Entropy, Inference, and Channel Capacity

Sean Meyn

Department of Electrical and Computer Engineering
University of Illinois
and the Coordinated Science Laboratory

NSF support: ECS 02-17836, ITR 00-85929 and CCF 00-49089

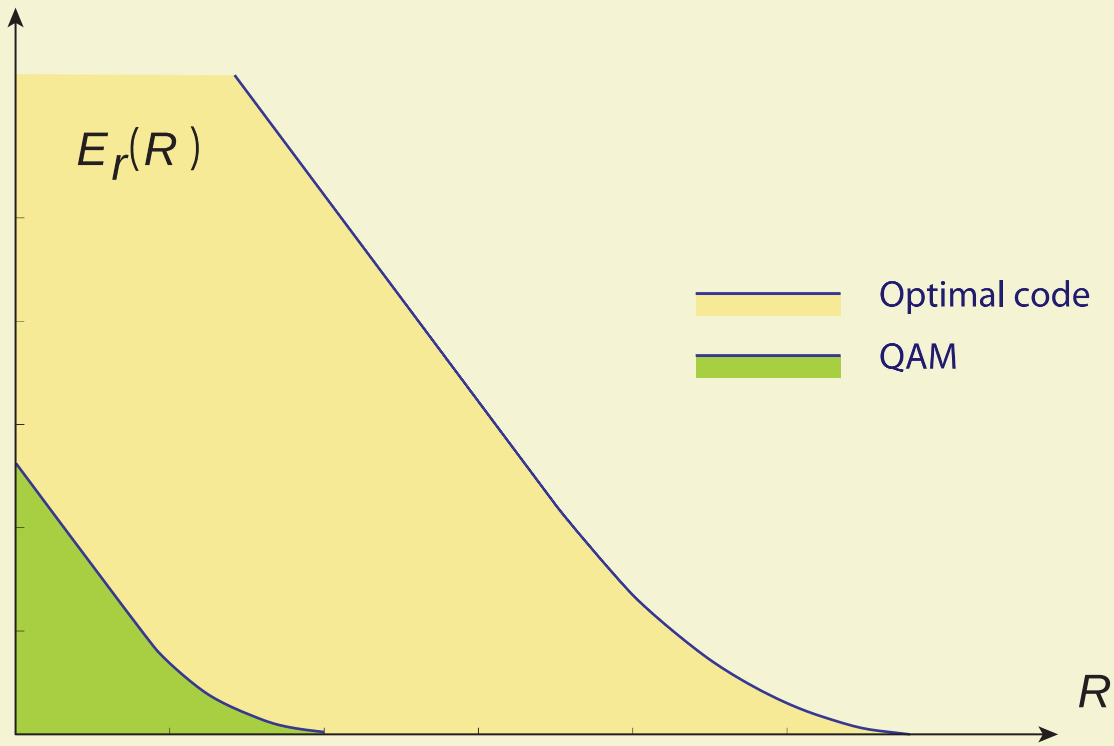
Overview

Hypothesis testing and channel coding

Structure of optimal codes

Error exponents

Algorithms



Outline (today)

Introduction

Relative entropy & Large deviations

Hypothesis testing

Channel capacity

Conclusions

Memoryless Channel Model

Memoryless channel with input sequence X, output sequence Y

Channel kernel $P(dy \mid x) = \mathsf{P}\{Y_t \quad dx \mid X_t = x\}$

If X is i.i.d. with marginal distribution μ

Then, Y is i.i.d. with marginal distribution π

$$\pi(\cdot) = {}^{\square} \quad P(\cdot \mid x)\mu(dx)$$

Random codebook

Channel kernel $\quad P(dy \mid x) = \mathsf{P}\{Y_t \quad dx \mid X_t = x\}$

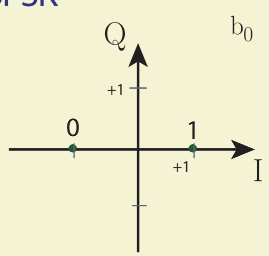
N-dimensional code words $X^i, \qquad i = 1, 2, \ldots, e^{NR}$

