



The interplay between error, total variation, entropy and guessing

—some cryptographic applications

Séminaire ICE, Jan. 3rd, 2023

Olivier Rioul

Télécom Paris,
Institut Polytechnique de Paris, France

<olivier.rioul@telecom-paris.fr>





What is Randomness?

For an idea we are all familiar with, randomness is surprisingly hard to formally define.

Martin Hairer, Fields 2014

Probabilistic Randomness (Probability Theory)

FONDEMENTS DU CALCUL DES PROBABILITÉS

ERGEBNISSE DER MATHEMATIK
UND IHRER GRENZGEBIETE
HERAUSGEGEBEN VON DER SCHRIFTLEITUNG
DES
„ZENTRALBLATT FÜR MATHEMATIK“
ZWEITER BAND

GRUNDBEGRIFFE DER
WAHRSCHEINLICHKEITS-
RECHNUNG

VON
A. KOLMOGOROFF



BERLIN
VERLAG VON JULIUS SPRINGER
1933

PAR

A. KOLMOGOROV

Édition française

Traduction et notes: Olivier RIOUL



PARIS
SPARTACUS-IDH • CASSINI
2023

Préface: Martin HAIRER

Probability Distribution vs. Random Variable

$$p = (p_1, p_2, \dots, p_M) \quad p_i \geq 0, \quad \sum_i p_i = 1$$

vs.

M -ary random variable $X : \Omega \rightarrow \mathcal{X} \quad |\mathcal{X}| = M$

Link:

$$p_i = \mathbb{P}(X = x_i) \text{ where } \mathcal{X} = \{x_1, x_2, \dots, x_M\}$$

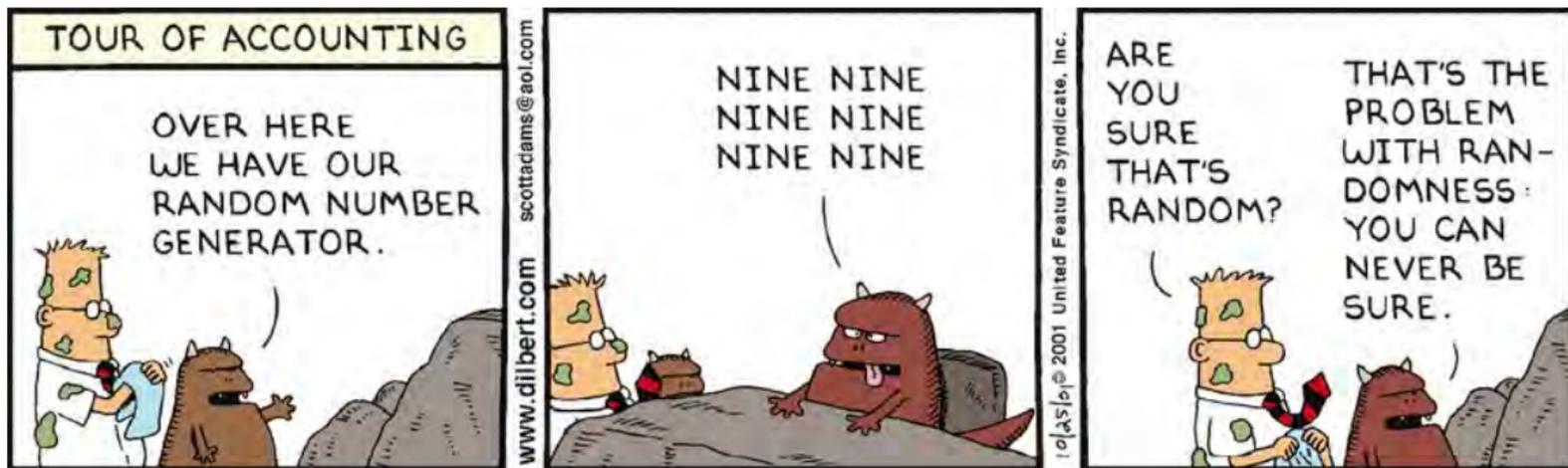
Uniform vs. Deterministic

Law of insufficient reason (“ideal randomness”)

$$p = \left(\frac{1}{M}, \frac{1}{M}, \dots, \frac{1}{M} \right)$$

vs. deterministic:

$$p = (1, 0, 0, \dots, 0)$$



Applications

Things that should be “random”:

- identifiers,
- cryptographic keys,
- signatures
- ... or any type of intended secret

for application to

- pseudo-random bit generators
- cipher security
- randomness extractors
- hash functions
- physically unclonable functions
- true random number generators

What does it mean that X or p is **sufficiently random** ?

Synonyms?

“Sufficiently”

- random
- entropic
- uncertain
- unpredictable
- hard to guess
- surprising



does it mean that we should maximize entropy?

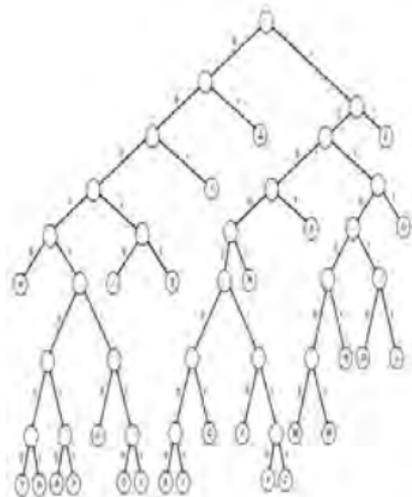
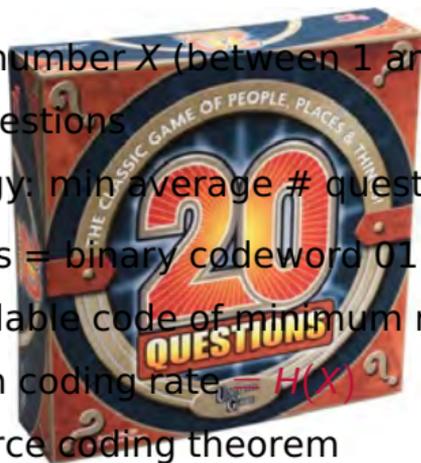
Entropy (Shannon, 1945, 1948)

$$H(pX) = p_1 \log \frac{1}{p_1} + p_2 \log \frac{1}{p_2} + \dots + p_M \log \frac{1}{p_M}$$

- minimum = 0 when X is **deterministic** (not random at all)
- maximum = $\log M$ when p is **uniform** (most random, cf. Laplace's principle of "insufficient" reason)
- logarithmic "measure of information"
- base of log = information unit (nat, dit, bit, ...) or... (**International standard ISO/IEC 80000-13**) the **Shannon** (symbol Sh)

“Game of 20 questions”: Operational Definition of Entropy

- you think of a number X (between 1 and M)
- I ask yes/no questions
- optimal strategy: min average # questions ?
- list of questions = binary codeword 0110111000...
- uniquely decodable code of minimum rate: Huffman code
- lower bound on coding rate = $H(X)$
- Shannon's source coding theorem



Guessing Entropy (Massey, 1994)

