

On information divergence measures,
surrogate loss functions and
decentralized hypothesis testing

XuanLong Nguyen, Martin J. Wainwright and Michael I. Jordan

Department of EECS and Department of Statistics

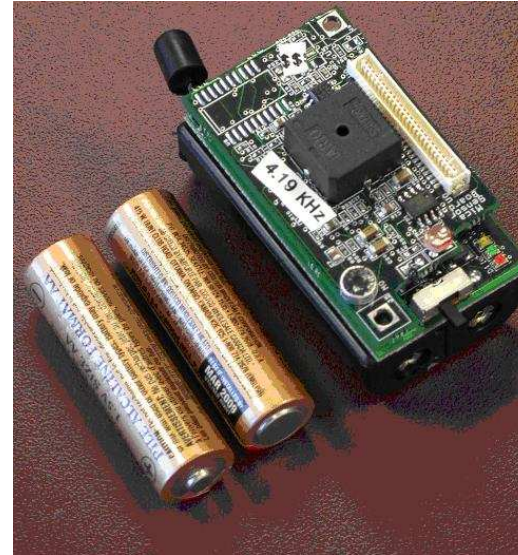
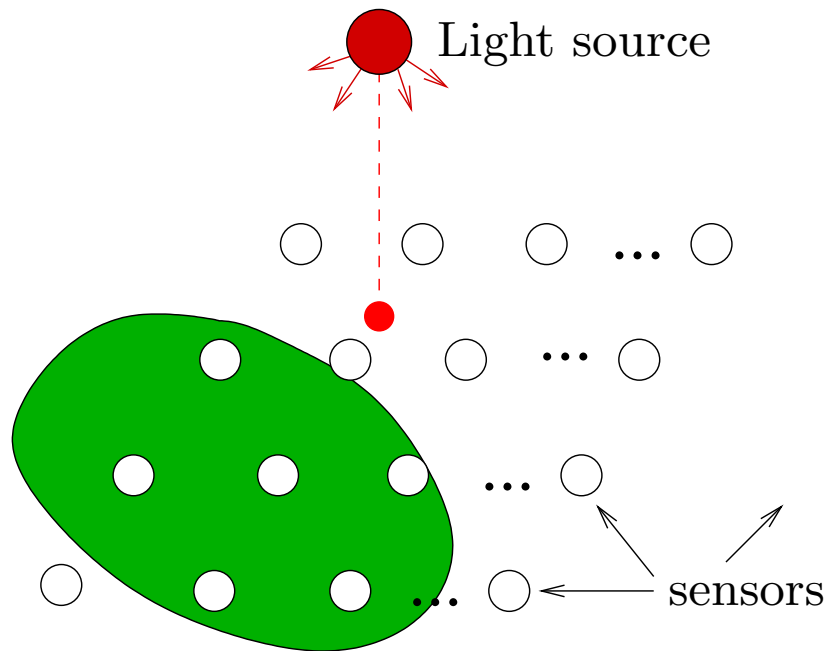
UC Berkeley, CA 94720

{xuanlong,wainwrig,jordan}@eecs.berkeley.edu

Introduction

- modern signal processing applications (e.g., databases, video, sensor nets) lead to large volumes of data
- (super)abundance of data \implies serious need for compression
- standard information-theoretic approaches to compression: based on classical distortion measures (e.g., Euclidean distance, Hamming distortion) (Shannon, 1948)
- traditional approach ignores the *functional role of data*
- more focused goal: perform compression while preserving information that is **relevant to an underlying statistical problem**

