

f -Divergences and Surrogate Loss Functions

Michael I. Jordan

Department of Statistics

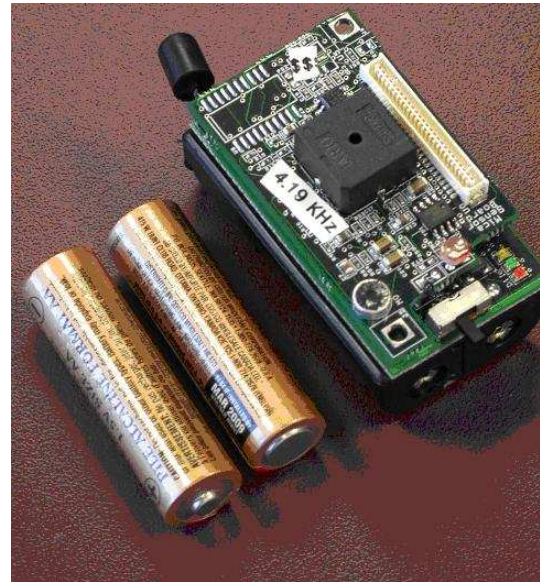
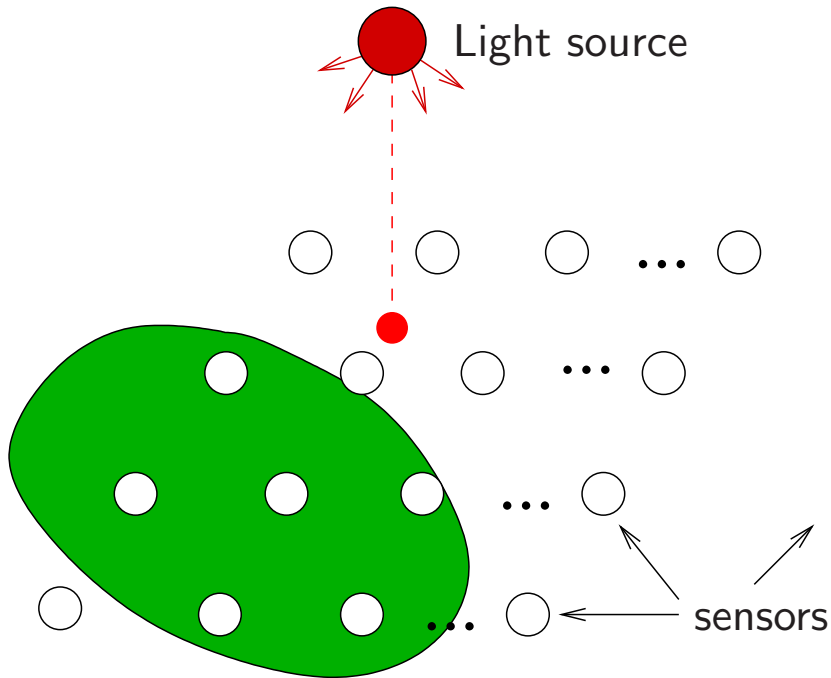
Department of EECS

University of California, Berkeley

<http://www.cs.berkeley.edu/~jordan>

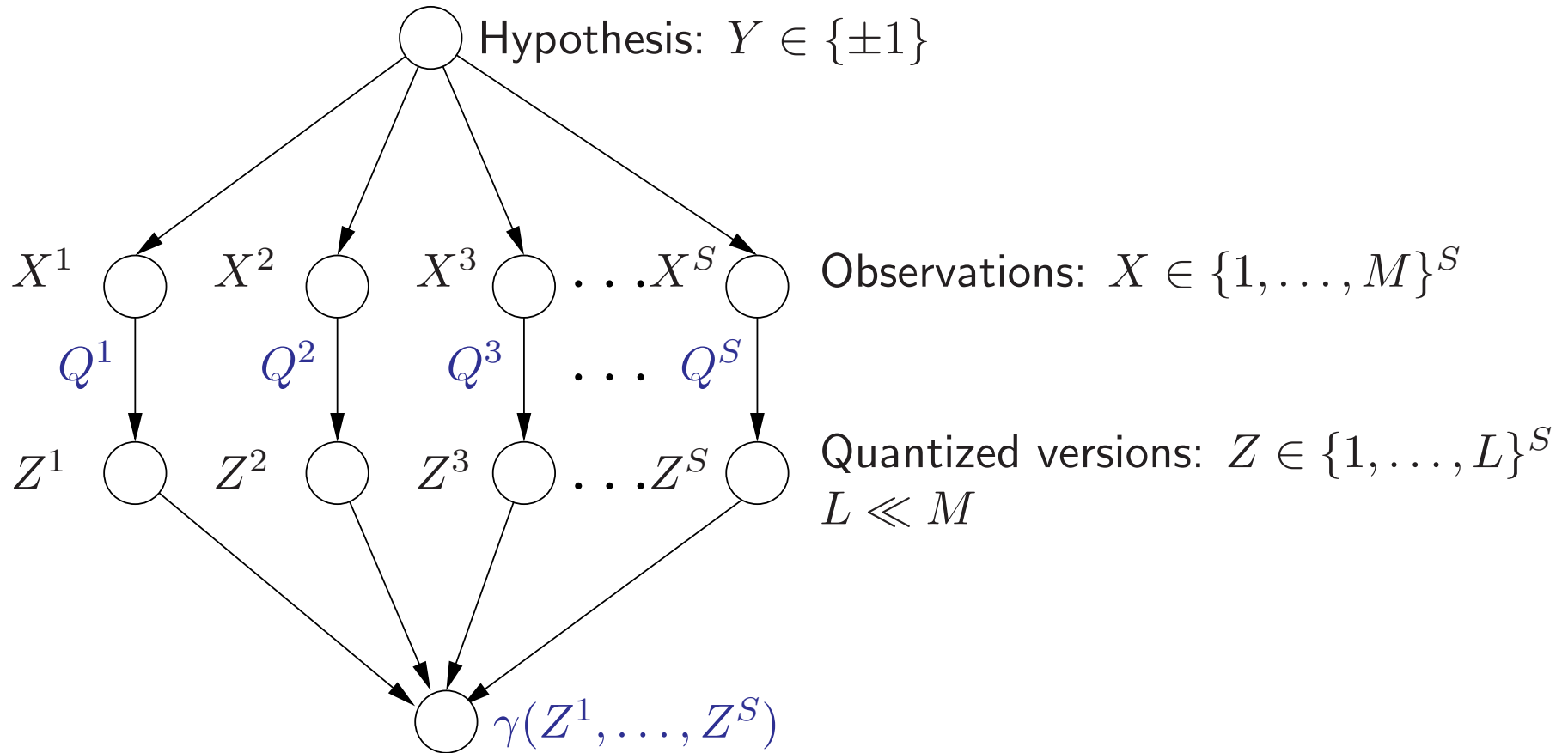
With: XuanLong Nguyen and Martin Wainwright