- you think of a number X (between 1 and M)
- I ask yes/no questions of the form “is $X = x$?”
- optimal strategy: min average # questions ?
- most probable first — probability $p_{(1)}$ for 1 guess
- second most probable second — probability $p_{(2)}$ for 2 guesses
- etc.
- **Guessing entropy**: expectation of the ordered distribution

$$G(X) = p_{(1)} + 2p_{(2)} + \dots + Mp_{(M)}$$

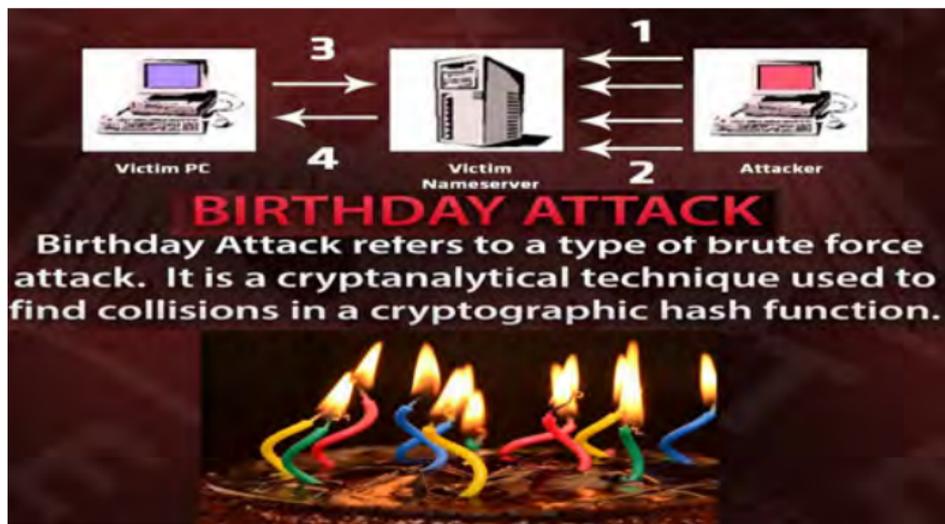
- Much higher than entropy: $G(X) \geq \frac{\exp(2H(X))}{e} + \frac{1}{2}$ (Rioul, 2022)



Coincidence Index / Collision

- X = identifier, signature, fingerprint from some randomized algorithm (e.g., hash).
- should be “unique”
- high probability of non coincidence: for any X' i.i.d. copy of X

$$R_2(X) = \mathbb{P}(X \neq X') = 1 - p_1^2 - p_2^2 - \dots - p_M^2$$



Estimation and Error

- estimate X : minimize the probability of error $\mathbb{P}_e = \min \mathbb{P}(X \neq \hat{X})$
- descending order (most probable to least probable):

$$p_{(1)} \geq p_{(2)} \geq \dots \geq p_{(M)}$$

- ensure high probability of error (worst case security):

$$\mathbb{P}_e(X) = 1 - p_{(1)} = p_{(2)} + p_{(3)} + \dots + p_{(M)}$$

Generalization: α -Entropy (Rényi, 1961)

$$H_\alpha(X) = \frac{1}{1-\alpha} \log(p_1^\alpha + p_2^\alpha + \dots + p_M^\alpha) = \frac{\alpha}{1-\alpha} \log \|p\|_\alpha$$

- $\alpha = 1/2$: asymptotically guessing entropy (Arikan, 1996)

$$\log G(X) \approx H_{1/2}(X)$$

- limiting case $\alpha \rightarrow 1$: $H_1(X) = H(X)$ (Shannon entropy)
- $\alpha = 2$: collision entropy

$$H_2(X) = \log \frac{1}{\mathbb{P}(X = X')} = \log \frac{1}{1 - R_2(X)}$$

- limiting case $\alpha \rightarrow \infty$: min-entropy

$$H_\infty(X) = \log \frac{1}{p_{(1)}} = \log \frac{1}{1 - \mathbb{P}_e(X)}$$

- $H \geq H_2 \geq H_\infty$

Remoteness from Uniform

- compare p with $u = (\frac{1}{M}, \frac{1}{M}, \dots, \frac{1}{M})$ uniform distribution (most random)
- entropy:

$$H(p) = \log M - D(p||u)$$

where $0 \leq D(p||u) \leq \log M$ is the KL-divergence.

- α -entropy

$$H_\alpha(p) = \log M - D_\alpha(p||u)$$

where $0 \leq D_\alpha(p||u) \leq \log M$ is the Rényi divergence.

- non-collision index:

$$R_2(p) = 1 - \frac{1}{M} - \|p - u\|_2^2$$

- **statistical randomness** R :

$$R(p) = 1 - \frac{1}{M} - \Delta(p, u) = 1 - \frac{1}{M} - \frac{1}{2}\|p - u\|_1$$

where $0 \leq \Delta(p, u) \leq 1$ is the **statistical distance**.

Statistical Randomness

- Statistical (Total Variation) Distance:

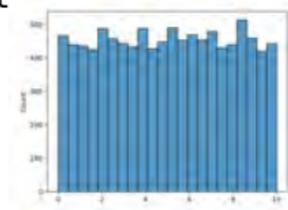
$$\begin{aligned}\Delta(p, q) &= \max_{\text{event } T} |p(T) - q(T)| \in [0, 1] \\ &= \frac{1}{2} (|p_1 - q_1| + |p_2 - q_2| + \dots + |p_M - q_M|) = \frac{1}{2} \|p - q\|_1\end{aligned}$$

- If Δ is small, p and q become **undistinguishable** under **any** statistical test $X \in T$
 - If $|\mathbb{P}(X \in T) - \mathbb{Q}(X \in T)|$ is small, type-I or type-II errors have total probability $\mathbb{P}(X \notin T) + \mathbb{Q}(X \in T) \approx 1$, hence p and q are statistically equivalent

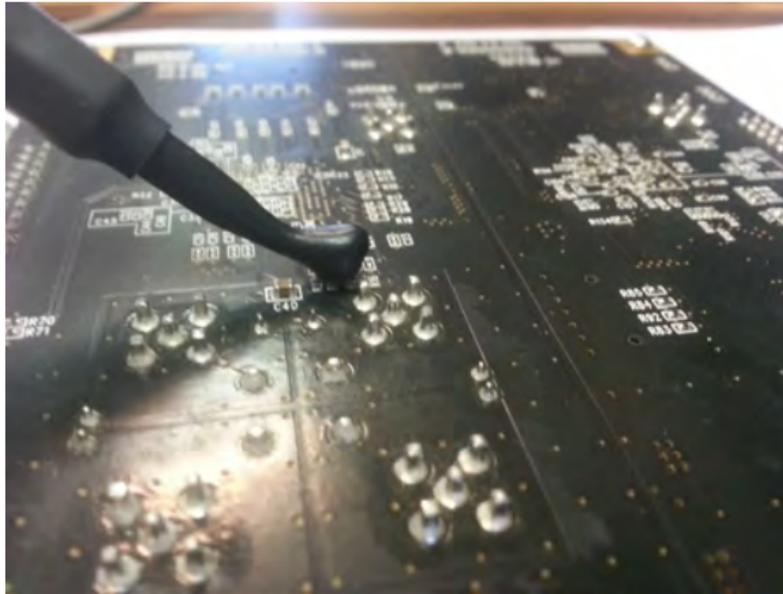
- **Statistical Randomness:**

$$R(X) = 1 - \frac{1}{M} - \Delta(p, u)$$

p undistinguishable from uniform if $R \approx 1 - \frac{1}{M}$.



Observation Y



Y: physical leakage, disclosed (sensitive) data, etc.

compress/guess/estimate/predict X with side information Y:

What is (what should be) a Randomness Measure?

