Surrogate loss functions, divergences and decentralized detection

XuanLong Nguyen
Department of Electrical Engineering and Computer Science
U.C. Berkeley

Advisors: Michael Jordan & Martin Wainwright

Talk outline

- nonparametric decentralized detection algorithm
  - use of surrogate loss functions and marginalized kernels
  - use of convex analysis
- study of surrogate loss functions and divergence functionals
  - correspondence of losses and divergences
  - M-estimator of divergences (e.g., Kullback-Leibler divergence)

Decentralized decision-making problem

learning both classifier and experiment

- covariate vector X and hypothesis (label) $Y = \pm 1$
- we do not have access directly to X in order to determine Y
- learn jointly the mapping (Q, γ)

$$X \xrightarrow{Q} Z \xrightarrow{\gamma} Y$$

Decentralized decision-making problem

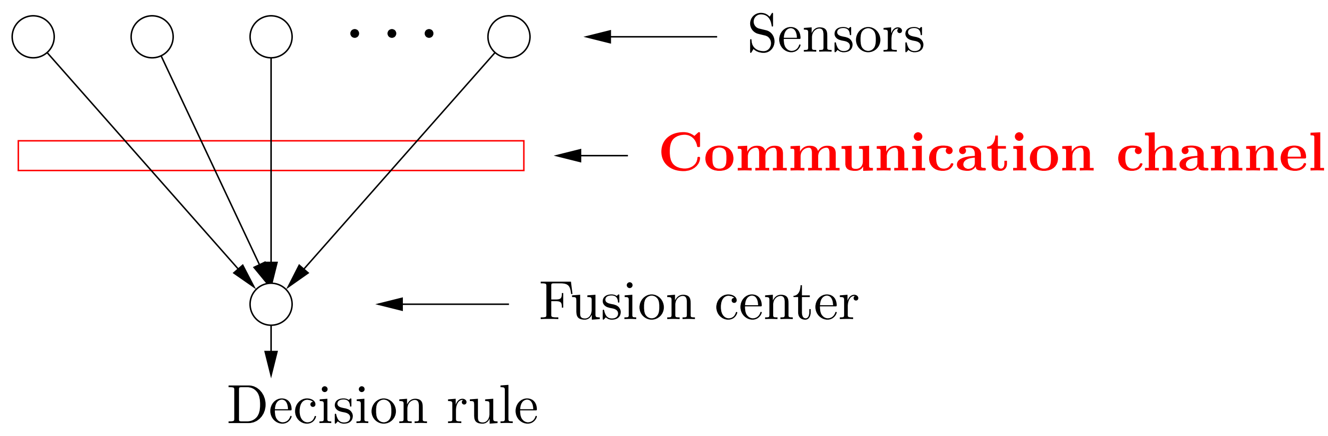
learning both classifier and experiment

- covariate vector X and hypothesis (label) $Y = \pm 1$
- we do not have access directly to X in order to determine Y
- learn jointly the mapping (Q, γ)

$$X \xrightarrow{Q} Z \xrightarrow{\gamma} Y$$

- roles of "experiment" Q:
  - due to data collection constraints (e.g., decentralization)
  - data transmission constraints
  - choice of variates (feature selection)
  - dimensionality reduction scheme

A decentralized detection system

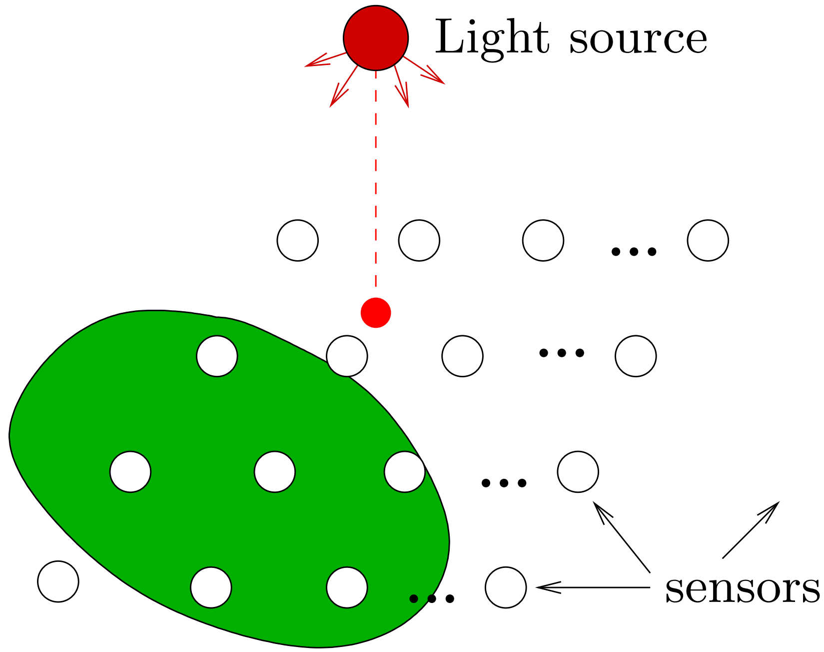


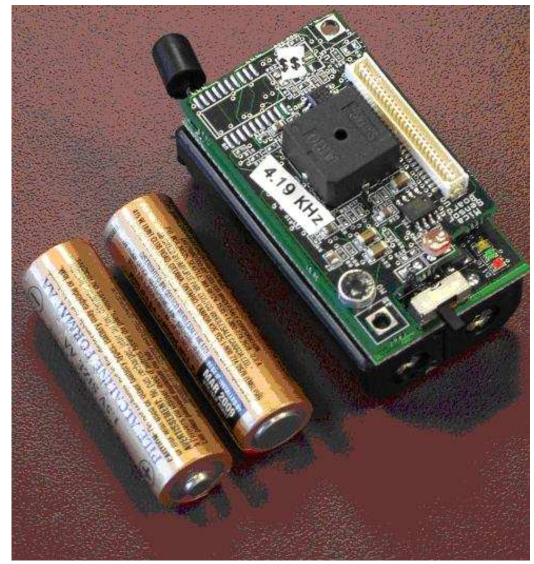
- **Decentralized setting:** Communication constraints between sensors and fusion center (e.g., bit constraints)

- **Goal:** Design decision rules for sensors and fusion center

- **Criterion:** Minimize *probability of incorrect detection*

Concrete example – wireless sensor network





Set-up:

- wireless network of tiny sensor motes, each equiped with light/ humidity/ temperature sensing capabilities
- measurement of signal strength ([0–1024] in magnitude, or 10 bits)

Goal: is there a forest fire in a certain region?

Related work

- Classical work on classification/detection:
  - completely centralized
  - no consideration of communication-theoretic infrastructure

Related work

- Classical work on classification/detection:
  - completely centralized
  - no consideration of communication-theoretic infrastructure

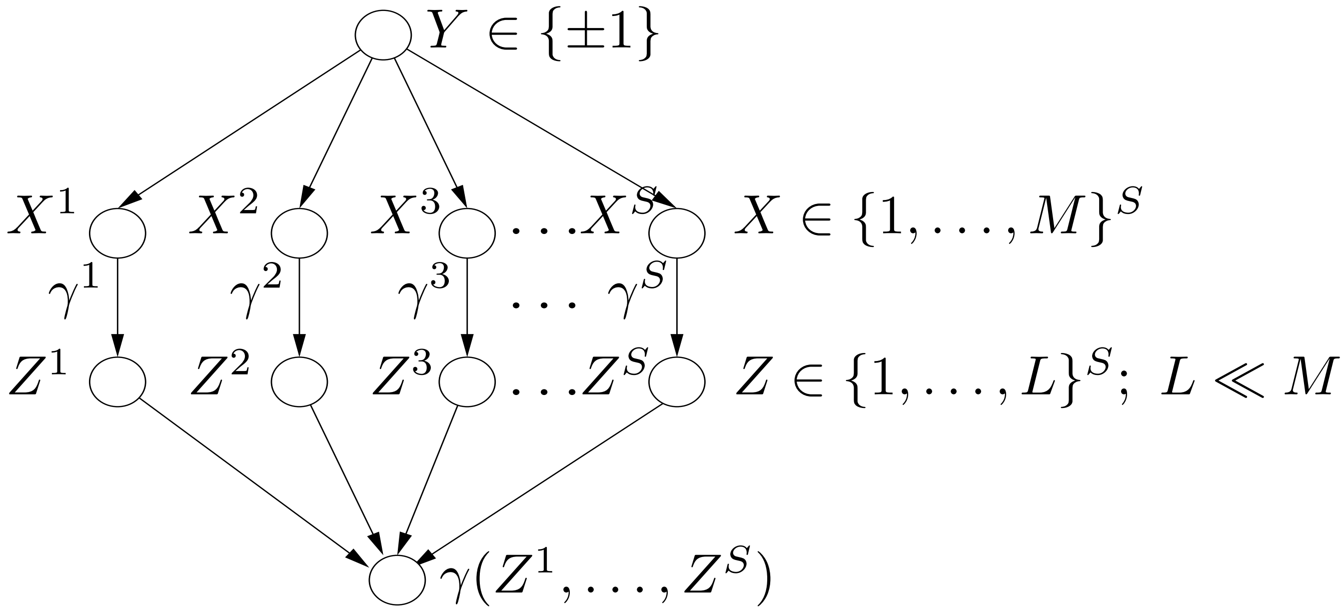
- Decentralized detection in signal processing (e.g., Tsitsiklis, 1993)
  - joint distribution assumed to be known
  - locally-optimal rules under conditional independence assumptions (i.e., Naive Bayes)

