Information, Divergence and Risk for Binary Classification

Mark Reid* [mark.reid@anu.edu.au]
Research School of Information Science and Engineering
The Australian National University, Canberra, ACT, Australia

THE AUSTRALIAN NATIONAL UNIVERSITY

Machine Learning Summer School
Thursday, 29th January 2009

MLSS.cc

*Joint work with **Robert Williamson**

Brooke Taylor
(1685-1731)

Johan Jensen
(1859-1925)

Taylor & Jensen's

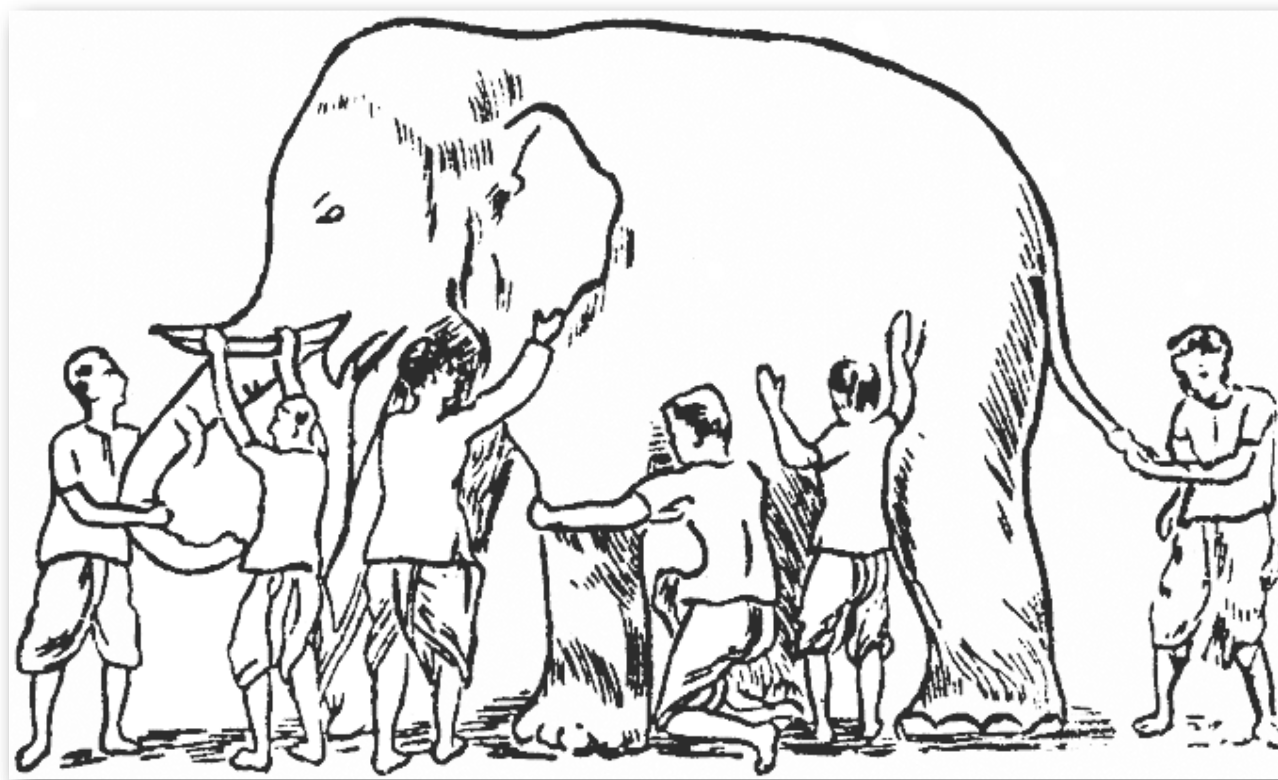
Most Excellent Adventure

through

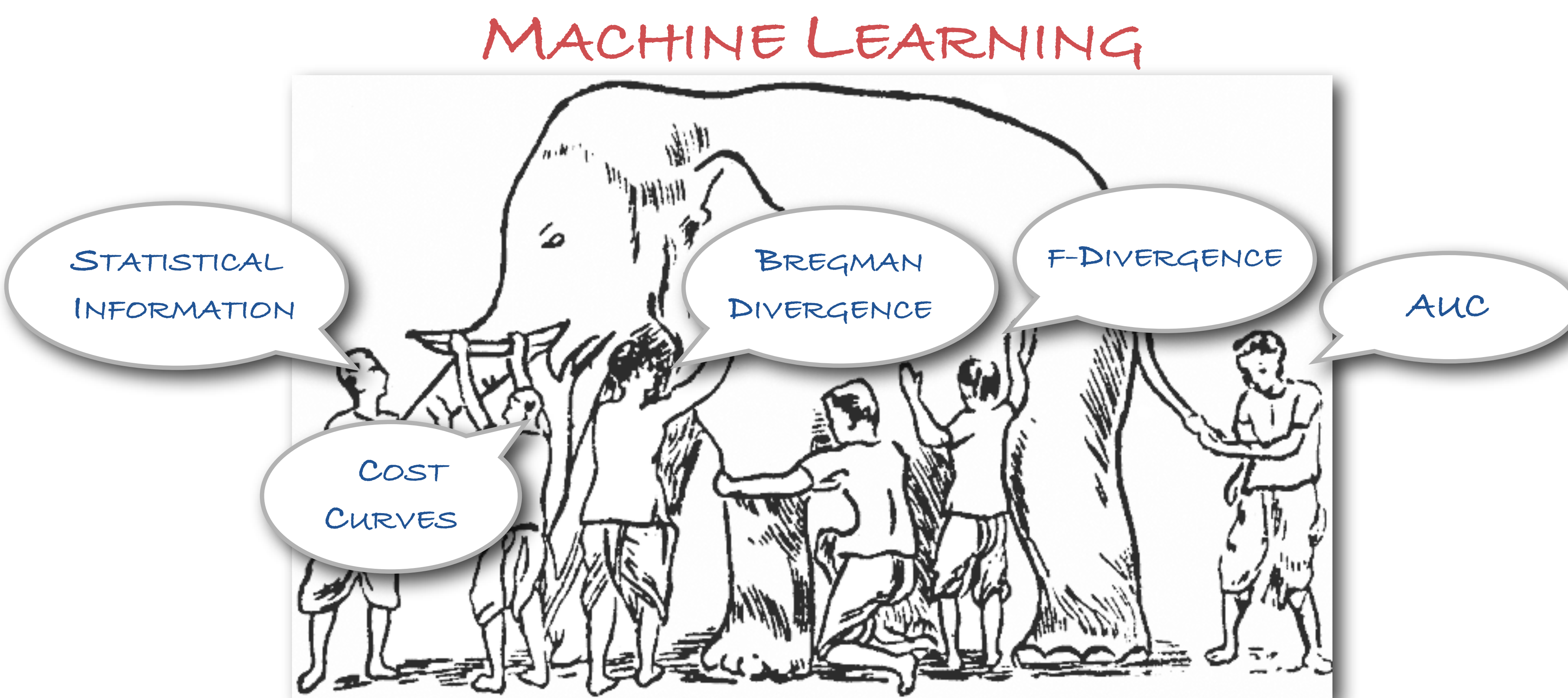
Statistical Learning Theory

Introduction

The Blind Men & The Elephant



The Blind Men & The Elephant



Mathematics is the art of giving the same name to different things.

Jules Henri Poincaré (1854-1912)

What’s in it for me?

What to expect

- Definitions
- Relationships
- Representations
- Proofs

What’s in it for me?

What to expect

- Definitions
- Relationships
- Representations
- Proofs

What <u>not</u> to expect

- Algorithms
- Models
- Data
- Technicalities

What's in it for me?

What to expect

- Definitions
- Relationships
- Representations
- Proofs

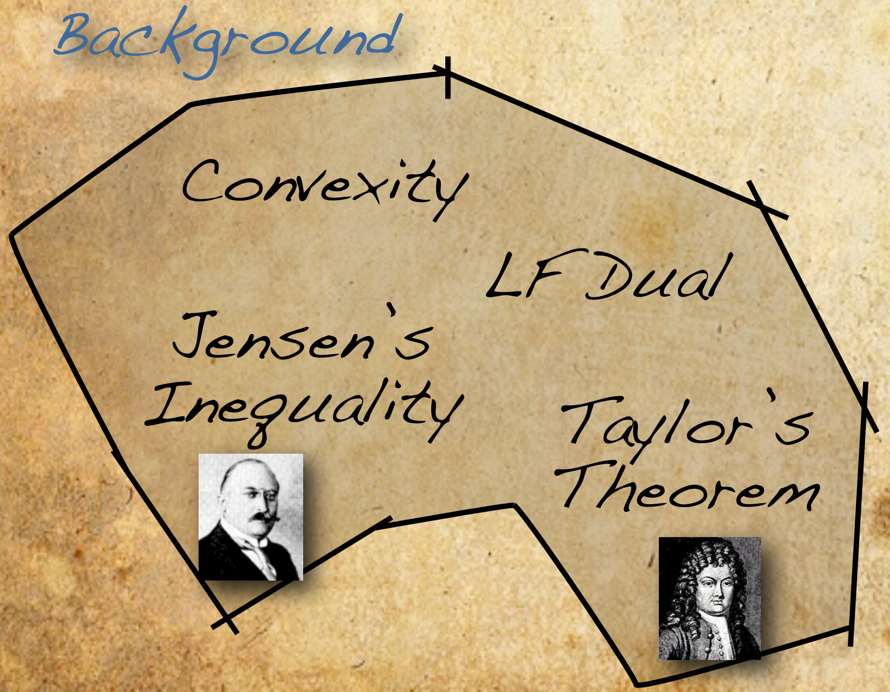
Theory

What <u>not</u> to expect

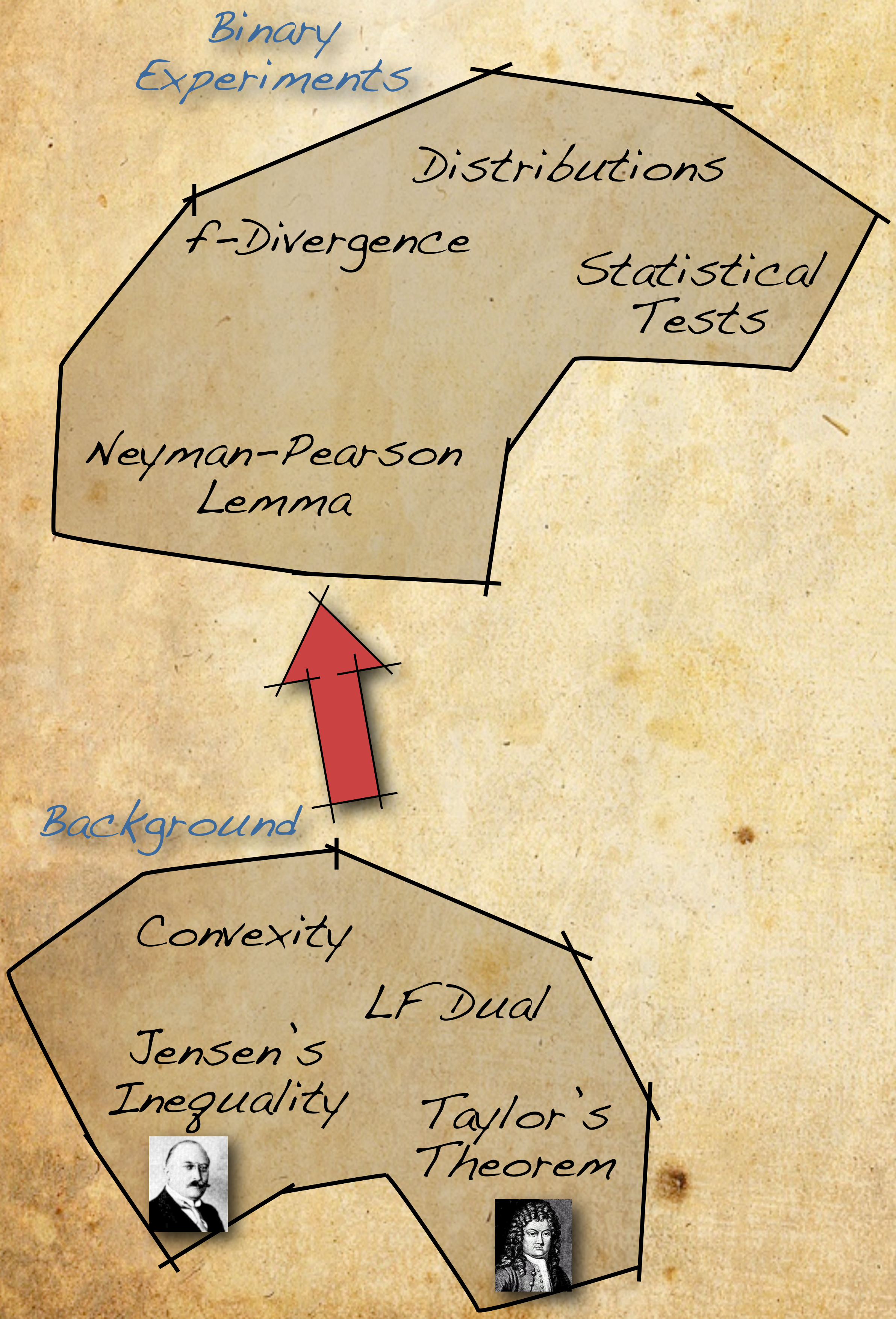
- Algorithms
- Models
- Data
- Technicalities

Practice

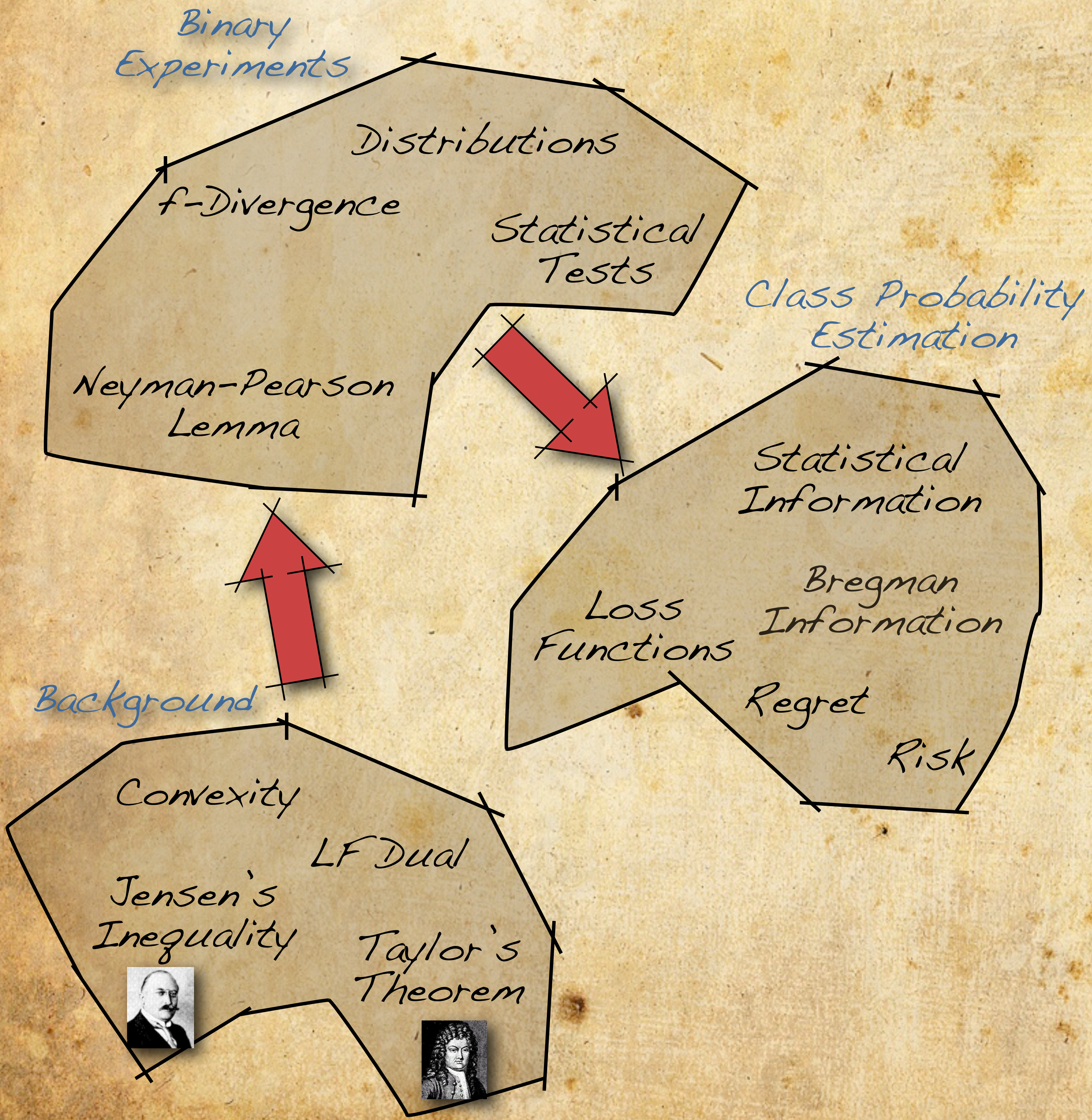
Terra Statistica



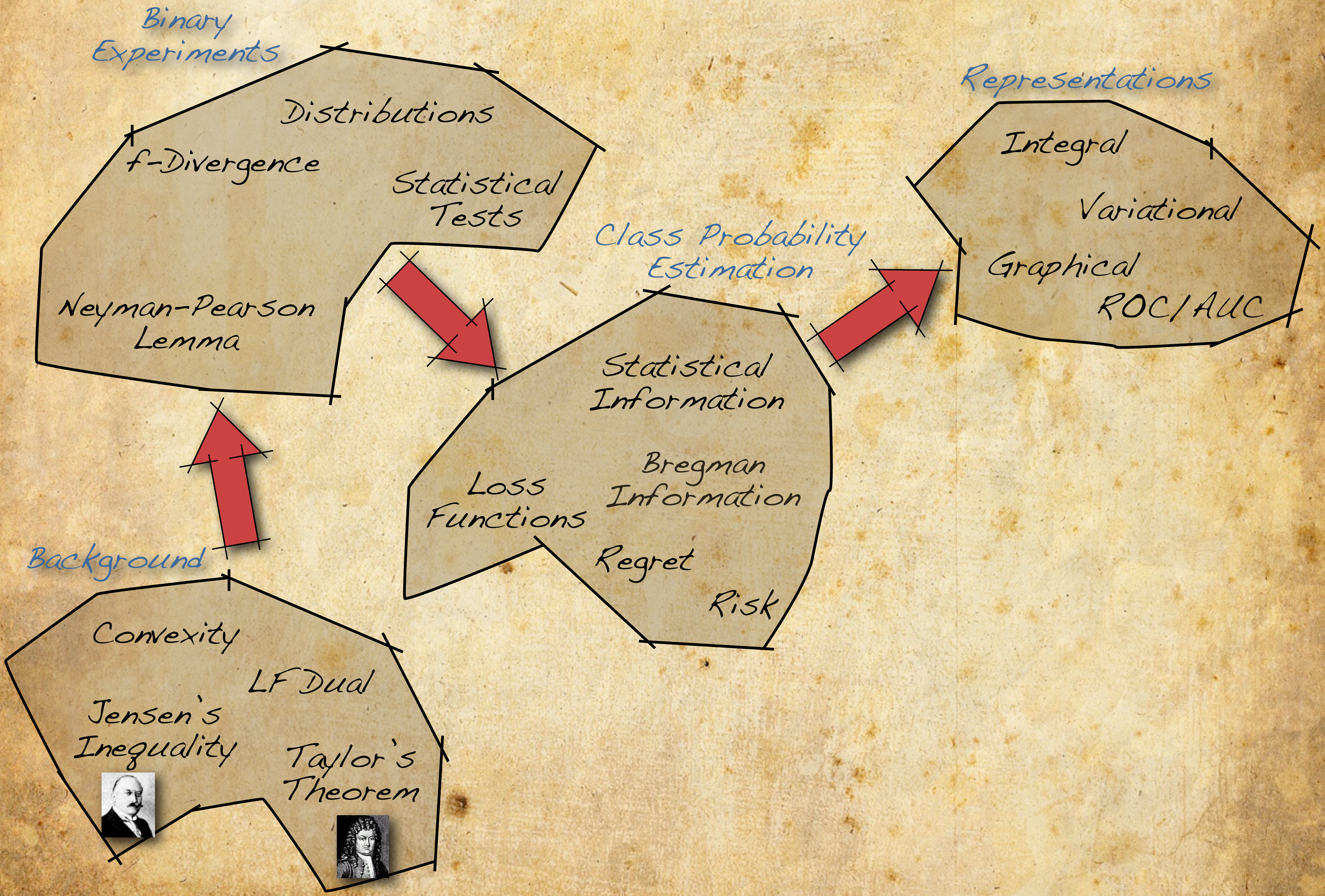
Terra Statistica

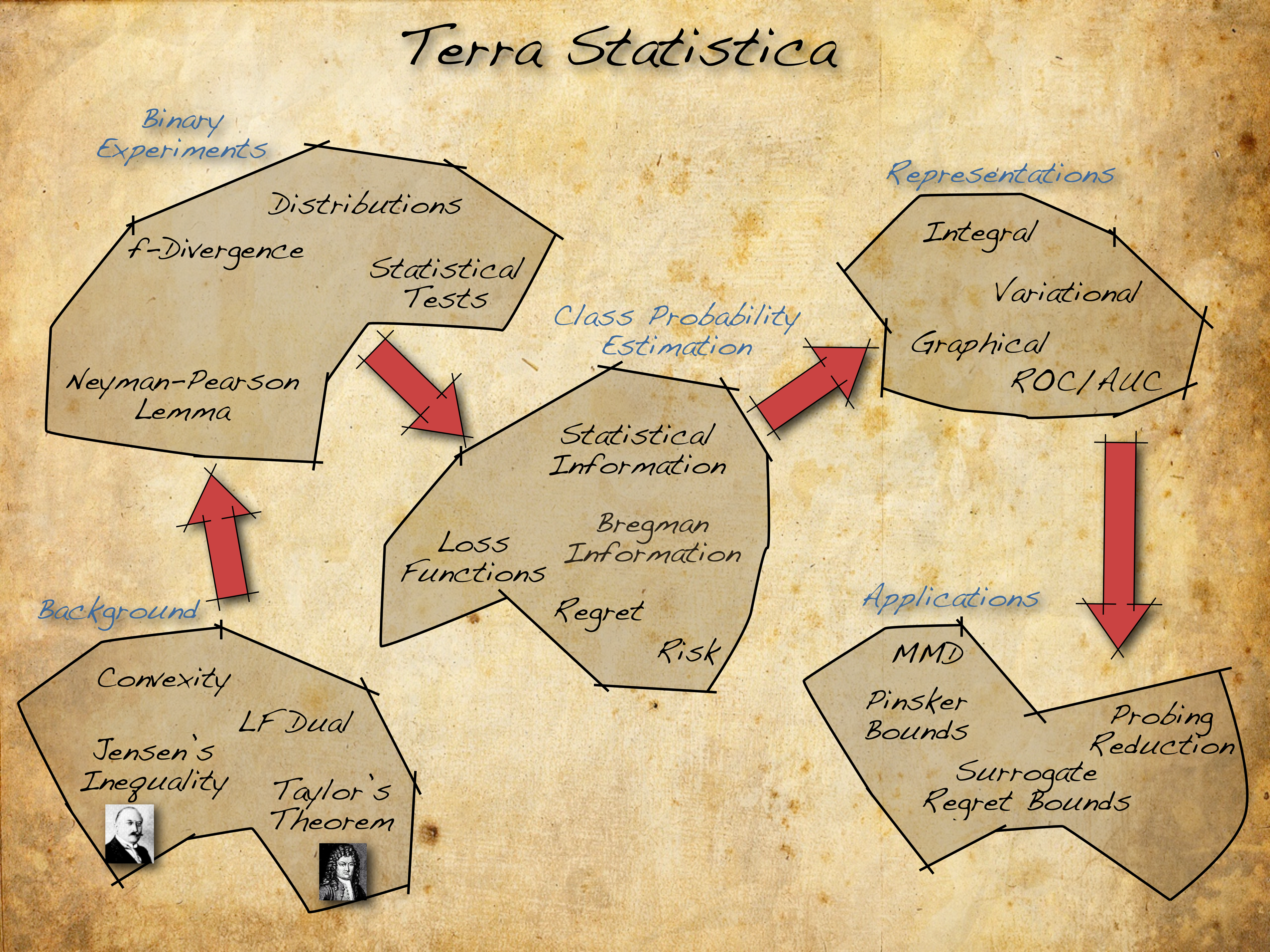


Terra Statistica



Terra Statistica




Terra Statistica
Binary Experiments
Distributions
f-Divergence
Statistical Tests
Neyman-Pearson Lemma
Class Probability Estimation
Statistical Information
Bregman Information
Loss Functions
Regret
Risk
Representations
Integral
Variational
Graphical
ROC/AUC
Background
Convexity
LF Dual
Jensen's Inequality
Taylor's Theorem
Applications
MMD
Pinsker Bounds
Probing Reduction
Surrogate Regret Bounds

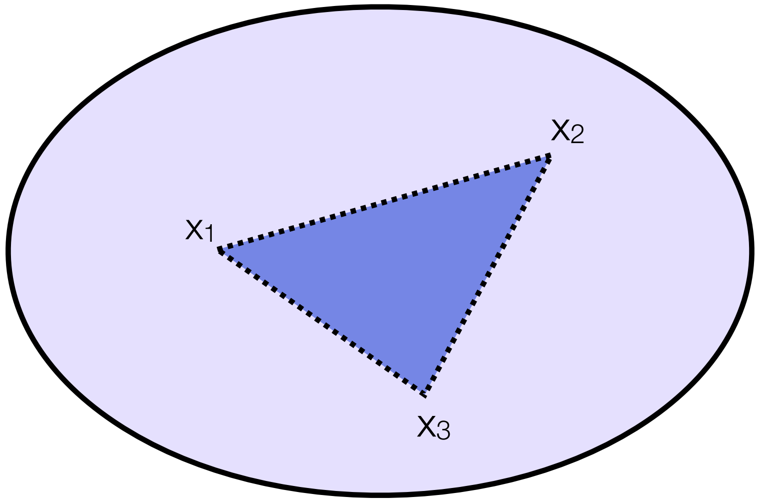
Part I: Convexity, Binary Experiments & Classification

Convexity

Convex Sets

- We say $\mathcal{S} \subseteq \mathbb{R}^d$ is a **convex set** if it is closed under convex combination. That is, for any n, any $\mathbf{x}_1, \ldots, \mathbf{x}_n \subset \mathcal{S}$ and weights $\lambda_1, \ldots, \lambda_n \geq 0$ such that $\sum_{i=1}^{n} \lambda_i = 1$

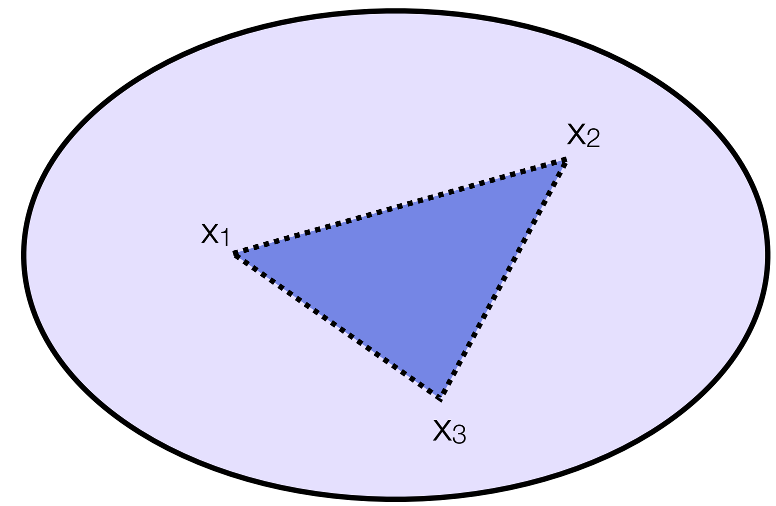
$$\sum_{i=1}^{n} \lambda_i \mathbf{x}_i \in \mathcal{S}$$



Convex Sets

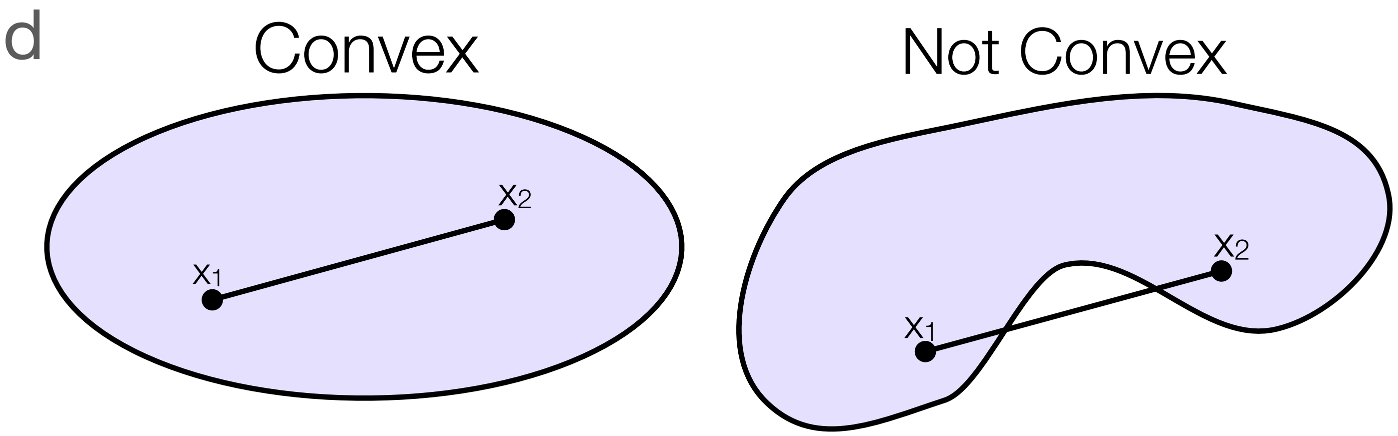
- We say $\mathcal{S} \subseteq \mathbb{R}^d$ is a **convex set** if it is closed under convex combination. That is, for any n, any $\mathbf{x}_1, \ldots, \mathbf{x}_n \subset \mathcal{S}$ and weights $\lambda_1, \ldots, \lambda_n \geq 0$ such that $\sum_{i=1}^{n} \lambda_i = 1$

$$\sum_{i=1}^{n} \lambda_i \mathbf{x}_i \in \mathcal{S}$$



- Suffices to show for all $\lambda \in [0, 1]$ and $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}$ that

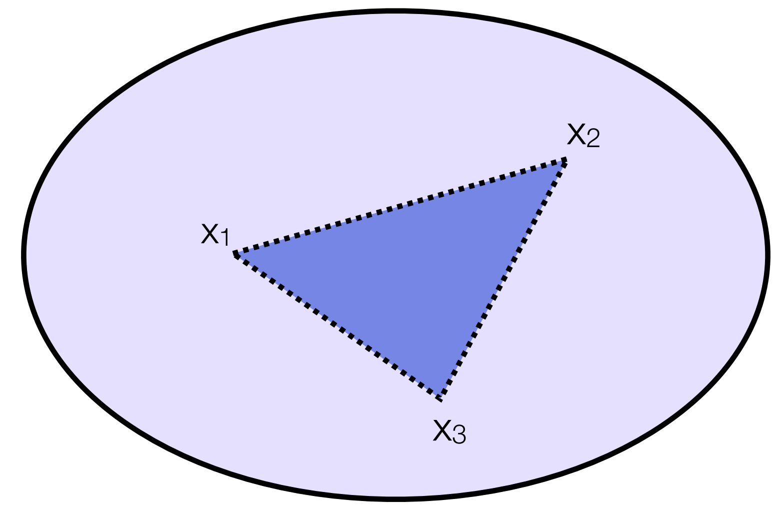
$$\lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2 \in \mathcal{S}$$



Convex Sets

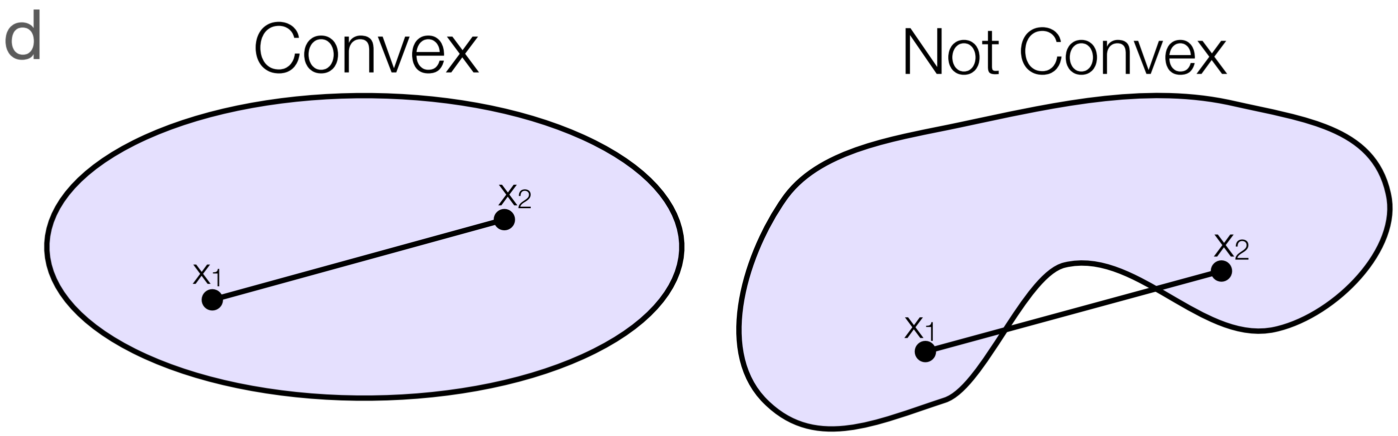
- We say $\mathcal{S} \subseteq \mathbb{R}^d$ is a **convex set** if it is closed under convex combination. That is, for any n, any $\mathbf{x}_1, \ldots, \mathbf{x}_n \subset \mathcal{S}$ and weights $\lambda_1, \ldots, \lambda_n \geq 0$ such that $\sum_{i=1}^n \lambda_i = 1$

$$\sum_{i=1}^{n} \lambda_i \mathbf{x}_i \in \mathcal{S}$$



- Suffices to show for all $\lambda \in [0, 1]$ and $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}$ that

$$\lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2 \in \mathcal{S}$$



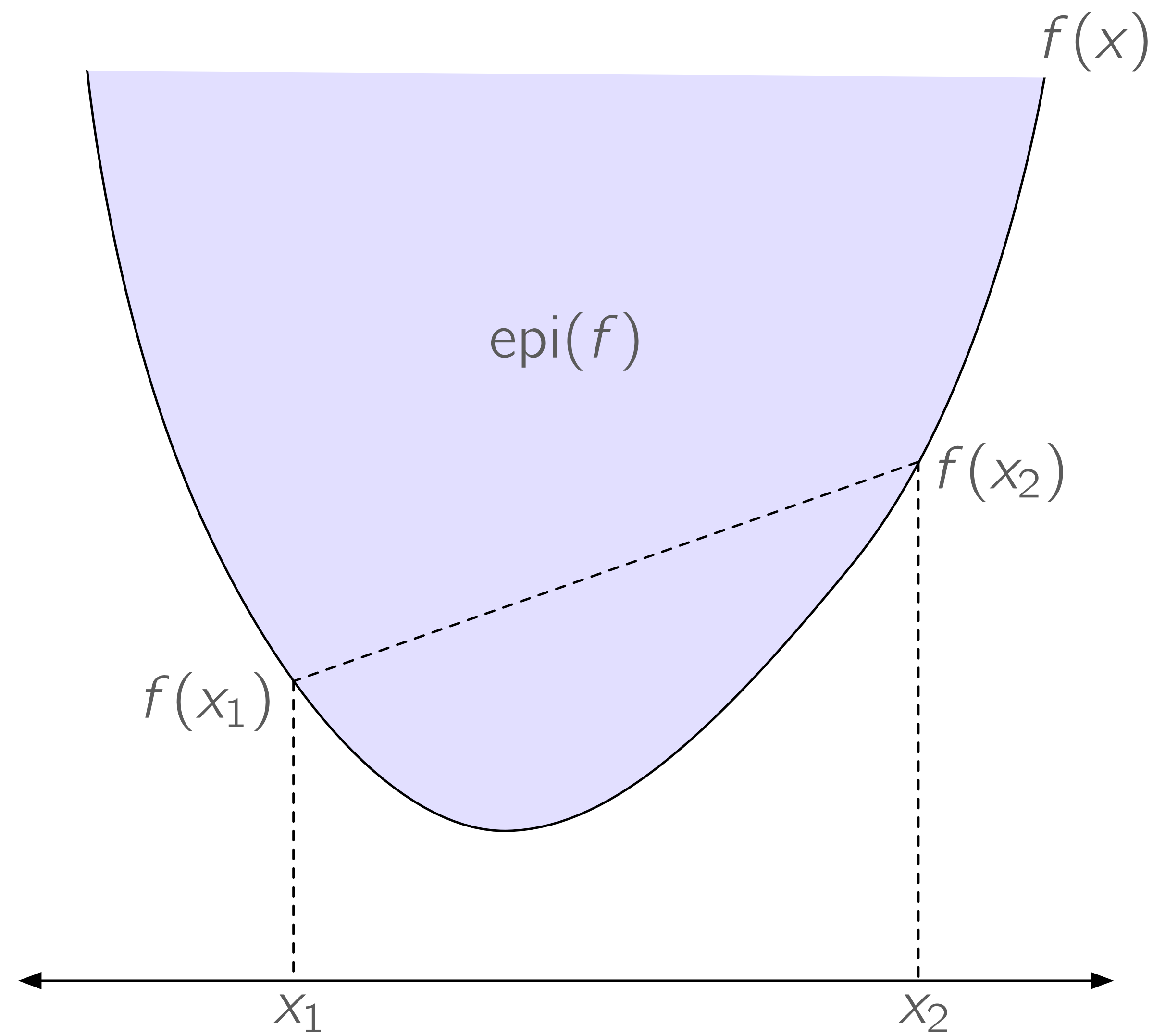
- Convex = “closed under expectation”

Convex Functions

- The **epigraph** of a function is the set of points that lie above it:

$$\mathrm{epi}(f) := \{(\mathbf{x}, y) : \mathbf{x} \in \mathbb{R}^d, y \geq f(\mathbf{x})\}$$

