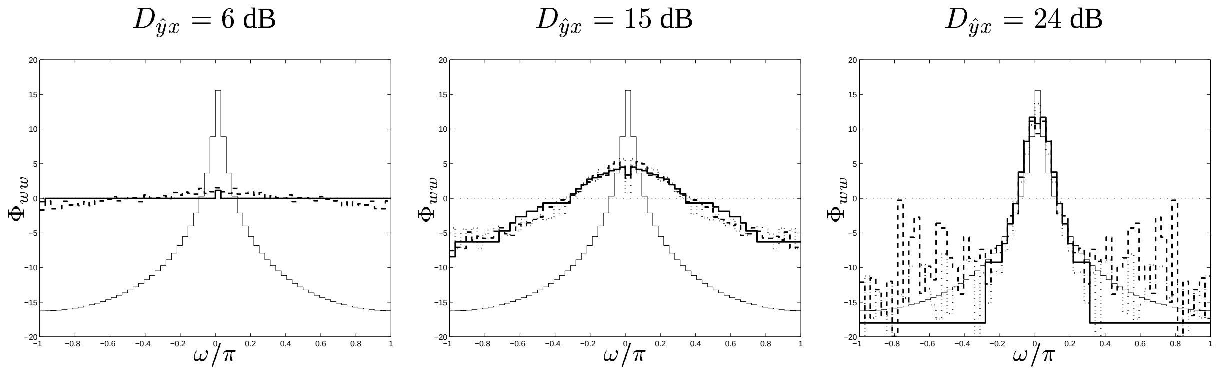
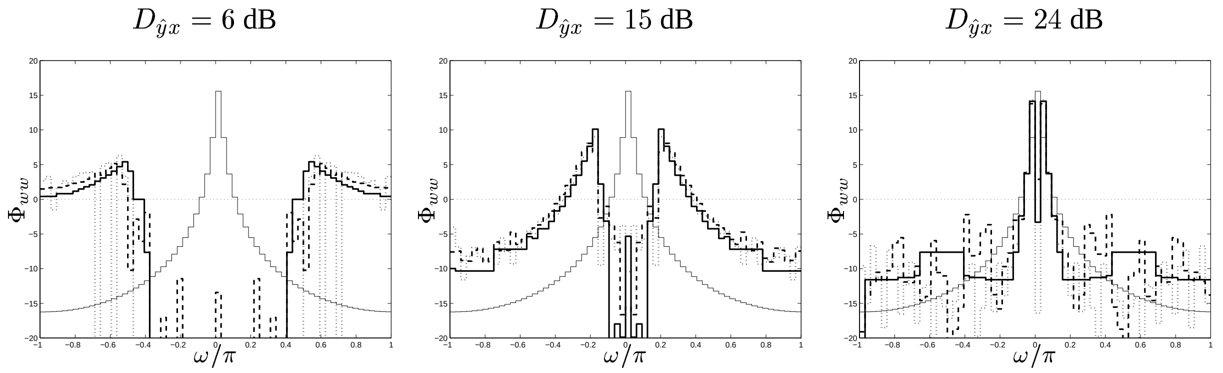


Figure 7: Performance of PSC-compliant, white, and optimized watermarks subject to the optimum attack.



$D_{\hat{y}x}$ [dB]	6 dB	15 dB	24 dB
GMA	0.189	0.0123	2.38×10^{-4}
GA/normal	0.190	0.0123	2.02×10^{-4}
GA/scaled	0.189	0.0122	2.38×10^{-4}
White	0.189	0.00818	2.05×10^{-5}
PSC	0.0519	0.00509	2.19×10^{-4}

Figure 8: Example optimized watermark power spectra for selected attack distortions $D_{\hat{y}x}$ when $a = 1$ (reception-with-original/optimal blind reception). Results for GMA [solid curve], GA/normal [dashed curve], and GA/scaled [dotted curve] are shown. For reference, thin curves corresponding to white [dotted] and PSC-compliant [solid] power spectra are also given. The accompanying table gives the corresponding capacity values C.



