DS603: Robust Machine Learning	Autumn 2025
Lecture 6 — 22/08/2025	
Lecturer: Arjun Bhagoji	*Scribe: Ritik*

6.1 Revisiting the standard supervised learning setup

6.1.1 Training vs. Testing

During **training/estimation**: We have data points

$$z_i = (x_i, y_i) \sim \mathbb{P}^*, \quad i = 1, \ldots, n$$

where $\mathbb{P}^*$ is the true data distribution. Using this training set, we contruct the **empirical distribution** which is defined as

$$\hat{\mathbb{P}}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{z_i}.$$

Here, δ_{z_i}, is the **Dirac measure** [3] at the point z_i (a probability mass concentrated at that sample).

During **testing/inference**: Samples $z = (x, y)$ are assumed to be drawn from the true distribution $\mathbb{P}^*$.

6.1.2 Dirac Measure

For a measurable space $(X, \mathscr{F})$, where $\mathscr{F}$ is an appropriately defined σ-algebra, the Dirac measure at a point $x \in X$ is defined as

$$\delta_x(A) = \mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A, \end{cases} \quad \forall A \in \mathscr{F}.$$

Note: The Dirac measure is a *probability measure*. For any measurable function f,

$$\int_x f(y)\, d\delta_x(y) = f(x).$$

See Appendix A for proof.

HW: Is $\hat{\mathbb{P}}_n$ a discrete measure?

Yes it is a discrete measure, because it is a finite sum of weighted Dirac measures. A general discrete measure can be expressed as

$$\mu = \sum_i a_i \delta_{x_i}, \quad a_i \geqslant 0.$$

The empirical distribution $\hat{\mathbb{P}}_n$ is a discrete measure with uniform weights $a_i = \frac{1}{n}$ and also note $\sum_i a_i = 1$.

6.1.3 Learning via Optimization

The **true (or population) risk minimization** problem is

$$\min_{\theta\in\Theta} \mathbb{E}_{z\sim\mathbb{P}^*}\big[\ell(\theta,z)\big] = \min_{\theta} \int_z \ell(\theta,z)\, d\mathbb{P}^*(z).$$

[what is? - AB]That is **expected loss** under the true distribution $\mathbb{P}^*$. Since $\mathbb{P}^*$ is unknown, we approximate using the **empirical risk**:

$$\min_{\theta\in\Theta} \mathbb{E}_{z\sim\hat{\mathbb{P}}_n}\big[\ell(\theta,z)\big] = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta,z_i).$$

This is the principle of **Empirical Risk Minimization (ERM)** [5].

6.1.4 Parameter Estimation

The goal is to learn $\hat{\theta} : \mathscr{P}(z) \to \Theta$ such that [What is $\mathscr{P}$ - AB]

$$\mathbb{E}_{z\sim\mathbb{P}^*}\big[\ell(\hat{\theta}(\hat{\mathbb{P}}_n),z)\big]$$

is minimized. In general, this is the framework of M-estimation [6] (a generalization of maximum-likelihood estimation), where different choices of ℓ lead to different statistics of the underlying distribution such as the median arising from $\ell = |\theta - z|$.

Example (Squared Loss): Using squared loss

$$\ell(\theta,z) = \|\theta - z\|^2,$$

the optimal θ corresponds to the **mean** of the distribution.