- All candidates share many properties:

$$H, H_\infty, R_2 (H_2), H_\alpha, G, G_\rho, \mathbb{P}_e, \mathbb{P}_e^m, R \dots$$

- Axiomatic approach** encompassing all randomness measures $\mathfrak{R}(X)$

Equivalence Axiom : Equivalent variables $X = f(Y)$ & $Y = g(X)$ are equally random

$$X \equiv Y \implies \mathfrak{R}(X) = \mathfrak{R}(Y)$$

Knowledge Axiom : Knowledge reduces randomness (on average)

$$\mathfrak{R}(X|Y) = \mathbb{E}_y \mathfrak{R}(X|y) \leq \mathfrak{R}(X)$$

Equivalence axiom \equiv Symmetry

Equivalence Axiom : Equivalent variables ($X = f(Y)$ and $Y = g(X)$) are equally random

$$X \equiv Y \implies \mathfrak{R}(X) = \mathfrak{R}(Y)$$

means

$\mathfrak{R}(X) = \mathfrak{R}(p_1, p_2, \dots, p_M)$ **symmetric** in p_1, p_2, \dots, p_M (invariant by permutation)

Examples: $H = \sum_k p_k \log \frac{1}{p_k}$, $\mathbb{P}_e = 1 - p_{(1)}$, etc.

Knowledge axiom \equiv Concavity

Knowledge Axiom : Knowledge reduces randomness (on average)

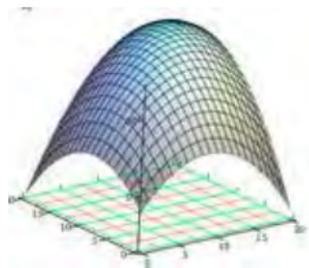
$$\mathfrak{R}(X|Y) = \mathbb{E}_y \mathfrak{R}(X|y) \leq \mathfrak{R}(X)$$

$$\iff \mathbb{E}_y \mathfrak{R}(p_{X|y}) \leq \mathfrak{R}(p_X) = \mathfrak{R}(\mathbb{E}_y p_{X|y})$$

(Jensen's inequality)

means

$$\mathfrak{R}(X) = \mathfrak{R}(p_1, p_2, \dots, p_M) \text{ **concave** in } p_1, p_2, \dots, p_M$$



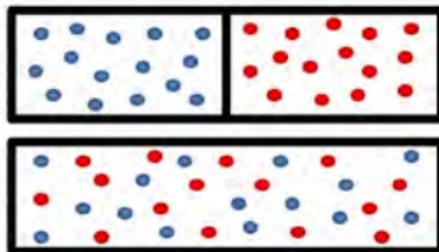
Examples: $H(X|Y) \leq H(X)$ with difference $I(X; Y)$, $G(X|Y) \leq G(X)$, $\mathbb{P}_e(X|Y) \leq \mathbb{P}_e(X)$, etc.

“Mixing” Increases Randomness

Mixing two distributions p, q of equal randomness $\mathfrak{R}(p) = \mathfrak{R}(q) = \mathfrak{R}$ results in a more random distribution $\lambda p + \bar{\lambda}q$ ($\lambda + \bar{\lambda} = 1$)

$$\mathfrak{R}(\lambda p + \bar{\lambda}q) \geq \lambda \mathfrak{R}(p) + \bar{\lambda} \mathfrak{R}(q) = \mathfrak{R}$$

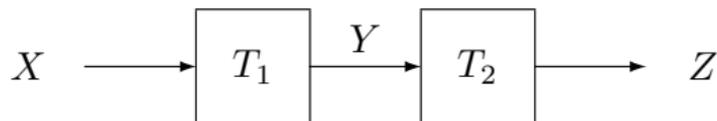
thermodynamical interpretation:



mixing two gases of equal entropy results in a gas with higher entropy

(Stochastic) Data Processing Inequality

If $X - Y - Z$ forms a Markov chain:



then

$$\mathfrak{R}(X|Y) \leq \mathfrak{R}(X|Z)$$

Processing knowledge cannot decrease randomness.

Example: $H(X|Y) \leq H(X|Z) \iff I(X; Z) \leq I(X; Y)$ (processing cannot increase information)

Robin Hood vs. Sheriff of Nottingham



“When Robin and his merry hoods performed an operation in the woods they took from the rich and gave to the poor. The Robin Hood principle asserts that this decreases inequality (subject only to the obvious constraint that you don’t take too much from the rich and turn them into poor.)”

Robin Hood Operation on $p = (p_1, p_2, \dots, p_M)$

- change only two components p_i and p_j (keeping their sum $p_i + p_j$ constant)
 - either $|p_i - p_j|$ decreases: elementary *Robin Hood* operation (more equal)
 - or $|p_i - p_j|$ increases: elementary *Sheriff of Nottingham* operation
- Define $p \preceq q$ if
 - q is obtained from p by finitely many Robin Hood operations: **equalize**
 - i.e., p is obtained from q by finitely many Sheriff of Nottingham operations

This is *majorization theory* (revisited): $p \preceq q \iff q$ is majorized by p

Properties

Lemma (Majorization (least random))

$$p \preceq (\underbrace{P, P, \dots, P}_{\lfloor 1/P \rfloor \text{ terms}}, \underbrace{r}_{\text{remainder}}, 0, \dots, 0) \quad \text{for any } p \text{ with } \max p \leq P$$

$\{1/P\}$

In particular

$$p \preceq (1, 0, 0, \dots, 0) \text{ for any } p$$

Proof.

If p has $0 < p_i, p_j < P$ then by a suitable SN-operation, one of these two can be made $= 0$ or P , reducing # probabilities $\in (0, P)$ by 1. Then continue with a finite # steps. \square

Lemma (Minorization (most random))

$$u = \left(\frac{1}{M}, \frac{1}{M}, \dots, \frac{1}{M}\right) \preceq p \quad \text{for any } p$$

Proof.

If there is one $p_i > \frac{1}{M}$ and one $p_j < \frac{1}{M}$, by a suitable RH-operation, one of these two can be made $= \frac{1}{M}$, reducing # probabilities $\neq \frac{1}{M}$ by 1. Then continue with a finite #

Schur Concavity

Theorem (More Equal \implies More Random)

$$X \preceq Y \implies \mathfrak{R}(X) \leq \mathfrak{R}(Y)$$

Proof.

Any Robin Hood operation can be written as $(p_i, p_j) \mapsto (\lambda p_i + \bar{\lambda} p_j, \lambda p_j + \bar{\lambda} p_i)$. But then

$$\mathfrak{R}(\lambda p_i + \bar{\lambda} p_j, \lambda p_j + \bar{\lambda} p_i) \geq \lambda \varphi(p_i, p_j) + \bar{\lambda} \varphi(p_j, p_i) = \mathfrak{R}(p_i, p_j)$$

by concavity and symmetry. □

Min/Max Random

Since

$$(1, 0, 0, \dots, 0) \preceq p \preceq \left(\frac{1}{M}, \frac{1}{M}, \dots, \frac{1}{M}\right),$$

randomness $\mathfrak{R}(X)$ is

- minimum for deterministic X (least random)
- maximum for uniformly distributed X (most random)

Examples: $H, H_\alpha, G, G_\rho, \mathbb{P}_e, R, \dots$

(Deterministic) Data Processing Inequality

Theorem

$$\mathfrak{R}(f(X)) \leq \mathfrak{R}(X)$$

Proof.