Concrete motivation: Wireless sensor networks

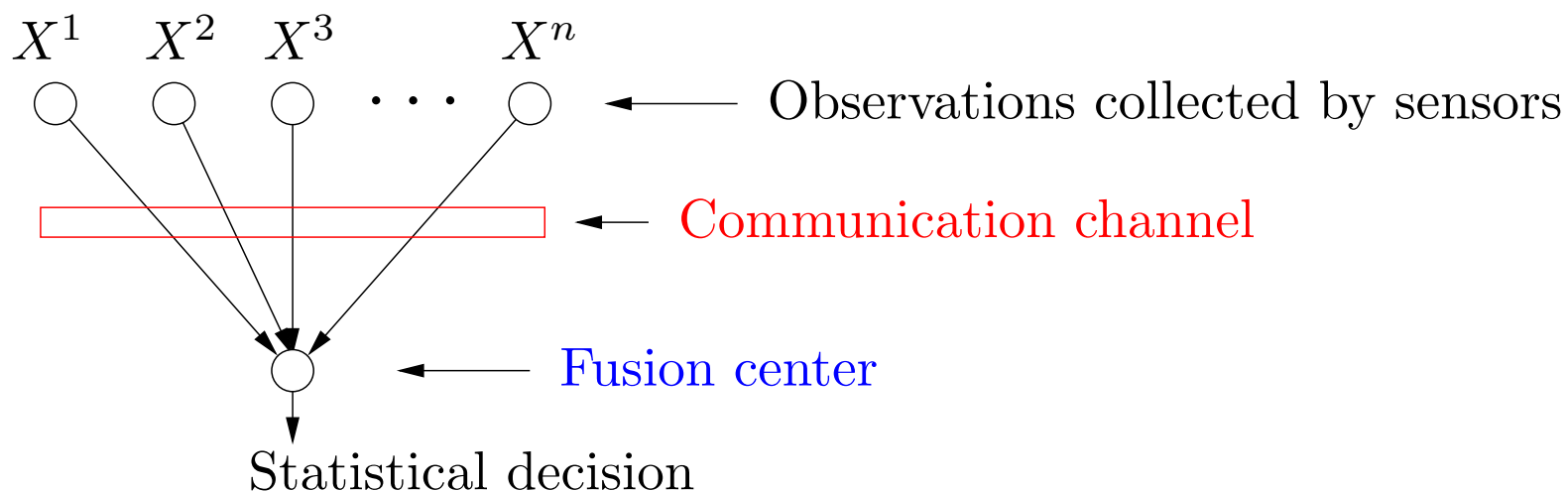


Set-up:

- Wireless network of motes equipped with sensors (e.g., light, heat, sound)
- Limited battery: can only transmit quantized observations

Goal: Perform a hypothesis test: e.g., is there a forest fire?

Decentralized hypothesis testing



- classical statistical procedures (e.g., hypothesis testing):
operate directly on observations X^1, \dots, X^n
- **Decentralization:**
 - sensors can relay only **quantized** versions of observations
 - statistical inference performed at a **central fusion center**
- **Goal:** Design compression or quantization rules so as to optimize some statistical criterion (e.g., Bayes error)

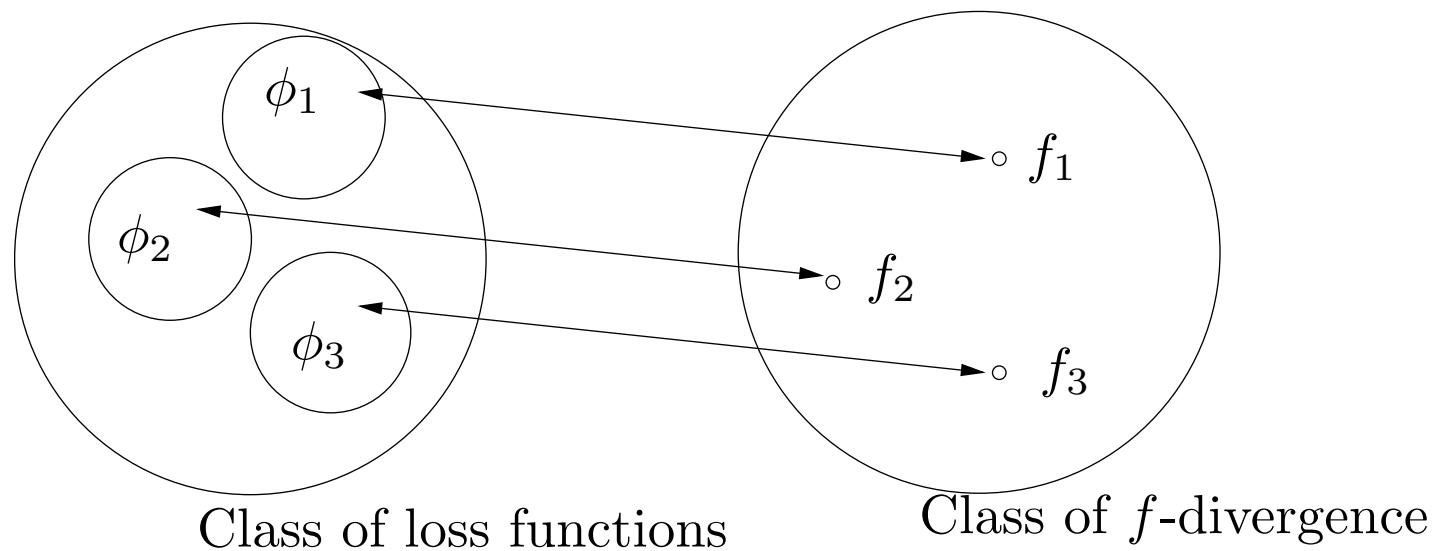
Related lines of work

- Decentralized hypothesis testing in statistical signal processing (Tenney & Sandell, 1981; Tsitsiklis, 1993)
 - joint distribution assumed to be known
 - locally-optimal rules under conditional independence assumptions
- Classical approaches based on f -divergence (Ali-Silvey distance)
 - Hellinger distance, Chernoff distance, KL divergence (e.g., Kailath, 1967; Longo et al, 1990, Chamberland & Veeravalli, 2003)
 - roles in asymptotic probability error rate
 - decision-theoretic insight lacking
- **Issue:** difficult to obtain the joint distribution over data (e.g., sensor networks)

Our contribution

- learning approach:
 - no assumption about the joint distribution over data
 - only have training data
- use of surrogate convex loss functions for 0-1 loss
 - loss function \equiv cost for making wrong decision
 - Bayes error \equiv expectation of 0-1 loss
- link between surrogate losses and distance measures
 - leads to Bayes consistent learning procedure

