Introduction to Distributionally Robust Optimization (DRO)

Stephan Rabanser

stephan@cs.toronto.edu

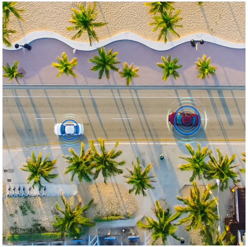
University of Toronto
Department of Computer Science

Vector Institute for
Artificial Intelligence

March 2, 2022

Motivation

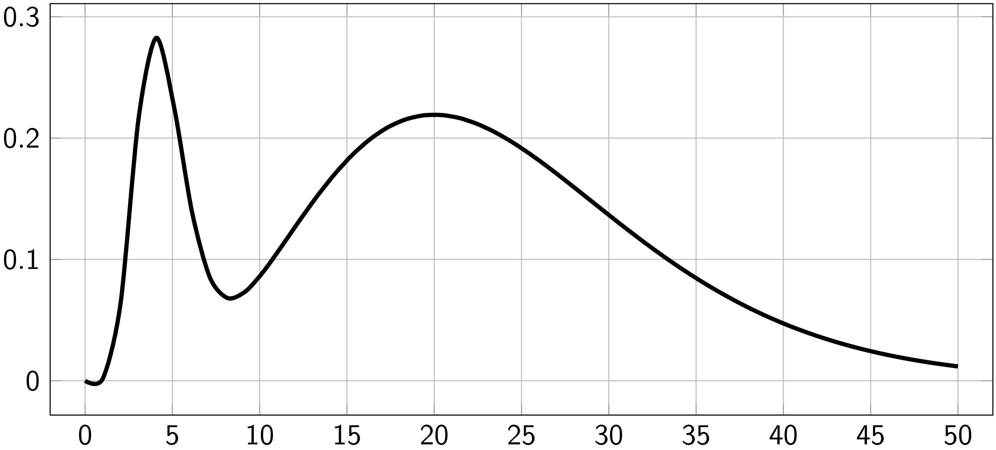
Machine Learning systems are becoming ubiquitous.



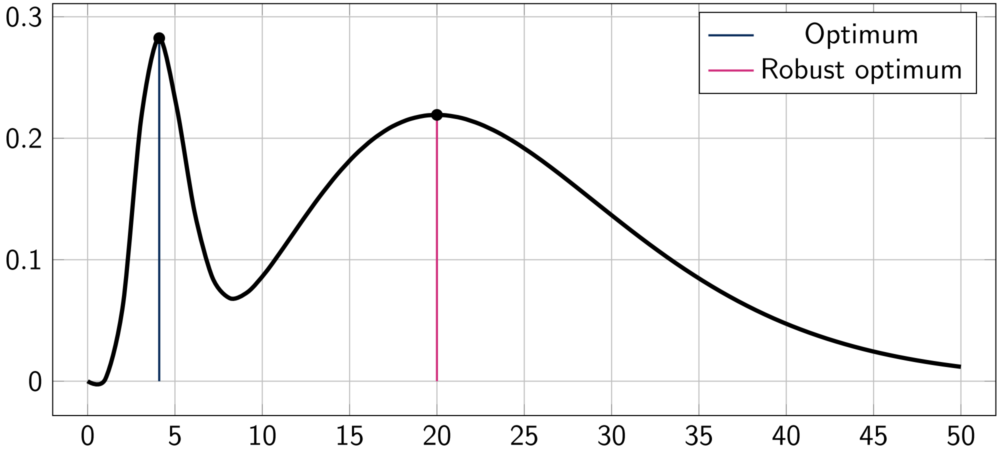

We need a thorough understanding of the robustness properties of ML algorithms to ensure safe deployment, especially in high-stakes decision-making systems.

