Differential Geometric Regularization for Supervised Learning of Classifiers

Qinxun Bai[1], Steven Rosenberg[2], Zheng Wu[3], Stan Sclaroff[1]

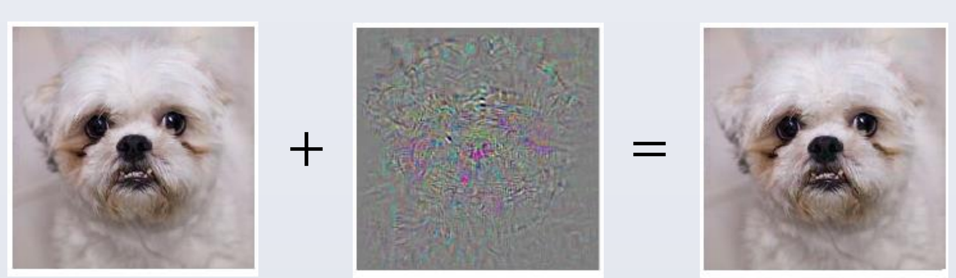
[1]Computer Science, [2]Mathematics & Statistics, Boston University [3]The Mathworks Inc.

Motivation

Visual Recognition – Supervised Learning of Classifiers

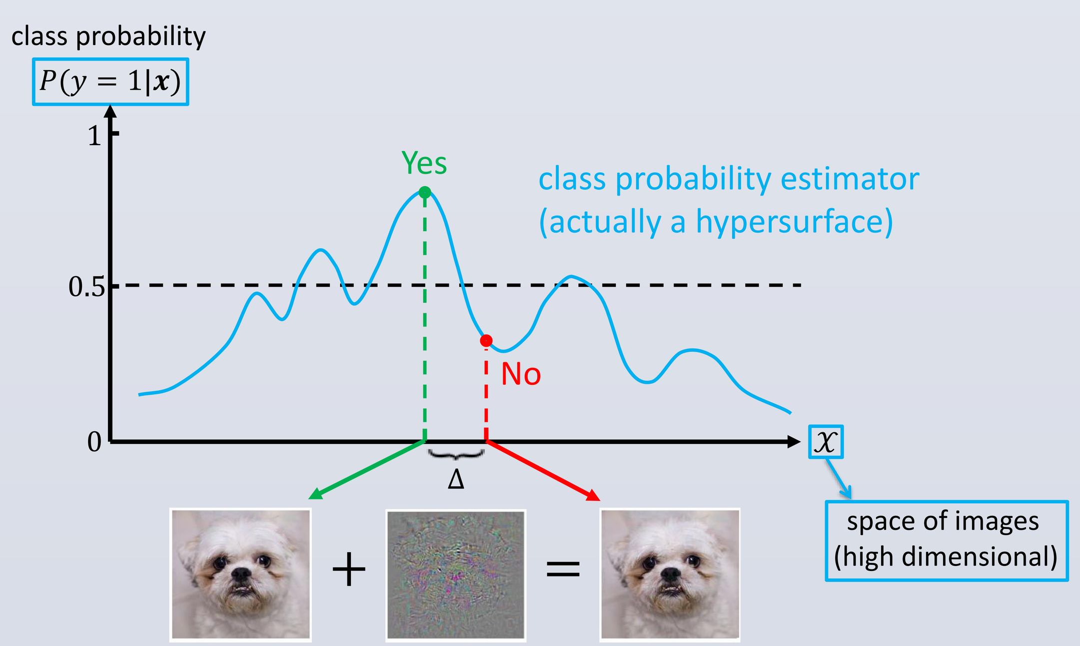
State-of-the-art on ImageNet Challenge: human level classification accuracy

Counter-intuitive Observations



Fool DNN by hardly perceptible perturbation [*Szegedy et at. 2013*]

Rapid Local Oscillation



Smoothness vs. Mean Curvature

Smoothness by functional norms:

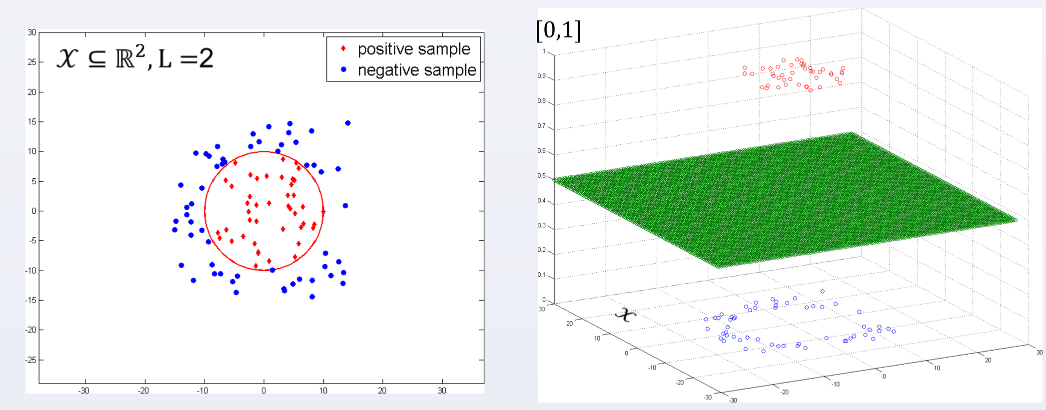
- Not specifically tailored to measure local oscillation
- Overkill the hypothesis space
- Sculpturing with an axe? Need a sculptor's knife!