N-dimensional output Y received: i.i.d.,
with marginal distribution π

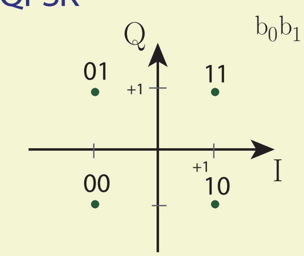
IEEE Std 802.11a -1999

SUPPLEMENT TO IEEE STANDARD FOR INFORMATION TECHNOLOGY

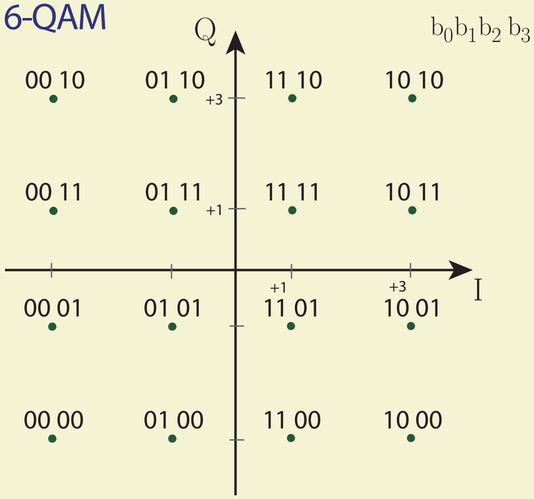
BPSK



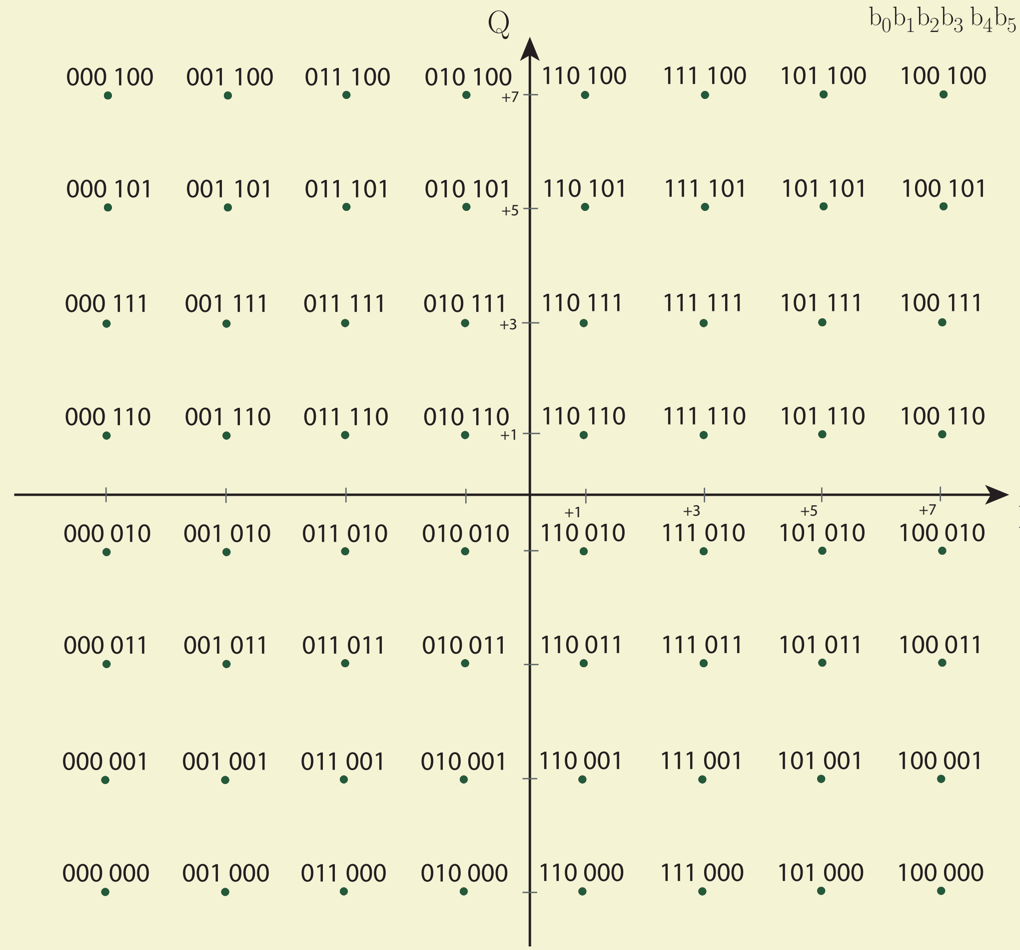
QPSK



16-QAM



64-QAM



Questions & Objectives

1. What is the structure of optimal μ ?

2. Construct algorithms based on this structure

3. Worst-case modeling to *simplify* code construction

4. Decoding algorithms and evaluation

Questions & Objectives

1. What is the structure of optimal μ ?

2. Construct algorithms based on this structure

3. Worst-case modeling to *simplify* code construction

4. Decoding algorithms and evaluation

Methodology & Viewpoint:

Hypothesis testing

Large deviations

Convex & linear optimization theory

Example: Rayleigh Channel $Y = AX + N$

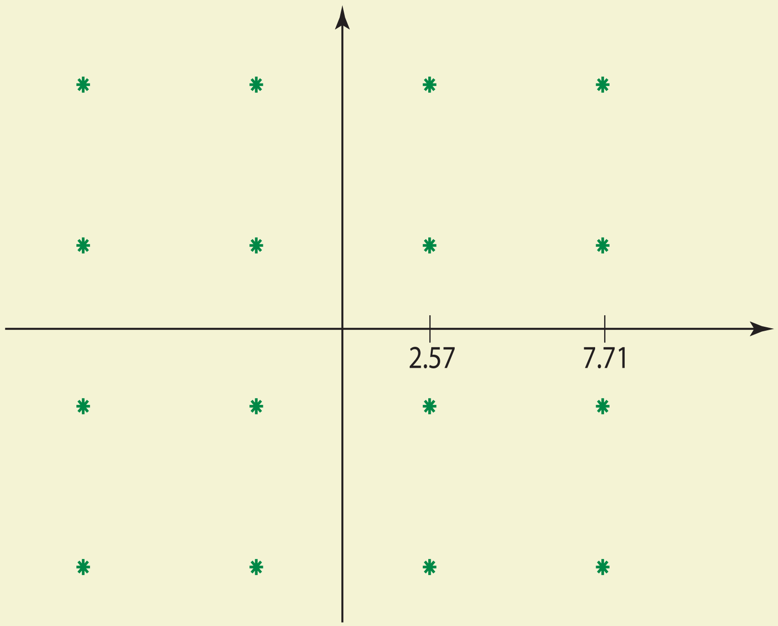
A and N are *i.i.d.* and mutually independent:

$$\sigma_A^2 = 1,\ \sigma_N^2 = 1, \text{ and } \sigma_P^2 = 26.4 \text{ (SNR=14.2 dB)}$$

Example: Rayleigh Channel $Y = AX + N$

A and N are *i.i.d.* and mutually independent:

$$\sigma_A^2 = 1,\ \sigma_N^2 = 1, \text{ and } \sigma_P^2 = 26.4 \text{ (SNR=14.2 dB)}$$



16-point QAM

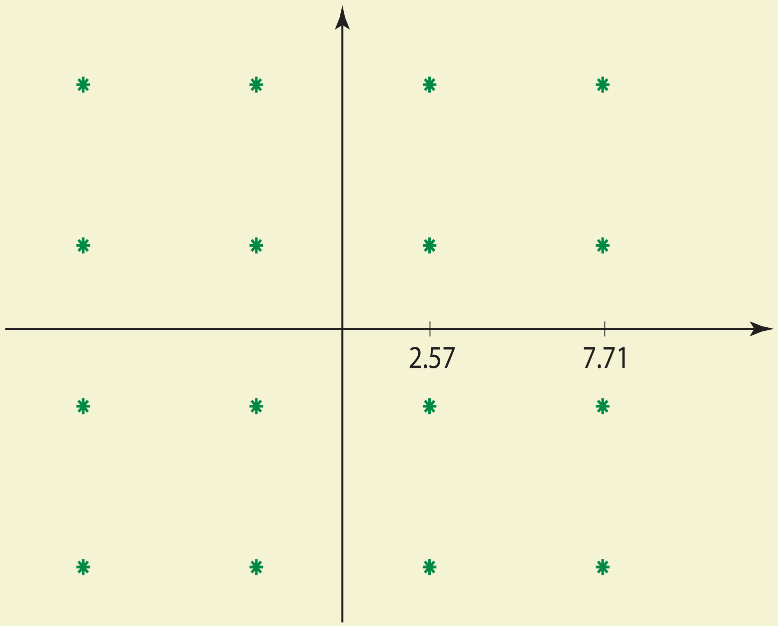
Standard: 16-point QAM

Rate: $I = 0.2$ nats/symbol.

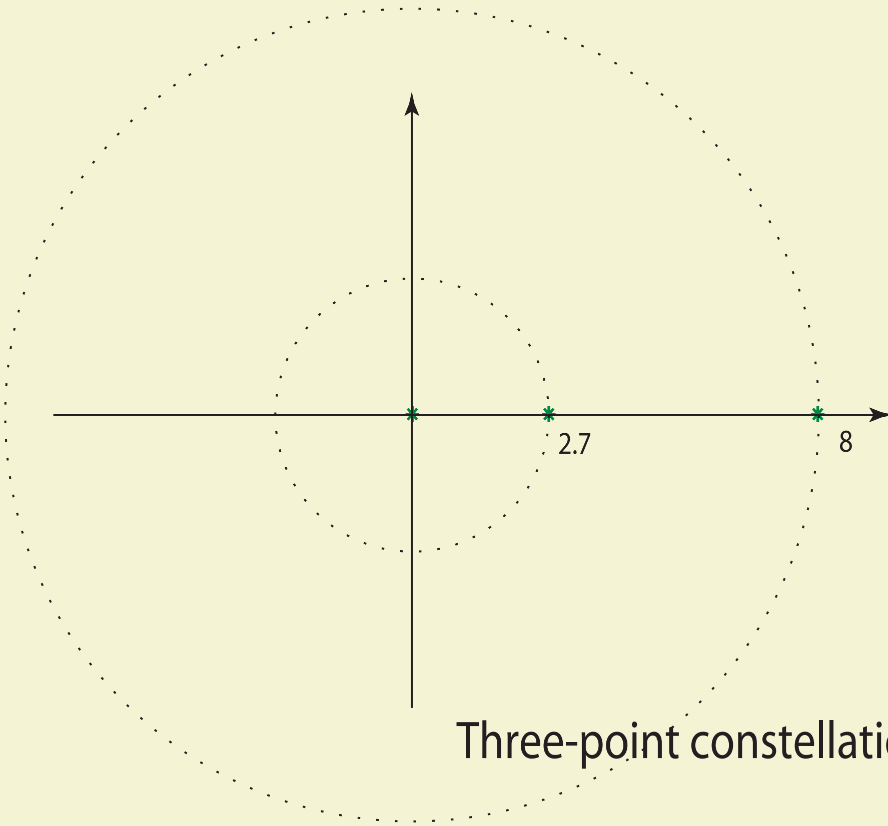
Example: Rayleigh Channel $Y = AX + N$

A and N are *i.i.d.* and mutually independent:

$$\sigma_A^2 = 1,\ \sigma_N^2 = 1, \text{ and } \sigma_P^2 = 26.4 \text{ (SNR=14.2 dB)}$$