$D_{\hat{y}x}$ [dB]	6 dB	15 dB	24 dB
GMA	0.0169	2.32×10^{-3}	5.72×10^{-5}
GA/normal	0.0171	2.21×10^{-3}	5.33×10^{-5}
GA/scaled	0.0156	2.16×10^{-3}	5.90×10^{-5}
White	0.0111	5.71×10^{-4}	2.51×10^{-6}
PSC	0.000632	2.11×10^{-4}	2.84×10^{-5}

Figure 9: Example optimized watermark power spectra for selected attack distortions $D_{\hat{y}x}$ when $a = 0$ (conventional blind reception). Results for GMA [solid curve], GA/normal [dashed curve], and GA/scaled [dotted curve] are shown. For reference, thin curves corresponding to white [dotted] and PSC-compliant [solid] power spectra are also given. The accompanying table gives the corresponding capacity values C.

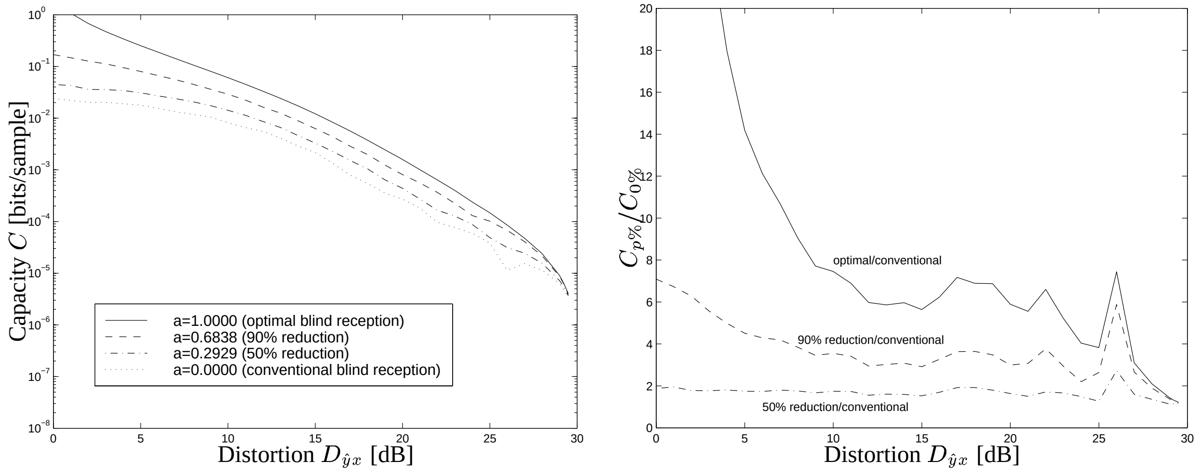


Figure 10: Performance comparison for blind watermarking schemes depending upon the original-interference suppression factor a.