Motivating Example: Decentralized Detection



- Wireless network of motes equipped with sensors (e.g., light, heat, sound)
- Limited battery: can only transmit quantized observations
- Is the light source above the green region?

Decentralized Detection



Decentralized Detection (cont.)

- General set-up:
 - data are (X, Y) pairs, assumed sampled i.i.d. for simplicity, where $Y \in \{0, 1\}$
 - given X , let $Z = Q(X)$ denote the covariate vector, where $Q \in \mathcal{Q}$, where \mathcal{Q} is some set of random mappings (can be viewed as an experimental design)
 - consider a family $\{\gamma(\cdot)\}$, where γ is a discriminant function lying in some (nonparametric) family Γ
- Problem: Find the decision $(Q; \gamma)$ that minimizes the probability of error $P(Y \neq \gamma(Z))$
- Applications include:
 - decentralized compression and detection
 - feature extraction, dimensionality reduction
 - problem of sensor placement

Perspectives

- *Signal processing literature*
 - everything is assumed known except for Q —the problem of “decentralized detection” is to find Q
 - this is done via the maximization of an “ f -divergence” (e.g., Hellinger distance, Chernoff distance)
 - basically a heuristic literature
- *Machine learning literature*
 - Q is assumed known and the problem is to find γ
 - this is done via the minimization of an “surrogate loss function” (e.g., boosting, logistic regression, support vector machine)
 - decision-theoretic flavor; consistency results

f -divergences (Ali-Silvey Distances)

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

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

Why the f -divergence?

- A classical theorem due to Blackwell (1951): *If a procedure A has a smaller f -divergence than a procedure B (for some fixed f), then there exist some set of prior probabilities such that procedure A has a smaller probability of error than procedure B*
- Given that it is intractable to minimize probability of error, this result has motivated (many) authors in signal processing to use f -divergences as surrogates for probability of error
- I.e., choose a quantizer Q by maximizing an f -divergence between $P(Z|Y = 1)$ and $P(Z|Y = -1)$
 - Hellinger distance (Kailath 1967; Longo et al, 1990)
 - Chernoff distance (Chamberland & Veeravalli, 2003)
- Supporting arguments from asymptotics
 - Kullback-Leibler divergence in the Neyman-Pearson setting
 - Chernoff distance in the Bayesian setting

Machine Learning Perspective

- *Decision-theoretic*: based on a loss function $\phi(Y, \gamma(Z))$

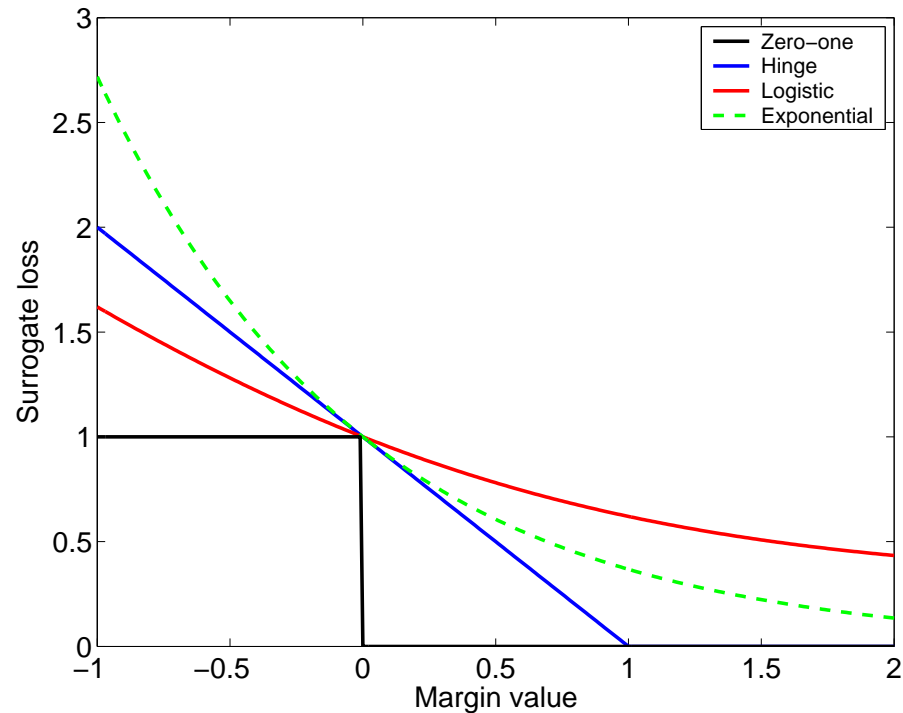
- E.g., 0-1 loss:

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

which can be written in the binary case as $\phi(Y, \gamma(Z)) = \mathbb{I}(Y \gamma(Z) < 0)$

- The main focus is on estimating γ ; the problem of estimating Q by minimizing the loss function is only occasionally addressed
- It is intractable to minimize 0-1 loss, so consider minimizing a **surrogate loss functions** that is a convex upper bound on the 0-1 loss

Margin-Based Surrogate Loss Functions



- Define a convex surrogate in terms of the **margin** $u = y\gamma(z)$
 - hinge loss: $\phi(u) = \max(0, 1 - u)$ support vector machine
 - exponential loss: $\phi(u) = \exp(-u)$ boosting
 - logistic loss: $\phi(u) = \log[1 + \exp(-u)]$ logistic regression

Estimation Based on a Convex Surrogate Loss

- Estimation procedures used in the machine learning literature are generally M -estimators (“empirical risk minimization”)
- Given i.i.d. training data $(x_1, y_1), \dots, (x_n, y_n)$
- Find a classifier γ that minimizes the empirical expectation of the surrogate loss:

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

where the convexity of ϕ makes this feasible in practice and in theory

Some Theory for Surrogate Loss Functions

(Bartlett, Jordan, & McAuliffe, JASA 2005)

- ϕ must be **classification-calibrated**, i.e., for any $a, b \geq 0$ and $a \neq b$,

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

(essentially a form of Fisher consistency that is appropriate for classification)

- This is necessary and sufficient for Bayes consistency; we take it as the definition of a “surrogate loss function” for classification
- In the convex case, ϕ is classification-calibrated *iff* differentiable at 0 and $\phi'(0) < 0$

Outline

- A precise link between surrogate convex losses and f -divergences
 - we establish a constructive and many-to-one correspondence
- A notion of **universal equivalence** among convex surrogate loss functions
- An application: Proof of consistency for the choice of a (Q, γ) pair using any convex surrogate for the 0-1 loss

Setup

- We want to 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:

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

Profiling

- Optimize out over γ (for each z) and define:

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

- For example, for 0-1 loss, we easily obtain $\gamma(z) = \text{sign}(\mu(z) - \pi(z))$. Thus:

$$\begin{aligned} R_{0-1}(Q) &= \sum_{z \in \mathcal{Z}} \min\{\mu(z), \pi(z)\} \\ &= \frac{1}{2} - \frac{1}{2} \sum_{z \in \mathcal{Z}} |\mu(z) - \pi(z)| \\ &= \frac{1}{2}(1 - V(\mu, \pi)) \end{aligned}$$

where $V(\mu, \pi)$ is the variational distance.

- I.e., optimizing out a ϕ -risk yields an f -divergence. Does this hold more generally?