Image credit: `unsplash.com`

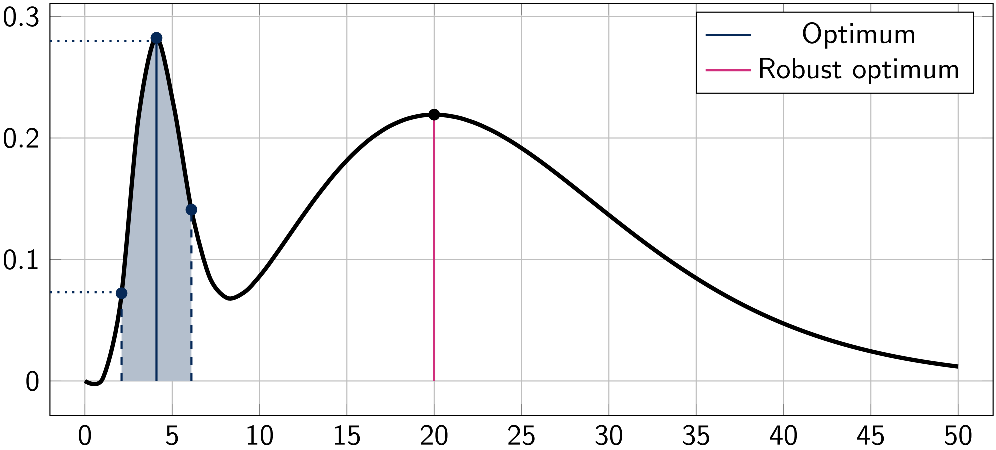
Warmup: Robust Optimization Example



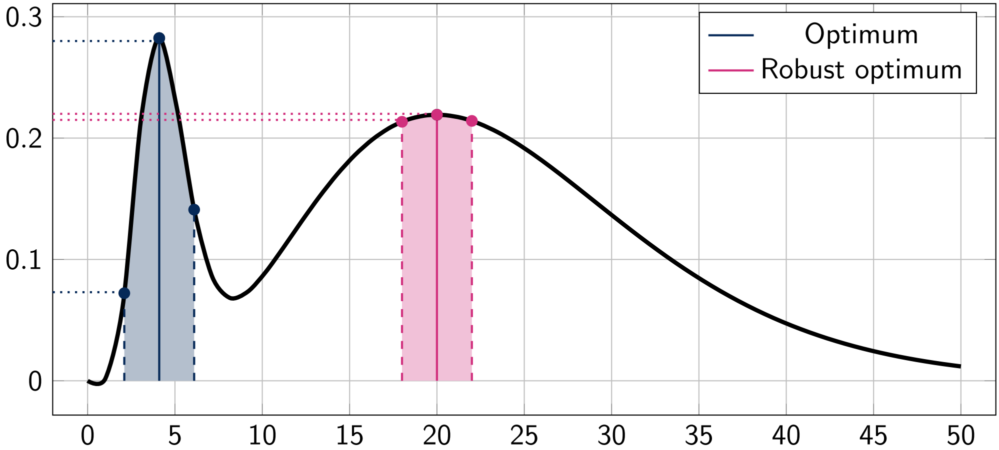
Warmup: Robust Optimization Example



Warmup: Robust Optimization Example



Warmup: Robust Optimization Example



Risk in Supervised Learning

Setup

- Dataset $D_p = \{(x_i, y_i)\}_{i=1}^N$ where $(x, y) \sim p$ over $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$ with $x \in \mathcal{X}$, $y \in \mathcal{Y}$.
- Prediction function $h_\theta(x) : \mathcal{X} \to \mathcal{Y}$ producing labels $\hat{y} = h_\theta(x)$ with $h_\theta(\cdot) \in \mathcal{H}$.
- Loss function $\ell(\hat{y}, y)$ measuring prediction quality of $h_\theta(x)$.

Risk in Supervised Learning

Setup

- Dataset $D_p = \{(x_i, y_i)\}_{i=1}^N$ where $(x, y) \sim p$ over $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$ with $x \in \mathcal{X}$, $y \in \mathcal{Y}$.
- Prediction function $h_\theta(x) : \mathcal{X} \to \mathcal{Y}$ producing labels $\hat{y} = h_\theta(x)$ with $h_\theta(\cdot) \in \mathcal{H}$.
- Loss function $\ell(\hat{y}, y)$ measuring prediction quality of $h_\theta(x)$.