Link between convex surrogate loss and f -divergence

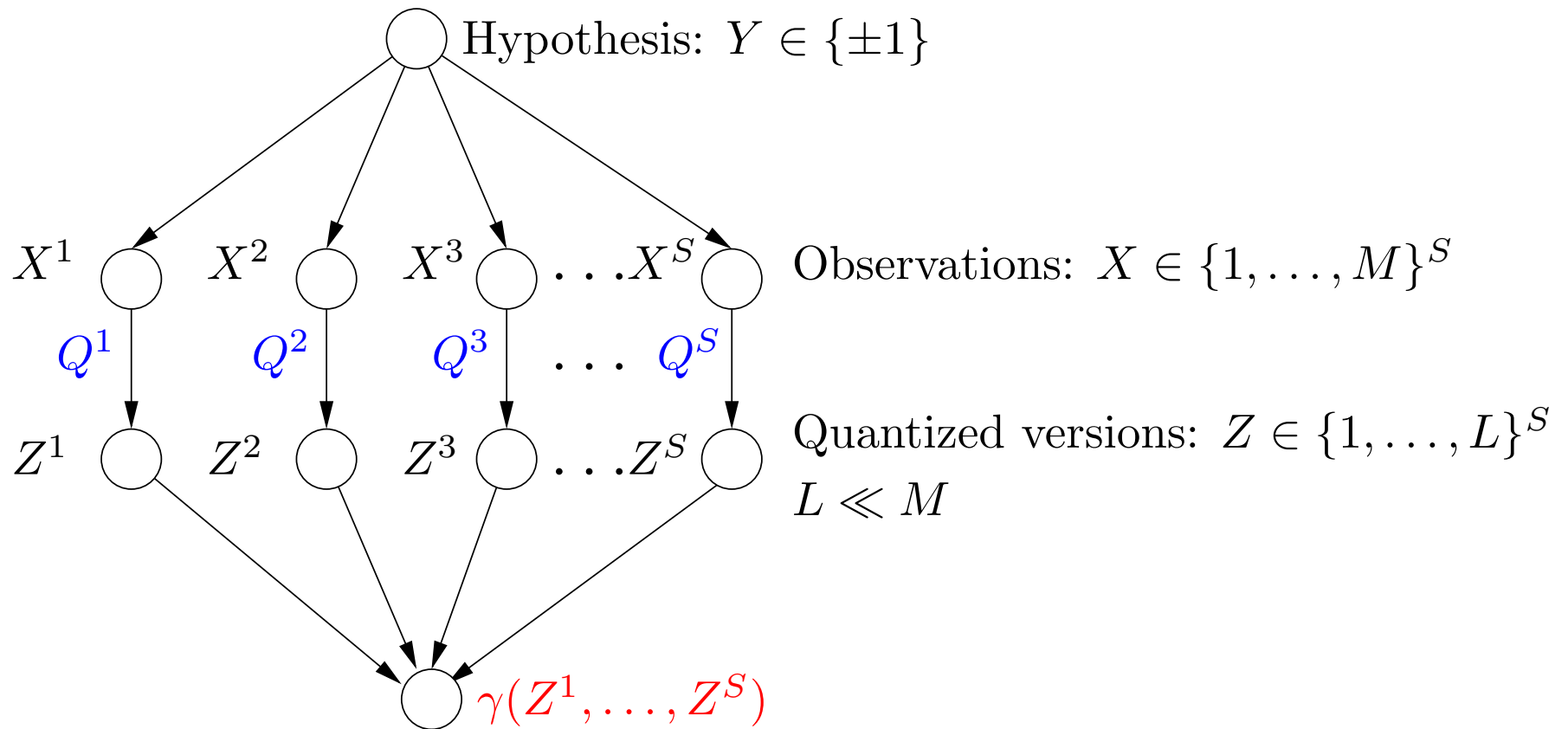


- two different mathematical objects:
decision-theoretic vs. **asymptotic** error
- cross-fertilization: relations among divergence measures
 \iff surrogate convex losses (e.g., inequalities)
- interesting equivalence classes of losses or divergences

Overview of talk

- formal set-up of decentralized hypothesis testing
- f -divergence measures in statistical signal processing
- surrogate convex loss functions in learning methods
- precise link between a class of surrogate convex loss and a class of f -divergence
 - establish constructive and many-to-one correspondence
- notion of universal equivalence among convex surrogate loss functions
 - establish universal consistency in learning procedure for (quantizer, classifier) pair using convex surrogate to 0-1 loss

Decentralized hypothesis testing



Problem: Find the decision rules $(Q; \gamma)$ so as to minimize the *prob. of incorrect decision*:

$$P(Y \neq \gamma(Z))$$

Classical approaches

- Maximize a distance measure between $P(Z|Y = 1)$ and $P(Z|Y = -1)$
 - Hellinger distance (Kailath 1967; Longo et al, 1990)
 - Chernoff distance (Chamberland & Veeravalli, 2003)
- Motivation: link between various distance measures and the **asymptotic error rate**
 - Kullback-Leibler divergence as in Neyman-Pearson setting
 - Chernoff distance as in a Bayesian setting

Ali-Silvey distance (f-divergence)

The **f-divergence** between measures μ and π is given by

$$I_f(\mu, \pi) := \sum_z \pi(z) f\left(\frac{\mu(z)}{\pi(z)}\right).$$

where $f : [0, +\infty) \rightarrow \mathbb{R} \cup \{+\infty\}$ is a continuous convex function

- **Kullback-Leibler** divergence: $f(u) = u \log u$.