Applying f consists of successive gatherings of two distinct values x_i, x_j of x in the same preimage of $y = f(x)$. Each such gathering corresponds to a SN-operation $(p_i, p_j) \mapsto (p_i + p_j, 0)$ and the result follows by Schur concavity. □

Example: $H(f(X)) = I(f(X); f(X)) \leq I(X; f(X)) \leq H(X)$ by data processing inequality since $X - f(X) - f(X)$ is a Markov chain.

Addition Increases Randomness

Theorem

For any additional random variable Y ,

$$\mathfrak{R}(X) \leq \mathfrak{R}(X, Y)$$

This is equivalent to the deterministic data processing inequality.

Proof.

Take $f(x, y) = x$ in the deterministic DPI. Conversely, apply it to $(f(X), X)$. □

Example: $H(X, Y) \geq H(X)$ with difference $H(Y|X)$.

Robert Fano and Mark Pinsker in the 1960s



USA

USSR

Interplay Between Entropies and Estimation Error: Fano Inequalities

- original inequality $H \leq h(\mathbb{P}_e) + \mathbb{P}_e \log(M - 1)$ (Fano, 1961) for X and $X|Y$
- Fano-type inequality: upper bound \mathfrak{R} in terms of \mathbb{P}_e
- reverse Fano inequality: lower bound \mathfrak{R} in terms of \mathbb{P}_e

Theorem (Optimal Reverse Fano and Fano Inequalities)

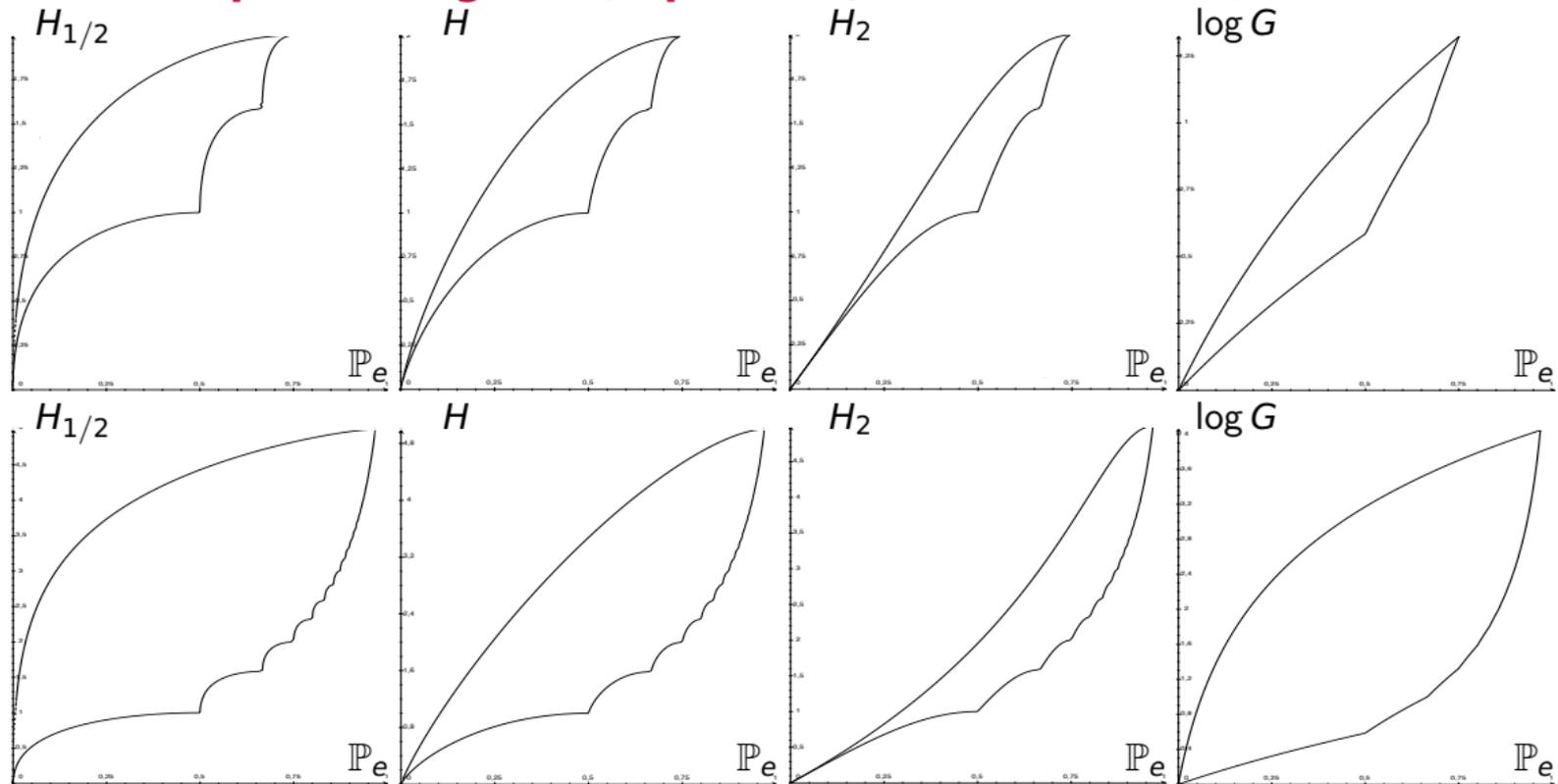
$$\mathfrak{R}(1 - \mathbb{P}_e, \dots, 1 - \mathbb{P}_e, r, 0, \dots, 0) \leq \mathfrak{R}(X) \leq \mathfrak{R}(1 - \mathbb{P}_e, \frac{\mathbb{P}_e}{M-1}, \dots, \frac{\mathbb{P}_e}{M-1})$$

Proof.

Let $\mathbb{P}_s = \max p = 1 - \mathbb{P}_e$, apply Schur concavity to

$$\underbrace{(\mathbb{P}_s, \dots, \mathbb{P}_s, r, 0, \dots, 0)}_{\text{Majorization Lemma}} \preceq p \preceq \underbrace{(\mathbb{P}_s, \frac{\mathbb{P}_e}{M-1}, \dots, \frac{\mathbb{P}_e}{M-1})}_{\text{Minorization Lemma}}$$

Optimal Regions (top: $M = 4$; bottom: $M = 32$)



Interplay Between Entropies and Total Variation: Pinsker Inequalities

- original inequality $D(p\|q) \geq 2(\log e)\Delta(p\|q)^2$ (Pinsker & many others, circa 1960), that is, for $q = u$, $H \leq \log M - 2(\log e)(1 - 1/M - R)^2$
- “Pinsker-type” inequality: upper bound \mathfrak{R} in terms of R
- “reverse Pinsker” inequality: lower bound \mathfrak{R} in terms of R
- Most known Pinsker/reverse Pinsker inequalities are suboptimal in this sense

Theorem (Optimal “reverse Pinsker” and “Pinsker”)

For *any* S -concave $\mathfrak{R}(p)$ and any K between # probabilities $> \frac{1}{M}$ and # probabilities $\geq \frac{1}{M}$