Overview of our approach

- Treat as a nonparametric estimation (learning) problem
  - under constraints from a distributed system

- Use kernel methods and convex surrogate loss functions
  - tools from convex optimization to derive an efficient algorithm

Problem set-up



Problem: Given training data $(x_i, y_i)_{i=1}^n$, find the decision rules $(\gamma^1, \ldots, \gamma^s; \gamma)$ so as to minimize the detection error probability:

$$P(Y \neq \gamma(Z^1, \ldots, Z^s)).$$

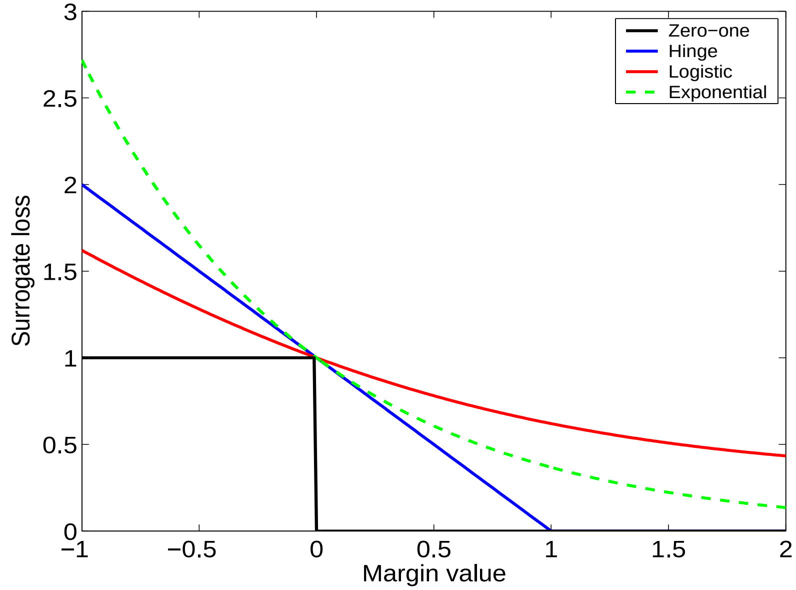
Kernel methods for classification

- Classification: Learn $\gamma(z)$ that predicts label y
- $K(z, z')$ is a *symmetric positive semidefinite* kernel function
- *feature space* $\mathcal{H}$ in which K acts as an inner product, i.e., $K(z, z') = \langle \Psi(z), \Psi(z') \rangle$
- Kernel-based algorithm finds linear function in $\mathcal{H}$, i.e.

$$\gamma(z) = \langle \mathrm{w}, \Psi(z) \rangle = \sum_{i=1}^{n} \alpha_i K(z_i, z)$$

- Advantages:
  - kernel function classes are sufficiently rich for many applications
  - optimizing over kernel function classes is computionally efficient

Convex surrogate loss function ϕ to 0-1 loss

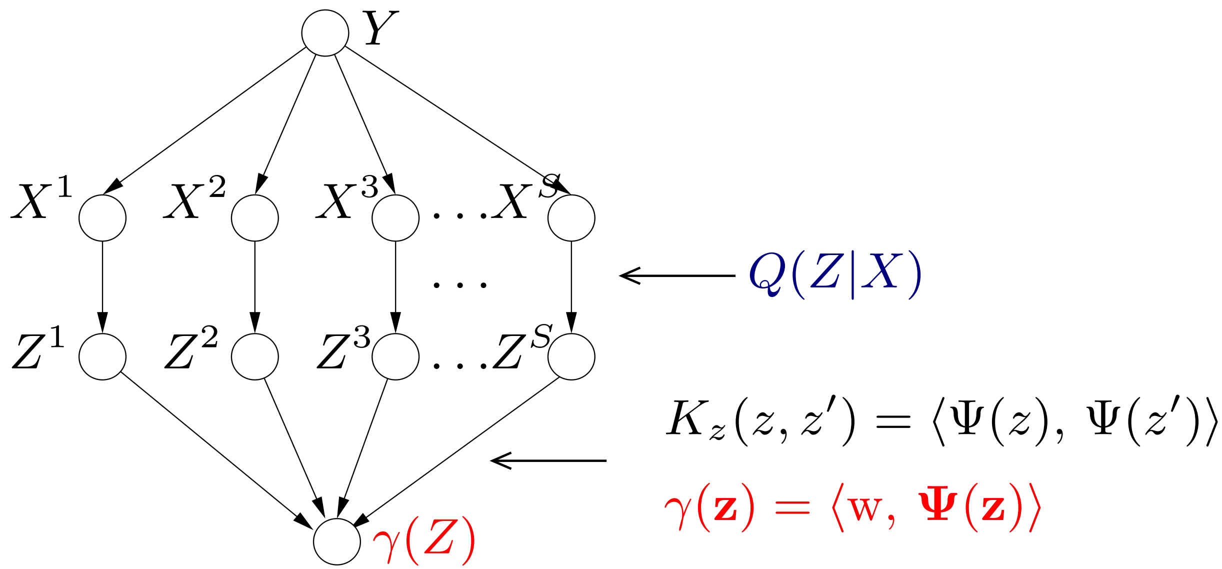


- minimizing (regularized) empirical ϕ-risk $\hat{E}\phi(Y\gamma(Z))$:

$$\min_{\gamma \in \mathcal{H}} \sum_{i=1}^{n} \phi(y_i \gamma(z_i)) + \frac{\lambda}{2} \|\gamma\|^2,$$

- $(z_i, y_i)_{i=1}^n$ are training data in $\mathcal{Z} \times \{\pm 1\}$
- ϕ is a convex *loss function* (surrogate to non-convex 0-1 loss)

Stochastic decision rules at each sensor



- Approximate deterministic sensor decisions by stochastic rules $Q(Z|X)$
- Sensors do not communicate directly $\Longrightarrow$ factorization: $Q(Z|X) = \prod_{t=1}^{S} Q^t(Z^t|X^t)$
- The overall decision rule is represented by $\begin{cases} \mathbf{Q} = \prod \mathbf{Q^t}, \\ \gamma(\mathbf{z}) = \langle \mathrm{w}, \boldsymbol{\Psi}(\mathbf{z}) \rangle \end{cases}$

High-level strategy:

Joint optimization

- Minimize over (Q, γ) an empirical version of $\mathbb{E}\phi(Y\gamma(Z))$
- Joint minimization:
  - fix Q, optimize over γ: A simple convex problem
  - fix γ, perform a gradient update for Q, sensor by sensor