Some Examples

- **hinge loss:**

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

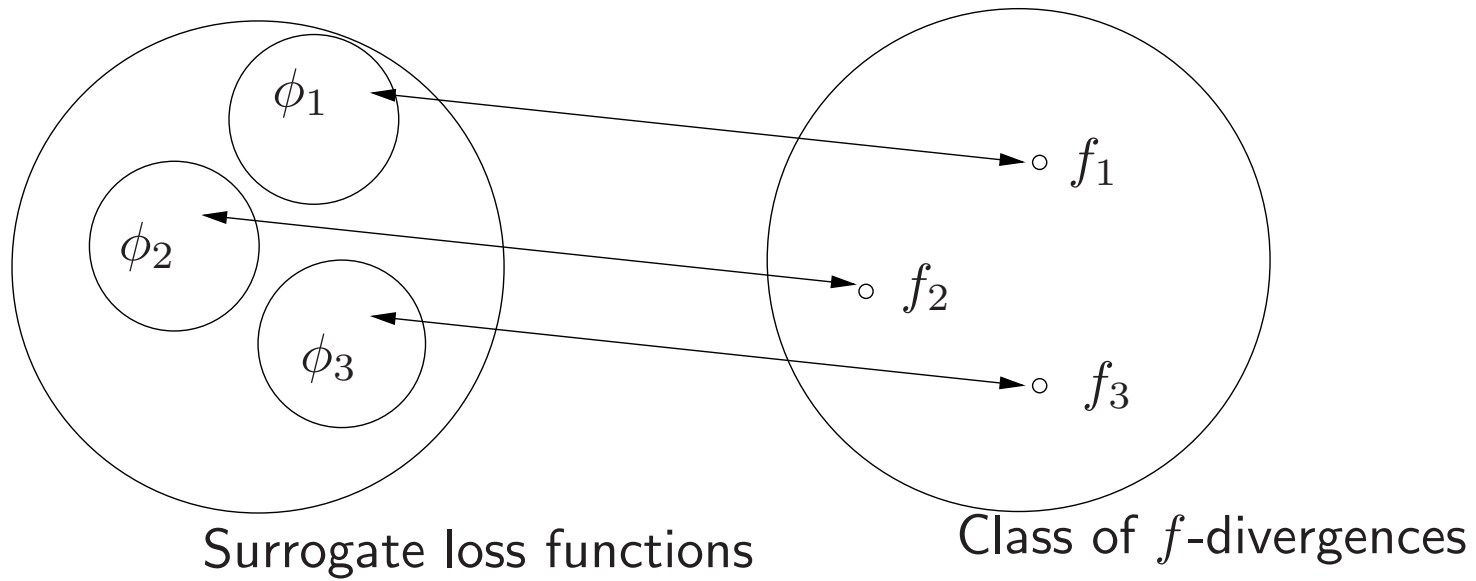
- **exponential loss:**

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

- **logistic loss:**

$$R_{log}(Q) = \log 2 - D\left(\mu \parallel \frac{\mu + \pi}{2}\right) - D\left(\pi \parallel \frac{\mu + \pi}{2}\right) \quad (\text{capacitory discrimination})$$

Link between ϕ -losses and f -divergences



Conjugate Duality

- Recall the notion of *conjugate duality* (Rockafellar): For a lower-semicontinuous convex function $f : \mathbb{R} \rightarrow \mathbb{R} \cup \{\infty\}$, the conjugate dual $f^* : \mathbb{R} \rightarrow \mathbb{R} \cup \{\infty\}$ is defined as

$$f^*(u) = \sup_{v \in \mathbb{R}} \{uv - f(v)\},$$

which is necessarily a convex function.

- Define

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

Link between ϕ -losses and f -divergences

Theorem 1. (a) For any margin-based surrogate loss function ϕ , there is an f -divergence such that $R_\phi(Q) = -I_f(\mu, \pi)$ for some lower-semicontinuous convex function f .

In addition, if ϕ is continuous and satisfies a (weak) regularity condition, then the following properties hold:

- (i) Ψ is a decreasing and convex function.
 - (ii) $\Psi(\Psi(\beta)) = \beta$ for all $\beta \in (\beta_1, \beta_2)$.
 - (iii) There exists a point u^* such that $\Psi(u^*) = u^*$.
- (b) Conversely, if f is a lower-semicontinuous convex function satisfying conditions (i–iii), there exists a decreasing convex surrogate loss ϕ that induces the corresponding f -divergence

The Easy Direction: $\phi \rightarrow f$

- Recall

$$R_\phi(\gamma, Q) = \sum_{z \in \mathcal{Z}} \phi(\gamma(z))\mu(z) + \phi(-\gamma(z))\pi(z)$$

- Optimizing out $\gamma(z)$ for each z :

$$R_\phi(Q) = \sum_{z \in \mathcal{Z}} \inf_{\alpha} \phi(\alpha)\mu(z) + \phi(-\alpha)\pi(z) = \sum_z \pi(z) \inf_{\alpha} \left(\phi(-\alpha) + \phi(\alpha) \frac{\mu(z)}{\pi(z)} \right)$$

- For each z let $u = \frac{\mu(z)}{\pi(z)}$, define:

$$f(u) := - \inf_{\alpha} (\phi(-\alpha) + \phi(\alpha)u)$$

- f is a convex function
- we have

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

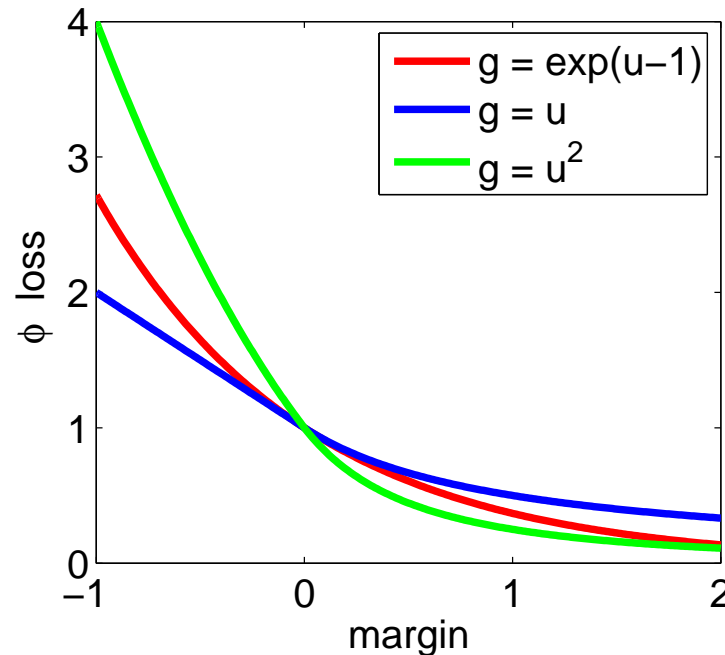
The $f \rightarrow \phi$ Direction Has a Constructive Consequence

- Any continuous loss function ϕ that induces an f -divergence must be of the form