Goal: By employing a learning algorithm $L : \mathcal{D} \to \mathcal{H}$ we want to produce a prediction function $h_\theta(\cdot)$ performing well on unseen test data $D'_p = \{(x_j, y_j)\}_{j=1}^M$, $(x, y) \sim p$, $D'_p \cap D_p = \varnothing$ as measured by our loss function $\ell(\cdot, \cdot)$.

UNIVERSITY OF TORONTO VECTOR INSTITUTE

Risk in Supervised Learning

Setup

- Dataset $D_p = \{(x_i, y_i)\}_{i=1}^N$ where $(x, y) \sim p$ over $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$ with $x \in \mathcal{X}$, $y \in \mathcal{Y}$.
- Prediction function $h_\theta(x) : \mathcal{X} \to \mathcal{Y}$ producing labels $\hat{y} = h_\theta(x)$ with $h_\theta(\cdot) \in \mathcal{H}$.
- Loss function $\ell(\hat{y}, y)$ measuring prediction quality of $h_\theta(x)$.

Goal: By employing a learning algorithm $L : \mathcal{D} \to \mathcal{H}$ we want to produce a prediction function $h_\theta(\cdot)$ performing well on unseen test data $D'_p = \{(x_j, y_j)\}_{j=1}^M$, $(x, y) \sim p$, $D'_p \cap D_p = \varnothing$ as measured by our loss function $\ell(\cdot, \cdot)$.

(True) Risk

$$\mathcal{R}(h_\theta) := \mathbb{E}_{p(x,y)}[\ell(h_\theta(x), y)] = \int_{\mathcal{Y}} \int_{\mathcal{X}} p(x, y)\ell(h_\theta(x), y)dxdy$$

Empirical Risk Minimization (ERM)

$p(x, y)$ is typically not known or intractable to compute and as a result $\mathcal{R}(h_\theta)$ cannot be computed. But we can empirically approximate $\mathcal{R}(h_\theta)$ as $\hat{\mathcal{R}}(h_\theta)$ using samples from $p(x, y)$ (i.e. using D_p):

$$\mathcal{R}(h_\theta) := \mathbb{E}_{p(x,y)}[\ell(h_\theta(x), y)] \qquad \hat{\mathcal{R}}(h_\theta) := \frac{1}{N}\sum_{i=1}^{N} \ell(h_\theta(x_i), y_i)$$

Empirical Risk Minimization (ERM)

$p(x, y)$ is typically not known or intractable to compute and as a result $\mathcal{R}(h_\theta)$ cannot be computed. But we can empirically approximate $\mathcal{R}(h_\theta)$ as $\hat{\mathcal{R}}(h_\theta)$ using samples from $p(x, y)$ (i.e. using D_p):

$$\mathcal{R}(h_\theta) := \mathbb{E}_{p(x,y)}[\ell(h_\theta(x), y)] \qquad \hat{\mathcal{R}}(h_\theta) := \frac{1}{N}\sum_{i=1}^{N} \ell(h_\theta(x_i), y_i)$$

Due to the law of large numbers we expect an increasingly better approximation of $\mathcal{R}(h_\theta)$ by $\hat{\mathcal{R}}(h_\theta)$ as more samples are provided to the learning algorithm L:

$$\hat{\mathcal{R}}(h_\theta) \approx \mathcal{R}(h_\theta) \qquad \hat{\mathcal{R}}(h_\theta) \xrightarrow{N\to\infty} \mathcal{R}(h_\theta) \qquad \arg\min_{h_\theta \in \mathcal{H}} \hat{\mathcal{R}}(h_\theta) \approx \arg\min_{h_\theta \in \mathcal{H}} \mathcal{R}(h_\theta)$$

The Problem of Distributional Shift

Revisiting our goal

Goal: By employing a learning algorithm $L : \mathcal{D} \to \mathcal{H}$ we want to produce a prediction function $h_\theta(\cdot)$ performing well on unseen test data $D'_p = \{(x_j, y_j)\}_{j=1}^M$, $(x, y) \sim p$, $D'_p \cap D_p = \varnothing$ as measured by our loss function $\ell(\cdot, \cdot)$.