Approximating empirical ϕ-risk

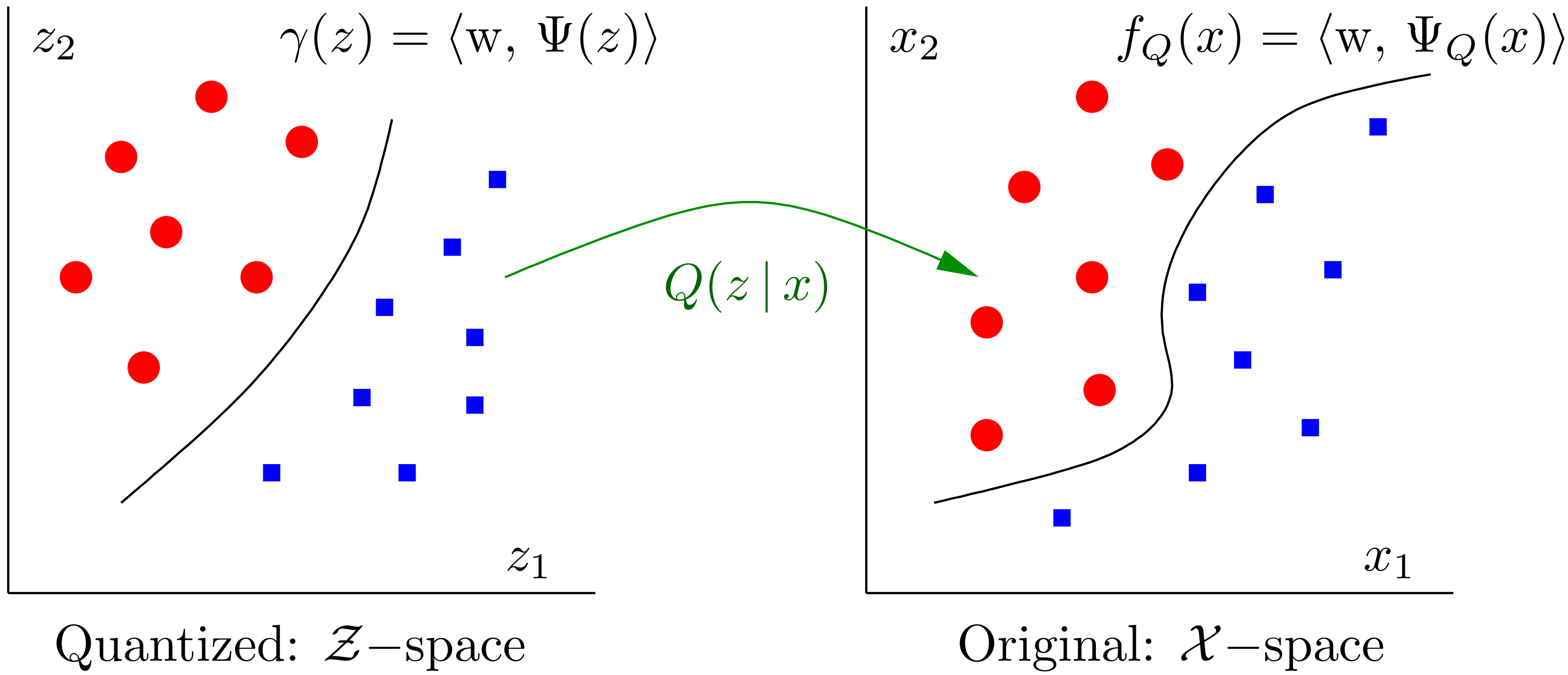
- The regularized empirical ϕ-risk $\hat{\mathbb{E}}\phi(Y\gamma(Z))$ has the form:

$$G_0 = \sum_z \sum_{i=1}^n \phi(y_i\gamma(z))Q(z|x_i) + \frac{\lambda}{2}||\mathrm{w}||^2$$

- **Challenge:** even evaluating G_0 at a single point is intractable
  Requires summing over L^S possible values for z

- **Idea**:
  - approximate G_0 by another objective function G
  - G is ϕ-risk of a "marginalized" feature space
  - $G_0 \equiv G$ for deterministic Q

"Marginalizing" over feature space



Stochastic decision rule $Q(z \mid x)$:

- maps between $\mathcal{X}$ and $\mathcal{Z}$
- induces marginalized feature map Ψ_Q from base map Ψ (or marginalized kernel K_Q from base kernel K)

Marginalized feature space $\{\Psi_Q(x)\}$

Marginalized feature space $\{\Psi_Q(x)\}$

- Define a new feature space $\Psi_Q(x)$ and a linear function over $\Psi_Q(x)$:

$$\begin{cases} \Psi_Q(x) = \sum_z Q(z|x)\Psi(z) & \Longleftarrow \text{ Marginalization over } z \\ f_Q(x) = \langle w,\, \Psi_Q(x)\rangle \end{cases}$$

Marginalized feature space $\{\Psi_Q(x)\}$

- Define a new feature space $\Psi_Q(x)$ and a linear function over $\Psi_Q(x)$:

$$\begin{cases} \Psi_Q(x) = \sum_z Q(z|x)\Psi(z) & \Longleftarrow \text{ Marginalization over } z \\ f_Q(x) = \langle w, \Psi_Q(x)\rangle \end{cases}$$

- The alternative objective function G is the ϕ-risk for f_Q:

$$G = \sum_{i=1}^{n} \phi(y_i f_Q(x_i)) + \frac{\lambda}{2}\|\mathrm{w}\|^2$$

Marginalized feature space $\{\Psi_Q(x)\}$

- Define a new feature space $\Psi_Q(x)$ and a linear function over $\Psi_Q(x)$:

$$\begin{cases} \Psi_Q(x) = \sum_z Q(z|x)\Psi(z) & \Longleftarrow \text{ Marginalization over } z \\ f_Q(x) = \langle w,\, \Psi_Q(x)\rangle \end{cases}$$

- The alternative objective function G is the ϕ-risk for f_Q:

$$G = \sum_{i=1}^{n} \phi(y_i f_Q(x_i)) + \frac{\lambda}{2}\|\mathrm{w}\|^2$$

- $\Psi_Q(x)$ induces a marginalized kernel over $\mathcal{X}$:

$$K_Q(x, x') := \langle \Psi_Q(x),\, \Psi_Q(x')\rangle = \sum_{z,z'} Q(z|x)Q(z'|x')\, K_z(z, z')$$

$\Rightarrow$ Marginalization taken over message z conditioned on sensor signal x

Marginalized kernels

- Have been used to derive kernel functions from generative models (e.g. Tsuda, 2002)

- Marginalized kernel $K_Q(x, x')$ is defined as:

$$K_Q(x, x') := \sum_{z,z'} \underbrace{Q(z|x)Q(z'|x')}_{\text{Factorized distributions}} \underbrace{K_z(z, z')}_{\text{Base kernel}} ,$$

- If $K_z(z, z')$ is decomposed into smaller components of z and z', then $K_Q(x, x')$ can be computed efficiently (in polynomial-time)

Centralized and decentralized function

- Centralized decision function obtained by minimizing ϕ-risk:

$$f_Q(x) = \langle \mathrm{w},\, \Psi_Q(x) \rangle$$

  - f_Q has direct access to sensor signal x

Centralized and decentralized function

- Centralized decision function obtained by minimizing ϕ-risk:

$$f_Q(x) = \langle \mathrm{w},\, \Psi_Q(x) \rangle$$

  - f_Q has direct access to sensor signal x

- Optimal w also define decentralized decision function:

$$\gamma(z) = \langle \mathrm{w},\, \Psi(z) \rangle$$

  - γ has access only to quantized version z

Centralized and decentralized function

- Centralized decision function obtained by minimizing ϕ-risk:

$$f_Q(x) = \langle \mathrm{w}, \, \Psi_Q(x) \rangle$$

  - f_Q has direct access to sensor signal x

- Optimal w also define decentralized decision function:

$$\gamma(z) = \langle \mathrm{w}, \, \Psi(z) \rangle$$

  - γ has access only to quantized version z

- Decentralized γ behaves *on average* like the centralized f_Q:

$$f_Q(x) = \mathbb{E}[\gamma(Z)|x]$$

Optimization algorithm

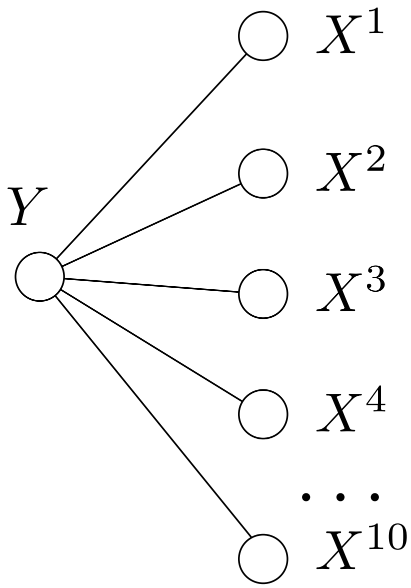
Goal: Solve the problem:

$$\inf_{\mathrm{w};Q} G(\mathrm{w};Q) \;:=\; \sum_i \phi\Big(y_i\langle \mathrm{w}, \sum_z Q(z|x_i)\Psi(z)\rangle\Big) + \frac{\lambda}{2}||\mathrm{w}||^2$$