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

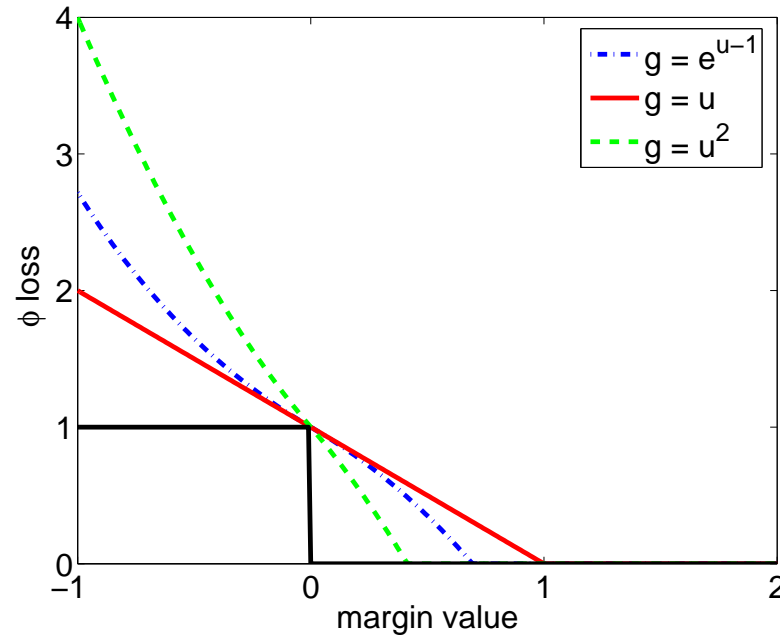
where $g : [u^*, +\infty) \rightarrow \overline{\mathbb{R}}$ is some increasing continuous and convex function such that $g(u^*) = u^*$, and g is right-differentiable at u^* with $g'(u^*) > 0$.

Example – Hellinger distance



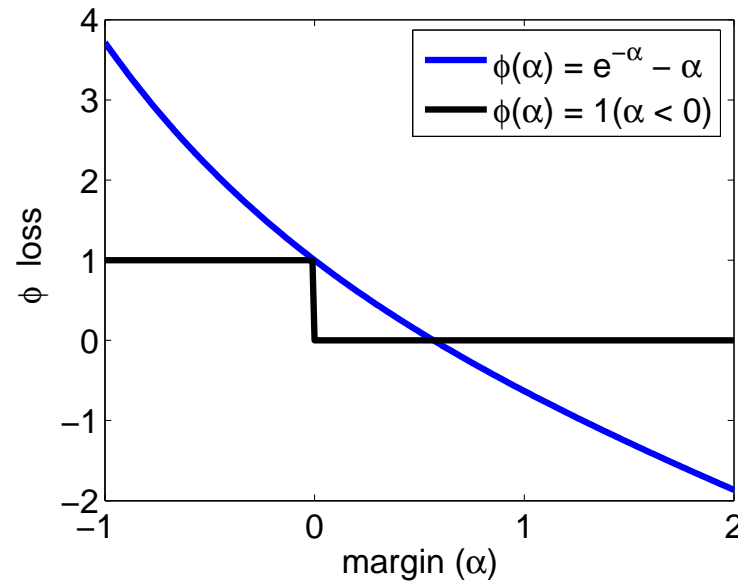
- Hellinger distance corresponds to an f -divergence with $f(u) = -2\sqrt{u}$
- Recover immediate function $\Psi(\beta) = f^*(-\beta) = \begin{cases} 1/\beta & \text{when } \beta > 0 \\ +\infty & \text{otherwise.} \end{cases}$
- Choosing $g(u) = e^{u-1}$ yields $\phi(\alpha) = \exp(-\alpha) \Rightarrow$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

Example – Variational distance



- Variational distance corresp. to an f -divergence with $f(u) = -2 \min\{u, 1\}$
- Recover immediate function $\Psi(\beta) = f^*(-\beta) = \begin{cases} (2 - \beta)_+ & \text{when } \beta > 0 \\ +\infty & \text{otherwise.} \end{cases}$
- Choosing $g(u) = u$ yields $\phi(\alpha) = (1 - \alpha)_+$ \Rightarrow hinge loss

Example – Kullback-Leibler divergence



- There is no corresponding ϕ loss for either $D(\mu||\pi)$ or $D(\pi||\mu)$
- But the *symmetrized* KL divergence $D(\mu||\pi) + D(\pi||\mu)$ is realized by

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

Bayes Consistency for Choice of (Q, λ)

- Recall that from the 0-1 loss, we obtain the variational distance as the corresponding f -divergence, where $f(u) = \min\{u, 1\}$.
- Consider a broader class of f -divergences defined by:

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

- And consider the set of (continuous, convex and classification-calibrated) ϕ -losses that can be obtained (via Theorem 1) from these f -divergences
- We will provide conditions under which such ϕ -losses yield Bayes consistency for procedures that jointly choose (Q, λ)
- (And later we will show that *only* such ϕ -losses yield Bayes consistency)

Setup

- Consider sequences of increasing compact function classes $\mathcal{C}_1 \subseteq \dots \subseteq \Gamma$ and $\mathcal{D}_1 \subseteq \dots \subseteq \mathcal{Q}$
- Assume there exists an oracle that outputs an optimal solution to:

$$\min_{(\gamma, Q) \in (\mathcal{C}_n, \mathcal{D}_n)} \hat{R}_\phi(\gamma, Q) = \min_{(\gamma, Q) \in (\mathcal{C}_n, \mathcal{D}_n)} \frac{1}{n} \sum_{i=1}^n \sum_{z \in \mathcal{Z}} \phi(Y_i \gamma(z)) Q(z|X_i)$$

and let (γ_n^*, Q_n^*) denote one such solution.