Mean Curvature of the hypersurface:

- In differential geometric sense
- A specific measure of the amount of local oscillation
- Generalizes to high dimensional space

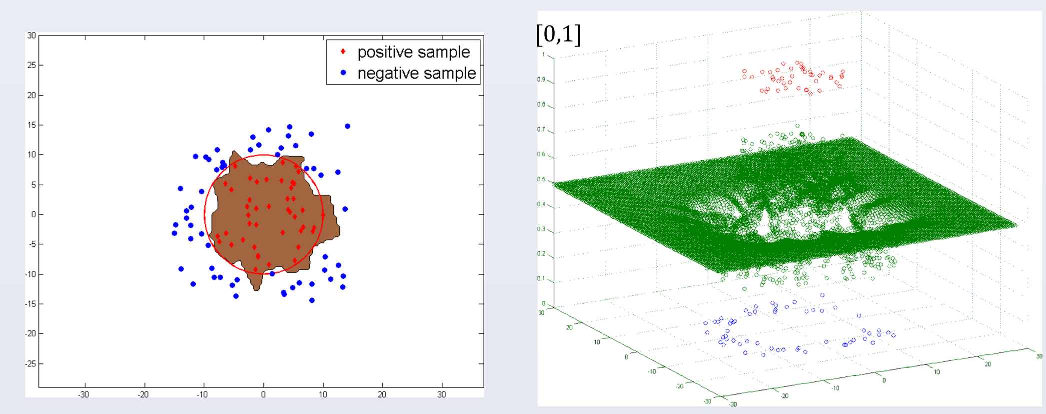
Main Idea

Physical Model

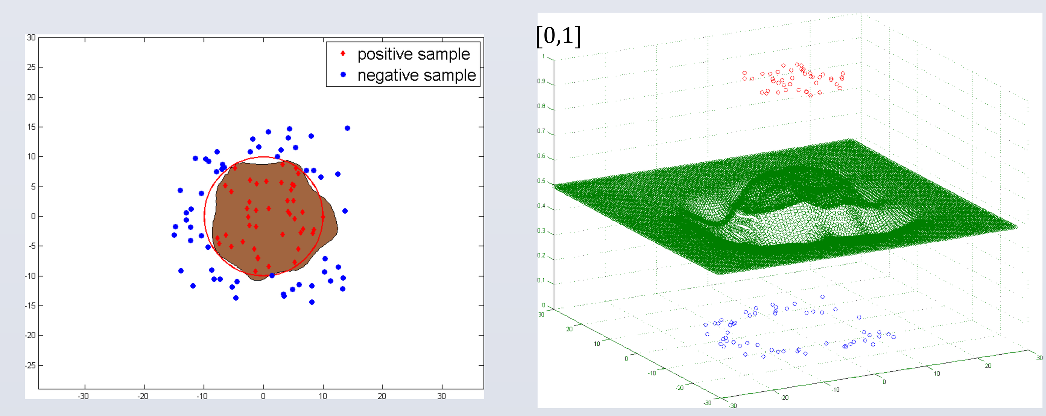


ground truth boundary

initial hypersurface



Hypersurface deforms towards training data as if attracted by gravitational force due to point masses centered at training data



In the mean time, hypersurface remains as tight as possible as if in the presence of surface tension

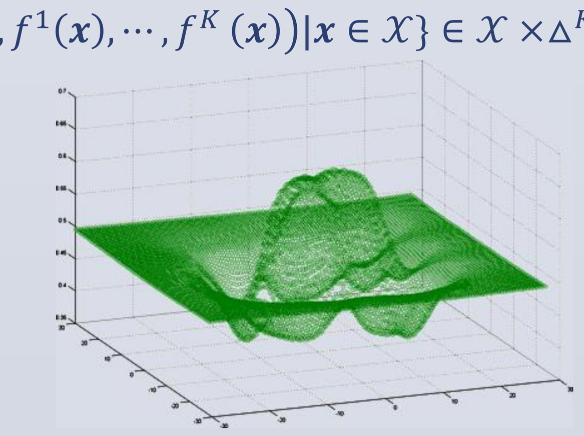
Formal Setup

Learn a function f: $\mathcal{X} \longrightarrow \Delta^K$ as an estimator of $P(y|x)$