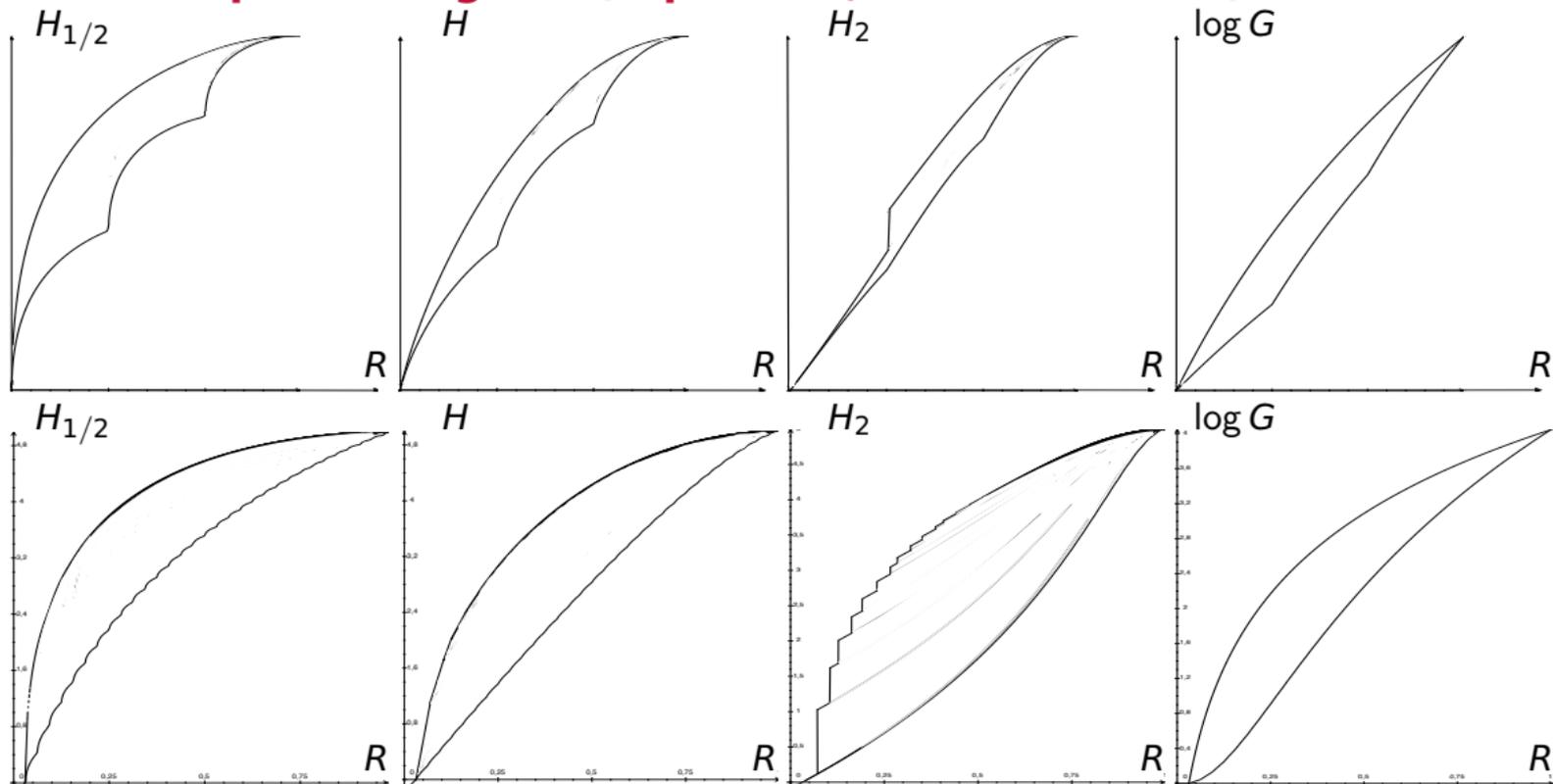
$$\mathfrak{R}(1 - R + \frac{1}{M}, \frac{1}{M}, \dots, \frac{1}{M}, r, 0, \dots, 0) \leq \mathfrak{R}(p) \leq \mathfrak{R}(\frac{1}{M} + \frac{1-R}{K}, \dots, \frac{1}{M} + \frac{1-R}{K}, \frac{1}{M} - \frac{1-R}{M-K}, \dots, \frac{1}{M} - \frac{1-R}{M-K})$$

Proof.

Let $\Delta = 1 - R$, apply S -concavity to

$$\underbrace{\left(\Delta + \frac{1}{M}, \frac{1}{M}, \dots, \frac{1}{M}, r, 0, \dots, 0 \right)}_{\text{Majorization Lemma}} \succcurlyeq p \succcurlyeq \underbrace{\left(\frac{1}{M} + \frac{\Delta}{K}, \dots, \frac{1}{M} + \frac{\Delta}{K}, \frac{1}{M} - \frac{\Delta}{M-K}, \dots, \frac{1}{M} - \frac{\Delta}{M-K} \right)}_{\text{Minorization Lemma}}$$

Optimal Regions (top: $M = 4$; bottom: $M = 32$)



Application 1: Universal Hash Functions are Good Randomness Extractors

- secret key X : an adversary knows part of the key, but we don't know which bits are left over.
 $\implies X$ has **low quality** randomness, only H_2 bits of (collision) entropy is remaining
- pick a uniformly distributed **hash function** h (seed on ℓ bits), independent of X



Universal hash function: $\mathbb{P}(h(X) = h(X'), X \neq X') \approx 2^{-m}$ (lowest collision)

- **Aim:** $h(X)$ extracts $m < H_2$ bits of **high quality** randomness (uniform) from X , even **knowing** the seed h (which can be publicly available, or recycled).
 $\implies h(X)$ can still be used as a secret key (on $< H_2$ bits).

That is, we want $(h, h(X))$ to be (jointly) **maximally random** $\approx \ell + m$ bits ($M = 2^{\ell+m}$)

\implies high statistical randomness, low statistical distance Δ to the uniform.

Solution: Optimal Pinsker inequality applied to $(h, h(X))$

Closed-form optimal Pinsker inequality (for M even and $\Delta \leq 1/2$)

$$\Delta \leq \frac{1}{2} \sqrt{M2^{-H_2} - 1}$$

Now

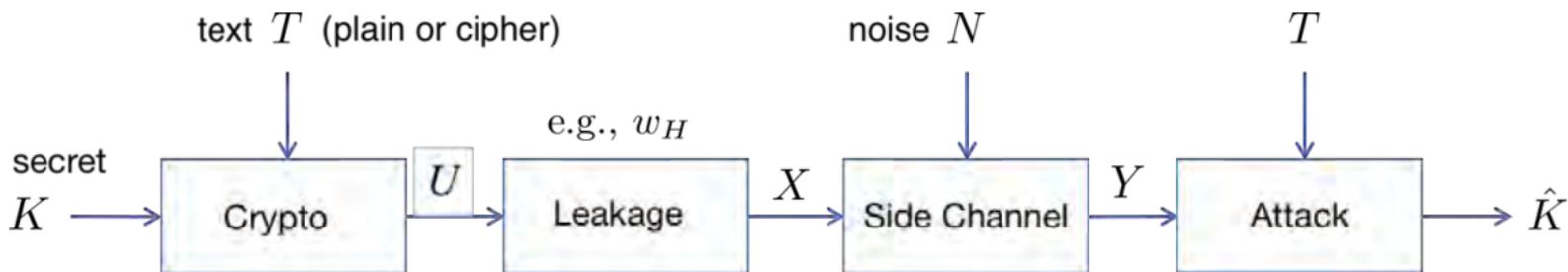
$$\begin{aligned} 2^{-H_2(h, h(X))} &= \mathbb{P}((h, h(X)) = (h', h'(X'))) \\ &= 2^{-\ell} \mathbb{P}(h(X) = h(X')) \\ &= 2^{-\ell} (\mathbb{P}(h(X) = h(X'), X \neq X') + \mathbb{P}(X = X')) \\ &= 2^{-\ell} (2^{-m} + 2^{-H_2(X)}) \end{aligned}$$

by universality, so

$$\Delta \leq \frac{1}{2} \sqrt{2^{\ell+m} (2^{-\ell} (2^{-m} + 2^{-H_2(X)})) - 1} = 2^{(m-H_2)/2-1}$$

very small since if m sufficiently $\ll H_2$.

