

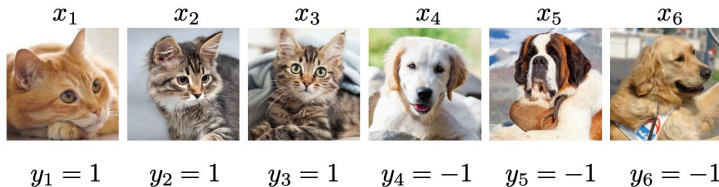
A12:Tackling Distribution Shifts via Test-Time Adaptation and Optimization

Instructor: Jun-Kun Wang (Assistant Professor at ECE and HDSI)

Training is Optimization

Supervised learning

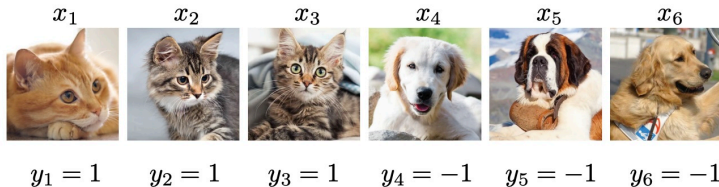
- n observations: $(x_i \in \mathbb{R}^d, y_i \in \{-1, +1\}), i = 1, \dots, n$.
- Prediction function: $h(x_i; w) \in \mathbb{R}$ parametrized by $w \in W$.



Training is Optimization

Supervised learning

- n observations: $(x_i \in \mathbb{R}^d, y_i \in \{-1, +1\}), i = 1, \dots, n$.
- Prediction function: $h(x_i; w) \in \mathbb{R}$ parametrized by $w \in W$.



Regularized Empirical Risk

$$\min_{w \in W} \underbrace{\frac{1}{n} \sum_{i=1}^n \ell(x_i, y_i; h(x_i; w))}_{\text{Empirical Risk}} + \underbrace{\lambda \phi(w)}_{\text{Regularization}}$$

Mathematical Background and Gradient Flow

(Derivative) For a function $g(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ and $x \in \mathbb{R}$, consider

$$\lim_{\delta \rightarrow 0} \frac{g(x + \delta) - g(x)}{\delta}.$$

Review: Calculus

(Derivative) For a function $g(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ and $x \in \mathbb{R}$, consider

$$\lim_{\delta \rightarrow 0} \frac{g(x + \delta) - g(x)}{\delta}.$$

The function $g(\cdot)$ is said to be “differentiable” if this limit exists for all $x \in \mathbb{R}$. In that case, the limit is called the “derivative” of $g(\cdot)$.

Review: Calculus

(**Derivative**) For a function $g(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ and $x \in \mathbb{R}$, consider

$$\lim_{\delta \rightarrow 0} \frac{g(x + \delta) - g(x)}{\delta}.$$

We denote the derivative as

Review: Calculus

(Gradient) For a differentiable function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^d$, the gradient is

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix},$$

where

$$\frac{\partial f}{\partial x_1} = \lim_{\delta \rightarrow 0} \frac{f(x_1 + \delta; x_2; \dots; x_d) - f(x_1; x_2; \dots; x_d)}{\delta}.$$

Review: Calculus

Definition

(**Hessian**) For a twice continuously differentiable function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^d$, the Hessian matrix of $f(\cdot)$ at \mathbf{x} is defined by

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix} \in \mathbb{R}^{d \times d}$$

Exercise

Let $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by $f(\mathbf{x}) = x_1^2 x_2$. Then

$$\nabla f(\mathbf{x}) =$$

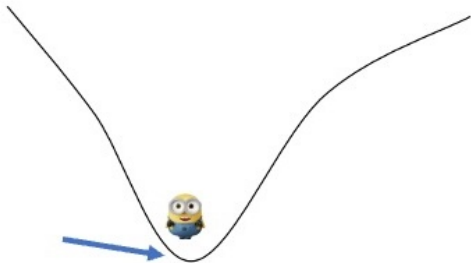
and

$$\nabla^2 f(\mathbf{x}) =$$

Convergence Rate

Finding the minimizer of a function

$$\min_{x \in \mathbb{R}^d} f(x) \quad (1)$$



Optimality Gap

Definition

(Optimality Gap): Given a function f such that $f: \mathbb{R}^d \rightarrow \mathbb{R}$, the optimality gap is the difference between the value of f at $\mathbf{x}_k \in \mathbb{R}^d$ at some time point k and the optimal value, i.e.