The Problem of Distributional Shift

Revisiting our goal

Goal: By employing a learning algorithm $L : \mathcal{D} \to \mathcal{H}$ we want to produce a prediction function $h_\theta(\cdot)$ performing well on unseen test data $D'_p = \{(x_j, y_j)\}_{j=1}^M$, $(x, y) \sim p$, $D'_p \cap D_p = \varnothing$ as measured by our loss function $\ell(\cdot, \cdot)$.

A more realistic scenario

$$D'_q = \{(x_j, y_j)\}_{j=1}^M \qquad (x, y) \sim q, \; 0 \leq d(p, q) \leq \delta \qquad D'_q \cap D_p = \varnothing$$

$d(p, q)$ is a divergence measure between training distribution p and testing distribution q and is bounded by δ.

UNIVERSITY OF TORONTO VECTOR INSTITUTE

Risk Minimization vs Distributionally Robust Optimization

Risk Minimization

$$\arg\min_{h_\theta \in \mathcal{H}} \mathbb{E}_{p(x,y)}[\ell(h_\theta(x), y)]$$

Distributionally Robust Optimization

$$\arg\min_{h_\theta \in \mathcal{H}} \max_{q \in \mathcal{Q}_p} \mathbb{E}_{q(x,y)}[\ell(h_\theta(x), y)]$$

$$\mathcal{Q}_p = \{q \ll p \mid d(p, q) \leq \delta\}$$

UNIVERSITY OF TORONTO VECTOR INSTITUTE

Risk Minimization vs Distributionally Robust Optimization

Risk Minimization

$$\arg\min_{h_\theta \in \mathcal{H}} \mathbb{E}_{p(x,y)}[\ell(h_\theta(x), y)]$$

Distributionally Robust Optimization

$$\arg\min_{h_\theta \in \mathcal{H}} \max_{q \in \mathcal{Q}_p} \mathbb{E}_{q(x,y)}[\ell(h_\theta(x), y)]$$

$$\mathcal{Q}_p = \{q \ll p \mid d(p,q) \leq \delta\}$$

Risk Minimization vs Distributionally Robust Optimization

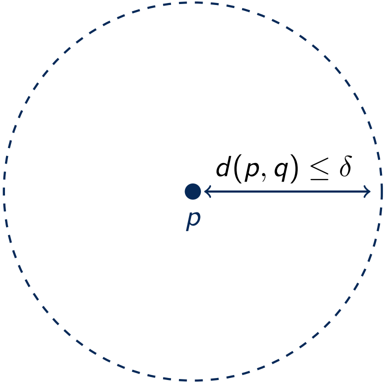
Risk Minimization

$$\arg\min_{h_\theta \in \mathcal{H}} \mathbb{E}_{p(x,y)}[\ell(h_\theta(x), y)]$$

Distributionally Robust Optimization

$$\arg\min_{h_\theta \in \mathcal{H}} \max_{q \in \mathcal{Q}_p} \mathbb{E}_{q(x,y)}[\ell(h_\theta(x), y)]$$

$$\mathcal{Q}_p = \{q \ll p \mid d(p,q) \leq \delta\}$$



Risk Minimization vs Distributionally Robust Optimization

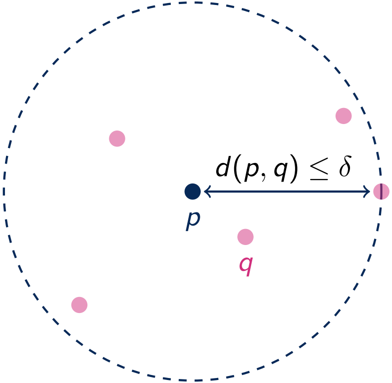
Risk Minimization

$$\arg\min_{h_\theta \in \mathcal{H}} \mathbb{E}_{p(x,y)}[\ell(h_\theta(x), y)]$$