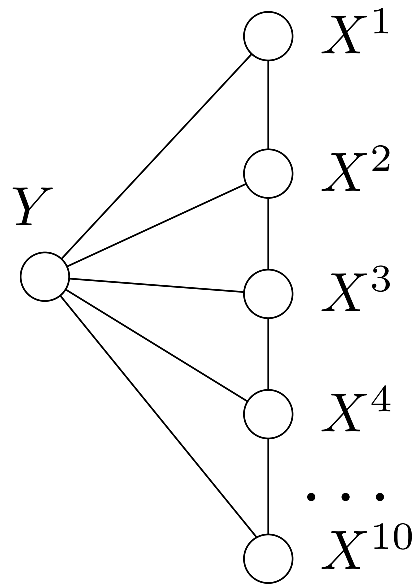
- Finding optimal weight vector:
  - G is convex in w with Q fixed
  - solve dual problem (quadratic convex program) to obtain optimal w(Q)
- Finding optimal decision rules:
  - G is convex in Q^t with w and all other $\{Q^r, r \neq t\}$ fixed
  - efficient computation of *subgradient* for G at optimal $(\mathrm{w}(Q), Q)$

Overall: Efficient joint minimization by blockwise coordinate descent

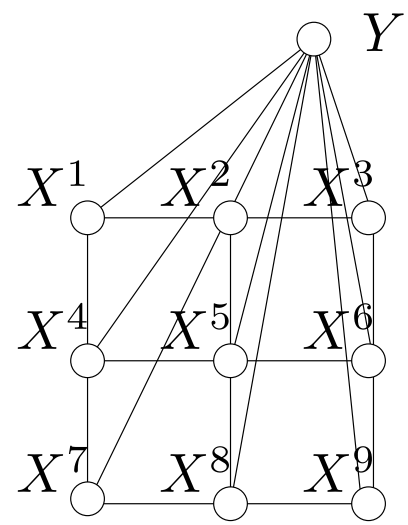
Simulated sensor networks



Naive Bayes net

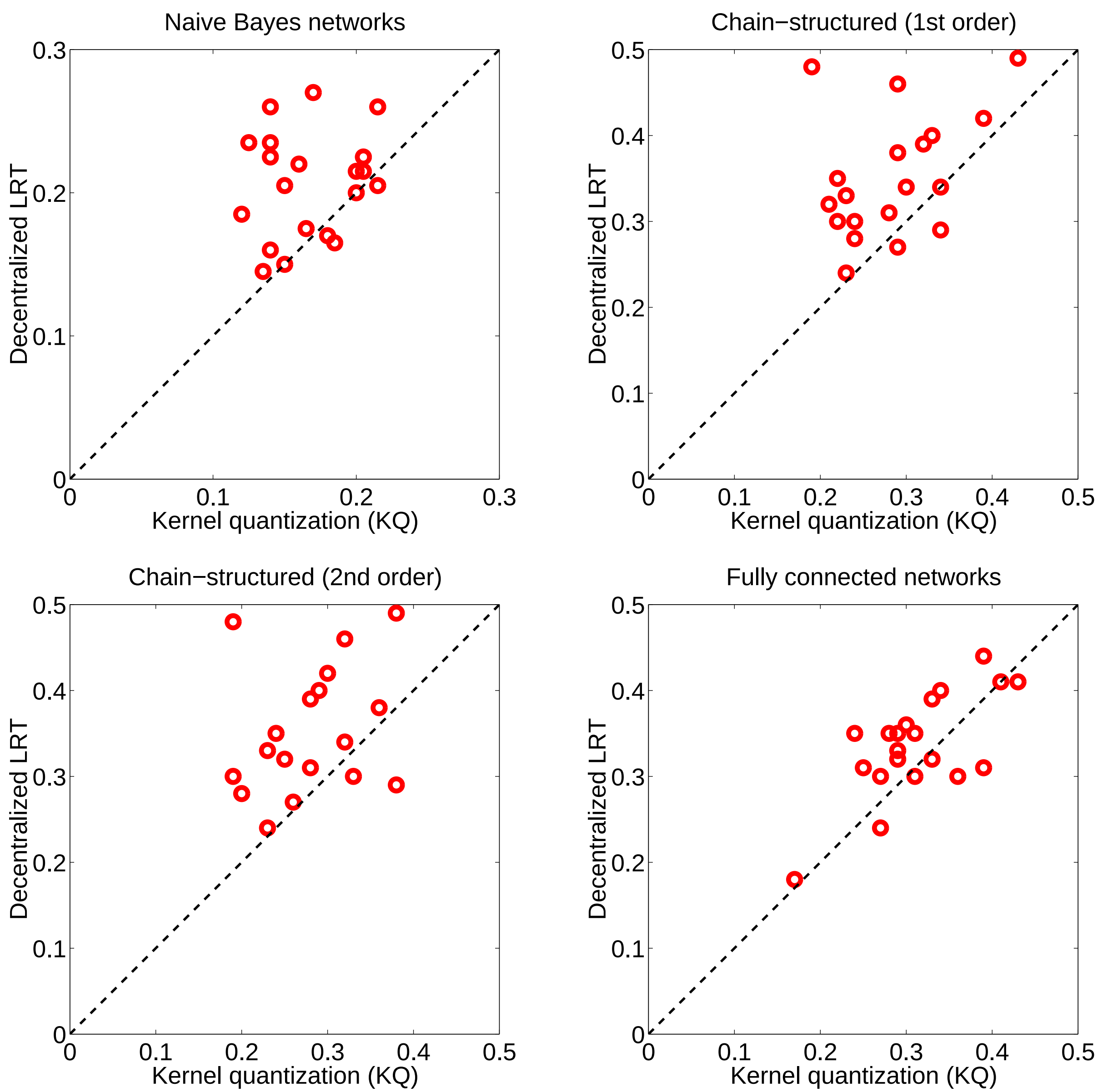


Chain-structured network

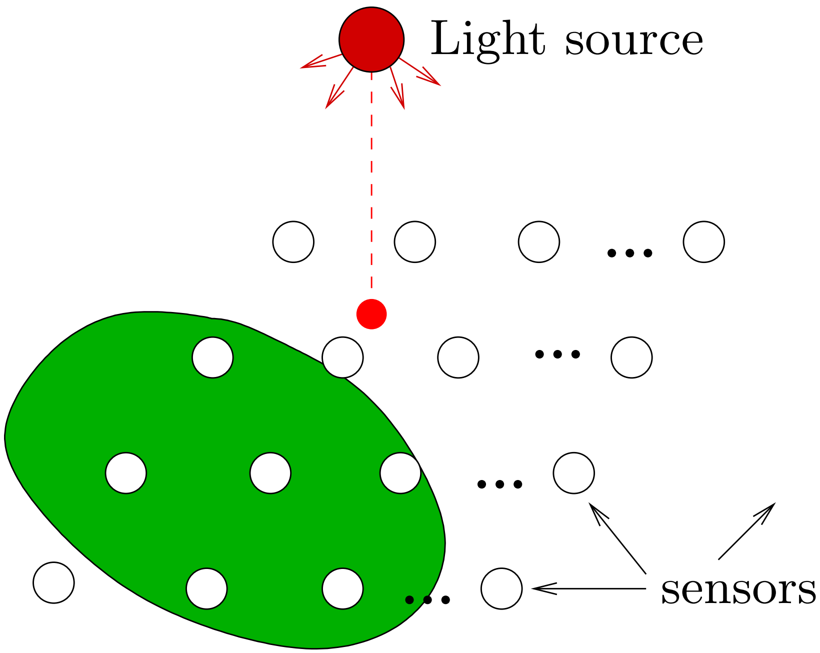


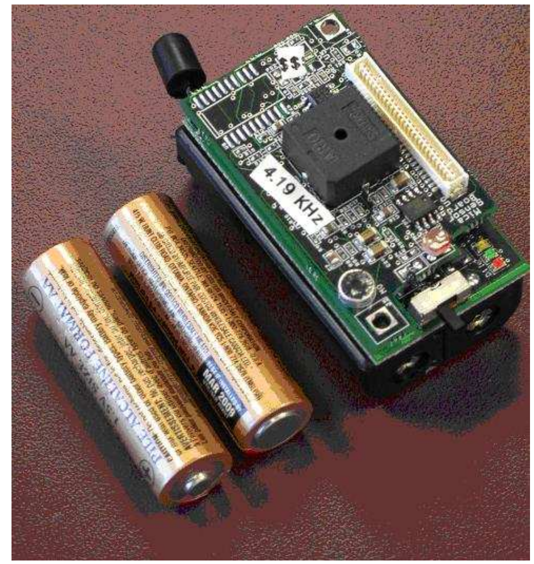
Spatially-dependent network

Kernel Quantization vs. Decentralized LRT



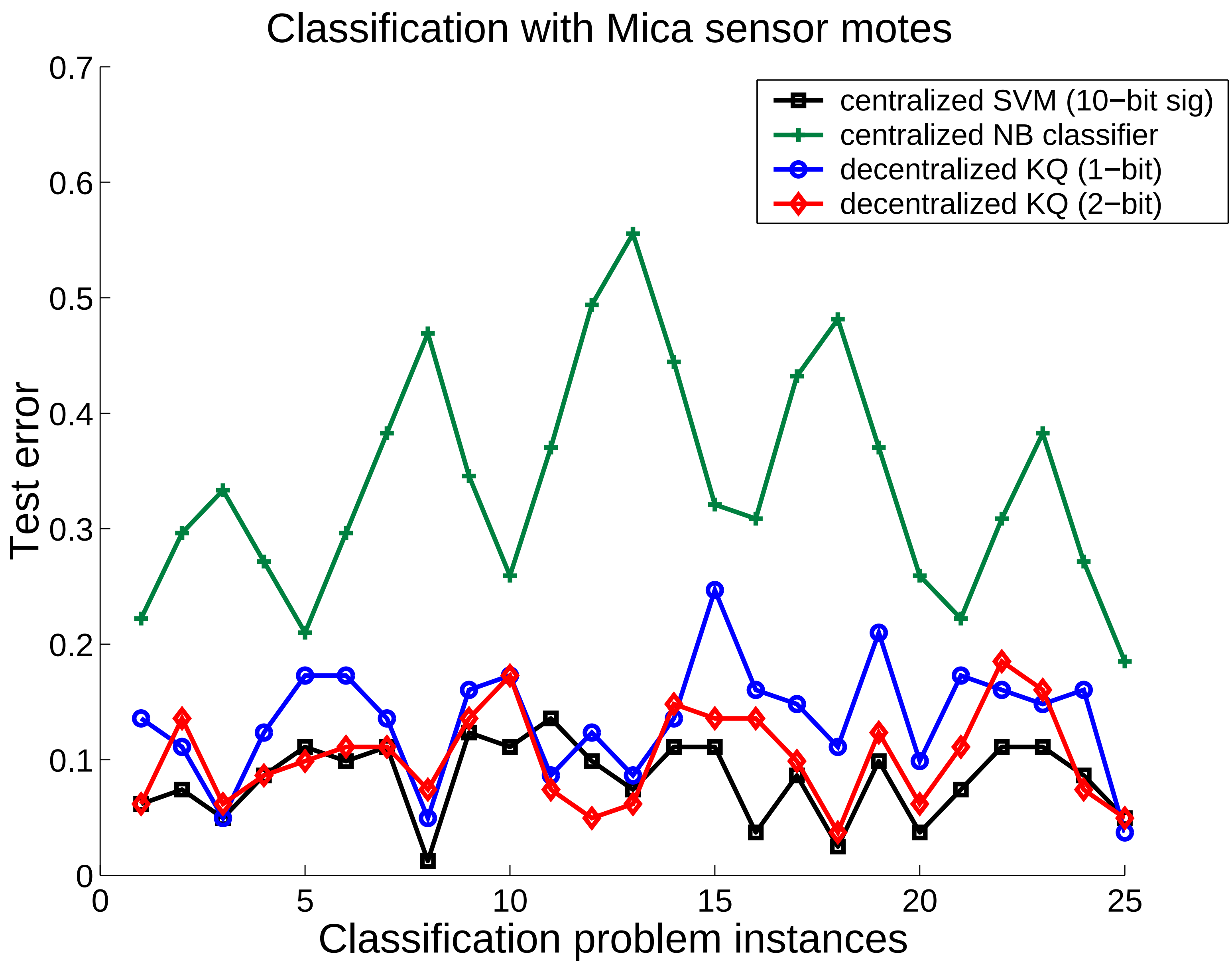
Wireless network with tiny Berkeley motes





- $5 \times 5 = 25$ tiny sensor motes, each equipped with a light receiver
- Light signal strength requires **10-bit** ([0–1024] in magnitude)
- Perform classification with respect to different regions
- Each problem has 25 training positions, 81 test positions

  (Data collection courtesy Bruno Sinopoli)


Classification with Mica sensor motes
Test error
Classification problem instances
centralized SVM (10−bit sig)
centralized NB classifier
decentralized KQ (1−bit)
decentralized KQ (2−bit)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
5
10
15
20
25

Outline

- nonparametric decentralized detection algorithm
  - use of surrogate loss functions and marginalized kernels

- study of surrogate loss functions and divergence functionals
  - correspondence of losses and divergences
  - M-estimator of divergences (e.g., Kullback-Leibler divergence)

Consistency question

- recall that our decentralized algorithm essentially solves