- Let R_{Bayes}^* denote the minimum Bayes risk:

$$R_{Bayes}^* := \inf_{(\gamma, Q) \in (\Gamma, \mathcal{Q})} R_{Bayes}(\gamma, Q).$$

- Excess Bayes risk: $R_{Bayes}(\gamma_n^*, Q_n^*) - R_{Bayes}^*$

Setup

- *Approximation error:*

$$\mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n) = \inf_{(\gamma, Q) \in (\mathcal{C}_n, \mathcal{D}_n)} \{R_\phi(\gamma, Q)\} - R_\phi^*$$

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

- *Estimation error:*

$$\mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) = \mathbb{E} \sup_{(\gamma, Q) \in (\mathcal{C}_n, \mathcal{D}_n)} \left| \hat{R}_\phi(\gamma, Q) - R_\phi(\gamma, Q) \right|$$

where the expectation is taken with respect to the measure $\mathbb{P}^n(X, Y)$

Bayes Consistency for Choice of (Q, λ)

Theorem 2.

Under the stated conditions:

$$R_{Bayes}(\gamma_n^*, Q_n^*) - R_{Bayes}^* \leq \frac{2}{c} \left\{ 2\mathcal{E}_1(\mathcal{C}_n, \mathcal{D}_n) + \mathcal{E}_0(\mathcal{C}_n, \mathcal{D}_n) + 2M_n \sqrt{2 \frac{\ln(2/\delta)}{n}} \right\}$$

- Thus, under the usual kinds of conditions that drive approximation and estimation error to zero, and under the additional condition on ϕ :

$$M_n := \max_{y \in \{-1, +1\}} \sup_{(\gamma, Q) \in (\mathcal{C}_n, \mathcal{D}_n)} \sup_{z \in \mathcal{Z}} |\phi(y\gamma(z))| < +\infty,$$

we obtain Bayes consistency (for the class of ϕ obtained from $f(u) = -c \min\{u, 1\} + au + b$)

Universal Equivalence of Loss Functions

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

$$\phi_1 \stackrel{u}{\approx} \phi_2$$

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

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

An Equivalence Theorem

Theorem 3.

$$\phi_1 \stackrel{u}{\approx} \phi_2$$

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; \Rightarrow is not
- In particular, surrogate losses universally equivalent to 0-1 loss are those whose induced f divergence has the form:

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

- Thus we see that *only* such losses yield Bayes consistency for procedures that jointly choose (Q, λ)

Estimation of Divergences

- Given i.i.d. $\{x_1, \dots, x_n\} \sim \mathbb{Q}$, $\{y_1, \dots, y_n\} \sim \mathbb{P}$
 - \mathbb{P}, \mathbb{Q} are unknown multivariate distributions with densities p_0, q_0 wrt Lebesgue measure μ on \mathbb{R}^d
- Consider the problem of estimating a divergence; e.g., KL divergence:
 - Kullback-Leibler (KL) divergence functional

$$D_K(\mathbb{P}, \mathbb{Q}) = \int p_0 \log \frac{p_0}{q_0} d\mu$$

Existing Work

- Relations to entropy estimation
 - large body of work on functional of one density (Bickel & Ritov, 1988; Donoho & Liu 1991; Birgé & Massart, 1993; Laurent, 1996 and so on)
- KL is a functional of two densities
- Very little work on nonparametric divergence estimation, especially for high-dimensional data
- Little existing work on estimating density ratio per se

Main Idea

- Variational representation of f -divergences:

Lemma 4. *Letting \mathcal{F} be any function class in $\mathcal{X} \rightarrow \mathbb{R}$, there holds:*

$$D_\phi(\mathbb{P}, \mathbb{Q}) \geq \sup_{f \in \mathcal{F}} \int f d\mathbb{Q} - \phi^*(f) d\mathbb{P},$$

with equality if $\mathcal{F} \cap \partial\phi(q_0/p_0) \neq \emptyset$.

ϕ^* denotes the conjugate dual of ϕ

- Implications:
 - obtain an M-estimation procedure for divergence functional
 - also obtain the likelihood ratio function $d\mathbb{P}/d\mathbb{Q}$
 - how to choose \mathcal{F}
 - how to implement the optimization efficiently
 - convergence rate?

Kullback-Leibler Divergence

- For the Kullback-Leibler divergence:

$$D_K(\mathbb{P}, \mathbb{Q}) = \sup_{g>0} \int \log g \, d\mathbb{P} - \int g d\mathbb{Q} + 1.$$

- Furthermore, the supremum is attained at $g = p_0/q_0$.