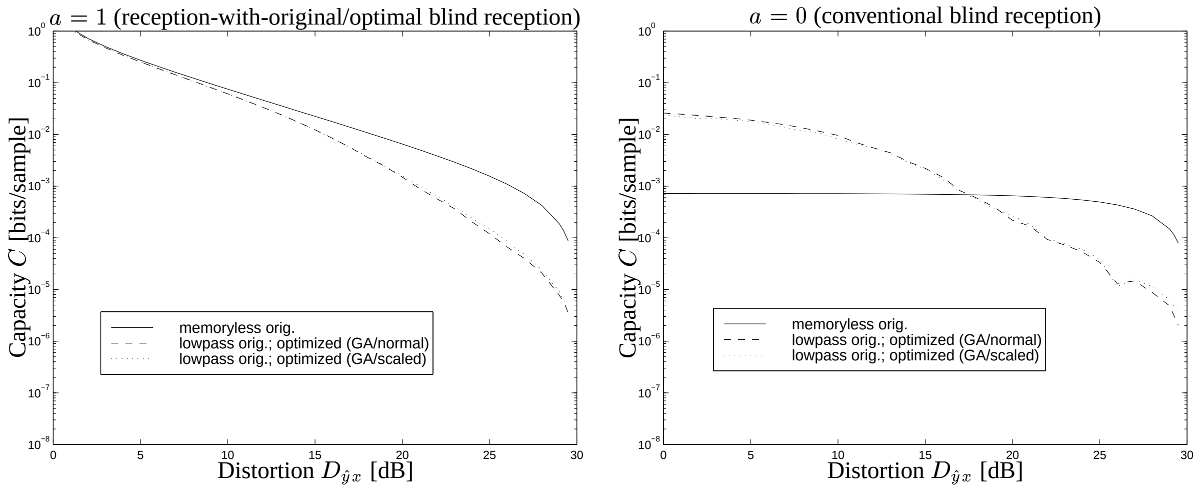


Figure 11: Performance comparison for memoryless original $\mathbf{x}[\vec{n}]$ and lowpass original $\mathbf{x}[\vec{n}]$ (AR(1), $a_1 = 0.95$).

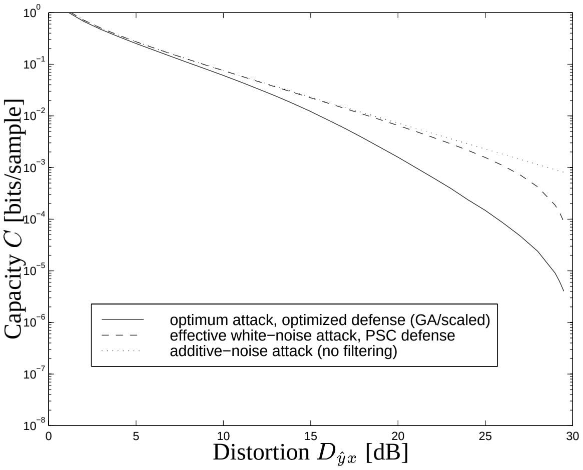


Figure 12: Performance of various attacks and their defenses. The original-interference suppression factor $a = 1$.

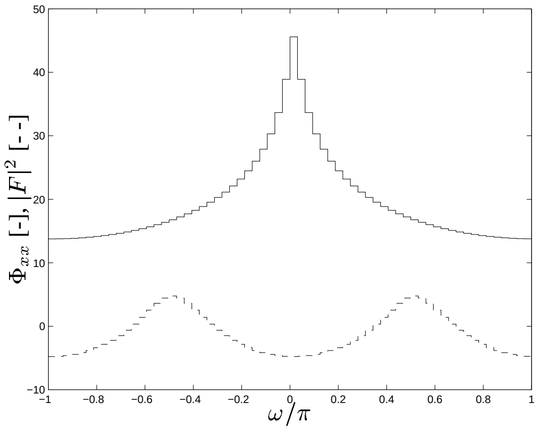
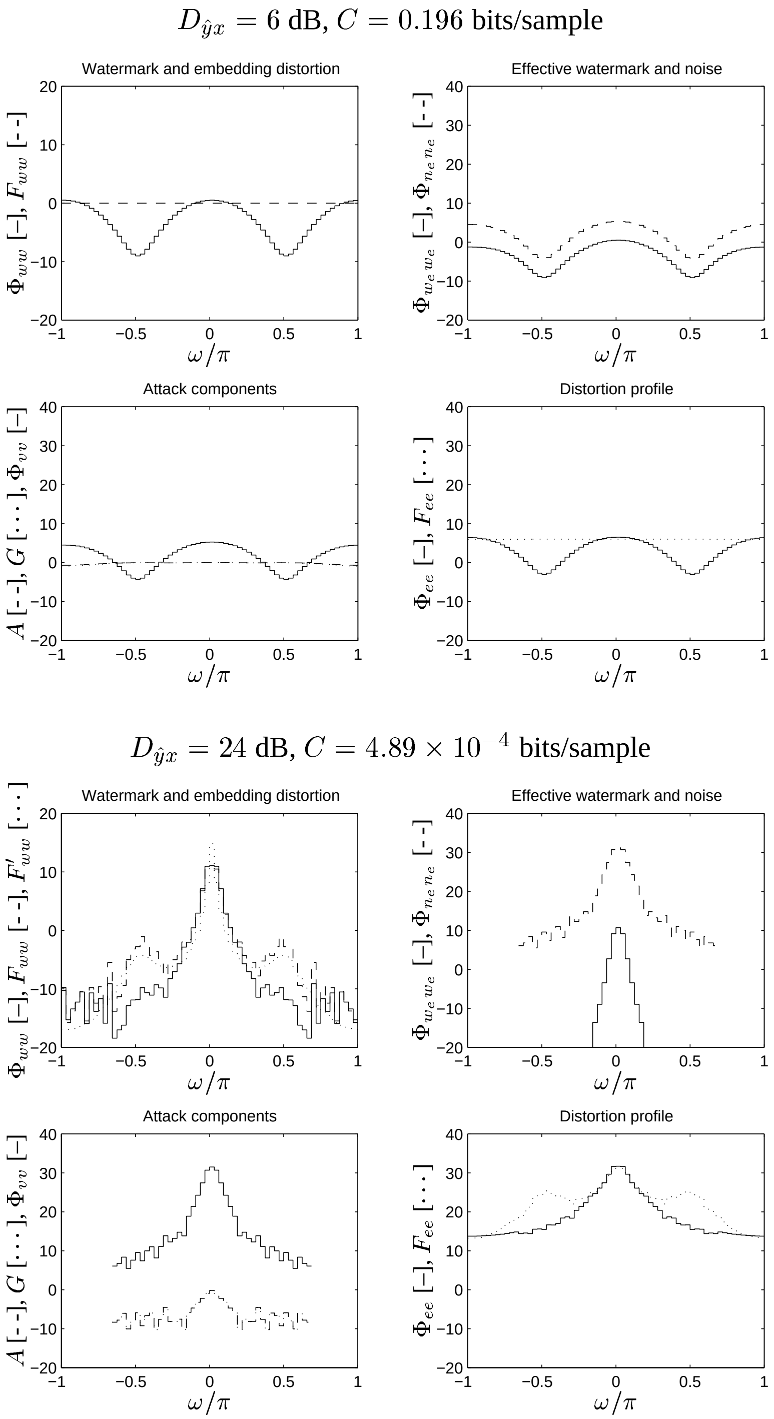