$$I_f(\mu, \pi) = \sum_z \mu(z) \log \frac{\mu(z)}{\pi(z)}.$$

- **variational** distance: $f(u) = |u - 1|$.

$$I_f(\mu, \pi) := \sum_z |\mu(z) - \pi(z)|.$$

- **Hellinger** distance: $f(u) = \frac{1}{2}(\sqrt{u} - 1)^2$.

$$I_f(\mu, \pi) := \sum_{z \in \mathcal{Z}} (\sqrt{\mu(z)} - \sqrt{\pi(z)})^2.$$

Loss functions

- function defined on the hypothesis Y and decision $\gamma(Z)$: $\phi(Y, \gamma(Z))$

- 0-1 loss:

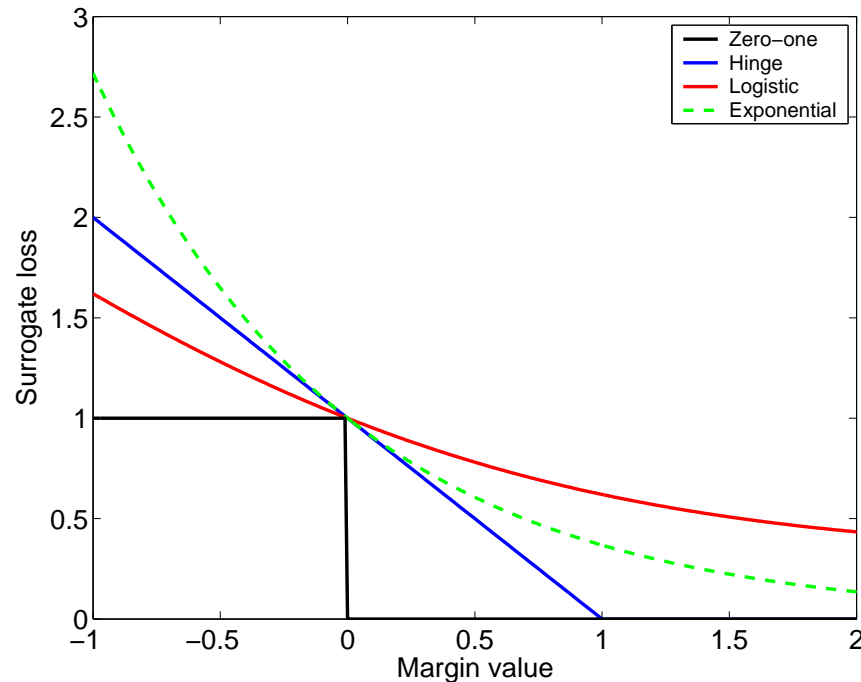
$$\phi(Y, \gamma(Z)) = \begin{cases} 1 & \text{if } Y \neq \gamma(Z) \\ 0 & \text{otherwise} \end{cases}$$

- for binary case, i.e., $Y \in \{1, -1\}$, 0-1 loss can be written as:

$$\phi(Y, \gamma(Z)) = \mathbb{I}(Y\gamma(Z) < 0)$$

- margin-based convex surrogate loss: ϕ defined on the *margin* $Y\gamma(Z)$

Margin-based surrogate loss functions



- convex surrogates ϕ are approximations to 0 – 1 loss
- defined in terms of *margin* $t = y\gamma(z)$
 - hinge loss: $\phi(t) = \max(0, 1 - t)$ support vector machine
 - exponential loss: $\phi(t) = \exp(-t)$ boosting
 - logistic loss: $\phi(t) = \log[1 + \exp(-t)]$ regularized log. regression

Learning methods based on convex surrogate losses

- (X, Y) be distributed by (possibly unknown) distribution $P(X, Y)$
 - Y is a random variable in $\{-1, +1\}$
- given i.i.d. training data $(x_1, y_1), \dots, (x_n, y_n)$
- **Learning problem:** Find a classifier $\gamma : \mathcal{X} \rightarrow \mathbb{R}$ to minimize the *probability of misclassification*:

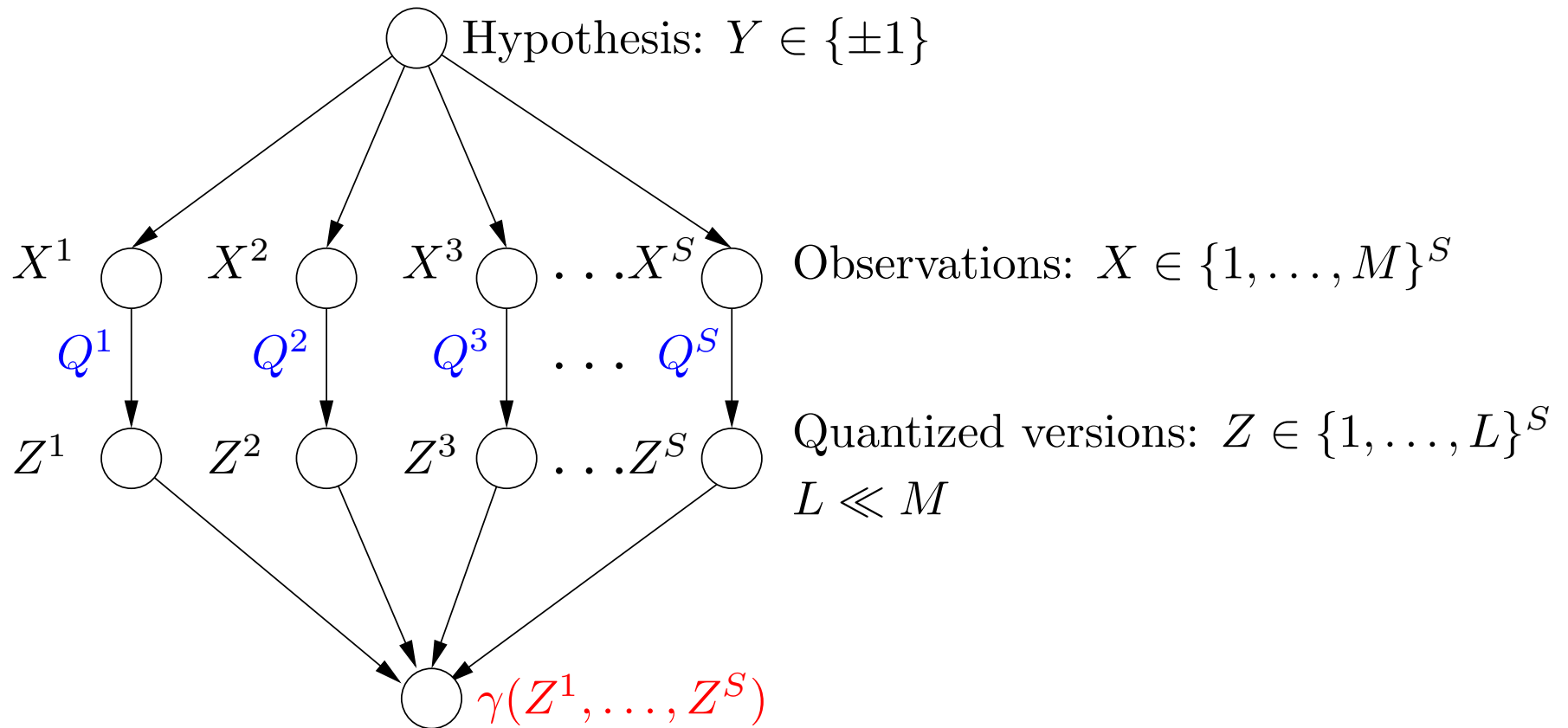
$$P(Y \neq \text{sign}(\gamma(X))) = \mathbb{E}\mathbb{I}_{(Y\gamma(X) < 0)}$$

- **Approach:** Find a classifier γ to minimize an empirical version of expected loss:

$$\hat{\mathbb{E}}\phi(Y\gamma(X)) := \frac{1}{n} \sum_{i=1}^n \phi(y_i \gamma(x_i))$$

- ϕ is a **convex surrogate loss** for 0-1 loss function

Decentralized hypothesis testing



Problem: Given training data $(x_i, y_i)_{i=1}^n$, find the decision rules $(Q; \gamma)$ so as to minimize the *prob. of incorrect decision*:

$$P(Y \neq \gamma(Z))$$

Decision-theoretic perspective with surrogate losses

Approach: Find (Q, γ) to minimize the ϕ -risk

$$R_\phi(\gamma, Q) = \mathbb{E}\phi(Y\gamma(Z))$$

- Define:

$$\mu(z) = P(Y = 1, z) = p \int_x Q(z|x) dP(x|Y = 1)$$

$$\pi(z) = P(Y = -1, z) = q \int_x Q(z|x) dP(x|Y = -1).$$

- ϕ -risk can be represented as:

$$R_\phi(\gamma, Q) = \sum_z \phi(\gamma(z))\mu(z) + \phi(-\gamma(z))\pi(z)$$

- This allows us to choose the optimal value for γ for each z to obtain:

$$R_\phi(Q) := \inf_{\gamma \in \Gamma} R_\phi(\gamma, Q)$$

Optimizing out surrogate loss functions

- **0-1 loss:**

$$R_{bayes}(Q) = \frac{1}{2} - \frac{1}{2} \sum_{z \in \mathcal{Z}} |\mu(z) - \pi(z)| \quad \Rightarrow \text{variational distance}$$

- **hinge loss:**

$$R_{hinge}(Q) = 1 - V(\mu, \pi) = 2R_{bayes}(Q) \quad \Rightarrow \text{variational distance}$$

- **exponential loss:**

$$R_{exp}(Q) = 1 - \sum_{z \in \mathcal{Z}} (\sqrt{\mu(z)} - \sqrt{\pi(z)})^2 \quad \Rightarrow \text{Hellinger distance}$$

- **logistic loss:**

$$R_{log}(Q) = \log 2 - KL(\mu || \frac{\mu + \pi}{2}) - KL(\pi || \frac{\mu + \pi}{2}) \Rightarrow \text{capacitory dis. distance}$$