- A function is **convex** if its epigraph is a convex set
  - A convex function is necessarily continuous



Taylor's Theorem

Integral Form of Taylor Expansion

- Let $[t_0, t]$ be an interval on which *f* is twice differentiable. Then

$$f(t) = f(t_0) + (t - t_0)f'(t_0) + \int_{t_0}^{t} (t - s)\, f''(s)\, ds$$

Taylor's Theorem

Integral Form of Taylor Expansion

- Let $[t_0, t]$ be an interval on which *f* is twice differentiable. Then

$$f(t) = f(t_0) + (t - t_0)f'(t_0) + \int_{t_0}^{t} (t - s)\, f''(s)\, ds$$

Corollary

- Let *f* be twice differentiable on [*a*,*b*]. Then, for all *t* in [*a*,*b*],

$$f(t) = f(t_0) + (t - t_0)f'(t_0) + \int_{a}^{b} g(t, s)\, f''(s)\, ds$$

where

$$g(t,s) = \begin{cases} (t-s) & t_0 \le s < t \\ (s-t) & t \le s < t_0 \\ 0 & \text{otherwise} \end{cases}$$

- Differentiability can be removed if *f'* and *f''* are interpreted distributionally

Integral Form of the Taylor Expansion

$$f(t) = f(t_0) + (t - t_0)f'(t_0) + \int_a^b g(t,s)\, f''(s)\, ds$$

where

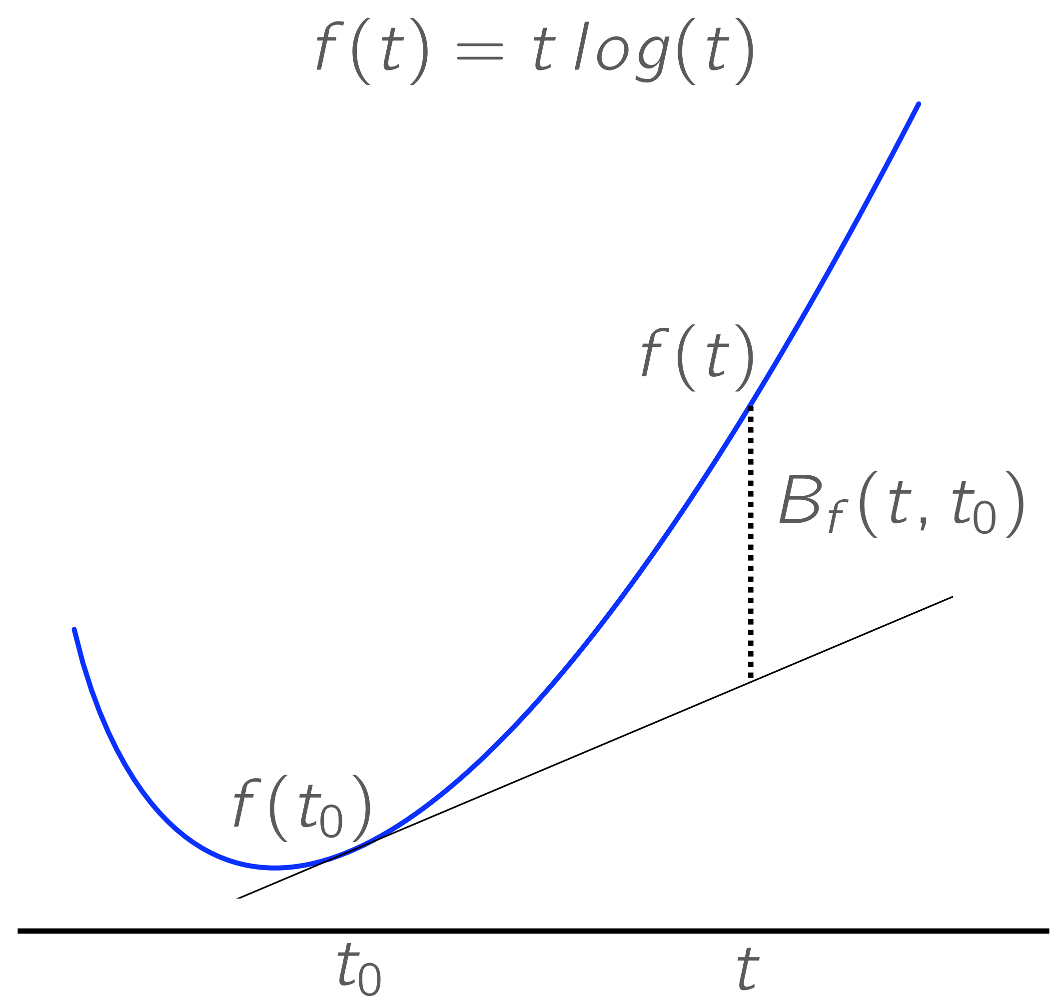
$$g(t,s) = (t - s)[\![t_0 \leq s < t]\!] + (s - t)[\![t \leq s < t_0]\!]$$

$$[\![p]\!] = \begin{cases} 1, & p \text{ is true} \\ 0, & \text{otherwise} \end{cases}$$

Bregman Divergence

- A **Bregman divergence** is a general class of “distance” measures defined using convex functions

$$B_f(t, t_0) := f(t) - f(t_0) - \langle t - t_0, \nabla f(t_0) \rangle$$



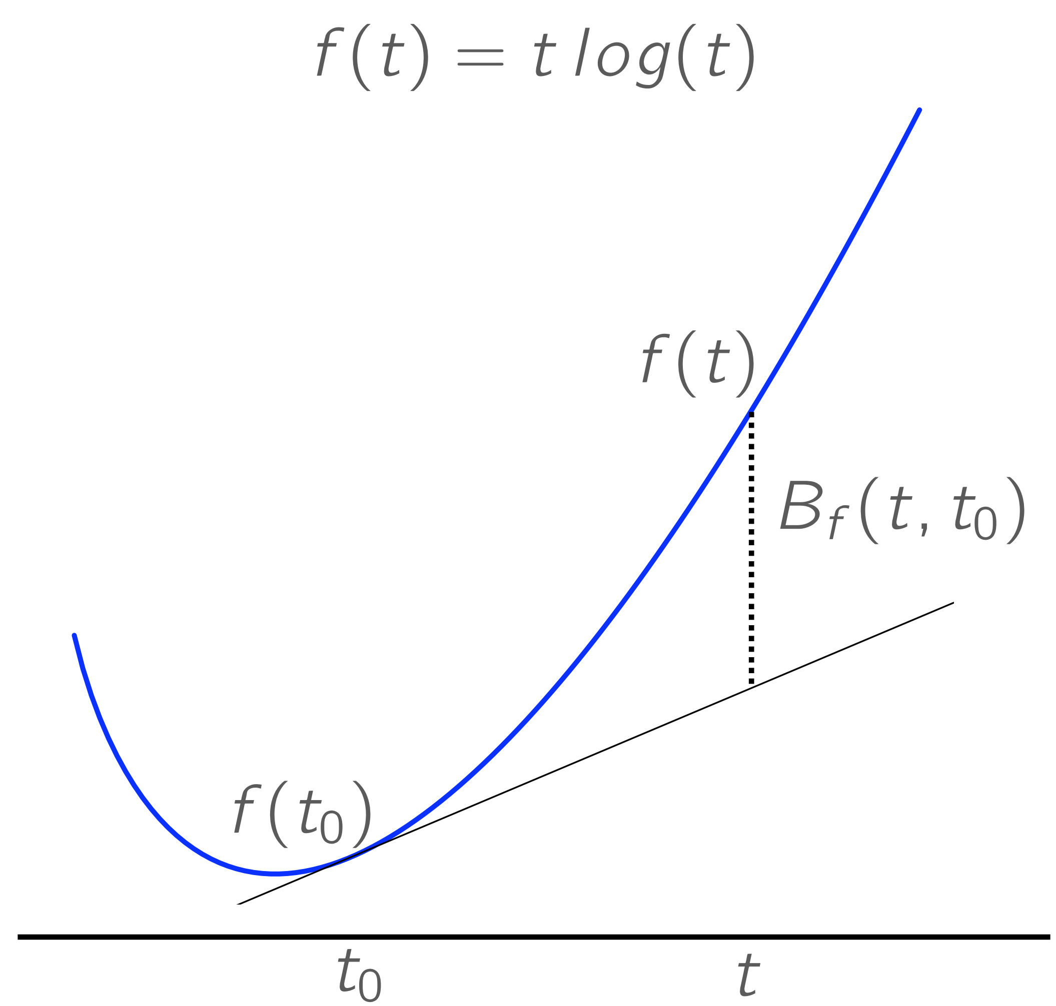
Bregman Divergence

- A **Bregman divergence** is a general class of “distance” measures defined using convex functions

$$B_f(t, t_0) := f(t) - f(t_0) - \langle t - t_0, \nabla f(t_0) \rangle$$

- In 1-d case, $B_f(t, t_0)$ is the non-linear part of the Taylor expansion of f

$$B_f(t, t_0) := \int_{t_0}^{t} (t - s)\, f''(s)\, ds$$



Jensen's Inequality

Jensen Gap

- For convex $f : \mathbb{R} \to \mathbb{R}$ and distribution *P* define

$$\mathbb{J}_P[f(x)] := \mathbb{E}_P\left[f(x)\right] - f\left(\mathbb{E}_P[x]\right)$$

Jensen's Inequality

Jensen Gap

- For convex $f : \mathbb{R} \to \mathbb{R}$ and distribution *P* define

$$\mathbb{J}_P[f(x)] := \mathbb{E}_P\left[f(x)\right] - f\left(\mathbb{E}_P[x]\right)$$

Jensen's Inequality

- The Jensen Gap is non-negative for all *P* **if and only if** *f* is convex

Jensen's Inequality

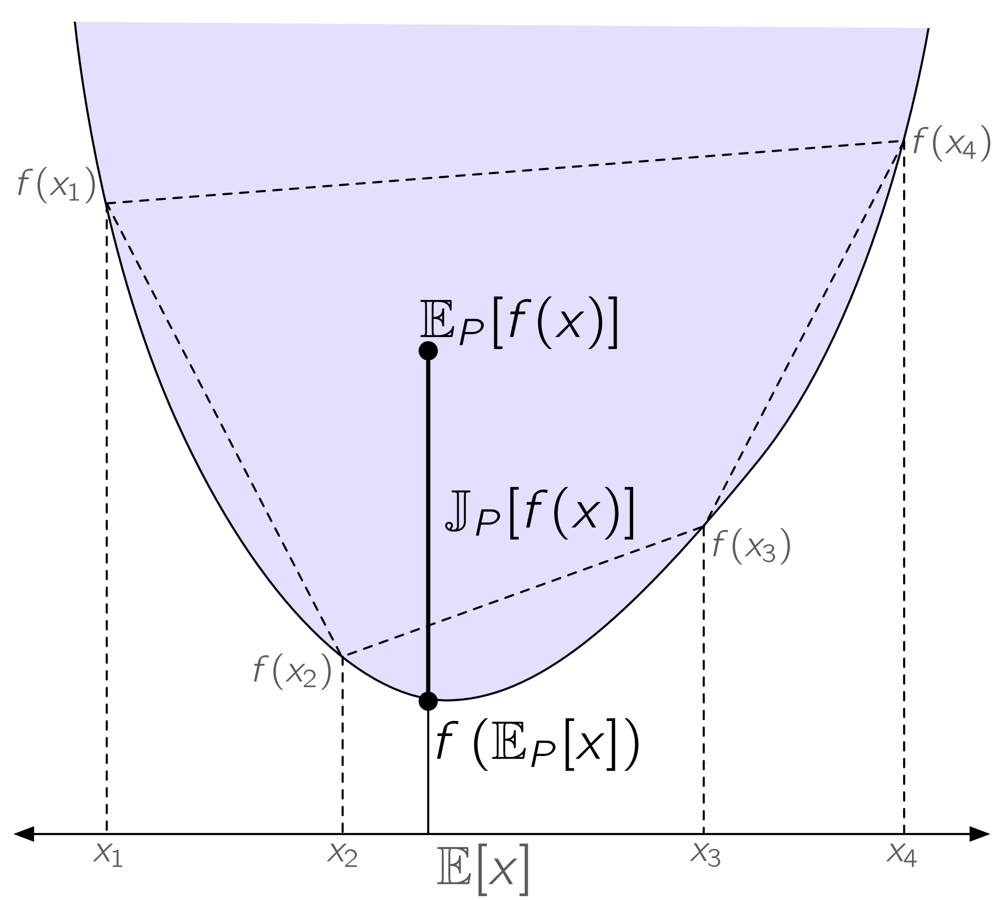
Jensen Gap

- For convex $f : \mathbb{R} \to \mathbb{R}$ and distribution P define

$$\mathbb{J}_P[f(x)] := \mathbb{E}_P\left[f(x)\right] - f\left(\mathbb{E}_P[x]\right)$$

Jensen's Inequality

- The Jensen Gap is non-negative for all P **if and only if** f is convex

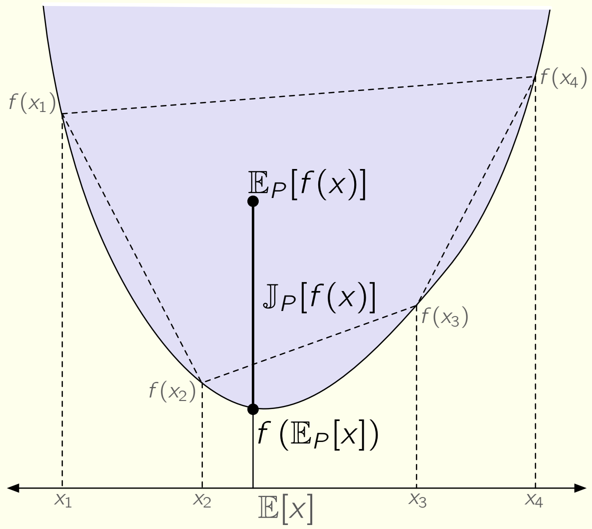


Jensen's Inequality

$$\mathbb{J}_P[f(x)] := \mathbb{E}_P\left[f(x)\right] - f\left(\mathbb{E}_P[x]\right) \geq 0$$

if and only if

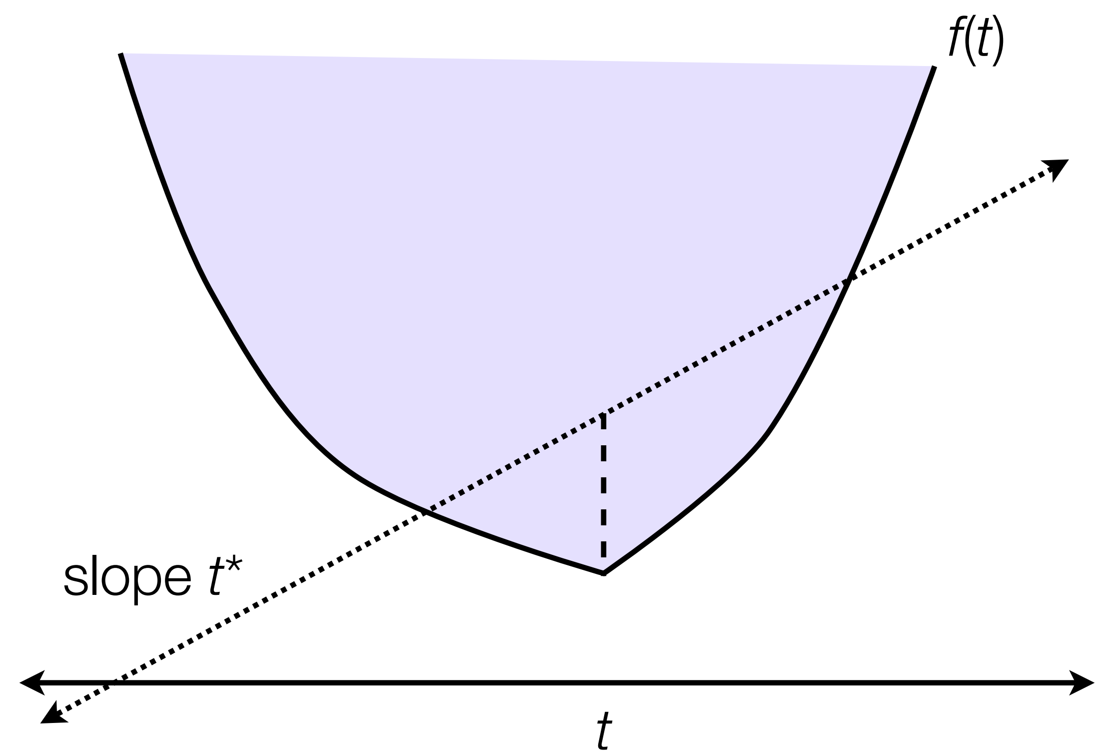
f is convex



The Legendre-Fenchel Transform

- The LF Transform generalises the notion of a derivative to non-differentiable functions

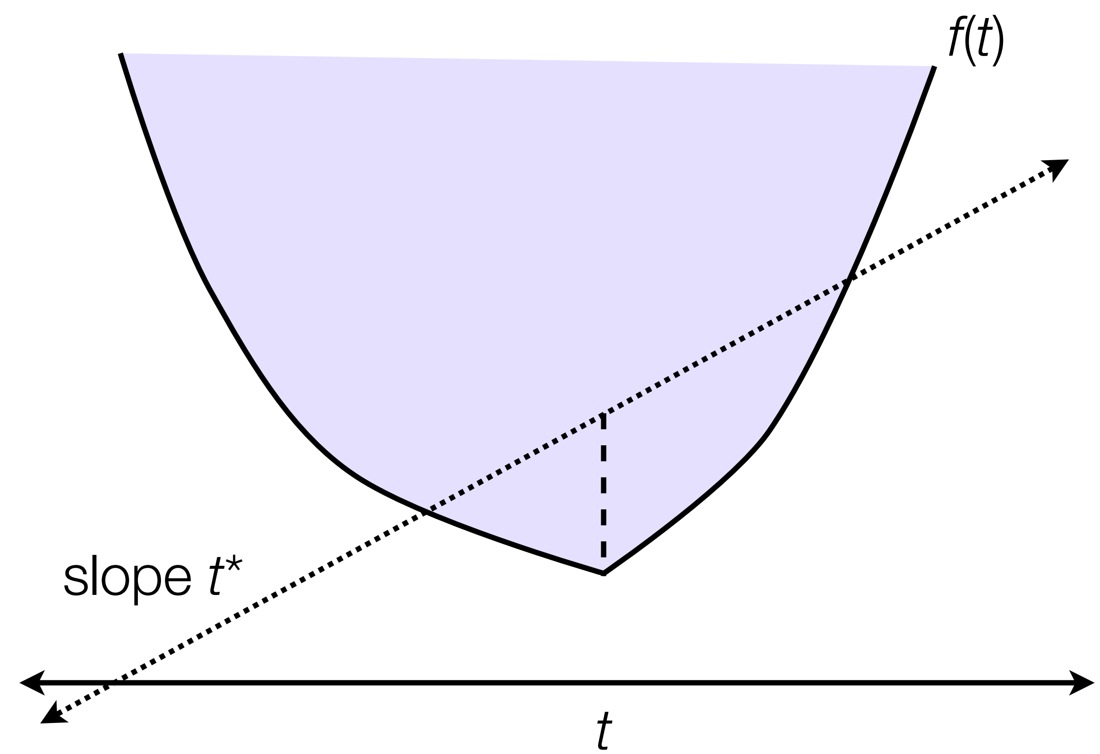
$$f^*(t^*) = \sup_{t \in \mathbb{R}^d} \{\langle t, t^* \rangle - f(t)\}$$



The Legendre-Fenchel Transform

- The LF Transform generalises the notion of a derivative to non-differentiable functions

$$f^*(t^*) = \sup_{t \in \mathbb{R}^d} \{\langle t, t^* \rangle - f(t)\}$$



- The **double LF transform** or **biconjugate**

$$f^{**}(t) = \sup_{t^* \in \mathbb{R}^d} \{\langle t^*, t \rangle - f^*(t^*)\}$$

is **involutive** for convex *f.* That is,

$$f^{**}(t) = f(t)$$

Representations of Convex Functions

Integral Representation

- Via Taylor's Theorem

$$f(t) = \Lambda_f(t) + \int_a^b g(t,s)\, f''(s)\, ds$$

where

$$\Lambda_f(t) = f(t_0) + f'(t_0)(t - t_0)$$

$$g(t,s) = \begin{cases} (t-s)_+ & s \geq t_0 \\ (s-t)_+ & s < t_0 \end{cases}$$

Variational Representation

- Via Fenchel Dual

$$f(t) = \sup_{t^* \in \mathbb{R}} \{t.t^* - f^*(t^*)\}$$

where

$$f^*(t) = \sup_{t \in \mathbb{R}} \{t.t^* - f(t)\}$$

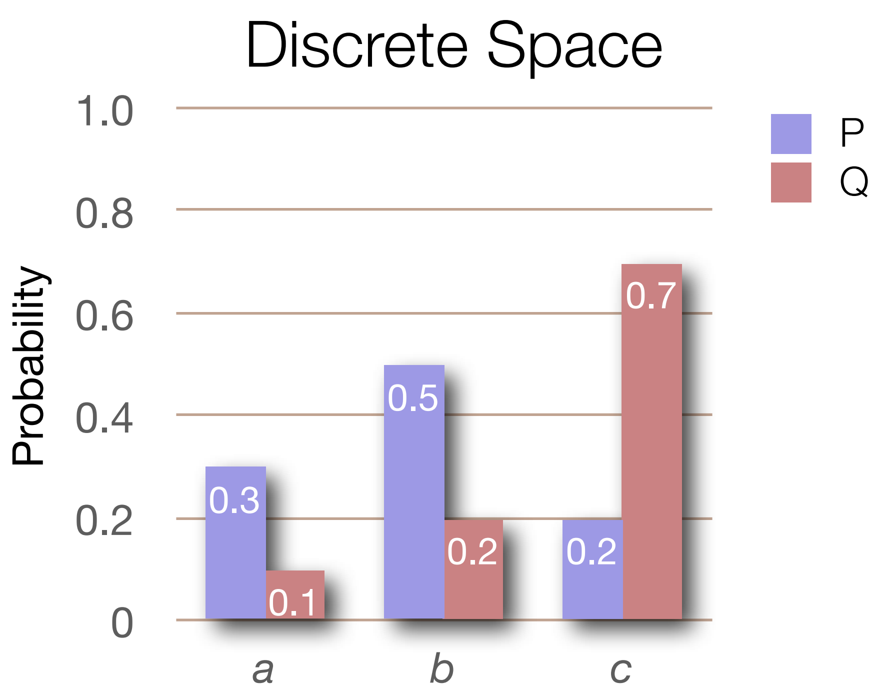
Terra Statistica

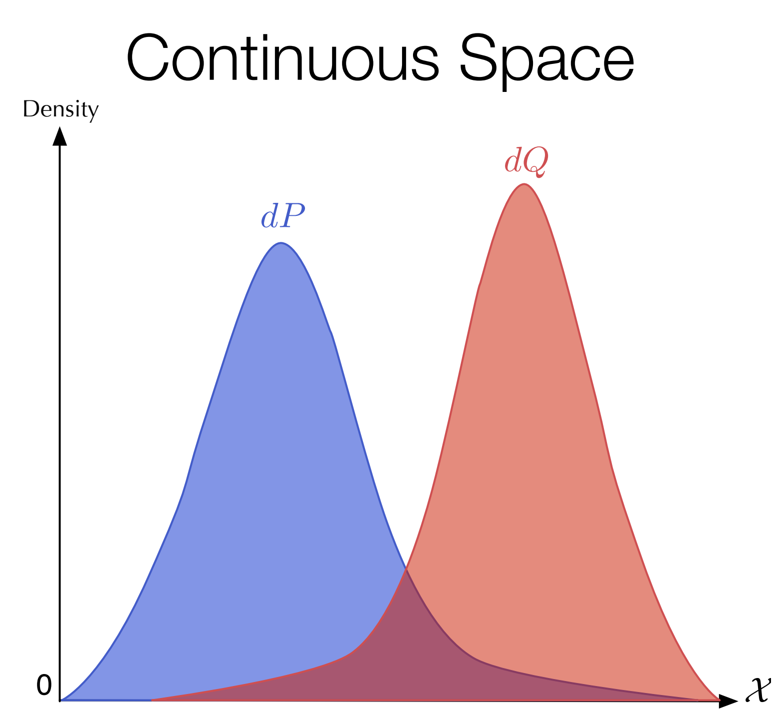
Terra Statistica

Binary Experiments and Measures of Divergence

Binary Experiments

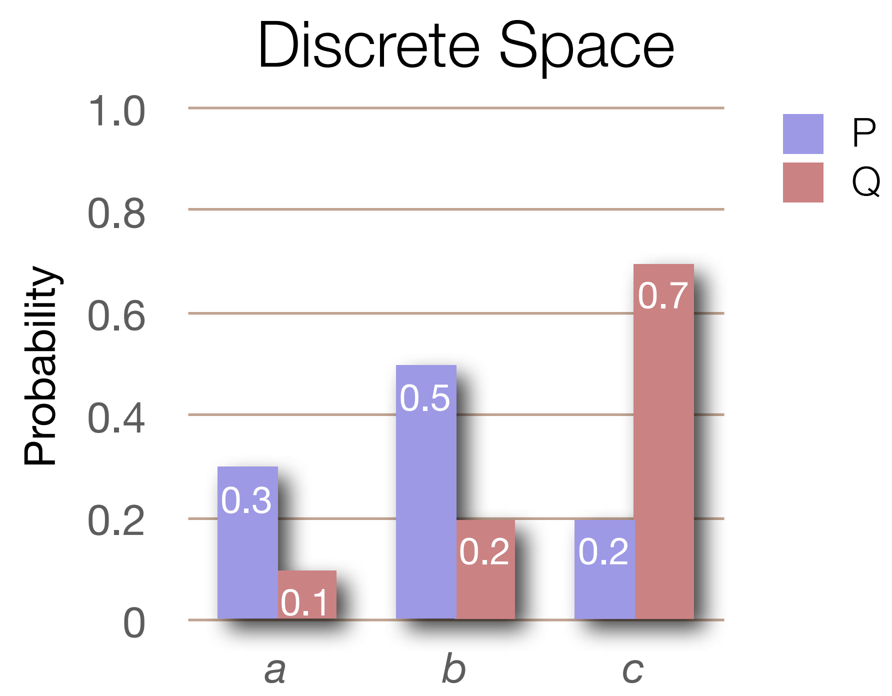
- A **binary experiment** is a pair of distributions (*P*,*Q*) over the same space $\mathcal{X}$
- We will think of P as the *positive* and Q as the negative *distribution*

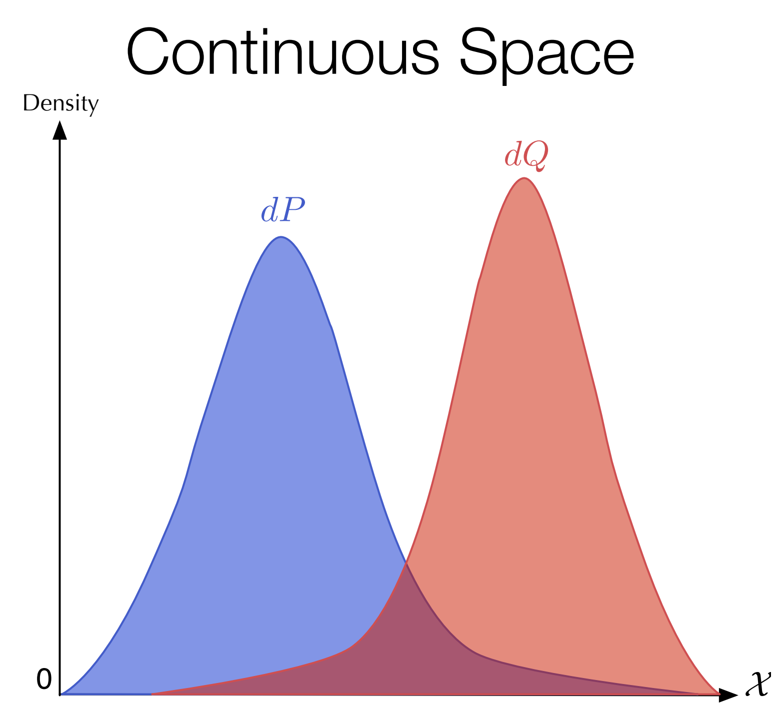




Binary Experiments

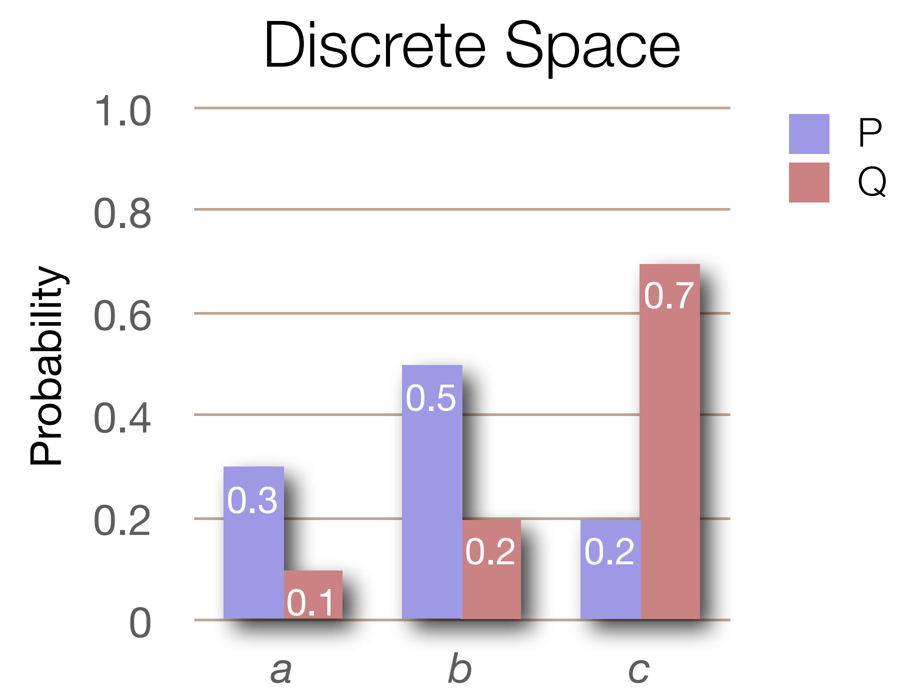
- A **binary experiment** is a pair of distributions (P,Q) over the same space $\mathcal{X}$
- We will think of P as the *positive* and Q as the negative *distribution*
- Given samples from $\mathcal{X}$, how can we tell if they came from P or Q?
    - Hypothesis Testing

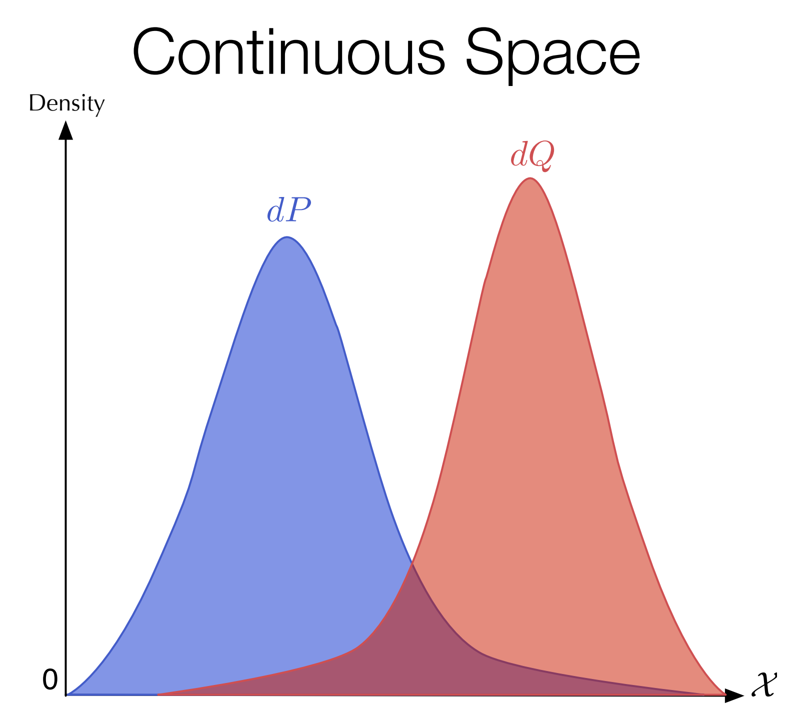




Binary Experiments

- A **binary experiment** is a pair of distributions (P,Q) over the same space $\mathcal{X}$
- We will think of P as the *positive* and Q as the negative *distribution*
- Given samples from $\mathcal{X}$, how can we tell if they came from P or Q?
  - Hypothesis Testing
- The "further apart" P and Q are the easier this will be
  - How do we define distance for distributions?