Distributionally Robust Optimization

$$\arg\min_{h_\theta \in \mathcal{H}} \max_{q \in \mathcal{Q}_p} \mathbb{E}_{q(x,y)}[\ell(h_\theta(x), y)]$$

$$\mathcal{Q}_p = \{q \ll p \mid d(p,q) \leq \delta\}$$



Risk Minimization vs Distributionally Robust Optimization

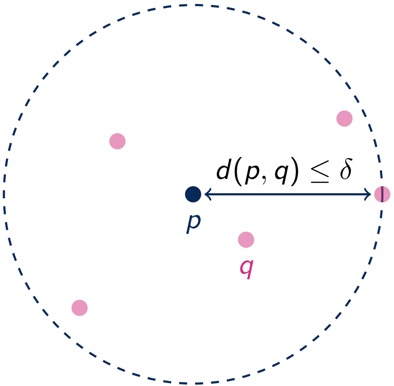
Risk Minimization

$$\arg\min_{h_\theta \in \mathcal{H}} \mathbb{E}_{p(x,y)}[\ell(h_\theta(x), y)]$$

Distributionally Robust Optimization

$$\arg\min_{h_\theta \in \mathcal{H}} \max_{q \in \mathcal{Q}_p} \mathbb{E}_{q(x,y)}[\ell(h_\theta(x), y)]$$

$$\mathcal{Q}_p = \{q \ll p \mid d(p, q) \leq \delta\}$$



Important: The distribution q that leads to the worst-case DRO loss does not necessarily correspond to be the distribution that maximizes $d(p, q)$!

Divergences Between Probability Distributions

Integral Probability Metrics: $p - q$

$$d_{\mathcal{F}}(p, q) = \sup_{g \in \mathcal{F}} |\mathbb{E}_{X \sim p}[g(X)] - \mathbb{E}_{X' \sim q}[g(X')]|$$

Divergences Between Probability Distributions

Integral Probability Metrics: $p - q$

$$d_{\mathcal{F}}(p, q) = \sup_{g \in \mathcal{F}} \left| \mathbb{E}_{X \sim p}[g(X)] - \mathbb{E}_{X' \sim q}[g(X')] \right|$$

Wasserstein Distance

Maximum Mean Discrepancy

Dudley Metric

Divergences Between Probability Distributions

Integral Probability Metrics: $p - q$

$$d_{\mathcal{F}}(p,q) = \sup_{g \in \mathcal{F}} |\mathbb{E}_{X \sim p}[g(X)] - \mathbb{E}_{X' \sim q}[g(X')]|$$

- Wasserstein Distance
- Maximum Mean Discrepancy
- Dudley Metric

ϕ-divergences (f-divergences): $\frac{p}{q}$

$$d_{\phi}(p,q) = \int_{\mathcal{X}} q(x)\phi\left(\frac{p(x)}{q(x)}\right)dx$$

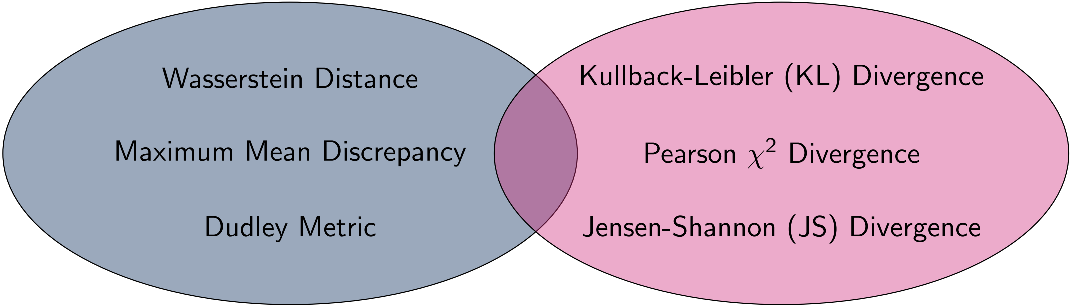
Divergences Between Probability Distributions

Integral Probability Metrics: $p - q$

$d_{\mathcal{F}}(p, q) = \sup_{g \in \mathcal{F}} |\mathbb{E}_{X \sim p}[g(X)] - \mathbb{E}_{X' \sim q}[g(X')]|$