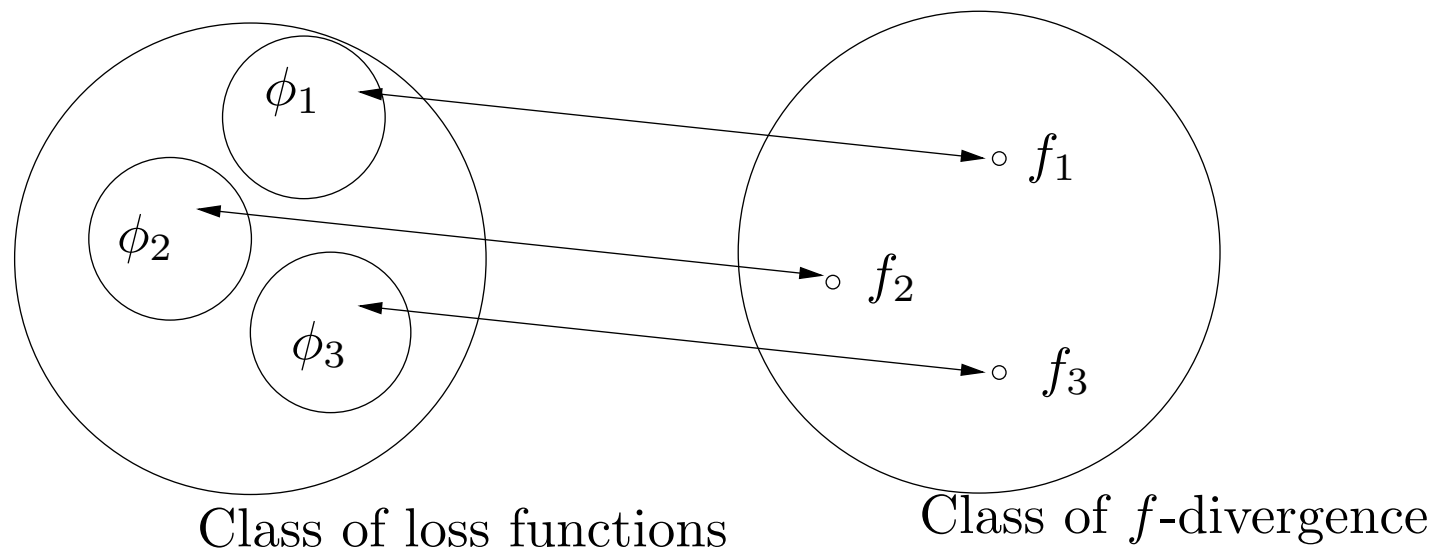
Properties of surrogate loss functions

- First, ϕ is continuous and convex.
- Second, ϕ must be classification-calibrated, i.e., for any $a, b \geq 0$ and $a \neq b$,

$$\inf_{\alpha: \alpha(a-b) < 0} \phi(\alpha)a + \phi(-\alpha)b > \inf_{\alpha \in \mathbb{R}} \phi(\alpha)a + \phi(-\alpha)b$$

- ϕ is classification-calibrated *iff* differentiable at 0 and $\phi'(0) < 0$

Link between surrogate loss and f -divergence



Proposition: For each Q , let γ_Q denote the optimal decision rule. The ϕ -risk for (Q, γ_Q) is an f -divergence for some convex f :

$$R_\phi(Q) = -I_f(\mu, \pi)$$

Nec. and suff. conditions for realizable f -divergence

- Only *symmetric* f -divergences are realizable by margin-based ϕ loss
- Our strategy: related ϕ and f by an intermediate function Ψ :

$$\phi \iff \Psi \iff f$$

- Define Ψ from ϕ :

$$\Psi(\beta) := \begin{cases} \phi(-\phi^{-1}(\beta)) & \text{if } \phi^{-1}(\beta) \in \mathbb{R}, \\ +\infty & \text{otherwise.} \end{cases}$$

- Ψ and f are related by

$$f(u) = \Psi^*(-u)$$

– Ψ^* denotes the conjugate dual of convex function f

- $\Psi(\beta)$ has to satisfy a number of necessary conditions

Recover ϕ loss for an f -divergence

Theorem: Given an f -divergence, define

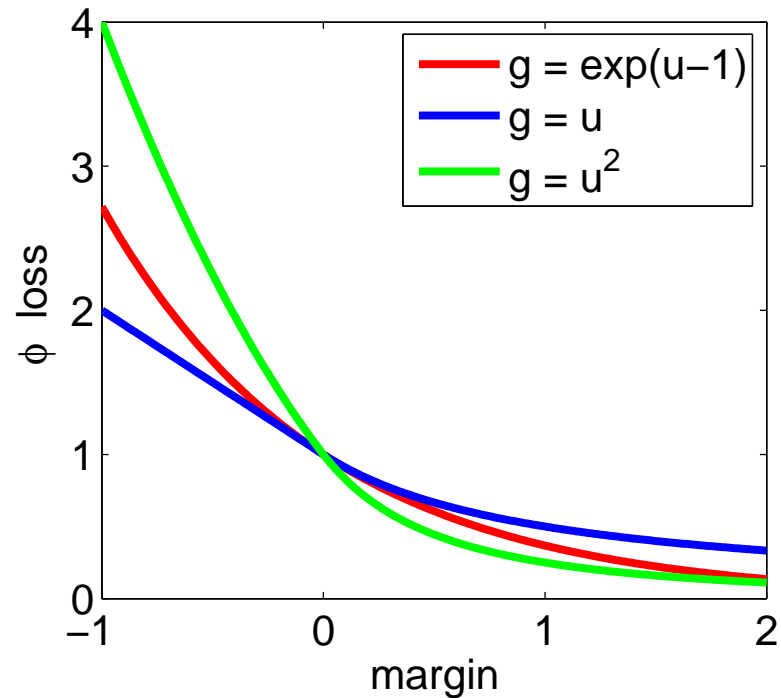
$$\Psi(\beta) = f^*(-\beta)$$

If Ψ satisfies certain sufficient conditions, then all corresponding ϕ loss are of the following form:

$$\phi(\alpha) = \begin{cases} \Psi(g(\alpha + u^*)) & \text{if } \alpha \geq 0 \\ g(-\alpha + u^*) & \text{otherwise} \end{cases}$$