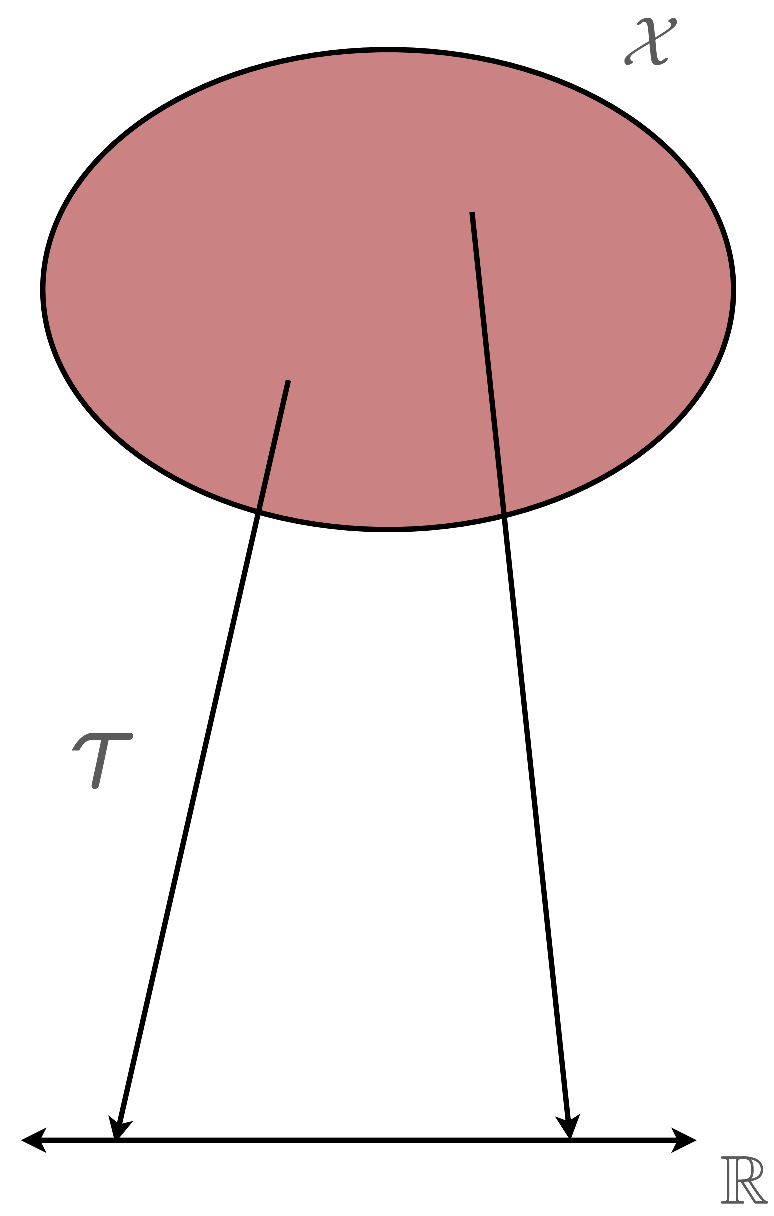


Test Statistics

- We would like our distances to not be dependent on the topology of the underlying space

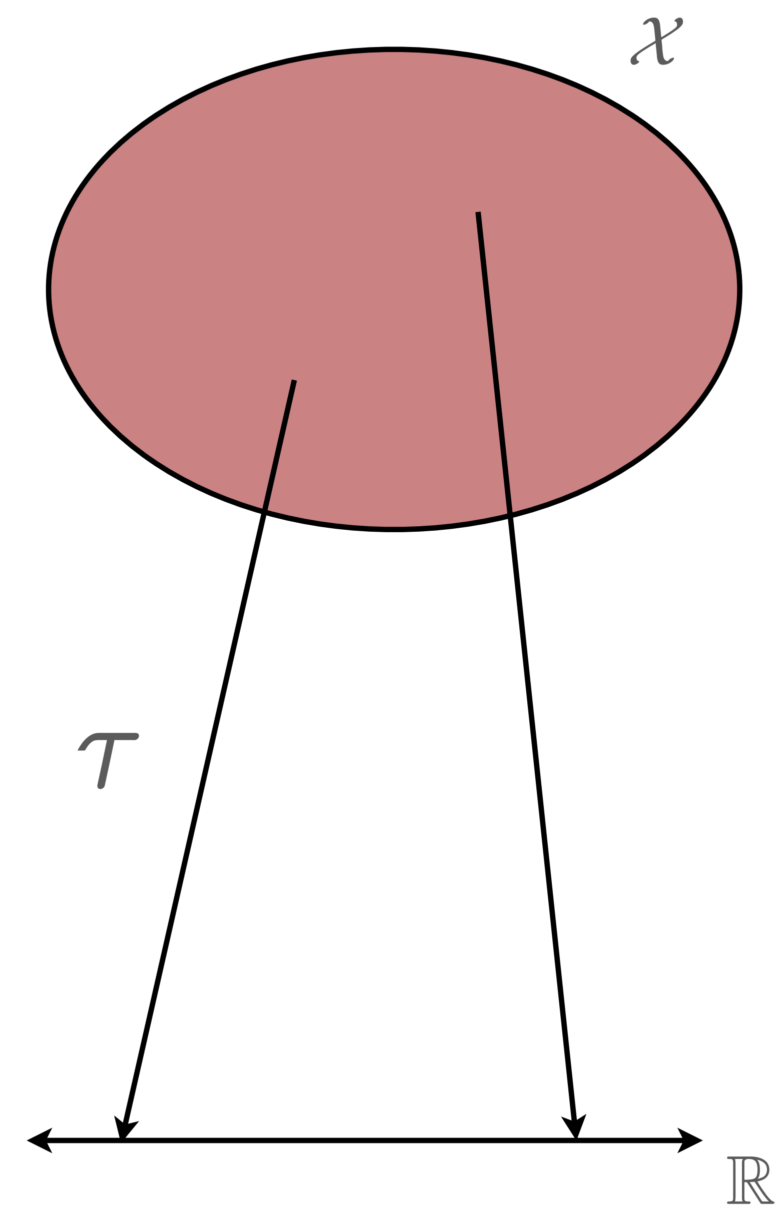
τ

$\mathcal{X}$



Test Statistics

- We would like our distances to not be dependent on the topology of the underlying space
- A **test statistic** τ maps each point in $\mathcal{X}$ to a point on the real line
  - Usually a function of the distribution

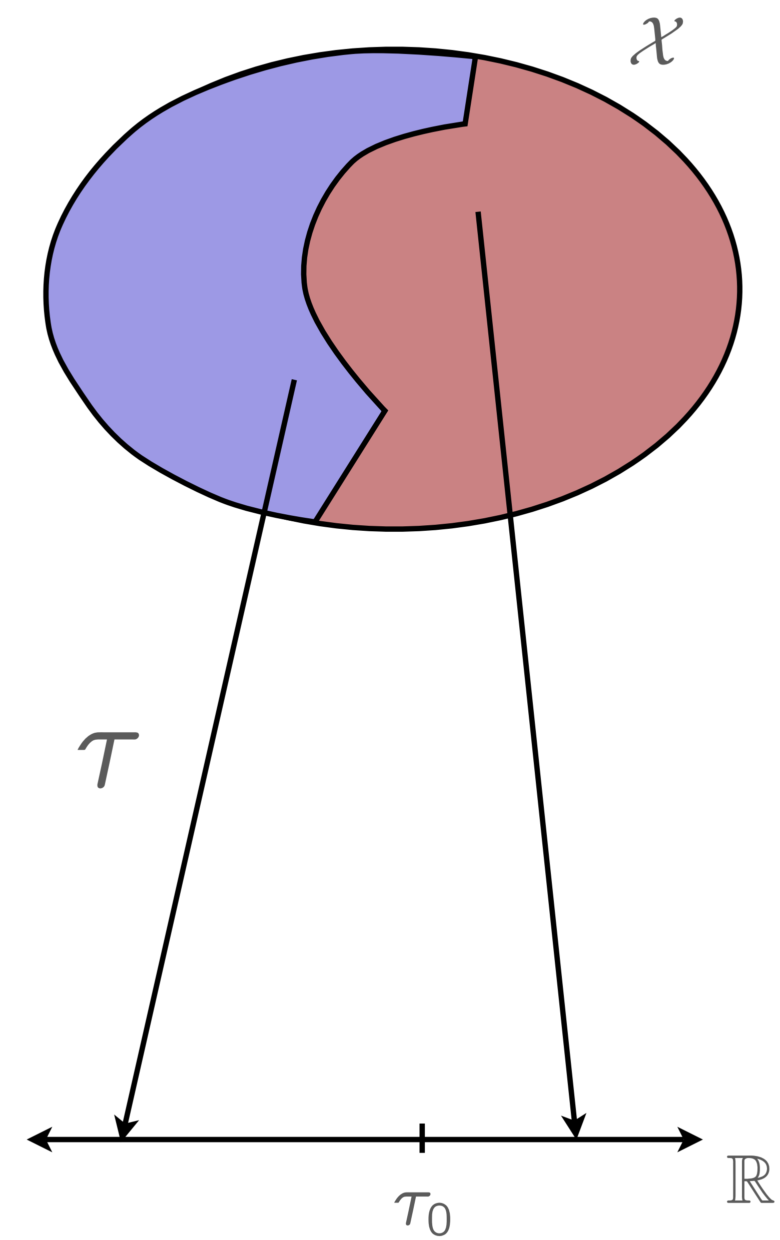


Test Statistics

- We would like our distances to not be dependent on the topology of the underlying space
- A **test statistic** τ maps each point in $\mathcal{X}$ to a point on the real line
  - Usually a function of the distribution
- A **statistical test** can be obtained by thresholding a test statistic

$$r(x) = [\![\tau(x) \geq \tau_0]\!]$$

- Each threshold partitions space into positive and negative parts



Statistical Power and Size

Contingency Table

- True Positive Rate $P(\tau \geq \tau_0)$ = "Power"
- False Positive Rate $Q(\tau \geq \tau_0)$ = "Size"
- True Negative Rate $Q(\tau < \tau_0)$
- False Negative Rate $P(\tau < \tau_0)$

		Actual Class	
		+	−
Predicted Class	+	True Positives **TP**	False Positives **FP**
	−	False Negatives **FN**	True Negatives **TN**

The Neyman-Pearson Lemma

Likelihood ratio

$$\tau^*(x) = \frac{dP}{dQ}(x)$$

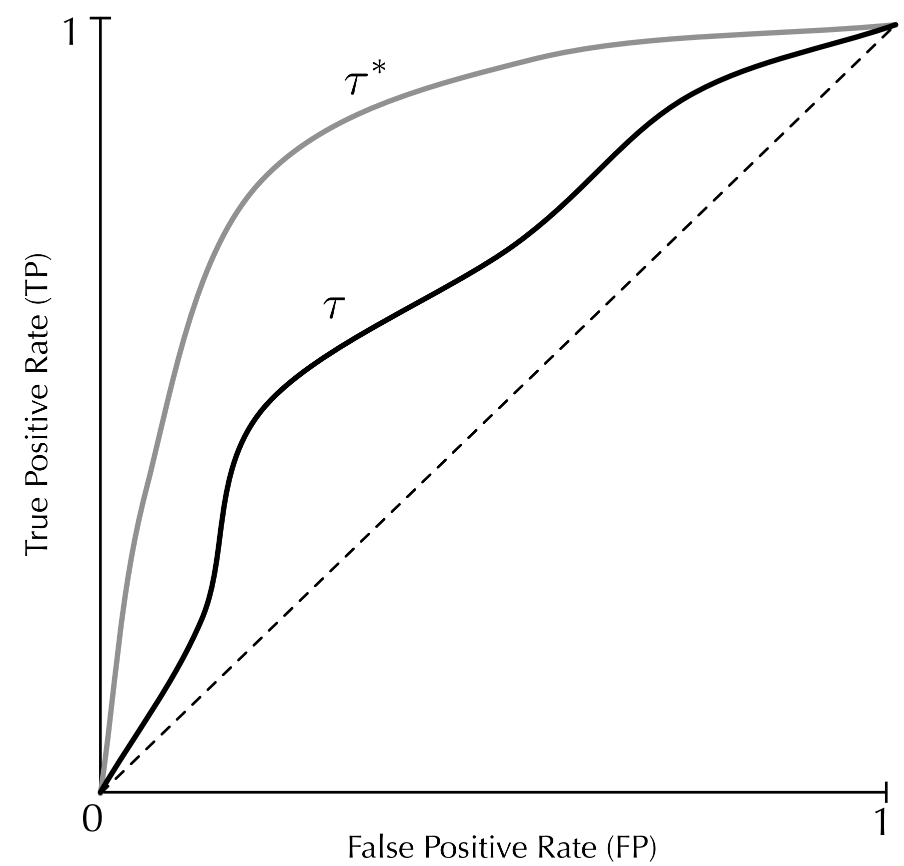
The Neyman-Pearson Lemma

Likelihood ratio

$$\tau^*(x) = \frac{dP}{dQ}(x)$$

Neyman-Pearson Lemma (1933)

- The the likelihood ratio is the **uniformly most powerful** (UMP) statistical test
  - Always has the largest TP Rate for any given FP rate



Csiszár f-Divergence

- **f-divergence of P from Q** is the Q-average of the likelihood ratio transformed by the function *f*
  - f can be seen as a penalty for dP(x) ≠ dQ(x)

$$
\begin{aligned}
\mathbb{I}_f(P, Q) &= \mathbb{E}_Q\left[f\left(\tau^*\right)\right] \\
&= \int_{\mathcal{X}} f\left(\frac{dP}{dQ}\right) dQ
\end{aligned}
$$

Csiszár f-Divergence

- **f-divergence of P from Q** is the Q-average of the likelihood ratio transformed by the function *f*
    - f can be seen as a penalty for dP(x) ≠ dQ(x)

$$\mathbb{I}_f(P,Q) = \mathbb{E}_Q\left[f\left(\tau^*\right)\right] = \int_{\mathcal{X}} f\left(\frac{dP}{dQ}\right)\, dQ$$

- To be a divergence, we want
    - $\mathbb{I}_f(P,Q) \geq 0$ for all P, Q
    - $\mathbb{I}_f(Q,Q) = 0$ for all Q

$$\begin{aligned}\mathbb{I}_f(P,Q) &= \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] \\ &\geq f\left(\mathbb{E}_Q\left[\frac{dP}{dQ}\right]\right) \\ &= f(1)\end{aligned}$$

Csiszár f-Divergence

- **f-divergence of P from Q** is the Q-average of the likelihood ratio transformed by the function *f*
  - f can be seen as a penalty for dP(x) ≠ dQ(x)
- To be a divergence, we want
  - $\mathbb{I}_f(P,Q) \geq 0$ for all P, Q
  - $\mathbb{I}_f(Q,Q) = 0$ for all Q
- Jensen's inequality requries
  - f convex
  - f(1) = 0

$$\begin{aligned} \mathbb{I}_f(P,Q) &= \mathbb{E}_Q\left[f\left(\tau^*\right)\right] \\ &= \int_{\mathcal{X}} f\left(\frac{dP}{dQ}\right) dQ \end{aligned}$$

$$\begin{aligned} \mathbb{I}_f(P,Q) &= \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] \\ &\geq f\left(\mathbb{E}_Q\left[\frac{dP}{dQ}\right]\right) \\ &= f(1) \end{aligned}$$

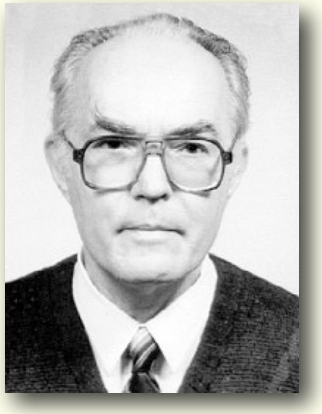
$$\mathbb{I}_f(P,Q) = \mathbb{J}_Q\left[f\left(\frac{dP}{dQ}\right)\right] \geq 0$$

"Jensen Gap"

Csiszár f-Divergence

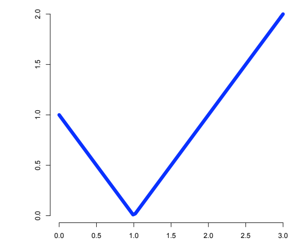
$$\begin{aligned}\mathbb{I}_f(P,Q) &= \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] - f\left(\mathbb{E}_Q\left[\frac{dP}{dQ}\right]\right)\\ &= \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right]\end{aligned}$$

A Jensen Gap where $f(1) = 0$

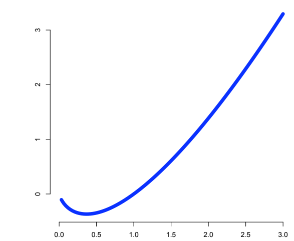


Examples

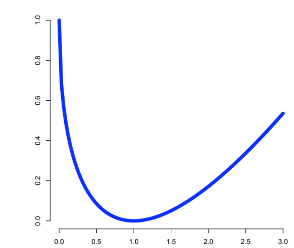
- **Variational** $f(t) = |t-1|$



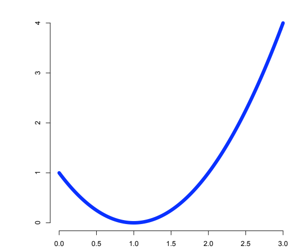
- **KL-Divergence** $f(t) = t \ln t$



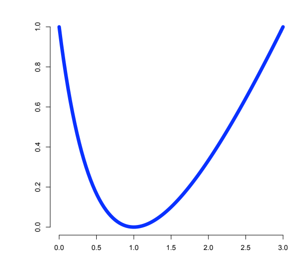
- **Hellinger** $f(t) = (\sqrt{t} - 1)^2$



- **Pearson** χ^2 $f(t) = (t-1)^2$



- **Triangular** $f(t) = \dfrac{(t-1)^2}{t+1}$

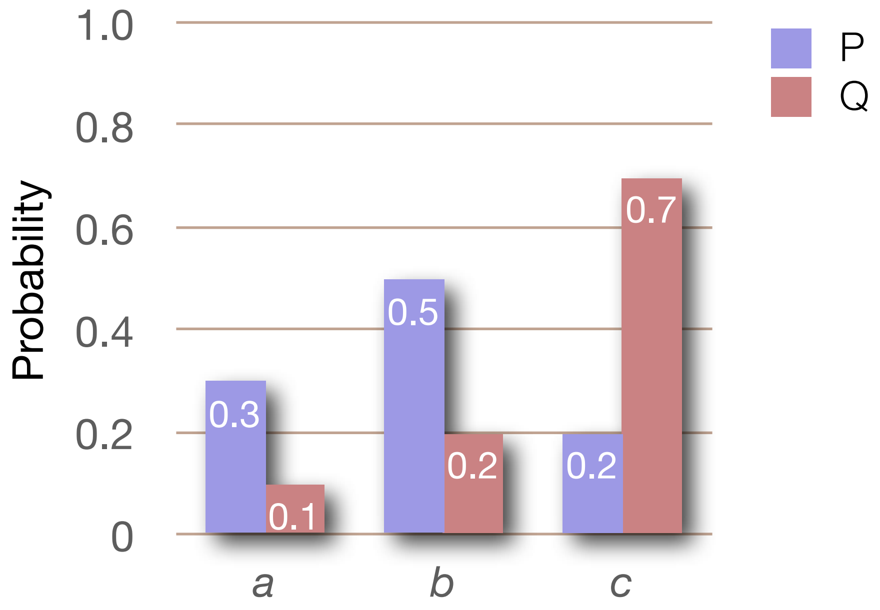


Examples

Variational Divergence

$$\sum_{x \in \{a,b,c\}} \left| \frac{P(x)}{Q(x)} - 1 \right| Q(x)$$

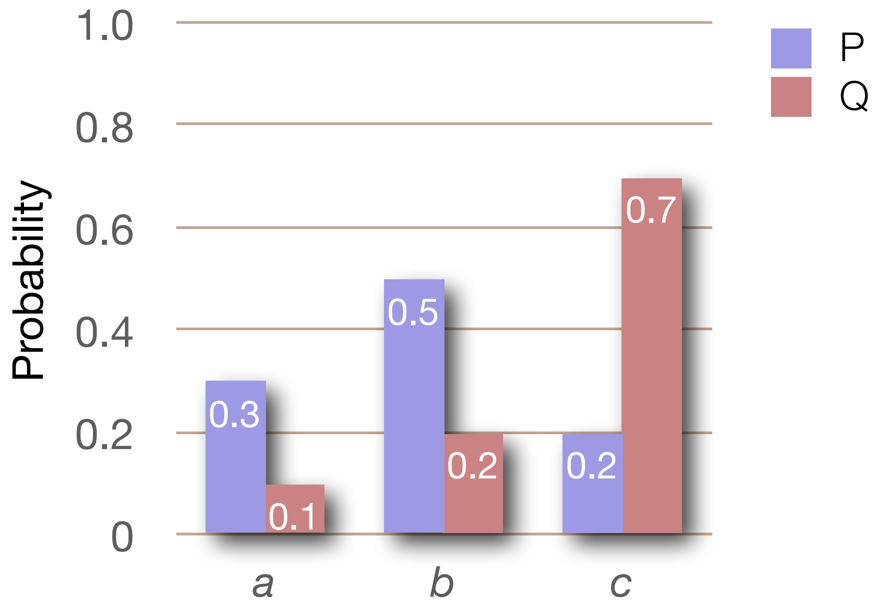
$$\begin{aligned}
&= |.3 - .1| + |.5 - .2| + |.2 - .7| \\
&= .2 + .3 + .5 \\
&= 1
\end{aligned}$$



Examples

Variational Divergence

$$\sum_{x\in\{a,b,c\}} \left|\frac{P(x)}{Q(x)} - 1\right| Q(x)$$
$$= \quad |.3 - .1| + |.5 - .2| + |.2 - .7|$$
$$= \quad .2 + .3 + .5$$
$$= \quad 1$$



KL Divergence

$$\sum_{x\in\{a,b,c\}} \frac{P(x)}{Q(x)} \ln\left(\frac{P(x)}{Q(x)}\right) Q(x)$$
$$= \quad .3\ln(3) + .5\ln(2.5) + .2\ln(2/7)$$
$$\approx \quad .43$$

Examples

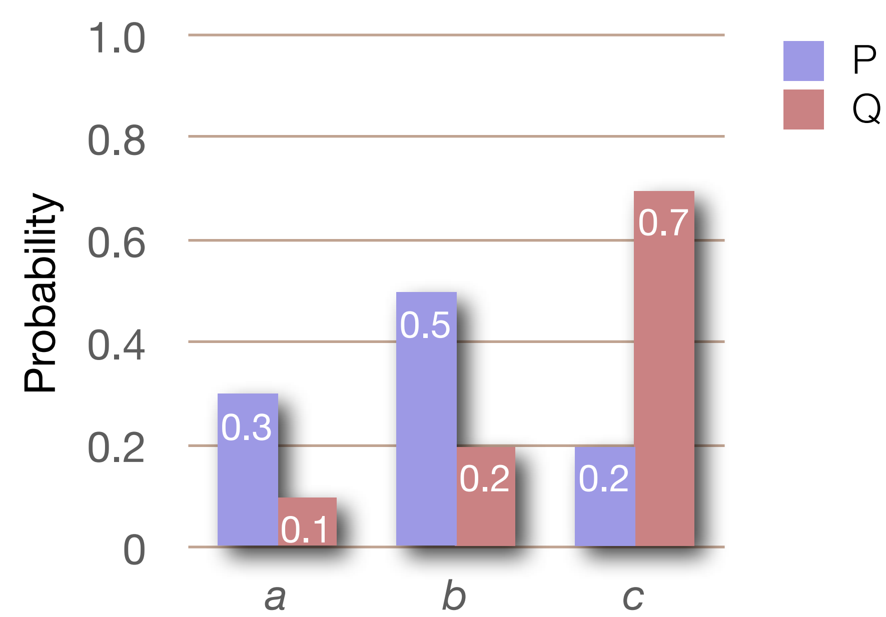
Variational Divergence

$$\sum_{x\in\{a,b,c\}} \left|\frac{P(x)}{Q(x)} - 1\right| Q(x)$$

$$= \quad |.3 - .1| + |.5 - .2| + |.2 - .7|$$

$$= \quad .2 + .3 + .5$$

$$= \quad 1$$

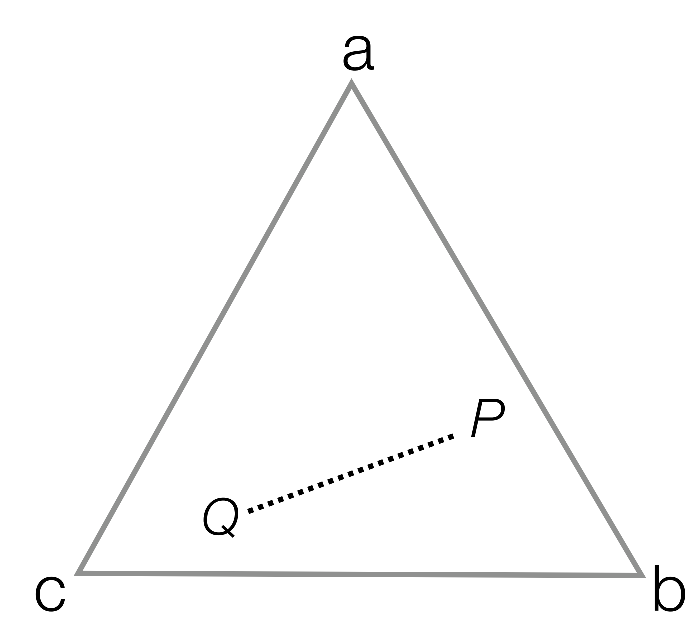


KL Divergence

$$\sum_{x\in\{a,b,c\}} \frac{P(x)}{Q(x)} \ln\left(\frac{P(x)}{Q(x)}\right) Q(x)$$

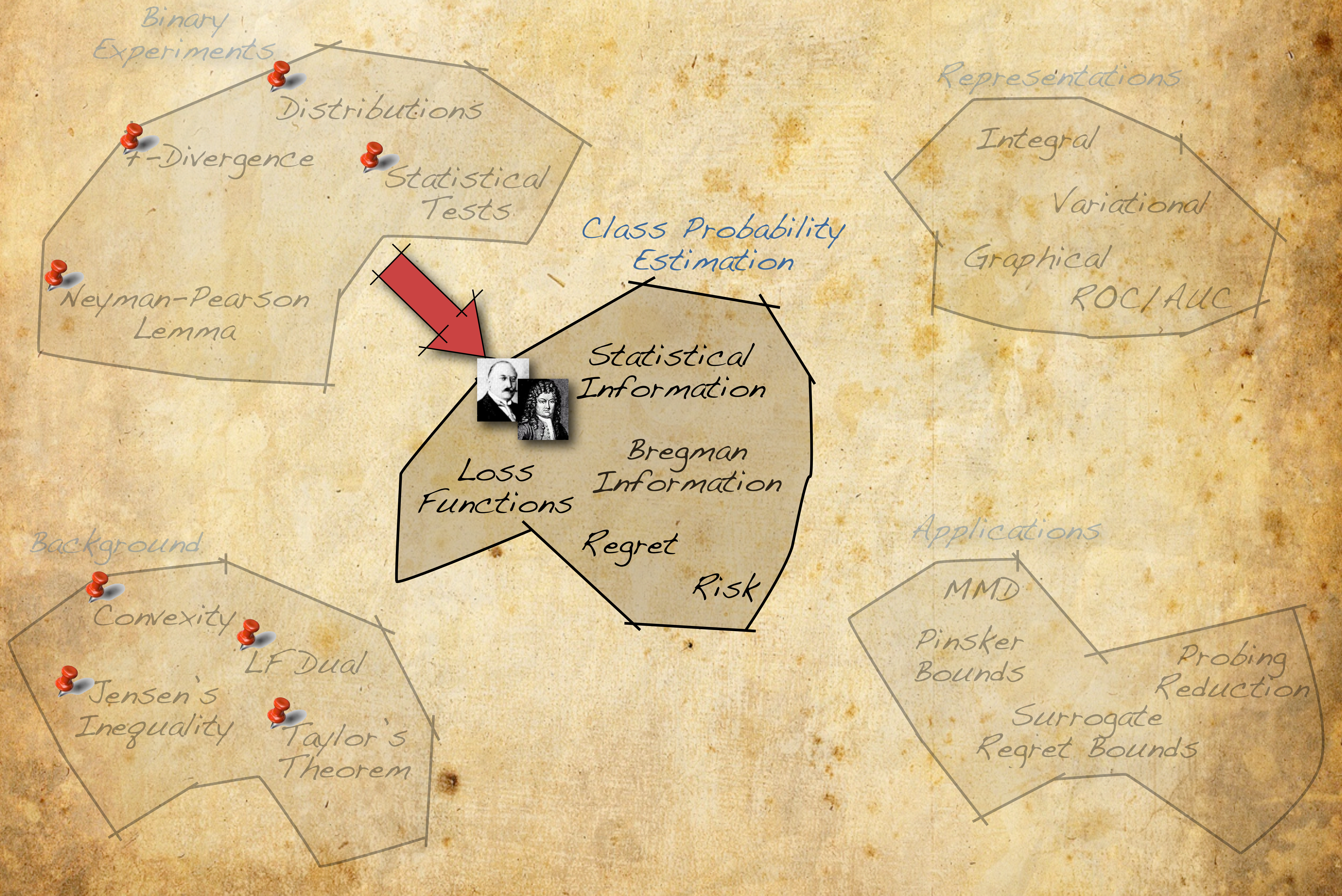
$$= \quad .3\ln(3) + .5\ln(2.5) + .2\ln(2/7)$$

$$\approx \quad .43$$



Terra Statistica

Terra Statistica



Classification and Probability Estimation

From Hypothesis Testing to Classification

Hypothesis Testing

- Instances are either drawn from P or Q exclusively
  - The aim is to correctly decide which
- Assumed
  - Binary Experiment (P,Q)
- Imposed
  - Measure of divergence

From Hypothesis Testing to Classification

Hypothesis Testing

- Instances are either drawn from P or Q exclusively
  - The aim is to correctly decide which
- Assumed
  - Binary Experiment (P,Q)
- Imposed
  - Measure of divergence

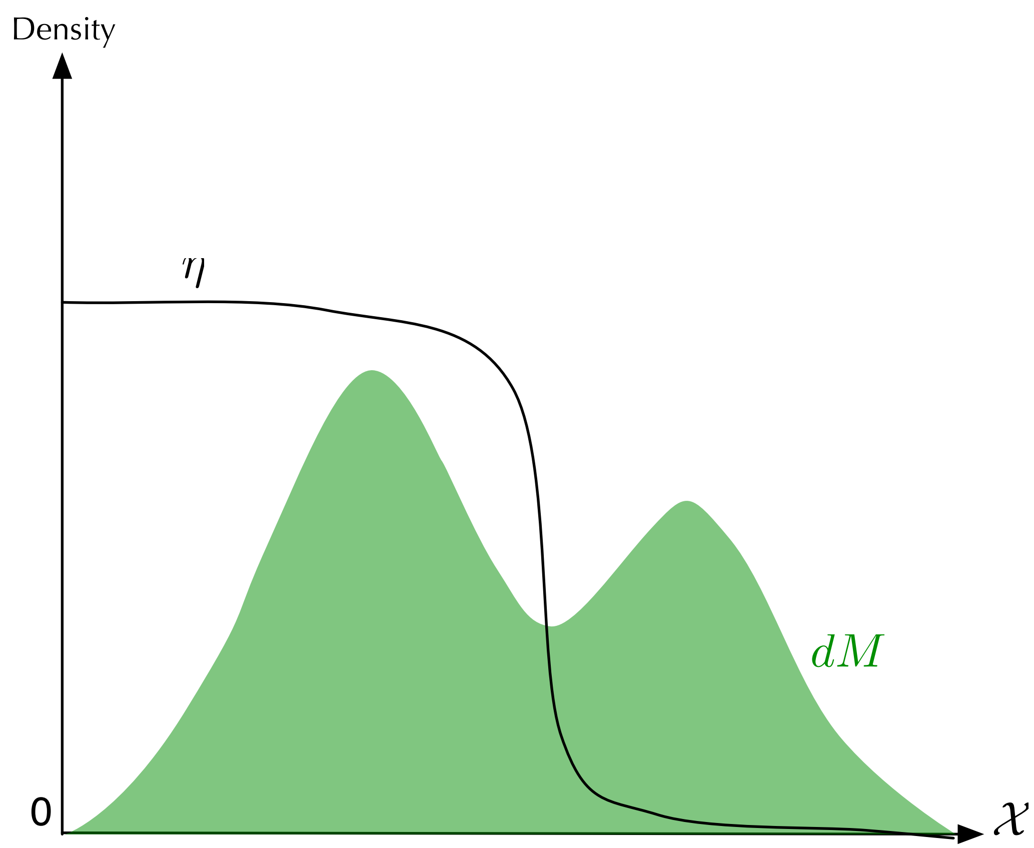
Classification / Prob. Estimation

- Instances are drawn from a **mixture** of P and Q
  - The aim is to correctly decide which **for each instance**
- Assumed
  - Binary Mixture (π,P,Q)
- Imposed
  - Misclassification penalty

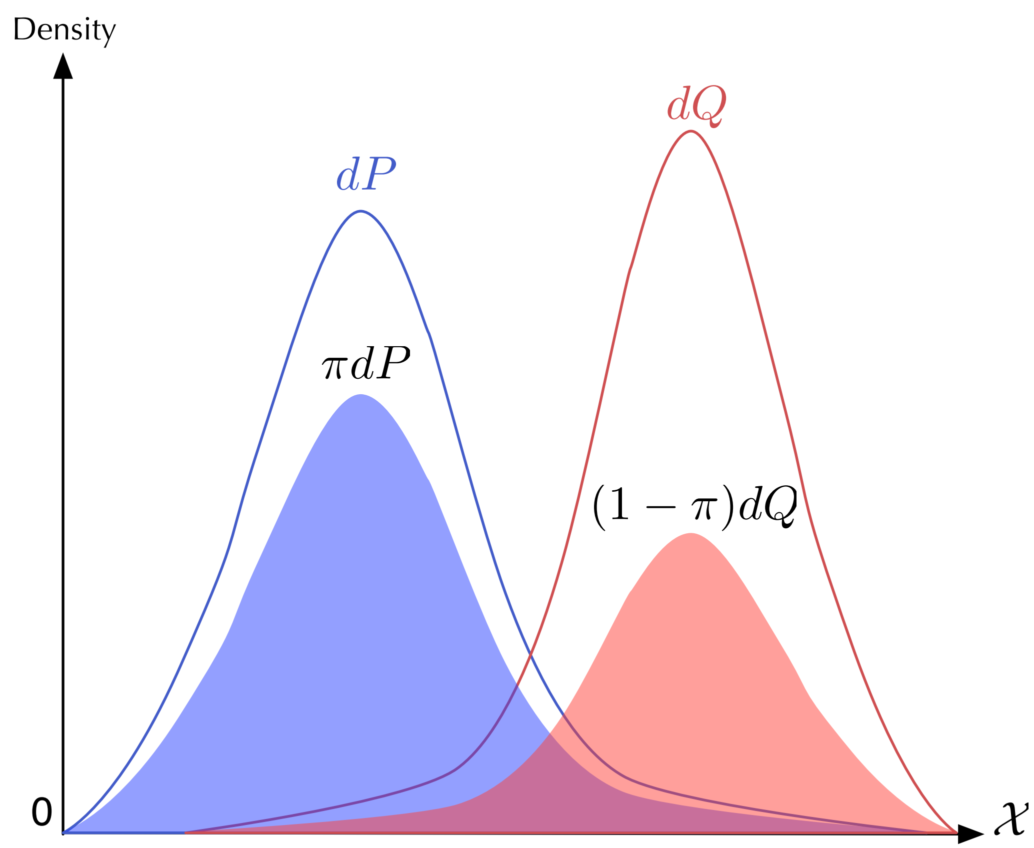
Generative and Discriminative Views

Joint Distribution

$$(\eta, M) \longleftarrow \mathbb{P}_{\mathcal{X}\times\mathcal{Y}} \longrightarrow (\pi, P, Q)$$



Discriminative



Generative

Generative and Discriminative Views

Joint Distribution

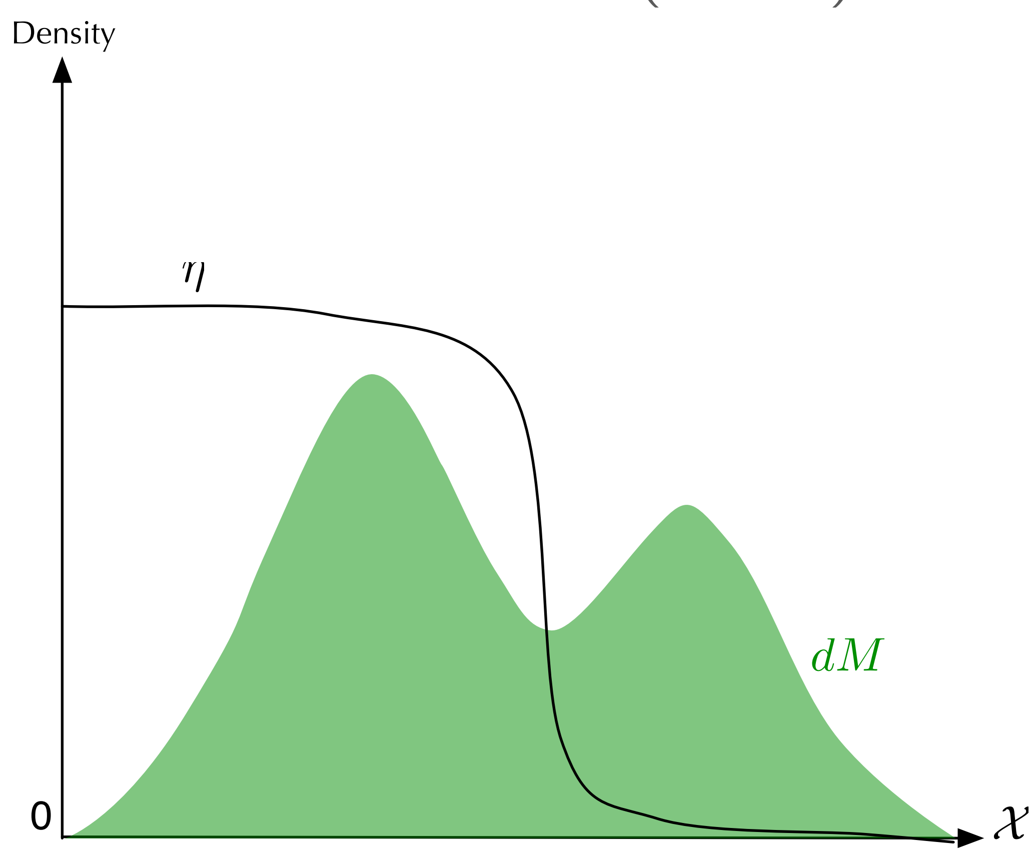
$$(\eta, M) \longleftarrow \mathbb{P}_{\mathcal{X}\times\mathcal{Y}} \longrightarrow (\pi, P, Q)$$

$$\eta = \pi \frac{dP}{dM}$$

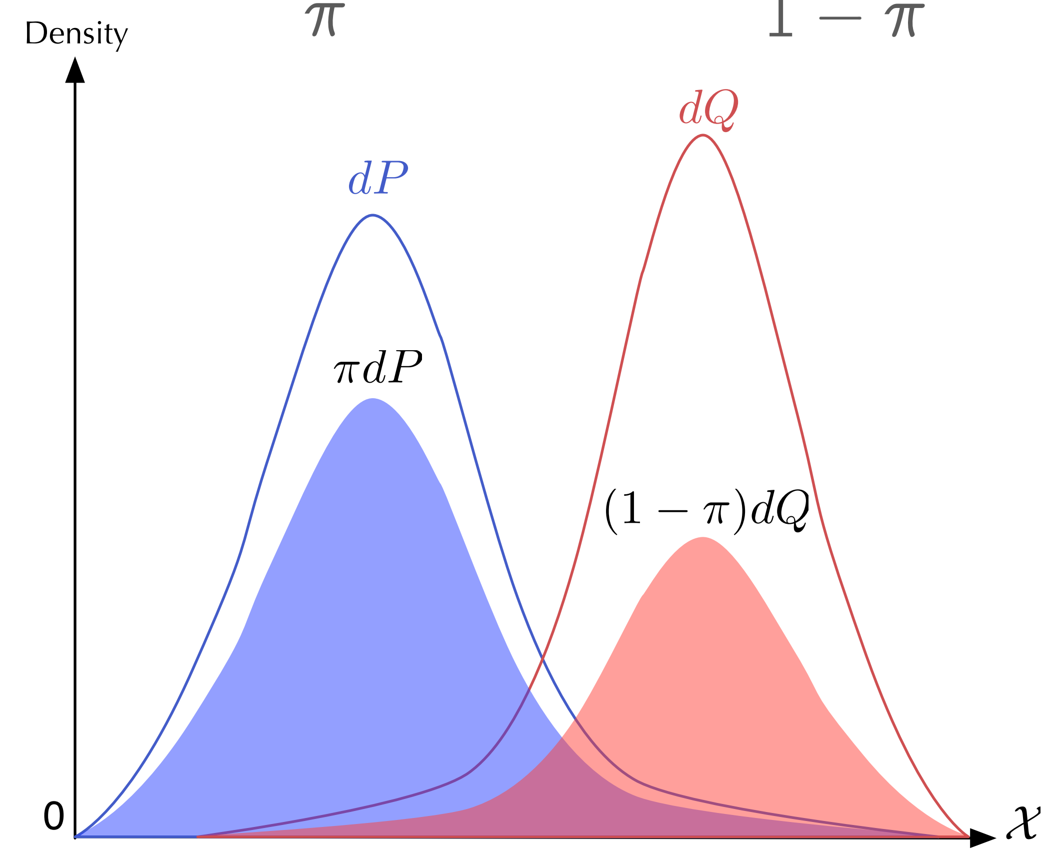
$$dM = \pi dP + (1-\pi) dQ$$

$$\pi = \mathbb{E}_M[\eta]$$

$$dP = \frac{\eta}{\pi} dM \qquad dQ = \frac{1-\eta}{1-\pi} dM$$



Discriminative



Generative

Generative and Discriminative Views

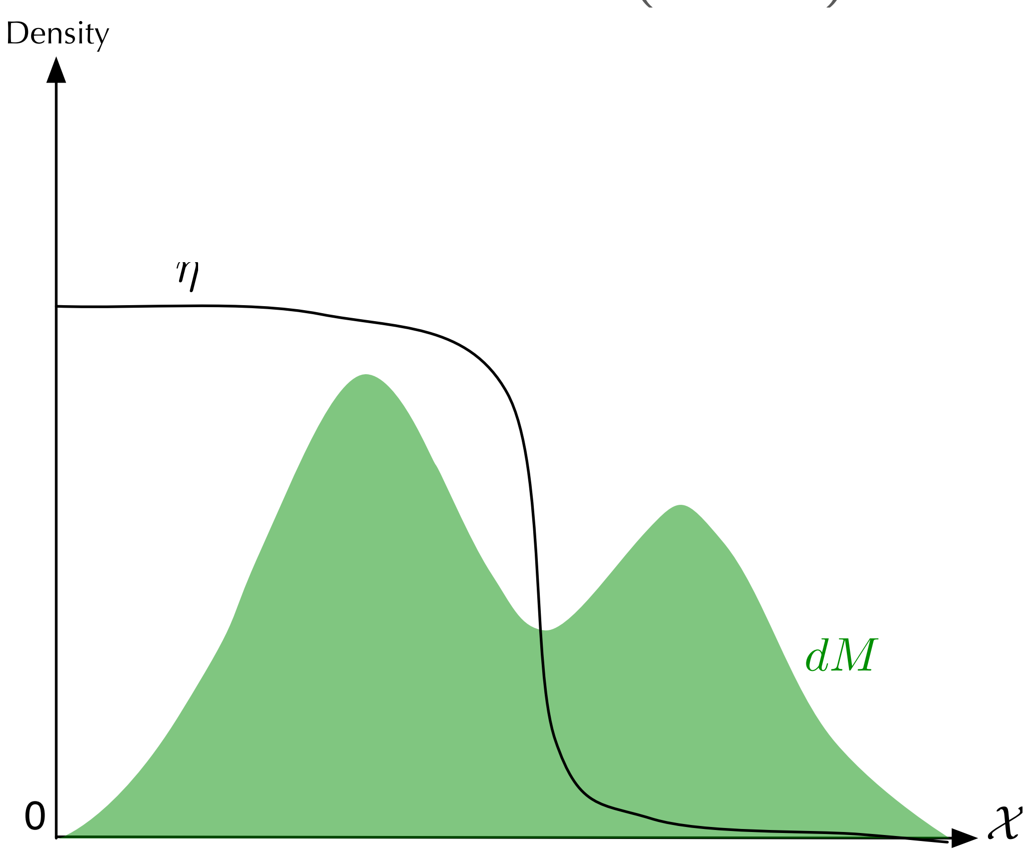
Joint Distribution

$$(\eta, M) \longleftarrow \mathbb{P}_{\mathcal{X}\times\mathcal{Y}} \longrightarrow (\pi, P, Q)$$