M-Estimation Procedure

- Let \mathcal{G} be a function class: $\mathcal{X} \rightarrow \mathbb{R}_+$
- $\int d\mathbb{P}_n$ and $\int d\mathbb{Q}_n$ denote the expectation under empirical measures \mathbb{P}_n and \mathbb{Q}_n , respectively
- One possible estimator has the following form:

$$\hat{D}_K = \sup_{g \in \mathcal{G}} \int \log g d\mathbb{P}_n - \int g d\mathbb{Q}_n + 1.$$

- Supremum is attained at \hat{g}_n , which estimates the likelihood ratio p_0/q_0

Convex Empirical Risk with Penalty

- In practice, control the size of the function class \mathcal{G} by using a penalty
- Let $I(g)$ be a measure of complexity for g
- Decompose \mathcal{G} as follows:

$$\mathcal{G} = \cup_{1 \leq M \leq \infty} \mathcal{G}_M,$$

where \mathcal{G}_M is restricted to g for which $I(g) \leq M$.

- The estimation procedure involves solving:

$$\hat{g}_n = \operatorname{argmin}_{g \in \mathcal{G}} \int g d\mathbb{Q}_n - \int \log g d\mathbb{P}_n + \frac{\lambda_n}{2} I^2(g).$$

Convergence Rates

Theorem 5. *When λ_n vanishes sufficiently slowly:*

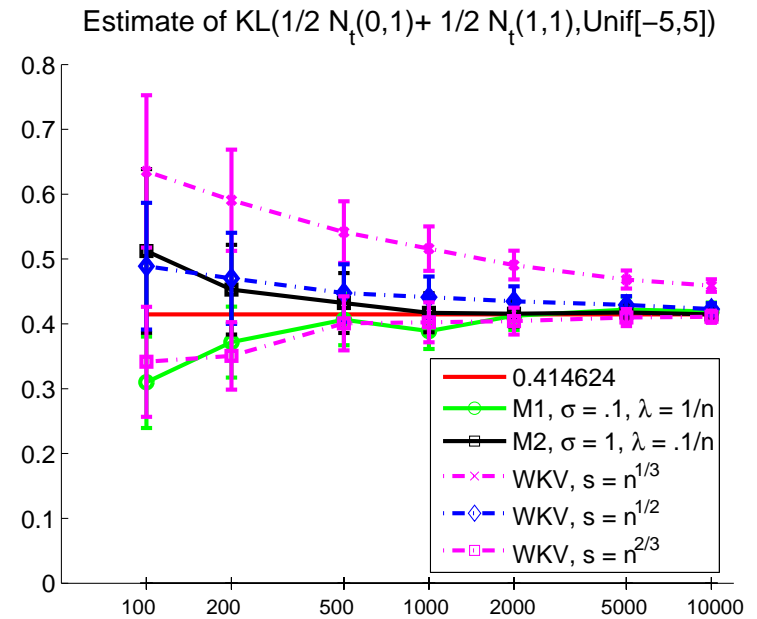
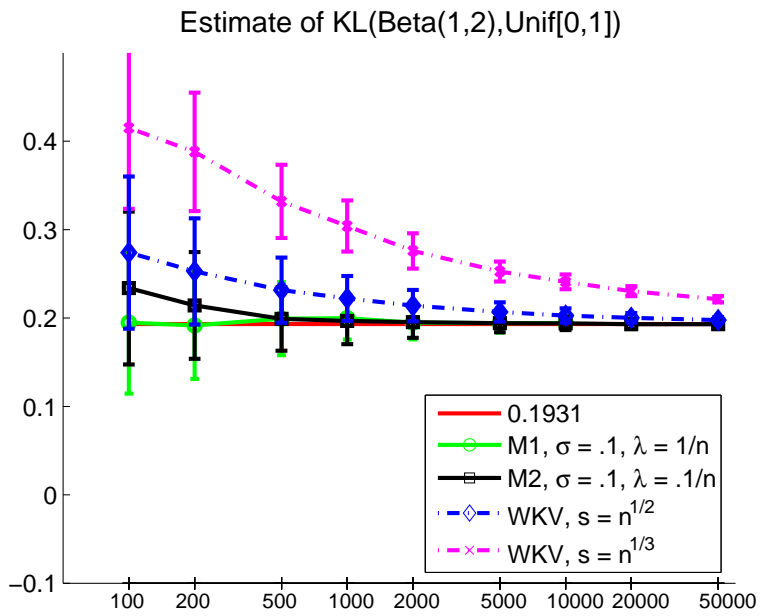
$$\lambda_n^{-1} = O_P(n^{2/(2+\gamma)})(1 + I(g_0)),$$

then under \mathbb{P} :

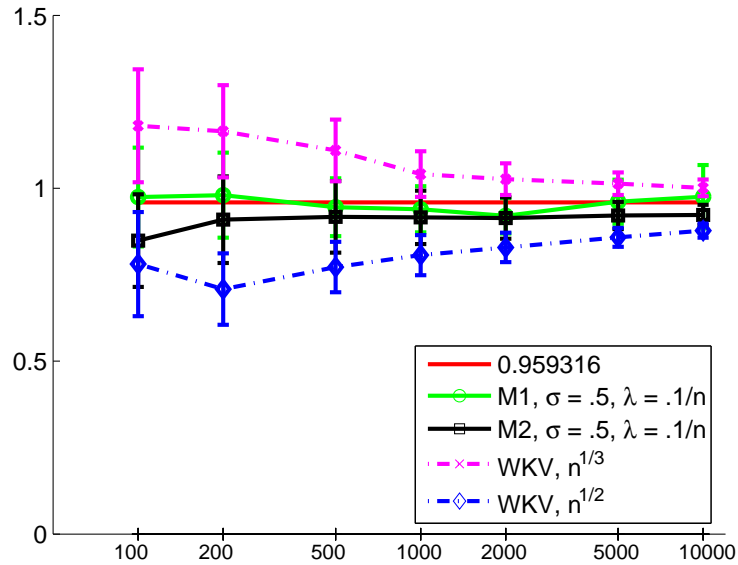
$$h_{\mathbb{Q}}(g_0, \hat{g}_n) = O_P(\lambda_n^{1/2})(1 + I(g_0))$$