$$\min_{\gamma, Q} \hat{\mathbb{E}} \phi(Y, \gamma(Z))$$

- does this also imply optimality in the sense of 0-1 loss?

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

Consistency question

- recall that our decentralized algorithm essentially solves

$$\min_{\gamma, Q} \hat{\mathbb{E}} \phi(Y, \gamma(Z))$$

- does this also imply optimality in the sense of 0-1 loss?

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

- answers:
  - hinge loss yields consistent estimates
  - all losses corresponding to variational distance yield consistency and we can identify all of them
  - exponential loss, logistic loss do not
- the gist lies in the *correspondence between loss functions and divergence functionals*

Divergence between two distributions

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

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

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

Divergence between two distributions

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

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

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

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

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

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

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

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

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

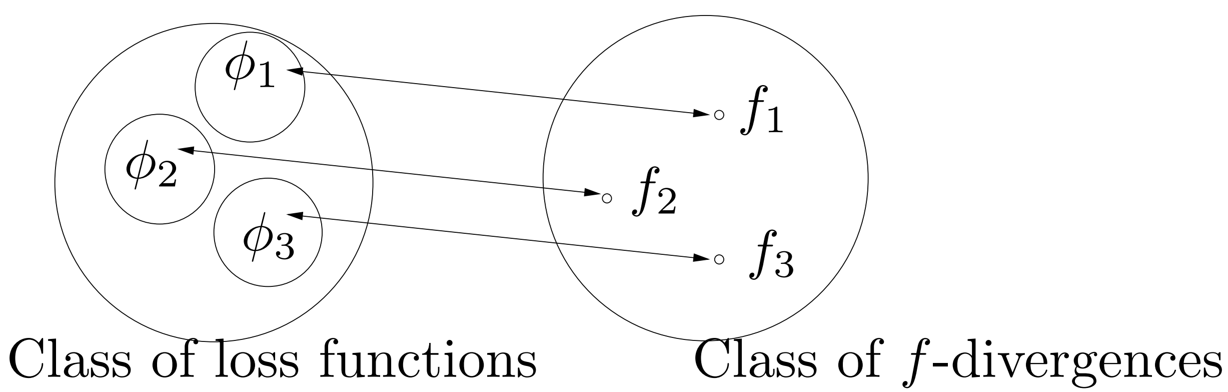
Surrogate loss and f-divergence

Map Q induces measures on Z:

$$\mu(z) := P(Y = 1, z); \quad \pi(z) := P(Y = -1, z)$$

Theorem: Fixing Q, the optimal risk for each ϕ loss is an f-*divergence* for some convex f, and vice versa:

$$R_\phi(Q) = -I_f(\mu, \pi), \quad \text{where} \quad R_\phi(Q) := \min_\gamma \mathbb{E}\phi(Y, \gamma(Z))$$