Application 2: Information Leakage over Side Channel



Framework of **[Cherisey-Guilley-Rioul-Piantanida'19]**:

- AES-256 implementation with many (q) measurement traces
- Hamming weight leakage model $Y_i = w_H(S(T_i \oplus K)) + N_i \quad (i = 1, 2, \dots, q)$
- **upper bound** success rate \mathbb{P}_S as a function of q
- **lower bound** # traces q_{\min} needed to achieve a given success \mathbb{P}_S
- compare to optimal (maximum likelihood) attacks giving $\mathbb{P}_S(K|Y)$

Solution: Optimal Fano Inequality for α -Information

$X - Y - \hat{X}$ with M -ary X , probability of success $\mathbb{P}_s = \mathbb{P}(\hat{X} = X)$

- X is a sensitive data (depending on a secret);
- $P_{Y|X}$ is a “side-channel” through which information leaks
- Y is disclosed to the attacker (measurements by probes/sniffers...)
- $P_{\hat{X}|Y}$ is the **attack** (MAP rule maximizes probability of success)

$$I_\alpha(X; Y) \underset{DPI}{\geq} I_\alpha(X, \hat{X}) = D_\alpha(p_{X, \hat{X}} \| p_X q_{\hat{X}}^*) \underset{DPI}{\geq} d_\alpha(\mathbb{P}_s(X|Y) \| \mathbb{P}'_s) \underset{dpi}{\geq} d_\alpha(\mathbb{P}_s(X|Y) \| \mathbb{P}_s(X))$$

where $\mathbb{P}'_s = \sum_x p_X(x) q_{\hat{X}}^*(x) \leq \max_x p_X(x) = \mathbb{P}_s(X)$.

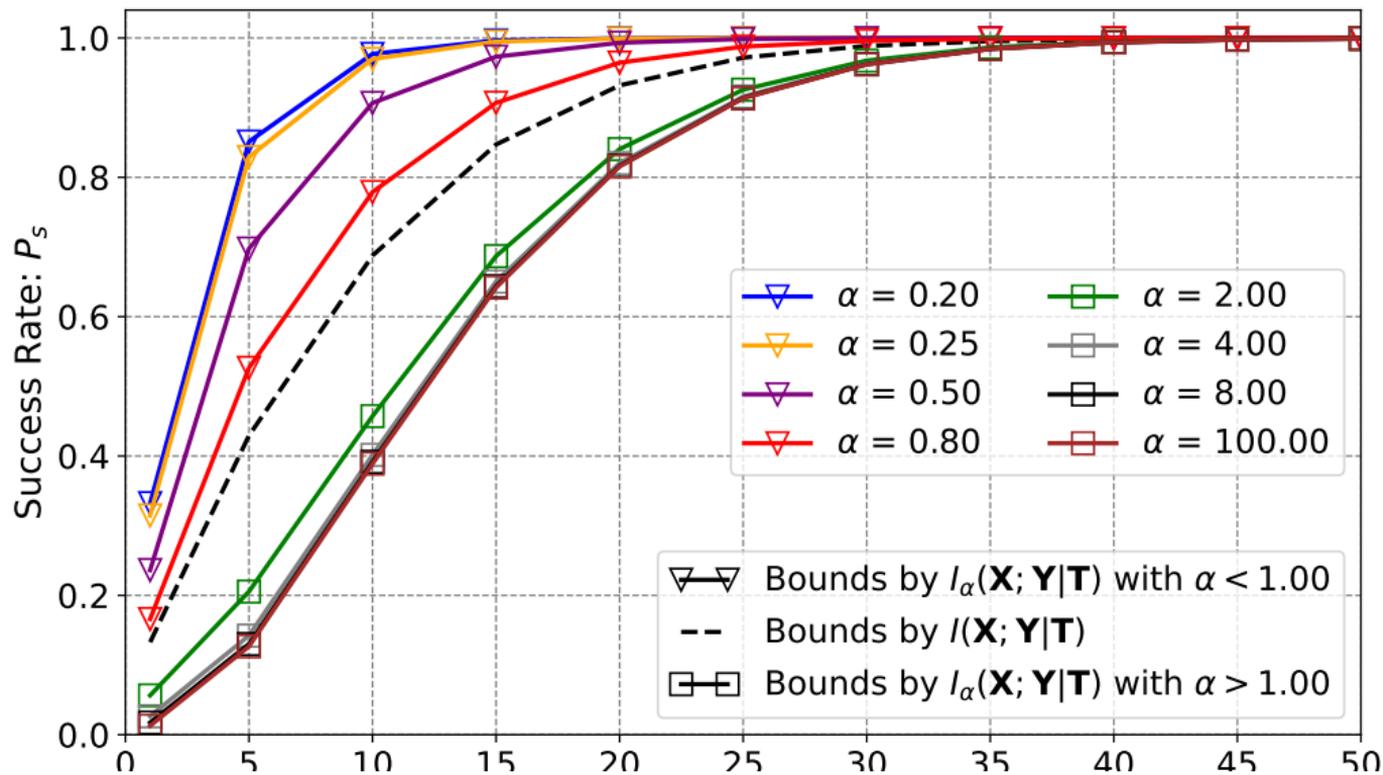
α -Fano's Inequality [Rioul'21]