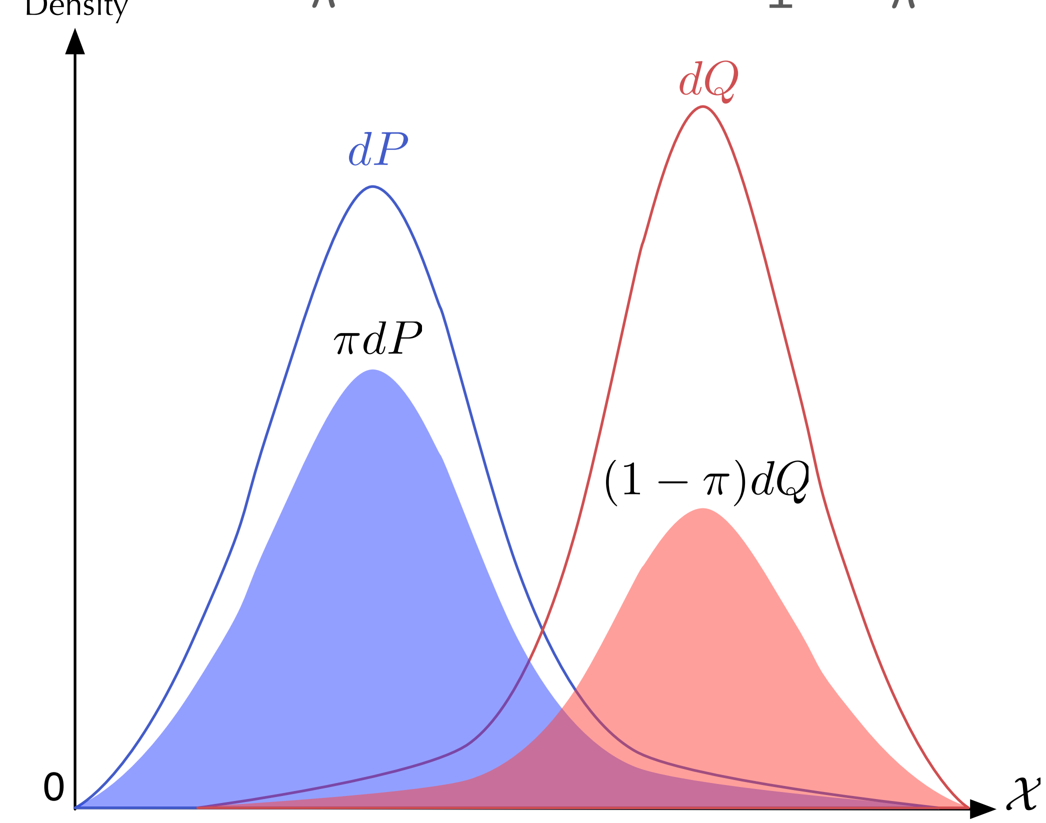
Bayes' Rule

$$\eta = \pi \frac{dP}{dM} \qquad \pi = \mathbb{E}_M[\eta]$$

$$dM = \pi dP + (1-\pi) dQ \qquad dP = \frac{\eta}{\pi} dM \qquad dQ = \frac{1-\eta}{1-\pi} dM$$



Discriminative



Generative

Loss, Risk and Regret

Loss

- Penalty $\ell(y, \hat{\eta})$ for guessing $\hat{\eta}$ when true class is *y*
  - Classification $\hat{\eta} \in \{0, 1\}$
  - Prob. Estimation $\hat{\eta} \in [0, 1]$

Loss, Risk and Regret

Loss

- Penalty $\ell(y, \hat{\eta})$ for guessing $\hat{\eta}$ when true class is *y*
  - Classification $\hat{\eta} \in \{0, 1\}$
  - Prob. Estimation $\hat{\eta} \in [0, 1]$

Point-wise Risk

- Expected point-wise loss

$$
\begin{aligned}
L &: [0,1] \times [0,1] \to \mathbb{R} \\
L(\eta, \hat{\eta}) &= \mathbb{E}_{Y \sim \eta}[\ell(Y, \hat{\eta})] \\
&= (1-\eta)\ell(0, \hat{\eta}) + \eta\ell(1, \hat{\eta})
\end{aligned}
$$

Loss, Risk and Regret

Loss

- Penalty $\ell(y, \hat{\eta})$ for guessing $\hat{\eta}$ when true class is *y*
  - Classification $\hat{\eta} \in \{0, 1\}$
  - Prob. Estimation $\hat{\eta} \in [0, 1]$

Point-wise Risk

- Expected point-wise loss

$$L : [0,1] \times [0,1] \to \mathbb{R}$$

$$\begin{aligned} L(\eta, \hat{\eta}) &= \mathbb{E}_{Y \sim \eta}[\ell(Y, \hat{\eta})] \\ &= (1-\eta)\ell(0, \hat{\eta}) + \eta\ell(1, \hat{\eta}) \end{aligned}$$

Risk

- Average point-wise risk

$$\mathbb{L} : [0,1]^{\mathcal{X}} \to \mathbb{R}$$

$$\mathbb{L}(\hat{\eta}) = \mathbb{E}_M[L(\eta, \hat{\eta})]$$

Loss, Risk and Regret

Loss

- Penalty $\ell(y, \hat{\eta})$ for guessing $\hat{\eta}$ when true class is *y*
  - Classification $\hat{\eta} \in \{0, 1\}$
  - Prob. Estimation $\hat{\eta} \in [0, 1]$

Point-wise Risk

- Expected point-wise loss

$$L : [0, 1] \times [0, 1] \to \mathbb{R}$$

$$\begin{aligned} L(\eta, \hat{\eta}) &= \mathbb{E}_{Y \sim \eta}[\ell(Y, \hat{\eta})] \\ &= (1 - \eta)\ell(0, \hat{\eta}) + \eta\ell(1, \hat{\eta}) \end{aligned}$$

Risk

- Average point-wise risk

$$\mathbb{L} : [0, 1]^{\mathcal{X}} \to \mathbb{R}$$

$$\mathbb{L}(\hat{\eta}) = \mathbb{E}_M[L(\eta, \hat{\eta})]$$

Bayes Risk

$$\underline{L}(\eta) = \inf_{\hat{\eta} \in [0,1]} L(\eta, \hat{\eta})$$

$$\underline{\mathbb{L}} = \inf_{\hat{\eta} \in [0,1]^{\mathcal{X}}} \mathbb{L}(\hat{\eta})$$

Loss, Risk and Regret

Loss

- Penalty $\ell(y, \hat{\eta})$ for guessing $\hat{\eta}$ when true class is *y*
  - Classification $\hat{\eta} \in \{0, 1\}$
  - Prob. Estimation $\hat{\eta} \in [0, 1]$

Point-wise Risk

- Expected point-wise loss

$$L : [0,1] \times [0,1] \to \mathbb{R}$$

$$\begin{aligned} L(\eta, \hat{\eta}) &= \mathbb{E}_{Y \sim \eta}[\ell(Y, \hat{\eta})] \\ &= (1-\eta)\ell(0, \hat{\eta}) + \eta\ell(1, \hat{\eta}) \end{aligned}$$

Risk

- Average point-wise risk

$$\mathbb{L} : [0,1]^{\mathcal{X}} \to \mathbb{R}$$

$$\mathbb{L}(\hat{\eta}) = \mathbb{E}_M[L(\eta, \hat{\eta})]$$

Bayes Risk

$$\underline{L}(\eta) = \inf_{\hat{\eta} \in [0,1]} L(\eta, \hat{\eta})$$

$$\underline{\mathbb{L}} = \inf_{\hat{\eta} \in [0,1]^{\mathcal{X}}} \mathbb{L}(\hat{\eta})$$

Regret

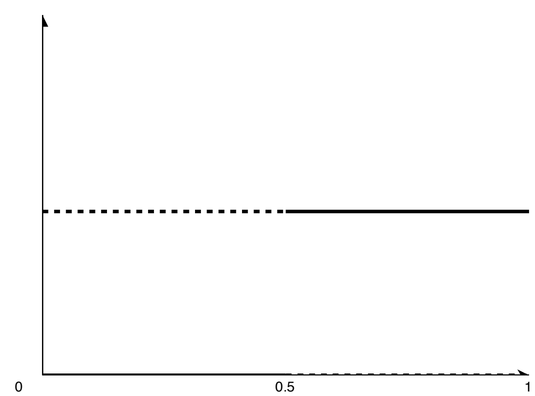
$$B(\eta, \hat{\eta}) = L(\eta, \hat{\eta}) - \underline{L}(\eta)$$

$$\mathbb{B}(\hat{\eta}) = \mathbb{L}(\hat{\eta}) - \underline{\mathbb{L}}$$

Loss Examples

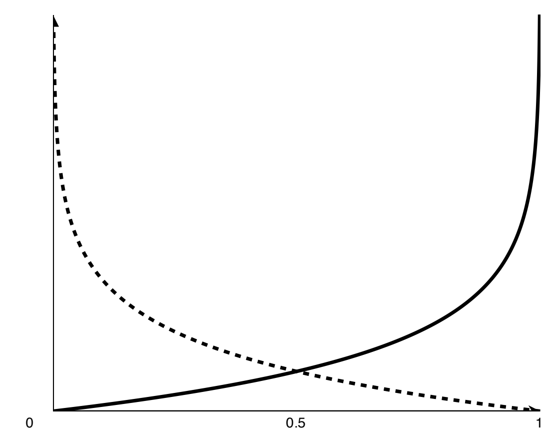
0-1 Misclassification Loss

$$\ell(y, \hat{\eta}) = [\![y \neq [\![\hat{\eta} > 0.5]\!]]\!]$$



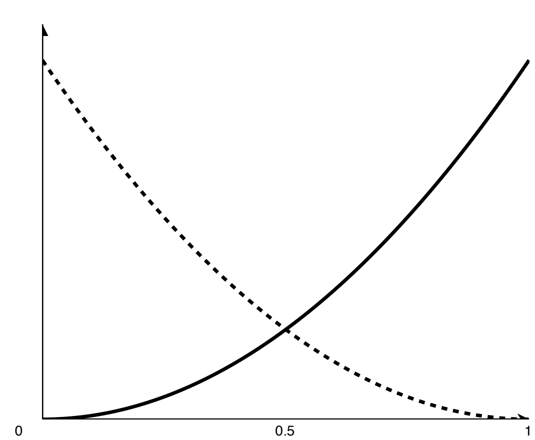
Log Loss

$$\ell(y, \hat{\eta}) = -y \log(\hat{\eta}) - (1 - y) \log(1 - \hat{\eta})$$



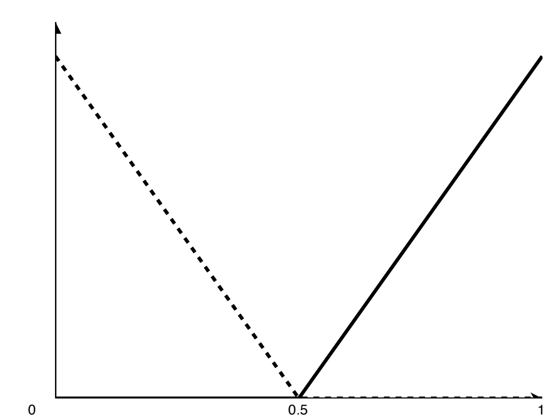
Square Loss

$$\ell(y, \hat{\eta}) = (y - \hat{\eta})^2$$



Hinge Loss

$$\ell(y, \hat{\eta}) = y(0.5 - \hat{\eta})_+ + (1 - y)(\hat{\eta} - 0.5)_+$$



Fisher Consistency & Proper Losses

Fisher Consistency

- Point-wise risk for a loss ℓ is minimised by true probability

$$L(\eta, \eta) = \inf_{\hat{\eta} \in [0,1]} L(\eta, \hat{\eta}) = \underline{L}(\eta)$$

- **Strict** consistency requires η to be the unique minimiser

Fisher Consistency & Proper Losses

Fisher Consistency

- Point-wise risk for a loss ℓ is minimised by true probability

$$L(\eta, \eta) = \inf_{\hat{\eta} \in [0,1]} L(\eta, \hat{\eta}) = \underline{L}(\eta)$$

- **Strict** consistency requires η to be the unique minimiser

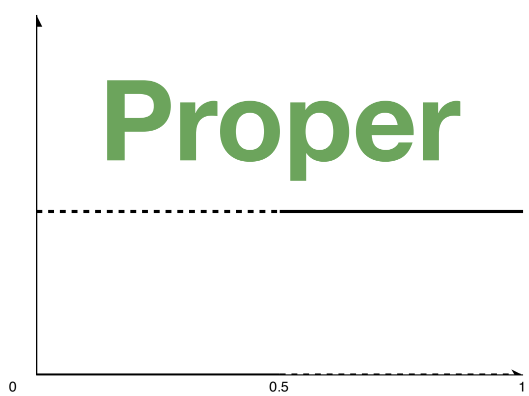
Proper Losses

- A loss ℓ is called (strictly) **proper** if it is (strictly) Fisher consistent
- In economics they are known as "proper scoring rules"
  - Shuford *et al.* (1966)
  - Savage (1971)
  - Schervish (1989)
  - Buja *et al.* (2005)
  - Lambert *et al.* (2008)

Examples of Proper Losses

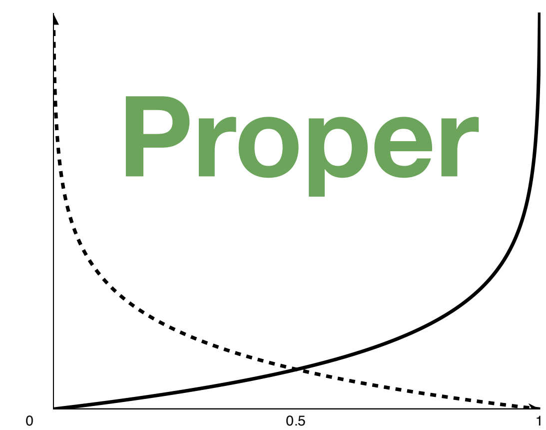
0-1 Misclassification Loss

$$\ell(y, \hat{\eta}) = [\![y \neq [\![\hat{\eta} > 0.5]\!]]\!]$$



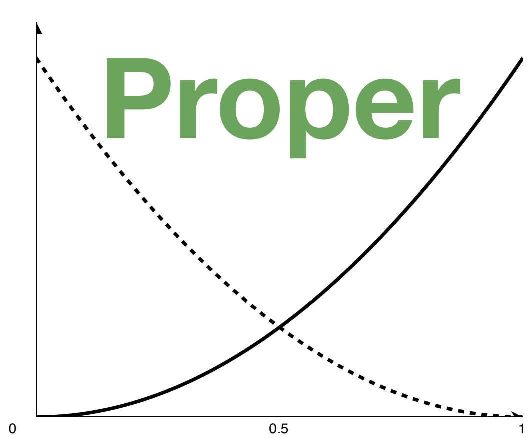
Log Loss

$$\ell(y, \hat{\eta}) = -y \log(\hat{\eta}) - (1-y)\log(1-\hat{\eta})$$



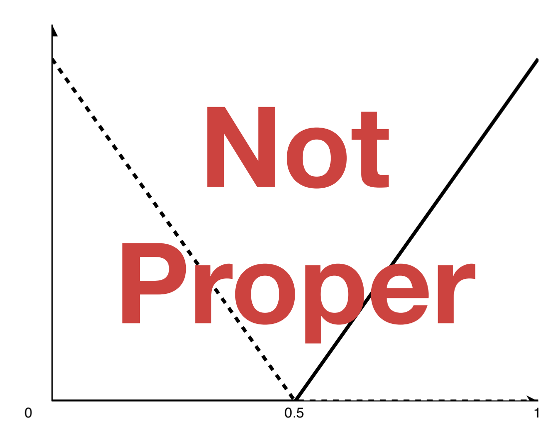
Square Loss

$$\ell(y, \hat{\eta}) = (y - \hat{\eta})^2$$



Hinge Loss

$$\ell(y, \hat{\eta}) = y(0.5 - \hat{\eta})_+ + (1-y)(\hat{\eta} - 0.5)_+$$

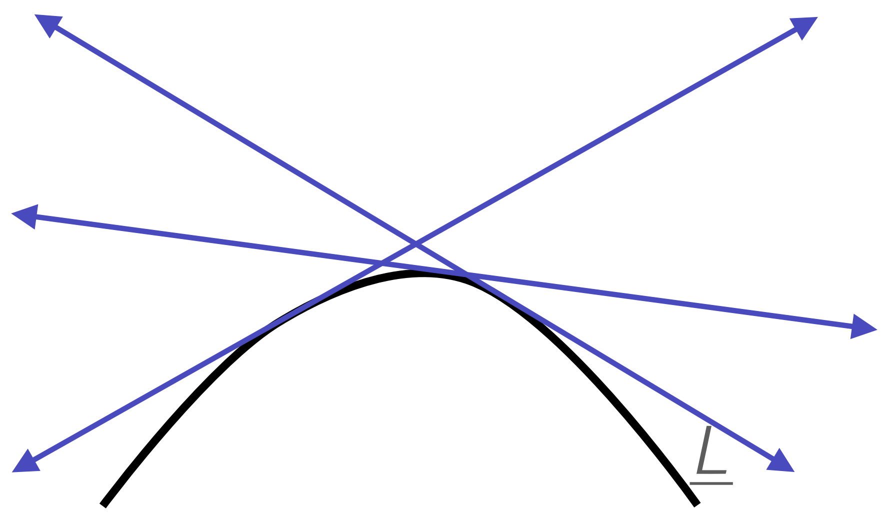


Properties of Proper Losses

Concave Bayes Risk

- Lower envelope of lines

$$\underline{L}(\eta) = \inf_{\hat{\eta}} (1-\eta)\ell(0,\hat{\eta}) + \eta\ell(1,\hat{\eta})$$

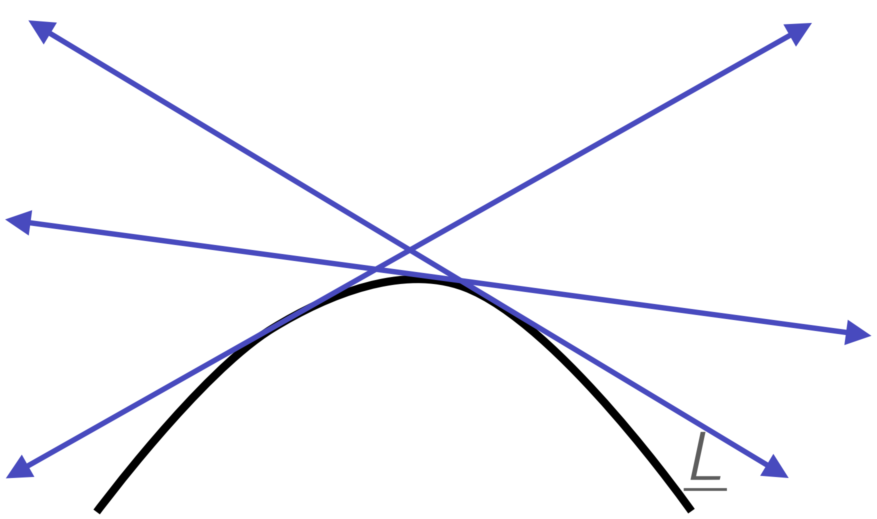


Properties of Proper Losses

Concave Bayes Risk

- Lower envelope of lines

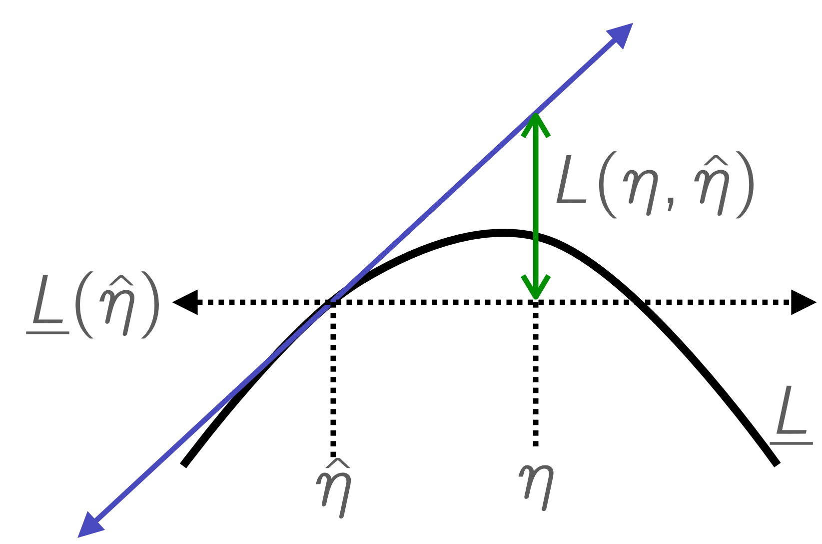
$$\underline{L}(\eta) = \inf_{\hat{\eta}} (1-\eta)\ell(0,\hat{\eta}) + \eta\ell(1,\hat{\eta})$$



Savage's Theorem

- Loss ℓ is proper **iff** its Bayes risk $\underline{L}$ is concave
- Relates Bayes risk and risk without optimisation

$$\begin{aligned} L(\eta,\hat{\eta}) &= \underline{L}(\hat{\eta}) - (\hat{\eta}-\eta)\underline{L}'(\hat{\eta}) \\ &= \underline{L}(\hat{\eta}) + (\eta-\hat{\eta})\underline{L}'(\hat{\eta}) \end{aligned}$$



Savage's Theorem

A loss is proper

if and only if

its point-wise Bayes risk is concave

Furthermore

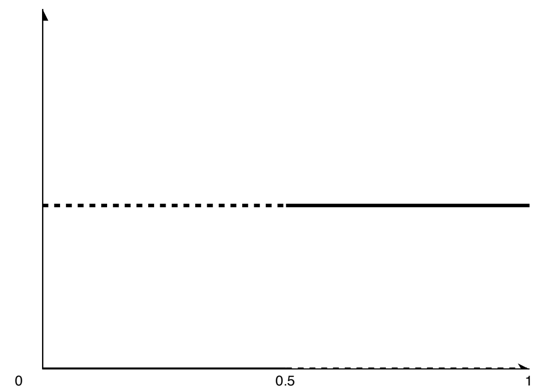
$$L(\eta, \hat{\eta}) = \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta})$$

[Savage, 1971]

Examples

0-1 Misclassification Loss

$$\ell(y, \hat{\eta}) = [\![y \neq [\![\hat{\eta} > 0.5]\!]]\!]$$

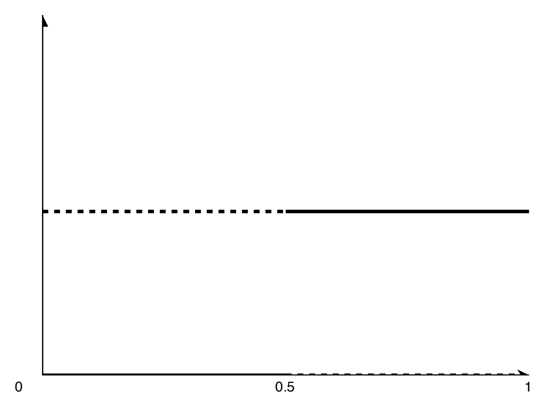


Log Loss

Examples

0-1 Misclassification Loss

$$\ell(y, \hat{\eta}) = [\![y \neq [\![\hat{\eta} > 0.5]\!]]\!]$$



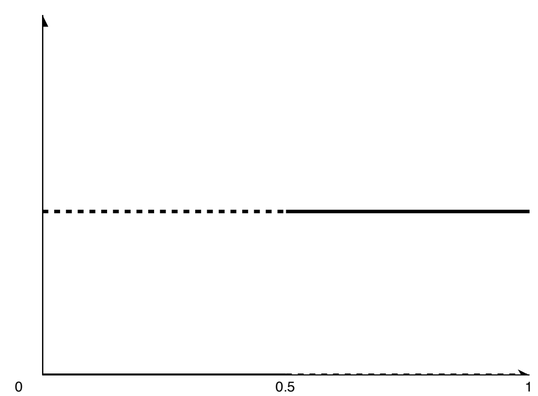
$$L(\eta, \hat{\eta}) = \begin{cases} (1-\eta) & \hat{\eta} > .5 \\ \eta & \hat{\eta} \leq .5 \end{cases}$$

Log Loss

Examples

0-1 Misclassification Loss

$$\ell(y, \hat{\eta}) = [\![y \neq [\![\hat{\eta} > 0.5]\!]]\!]$$



$$L(\eta, \hat{\eta}) = \begin{cases} (1-\eta) & \hat{\eta} > .5 \\ \eta & \hat{\eta} \leq .5 \end{cases}$$

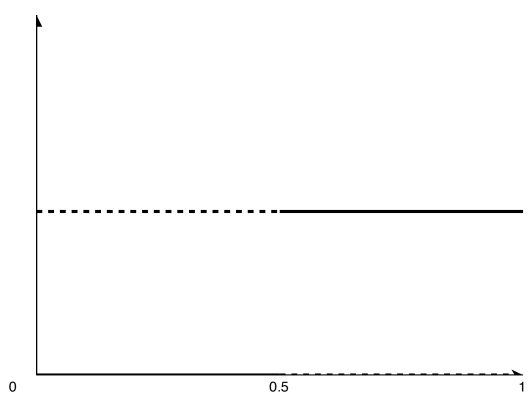
$$\underline{L}(\eta) = L(\eta, \eta) = \begin{cases} (1-\eta) & \eta > .5 \\ \eta & \eta \leq .5 \end{cases}$$

Log Loss

Examples

0-1 Misclassification Loss

$$\ell(y, \hat{\eta}) = [\![y \neq [\![\hat{\eta} > 0.5]\!]]\!]$$



$$L(\eta, \hat{\eta}) = \begin{cases} (1-\eta) & \hat{\eta} > .5 \\ \eta & \hat{\eta} \leq .5 \end{cases}$$

$$\underline{L}(\eta) = L(\eta, \eta) = \begin{cases} (1-\eta) & \eta > .5 \\ \eta & \eta \leq .5 \end{cases}$$

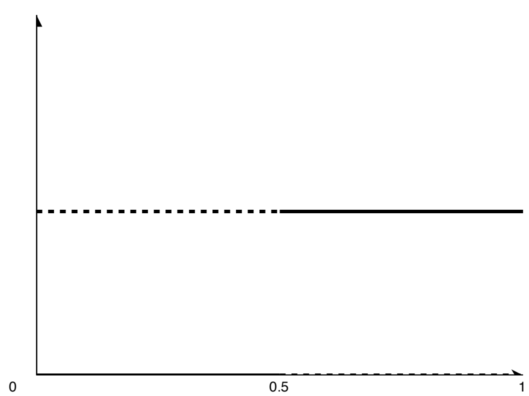
$$\underline{L}'(\eta) = \begin{cases} -1 & \eta > .5 \\ 1 & \eta \leq .5 \end{cases}$$

Log Loss

Examples

0-1 Misclassification Loss

$$\ell(y, \hat{\eta}) = [\![y \neq [\![\hat{\eta} > 0.5]\!]]\!]$$



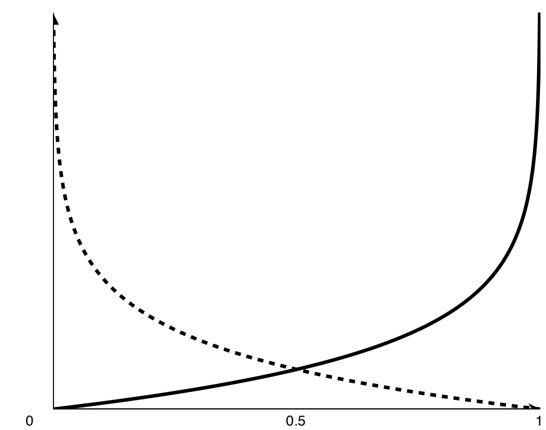
$$L(\eta, \hat{\eta}) = \begin{cases} (1-\eta) & \hat{\eta} > .5 \\ \eta & \hat{\eta} \leq .5 \end{cases}$$

$$\underline{L}(\eta) = L(\eta, \eta) = \begin{cases} (1-\eta) & \eta > .5 \\ \eta & \eta \leq .5 \end{cases}$$

$$\underline{L}'(\eta) = \begin{cases} -1 & \eta > .5 \\ 1 & \eta \leq .5 \end{cases}$$

Log Loss

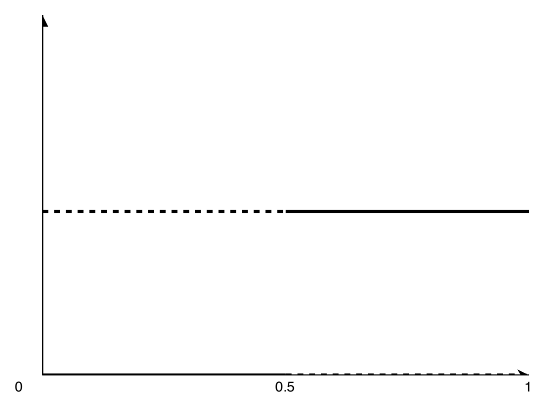
$$\ell(y, \hat{\eta}) = -y \log(\hat{\eta}) - (1-y) \log(1-\hat{\eta})$$



Examples

0-1 Misclassification Loss

$$\ell(y,\hat{\eta}) = [\![y \neq [\![\hat{\eta} > 0.5]\!]]\!]$$



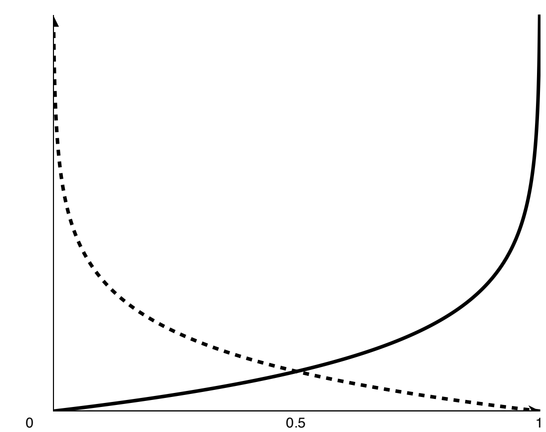
$$L(\eta,\hat{\eta}) = \begin{cases} (1-\eta) & \hat{\eta} > .5 \\ \eta & \hat{\eta} \leq .5 \end{cases}$$

$$\underline{L}(\eta) = L(\eta,\eta) = \begin{cases} (1-\eta) & \eta > .5 \\ \eta & \eta \leq .5 \end{cases}$$

$$\underline{L}'(\eta) = \begin{cases} -1 & \eta > .5 \\ 1 & \eta \leq .5 \end{cases}$$

Log Loss

$$\ell(y,\hat{\eta}) = -y\log(\hat{\eta}) - (1-y)\log(1-\hat{\eta})$$

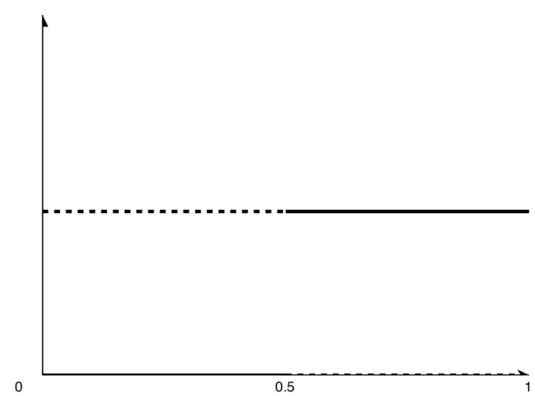


$$L(\eta,\hat{\eta}) = -\eta\log(\hat{\eta}) - (1-\eta)\log(1-\hat{\eta})$$

Examples

0-1 Misclassification Loss

$$\ell(y, \hat{\eta}) = [\![y \neq [\![\hat{\eta} > 0.5]\!]]\!]$$



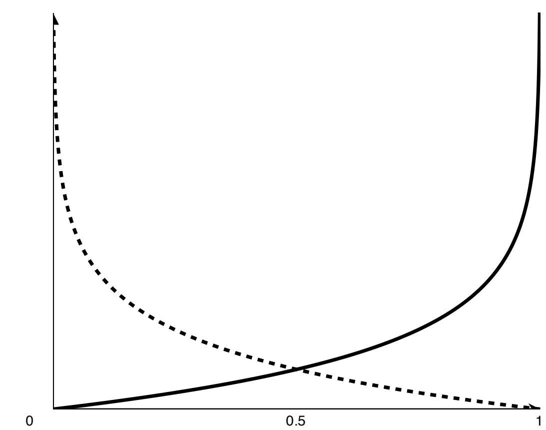
$$L(\eta, \hat{\eta}) = \begin{cases} (1-\eta) & \hat{\eta} > .5 \\ \eta & \hat{\eta} \leq .5 \end{cases}$$

$$\underline{L}(\eta) = L(\eta, \eta) = \begin{cases} (1-\eta) & \eta > .5 \\ \eta & \eta \leq .5 \end{cases}$$

$$\underline{L}'(\eta) = \begin{cases} -1 & \eta > .5 \\ 1 & \eta \leq .5 \end{cases}$$

Log Loss

$$\ell(y, \hat{\eta}) = -y \log(\hat{\eta}) - (1-y)\log(1-\hat{\eta})$$



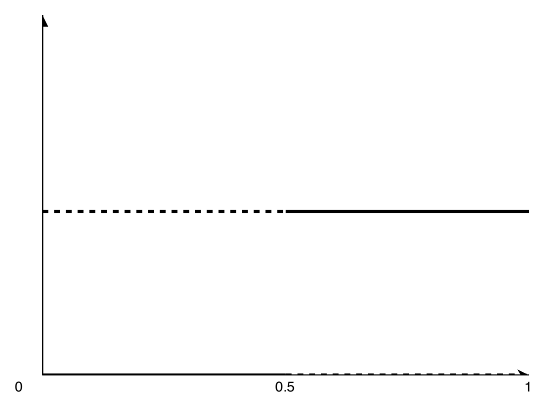
$$L(\eta, \hat{\eta}) = -\eta \log(\hat{\eta}) - (1-\eta)\log(1-\hat{\eta})$$

$$\underline{L}(\eta) = -\eta \log(\eta) - (1-\eta)\log(1-\eta)$$

Examples

0-1 Misclassification Loss

$$\ell(y, \hat{\eta}) = [\![y \neq [\![\hat{\eta} > 0.5]\!]]\!]$$



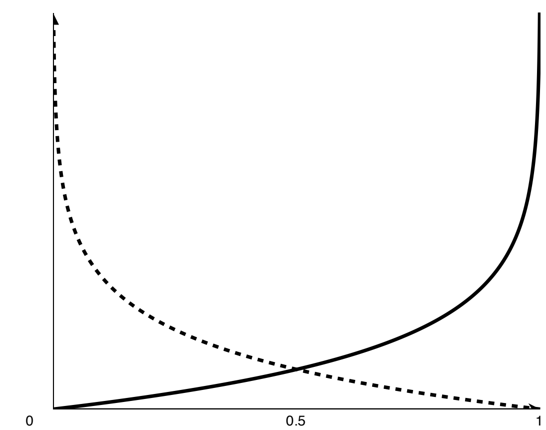
$$L(\eta, \hat{\eta}) = \begin{cases} (1-\eta) & \hat{\eta} > .5 \\ \eta & \hat{\eta} \leq .5 \end{cases}$$

$$\underline{L}(\eta) = L(\eta, \eta) = \begin{cases} (1-\eta) & \eta > .5 \\ \eta & \eta \leq .5 \end{cases}$$

$$\underline{L}'(\eta) = \begin{cases} -1 & \eta > .5 \\ 1 & \eta \leq .5 \end{cases}$$

Log Loss

$$\ell(y, \hat{\eta}) = -y \log(\hat{\eta}) - (1-y) \log(1-\hat{\eta})$$



$$L(\eta, \hat{\eta}) = -\eta \log(\hat{\eta}) - (1-\eta) \log(1-\hat{\eta})$$

$$\underline{L}(\eta) = -\eta \log(\eta) - (1-\eta) \log(1-\eta)$$

$$\begin{aligned} \underline{L}'(\eta) &= -1 - \log(\eta) + 1 + \log(1-\eta) \\ &= \log\left(\frac{1-\eta}{\eta}\right) \end{aligned}$$

Proper Point-wise Bayes Risks

Given a proper loss,
its point-wise Bayes risk
is easy to compute

$$\underline{L}(\eta) = L(\eta, \eta)$$

Information

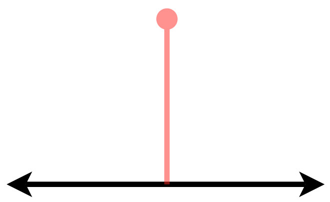
Where is the wisdom
we have lost in knowledge?

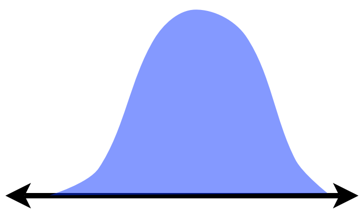
Where is the knowledge
we have lost in information?