"Unrolling" divergences by convex duality

- Legendre-Fenchel convex duality: $f(u) = \sup_{v \in \mathbb{R}} uv - f^*(v)$, where f^* is the convex conjugate of f

$$I_f(\mu, \pi) = \int \pi f\left(\frac{\mu}{\pi}\right) d\nu$$

"Unrolling" divergences by convex duality

- Legendre-Fenchel convex duality: $f(u) = \sup_{v \in \mathbb{R}} uv - f^*(v)$, where f^* is the convex conjugate of f

$$I_f(\mu, \pi) = \int \pi f\left(\frac{\mu}{\pi}\right) \, d\nu$$

$$= \int \pi \sup_{\gamma} (\gamma \mu / \pi - f^*(\gamma)) \, d\nu$$

“Unrolling” divergences by convex duality

- Legendre-Fenchel convex duality: $f(u) = \sup_{v \in \mathbb{R}} uv - f^*(v)$, where f^* is the convex conjugate of f

$$I_f(\mu, \pi) = \int \pi f\left(\frac{\mu}{\pi}\right) \, d\nu$$

$$= \int \pi \sup_{\gamma} (\gamma\mu/\pi - f^*(\gamma)) \, d\nu$$

$$= \sup_{\gamma} \int \gamma\mu - f^*(\gamma)\pi \, d\nu$$

“Unrolling” divergences by convex duality

- Legendre-Fenchel convex duality: $f(u) = \sup_{v \in \mathbb{R}} uv - f^*(v)$, where f^* is the convex conjugate of f

$$I_f(\mu, \pi) = \int \pi f\left(\frac{\mu}{\pi}\right) \, d\nu$$

$$= \int \pi \sup_{\gamma} (\gamma\mu/\pi - f^*(\gamma)) \, d\nu$$

$$= \sup_{\gamma} \int \gamma\mu - f^*(\gamma)\pi \, d\nu$$

$$= -\inf_{\gamma} \int f^*(\gamma)\pi - \gamma\mu \, d\nu$$

"Unrolling" divergences by convex duality

- Legendre-Fenchel convex duality: $f(u) = \sup_{v \in \mathbb{R}} uv - f^*(v)$, where f^* is the convex conjugate of f

$$I_f(\mu, \pi) = \int \pi f\left(\frac{\mu}{\pi}\right) d\nu$$

$$= \int \pi \sup_{\gamma}(\gamma\mu/\pi - f^*(\gamma)) \, d\nu$$

$$= \sup_{\gamma} \int \gamma\mu - f^*(\gamma)\pi \, d\nu$$

$$= -\inf_{\gamma} \int f^*(\gamma)\pi - \gamma\mu \, d\nu$$

- The last quantity can be viewed as a risk functional with respect to loss functions $f^*(\gamma)$ and $-\gamma$

Examples

- **0-1 loss**:

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

- **hinge loss**:

  $R_{hinge}(Q) = 2R_{bayes}(Q)$ $\Rightarrow$ variational distance

- **exponential loss**:

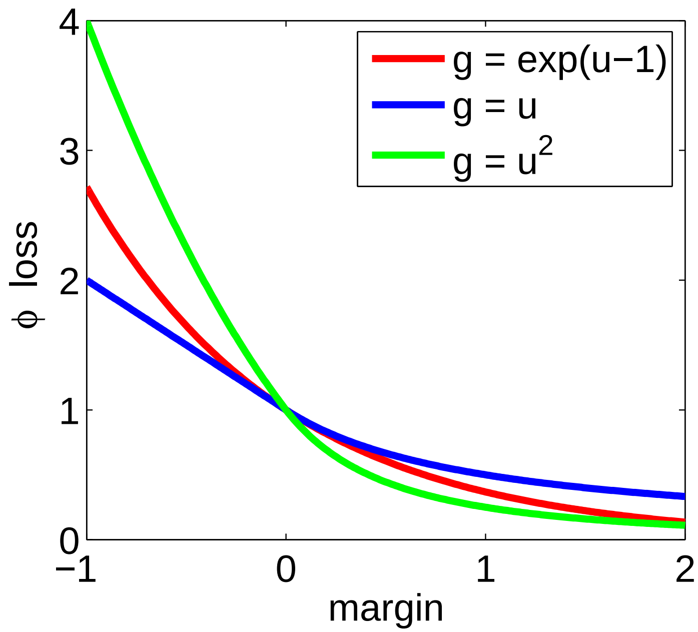
  $R_{exp}(Q) = 1 - \sum_{z \in \mathcal{Z}} (\mu(z)^{1/2} - \pi(z)^{1/2})^2$ $\Rightarrow$ Hellinger distance

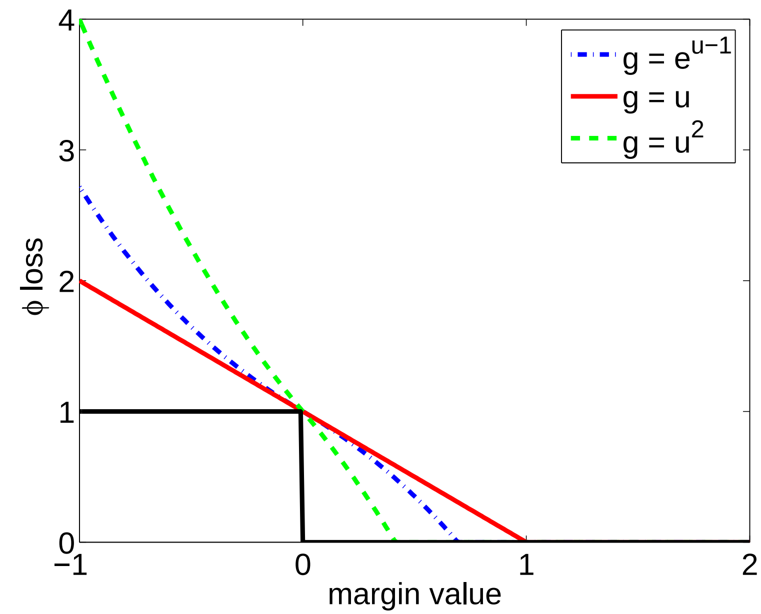
- **logistic loss**:

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

Examples

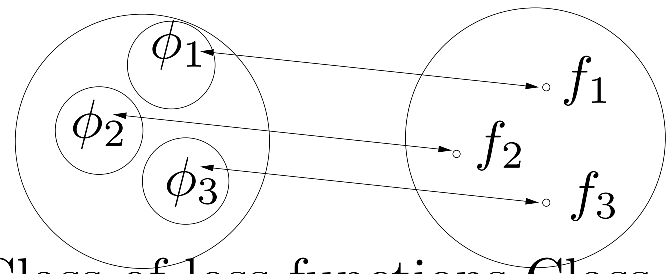
Equivalent surrogage losses corresponding to Hellinger distance (left) and variational distance (right)





the part after 0 is a fixed map of the part before 0!