ϕ-divergences (f-divergences): $\frac{p}{q}$

$d_\phi(p, q) = \int_{\mathcal{X}} q(x)\phi(\frac{p(x)}{q(x)})dx$



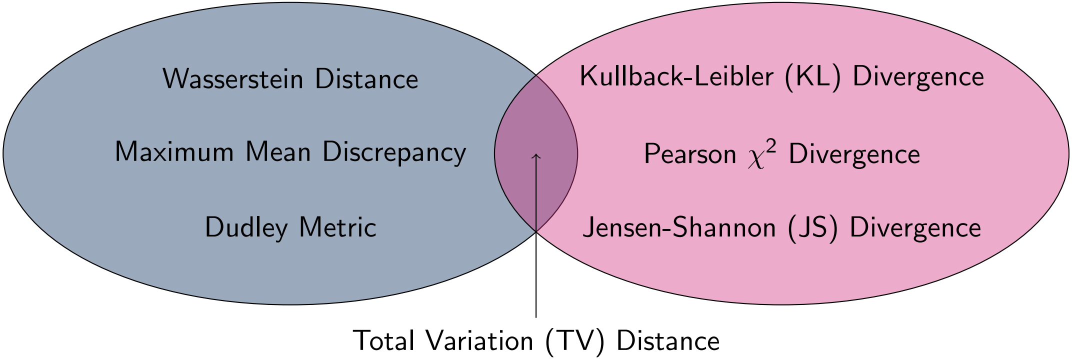
Divergences Between Probability Distributions

Integral Probability Metrics: $p - q$

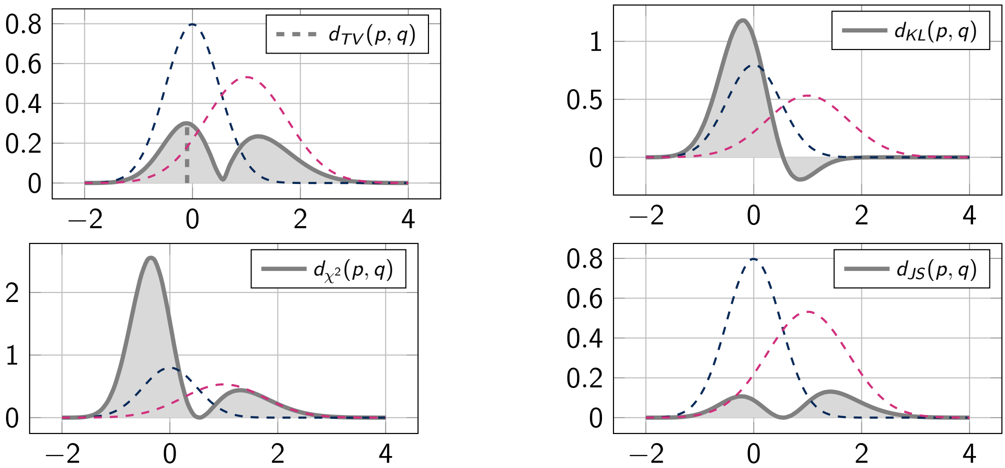
ϕ-divergences (f-divergences): $\frac{p}{q}$

$$d_{\mathcal{F}}(p, q) = \sup_{g \in \mathcal{F}} \left|\mathbb{E}_{X \sim p}[g(X)] - \mathbb{E}_{X' \sim q}[g(X')]\right|$$

$$d_{\phi}(p, q) = \int_{\mathcal{X}} q(x)\phi\left(\frac{p(x)}{q(x)}\right)dx$$



ϕ-divergences: Choices for $\phi(\cdot)$



Application: ERM Generalization and Regularization

Recall the ERM definition:

$$\hat{\mathcal{R}}_\lambda(h_\theta) := \frac{1}{N}\sum_{i=1}^{N} \ell(h_\theta(x_i), y_i) + \underbrace{\lambda\Omega(\theta)}_{\text{regularizer}}$$

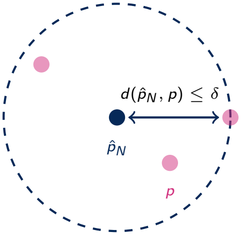
By regularizing, we reduce overfitting on the sample distribution $\hat{p}_N$ and enable generalization to p.