T.S. Eliot (1988-1965)

Statistical Information

- Let U measure the "**uncertainty**" of a distribution ξ.
    - When ξ is peaked its uncertainty is small

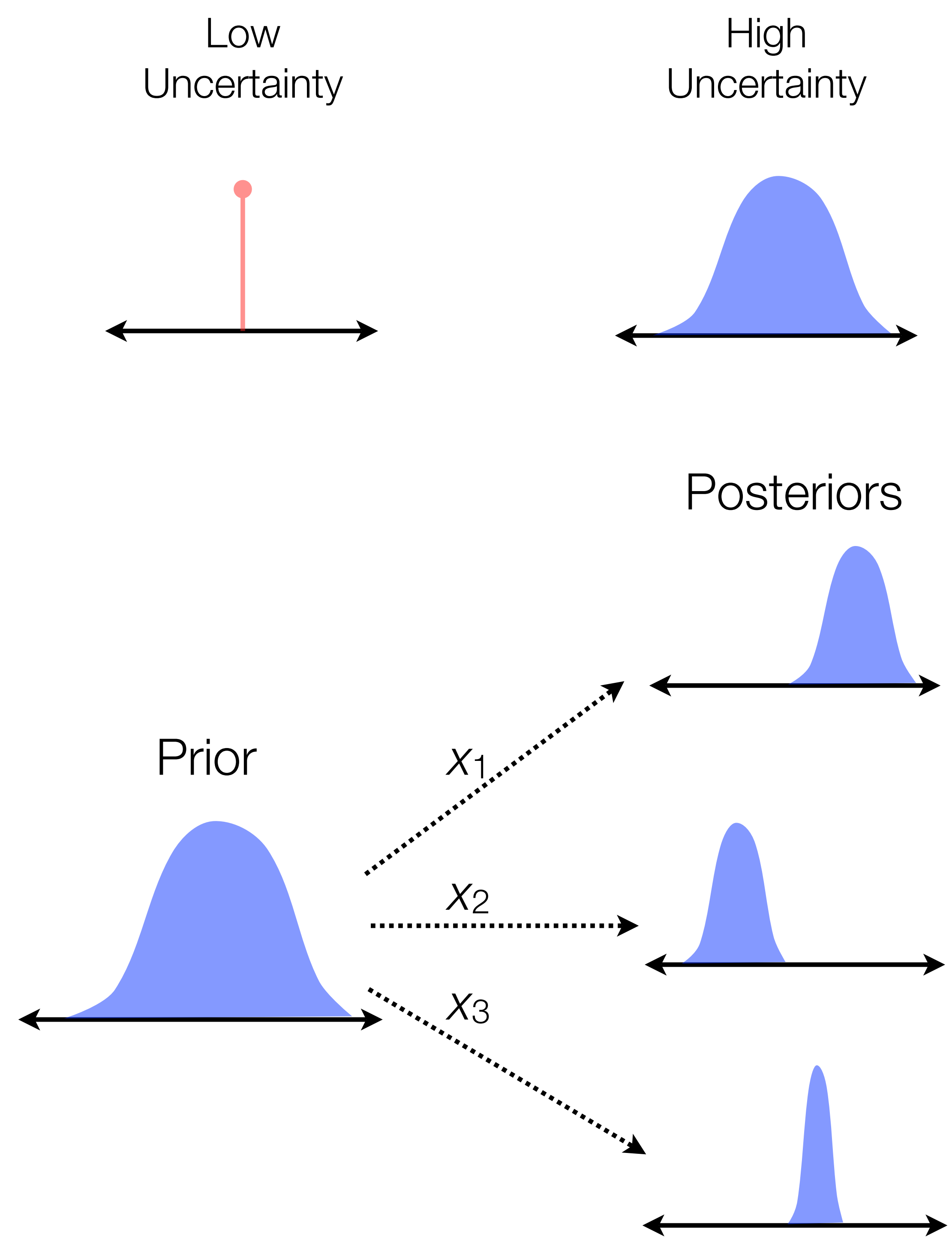




[De Groot, 1962]

Statistical Information

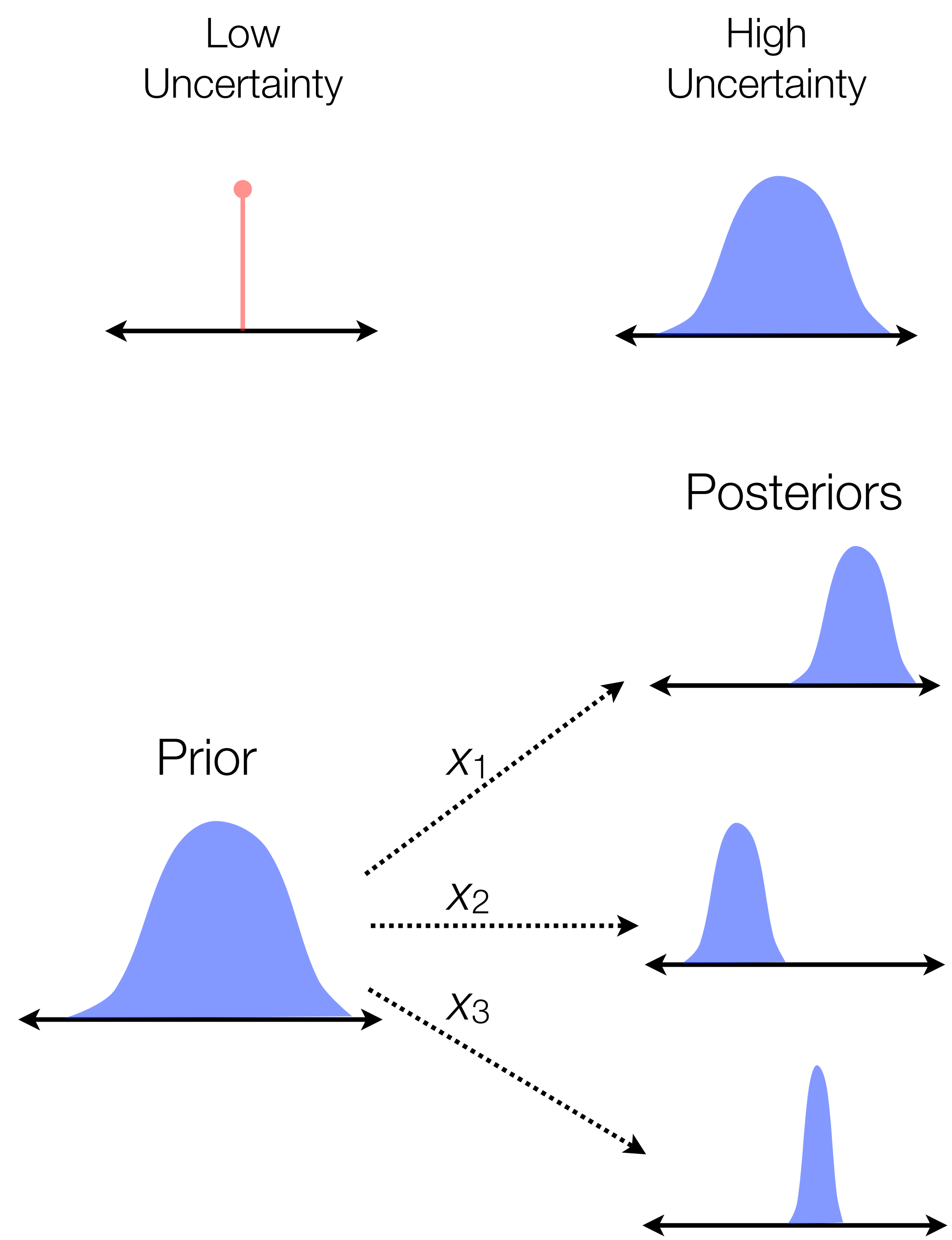
- Let U measure the "**uncertainty**" of a distribution ξ.
  - When ξ is peaked its uncertainty is small
- Assume π is a prior for ξ(x) — the posterior distribution after seeing x
  - Reduction in uncertainty is $\Delta U(\pi, \xi(x)) = U(\pi) - U(\xi(x))$



[De Groot, 1962]

Statistical Information

- Let U measure the "**uncertainty**" of a distribution ξ.
  - When ξ is peaked its uncertainty is small
- Assume π is a prior for ξ(*x*) — the posterior distribution after seeing *x*
  - Reduction in uncertainty is
    $$\Delta U(\pi, \xi(x)) = U(\pi) - U(\xi(x))$$
- The **statistical information** is the expected reduction in uncertainty for ξ when $X \sim M$ and $\pi := \mathbb{E}_M[\xi(X)]$
  $$\Delta \mathbb{U}(\xi, M) = \mathbb{E}_M[U(\pi) - U(\xi(X))]$$



[De Groot, 1962]

Statistical Information

- Observations can “at worst, contain no information ... typically [do] contain some information”

$$\Delta\mathbb{U}(\xi, M) \geq 0$$

$$\begin{aligned} \mathbb{E}_M[U(\pi) - U(\xi(X))] &\geq 0 \\ U(\mathbb{E}_M[\xi(X)] - \mathbb{E}_M[U(\xi(X))] &\geq 0 \\ \mathbb{J}_M[-U(\xi(X))] &\geq 0 \end{aligned}$$

Statistical Information

- Observations can "at worst, contain no information ... typically [do] contain some information"

$$\Delta\mathbb{U}(\xi, M) \geq 0$$

- By Jensen's inequality, information is non-negative **iff** the uncertainty function *U* is **concave**

$$\begin{aligned}
\mathbb{E}_M[U(\pi) - U(\xi(X))] &\geq 0 \\
U(\mathbb{E}_M[\xi(X)] - \mathbb{E}_M[U(\xi(X))] &\geq 0 \\
\mathbb{J}_M[-U(\xi(X))] &\geq 0
\end{aligned}$$

Statistical Information

- Observations can "at worst, contain no information ... typically [do] contain some information"

$$\Delta\mathbb{U}(\xi, M) \geq 0$$

- By Jensen's inequality, information is non-negative **iff** the uncertainty function *U* is **concave**

- Very general definition of information
  - e.g., Shannon information

$$U(p) = -\sum_i p_i \log p_i$$

$$\begin{aligned} \mathbb{E}_M[U(\pi) - U(\xi(X))] &\geq 0 \\ U(\mathbb{E}_M[\xi(X)] - \mathbb{E}_M[U(\xi(X))] &\geq 0 \\ \mathbb{J}_M[-U(\xi(X))] &\geq 0 \end{aligned}$$

Statistical Information

Prior Uncertainty Posterior Uncertainty

$$\mathbb{J}_M[-U(\xi(X))] = U(\mathbb{E}_M[\xi(X)]) - \mathbb{E}_M[U(\xi(X))] \geq 0$$

if and only if

U is concave

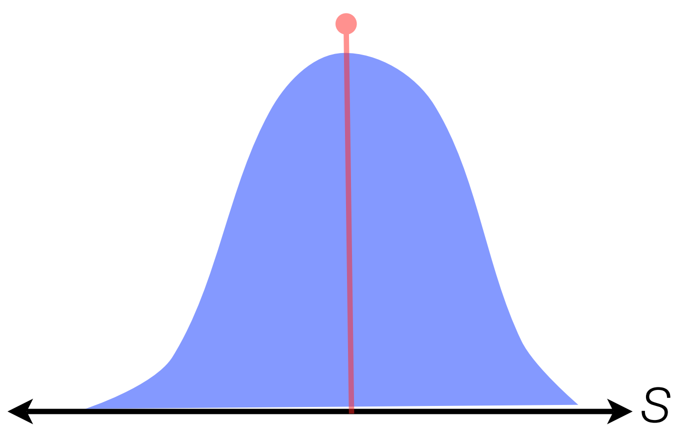
(another Jensen Gap)

[De Groot, 1962]

Bregman Information

- A recent, alternative formulation of information used to motivate clustering with Bregman divergences
  - Given a random variable S, its Bregman information is the minimum expected divergence from a single point in its domain
  - This single point is always the mean of S

$$\begin{aligned} \mathbb{B}_f(S) &:= \inf_{s \in \mathcal{S}} \mathbb{E}_{S\sim\sigma}[B_f(S,s)] \\ &= \mathbb{E}_{S\sim\sigma}[B_f(S,\mathbb{E}_\sigma[S])] \end{aligned}$$



[Banerjee et al., 2005]

Mathematics is the art of giving the same name to different things.

Jules Henri Poincaré (1854-1912)

Part II: Relationships and Representations

Terra Statistica

Terra Statistica

The acts of the mind, wherein it exerts its power over simple ideas, are chiefly these three:

1. **Combining** several **simple ideas into one compound one**, and thus all complex ideas are made.

2. The second is **bringing two ideas**, whether simple or complex, **together**, and setting them by one another **so as to take a view of them at once**, without uniting them into one, by which it gets all its ideas of relations.

3. The third is **separating** them **from all other ideas** that accompany them in their real existence: this is called abstraction, and thus all its general ideas are made.

John Locke (1632-1704)

The acts of the mind, wherein it exerts its power over simple ideas, are chiefly these three:

1. **Combining** several **simple ideas into one compound one**, and thus all complex ideas are made.

2. The second is **bringing two ideas**, whether simple or complex, **together**, and setting them by one another **so as to take a view of them at once**, without uniting them into one, by which it gets all its ideas of relations.

3. The third is **separating** them **from all other ideas** that accompany them in their real existence: this is called abstraction, and thus all its general ideas are made.

John Locke (1632-1704)

The acts of the mind, wherein it exerts its power over simple ideas, are chiefly these three:

1. **Combining** several **simple ideas into one compound one**, and thus all complex ideas are made.

2. The second is **bringing two ideas**, whether simple or complex, **together**, and setting them by one another **so as to take a view of them at once**, without uniting them into one, by which it gets all its ideas of relations.

3. The third is **separating** them **from all other ideas** that accompany them in their real existence: this is called abstraction, and thus all its general ideas are made.

John Locke (1632-1704)

Relationships

Regret and Bregman Divergence

Binary Mixtures (Review)

- Positive/Negative class distributions (P,Q)
- Mixture $M = \pi P + (1-\pi)Q$
- Conditional Positive Class Probability $\eta(x) = \pi \, dP/dM$

[Buja et al., 2005]

Regret and Bregman Divergence

Binary Mixtures (Review)

- Positive/Negative class distributions (P,Q)
- Mixture $M = \pi P + (1-\pi)Q$
- Conditional Positive Class Probability $\eta(x) = \pi\, dP/dM$

Proper Losses (Review)

- Fisher consistent $\underline{L}(\eta) = L(\eta,\eta)$
- Loss function is proper **iff** $\underline{L}$ is concave (Savage's Theorem)

$$L(\eta, \hat{\eta}) = \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta})$$

[Buja et al., 2005]

Regret and Bregman Divergence

Binary Mixtures (Review)

- Positive/Negative class distributions (*P*,*Q*)
- Mixture $M = \pi P + (1-\pi)Q$
- Conditional Positive Class Probability $\eta(x) = \pi \, dP/dM$

Proper Losses (Review)

- Fisher consistent $\underline{L}(\eta) = L(\eta,\eta)$
- Loss function is proper **iff** $\underline{L}$ is concave (Savage's Theorem)

$$L(\eta, \hat{\eta}) = \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta})$$

Bregman Divergence (Review)

- For convex *f*

$$B_f(t, t_0) = f(t) - f(t_0) - (t - t_0)f'(t)$$

[Buja et al., 2005]

Regret and Bregman Divergence

Binary Mixtures (Review)

- Positive/Negative class distributions (*P*,*Q*)
- Mixture $M = \pi P + (1-\pi)Q$
- Conditional Positive Class Probability $\eta(x) = \pi \, dP/dM$

Proper Losses (Review)

- Fisher consistent $\underline{L}(\eta) = L(\eta,\eta)$
- Loss function is proper **iff** $\underline{L}$ is concave (Savage's Theorem)

$$L(\eta, \hat{\eta}) = \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta})$$

Bregman Divergence (Review)

- For convex *f*

$$B_f(t, t_0) = f(t) - f(t_0) - (t - t_0)f'(t)$$

Bregman Divergence for Estimates

- Let $f = -\underline{L}$. Then *f* is convex and

$$\begin{aligned} B_f(\eta, \hat{\eta}) &= -\underline{L}(\eta) + \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta}) \\ &= L(\eta, \hat{\eta}) - \underline{L}(\eta) \end{aligned}$$

[Buja et al., 2005]

Point-wise Regret is a Bregman Divergence

$$B_f(\eta, \hat{\eta}) = L(\eta, \hat{\eta}) - \underline{L}(\eta)$$

for $f = -\underline{L}$

[Buja *et al*., 2005]

Bregman and Statistical Information

Bregman Info = Statistical Info

- Binary mixture (π, *P*, *Q*) = (η, *M*)

$$\mathbb{B}_f(\eta(X)) = \Delta\mathbb{U}(\eta, M)$$

when $f = -U$

Bregman and Statistical Information

Bregman Info = Statistical Info

- Binary mixture (π, *P*, *Q*) = (η, *M*)

$$\mathbb{B}_f(\eta(X)) = \Delta\mathbb{U}(\eta, M)$$

when $f = -U$

Proof

$$\mathbb{B}_f(\eta(X)) \quad = \quad \mathbb{E}_M[B_f(\eta(X), \mathbb{E}_M[\eta(X)])]$$

Bregman and Statistical Information

Bregman Info = Statistical Info

- Binary mixture (π, *P*, *Q*) = (η, *M*)

$$\mathbb{B}_f(\eta(X)) = \Delta\mathbb{U}(\eta, M)$$

when *f* = -*U*

Proof

$$\begin{aligned}\mathbb{B}_f(\eta(X)) &= \mathbb{E}_M[B_f(\eta(X), \mathbb{E}_M[\eta(X)])] \\ &= \mathbb{E}_M[f(\eta(X)) - f(\pi) \\ &\quad -(\eta(X) - \pi)f'(\pi)]\end{aligned}$$

Bregman and Statistical Information

Bregman Info = Statistical Info

- Binary mixture (π, P, Q) = (η, M)

$$\mathbb{B}_f(\eta(X)) = \Delta\mathbb{U}(\eta, M)$$

when $f = -U$

Proof

$$\begin{aligned}
\mathbb{B}_f(\eta(X)) &= \mathbb{E}_M[B_f(\eta(X), \mathbb{E}_M[\eta(X)])] \\
&= \mathbb{E}_M[f(\eta(X)) - f(\pi) \\
&\quad -(\eta(X) - \pi)f'(\pi)] \\
&= \mathbb{E}_M[f(\eta(X))] - f(\pi) - 0
\end{aligned}$$

Bregman and Statistical Information

Bregman Info = Statistical Info

- Binary mixture (π, *P*, *Q*) = (η, *M*)

$$\mathbb{B}_f(\eta(X)) = \Delta\mathbb{U}(\eta, M)$$

when $f = -U$

Proof

$$\begin{aligned}
\mathbb{B}_f(\eta(X)) &= \mathbb{E}_M[B_f(\eta(X), \mathbb{E}_M[\eta(X)])] \\
&= \mathbb{E}_M[f(\eta(X)) - f(\pi) \\
&\quad -(\eta(X) - \pi)f'(\pi)] \\
&= \mathbb{E}_M[f(\eta(X))] - f(\pi) - 0 \\
&= U(\pi) - \mathbb{E}_M[U(\eta(X))]
\end{aligned}$$

Bregman and Statistical Information

Bregman Info = Statistical Info

- Binary mixture (π, *P*, *Q*) = (η, *M*)

$$\mathbb{B}_f(\eta(X)) = \Delta\mathbb{U}(\eta, M)$$

when $f = -U$

Proof

$$\begin{aligned}
\mathbb{B}_f(\eta(X)) &= \mathbb{E}_M[B_f(\eta(X), \mathbb{E}_M[\eta(X)])] \\
&= \mathbb{E}_M[f(\eta(X)) - f(\pi) \\
&\quad -(\eta(X) - \pi)f'(\pi)] \\
&= \mathbb{E}_M[f(\eta(X))] - f(\pi) - 0 \\
&= U(\pi) - \mathbb{E}_M[U(\eta(X))] \\
&= U(\mathbb{E}_M[\eta(X)]) - \mathbb{E}_M[U(\eta(X))] \\
&= \Delta\mathbb{U}(\eta, M)
\end{aligned}$$

Bregman and Statistical Information

Bregman Info = Statistical Info

- Binary mixture (π, *P*, *Q*) = (η, *M*)

$$\mathbb{B}_f(\eta(X)) = \Delta\mathbb{U}(\eta, M)$$

when *f* = -*U*

Proof

$$\begin{aligned}
\mathbb{B}_f(\eta(X)) &= \mathbb{E}_M[B_f(\eta(X), \mathbb{E}_M[\eta(X)])] \\
&= \mathbb{E}_M[f(\eta(X)) - f(\pi) \\
&\quad -(\eta(X) - \pi)f'(\pi)] \\
&= \mathbb{E}_M[f(\eta(X))] - f(\pi) - 0 \\
&= U(\pi) - \mathbb{E}_M[U(\eta(X))] \\
&= U(\mathbb{E}_M[\eta(X)]) - \mathbb{E}_M[U(\eta(X))] \\
&= \Delta\mathbb{U}(\eta, M)
\end{aligned}$$

Information and Proper Losses

- Savage's Theorem implies $\underline{L}$ is concave for proper scoring rules
  - Choosing $U = \underline{L}$ gives a measure of information in the mixture (π, *P*, *Q*) = (η, *M*)

$$\begin{aligned}
\Delta\underline{\mathbb{L}}(\eta, M) &= \mathbb{E}_M[\underline{L}(\pi) - \underline{L}(\eta)] \\
&= \underline{\mathbb{L}}(\pi, M) - \underline{\mathbb{L}}(\eta, M)
\end{aligned}$$

- Maximum reduction in risk obtained by knowing posterior

Bregman Info = Statistical Info

$$\mathbb{B}_f(\eta(X)) = \Delta\mathbb{U}(\eta, M) = \Delta\mathbb{L}(\eta, M)$$

for $f = -U = -\underline{L}$

Can be interpreted as
maximal reduction in risk

Statistical Information and f-Divergence

Binary Mixtures & Experiments

- (P,Q) vs. (π, P, Q) = (η, M)
- For each π there is a mapping between dP/dQ and η

$$\begin{aligned} \eta &= \frac{\pi dP}{dM} \\ &= \frac{\pi dP}{\pi dP + (1-\pi) dQ} \\ &= \frac{\lambda}{\lambda + 1} \end{aligned}$$

where $\lambda = \frac{\pi}{(1-\pi)} \frac{dP}{dQ}$

f-Divergence to Information

- If
  then

  for all binary mixtures (π, P, Q)

Information to f-Divergence

- If
  then

  for all binary mixtures (π, P, Q)

[Österreicher & Vajda, 1993]

Statistical Information and f-Divergence

Binary Mixtures & Experiments

- (P,Q) vs. (π, P, Q) = (η, M)
- For each π there is a mapping between dP/dQ and η

$$\begin{aligned}\eta &= \frac{\pi dP}{dM} \\ &= \frac{\pi dP}{\pi dP + (1-\pi)dQ} \\ &= \frac{\lambda}{\lambda+1}\end{aligned}$$

where $\lambda = \frac{\pi}{(1-\pi)} \frac{dP}{dQ}$

$$\frac{dP}{dQ} = \frac{(1-\pi)}{\pi} \frac{\eta}{(1-\eta)}$$

f-Divergence to Information

- If
  then

  for all binary mixtures (π, P, Q)

Information to f-Divergence

- If
  then

  for all binary mixtures (π, P, Q)

[Österreicher & Vajda, 1993]

Statistical Information and f-Divergence

Binary Mixtures & Experiments

- (P,Q) vs. (π, P, Q) = (η, M)
- For each π there is a mapping between dP/dQ and η

$$\eta = \frac{\pi dP}{dM} = \frac{\pi dP}{\pi dP + (1-\pi)dQ} = \frac{\lambda}{\lambda+1}$$

where $\lambda = \frac{\pi}{(1-\pi)}\frac{dP}{dQ}$

$$\frac{dP}{dQ} = \frac{(1-\pi)}{\pi}\frac{\eta}{(1-\eta)}$$

f-Divergence to Information

- If $f^{\pi}(t) = \underline{L}(\pi) - (\pi t + 1 - \pi)\underline{L}\left(\frac{\pi t}{\pi t + 1 - \pi}\right)$ then
$$\mathbb{I}_{f^{\pi}}(P,Q) = \Delta\underline{\mathbb{L}}(\eta, M)$$
for all binary mixtures (π, P, Q)

[Österreicher & Vajda, 1993]

Statistical Information and f-Divergence

Binary Mixtures & Experiments

- (P,Q) vs. (π, P, Q) = (η, M)
- For each π there is a mapping between dP/dQ and η

$$\eta = \frac{\pi dP}{dM} = \frac{\pi dP}{\pi dP + (1-\pi) dQ} = \frac{\lambda}{\lambda + 1}$$

where $\lambda = \frac{\pi}{(1-\pi)} \frac{dP}{dQ}$

$$\frac{dP}{dQ} = \frac{(1-\pi)}{\pi} \frac{\eta}{(1-\eta)}$$

f-Divergence to Information

- If $f^\pi(t) = \underline{L}(\pi) - (\pi t + 1 - \pi)\underline{L}\left(\frac{\pi t}{\pi t + 1 - \pi}\right)$ then
$$\mathbb{I}_{f^\pi}(P,Q) = \Delta\underline{\mathbb{L}}(\eta, M)$$
for all binary mixtures (π, P, Q)

Information to f-Divergence

- If $\underline{L}^\pi(\eta) = -\frac{1-\eta}{1-\pi} f\left(\frac{1-\pi}{\pi}\frac{\eta}{1-\eta}\right)$ then
$$\mathbb{I}_f(P,Q) = \Delta\underline{\mathbb{L}}^\pi(\eta, M)$$
for all binary mixtures (π, P, Q)

[Österreicher & Vajda, 1993]

f-Divergence = Statistical Info

$$\mathbb{I}_f(P, Q) = \Delta\underline{\mathbb{L}}^\pi(\eta, M)$$

for binary mixtures (π, P, Q)
when $f = -\underline{L}$

(plus a map to/from [0,1])

[Österreicher & Vajda, 1993]

The acts of the mind, wherein it exerts its power over simple ideas, are chiefly these three:

1. **Combining** several **simple ideas into one compound one**, and thus all complex ideas are made.

2. The second is **bringing two ideas**, whether simple or complex, **together**, and setting them by one another **so as to take a view of them at once**, without uniting them into one, by which it gets all its ideas of relations.

3. The third is **separating** them **from all other ideas** that accompany them in their real existence: this is called abstraction, and thus all its general ideas are made.

John Locke (1632-1704)

The acts of the mind, wherein it exerts its power over simple ideas, are chiefly these three:

1. **Combining** several **simple ideas into one compound one**, and thus all complex ideas are made.

2. The second is **bringing two ideas**, whether simple or complex, **together**, and setting them by one another **so as to take a view of them at once**, without uniting them into one, by which it gets all its ideas of relations.

3. The third is **separating** them **from all other ideas** that accompany them in their real existence: this is called abstraction, and thus all its general ideas are made.

John Locke (1632-1704)

Weighted Integral Representations

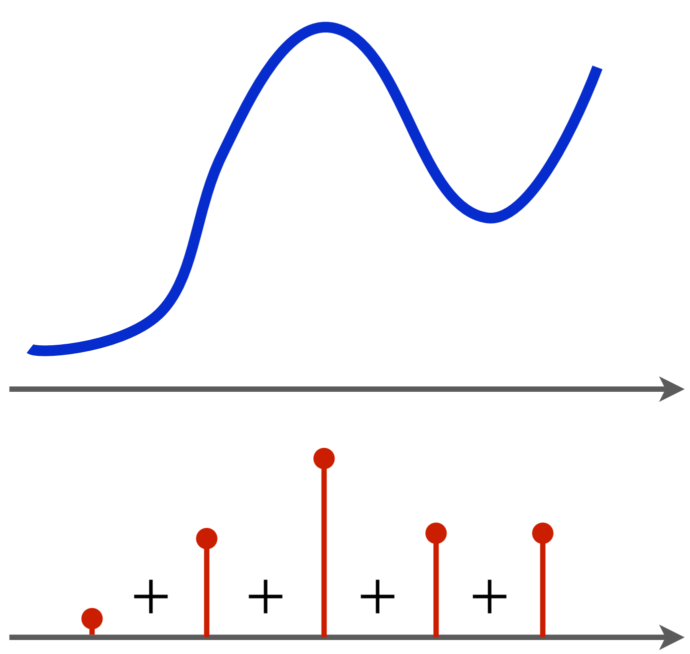
Representations of Functions

Functions as "Sums" of Points

- A function *f* can be described by its values at each point

$$f(x) = \sum_u f_u \, \delta_u(x)$$

where $\delta_u(x) := [\![u = x]\!]$



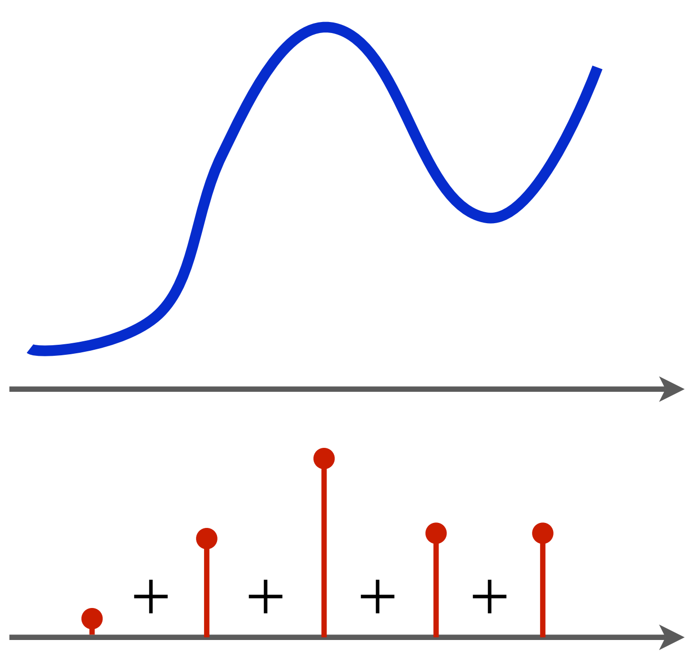
Representations of Functions

Functions as "Sums" of Points

- A function *f* can be described by its values at each point

$$f(x) = \sum_u f_u\, \delta_u(x)$$

where $\delta_u(x) := [\![u = x]\!]$

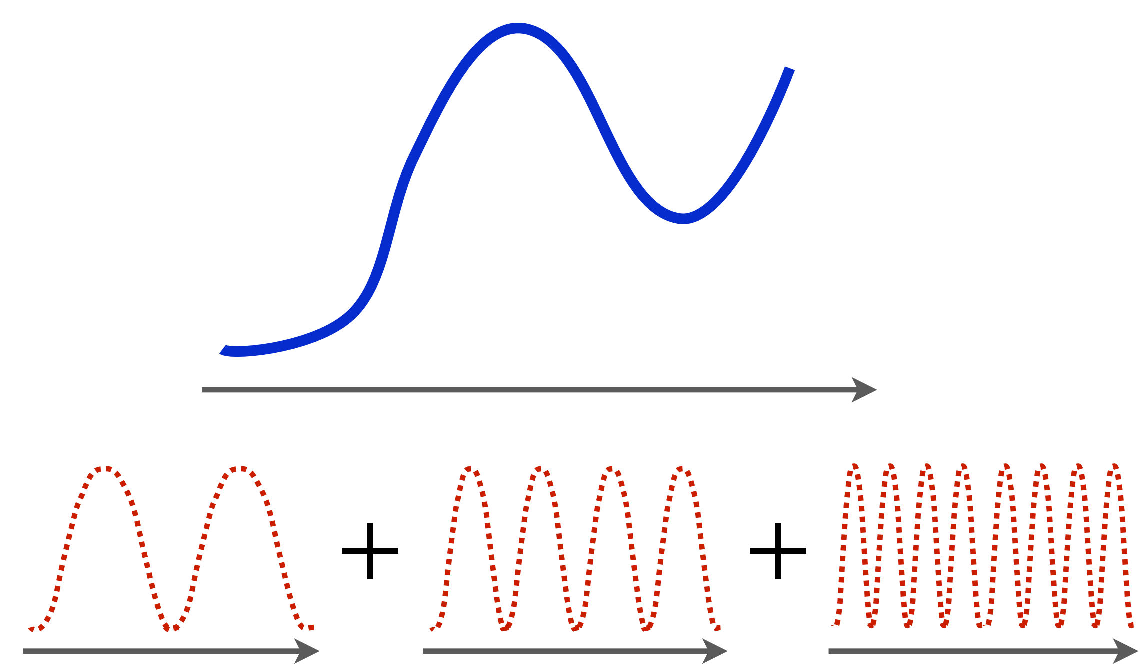


Functions as Sums of Functions

- Can also describe *f* as a sum of "simple" functions

$$f(x) = \sum_i w_i\, \phi_i(x)$$

(e.g., Fourier analysis)



Integral Representation of f-Divergence

Taylor Integral Representation

$$f(t) = \Lambda_f(t) + \int_a^b g_s(t)\, f''(s)\, ds$$

Linear Term Simple Weights

$$g_s(t) = [\![s \geq t_0]\!](t-s)_+ + [\![s < t_0]\!](s-t)_+$$

f-Divergence

$$\mathbb{I}_f(P,Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ} \right) \right]$$

[Liese & Vajda et al., 2006]

Integral Representation of f-Divergence

Taylor Integral Representation

$$f(t) = \underbrace{\Lambda_f(t)}_{\text{Linear Term}} + \int_a^b \underbrace{g_s(t)}_{\text{Simple}}\, \underbrace{f''(s)}_{\text{Weights}}\, ds$$

$$g_s(t) = [\![s \geq t_0]\!](t-s)_+ + [\![s < t_0]\!](s-t)_+$$

f-Divergence

$$\mathbb{I}_f(P,Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right]$$

Integral Representation I

$$\begin{aligned}
\mathbb{I}_f(P,Q) &= \mathbb{E}_Q\left[\int_0^\infty g_s\left(\frac{dP}{dQ}\right) f''(s)\, ds\right] \\
&= \int_0^\infty \mathbb{E}_Q\left[g_s\left(\frac{dP}{dQ}\right)\right] f''(s)\, ds \\
\mathbb{I}_f(P,Q) &= \int_0^\infty \mathbb{I}_{g_s}(P,Q)\, f''(s)\, ds
\end{aligned}$$

[Liese & Vajda et al., 2006]

Integral Representation of f-Divergence

Taylor Integral Representation

$$f(t) = \underbrace{\Lambda_f(t)}_{\text{Linear Term}} + \int_a^b \underbrace{g_s(t)}_{\text{Simple}}\, \underbrace{f''(s)}_{\text{Weights}}\, ds$$

$$g_s(t) = [\![s \geq t_0]\!](t-s)_+ + [\![s < t_0]\!](s-t)_+$$

f-Divergence

$$\mathbb{I}_f(P,Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right]$$

Integral Representation I

$$\begin{aligned}\mathbb{I}_f(P,Q) &= \mathbb{E}_Q\left[\int_0^\infty g_s\left(\frac{dP}{dQ}\right)\, f''(s)\, ds\right] \\ &= \int_0^\infty \mathbb{E}_Q\left[g_s\left(\frac{dP}{dQ}\right)\right]\, f''(s)\, ds\end{aligned}$$

$$\mathbb{I}_f(P,Q) = \int_0^\infty \mathbb{I}_{g_s}(P,Q)\, f''(s)\, ds$$

Integral Representation II

$$\begin{aligned}\mathbb{I}_f(P,Q) &= \int_0^1 \mathbb{I}_{g_{\frac{1-\pi}{\pi}}}(P,Q)\, f''\left(\tfrac{1-\pi}{\pi}\right)\pi^{-2}\, d\pi \\ &= \int_0^1 \mathbb{I}_{f_\pi}(P,Q)\, \gamma(\pi)\, d\pi\end{aligned}$$

$$\gamma(\pi) = \tfrac{1}{\pi^3} f''\left(\tfrac{1-\pi}{\pi}\right)$$

$$f_\pi(t) = \min(1-\pi, \pi) - \min(1-\pi, \pi t)$$

[Liese & Vajda et al., 2006]

Integral Representation of f-Divergence

$$\mathbb{I}_f(P,Q) \quad = \quad \int_0^1 \mathbb{I}_{f_\pi}(P,Q)\,\gamma(\pi)\,d\pi$$

Weight Function $\quad \gamma(\pi) = \frac{1}{\pi^3} f''\left(\frac{1-\pi}{\pi}\right)$

Primitives $\quad f_\pi(t) = \min(1-\pi,\pi) - \min(1-\pi,\pi t)$

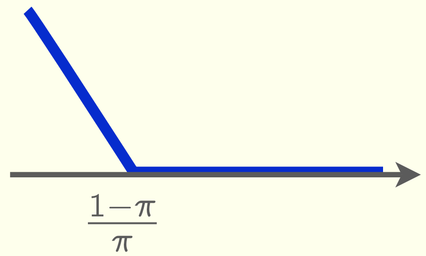
[Österreicher & Vajda, 1993]

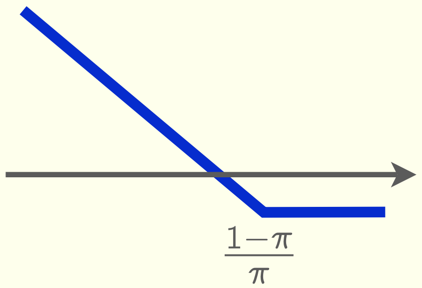
Integral Representation of f-Divergence

$$\mathbb{I}_f(P,Q) \quad = \quad \int_0^1 \mathbb{I}_{f_\pi}(P,Q)\,\gamma(\pi)\,d\pi$$

Weight Function $\gamma(\pi) = \frac{1}{\pi^3} f''\left(\frac{1-\pi}{\pi}\right)$

Primitives $f_\pi(t) = \min(1-\pi, \pi) - \min(1-\pi, \pi t)$





[Österreicher & Vajda, 1993]

Integral Representation of Proper Losses

Taylor Integral Representation

$$f(t) = \underset{\substack{\uparrow\\ \text{Linear Term}}}{\Lambda_f(t)} + \int_a^b \underset{\substack{\uparrow\\ \text{Simple}}}{g_s(t, t_0)}\, \underset{\substack{\uparrow\\ \text{Weights}}}{f''(s)}\, ds$$

$$g_s(t, t_0) = [\![s \geq t_0]\!](t - s)_+ + [\![s < t_0]\!](s - t)_+$$

[Shuford et al., 1966] [Schervish, 1989] [Buja et al., 2005] [Lambert et al., 2008]

Integral Representation of Proper Losses

Taylor Integral Representation

$$f(t) = \Lambda_f(t) + \int_a^b g_s(t, t_0)\, f''(s)\, ds$$

Linear Term Simple Weights

$$g_s(t, t_0) = [\![s \geq t_0]\!](t - s)_+ + [\![s < t_0]\!](s - t)_+$$

Savage's Theorem

- Given concave $\underline{L}$ the loss is

$$L(\eta, \hat{\eta}) = \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta})$$