Comparison of loss functions



Class of loss functions Class of f-divergences

- 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

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

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

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

An equivalence theorem

Theorem:

$$\phi_1 \overset{U}{\approx} \phi_2 \;\; \text{(or, equivalently)} \; f_1 \overset{U}{\approx} f_2$$

if and only if

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

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

- in particular, surrogate losses universally equivalent to $0-1$ loss are those whose induced f divergence has the form:

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

Empirical risk minimization procedure

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

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

- our procedure learns:

$$(\gamma_n^*, Q_n^*) := \text{argmin}_{(\gamma,Q)\in(\mathcal{C}_n,\mathcal{D}_n)} \hat{\mathbb{E}}\phi(Y\gamma(Z))$$

- let $R^*_{bayes} := \inf_{(\gamma,Q)\in(\Gamma,\mathcal{Q})} P(Y \neq \gamma(Z))$ $\Leftarrow$ optimal Bayes error
- our procedure is consistent if

$$R_{bayes}(\gamma_n^*, Q_n^*) - R^*_{bayes} \to 0$$

Consistency result

Theorem: If

- $\cup_{n=1}^{\infty}(\mathcal{C}_n, \mathcal{D}_n)$ is dense in the space of pairs 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.,

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

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

Outline

- nonparametric decentralized detection algorithm
  - use of surrogate loss functions and marginalized kernels

- study of surrogate loss functions and divergence functionals
  - correspondence of losses and divergences
  - M-estimator of divergences (e.g., Kullback-Leibler divergence)

Estimating divergence and likelihood ratio

- given i.i.d $\{x_1, \ldots, x_n\} \sim \mathbb{Q}$, $\{y_1, \ldots, y_n\} \sim \mathbb{P}$
- want to estimate two quantities
  - KL divergence functional

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

  - likelihood ratio function

$$g_0(.) = p_0(.)/q_0(.)$$

Variational characterization

- recall the correspondence:

$$\min_{\gamma} \mathbb{E}\phi(Y, \gamma(Z)) = -I_f(\mu, \pi)$$

- f-divergence can be estimated by minimizing over some associated ϕ-risk functional

Variational characterization

- recall the correspondence:

$$\min_{\gamma} \mathbb{E}\phi(Y, \gamma(Z)) = -I_f(\mu, \pi)$$

- f-divergence can be estimated by minimizing over some associated ϕ-risk functional
- 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 of $\mathcal{X} \to \mathbb{R}_+$
- $\int d\mathbb{P}_n$ and $\int d\mathbb{Q}_n$ denote the expectation under empirical measures $\mathbb{P}_n$ and $\mathbb{Q}_n$, respectively
- our 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 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:

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

Convergence analysis

- for KL divergence estimation, we study

$$|\hat{D}_K - D_K(\mathbb{P}, \mathbb{Q})|$$

- for the likelihood ratio estimation, we use Hellinger distance

$$h^2_{\mathbb{Q}}(g, g_0) := \frac{1}{2} \int (g^{1/2} - g_0^{1/2})^2 \, d\mathbb{Q}.$$

Assumptions for convergence analysis

- true likelihood ratio g_0 is bounded from below by some positive constant:

$$g_0 \geq \eta_0 > 0.$$

Note: we don't assume that $\mathcal{G}$ is bounded away from 0 (not yet)!

- uniform norm of $\mathcal{G}_M$ is Lipchitz with respect to the penalty measure $I(g)$: for any $M \geq 1$:

$$\sup_{g \in \mathcal{G}_M} |g|_\infty \leq cM.$$

- on the entropy of $\mathcal{G}$: For some $0 < \gamma < 2$,

$$\mathcal{H}_\delta^B(\mathcal{G}_M, L_2(\mathbb{Q})) = O(M/\delta)^\gamma.$$

Convergence rates

- when λ_n vanishes sufficiently slowly:

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

- then under $\mathbb{P}$:

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

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

Convergence rates

- when λ_n vanishes sufficiently slowly:

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

- then under $\mathbb{P}$:

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

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

- if $\mathcal{G}$ is bounded away from 0:

$$|D_K - \hat{D}_K| = O_P(\lambda_n^{1/2})(1 + I(g_0)).$$

$\mathcal{G}$ is RKHS function class

- $\{x_i\} \sim \mathbb{Q}$, $\{y_j\} \sim \mathbb{P}$
- $\mathcal{G}$ is a RKHS with Mercer kernel $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$
- $I(g) = \|g\|_{\mathcal{H}}$

$$\hat{g}_n = \mathrm{argmin}_{g \in \mathcal{G}} \int g d\mathbb{Q}_n - \int \log g \; d\mathbb{P}_n + \frac{\lambda_n}{2} \|g\|^2_{\mathcal{H}}$$

$\mathcal{G}$ is RKHS function class

- $\{x_i\} \sim \mathbb{Q}$, $\{y_j\} \sim \mathbb{P}$
- $\mathcal{G}$ is a RKHS with Mercer kernel $k(x, x') = \langle \Phi(x),\, \Phi(x') \rangle$
- $I(g) = \|g\|_{\mathcal{H}}$

$$\hat{g}_n = \mathrm{argmin}_{g \in \mathcal{G}} \int g d\mathbb{Q}_n - \int \log g \, d\mathbb{P}_n + \frac{\lambda_n}{2} \|g\|_{\mathcal{H}}^2$$

- Convex dual formulation:

$$\alpha := \mathrm{argmax} \frac{1}{n} \sum_{j=1}^{n} \log \alpha_j - \frac{1}{2\lambda_n} \| \sum_{j=1}^{n} \alpha_j \Phi(y_j) - \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \|^2$$

$$\hat{D}_K(\mathbb{P}, \mathbb{Q}) := -\frac{1}{n} \sum_{j=1}^{n} \log \alpha_j - \log n$$

$\log \mathcal{G}$ is RKHS function class

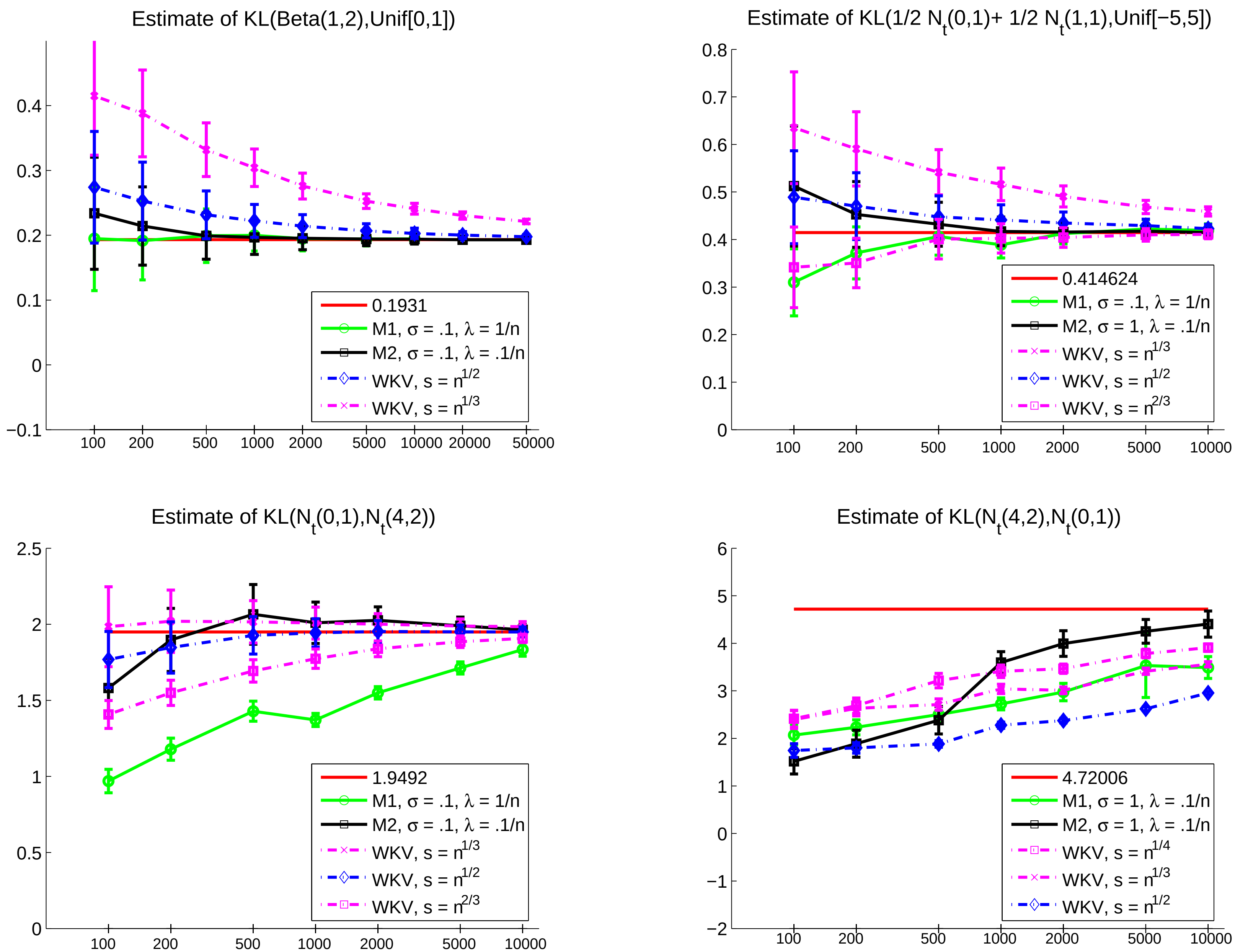
- $\{x_i\} \sim \mathbb{Q}$, $\{y_j\} \sim \mathbb{P}$
- $\log \mathcal{G}$ is a RKHS with Mercer kernel $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$
- $I(g) = \| \log g \|_{\mathcal{H}}$