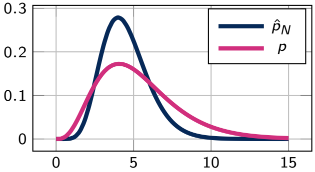
Application: ERM Generalization and Regularization

Recall the ERM definition:

$$\hat{\mathcal{R}}_\lambda(h_\theta) := \frac{1}{N}\sum_{i=1}^{N} \ell(h_\theta(x_i), y_i) + \underbrace{\lambda\Omega(\theta)}_{\text{regularizer}}$$

By regularizing, we reduce overfitting on the sample distribution $\hat{p}_N$ and enable generalization to p.





Application: ERM Generalization and Regularization

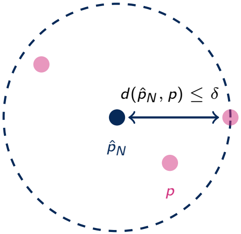
Recall the ERM definition:

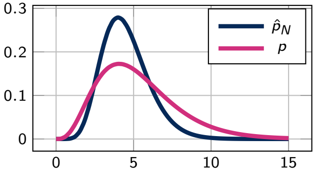
$$\hat{\mathcal{R}}_\lambda(h_\theta) := \frac{1}{N}\sum_{i=1}^{N} \ell(h_\theta(x_i), y_i) + \underbrace{\lambda\Omega(\theta)}_{\text{regularizer}}$$

By regularizing, we reduce overfitting on the sample distribution $\hat{p}_N$ and enable generalization to p.

Different divergences lead to different regularization:

- χ^2 penalizes $\mathbb{V}_{\hat{p}_N}[\ell(h_\theta(x), y)]$
- Wasserstein penalizes $||\nabla_x \ell(h_\theta(x), y)||$
- MMD penalizes $||\ell(h_\theta(x), y)||_{\mathcal{F}}$





UNIVERSITY OF TORONTO VECTOR INSTITUTE

Application: Distribution Shifts in General

Example setting: You are building a predictive model for house prices based on square meters.

- p: square meters distribution in the inner city
- q_S: square meters distribution in the city's **s**uburbs
- q_W: square meters distribution in the **w**hole city
- q_O: square meters distribution of an**o**ther city

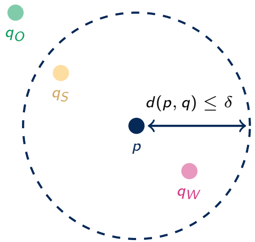
Goal: Generalize to the worst-case distribution within the city, i.e., q_S and q_W, but not to q_O.

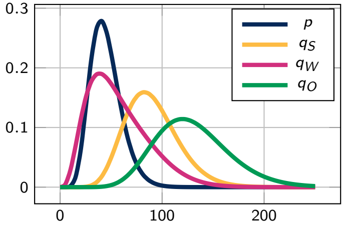
Application: Distribution Shifts in General

Example setting: You are building a predictive model for house prices based on square meters.

- p: square meters distribution in the inner city
- q_S: square meters distribution in the city's **s**uburbs
- q_W: square meters distribution in the **w**hole city
- q_O: square meters distribution of an**o**ther city

Goal: Generalize to the worst-case distribution within the city, i.e., q_S and q_W, but not to q_O.





UNIVERSITY OF TORONTO VECTOR INSTITUTE

Discussion: Practical Estimation of the DRO Objective

$$\arg\min_{h_\theta \in \mathcal{H}} \max_{q \in \mathcal{Q}_p} \mathbb{E}_{q(x,y)}[\ell(h_\theta(x), y)] \qquad \text{with} \qquad \mathcal{Q}_p = \{q \ll p \mid d(p,q) \leq \delta\}$$

Discussion: Practical Estimation of the DRO Objective

$$\arg\min_{h_\theta \in \mathcal{H}} \max_{q \in \mathcal{Q}_p} \mathbb{E}_{q(x,y)}[\ell(h_\theta(x), y)] \qquad \text{with} \qquad \mathcal{Q}_p = \{q \ll p \mid d(p, q) \leq \delta\}$$

Estimation from empirical data

1. Collect worst case test data D'_q.
2. Minimize empirical loss on worst case test data

$$\arg\min_{h_\theta \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \ell(h_\theta(x_i), y_i)$$

with $(x, y) \in D'_q$.