16-point QAM

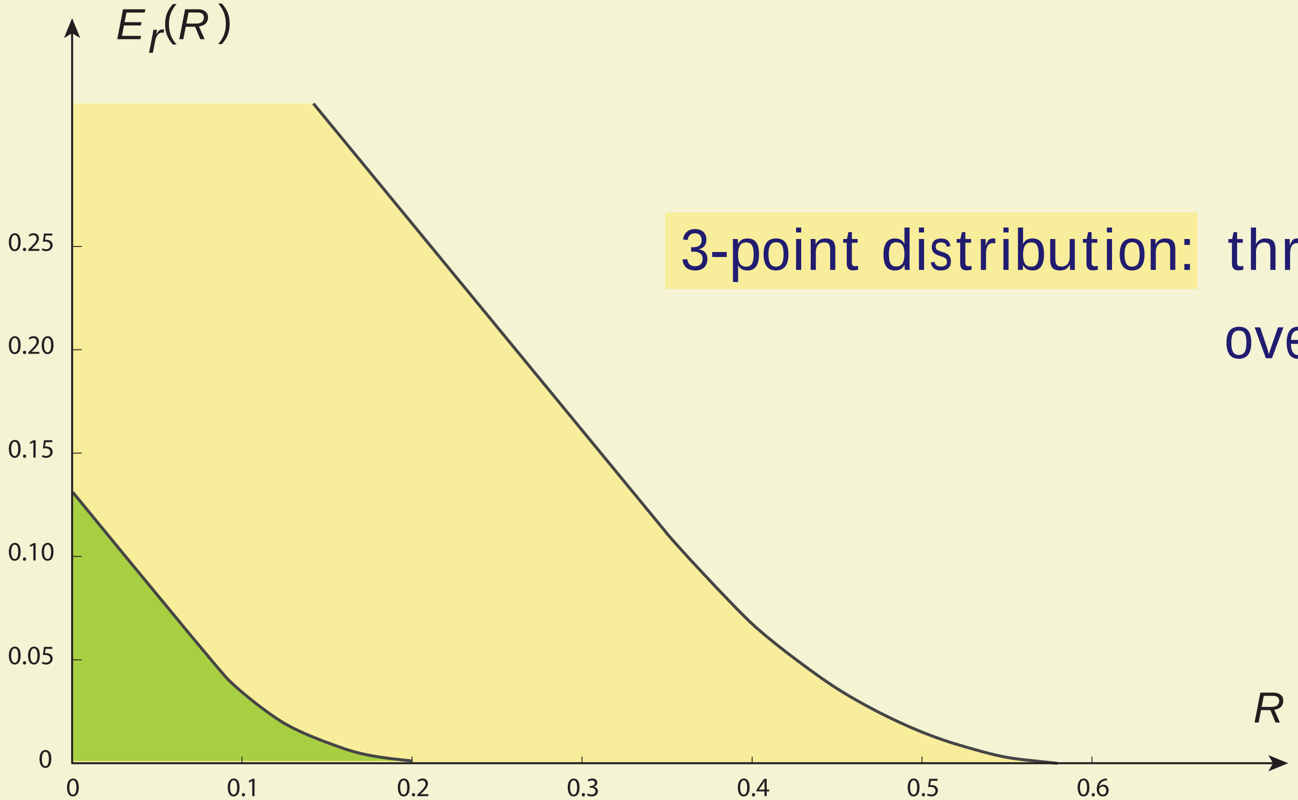


Three-point constellation

Example: Rayleigh Channel $Y = AX + N$

A and N are *i.i.d.* and mutually independent:

$$\sigma_A^2 = 1,\ \sigma_N^2 = 1, \text{ and } \sigma_P^2 = 26.4 \text{ (SNR=14.2 dB)}$$



3-point distribution: three-fold improvement over 16-point QAM

Outline

Large Deviations

$\boldsymbol{X} = \{X_1, X_2, \ldots\}$ a nice Markov chain on X, marginal distribution μ

Simulate a function $g\colon \mathsf{X} \to \mathbb{R}$

$$\hat{c}_n = n^{-1} \sum_{t=1}^{n} g(X_t)$$

Large Deviations

$\boldsymbol{X} = \{X_1, X_2, \ldots\}$ a nice Markov chain on X, marginal distribution μ

Simulate a function $g : \mathsf{X} \to \mathbb{R}$

$$\hat{c}_n = n^{-1} \sum_{t=1}^{n} g(X_t) \quad \to \quad c_0 = \mu(g)$$

Probability of over-estimate $c > c_0$

$$n^{-1} \log \mathsf{P}\Big\{ n^{-1} \sum_{t=1}^{n} g(X_t) \geq c \Big\} \to -\Lambda^*(c)$$

Large Deviations

$\boldsymbol{X} = \{X_1, X_2, \ldots\}$ a nice Markov chain on X, marginal distribution μ

Simulate a function $g : \mathsf{X} \to \mathbb{R}$

$$\hat{c}_n = n^{-1} \sum_{t=1}^{n} g(X_t) \quad \to \quad c_0 = \mu(g)$$

Probability of over-estimate $c > c_0 = \mu(g)$,

$$n^{-1} \log \mathsf{P}\Big\{ n^{-1} \sum_{t=1}^{n} g(X_t) \geq c \Big\} \to -\Lambda^*(c)$$

Rate function & log-moment generating function

$$\Lambda^*(c) = \sup_{\theta > 0} [\theta c - \Lambda(\theta)] \qquad \Lambda(\theta) = \lim_{n \to \infty} n^{-1} \log \mathsf{E}\Big[\exp\Big(\theta \sum_{t=1}^{n} g(X_t) \Big) \Big]$$

Hoeffding's Bound

$\boldsymbol{X} = \{X_1, X_2, \dots\}$ is i.i.d. on X $= [0, 1]$ $\qquad g(x) = x$

Marginal distribution μ unknown

$$\hat{c}_n = n^{-1} \sum_{t=1}^{n} X_t \quad \to \quad c_0 = \mu(g)$$

Worst-case rate function & log-moment generating function

$$\inf\{\Lambda_\mu(c) : \mu(g) = c_0\} \qquad \sup\{\Lambda_\mu(\theta) : \mu(g) = c_0\}$$

Hoeffding's Bound

$\boldsymbol{X} = \{X_1, X_2, \dots\}$ is i.i.d. on X $= [0, 1]$ $\quad g(x) = x$

Marginal distribution μ unknown

$$\hat{c}_n = n^{-1} \sum_{t=1}^{n} X_t \to c_0 = \mu(g)$$

Worst-case rate function & log-moment generating function

$$\inf\{\Lambda_\mu(c) : \mu(g) = c_0\} \qquad \sup\{\Lambda_\mu(\theta) : \mu(g) = c_0\}$$

Solution: μ is binary on $\{0, 1\}$

Bennett's Lemma

$\boldsymbol{X} = \{X_1, X_2, \ldots\}$ is i.i.d. on X $= [0, 1]$ Mean *and* variance given

Marginal distribution μ unknown $g(x) = x$

$$\hat{c}_n = n^{-1} \sum_{t=1}^{n} X_t$$

Worst-case rate function & log-moment generating function

$$\inf\{\Lambda_\mu(c) : \mu(g_i) = c_i,\ \ i = 1, 2\} \qquad \sup\{\Lambda_\mu(\theta) : \mu(g_i) = c_i,\ \ i = 1, 2\}$$