$$f(\mathbf{x}_k) - \min_{\mathbf{x}} f(\mathbf{x})$$

Gradient Descent

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}).$$

-
- 1: Input: an initial point $\mathbf{x}_0 \in \mathbf{dom} f$ and step size η .
 - 2: **for** $k = 1$ to K **do**
 - 3: $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \eta \nabla f(\mathbf{x}_k)$
 - 4: **end for**
 - 5: Return \mathbf{x}_{k+1} .
-

(Gradient Flow): Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be a smooth function. Gradient flow is a smooth curve $\mathbf{x}: \mathbb{R} \rightarrow \mathbb{R}^d$ such that



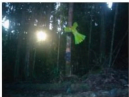

$$\frac{d\mathbf{x}(t)}{dt} = -\nabla f(\mathbf{x}(t))$$

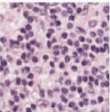
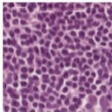
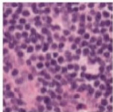
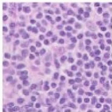
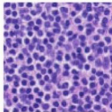
Gradient Flow is Gradient Descent as $\eta \rightarrow 0$

Our project:

Tackling Distribution Shifts via Test-Time Adaptation and Optimization

Machine learning has been used widely in
Data Science and Engineering

Train			Test (OOD)
$d = \text{Location 1}$	$d = \text{Location 2}$	$d = \text{Location 245}$	$d = \text{Location 246}$
			

Train			Val (OOD)	Test (OOD)
$d = \text{Hospital 1}$	$d = \text{Hospital 2}$	$d = \text{Hospital 3}$	$d = \text{Hospital 4}$	$d = \text{Hospital 5}$
$y = \text{Normal}$ 				

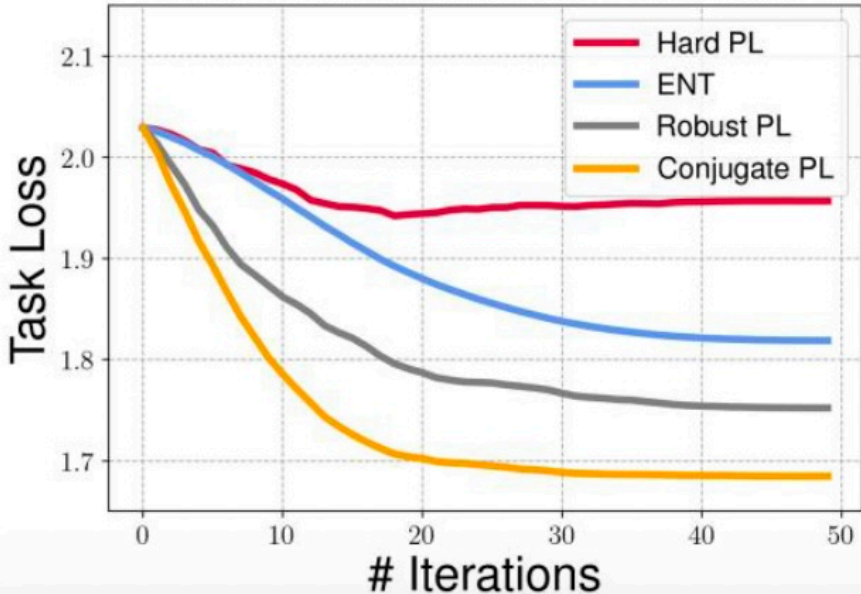
Needs
Adaptation
**Distribution
Shifts**

Source: Wilds: A Benchmark of in-the-Wild Distribution Shifts. Koh, Sagawa, et al. ICML 2021

8

Given a source model, adapt the model to a new domain, by using un-labeled samples of data from the new domain.

Task Loss Evaluated on Test Data



Research Group Meeting Schedule

[https://docs.google.com/document/d/](https://docs.google.com/document/d/1nEa5PQowFBSJhRoUYPjrQ-NKL4G6Qz5VCiv8AjwtMQCc/edit)

[1nEa5PQowFBSJhRoUYPjrQ-NKL4G6Qz5VCiv8AjwtMQCc/edit](https://docs.google.com/document/d/1nEa5PQowFBSJhRoUYPjrQ-NKL4G6Qz5VCiv8AjwtMQCc/edit)