Discussion: Practical Estimation of the DRO Objective

$$\arg\min_{h_\theta \in \mathcal{H}} \max_{q \in \mathcal{Q}_p} \mathbb{E}_{q(x,y)}[\ell(h_\theta(x), y)] \quad \text{with} \quad \mathcal{Q}_p = \{q \ll p \mid d(p,q) \leq \delta\}$$

Estimation from empirical data

1. Collect worst case test data D'_q.
2. Minimize empirical loss on worst case test data

$$\arg\min_{h_\theta \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \ell(h_\theta(x_i), y_i)$$

with $(x, y) \in D'_q$.

Estimation from theoretical framework

1. Estimate the distribution p from D.
2. Approximate p with a simpler distribution $\tilde{p}$ using VI.
3. Choose $d(\tilde{p}, q)$ and δ.
4. Either
   - directly minimize DRO objective; or
   - sample from worst q and empirically minimize DRO objective.

UNIVERSITY OF TORONTO VECTOR INSTITUTE

Thanks! :)

References

- https://www.youtube.com/watch?v=IgAPc0i0-9E
- https://web.stanford.edu/~yyye/MFDataScience2020.pdf
- https://simons.berkeley.edu/sites/default/files/docs/8841/simons-2.pdf

John C Duchi and Hongseok Namkoong, *Learning models with uniform performance via distributionally robust optimization*, The Annals of Statistics **49** (2021), no. 3, 1378–1406.

Hamed Rahimian and Sanjay Mehrotra, *Distributionally robust optimization: A review*, arXiv preprint arXiv:1908.05659 (2019).

Matthew Staib and Stefanie Jegelka, *Distributionally robust optimization and generalization in kernel methods*, Advances in Neural Information Processing Systems **32** (2019), 9134–9144.

Backup

The Connection Between Optimization and Uncertainty

Optimization Technique	Uncertainty Model
Deterministic	Point-forecast (no uncertainty)
Stochastic optimization	Expectation
Chance-constrained optimization	Probability distribution
Robust optimization	Worst-case deviation under unbounded divergence
Distributionally robust optimization	Worst-case deviation under bounded divergence

ϕ-divergences: Choices for $\phi(\cdot)$

$$\boxed{d_\phi(p,q) = \int_{\mathcal{X}} q(x)\phi(\frac{p(x)}{q(x)})dx \qquad \phi \text{ convex and } \phi(1) = 0}$$

TV distance: $\phi(x) = \frac{|x-1|}{2}$

$$d_{TV}(p,q) = \int_{\mathcal{X}} q(x)\frac{|\frac{p(x)}{q(x)} - 1|}{2}dx = \int_{\mathcal{X}} \frac{|p(x) - q(x)|}{2}dx$$

χ^2 divergence: $\phi(x) = (x-1)^2$

$$d_{\chi^2}(p,q) = \int_{\mathcal{X}} q(x)(\frac{p(x)}{q(x)} - 1)^2 dx = \ldots = \int_{\mathcal{X}} \frac{(p(x) - q(x))^2}{q(x)}dx$$

ϕ-divergences: Choices for $\phi(\cdot)$

$$\boxed{d_\phi(p,q) = \int_{\mathcal{X}} q(x)\phi(\frac{p(x)}{q(x)})dx \qquad \phi \text{ convex and } \phi(1) = 0}$$

KL divergence: $\phi(x) = x \log x$

$$d_{KL}(p,q) = \int_{\mathcal{X}} q(x)\frac{p(x)}{q(x)} \log(\frac{p(x)}{q(x)})dx = \int_{\mathcal{X}} p(x) \log(\frac{p(x)}{q(x)})dx$$

Jensen-Shannon divergence: $\phi(x) = \frac{1}{2}[(x+1)\log\left(\frac{2}{x+1}\right) + x \log x]$

$$d_{JS}(p,q) = \ldots = \frac{1}{2}d_{KL}(p, \frac{1}{2}(p+q)) + \frac{1}{2}d_{KL}(q, \frac{1}{2}(p+q))$$