[Shuford et al., 1966] [Schervish, 1989] [Buja et al., 2005] [Lambert et al., 2008]

Integral Representation of Proper Losses

Taylor Integral Representation

$$f(t) = \Lambda_f(t) + \int_a^b g_s(t, t_0)\, f''(s)\, ds$$

Linear Term Simple Weights

$$g_s(t, t_0) = [\![s \geq t_0]\!](t - s)_+ + [\![s < t_0]\!](s - t)_+$$

Savage's Theorem

- Given concave $\underline{L}$ the loss is

$$L(\eta, \hat{\eta}) = \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta})$$

Int. Representation of Bayes Risk

$$\begin{aligned} \underline{L}(\eta) &= \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta}) + \int_0^1 g_c(\eta, \hat{\eta})\, \underline{L}''(c)\, dc \\ &= L(\eta, \hat{\eta}) + \int_0^1 g_c(\eta, \hat{\eta})\, \underline{L}''(c)\, dc \end{aligned}$$

[Shuford et al., 1966] [Schervish, 1989] [Buja et al., 2005] [Lambert et al., 2008]

Integral Representation of Proper Losses

Taylor Integral Representation

$$f(t) = \underbrace{\Lambda_f(t)}_{\text{Linear Term}} + \int_a^b \underbrace{g_s(t, t_0)}_{\text{Simple}}\, \underbrace{f''(s)}_{\text{Weights}}\, ds$$

$$g_s(t, t_0) = [\![s \geq t_0]\!](t - s)_+ + [\![s < t_0]\!](s - t)_+$$

Int. Representation of Risk

$$L(\eta, \hat{\eta}) \quad = \quad \underline{L}(\eta) + \int_0^1 L_c(\eta, \hat{\eta})\, w(c)\, dc$$

$$L_c(\eta, \hat{\eta}) = [\![\eta > c \geq \hat{\eta}]\!](\eta - c) + [\![\hat{\eta} > c \geq \eta]\!](c - \eta)$$

$$w(c) = -\underline{L}''(c)$$

Savage's Theorem

- Given concave $\underline{L}$ the loss is

$$L(\eta, \hat{\eta}) = \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta})$$

Int. Representation of Bayes Risk

$$\begin{aligned} \underline{L}(\eta) &= \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta}) + \int_0^1 g_c(\eta, \hat{\eta})\, \underline{L}''(c)\, dc \\ &= L(\eta, \hat{\eta}) + \int_0^1 g_c(\eta, \hat{\eta})\, \underline{L}''(c)\, dc \end{aligned}$$

[Shuford et al., 1966] [Schervish, 1989] [Buja et al., 2005] [Lambert et al., 2008]

Integral Representation of Proper Losses

Taylor Integral Representation

$$f(t) = \underset{\substack{\uparrow \\ \text{Linear Term}}}{\Lambda_f(t)} + \int_a^b \underset{\substack{\uparrow \\ \text{Simple}}}{g_s(t, t_0)}\, \underset{\substack{\uparrow \\ \text{Weights}}}{f''(s)}\, ds$$

$$g_s(t, t_0) = [\![s \geq t_0]\!](t - s)_+ + [\![s < t_0]\!](s - t)_+$$

Int. Representation of Risk

$$L(\eta, \hat{\eta}) \quad = \quad \underline{L}(\eta) + \int_0^1 L_c(\eta, \hat{\eta})\, w(c)\, dc$$

$$L_c(\eta, \hat{\eta}) = [\![\eta > c \geq \hat{\eta}]\!](\eta - c) + [\![\hat{\eta} > c \geq \eta]\!](c - \eta)$$

$$w(c) = -\underline{L}''(c)$$

Savage's Theorem

- Given concave $\underline{L}$ the loss is

$$L(\eta, \hat{\eta}) = \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta})$$

Int. Representation of Loss

$$\ell(y, \hat{\eta}) = L(y, \hat{\eta}) \text{ for } y \in \{0, 1\}$$

- Assuming $\underline{L}(0) = \underline{L}(1) = 0$

$$\ell(y, \hat{\eta}) = \int_0^1 \ell_c(y, \hat{\eta})\, w(c)\, dc$$

Int. Representation of Bayes Risk

$$\begin{aligned} \underline{L}(\eta) &= \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta}) + \int_0^1 g_c(\eta, \hat{\eta})\, \underline{L}''(c)\, dc \\ &= L(\eta, \hat{\eta}) + \int_0^1 g_c(\eta, \hat{\eta})\, \underline{L}''(c)\, dc \end{aligned}$$

[Shuford et al., 1966] [Schervish, 1989] [Buja et al., 2005] [Lambert et al., 2008]

Integral Representation of Proper Losses

Taylor Integral Representation

$$f(t) = \underbrace{\Lambda_f(t)}_{\text{Linear Term}} + \int_a^b \underbrace{g_s(t, t_0)}_{\text{Simple}} \underbrace{f''(s)}_{\text{Weights}} \, ds$$

$$g_s(t, t_0) = [\![s \geq t_0]\!](t - s)_+ + [\![s < t_0]\!](s - t)_+$$

Savage's Theorem

- Given concave $\underline{L}$ the loss is

$$L(\eta, \hat{\eta}) = \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta})$$

Int. Representation of Bayes Risk

$$\begin{aligned} \underline{L}(\eta) &= \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta}) + \int_0^1 g_c(\eta, \hat{\eta}) \, \underline{L}''(c) \, dc \\ &= L(\eta, \hat{\eta}) + \int_0^1 g_c(\eta, \hat{\eta}) \, \underline{L}''(c) \, dc \end{aligned}$$

Int. Representation of Risk

$$L(\eta, \hat{\eta}) \quad = \quad \underline{L}(\eta) + \int_0^1 L_c(\eta, \hat{\eta}) \, w(c) \, dc$$

$$L_c(\eta, \hat{\eta}) = [\![\eta > c \geq \hat{\eta}]\!](\eta - c) + [\![\hat{\eta} > c \geq \eta]\!](c - \eta)$$

$$w(c) = -\underline{L}''(c)$$

Int. Representation of Loss

$$\ell(y, \hat{\eta}) = L(y, \hat{\eta}) \text{ for } y \in \{0, 1\}$$

- Assuming $\underline{L}(0) = \underline{L}(1) = 0$

$$\ell(y, \hat{\eta}) = \int_0^1 \ell_c(y, \hat{\eta}) \, w(c) \, dc$$

Cost-Weighted Loss

$$\ell_c(y, \hat{\eta}) = (1 - c)[\![y = 1]\!][\![c \geq \hat{\eta}]\!] + c[\![y = 0]\!][\![\hat{\eta} > c]\!]$$

[Shuford et al., 1966] [Schervish, 1989] [Buja et al., 2005] [Lambert et al., 2008]

Integral Representation of Proper Losses

$$\ell(y,\hat{\eta}) = \int_0^1 \ell_c(y,\hat{\eta})\, w(c)\, dc$$

Weight Function $w(c) = -\underline{L}''(c)$

Primitives $\ell_c(y,\hat{\eta}) = \underbrace{(1-c)}_{\text{Cost of False Negative}}[\![y=1]\!][\![c \geq \hat{\eta}]\!] + \underbrace{c}_{\text{Cost of False Positive}}[\![y=0]\!][\![\hat{\eta} > c]\!]$

[Schervish, 1989]

Integral Representation Corollaries

Point-wise Risk

$$\begin{aligned} L(\eta, \hat{\eta}) &= \mathbb{E}_{y\sim\eta}\left[\int_0^1 \ell_c(y, \hat{\eta})\, w(c)\, dc\right] \\ &= \int_0^1 L_c(\eta, \hat{\eta})\, w(c)\, dc \end{aligned}$$

Integral Representation Corollaries

Point-wise Risk

$$
\begin{aligned}
L(\eta, \hat{\eta}) &= \mathbb{E}_{y \sim \eta}\left[\int_0^1 \ell_c(y, \hat{\eta})\, w(c)\, dc\right] \\
&= \int_0^1 L_c(\eta, \hat{\eta})\, w(c)\, dc
\end{aligned}
$$

Point-wise Bayes Risk

$$\underline{L}(\eta) = \int_0^1 \underline{L}_c(\eta)\, w(c)\, dc$$

$$\underline{L}_c(\eta) = \min((1-\eta)c, (1-c)\eta)$$

Integral Representation Corollaries

Point-wise Risk

$$
\begin{aligned}
L(\eta, \hat{\eta}) &= \mathbb{E}_{y \sim \eta}\left[\int_0^1 \ell_c(y, \hat{\eta})\, w(c)\, dc\right] \\
&= \int_0^1 L_c(\eta, \hat{\eta})\, w(c)\, dc
\end{aligned}
$$

Point-wise Bayes Risk

$$\underline{L}(\eta) = \int_0^1 \underline{L}_c(\eta)\, w(c)\, dc$$

$$\underline{L}_c(\eta) = \min((1-\eta)c, (1-c)\eta)$$

Point-wise Regret

$$B(\eta, \hat{\eta}) = \int_{\min(\eta,\hat{\eta})}^{\max(\eta,\hat{\eta})} |\eta - c|\, w(c)\, dc$$

Integral Representation Corollaries

Point-wise Risk

$$\begin{aligned} L(\eta, \hat{\eta}) &= \mathbb{E}_{y\sim\eta}\left[\int_0^1 \ell_c(y, \hat{\eta})\, w(c)\, dc\right] \\ &= \int_0^1 L_c(\eta, \hat{\eta})\, w(c)\, dc \end{aligned}$$

Risk

$$\begin{aligned} \mathbb{L}(\eta, \hat{\eta}, M) &= \mathbb{E}_M[L(\eta, \hat{\eta})] \\ &= \int_0^1 \mathbb{L}_c(\hat{\eta})\, w(c)\, dc \end{aligned}$$

Point-wise Bayes Risk

$$\underline{L}(\eta) = \int_0^1 \underline{L}_c(\eta)\, w(c)\, dc$$

$$\underline{L}_c(\eta) = \min((1-\eta)c, (1-c)\eta)$$

Point-wise Regret

$$B(\eta, \hat{\eta}) = \int_{\min(\eta,\hat{\eta})}^{\max(\eta,\hat{\eta})} |\eta - c|\, w(c)\, dc$$

Integral Representation Corollaries

Point-wise Risk

$$L(\eta, \hat{\eta}) = \mathbb{E}_{y \sim \eta}\left[\int_0^1 \ell_c(y, \hat{\eta})\, w(c)\, dc\right] = \int_0^1 L_c(\eta, \hat{\eta})\, w(c)\, dc$$

Point-wise Bayes Risk

$$\underline{L}(\eta) = \int_0^1 \underline{L}_c(\eta)\, w(c)\, dc$$

$$\underline{L}_c(\eta) = \min((1-\eta)c, (1-c)\eta)$$

Point-wise Regret

$$B(\eta, \hat{\eta}) = \int_{\min(\eta, \hat{\eta})}^{\max(\eta, \hat{\eta})} |\eta - c|\, w(c)\, dc$$

Risk

$$\mathbb{L}(\eta, \hat{\eta}, M) = \mathbb{E}_M[L(\eta, \hat{\eta})] = \int_0^1 \mathbb{L}_c(\hat{\eta})\, w(c)\, dc$$

Bayes Risk

$$\underline{\mathbb{L}}(\eta, M) = \int_0^1 \underline{\mathbb{L}}_c(\eta, M)\, w(c)\, dc$$

Integral Representation Corollaries

Point-wise Risk

$$L(\eta, \hat{\eta}) = \mathbb{E}_{y \sim \eta}\left[\int_0^1 \ell_c(y, \hat{\eta})\, w(c)\, dc\right]$$
$$= \int_0^1 L_c(\eta, \hat{\eta})\, w(c)\, dc$$

Point-wise Bayes Risk

$$\underline{L}(\eta) = \int_0^1 \underline{L}_c(\eta)\, w(c)\, dc$$

$$\underline{L}_c(\eta) = \min((1-\eta)c, (1-c)\eta)$$

Point-wise Regret

$$B(\eta, \hat{\eta}) = \int_{\min(\eta,\hat{\eta})}^{\max(\eta,\hat{\eta})} |\eta - c|\, w(c)\, dc$$

Risk

$$\mathbb{L}(\eta, \hat{\eta}, M) = \mathbb{E}_M[L(\eta, \hat{\eta})]$$
$$= \int_0^1 \mathbb{L}_c(\hat{\eta})\, w(c)\, dc$$

Bayes Risk

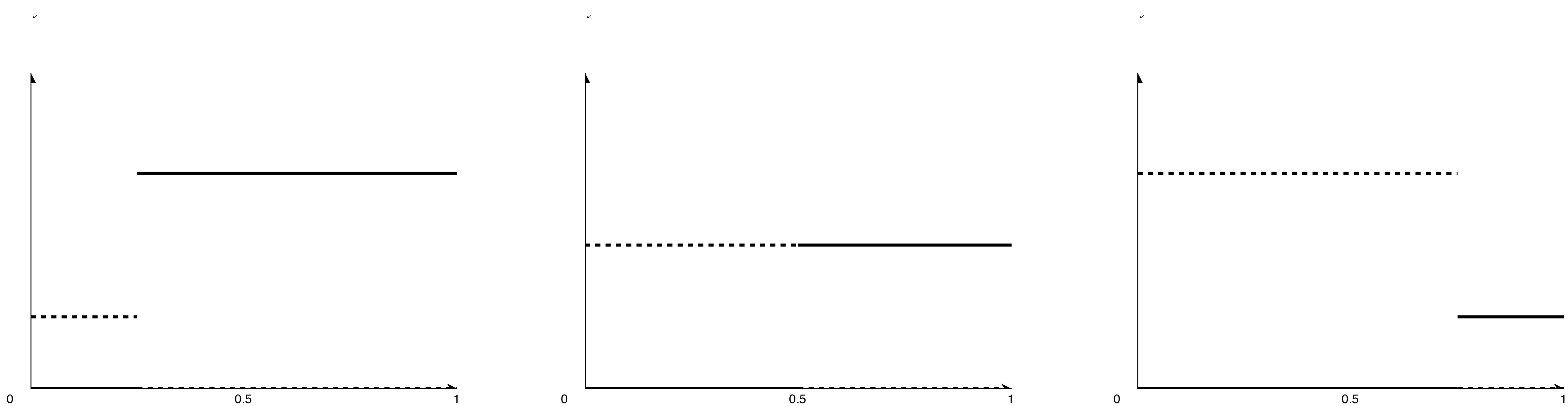
$$\underline{\mathbb{L}}(\eta, M) = \int_0^1 \underline{\mathbb{L}}_c(\eta, M)\, w(c)\, dc$$

Statistical Information

$$\Delta\underline{\mathbb{L}}(\eta, M) = \int_0^1 \Delta\underline{\mathbb{L}}_c(\eta, M)\, w(c)\, dc$$

Cost-Weighted Misclassification Loss

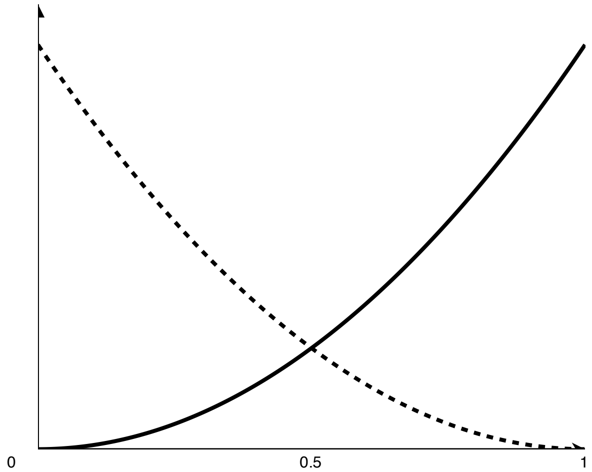
$$\ell_c(y, \hat{\eta}) = (1-c)[\![y=1]\!][\![c \geq \hat{\eta}]\!] + c[\![y=0]\!][\![\hat{\eta} > c]\!]$$



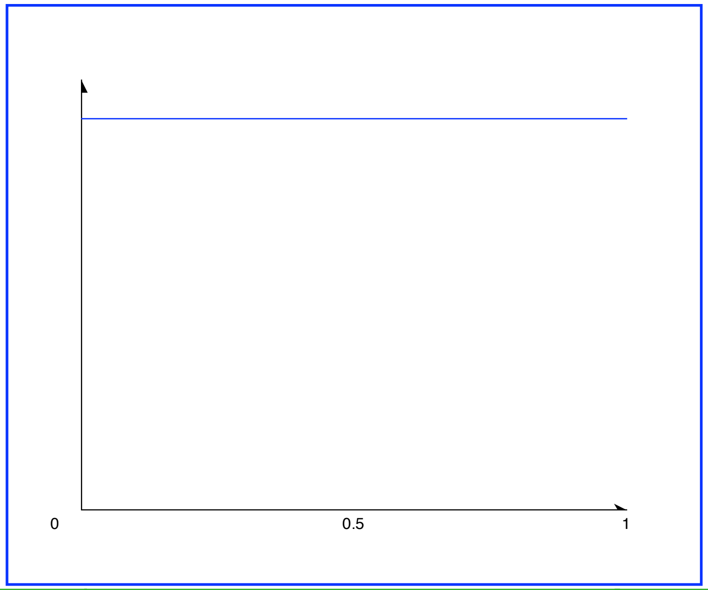
c = 0.25 c = 0.5 c = 0.75

Example - Square Loss

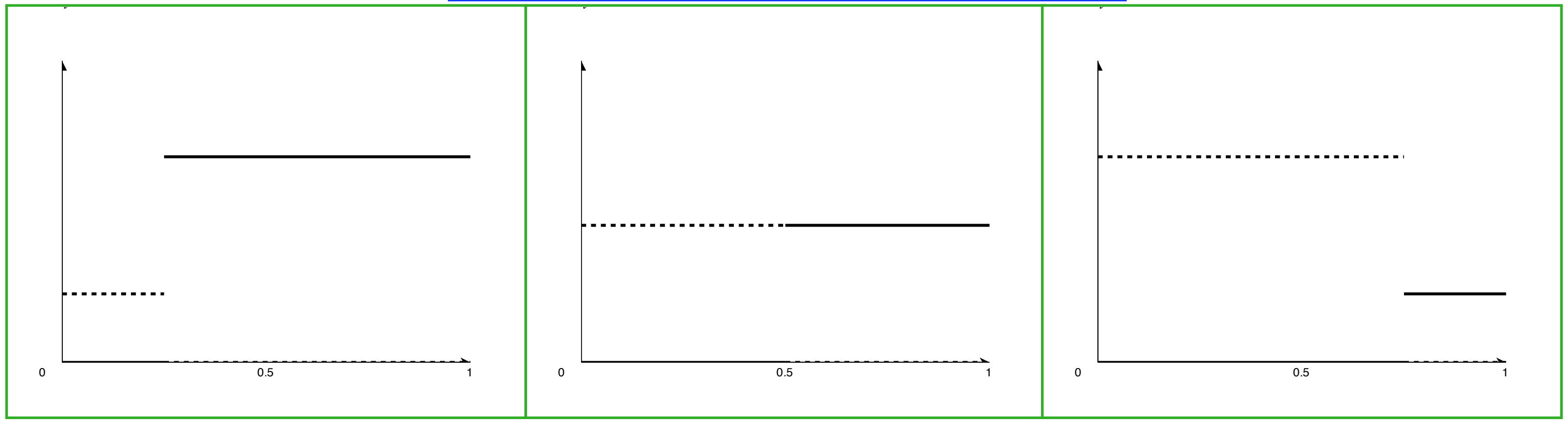
$$\ell(y, \hat{\eta}) = (y - \hat{\eta})^2$$



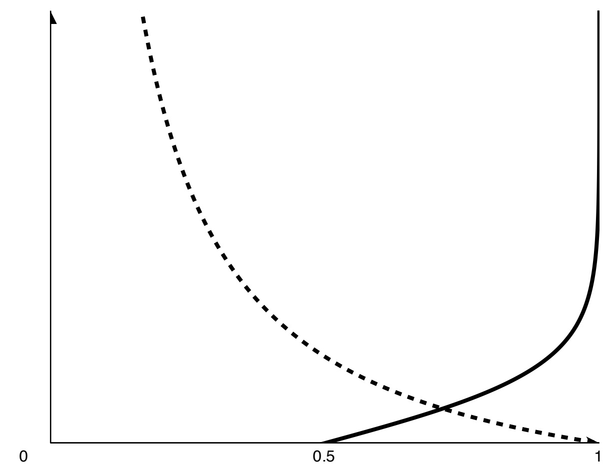
$$\ell(y, \hat{\eta}) = \int_0^1 \boxed{\ell_c(y, \hat{\eta})}\,\boxed{w(c)}\,dc$$



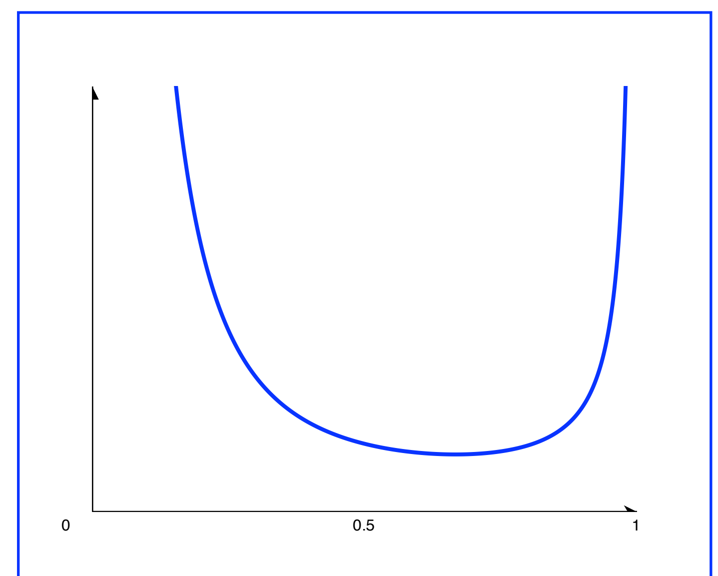
$$w(c) = 1$$



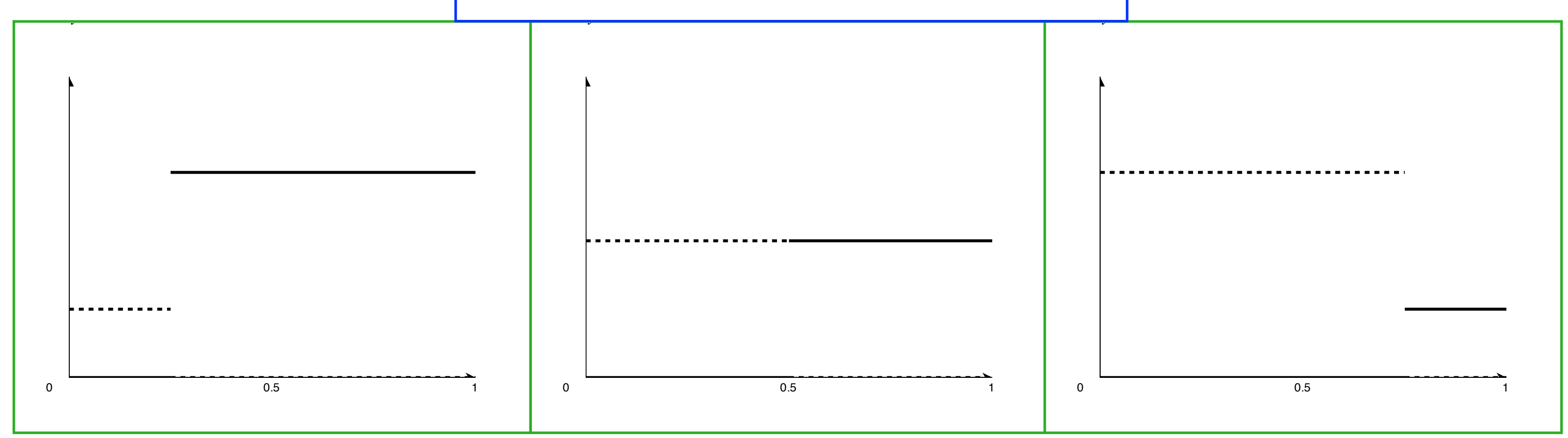
Example - Asymmetric Log Loss



$$\ell(y, \hat{\eta}) = \int_0^1 \boxed{\ell_c(y, \hat{\eta})}\,\boxed{w(c)}\,dc$$



$$w(c) = \frac{1}{c^2(1-c)}$$



Translating Weights

- The earlier connection between f-divergence and statistical information suggests that their weight functions are related

Primitives

$$\Delta\underline{\mathbb{L}} = \int_0^1 \Delta\underline{\mathbb{L}}_c \, w(c) \, dc \qquad \mathbb{I}_f = \int_0^1 \mathbb{I}_{f_\pi} \, \gamma(\pi) \, d\pi$$

Weights

Translating Weights

- The earlier connection between f-divergence and statistical information suggests that their weight functions are related
- Some straight-forward algebra gives and explicit translation
  - Dependence on prior π
  - Cubic term due to mapping from [0,∞) to [0,1]

Primitives

$$\Delta\underline{\mathbb{L}} = \int_0^1 \Delta\underline{\mathbb{L}}_c \, w(c)\, dc \qquad \mathbb{I}_f = \int_0^1 \mathbb{I}_{f_\pi} \, \gamma(\pi)\, d\pi$$

Weights

$$w_\pi(c) = \frac{\pi(1-\pi)}{\nu(\pi, c)^3}\gamma\left(\frac{(1-c)\pi}{\nu(\pi, c)}\right)$$

$$\nu(\pi, c) = (1-c)\pi + (1-\pi)c$$

Translating Weights

- The earlier connection between f-divergence and statistical information suggests that their weight functions are related
- Some straight-forward algebra gives and explicit translation
  - Dependence on prior π
  - Cubic term due to mapping from [0,∞) to [0,1]

Primitives

$$\Delta\underline{\mathbb{L}} = \int_0^1 \Delta\underline{\mathbb{L}}_c \, w(c) \, dc \qquad \mathbb{I}_f = \int_0^1 \mathbb{I}_{f_\pi} \, \gamma(\pi) \, d\pi$$

Weights

$$w_\pi(c) = \frac{\pi(1-\pi)}{\nu(\pi,c)^3}\gamma\left(\frac{(1-c)\pi}{\nu(\pi,c)}\right)$$

$$\gamma_\pi(c) = \frac{\pi^2(1-\pi)^2}{\nu(\pi,c)^3} w\left(\frac{(1-c)\pi}{\nu(\pi,c)}\right)$$

$$\nu(\pi,c) = (1-c)\pi + (1-\pi)c$$

Translating Weights

- The earlier connection between f-divergence and statistical information suggests that their weight functions are related
- Some straight-forward algebra gives and explicit translation
  - Dependence on prior π
  - Cubic term due to mapping from [0,∞) to [0,1]
- Cost-weighted loss relates to a prior-sensitive variational divergence

Primitives

$$\Delta\underline{\mathbb{L}} = \int_0^1 \Delta\underline{\mathbb{L}}_c \, w(c)\, dc \qquad \mathbb{I}_f = \int_0^1 \mathbb{I}_{f_\pi}\, \gamma(\pi)\, d\pi$$

Weights

$$w_\pi(c) = \frac{\pi(1-\pi)}{\nu(\pi,c)^3}\gamma\left(\frac{(1-c)\pi}{\nu(\pi,c)}\right)$$

$$\gamma_\pi(c) = \frac{\pi^2(1-\pi)^2}{\nu(\pi,c)^3} w\left(\frac{(1-c)\pi}{\nu(\pi,c)}\right)$$

$$\nu(\pi,c) = (1-c)\pi + (1-\pi)c$$

Graphical Representations

ROC Curves

- A threshold t is applied to a test statistic τ to create a statistical test
    - Contingency table for each test $\tau \geq t$

Predicted \ Actual	+	−
+	**TP**	**FP**
−	**FN**	**TN**

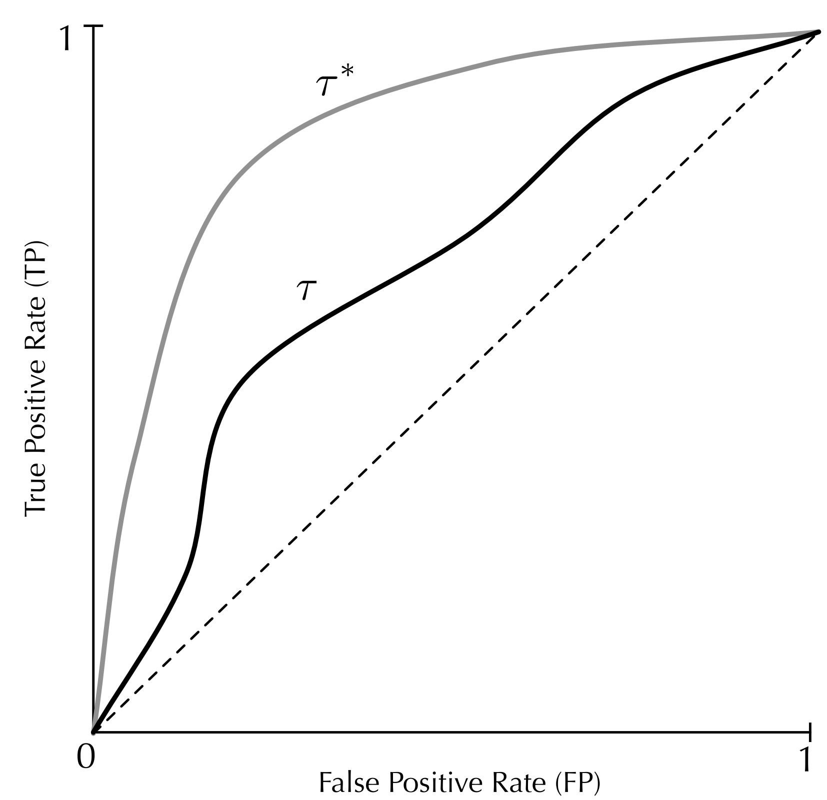
- Plotting

$$(TP, FP) = (P(\tau \geq t), Q(\tau \geq t))$$

as t varies gives an **ROC curve** for τ

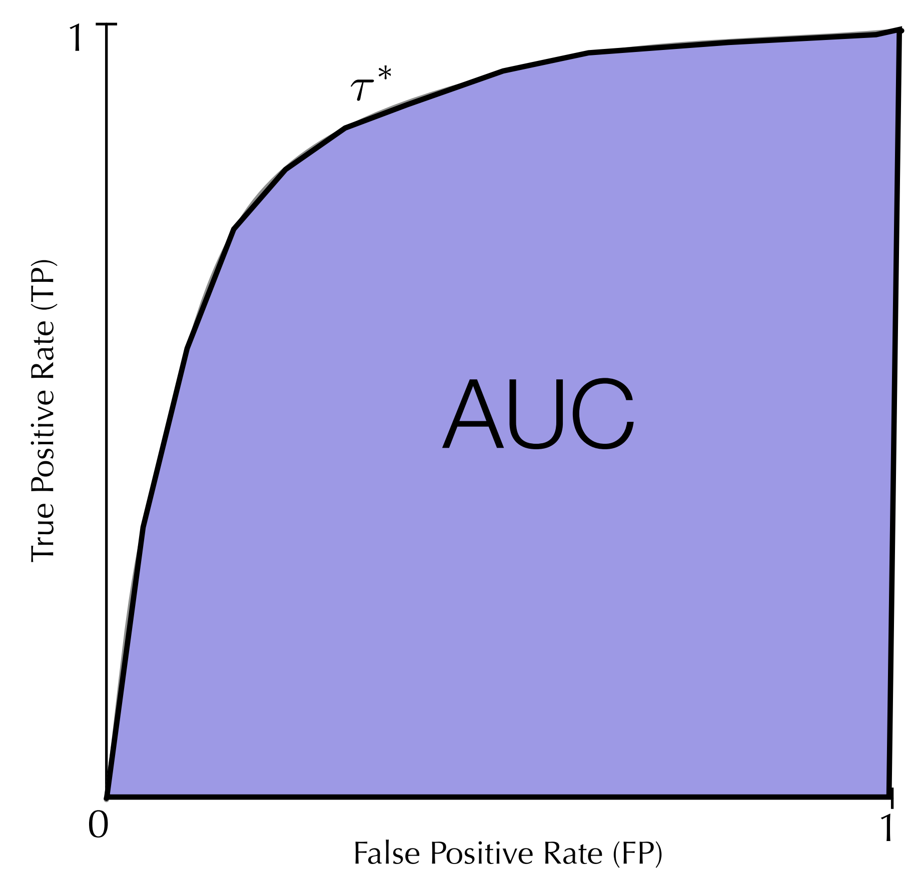
- NP Lemma implies that optimal ROC curve is obtained when

$$\tau^* = \frac{dP}{dQ}$$



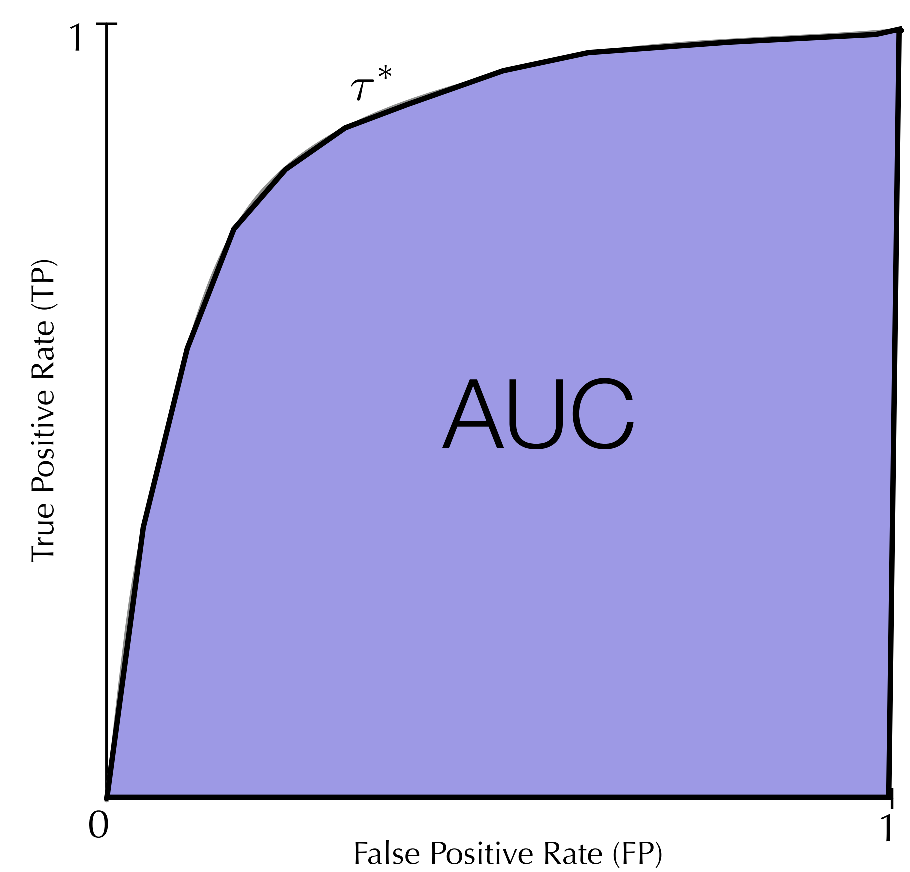
Area Under the ROC Curve (AUC)

- A natural measure of quality for a test statistic is the **area under the ROC curve**
- Ranking interpretation
  - Probability of misranking instance from Q ahead of one from P
  - Equivalent to the Mann-Whitney-Wilcoxon statistic



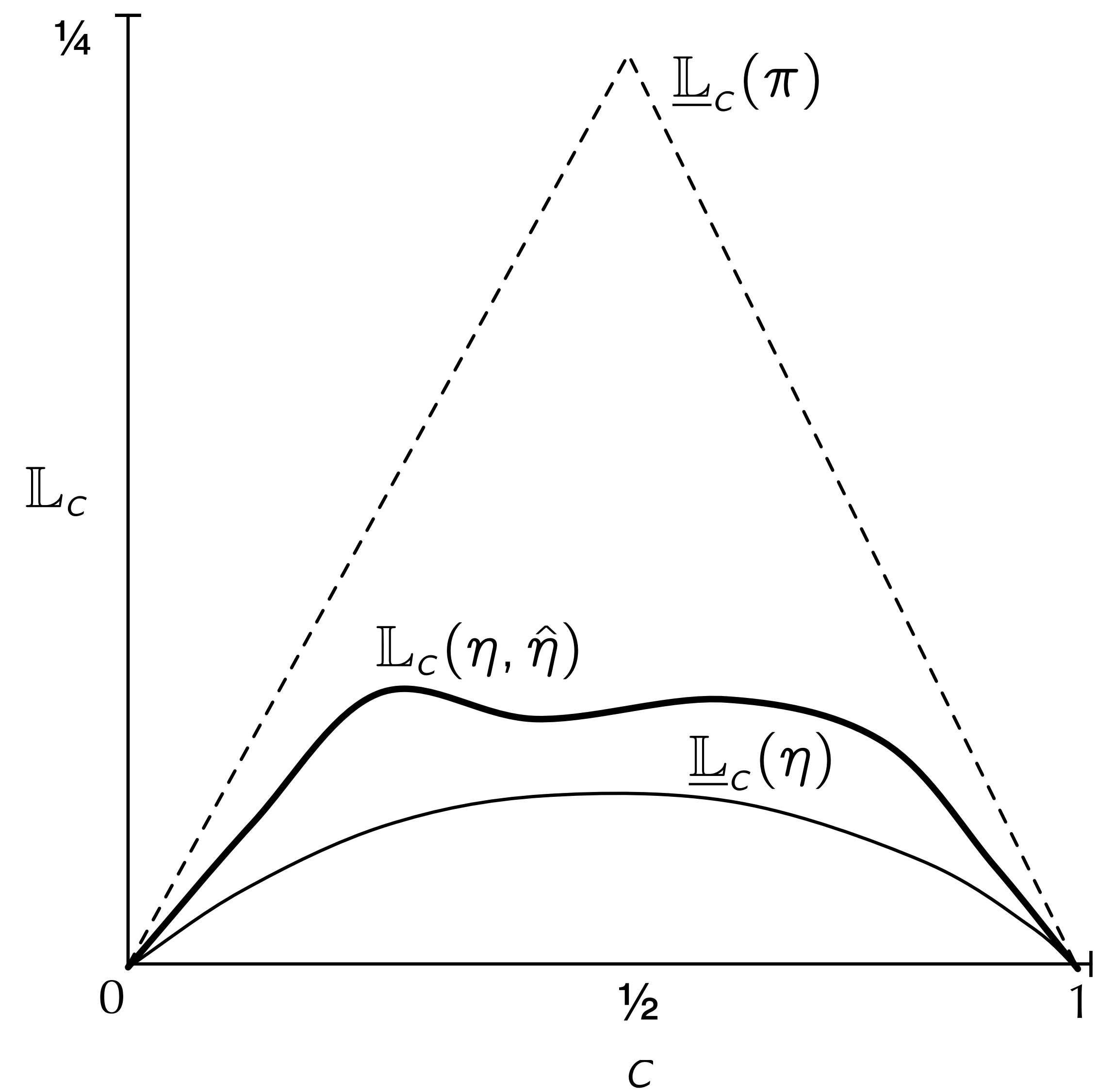
Area Under the ROC Curve (AUC)