Bennett's Lemma

$\boldsymbol{X} = \{X_1, X_2, \ldots\}$ is i.i.d. on $\mathsf{X} = [0, 1]$ Mean *and* variance given

Marginal distribution μ unknown $g(x) = x$

$$\hat{c}_n = n^{-1} \sum_{t=1}^{n} X_t$$

Worst-case rate function & log-moment generating function

$$\inf\{\Lambda_\mu(c) : \mu(g_i) = c_i, \ i = 1, 2\} \qquad \sup\{\Lambda_\mu(\theta) : \mu(g_i) = c_i, \ i = 1, 2\}$$

Solution: μ is binary on $\{x_0, 1\}$

Generalized Bennett's Lemma

$\boldsymbol{X} = \{X_1, X_2, \ldots\}$ is i.i.d. on X $= [0, 1]$ n moments g_i given

Marginal distribution μ unknown

$$\hat{c}_n = n^{-1} \sum_{t=1}^{n} g(X_t)$$

Worst-case moment generating function: $\lambda(\theta) = \mathrm{E}[e^{\theta g(X_t)}] = \langle \mu, e^{\theta g} \rangle$

Generalized Bennett's Lemma

$\boldsymbol{X} = \{X_1, X_2, \ldots\}$ is i.i.d. on $\mathsf{X} = [0, 1]$ n moments g_i given

Marginal distribution μ unknown

$$\hat{c}_n = n^{-1} \sum_{t=1}^{n} g(X_t)$$

Worst-case moment generating function: $\lambda(\theta) = \mathsf{E}[e^{\theta g(X_t)}] = \mu, e^{\theta g}$

Linear program over M:

$$\begin{aligned} &\max \quad \mu, e^{\theta g} \\ &\mathbf{s.\,t.} \quad \mu, g_i = c_i, \qquad i = 1, \ldots, n. \end{aligned}$$

μ is discrete

Sanov's Theorem

State space: X   Probability measures: M

Notation: $\langle \mu, g \rangle = \mu(g) := \int g(y)\,\mu(dy)$ μ a measure
g a function on X

Empirical measures:

$$L_n := \frac{1}{n}\sum_{t=0}^{n-1}\delta_{X_t} \qquad L_n \in M \quad \text{for} \quad n \geq 1$$

$$\langle L_n, g \rangle = \frac{1}{n}\sum_{t=0}^{n-1} g(X_t)$$

Sanov's Theorem

State space: X Probability measures: M

Notation: $\langle \mu, g \rangle = \mu(g) := \int g(y)\,\mu(dy)$ μ a measure
g a function on X

Empirical measures:

$$L_n := \frac{1}{n} \sum_{t=0}^{n-1} \delta_{X_t} \qquad L_n \in M \quad \text{for} \quad n \geq 1$$

Relative entropy:

$$D(\nu \,\|\, \mu) = \left\langle \nu, \log\left(\frac{d\nu}{d\mu}\right)\right\rangle = \int \log\left(\frac{d\nu}{d\mu}\right) \nu(dx)$$

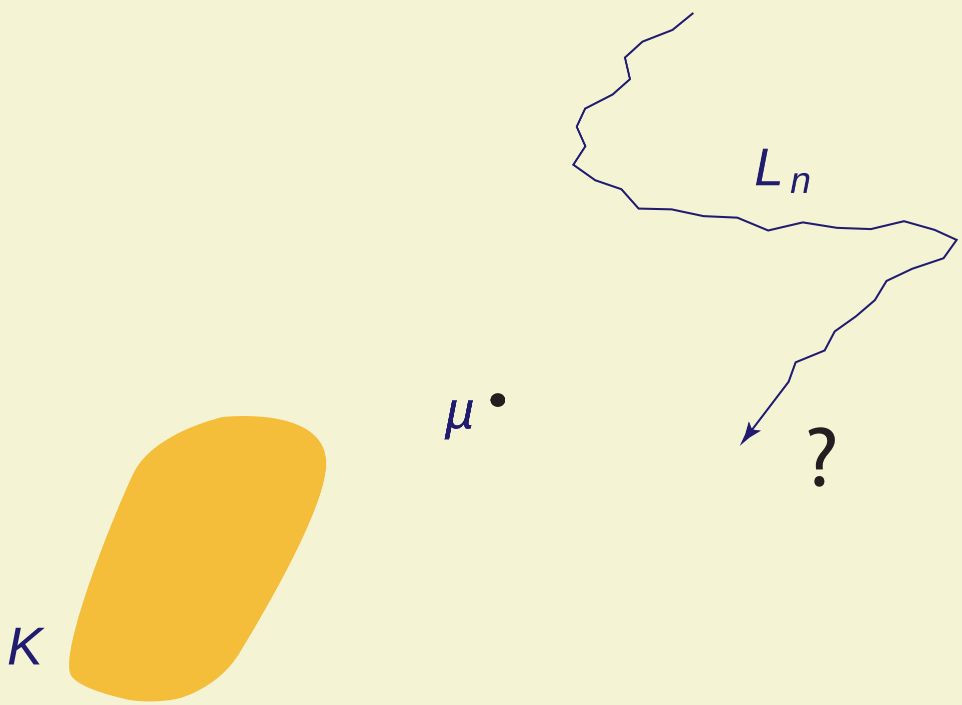
Sanov's Theorem

Law of large numbers: $L_n := \frac{1}{n}\sum_{t=0}^{n-1}\delta_{X_t}$ $\qquad L_n \to \mu, \quad n \to \infty$

Sanov's Theorem

Convex set of probability measures K M μ K

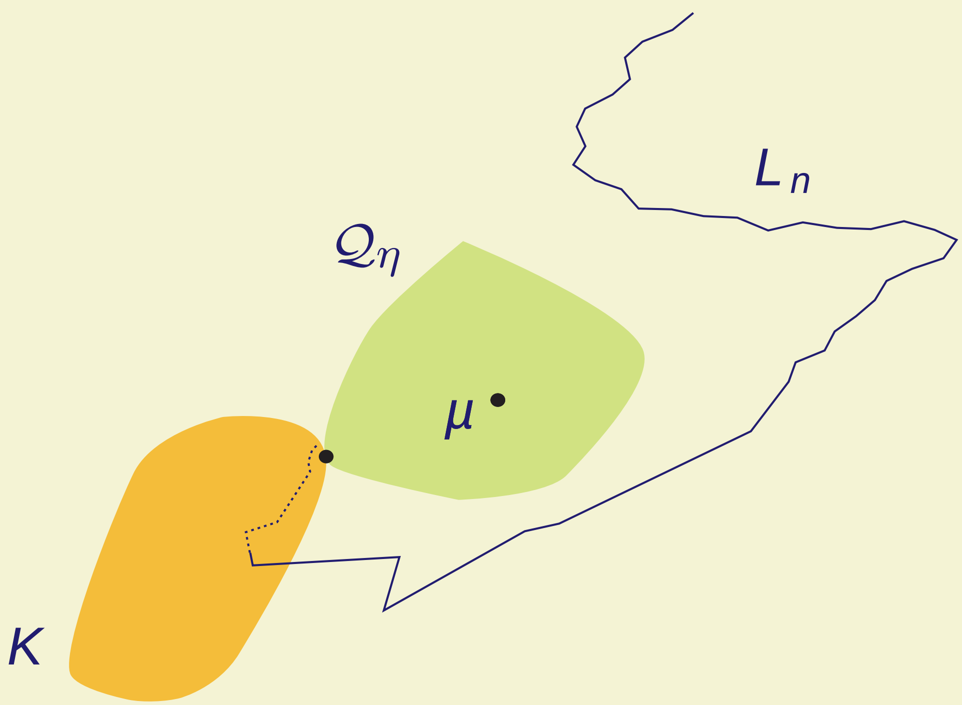
$$n^{-1} \log P\{L_n \quad K\} \to \; - \; ?$$