Figure 13: Piecewise-constant original power spectrum $\Phi_{xx}(\omega)$ and frequency-weighting function $|F(\omega)|^2$ used in experiments with frequency-weighted distortion.


$D_{\hat{y}x} = 6$ dB, $C = 0.196$ bits/sample

$D_{\hat{y}x} = 24$ dB, $C = 4.89 \times 10^{-4}$ bits/sample

Figure 14: Example optimized watermark power spectra from GA/scaled algorithm and attack behavior for frequency-weighted distortion. The frequency-weighted counterpart of $\Phi_{ww}(\omega)$, $F_{ww}(\omega) = |F(\omega)|^2\Phi_{ww}(\omega)$, appears in the upper-left graph for the two distortions shown; for the case $D_{\hat{y}x} = 24$ dB, the frequency-weighted PSC-compliant spectrum $F'_{ww}(\omega)$ is also shown as a dotted curve. The frequency-weighted version of the error power spectrum $\Phi_{ee}(\omega)$, $F_{ee}(\omega) = |F(\omega)|^2\Phi_{ee}(\omega)$, appears in the lower-right graph for each distortion. The original-interference suppression factor is $a = 1$.

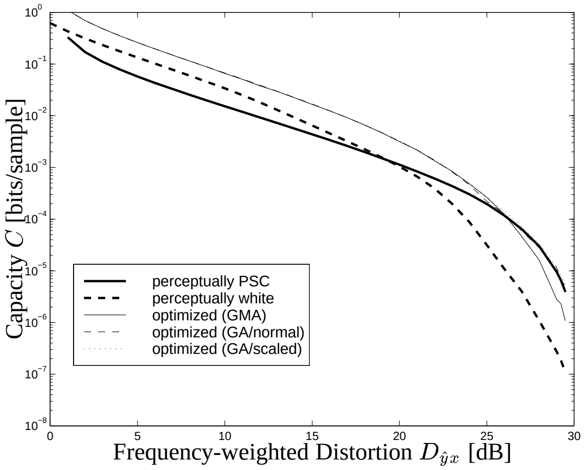


Figure 15: Performance of PSC-compliant, white, and optimized watermarks subject to the optimum attack. The original-interference suppression factor $a = 1$. The distortion measure is *frequency-weighted* MSE.