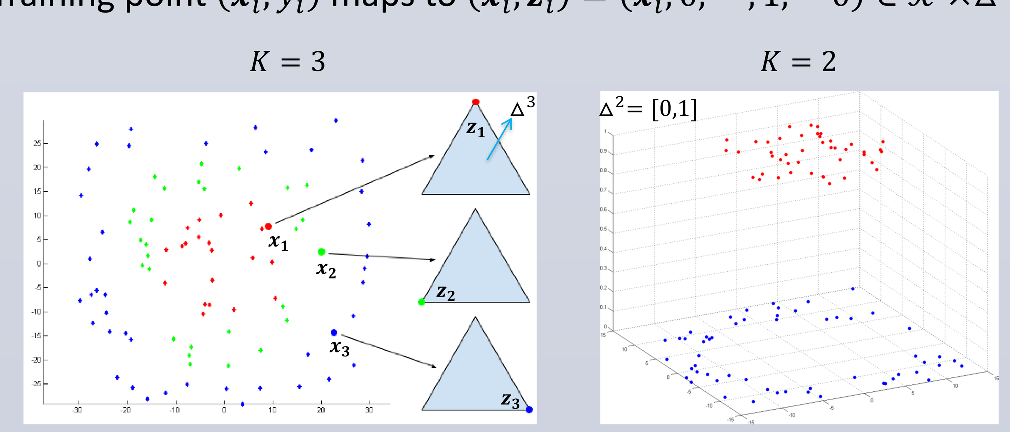
The hypersurface associated with f:

$$graph(f) = \{(x, f^1(x), \cdots, f^K(x)) | x \in \mathcal{X}\} \in \mathcal{X} \times \Delta^K$$

exploit the geometry of this hypersurface!



Training point (x_i, y_i) maps to $(x_i, z_i) = (x_i, 0, \cdots, \overset{y_i^{th}}{1}, \cdots 0) \in \mathcal{X} \times \Delta^K$



Regularized ERM Formulation

Minimize the regularized loss $\mathcal{P}$ in functional space $\mathcal{H}$

$$\min_{f \in \mathcal{H}} \mathcal{P}(f) = \min_{f \in \mathcal{H}} \{L(f) + \lambda G(f)\}$$

Data term: penalize the error of f in explaining the training data

Regularization term: penalize the volume of $graph(f)$

Solve for $\min_{f \in \mathcal{H}} \mathcal{P}(f)$

Solve iteratively by gradient flow: $\frac{df_t}{dt} = -\nabla \mathcal{P}$

- starting from neutral estimator $f_0 = (\frac{1}{K}, \cdots, \frac{1}{K})$
- evolve f_t towards $-\nabla \mathcal{P}_{f_t}$
- f_t will flow to a local minimum of $\mathcal{P}$

Computation of $\nabla \mathcal{P} = \nabla L + \lambda \nabla G$

Computing ∇L is easy

- *e.g.* back propagation for neural networks

Computing ∇G: mean curvature flow

- $G(f)$ measures the volume of $graph(f)$

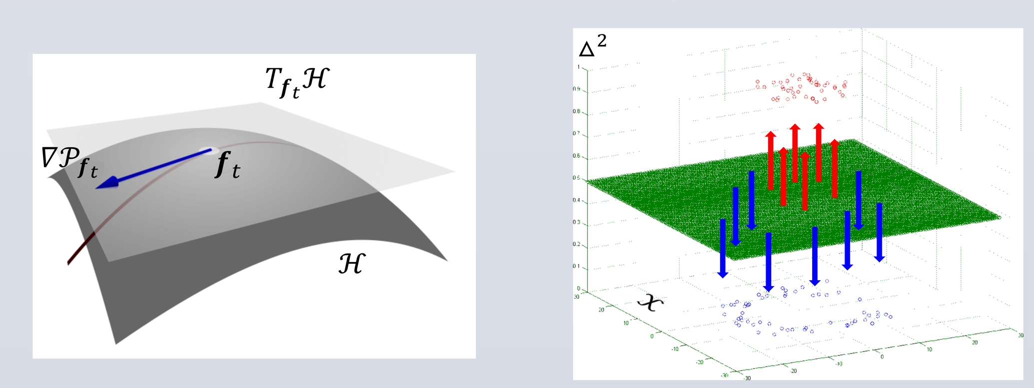
$G(f) = \int_{graph(f)} dvol = \int_{graph(f)} \sqrt{\det(g)} dx^1 \cdots dx^N$

where g is the Riemannian metric induced from $\mathbb{R}^{N+K}$

- **Our Theorem**:

need only 1st and 2nd partial derivatives of f, rest of computation is just matrix manipulations