- where u^* satisfies $\Psi(u^*) = u^*$,
- $g : [u^*, +\infty) \rightarrow \overline{\mathbb{R}}$ is any increasing continuous convex function such that $g(u^*) = u^*$
- g differentiable at u^* and $g'(u^*+) > 0$

Example – Hellinger distance

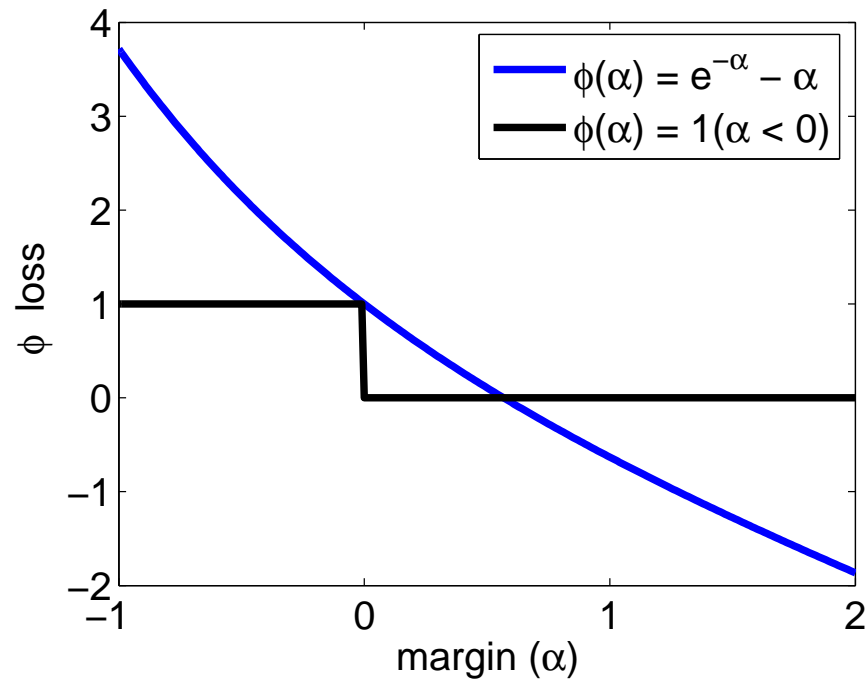


- Hellinger distance corresponds to an f -divergence with $f(u) = -2\sqrt{u}$
- Recover immediate function $\Psi(\beta) = f^*(-\beta) = \begin{cases} 1/\beta & \text{when } \beta > 0 \\ +\infty & \text{otherwise.} \end{cases}$
- Choosing $g(u) = e^{u-1}$ yields $\phi(\alpha) = \exp(-\alpha) \Rightarrow$ exponential loss