$$I(\hat{g}_n) = O_P(1 + I(g_0)).$$

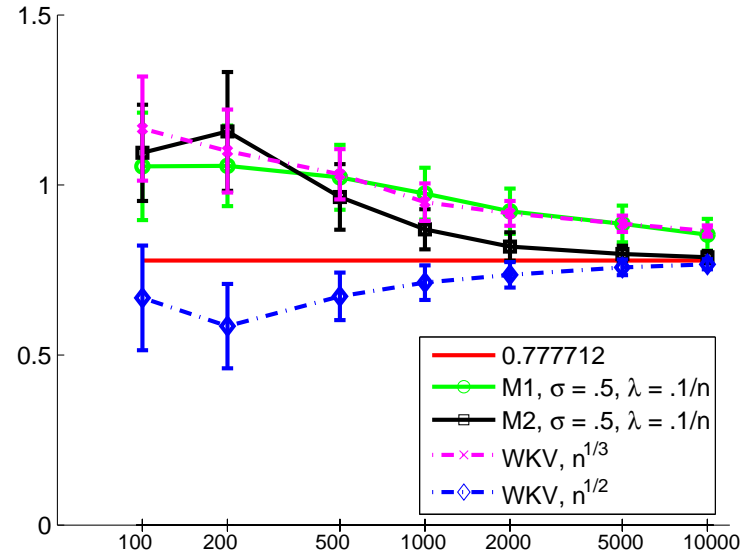
Results



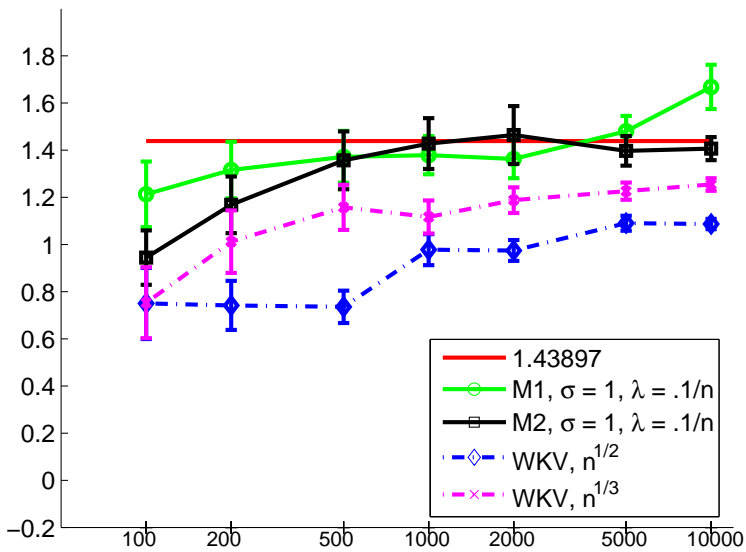
Estimate of $KL(N_t(0, I_2), N_t(1, I_2))$



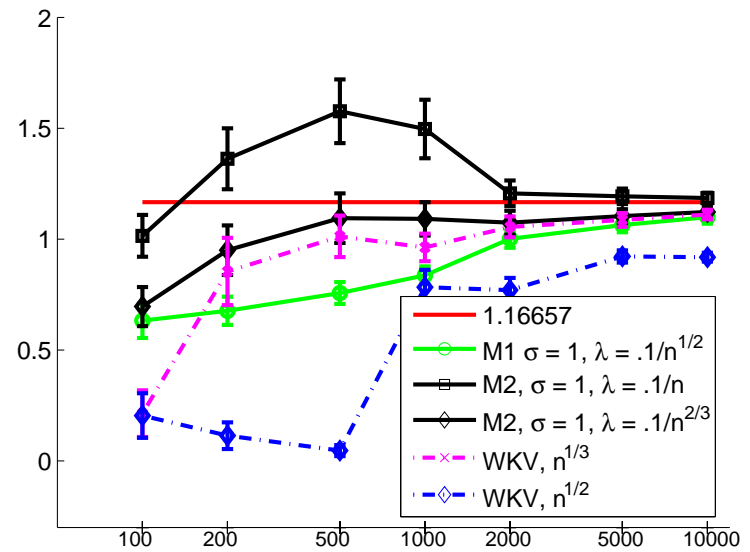
Estimate of $KL(N_t(0, I_2), \text{Unif}[-3, 3]^2)$



Estimate of $KL(N_t(0, I_3), N_t(1, I_3))$



Estimate of $KL(N_t(0, I_3), \text{Unif}[-3, 3]^3)$



Conclusions

- Formulated a precise link between f -divergences and surrogate loss functions
- Decision-theoretic perspective on f -divergences
- Equivalent classes of loss functions
- Can design new convex surrogate loss functions that are equivalent (in a deep sense) to 0-1 loss
 - Applications to the Bayes consistency of procedures that jointly choose an experimental design and a classifier
 - Applications to the estimation of divergences and entropy