$$I_\alpha(X; Y) \geq d_\alpha(\mathbb{P}_s(X|Y) \| \mathbb{P}_s(X))$$

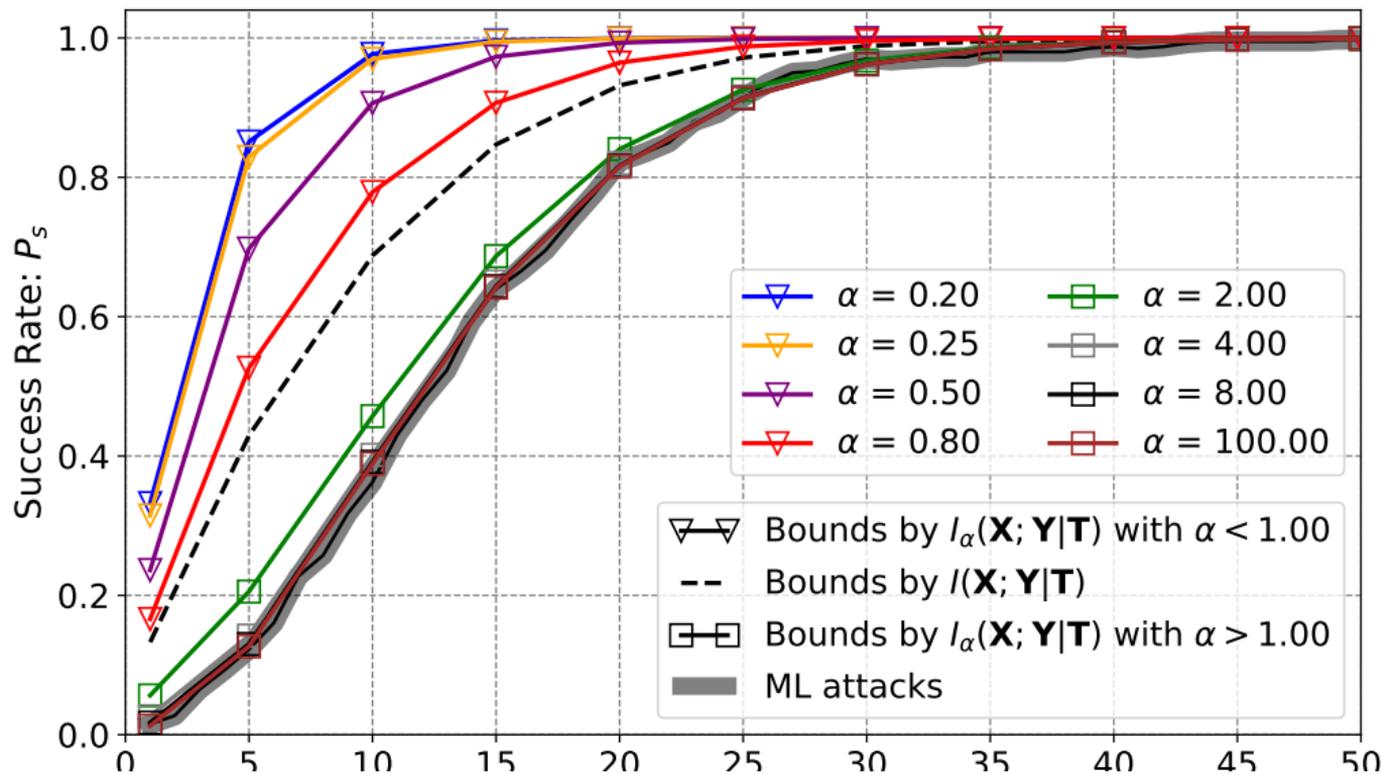
generalizes [HanVerdú'94] ($\alpha = 1$)

\implies implicit upper bound on $\mathbb{P}_s(X|Y)$ as a function of α -information.

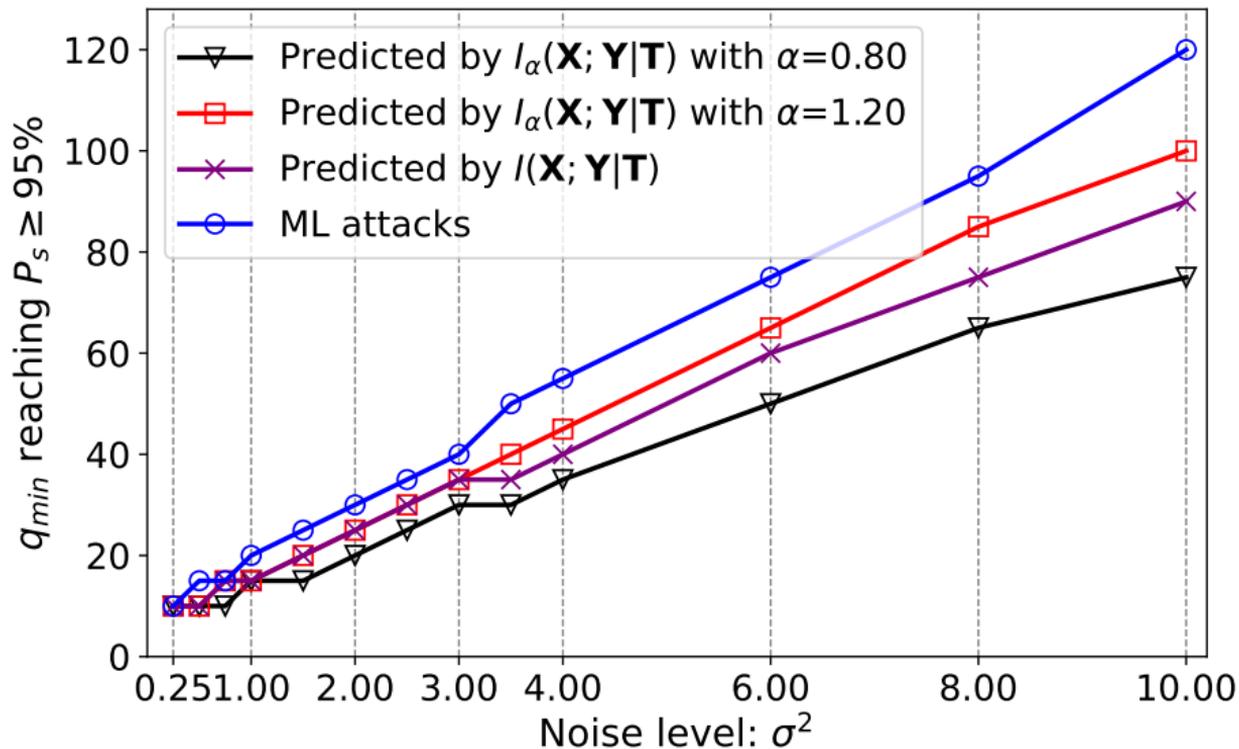
Upper Bounds on Success Rate \mathbb{P}_S



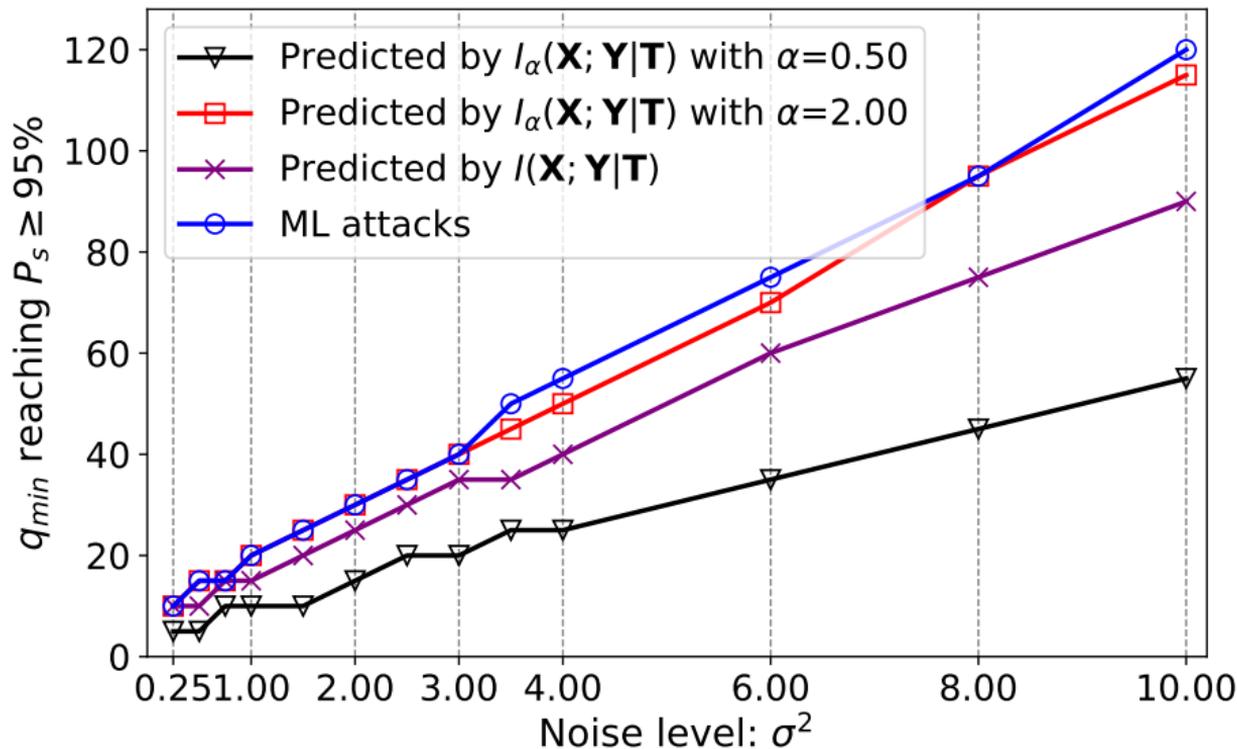
Upper Bounds on Success Rate \mathbb{P}_S