Sanov's Theorem

Convex set of probability measures $K \quad M \quad \mu \quad K$

$$n^{-1} \log \mathrm{P}\{L_n \quad K\} \to -\eta = -\inf_{\nu \quad K} J(\nu)$$

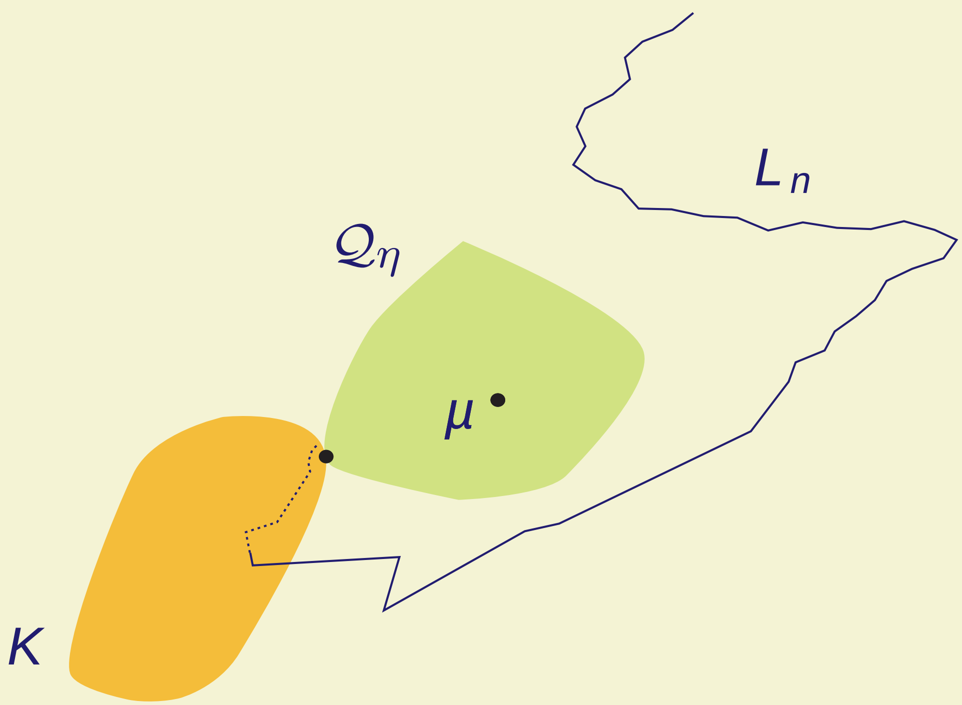


$$\mathcal{Q}_\eta = \{\nu : J(\nu) < \eta\}$$

Sanov's Theorem

i.i.d. source: $J(\nu) = D(\nu \quad \mu)$

Markov: $J(\nu) = \inf D(\nu \quad \check{P} \ \ \nu \quad P) \ : \ \check{P}$ tr. kernel with ν invariant



$\mathcal{Q}_\eta = \{\nu : J(\nu) < \eta\}$

Sanov's Theorem

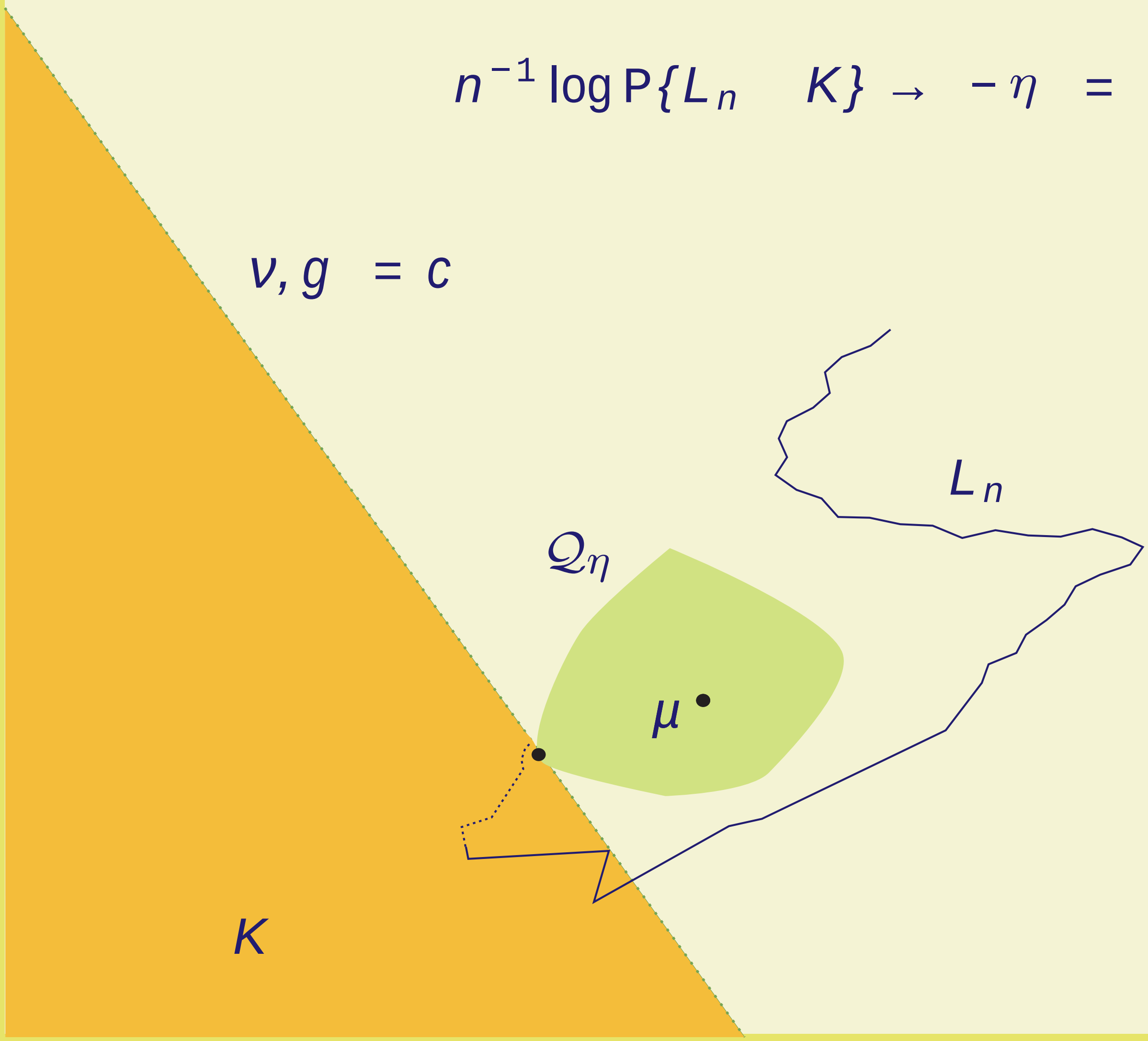
Example: $K = \{\nu : \nu, g \geq c\}$

$$n^{-1} \log \mathrm{P}\{L_n \quad K\} \to -\eta = -\inf_{\nu, g \geq c} J(\nu) = -\Lambda \ (c)$$

Sanov's Theorem

Example: $K = \{\nu : \nu, g \geq c\}$

$$n^{-1} \log P\{L_n \quad K\} \to -\eta = -\inf_{\nu, g \geq c} J(\nu) = -\Lambda(c)$$



$$\mathcal{Q}_\eta = \{\nu : J(\nu) < \eta\}$$

Outline

Neyman Pearson Hypothesis Testing

Observations $\boldsymbol{X} = \{X_t : t = 1, 2, \ldots N\}$

X i.i.d. with marginal π_j under H_j, $j = 0, 1$

Hypothesis test:

$$\phi(x) = 1 \text{ if } x \in A_1 \subset \mathcal{A}^N$$

Error Probabilities

$$\mathrm{P}_{e,0} = \mathrm{P}_0\{\phi(X) = 1\}, \quad \mathrm{P}_{e,1} = \mathrm{P}_1\{\phi(X) = 0\}$$