Figure Captions

Figure 1

Block diagram of watermark embedding, LSI filtering-additive noise attack, and watermark reception. The original $\mathbf{x}[\vec{n}]$ may not be physically available to the receiver, but the watermarking system may exploit knowledge of $\mathbf{x}[\vec{n}]$ during embedding so that it behaves as if it eliminates some or all of the interference from $\mathbf{x}[\vec{n}]$. Thus, the receiver can *effectively* remove $ag[\vec{n}] * \mathbf{x}[\vec{n}]$ from $\hat{\mathbf{y}}[\vec{n}]$.

Figure 2

Example power spectra for original and watermarks used in experiments. All power spectra are shown with a decibel scale.

Figure 3

Comparison of effective white-noise attack and optimum attack. Left: performance of various watermarks subject to the effective white-noise attack. Right: performance of PSC-compliant and white watermarks subject to either attack. For both graphs, the original-interference suppression factor is $a = 1$.

Figure 4

Example of attack behavior for white and PSC-compliant watermarks after optimum attack at low distortion. The original-interference suppression factor is $a = 1$. All power spectra are shown with a decibel scale.

Figure 5

Example of attack behavior for white and PSC-compliant watermarks after optimum attack at high distortion. The original-interference suppression factor is $a = 1$.

Figure 6

Performance of various fixed watermark power spectra subject to the optimum attack.

Figure 7

Performance of PSC-compliant, white, and optimized watermarks subject to the optimum attack.

Figure 8

Example optimized watermark power spectra for selected attack distortions $D_{\hat{y}x}$ when $a = 1$ (reception-with-original/optimal blind reception). Results for GMA [solid curve], GA/normal [dashed curve], and

GA/scaled [dotted curve] are shown. For reference, thin curves corresponding to white [dotted] and PSC-compliant [solid] power spectra are also given. The accompanying table gives the corresponding capacity values C.

Figure 9

Example optimized watermark power spectra for selected attack distortions $D_{\hat{y}x}$ when $a = 0$ (conventional blind reception). Results for GMA [solid curve], GA/normal [dashed curve], and GA/scaled [dotted curve] are shown. For reference, thin curves corresponding to white [dotted] and PSC-compliant [solid] power spectra are also given. The accompanying table gives the corresponding capacity values C.

Figure 10

Performance comparison for blind watermarking schemes depending upon the original-interference suppression factor a.

Figure 11

Performance comparison for memoryless original $\mathbf{x}[\vec{n}]$ and lowpass original $\mathbf{x}[\vec{n}]$ (AR(1), $a_1 = 0.95$).

Figure 12

Performance of various attacks and their defenses. The original-interference suppression factor $a = 1$.

Figure 13

Piecewise-constant original power spectrum $\Phi_{xx}(\omega)$ and frequency-weighting function $|F(\omega)|^2$ used in experiments with frequency-weighted distortion.

Figure 14

Example optimized watermark power spectra from GA/scaled algorithm and attack behavior for frequency-weighted distortion. The frequency-weighted counterpart of $\Phi_{ww}(\omega)$, $F_{ww}(\omega) = |F(\omega)|^2\Phi_{ww}(\omega)$, appears in the upper-left graph for the two distortions shown; for the case $D_{\hat{y}x} = 24$ dB, the frequency-weighted PSC-compliant spectrum $F'_{ww}(\omega)$ is also shown as a dotted curve. The frequency-weighted version of the error power spectrum $\Phi_{ee}(\omega)$, $F_{ee}(\omega) = |F(\omega)|^2\Phi_{ee}(\omega)$, appears in the lower-right graph for each distortion. The original-interference suppression factor is $a = 1$.

Figure 15

Performance of PSC-compliant, white, and optimized watermarks subject to the optimum attack. The original-interference suppression factor $a = 1$. The distortion measure is *frequency-weighted* MSE.