- A natural measure of quality for a test statistic is the **area under the ROC curve**
- Ranking interpretation
  - Probability of misranking instance from Q ahead of one from P
  - Equivalent to the Mann-Whitney-Wilcoxon statistic
- Is maximal AUC an f-divergence?
  - No...
  - ...but it is *V*(*P*x*Q*, *Q*x*P*)



Risk Curves

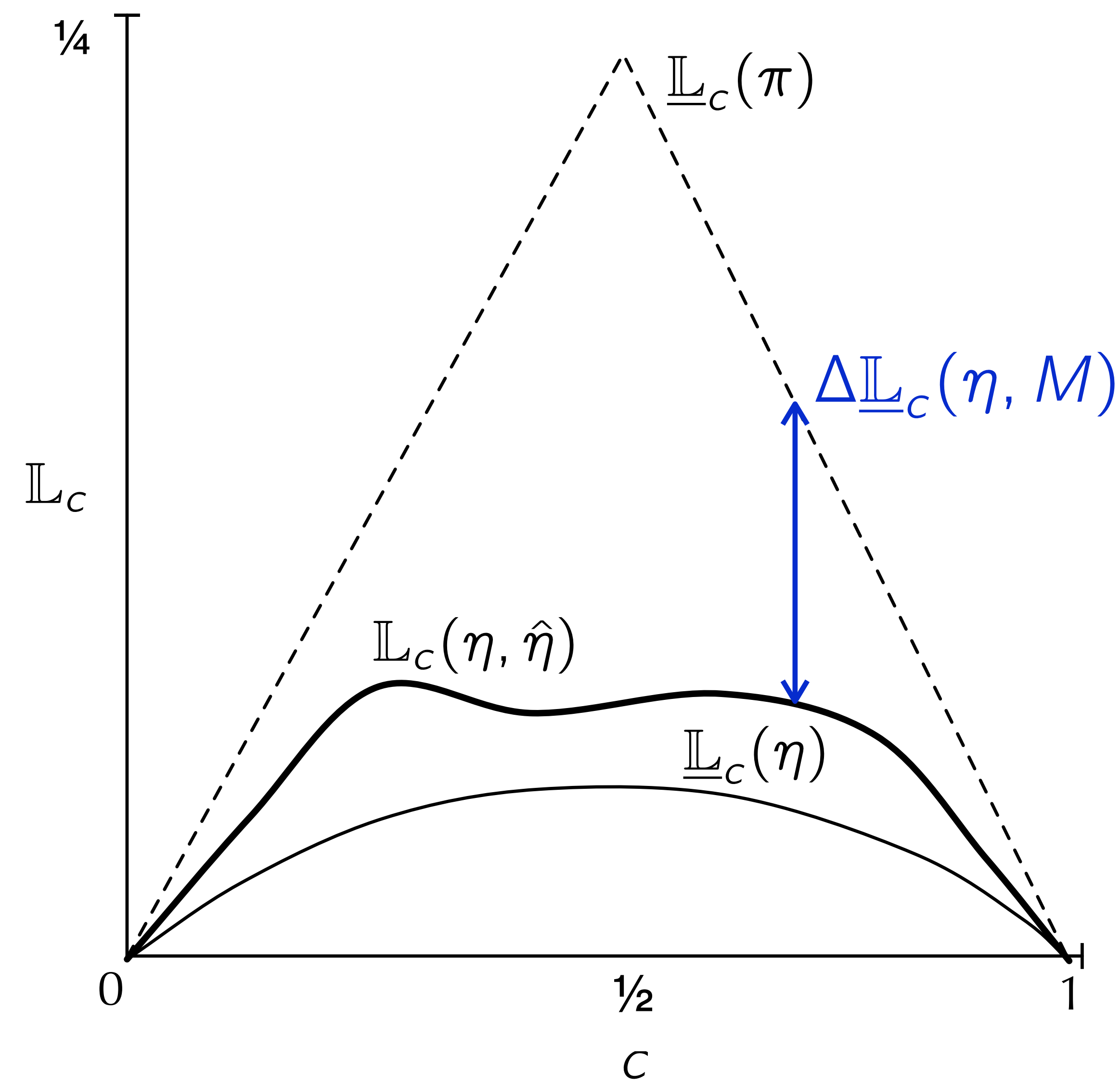
- A plot of cost-sensitive risk for each value of the cost parameter
  - Shape of curve dependent on mixing probability π



[Drummond & Holte, 2006]

Risk Curves

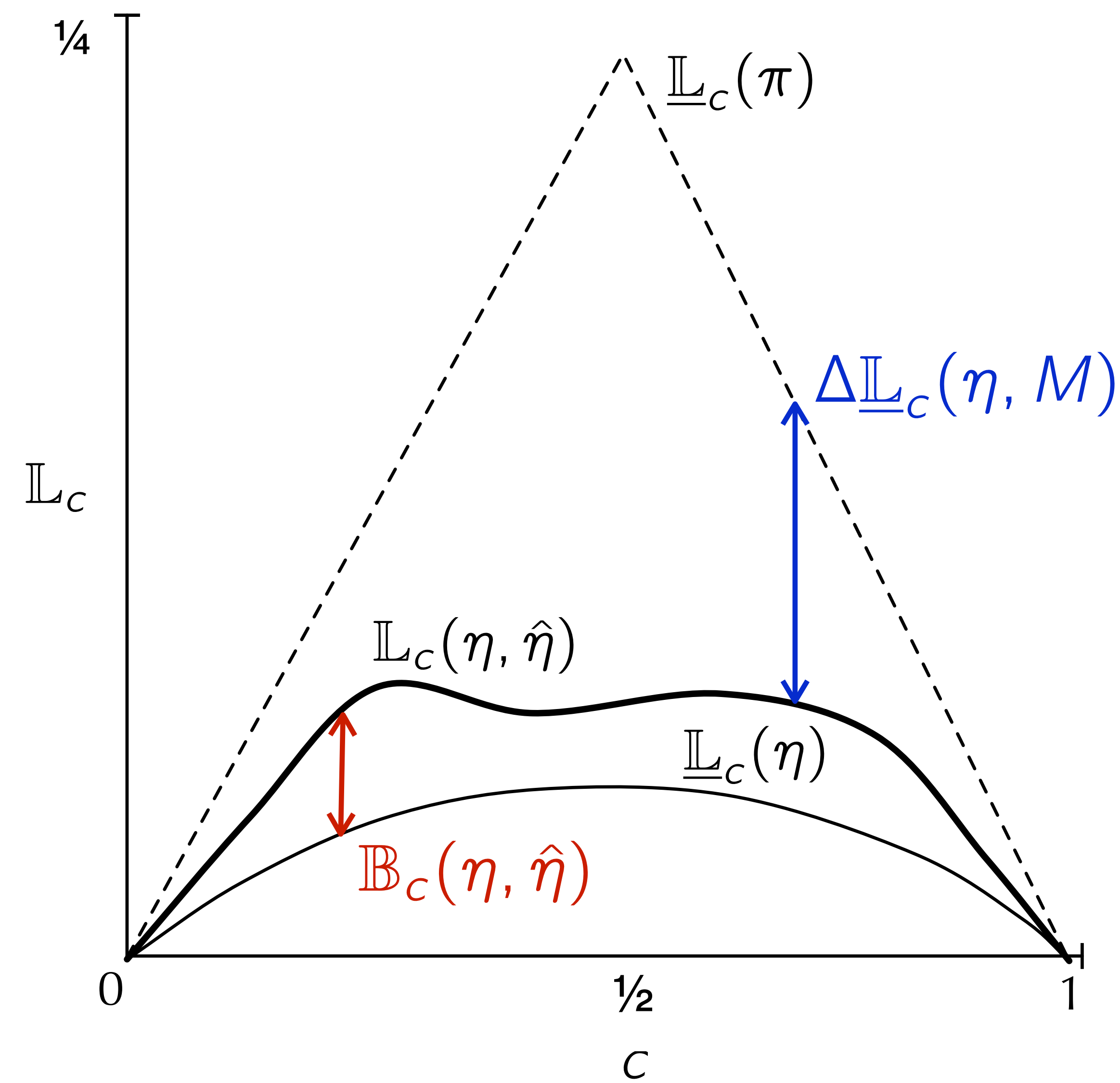
- A plot of cost-sensitive risk for each value of the cost parameter
  - Shape of curve dependent on mixing probability π
- Weighted area between bottom curve and "tent" is statistical information
  - Divergence bounds



[Drummond & Holte, 2006]

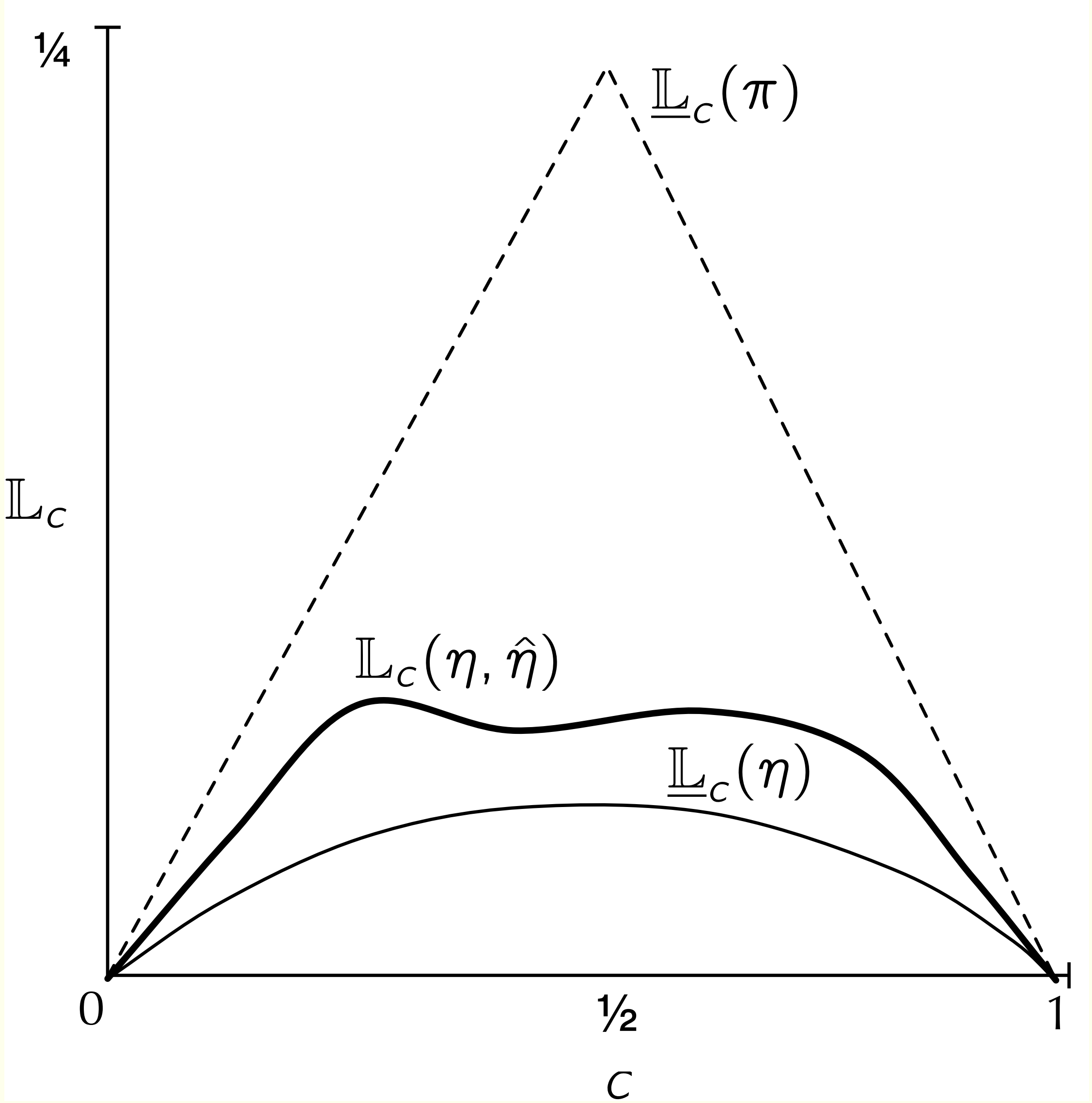
Risk Curves

- A plot of cost-sensitive risk for each value of the cost parameter
  - Shape of curve dependent on mixing probability π
- Weighted area between bottom curve and "tent" is statistical information
  - Divergence bounds
- Weighted area between two curves at bottom is regret
  - Surrogate loss bounds



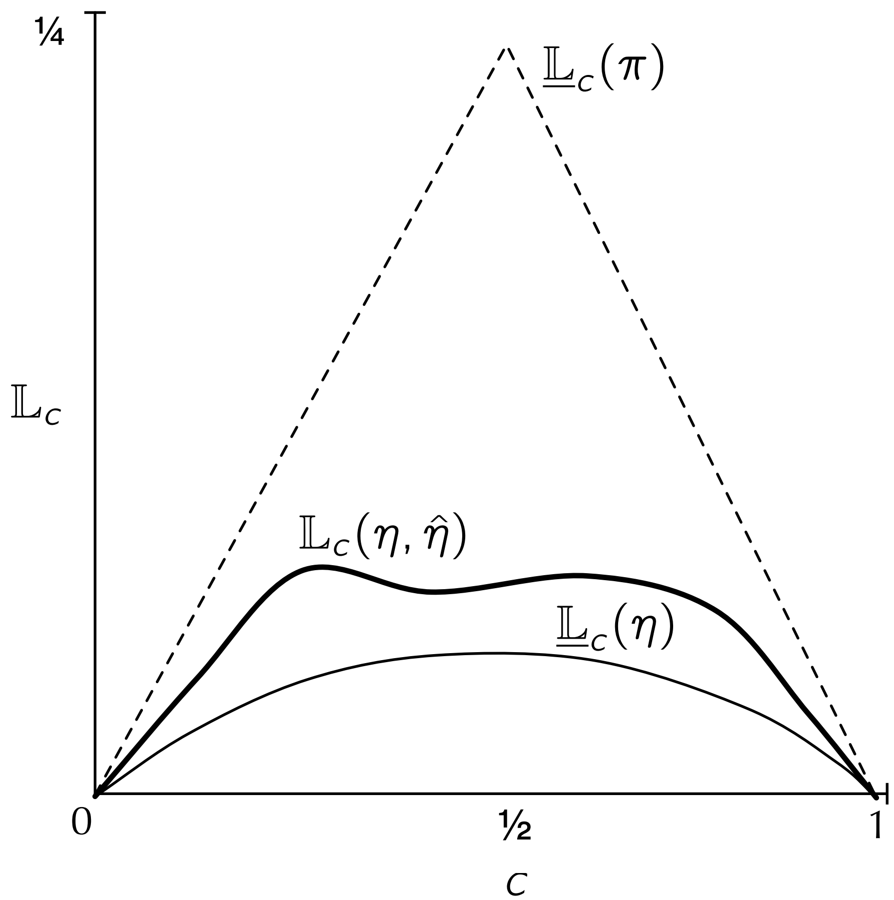
[Drummond & Holte, 2006]

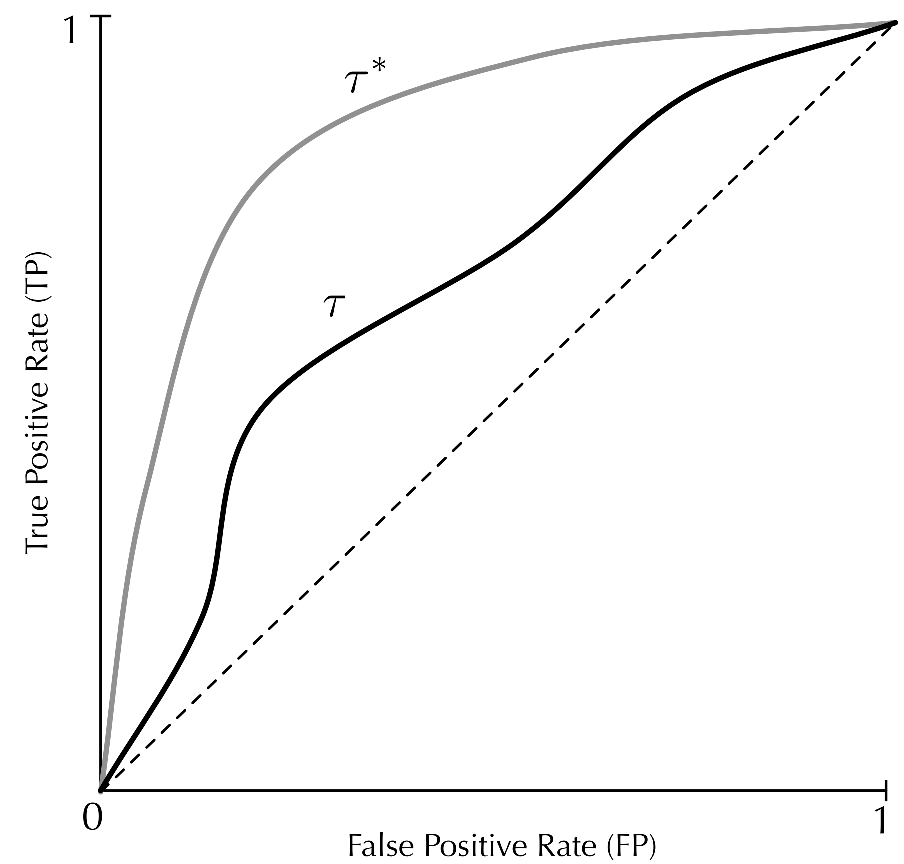
Risk Curves



¼
$\underline{\mathbb{L}}_c(\pi)$
$\mathbb{L}_c$
$\mathbb{L}_c(\eta, \hat{\eta})$
$\underline{\mathbb{L}}_c(\eta)$
0
½
1
c

ROC Curves to Risk Curves and Back





$$(FP, TP) \mapsto \mathbb{L}_c = (1-\pi)cFP + \pi(1-c)(1-TP)$$

$$(c, \mathbb{L}_c) \mapsto TP = \frac{(1-\pi)c}{(1-c)\pi}FP + \frac{(1-\pi)c - \mathbb{L}_c}{(1-c)\pi}$$

Variational Representations

Variational Form of f-Divergence

- Convex functions are invariant under the LF bidual

$$f(t) = f^{**}(t) = \sup_{t^* \in \mathbb{R}} \{t^*.t - f^*(t^*)\}$$

[Nguyen *et al.*, 2005]

Variational Form of f-Divergence

- Convex functions are invariant under the LF bidual

$$f(t) = f^{**}(t) = \sup_{t^* \in \mathbb{R}} \{t^*.t - f^*(t^*)\}$$

- Substitute into f-divergence definition

$$\begin{aligned}
\mathbb{I}_f(P, Q) &= \mathbb{E}_Q \left[\sup_{t^* \in \mathbb{R}} \left\{ t^* . \frac{dP}{dQ} - f^*(t^*) \right\} \right] \\
&= \int_{\mathcal{X}} \sup_{t^* \in \mathbb{R}} \{t^* dP - f^*(t^*) dQ\} \\
&= \sup_{r : \mathcal{X} \to \mathbb{R}} \int_{\mathcal{X}} r \, dP - f^*(r) \, dQ \\
&= \sup_{r : \mathcal{X} \to \mathbb{R}} \mathbb{E}_P[r] - \mathbb{E}_Q[f^*(r)]
\end{aligned}$$

[Nguyen *et al.*, 2005]

Variational Form of f-Divergence

- Convex functions are invariant under the LF bidual

$$f(t) = f^{**}(t) = \sup_{t^* \in \mathbb{R}} \{ t^*.t - f^*(t^*) \}$$

- Substitute into f-divergence definition

$$\begin{aligned}
\mathbb{I}_f(P, Q) &= \mathbb{E}_Q \left[\sup_{t^* \in \mathbb{R}} \left\{ t^* . \frac{dP}{dQ} - f^*(t^*) \right\} \right] \\
&= \int_{\mathcal{X}} \sup_{t^* \in \mathbb{R}} \{ t^* dP - f^*(t^*) dQ \} \\
&= \sup_{r:\mathcal{X} \to \mathbb{R}} \int_{\mathcal{X}} r \, dP - f^*(r) \, dQ \\
&= \sup_{r:\mathcal{X} \to \mathbb{R}} \mathbb{E}_P[r] - \mathbb{E}_Q[f^*(r)]
\end{aligned}$$

- Variational form does not use *dP/dQ*
    - Easier estimation

[Nguyen *et al.*, 2005]

Variational Representation of f-Divergence

$$\mathbb{I}_f(P, Q) \quad = \quad \sup_{r:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[r] - \mathbb{E}_Q[f^*(r)]$$

[Nguyen *et al.*, 2005]

The acts of the mind, wherein it exerts its power over simple ideas, are chiefly these three:

1. **Combining** several **simple ideas into one compound one**, and thus all complex ideas are made.

2. The second is **bringing two ideas**, whether simple or complex, **together**, and setting them by one another **so as to take a view of them at once**, without uniting them into one, by which it gets all its ideas of relations.

3. The third is **separating** them **from all other ideas** that accompany them in their real existence: this is called abstraction, and thus all its general ideas are made.

John Locke (1632-1704)

Part III: Bounds and Applications

Terra Statistica

Terra Statistica

In our theories, we rightly search for unification, but real life is both complicated and short, and we make no mockery of honest adhockery.

I.J. Good (1916-)

Maximum Mean Discrepancy

Maximum Mean Discrepancy (MMD)

- A special case of the variational form of f-divergence is when $f(t) = |t - 1|$
  - Restriction to [-1,1] occurs due to form of $f^*(t)$
- Assume r is from the unit ball in a RKHS for the kernel k with feature map ϕ and define
- Easy test statistic to estimate since

[Gretton *et al.*, 2007]

Maximum Mean Discrepancy (MMD)

- A special case of the variational form of f-divergence is when $f(t) = |t - 1|$
  - Restriction to [-1,1] occurs due to form of $f^*(t)$
- Assume r is from the unit ball in a RKHS for the kernel k with feature map ϕ and define
- Easy test statistic to estimate since

$$V(P, Q) = \sup_{r:\mathcal{X}\to[-1,1]} \mathbb{E}_P[r] - \mathbb{E}_Q[r]$$

$$f^*(t) = \begin{cases} t & t \in [-1, 1] \\ +\infty & \text{otherwise} \end{cases}$$

[Gretton *et al.*, 2007]

Maximum Mean Discrepancy (MMD)

- A special case of the variational form of f-divergence is when *f*(*t*) = |*t* - 1|
    - Restriction to [-1,1] occurs due to form of *f**(*t*)

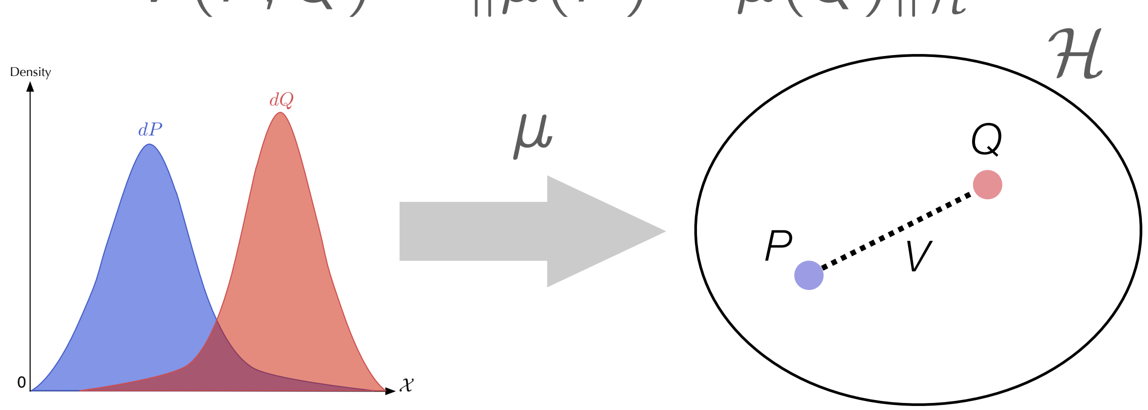
$$V(P,Q) = \sup_{r:\mathcal{X}\to[-1,1]} \mathbb{E}_P[r] - \mathbb{E}_Q[r]$$

$$f^*(t) = \begin{cases} t & t \in [-1,1] \\ +\infty & \text{otherwise} \end{cases}$$

- Assume *r* is from the unit ball in a RKHS for the kernel *k* with feature map ϕ and define

$$\mu[P] := \mathbb{E}_P[\phi(x)] = \mathbb{E}_P[k(x,\cdot)]$$

$$V(P,Q) = \|\mu(P) - \mu(Q)\|_{\mathcal{H}}$$



[Gretton *et al.*, 2007]

Maximum Mean Discrepancy (MMD)

- A special case of the variational form of f-divergence is when $f(t) = |t - 1|$

$$V(P,Q) = \sup_{r:\mathcal{X}\to[-1,1]} \mathbb{E}_P[r] - \mathbb{E}_Q[r]$$

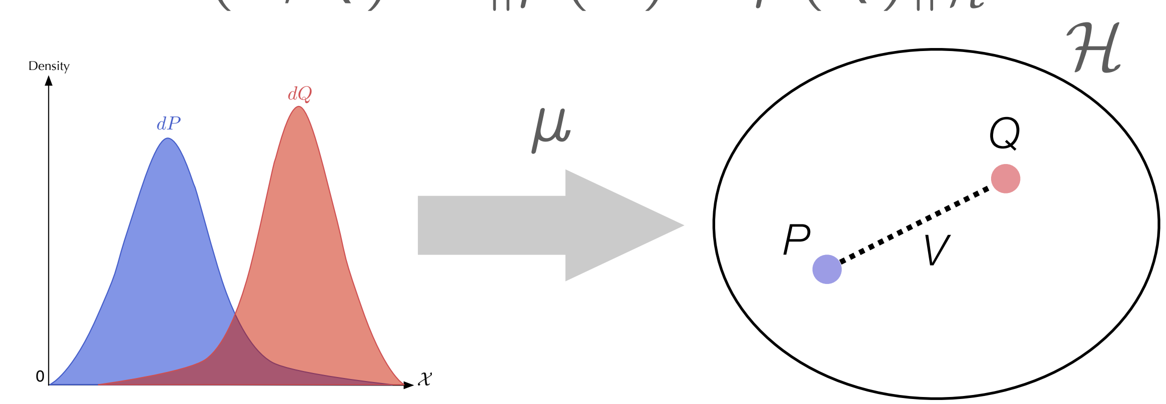
  - Restriction to [-1,1] occurs due to form of $f^*(t)$

$$f^*(t) = \begin{cases} t & t \in [-1, 1] \\ +\infty & \text{otherwise} \end{cases}$$

- Assume r is from the unit ball in a RKHS for the kernel k with feature map ϕ and define

$$\mu[P] := \mathbb{E}_P[\phi(x)] = \mathbb{E}_P[k(x,\cdot)]$$

$$V(P,Q) = \|\mu(P) - \mu(Q)\|_{\mathcal{H}}$$



- Easy test statistic to estimate since

$$\|\mu(P) - \mu(Q)\|_{\mathcal{H}} = \mathbb{E}_{P\times P} k(x,x') + \mathbb{E}_{Q\times Q} k(y,y') - 2\mathbb{E}_{P\times Q} k(x,y)$$

$$\approx \frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i,x_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i,y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i,y_j)$$

[Gretton *et al.*, 2007]

Generalised Pinsker Bounds

Pinsker's Inequality

Pinsker's Inequality

- A lower bound on KL divergence in terms of variational divergence

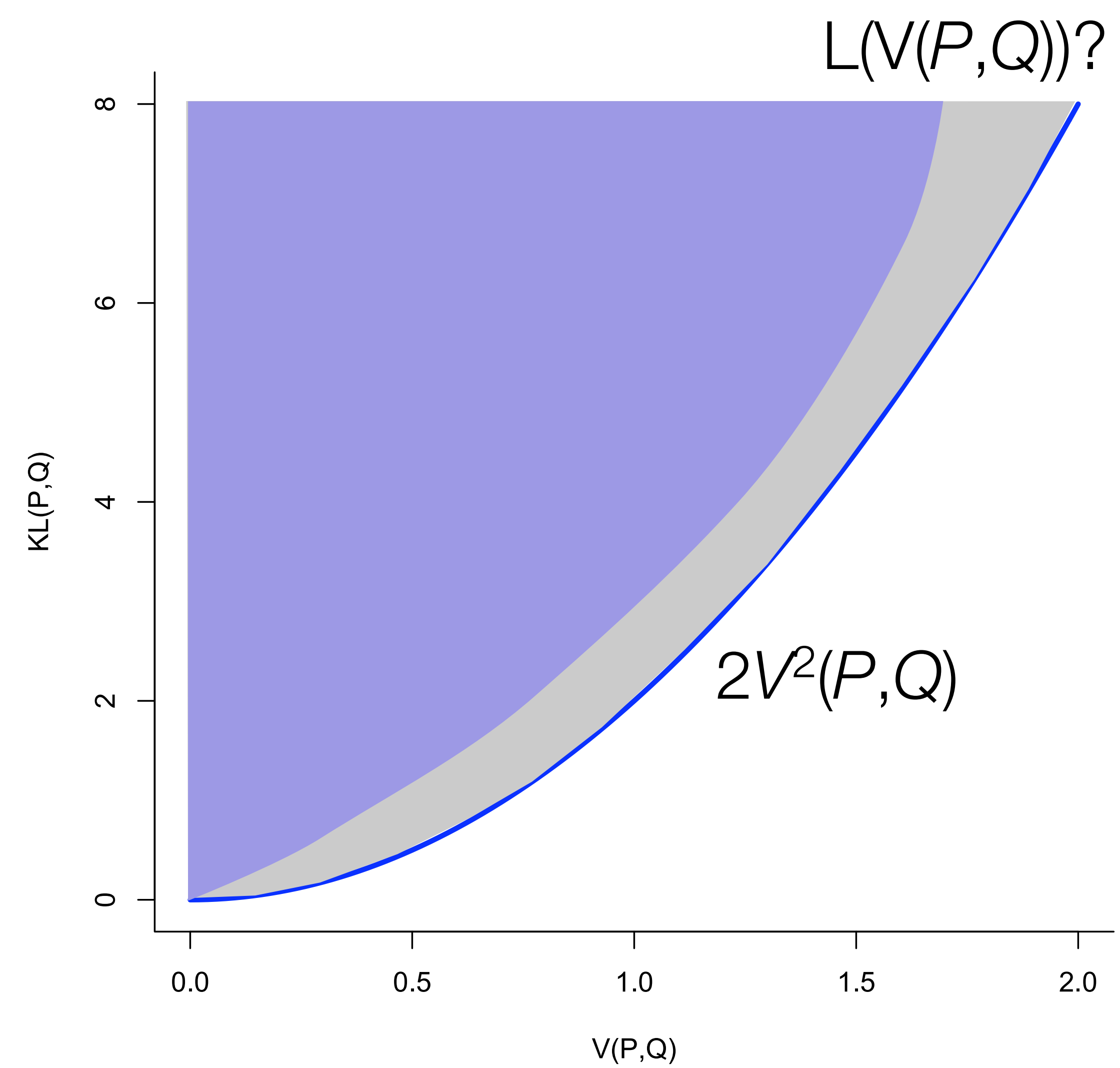
$$KL(P,Q) \geq 2V^2(P,Q)$$

- Information about the value of V constraints the possible values of KL

Better Pinsker Bounds

- The above inequality is not tight
- What we really want is

$$L(V) = \inf_{V(P,Q)=V} KL(P,Q)$$



Generalised Pinsker Inequalities

Primitive vs Composite

- V is "primitive"
- KL is "composite"

General Bound

- Can we get tight bounds for **any** f-divergence given V?
  - Yes we can!
- V gives "partial information" about separation of P and Q

Generalised Pinsker Inequalities

Primitive vs Composite

- V is "primitive"
- KL is "composite"

General Bound

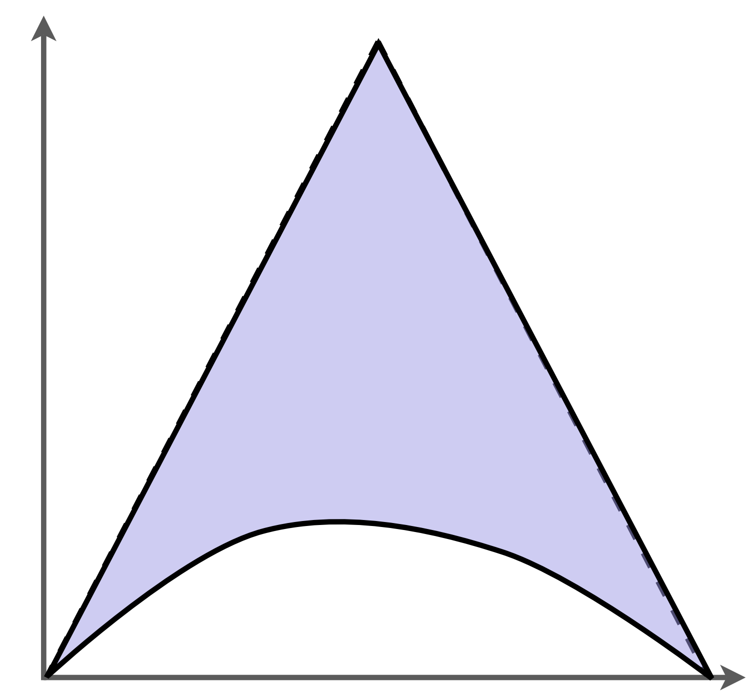
- Can we get tight bounds for **any** f-divergence given V?
  - Yes we can!
- V gives "partial information" about separation of P and Q

Divergence	**Variational Bound**
Hellinger	$h^2 \geq 2 - \sqrt{4 - V^2}$
Jeffreys	$J \geq 2V \ln\left(\frac{2+V}{2-V}\right)$
Symmetric χ^2	$\Psi \geq \frac{8V^2}{4-V^2}$
AG Mean	$T \geq \ln\left(\frac{4}{\sqrt{4-V^2}}\right) - \ln 2$
Pearson χ^2	$\chi^2 \geq \begin{cases} V^2 & V < 1 \\ \frac{V}{2-V} & V \geq 1 \end{cases}$

Generalised Pinsker Inequalities

Proof Sketch

- f-divergence is a weighted sum of primitive statistical information
  - This is just an area on a risk diagram
- Value at one point bounds the total area



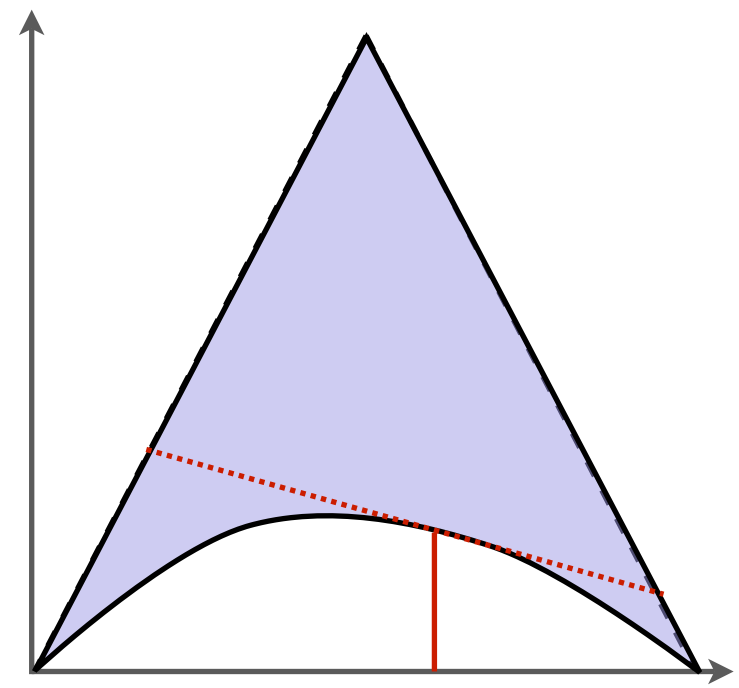
Going Further

- This proof is amenable to knowing multiple primitive values

Generalised Pinsker Inequalities

Proof Sketch

- f-divergence is a weighted sum of primitive statistical information
  - This is just an area on a risk diagram
- Value at one point bounds the total area



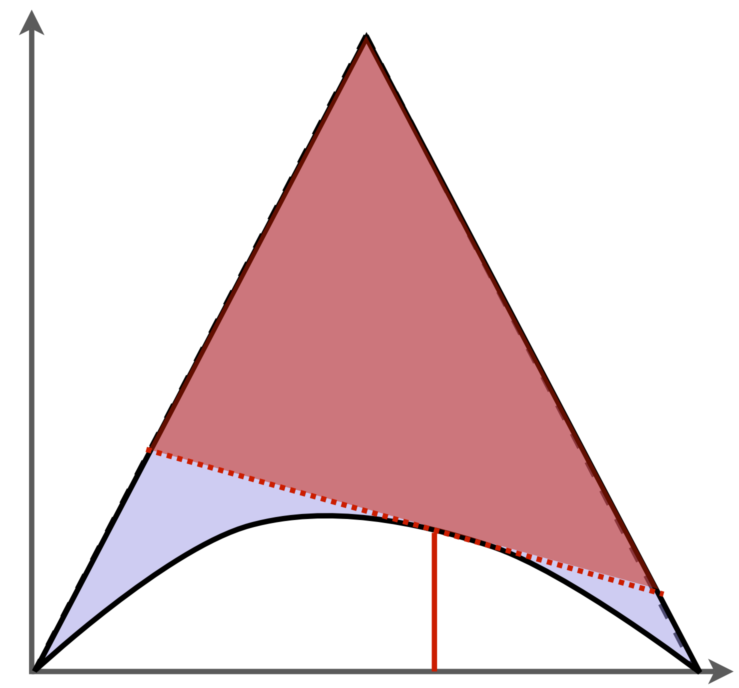
Going Further

- This proof is amenable to knowing multiple primitive values

Generalised Pinsker Inequalities

Proof Sketch

- f-divergence is a weighted sum of primitive statistical information
  - This is just an area on a risk diagram
- Value at one point bounds the total area