N-P Criterion: $\inf_{\phi} \mathrm{P}_{e,1}$ subject to $\mathrm{P}_{e,0} \leq e^{-N\eta}$

Neyman Pearson Hypothesis Testing

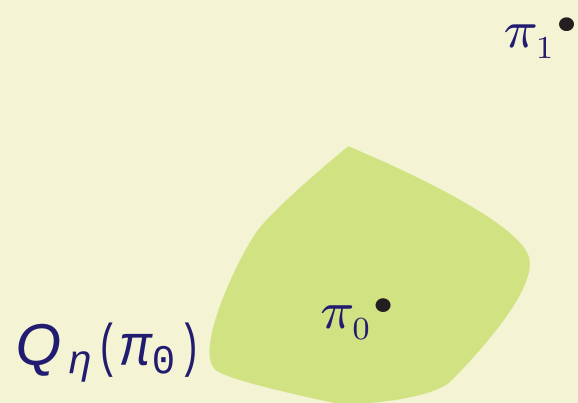
Observations $\boldsymbol{X} = \{X_t : t = 1, 2, \ldots N\}$

X i.i.d. with marginal π_j under H_j, $j = 0, 1$

Error Probabilities

$\mathrm{P}_{e,0} = \mathrm{P}_0\{\phi(X) = 1\}, \quad \mathrm{P}_{e,1} = \mathrm{P}_1\{\phi(X) = 0\}$

Solution: $\phi(X) = 0$ if $L_n \quad Q_\eta(\pi_0)$



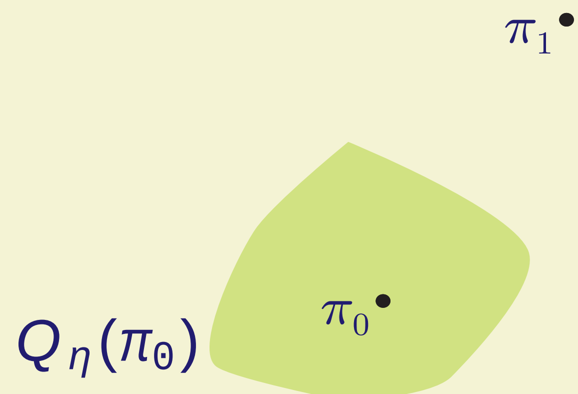
N-P Criterion: $\inf_{\phi} \mathrm{P}_{e,1}$ subject to $\mathrm{P}_{e,0} \leq e^{-N\eta}$

Neyman Pearson Hypothesis Testing

Solution: $\phi(X) = 0$ if $L_n \quad Q_\eta(\pi_0)$

$$\lim_{N \to \infty} N^{-1} \log \mathrm{P}_0\{\varphi_N = 1\} = -\eta$$

$$\lim_{N \to \infty} N^{-1} \log \mathrm{P}_1\{\varphi_N = 0\} = -\beta$$

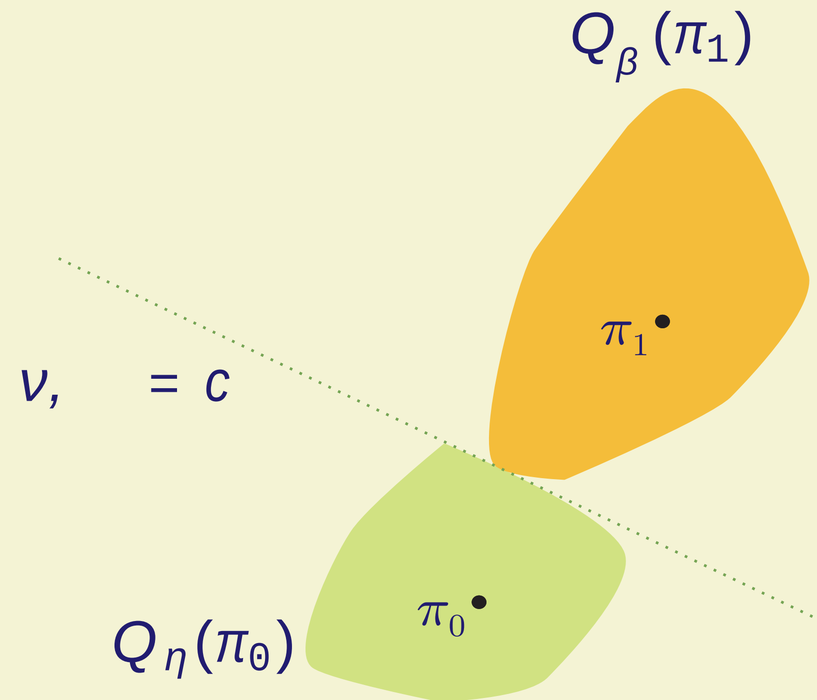


Neyman Pearson Hypothesis Testing

Solution: $\phi(X) = 0$ if $L_n \quad Q_\eta(\pi_0)$

$$\lim_{N \to \infty} N^{-1} \log \mathrm{P}_0\{\varphi_N = 1\} = -\eta$$

$$\lim_{N \to \infty} N^{-1} \log \mathrm{P}_1\{\varphi_N = 0\} = -\beta$$



$$\beta = \inf\{J_1(\nu) : J_0(\nu) \le \eta\}$$

$$= \inf\{\beta > 0 : Q_\beta(\pi_1) \cap Q_\eta(\pi_0) = \quad \}$$

Robust Neyman Pearson Hypothesis Testing

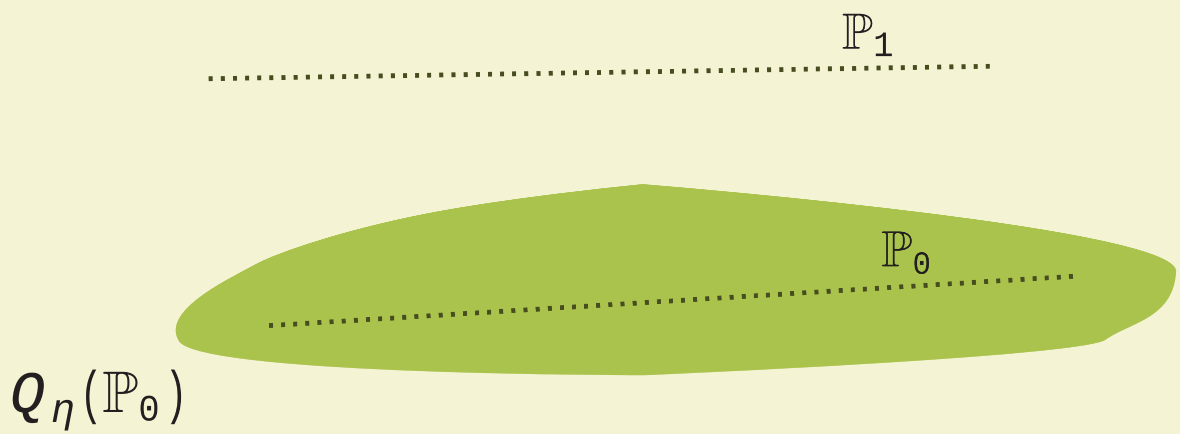
Uncertainty classes defined by moment constraints