6.2 Data-Driven Decision Making Cycle

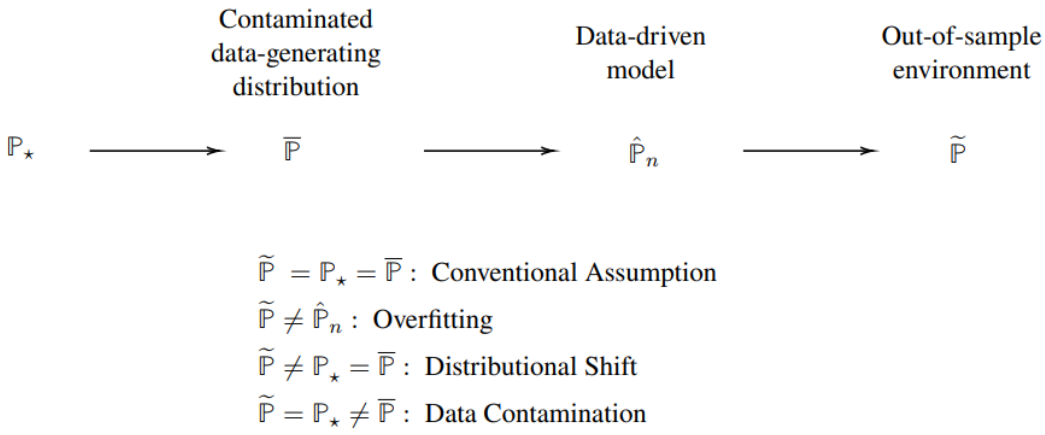

Figure 6.1. Data-Driven Decision Making Cycle

In the conventional data-driven decision-making cycle given in Figure 6.1 [add citation - AB], we observe n i.i.d. samples generated from the unknown data-generating distribution $\mathbb{P}^\star$, and build a model $\hat{\mathbb{P}}_n$ from these samples (model could be parametric or non-parametric). Based on $\hat{\mathbb{P}}_n$, we make a decision (e.g., parameter estimation), which is then deployed in an **out-of-sample environment** $\tilde{\mathbb{P}}$ that may or may not coincide with the original distribution $\mathbb{P}_\star$.

6.2.1 Potential Sources of Sub-optimal Decision-Making

Several factors can lead to suboptimal performance in this cycle:

1. **Overfitting** ($\tilde{\mathbb{P}} \neq \hat{\mathbb{P}}_n$): When the sample size n is not sufficiently large, the learned model $\hat{\mathbb{P}}_n$ may fit the training samples well but fail to generalize to $\tilde{\mathbb{P}}$, resulting in poor out-of-sample performance.
2. **Distributional Shift** ($\tilde{\mathbb{P}} \neq \mathbb{P}^\star = \overline{\mathbb{P}}$): In many real-world settings, the deployment distribution $\tilde{\mathbb{P}}$ deviates from the data-generating distribution $\mathbb{P}^\star$. This discrepancy may arise in scenarios such as:
   - *Adversarial deployment*: Malicious actors manipulate the data distribution to degrade model performance.
   - *Transfer learning*: Models must generalize to target datasets that differ from the source distribution.
3. **Data Contamination** ($\tilde{\mathbb{P}} = \mathbb{P}^\star \neq \overline{\mathbb{P}}$): The observed data may contain outliers or measurement errors during the data generation and collection process. Thus, the effective training distribution is a contaminated version $\overline{\mathbb{P}}$ of the true $\mathbb{P}_\star$. This discrepancy may also arise due to model poisoning.
4. **Backdoor Attacks** ($\tilde{\mathbb{P}} = \overline{\mathbb{P}} \neq \mathbb{P}^\star$): These occur when an adversary poisons the training data by inserting specific "triggers" (e.g., small patterns or perturbations) into a subset of examples with manipulated labels. The model thus learns a hidden rule: it behaves correctly on clean inputs, but produces attacker-chosen outputs whenever the trigger is present. In this setting, both training and deployment data distributions are corrupted with the trigger and differ from the true clean distribution. This makes detection difficult, as the poisoned model performs well on standard evaluation but fails under the triggered inputs.

6.2.2 Pre-decision vs. Post-decision Errors

- Cases (i) and (ii) (**overfitting and distributional shift**) occur in the **post-decision stage**, when the trained model is deployed in the out-of-sample environment.
- Cases (iii) and (iv) (**data contamination and backdoor attacks**) occurs in the **pre-decision stage**, during data generation and collection.

6.2.3 Data contamination methods

Let $\mathbb{P}^*$ denote the true underlying data-generating distribution. We are ultimately interested in learning a model $\hat{\mathbb{P}}_n$ that is close to $\mathbb{P}^*$. However, data contamination can occur at different stages of the pipeline as shown in Figure 6.2. In both cases, the learned model is constructed from corrupted information but is still intended to perform well or follow the true distribution $\mathbb{P}^*$.