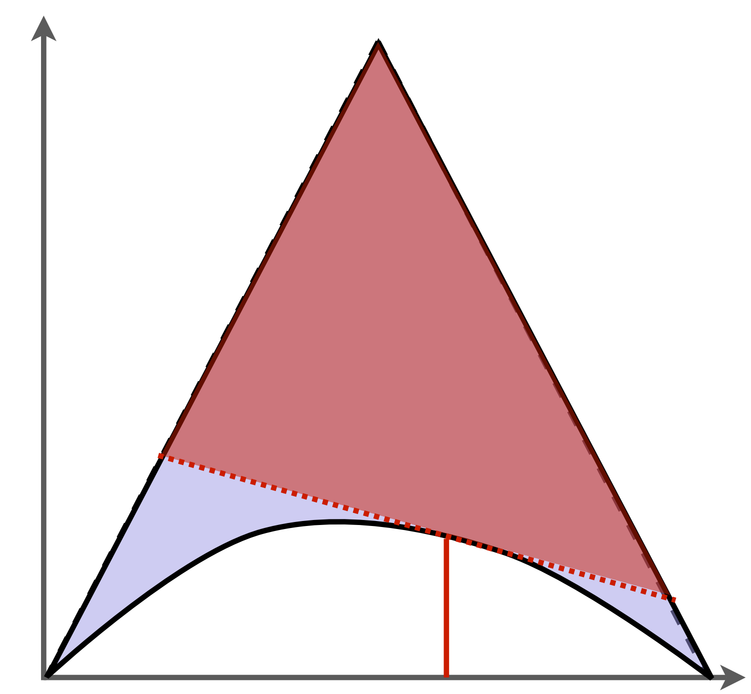
Going Further

- This proof is amenable to knowing multiple primitive values

Generalised Pinsker Inequalities

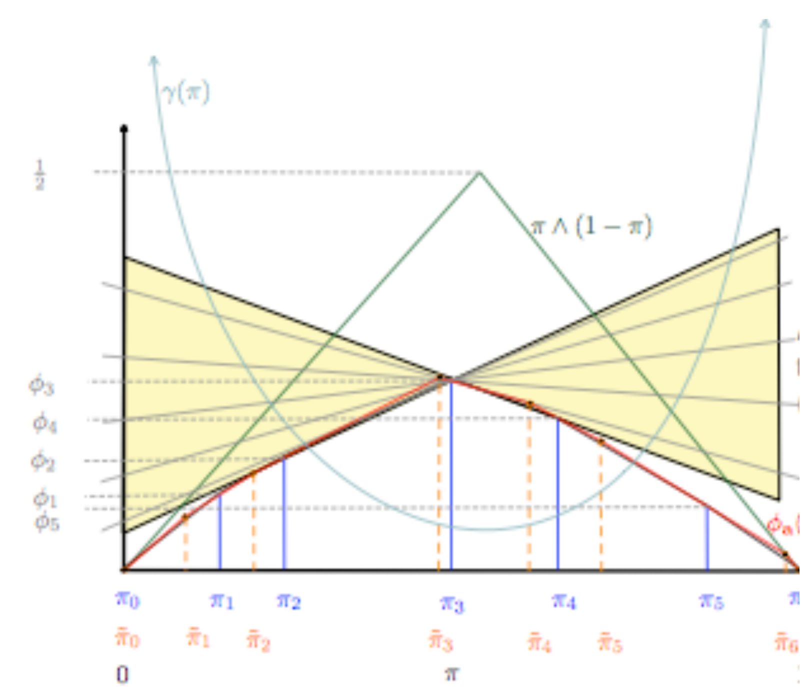
Proof Sketch

- f-divergence is a weighted sum of primitive statistical information
    - This is just an area on a risk diagram
- Value at one point bounds the total area



Going Further

- This proof is amenable to knowing multiple primitive values

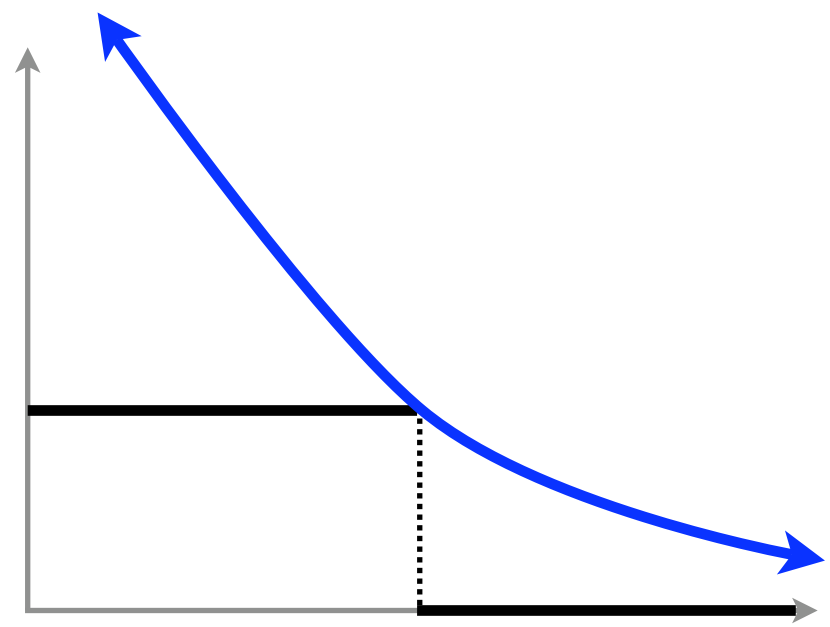


Surrogate Loss Bounds

Surrogate Loss

Surrogate Loss

- 0-1 loss is notoriously hard to optimise directly
- One solution is to optimise a **surrogate** - an upper bound on 0-1 loss



Surrogate Bounds

- Want guarantees that minimising the surrogate regret minimises the 0-1 regret

[Bartlett *et al.*, 2006]

Surrogate Loss Bounds

Main Result

- Suppose we know $B_{c_0}(\eta, \hat{\eta}) = \alpha$. Then for an arbitrary proper loss, its regret satisfies

$$B(\eta, \hat{\eta}) \geq \min(\psi(c_0, \alpha), \psi(c_0, -\alpha))$$

where $\psi(c_0, \alpha) = \underline{L}(c_0) - \underline{L}(c_0 - \alpha) + \alpha \underline{L}'(c_0)$

Surrogate Loss Bounds

Main Result

- Suppose we know $B_{c_0}(\eta, \hat{\eta}) = \alpha$. Then for an arbitrary proper loss, its regret satisfies

$$B(\eta, \hat{\eta}) \geq \min(\psi(c_0, \alpha), \psi(c_0, -\alpha))$$

where $\psi(c_0, \alpha) = \underline{L}(c_0) - \underline{L}(c_0 - \alpha) + \alpha \underline{L}'(c_0)$

Corollary

- For a symmetric loss where $\underline{L}(c-½) = \underline{L}(½-c)$, then if $B_{\frac{1}{2}}(\eta, \hat{\eta}) = \alpha$

$$B(\eta, \hat{\eta}) \geq \underline{L}(1/2) - \underline{L}(1/2 - \alpha)$$

Surrogate Bound Example

Exponential Loss

- Let $\ell(y, \hat{\eta}) = \begin{cases} \sqrt{\frac{\hat{\eta}}{1-\hat{\eta}}} & y = 0 \\ \sqrt{\frac{1-\hat{\eta}}{\hat{\eta}}} & y = 1 \end{cases}$

[Bartlett *et al.*, 2006]

Surrogate Bound Example

Exponential Loss

- Let $\ell(y, \hat{\eta}) = \begin{cases} \sqrt{\frac{\hat{\eta}}{1-\hat{\eta}}} & y = 0 \\ \sqrt{\frac{1-\hat{\eta}}{\hat{\eta}}} & y = 1 \end{cases}$

[Bartlett *et al.*, 2006]

Surrogate Bound Example

Exponential Loss

- Let $\ell(y, \hat{\eta}) = \begin{cases} \sqrt{\frac{\hat{\eta}}{1-\hat{\eta}}} & y = 0 \\ \sqrt{\frac{1-\hat{\eta}}{\hat{\eta}}} & y = 1 \end{cases}$
- Then $L(\eta, \hat{\eta}) = (1-\eta)\sqrt{\dfrac{\hat{\eta}}{1-\hat{\eta}}} + \eta\sqrt{\dfrac{1-\hat{\eta}}{\hat{\eta}}}$

[Bartlett *et al.*, 2006]

Surrogate Bound Example

Exponential Loss

- Let $\ell(y, \hat{\eta}) = \begin{cases} \sqrt{\frac{\hat{\eta}}{1-\hat{\eta}}} & y = 0 \\ \sqrt{\frac{1-\hat{\eta}}{\hat{\eta}}} & y = 1 \end{cases}$
- Then $L(\eta, \hat{\eta}) = (1 - \eta)\sqrt{\dfrac{\hat{\eta}}{1 - \hat{\eta}}} + \eta\sqrt{\dfrac{1 - \hat{\eta}}{\hat{\eta}}}$
- And so $\underline{L}(\eta) = 2\sqrt{\eta(1 - \eta)}$ which is symmetric

[Bartlett *et al.*, 2006]

Surrogate Bound Example

Exponential Loss

- Let $\ell(y, \hat{\eta}) = \begin{cases} \sqrt{\frac{\hat{\eta}}{1-\hat{\eta}}} & y = 0 \\ \sqrt{\frac{1-\hat{\eta}}{\hat{\eta}}} & y = 1 \end{cases}$
- Then $L(\eta, \hat{\eta}) = (1-\eta)\sqrt{\dfrac{\hat{\eta}}{1-\hat{\eta}}} + \eta\sqrt{\dfrac{1-\hat{\eta}}{\hat{\eta}}}$
- And so $\underline{L}(\eta) = 2\sqrt{\eta(1-\eta)}$ which is symmetric
- Thus, if $B_{\frac{1}{2}}(\eta, \hat{\eta}) = \alpha$ then the exponential regret satisfies

$$B(\eta, \hat{\eta}) \geq 1 - \sqrt{1 - 4\alpha^2}$$

[Bartlett *et al.*, 2006]

Surrogate Bound Example

Exponential Loss

- Let $\ell(y, \hat{\eta}) = \begin{cases} \sqrt{\frac{\hat{\eta}}{1-\hat{\eta}}} & y = 0 \\ \sqrt{\frac{1-\hat{\eta}}{\hat{\eta}}} & y = 1 \end{cases}$
- Then $L(\eta, \hat{\eta}) = (1-\eta)\sqrt{\dfrac{\hat{\eta}}{1-\hat{\eta}}} + \eta\sqrt{\dfrac{1-\hat{\eta}}{\hat{\eta}}}$
- And so $\underline{L}(\eta) = 2\sqrt{\eta(1-\eta)}$ which is symmetric
- Thus, if $B_{\frac{1}{2}}(\eta, \hat{\eta}) = \alpha$ then the exponential regret satisfies

$$B(\eta, \hat{\eta}) \geq 1 - \sqrt{1 - 4\alpha^2}$$

- And so

$$B_{\frac{1}{2}}(\eta, \hat{\eta}) \leq \frac{1}{2}\sqrt{(1 - B(\eta, \hat{\eta}))^2 - 1}$$

[Bartlett *et al.*, 2006]

Proof of Surrogate Loss Bound

- First recall that $B_{c_0}(\eta, \hat{\eta}) = |\eta - c_0| [\![\min(\eta, \hat{\eta}) \leq c_0 < \max(\eta, \hat{\eta})]\!]$

Proof of Surrogate Loss Bound

- First recall that $B_{c_0}(\eta, \hat{\eta}) = |\eta - c_0| [\![\min(\eta, \hat{\eta}) \leq c_0 < \max(\eta, \hat{\eta})]\!]$
- And so when $B_{c_0}(\eta, \hat{\eta}) = \alpha$ we know that

$$\eta = \begin{cases} c_0 + \alpha, & \hat{\eta} \leq c_0 < \eta \\ c_0 - \alpha, & \eta \leq c_0 < \hat{\eta} \end{cases}$$

Proof of Surrogate Loss Bound

- First recall that $B_{c_0}(\eta, \hat{\eta}) = |\eta - c_0| [\![\min(\eta, \hat{\eta}) \leq c_0 < \max(\eta, \hat{\eta})]\!]$
- And so when $B_{c_0}(\eta, \hat{\eta}) = \alpha$ we know that

$$\eta = \begin{cases} c_0 + \alpha, & \hat{\eta} \leq c_0 < \eta \\ c_0 - \alpha, & \eta \leq c_0 < \hat{\eta} \end{cases}$$

- For a general proper loss, recall its regret can be expressed as

$$B(\eta, \hat{\eta}) = \int_{\min(\eta,\hat{\eta})}^{\max(\eta,\hat{\eta})} |\eta - c| \, w(c) \, dc$$

Proof of Surrogate Loss Bound

- First recall that $B_{c_0}(\eta, \hat{\eta}) = |\eta - c_0| [\![\min(\eta, \hat{\eta}) \le c_0 < \max(\eta, \hat{\eta})]\!]$
- And so when $B_{c_0}(\eta, \hat{\eta}) = \alpha$ we know that

$$\eta = \begin{cases} c_0 + \alpha, & \hat{\eta} \le c_0 < \eta \\ c_0 - \alpha, & \eta \le c_0 < \hat{\eta} \end{cases}$$

- For a general proper loss, recall its regret can be expressed as

$$B(\eta, \hat{\eta}) = \int_{\min(\eta,\hat{\eta})}^{\max(\eta,\hat{\eta})} |\eta - c|\, w(c)\, dc$$

- In the first case, when $\hat{\eta} \le c_0 < \eta = c_0 + \alpha$ we see

$$\begin{aligned} B(\eta, \hat{\eta}) &= \int_{\hat{\eta}}^{\eta} (c_0 + \alpha - c)\, w(c)\, dc \\ &\ge \int_{c_0}^{c_0+\alpha} (c_0 + \alpha - c)\, w(c)\, dc \end{aligned}$$

Proof of Surrogate Loss Bound (continued)

- Thus, using $w(c) = -\underline{L}''(c)$, and integrating by parts, we see

$$
\begin{aligned}
B(\eta, \hat{\eta}) &\geq \int_{c_0}^{c_0+\alpha} (c_0 + \alpha - c)\, w(c)\, dc \\
&= -\int_{c_0}^{c_0+\alpha} (c_0 + \alpha - c)\, \underline{L}''(c)\, dc \\
&= -[(c_0 + \alpha - c)\underline{L}'(c)]_{c_0}^{c_0+\alpha} - \int_{c_0}^{c_0+\alpha} \underline{L}'(c)\, dc \\
&= \alpha \underline{L}'(c_0) - \underline{L}(c_0 + \alpha) + \underline{L}(c_0)
\end{aligned}
$$

- The case when $c_0 - \alpha = \eta \leq c_0 < \hat{\eta}$ is almost identical

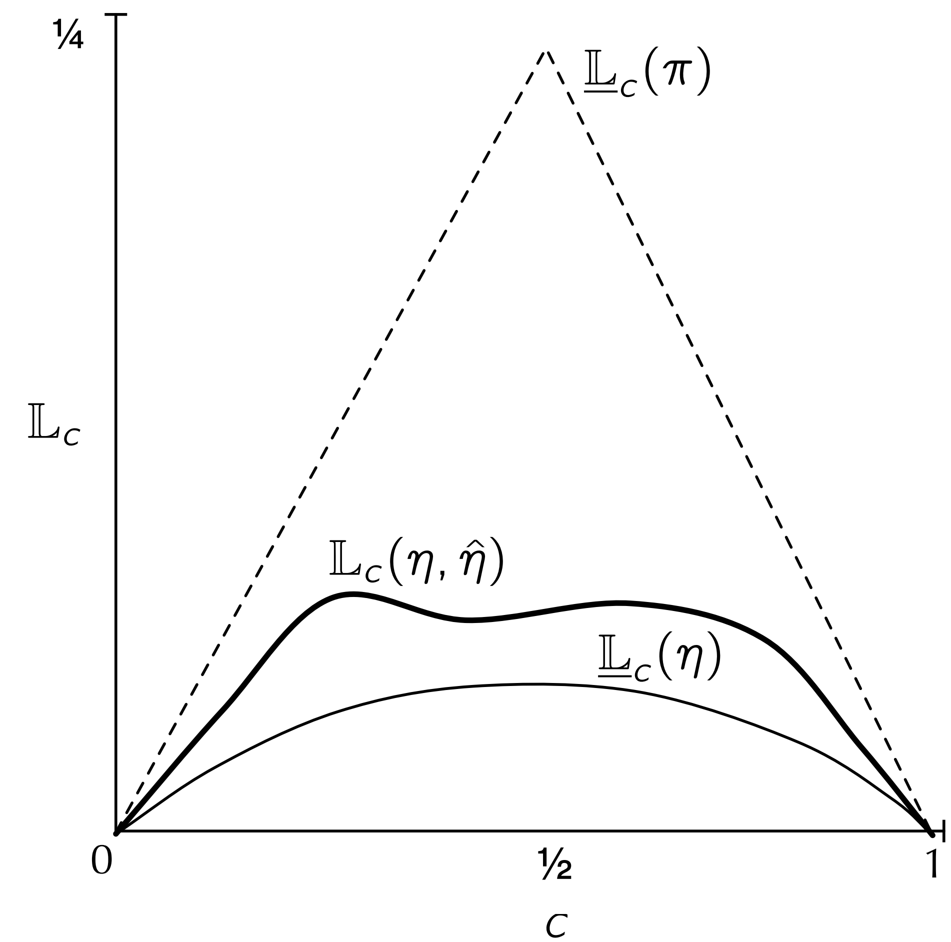
It is the snobbishness of the young to suppose that a theorem is trivial because the proof is trivial

Henry Whitehead (1904-1960)

f-Divergence Estimation

f-Divergence Estimation

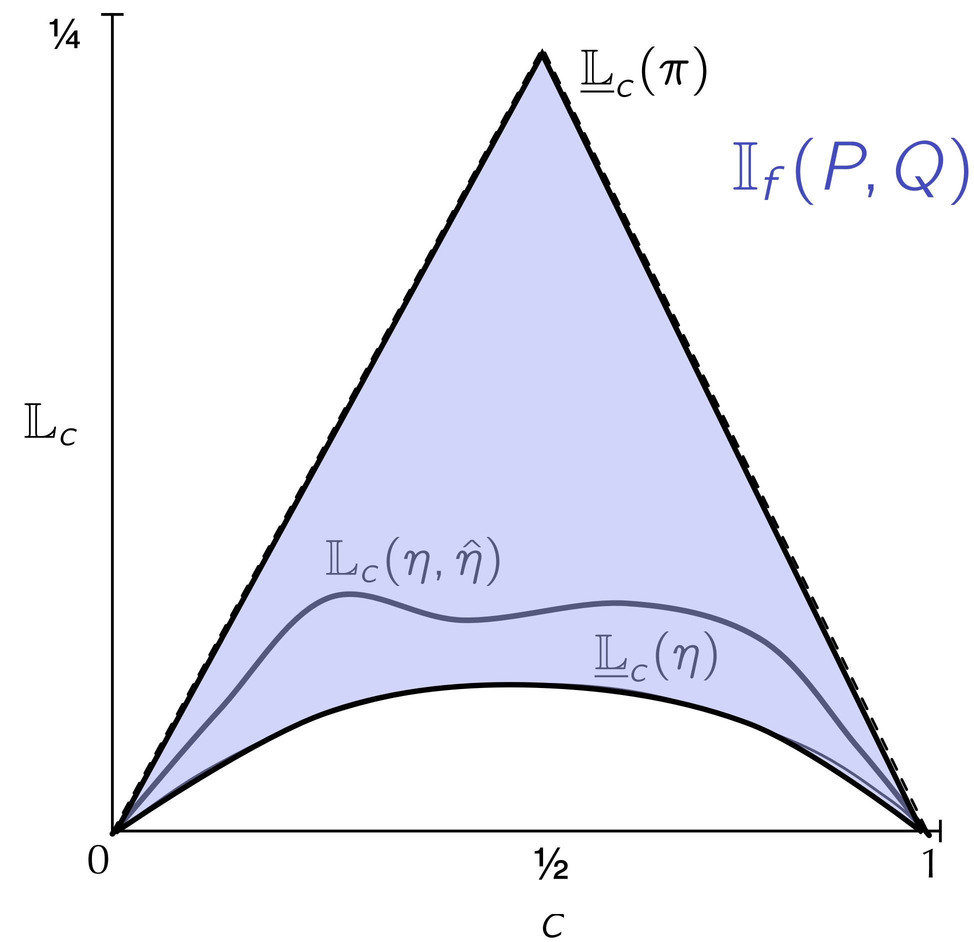
f-Divergence and Bayes Risk



f-Divergence Estimation

f-Divergence and Bayes Risk

- Recall that $\mathbb{I}_f(P, Q) = \underline{\mathbb{L}}(\pi, M) - \underline{\mathbb{L}}(\eta, M)$

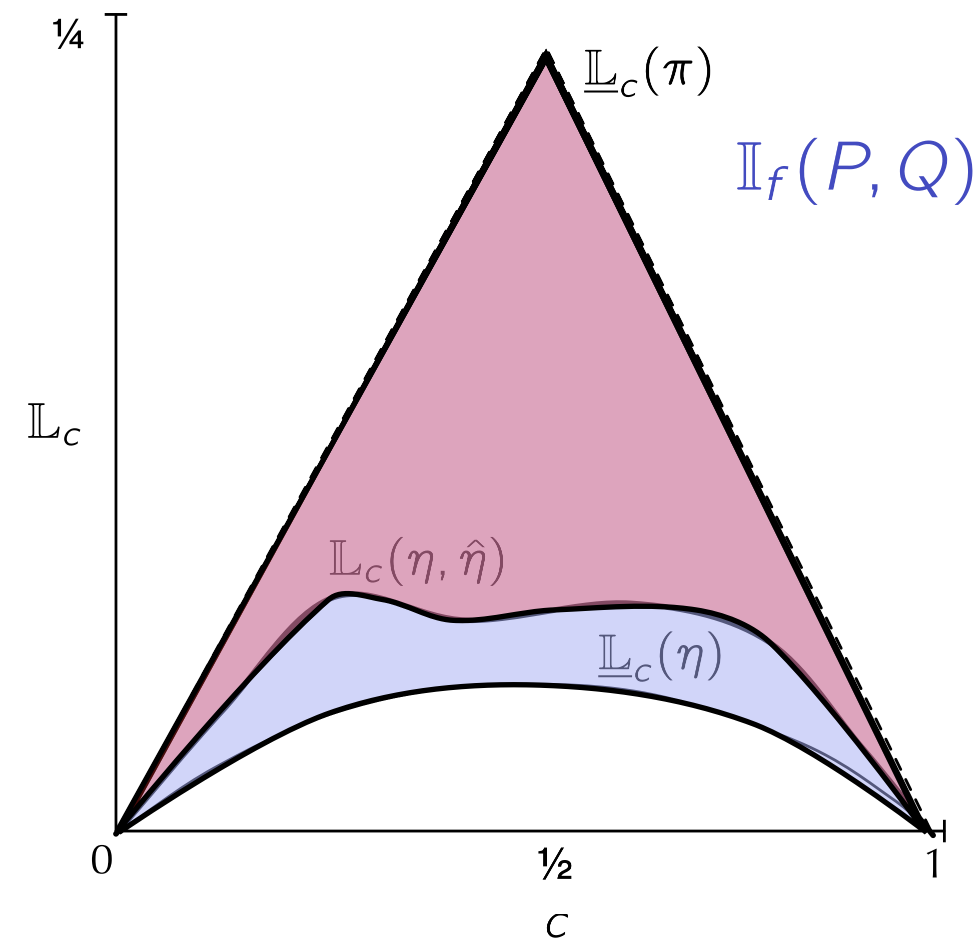


f-Divergence Estimation

f-Divergence and Bayes Risk

- Recall that $\mathbb{I}_f(P, Q) = \underline{\mathbb{L}}(\pi, M) - \underline{\mathbb{L}}(\eta, M)$
- For good estimators $\mathbb{L}(\eta, \hat{\eta}, M) \approx \underline{\mathbb{L}}(\eta, M)$ and so

$$\mathbb{I}_f(P, Q) \approx K - \mathbb{L}(\eta, \hat{\eta}, M)$$



f-Divergence Estimation

f-Divergence and Bayes Risk

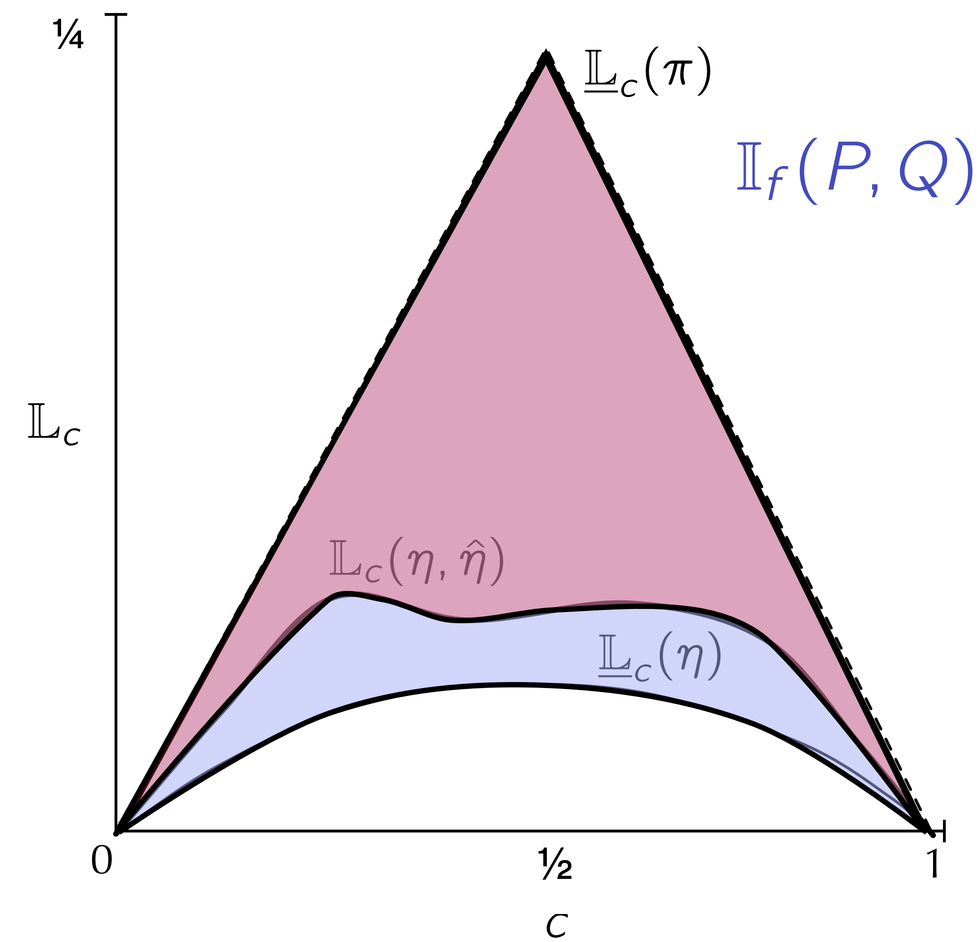
- Recall that $\mathbb{I}_f(P,Q) = \underline{\mathbb{L}}(\pi, M) - \underline{\mathbb{L}}(\eta, M)$
- For good estimators $\mathbb{L}(\eta, \hat{\eta}, M) \approx \underline{\mathbb{L}}(\eta, M)$ and so

$$\mathbb{I}_f(P,Q) \approx K - \mathbb{L}(\eta, \hat{\eta}, M)$$

- Furthermore,

$$\begin{aligned} \mathbb{L}(\eta, \hat{\eta}, M) &= \int_0^1 \mathbb{L}_c(\eta, \hat{\eta}, M)\, w(c)\, dc \\ &\approx \sum_{i=1}^{n} \mathbb{L}_{c_i}(\eta, \hat{\eta}, M) \end{aligned}$$

where the c_i are importance sampled using w



f-Divergence Estimation

f-Divergence and Bayes Risk

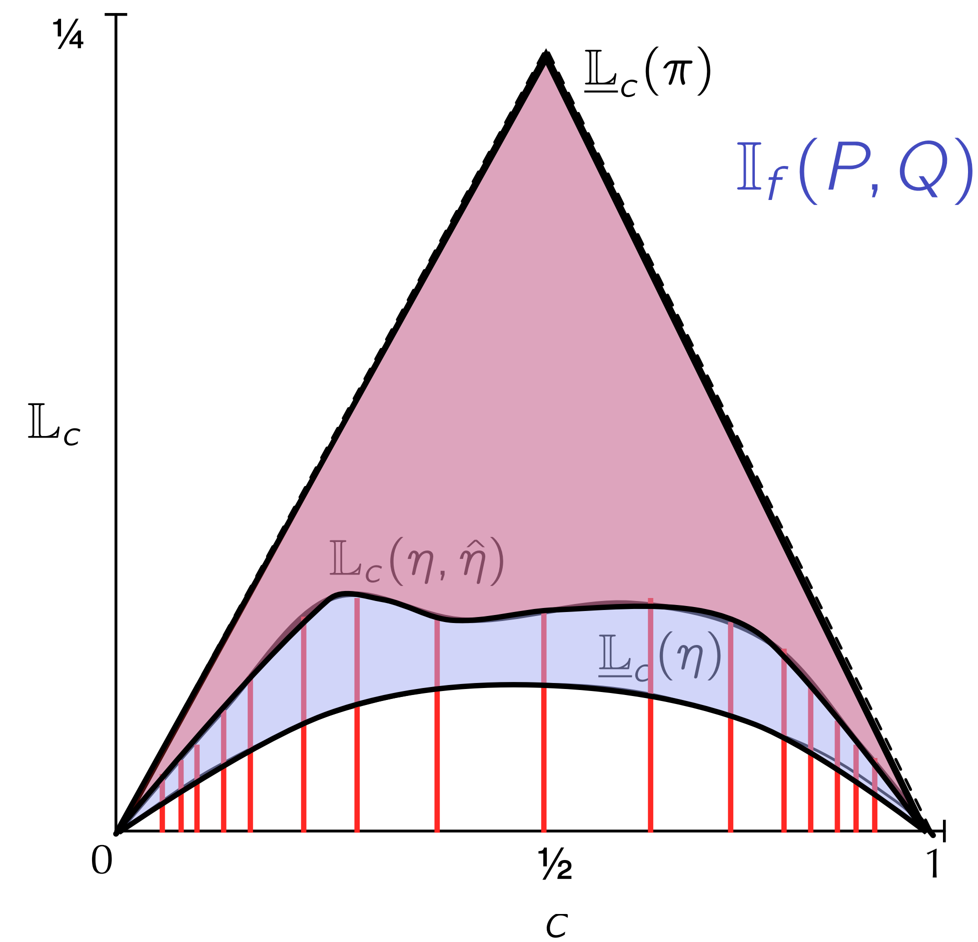
- Recall that $\mathbb{I}_f(P,Q) = \underline{\mathbb{L}}(\pi, M) - \underline{\mathbb{L}}(\eta, M)$
- For good estimators $\mathbb{L}(\eta, \hat{\eta}, M) \approx \underline{\mathbb{L}}(\eta, M)$ and so

$$\mathbb{I}_f(P,Q) \approx K - \mathbb{L}(\eta, \hat{\eta}, M)$$

- Furthermore,

$$\begin{aligned} \mathbb{L}(\eta, \hat{\eta}, M) &= \int_0^1 \mathbb{L}_c(\eta, \hat{\eta}, M)\, w(c)\, dc \\ &\approx \sum_{i=1}^{n} \mathbb{L}_{c_i}(\eta, \hat{\eta}, M) \end{aligned}$$

where the c_i are importance sampled using w



In theory, there is no difference between theory and practice. But, in practice, there is.

Jan L. A. van de Snepscheut (1953-1994)

Summary and Conclusions

Integral Form of the Taylor Expansion

$$f(t) = f(t_0) + (t - t_0)f'(t_0) + \int_a^b g(t,s)\, f''(s)\, ds$$

where $\quad g(t,s) = \begin{cases} (t-s) & t_0 \le s < t \\ (s-t) & t \le s < t_0 \end{cases}$

Jensen's Inequality

$$\mathbb{J}_P[f(x)] := \mathbb{E}_P\left[f(x)\right] - f\left(\mathbb{E}_P[x]\right) \ge 0$$

if and only if

f is convex

Summary - The Problems

Hypothesis Testing

- Given samples from P or Q **decide** whether samples were drawn from P or Q
    - Divergence / MMD

Summary - The Problems

Hypothesis Testing

- Given samples from P or Q **decide** whether samples were drawn from P or Q
  - Divergence / MMD

Classification

- Given samples from a π-mixture of P and Q **decide, for each instance** *x*, whether *x* was drawn from P or Q
  - 0-1 Misclassification Loss

Summary - The Problems

Hypothesis Testing

- Given samples from P or Q **decide** whether samples were drawn from P or Q
  - Divergence / MMD

Classification

- Given samples from a π-mixture of P and Q **decide, for each instance** *x*, whether *x* was drawn from P or Q
  - 0-1 Misclassification Loss

Probability Estimation

- Given samples from a π-mixture of P and Q **estimate**, for each instance x, **the probability** x was drawn from P (or Q)
  - Proper Scoring Rules

Summary - The Problems

Hypothesis Testing

- Given samples from P or Q **decide** whether samples were drawn from P or Q
  - Divergence / MMD

Classification

- Given samples from a π-mixture of P and Q **decide, for each instance** *x*, whether *x* was drawn from P or Q
  - 0-1 Misclassification Loss

Probability Estimation

- Given samples from a π-mixture of P and Q **estimate**, for each instance x, **the probability** x was drawn from P (or Q)
  - Proper Scoring Rules

Bipartite Ranking

- Given samples from a π-mixture of P and Q **sort** instances drawn from P ahead of those from Q
  - Area under ROC curve

Summary - The Representations

Weighted Integral Representation

- Taylor's Theorem

$$f(t) = \Lambda_f(t) + \int_a^b g_s(t)\, f''(s)\, ds$$

- f-Divergences

$$\mathbb{I}_f(P, Q) = \int_0^1 \mathbb{I}_{f_\pi}(P, Q)\, \gamma(\pi)\, d\pi$$

- Proper Scoring Rules

$$\ell_c(y, \hat{\eta}) = \int_0^1 \ell_c(y, \hat{\eta})\, w(c)\, dc$$

Variational Representation

- Legendre-Fenchel Dual

$$f(t) = f^{**}(t) = \sup_{t^* \in \mathbb{R}} \{t^*.t - f^*(t^*)\}$$

- f-Divergence

$$\mathbb{I}_f(P, Q) = \sup_{r:\mathcal{X} \to \mathbb{R}} \mathbb{E}_P[r] - \mathbb{E}_Q[f^*(r)]$$

Summary - The Relationships

Information

- Bregman Info = Stat Info
- Information is a Jensen gap

Divergence

- f-divergence is a Jensen gap

Risk and Regret

- Regret for proper losses is a Bregman divergence

Risk and Information

- Info = Max. reduction in risk

Information & Divergence

- Statistical Info = f-divergence (given mixing prior π)
- Explicit mapping of weights

Divergence and AUC

- Maximal AUC is not an f-divergence

Conclusions

Conclusions

Convexity and Expectations

- Convexity = Closure under expectation
- For Jensen Gaps
  - convexity => non-negativity

Conclusions

Convexity and Expectations

- Convexity = Closure under expectation
- For Jensen Gaps
  - convexity => non-negativity

Point-wise Bayes Risk

- Fundamental function in representation results
- Simple to derive from loss

Conclusions

Convexity and Expectations

- Convexity = Closure under expectation
- For Jensen Gaps
  - convexity => non-negativity

Point-wise Bayes Risk

- Fundamental function in representation results
- Simple to derive from loss

Divergence and Risk

- Two sides of the same coin

Conclusions

Convexity and Expectations

- Convexity = Closure under expectation
- For Jensen Gaps
  - convexity => non-negativity

Point-wise Bayes Risk

- Fundamental function in representation results
- Simple to derive from loss

Divergence and Risk

- Two sides of the same coin

Taylor Integral Expansion

- Implies weighted integral of piece-wise linear functions
  - Convexity => positive weights
  - Piece-wise linear = primitives

Conclusions

Convexity and Expectations

- Convexity = Closure under expectation
- For Jensen Gaps
  - convexity => non-negativity

Point-wise Bayes Risk

- Fundamental function in representation results
- Simple to derive from loss

Divergence and Risk

- Two sides of the same coin

Taylor Integral Expansion

- Implies weighted integral of piece-wise linear functions
  - Convexity => positive weights
  - Piece-wise linear = primitives

Problems, not just Techniques

- Insight by abstracting away from samples and understanding relationships

Fundamental progress has to do with the reinterpretation of basic ideas

Alfred North Whitehead (1961-1947)


Terra Statistica
Binary Experiments
Distributions
f-Divergence
Statistical Tests
Neyman-Pearson Lemma
Class Probability Estimation
Statistical
Thank You!
Bregman Information
Loss Functions
Regret
Risk
Representations
Integral
Variational
Graphical
ROC/AUC
Background
Convexity
LF Dual
Jensen's Inequality
Taylor's Theorem
Applications
MMD
Pinsker Bounds
Probing Reduction
Surrogate Regret Bounds

Thank You

Selected References

1. Reid and Williamson, **Information, Divergence and Risk for Binary Classification**, arXiv, 2009
2. Österreicher and Vajda, **Statistical Information and Divergence**, *Journal of Something or Other*, 1993
3. L. Savage, **On Measures of Uncertainty**, *Journal of Something*

Colophon

- Keynote 4 (with LinkBack plugin) using a modified Modern Portfolio theme
- OmniGraffle 5 for diagrams
- R for plots
- LaTeXiT for equations
- Text set in Helvetica Neue and equations in Computer Modern Bright [`\usepackage{cmbright}`]