$$\pi_0 \in \mathbb{P}_0 \qquad \pi_1 \in \mathbb{P}_1$$

Robust Neyman Pearson Hypothesis Testing

Uncertainty classes defined by moment constraints

$\pi_0 \in \mathbb{P}_0$ $\pi_1 \in \mathbb{P}_1$

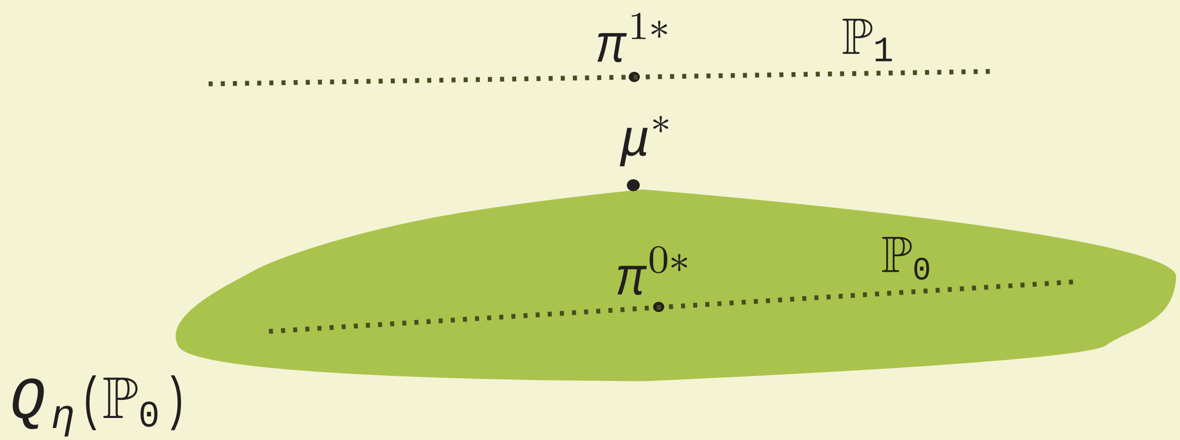


Robust Neyman Pearson Hypothesis Testing

Uncertainty classes defined by moment constraints

There exist $\pi_0^* \in \mathbb{P}_0, \pi_1^* \in \mathbb{P}_1$, and μ^* solving,

$$\beta^* = \inf_{\pi_1 \in \mathbb{P}_1} \inf_{\mu \in \mathcal{Q}_\eta(\mathbb{P}_0)} D(\mu \,\|\, \pi_1)$$

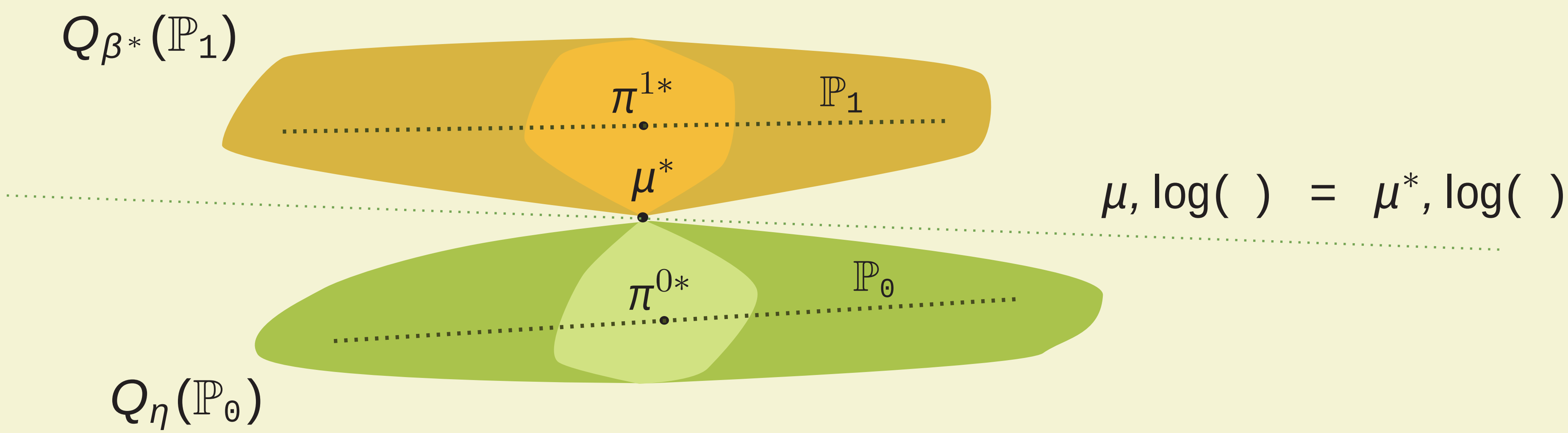


Robust Neyman Pearson Hypothesis Testing

Uncertainty classes defined by moment constraints

There exist $\pi_0^* \in \mathbb{P}_0, \pi_1^* \in \mathbb{P}_1$, and μ^* solving,

$$\beta^* = \inf_{\pi_1 \in \mathbb{P}_1} \inf_{\mu \in \mathcal{Q}_\eta(\mathbb{P}_0)} D(\mu \,\|\, \pi_1)$$



Optimizers again *discrete*

Outline

Channel Coding and Sanov's Theorem

Channel kernel $P(dy \mid x) = \mathrm{P}\{Y_t \in dx \mid X_t = x\}$

N-dimensional code words $X^i, \quad i = 1, 2, \ldots, e^{NR}$

N-dimensional output Y received

X is i.i.d. with marginal distribution μ

Y is i.i.d. with marginal distribution π

$$\pi(\cdot) = \int P(\cdot \mid x)\mu(dx)$$

Channel Coding and Sanov's Theorem

Channel kernel $\quad P(dy \mid x) = \mathsf{P}\{Y_t \quad dx \mid X_t = x\}$

N-dimensional code words $X^i, \qquad i = 1, 2, \ldots, e^{NR}$

N-dimensional output Y received

If i is the true codeword then
(X^i, Y) has marginal distribution

$$\mu \quad P(dx, dy) = \mu(dx) P(x, dy)$$

Otherwise, independence:

$$\mu \quad \pi(dx, dy) = \mu(dx) \pi(dy)$$

Channel Coding and Sanov's Theorem

Two hypotheses based on observations:

H_0 : $\mu \otimes \pi\,(dx, dy) = \mu(dx)\pi(dy)$

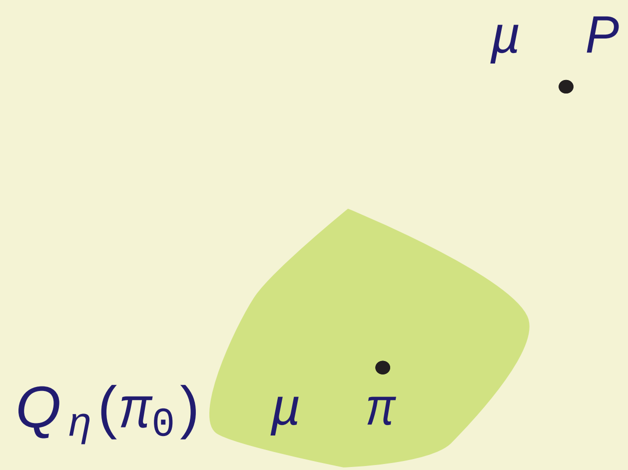
H_1 : $\mu \otimes P\,(dx, dy) = \mu(dx)P\,(x, dy)$

Channel Coding and Sanov's Theorem

Two hypotheses based on observations:

H_0 : $\mu \quad \pi(dx, dy) = \mu(dx)\pi(dy)$

H_1 : $\mu \quad P(dx, dy) = \mu(dx)P(x, dy)$



Solution: Reject codeword i ($\phi = 0$)

if $L_n \quad Q_\eta(\pi_0)$

Empirical distributions for
joint observations

(X^i, Y)