The Gradient $\nabla \mathcal{P}_{f_t}$



$\nabla \mathcal{P}_{f_t}$: tangent vector in $T_{f_t}\mathcal{H}$ ⟷ vector field on $graph(f_t)$

Geometric Foundation on $\mathcal{H}$

$\mathcal{H} = Maps(\mathcal{X}, \Delta^K)$, $\mathcal{H}' = Maps(\mathcal{X}, \mathbb{R}^K)$

Topology

- Frechet topology on $\mathcal{H}'$, and the induced topology on $\mathcal{H}$

i.e. two functions in $\mathcal{H}$ are close if the functions and all their partial derivatives are pointwise close

Riemannian metric

- Restrict the L^2 metric on $\mathcal{H}'$ to each tangent space $T_f\mathcal{H}$

$$\langle \phi_1, \phi_2 \rangle = \int_{\mathcal{X}} \phi_1(x)\phi_2(x) dvol_x$$

where $\phi_i \in \mathcal{H}'$ and $dvol_x$ is the volume form of the induced Riemannian metric on $graph(f)$.

Experiments

RBF Representation

Represent f as "softmax" output of RBFs

$f^j = \frac{\exp(h^j)}{\sum_{l=1}^K \exp(h^l)}$, $h^j = \sum_{i=1}^m a_i^j \varphi_i(x)$, for $j = 1, \cdots, K$

where $\varphi_i(x) = e^{-\frac{1}{c}\|x - x_i\|^2}$ is the RBF centered at x_i

Gradient update for $A = (a_i^l)$

$A \leftarrow A - \tau M^{-1} [\nabla \mathcal{P}_h(x_1), \cdots, \nabla \mathcal{P}_h(x_m)]^T$,

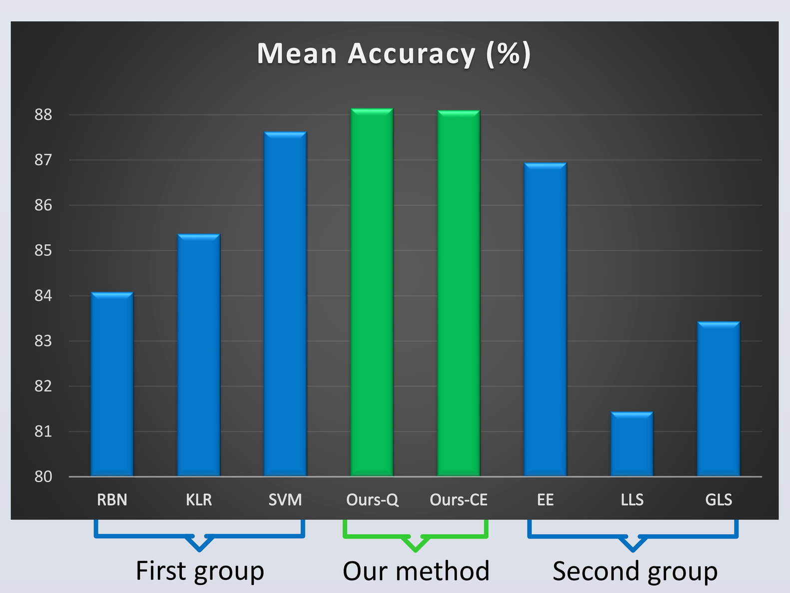
where $\nabla \mathcal{P}_h(x_i) = \left[\frac{\partial f}{\partial h}\right]_{x_i}^T \nabla \mathcal{P}_f(x_i)$, $M_{ij} = \varphi_j(x_i)$

Datasets from UCI Repository

- Four binary and four multiclass datasets
- Following the choice/setup of previous papers

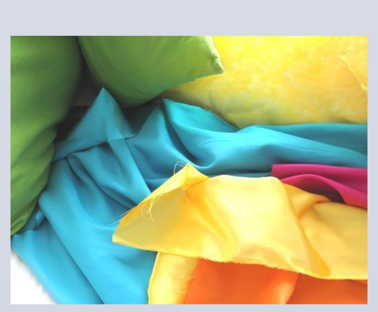
Comparing with two groups of classifiers

- RBF + functional norm regularization: RBN, SVM, KLR
- RBF + existing geometric regularization: LLS, GLS, EE

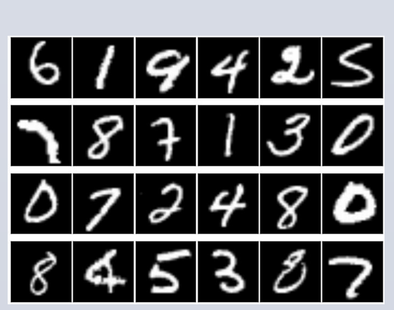


Real-world datasets – comparing with baseline

- Flickr Material Database (4096 dimensional feature)
- MNIST handwritten digits (60,000 samples)



Flickr Material Database



MNIST handwritten digits