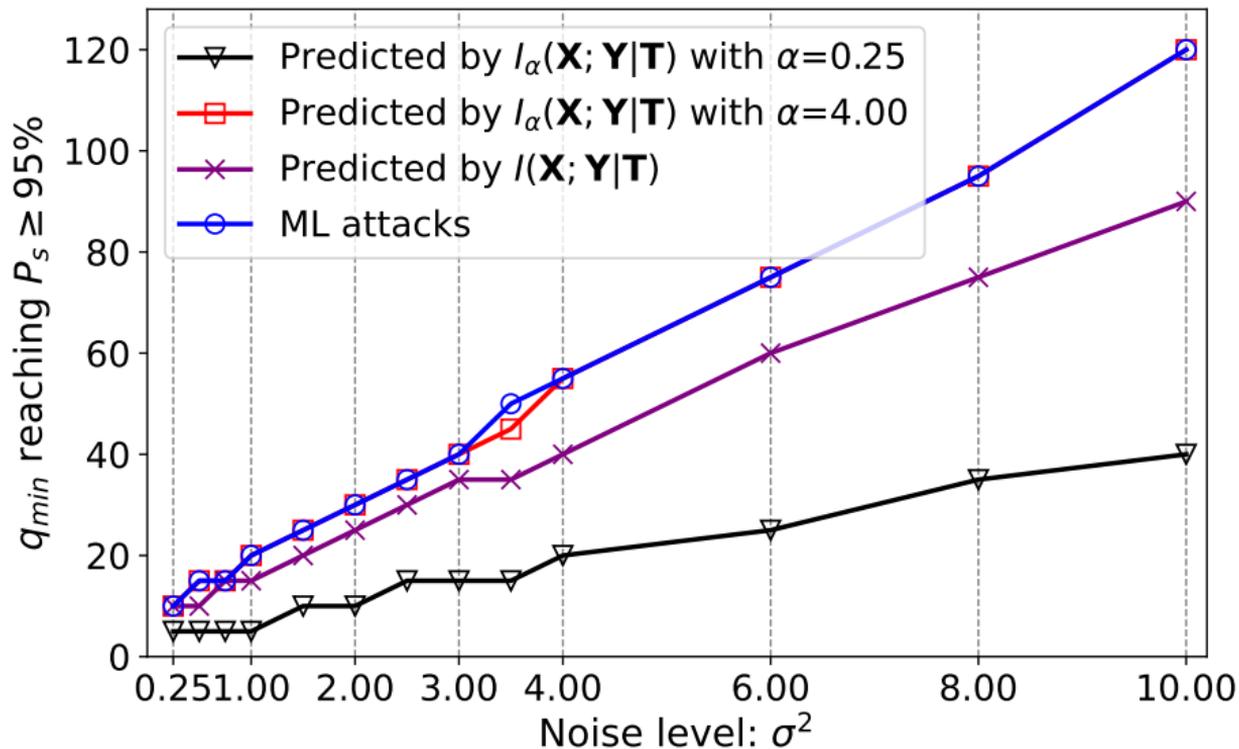
Lower Bounds on # of Traces to Achieve 95% Success



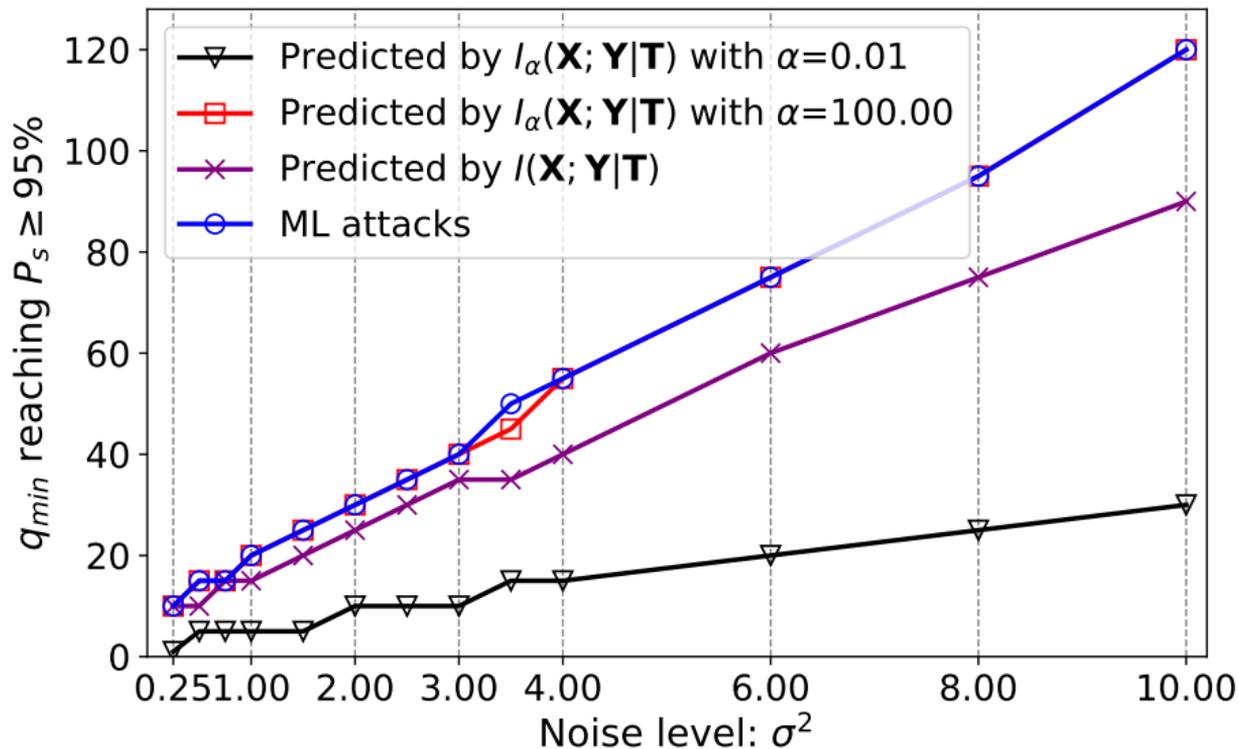
Lower Bounds on # of Traces to Achieve 95% Success



Lower Bounds on # of Traces to Achieve 95% Success



Lower Bounds on # of Traces to Achieve 95% Success





The interplay between error, total variation, entropy and guessing

—some cryptographic applications

Thank you!

Olivier Rioul

Télécom Paris,
Institut Polytechnique de Paris, France

<olivier.rioul@telecom-paris.fr>