Channel Coding and Sanov's Theorem

Solution: $\phi = 0$ if $L_n \quad Q_\eta(\pi_0)$

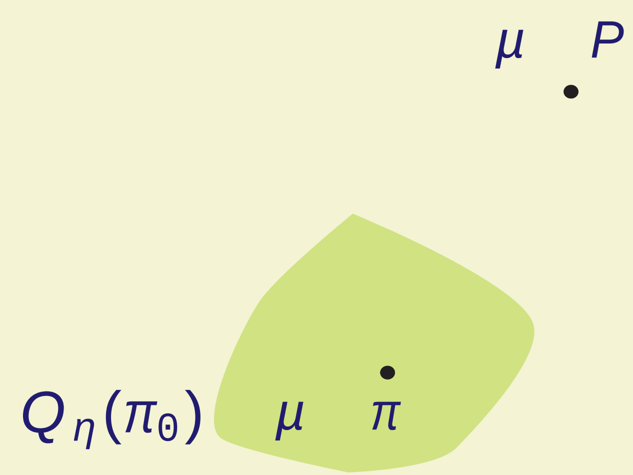
$$\lim_{N \to \infty} N^{-1} \log \mathrm{P}_0\{\varphi_N = 1\} = -\eta$$

The error probability $e^{-N\eta}$ must be multiplied by e^{NR}

For vanishing error,

$$e^{NR} \times e^{-N\eta} < 1$$

That is, $R < \eta$



Channel Coding and Sanov's Theorem

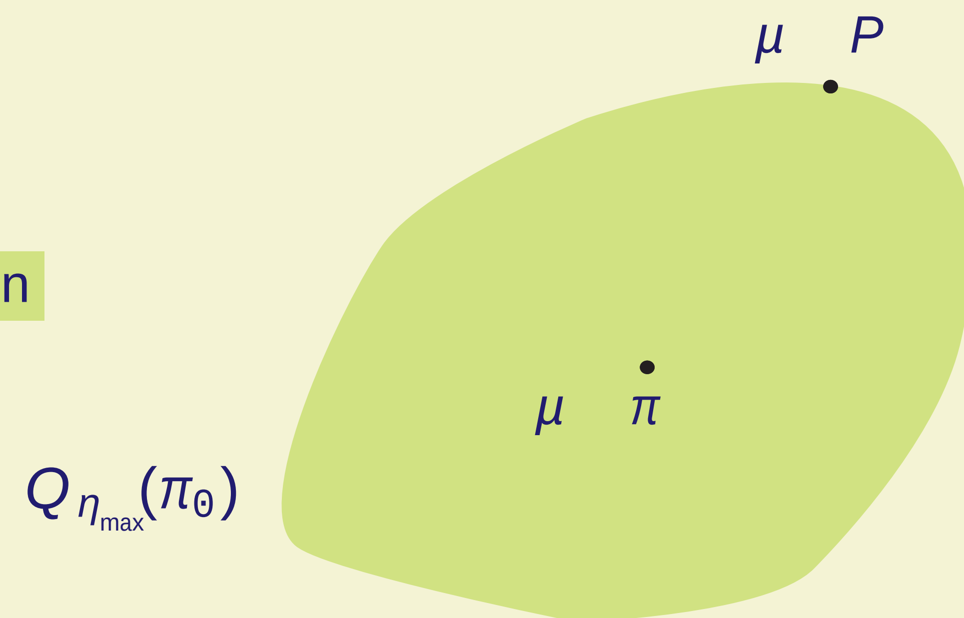
Solution: $\phi = 0$ if $L_n \quad Q_\eta(\pi_0)$

$$\lim_{N \to \infty} N^{-1} \log \mathrm{P}_0\{\varphi_N = 1\} = -\eta$$

The error probability $e^{-N\eta}$ must be multiplied by e^{NR}

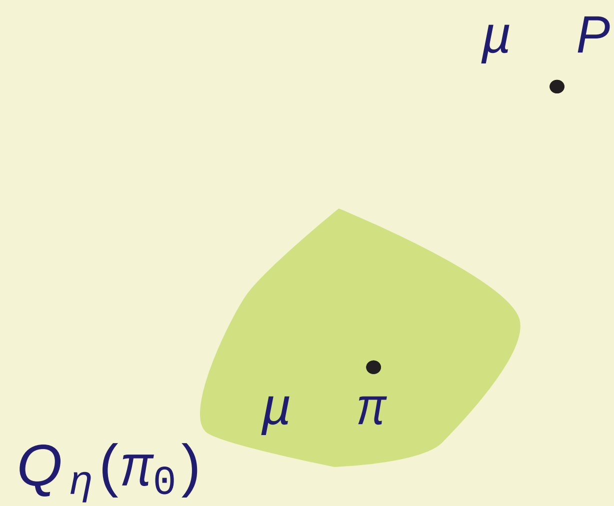
$$R < \eta_{\max} = D(\mu \quad P \;\; \mu \quad \pi)$$

$$= \text{mutual information}$$



Error Exponent

$$
\begin{aligned}
E(R,\mu) &= -\lim_{N \to \infty} N^{-1} \log \mathrm{P}\{\text{error}\} \\
&= -N^{-1} \log \{e^{NR} \times e^{-N\eta} \times e^{-N\beta}\} \\
&= -R + \eta + \beta \qquad \text{(some } \eta\text{, small } \beta\text{)}
\end{aligned}
$$



Error Exponent

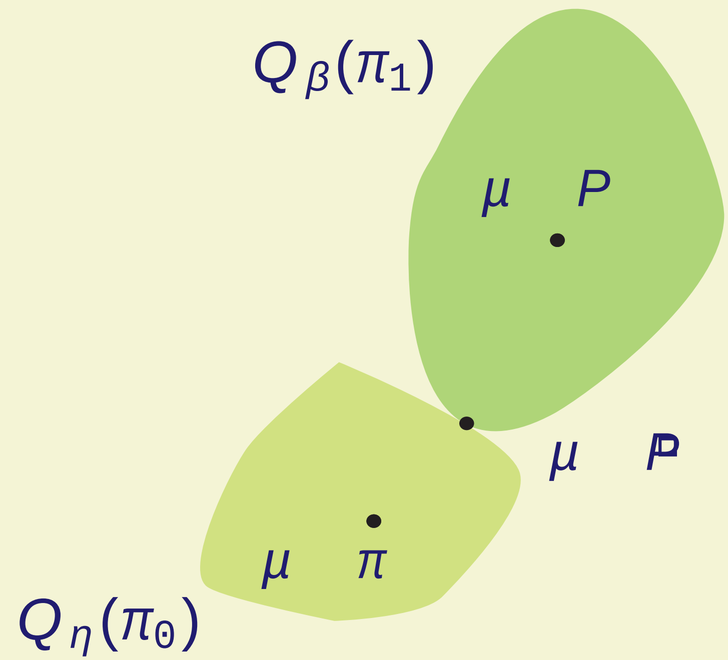
$$E(R,\mu) = -\lim_{N \to \infty} N^{-1} \log \mathrm{P}\{\text{error}\}$$

$$= -N^{-1} \log \{e^{NR} \times e^{-N\eta} \times e^{-N\beta}\}$$

$$= -R + \eta + \beta$$

Set $\eta = R$

$$\beta = E(R,\mu) = \inf_{P} D(\mu \quad P \;\; \mu \quad P)$$



Outline

Summary

Standard coding based on AWGN models

May be unrealistic in wireless models with fading

Discrete distributions arise in coding,
and other applications involving optimization over M

Extremal distributions arise in worst-case models

What's Next?

II Channel models

Convex optimization and channel coding

Cutting plane algorithm

III Worst-case models

Extremal distributions