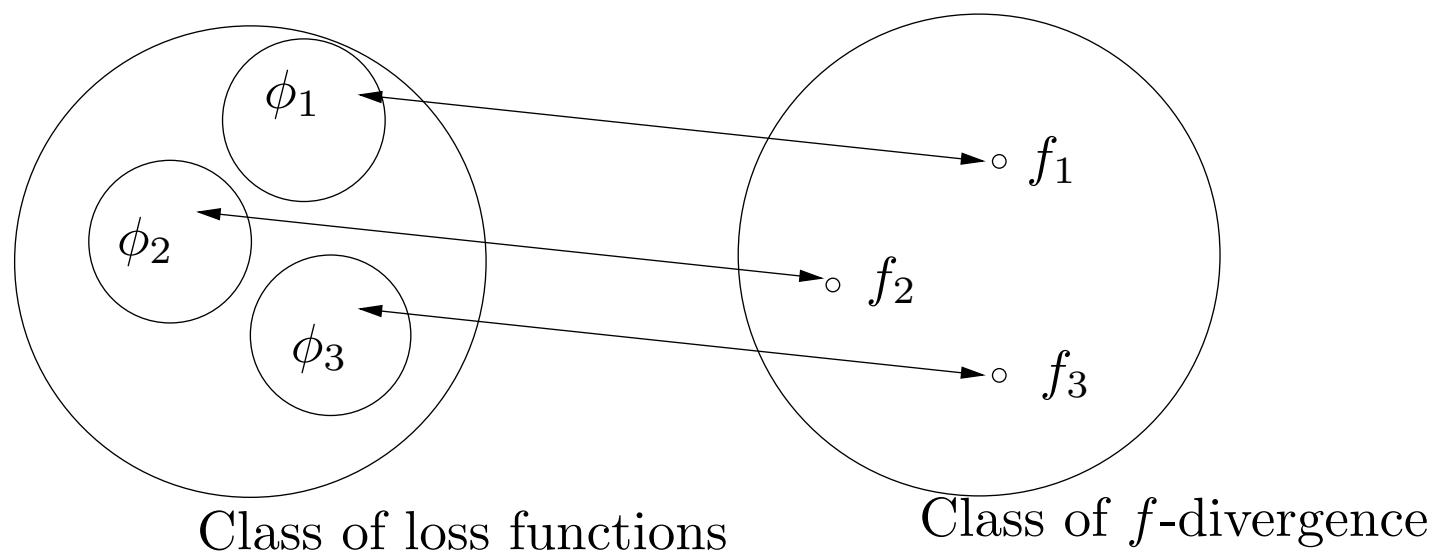
Example – Kullback-Leibler divergence

- there exist no corresponding ϕ loss for both $KL(\mu||\pi)$ and $KL(\pi||\mu)$
- symmetrized KL divergence $KL(\mu||\pi) + KL(\pi||\mu)$ is realized by

$$\phi(\alpha) = e^{-\alpha} - \alpha$$



Exploiting the link between surrogate loss and f -divergence



- establish useful relations between different loss functions (e.g., inequalities)
- design loss functions that are computationally useful
 - what are convex surrogate ϕ loss that are equivalent to 0-1 loss?

Universal equivalence of loss functions

- Consider two loss functions ϕ_1 and ϕ_2 , corresponding to f -divergence induced by f_1 and f_2
- ϕ_1 and ϕ_2 are **universally equivalent**, denoted by

$$\phi_1 \stackrel{u}{\approx} \phi_2 \quad (\text{or, equivalently}) \quad f_1 \stackrel{u}{\approx} f_2$$

if for **any** $P(X, Y)$ and quantization rules Q_A, Q_B , there holds:

$$R_{\phi_1}(Q_A) \leq R_{\phi_1}(Q_B) \Leftrightarrow R_{\phi_2}(Q_A) \leq R_{\phi_2}(Q_B).$$

i.e., \mathbb{R}_{ϕ_1} and \mathbb{R}_{ϕ_2} have the **same monotonic behavior** with respect to Q

An equivalence theorem

Theorem:

$$\phi_1 \stackrel{u}{\approx} \phi_2 \quad (\text{or, equivalently}) \quad f_1 \stackrel{u}{\approx} f_2$$

if and only if

$$f_1(u) = cf_2(u) + au + b$$

for constants $a, b \in \mathbb{R}$ and $c > 0$

- \Leftarrow is easy: $I_{f_1} = c I_{f_2} + a + b$
- in particular, surrogate losses universally equivalent to 0 – 1 loss are those whose induced f divergence has the form:

$$f(u) = c \min\{u, 1\} + au + b$$

Decentralized detection using empirical samples

- problem: find (Q, γ) that minimize the Bayes error

$$R_{\text{bayes}}(Q, \gamma) = P(Y \neq \gamma(Z))$$

- our previous work: minimizing an empirical version of the expected ϕ loss:

$$\hat{R}_\phi(Q, \gamma) = \hat{\mathbb{E}}\phi(Y\gamma(Z)) := \frac{1}{n} \sum_{i=1}^n \sum_z \phi(y_i \gamma(z)) Q(z|x_i)$$

- question:

is this learning procedure **optimal** (i.e., Bayes consistent)
for *any* underlying $P(X, Y)$?

Empirical risk minimization procedure

- let ϕ be a convex surrogate universally equivalent to 0 – 1 loss
- $(\mathcal{C}_n, \mathcal{D}_n)$ is a sequence of increasing function classes for (γ, Q)

$$(\mathcal{C}_1, \mathcal{D}_1) \subseteq (\mathcal{C}_2, \mathcal{D}_2) \subseteq \dots \subseteq (\Gamma, \mathcal{Q})$$

- our procedure learns:

$$(\gamma_n^*, Q_n^*) := \operatorname{argmin}_{(\gamma, Q) \in (\mathcal{C}_n, \mathcal{D}_n)} \hat{R}_\phi(\gamma, Q)$$

- let $R_{bayes}^* := \inf_{(\gamma, Q) \in (\Gamma, \mathcal{Q})} P(Y \neq \gamma(Z)) \quad \Leftarrow$ **optimal Bayes risk**
- $R_{bayes}(\gamma_n^*, Q_n^*) - R_{bayes}^*$ is called the **Bayes error** of our procedure

Consistency of empirical risk minimization procedure

Theorem: If

- $\cup_{n=1}^{\infty} (\mathcal{C}_n, \mathcal{D}_n)$ is dense in the space of measurable pair of classifier and quantizer $(\gamma, Q) \in (\Gamma, \mathcal{Q})$
- sequence $(\mathcal{C}_n, \mathcal{D}_n)$ increases in size sufficiently slowly

then our procedure is consistent, i.e.,

$$\lim_{n \rightarrow \infty} R_{bayes}(\gamma_n^*, Q_n^*) - R_{bayes}^* = 0 \quad \text{in probability.}$$

- proof exploits the universal equivalence of ϕ loss and 0 – 1 loss
- decomposition of ϕ risk into approximation error and estimation error

Conclusions

- precise link between f -divergence measures and surrogate convex loss functions
- decision-theoretic perspective to the use of f -divergence
- equivalent classes of loss functions
- design new convex surrogate loss functions that are equivalent to 0-1 loss
⇒ consistent learning procedure in decentralized hypothesis testing problem