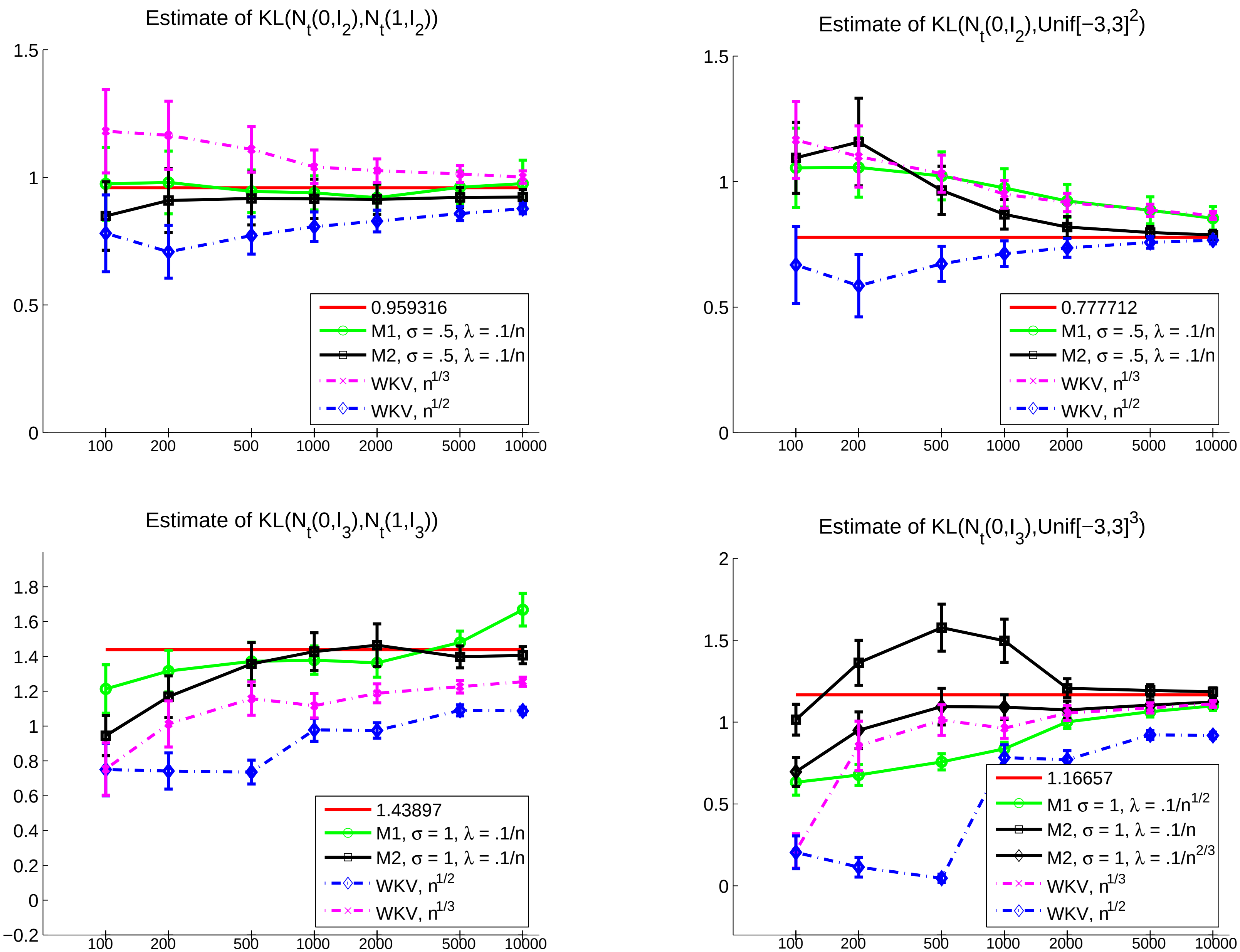
$$\hat{g}_n = \mathrm{argmin}_{g \in \mathcal{G}} \int g d\mathbb{Q}_n - \int \log g \, d\mathbb{P}_n + \frac{\lambda_n}{2} \| \log g \|_{\mathcal{H}}^2$$

- Convex dual formulation:

$$\alpha := \mathrm{argmax} \frac{1}{n} \sum_{i=1}^{n} \left(\alpha_i \log \alpha_i + \alpha_i \log \frac{n}{e} \right) - \frac{1}{2\lambda_n} \| \sum_{i=1}^{n} \alpha_i \Phi(x_i) - \frac{1}{n} \sum_{j=1}^{n} \Phi(y_j) \|^2$$

$$\hat{D}_K(\mathbb{P}, \mathbb{Q}) := 1 + \sum_{i=1}^{n} \alpha_i \log \alpha_i + \alpha_i \log \frac{n}{e}$$


Estimate of KL(Beta(1,2),Unif[0,1])
0.1931
M1, σ = .1, λ = 1/n
M2, σ = .1, λ = .1/n
WKV, s = $n^{1/2}$
WKV, s = $n^{1/3}$
Estimate of KL(1/2 N_t(0,1)+ 1/2 N_t(1,1),Unif[−5,5])
0.414624
M1, σ = .1, λ = 1/n
M2, σ = 1, λ = .1/n
WKV, s = $n^{1/3}$
WKV, s = $n^{1/2}$
WKV, s = $n^{2/3}$
Estimate of KL(N_t(0,1),N_t(4,2))
1.9492
M1, σ = .1, λ = 1/n
M2, σ = .1, λ = .1/n
WKV, s = $n^{1/3}$
WKV, s = $n^{1/2}$
WKV, s = $n^{2/3}$
Estimate of KL(N_t(4,2),N_t(0,1))
4.72006
M1, σ = 1, λ = .1/n
M2, σ = 1, λ = .1/n
WKV, s = $n^{1/4}$
WKV, s = $n^{1/3}$
WKV, s = $n^{1/2}$



Estimate of KL(N_t(0,I_2),N_t(1,I_2))
0.959316
M1, σ = .5, λ = .1/n
M2, σ = .5, λ = .1/n
WKV, n^1/3
WKV, n^1/2
Estimate of KL(N_t(0,I_2),Unif[−3,3]^2)
0.777712
M1, σ = .5, λ = .1/n
M2, σ = .5, λ = .1/n
WKV, n^1/3
WKV, n^1/2
Estimate of KL(N_t(0,I_3),N_t(1,I_3))
1.43897
M1, σ = 1, λ = .1/n
M2, σ = 1, λ = .1/n
WKV, n^1/2
WKV, n^1/3
Estimate of KL(N_t(0,I_3),Unif[−3,3]^3)
1.16657
M1 σ = 1, λ = .1/n^1/2
M2, σ = 1, λ = .1/n
M2, σ = 1, λ = .1/n^2/3
WKV, n^1/3
WKV, n^1/2

Conclusion

- nonparametric decentralized detection algorithm
  - use of surrogate loss functions and marginalized kernels

- study of surrogate loss functions and divergence functionals
  - correspondence of losses and divergences
  - M-estimator of divergences (e.g., Kullback-Leibler divergence)