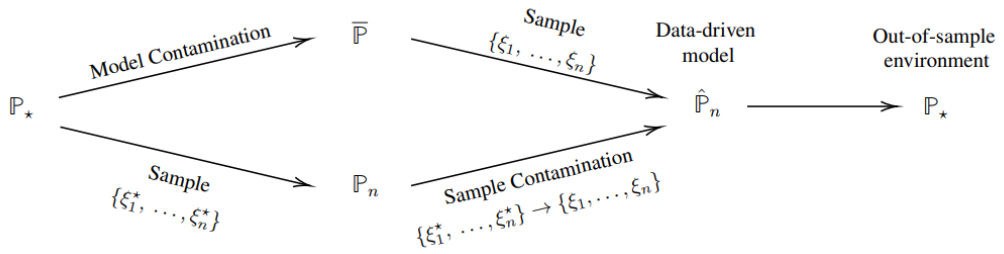

Figure 6.2. Decision Making Cycle in the Context of Robust Statistics

6.3 Distributionally Robust Optimization (DRO)

To address cases (i) and (ii), **distributionally robust optimization (DRO)** [4] is a principled framework that seeks to minimize discrepancies between the *in-sample expected loss* (under $\mathbb{P}^*$) and the *out-of-sample expected loss* (under $\tilde{\mathbb{P}}$). DRO explicitly accounts for distributional uncertainty to achieve more robust decision-making.

6.3.1 Stochastic Optimization Setup

- Random vector: $\xi \in \Xi \subseteq \mathbb{R}^d$, let $\vec{z} = \xi$, distributed according to the true distribution $\mathbb{P}^\star$.
- Model parameters: $\theta \in \Theta$ (assume finite-dimensional).
- Loss function: $\ell(\vec{z}, \vec{\theta})$ measures the cost of using parameter $\vec{\theta}$ on data realization $\vec{z}$. In a classification setting, $\vec{z} = (\vec{x}, y)$, and loss is measured by the discrepancy between $\hat{y}$ predicted using $\vec{\theta}$ and y.

6.3.2 DRO Formulation

Standard ERM may lead to poor out-of-sample performance due to overfitting, distributional shift, or data contamination. To ensure robustness, DRO introduces an **uncertainty set** $\mathscr{B}(\hat{\mathbb{P}}_n)$ around the empirical distribution, which captures possible deviations between in-sample and out-of-sample distributions. The DRO problem is then formulated as:

$$\min_{\vec{\theta} \in \Theta} \sup_{Q \in \mathscr{B}(\hat{\mathbb{P}}_n)} \mathbb{E}_Q\left[\ell(\vec{z}, \vec{\theta})\right] = \hat{L}_r(\vec{z}, \vec{\theta}), \tag{1}$$

where the inner sup considers the worst-case expected loss under all distributions Q in the uncertainty set $\mathscr{B}(\hat{\mathbb{P}}_n)$.

6.4 Distributional Uncertainty Sets

Key Question: How do we measure distributional uncertainty?
A key step in distributionally robust optimization (DRO) is the specification of the **distributional uncertainty set** $\mathscr{B}$. Different choices of $\mathscr{B}$ lead to different formulations of the DRO problem.

6.4.1 Moment-Based Uncertainty Sets

One of the earliest approaches [1] models uncertainty using **moment constraints**. The general DRO problem is

$$\min_{\theta\in\Theta} \sup_{\mathbb{Q}\in\mathscr{B}} \mathbb{E}_{\mathbb{Q}}[\ell(\theta,z)],$$

where the uncertainty set $\mathscr{B}$ is defined via constraints on the mean and covariance.

Construction of $\mathscr{B}$: Let $\hat{\mu}$ and $\hat{\Sigma}$ denote the empirical mean and covariance estimated from $\hat{\mathbb{P}}_n$. Given non-negative constants δ_1, δ_2, define

$$\mathscr{B} = \left\{ \mathbb{Q} \,:\, \left(\mathbb{E}_{\mathbb{Q}}[z] - \hat{\mu}\right)^{\top} \hat{\Sigma}^{-1} \left(\mathbb{E}_{\mathbb{Q}}[z] - \hat{\mu}\right) \leqslant \delta_1,\ \mathbb{E}_{\mathbb{Q}}\left[(z-\hat{\mu})^{\top}(z-\hat{\mu})\right] \preceq \delta_2 \hat{\Sigma} \right\}.$$

Note that the first condition is basically bounding the square of **Mahalanobis distance** to a constant.

Advantages: This formulation often admits a **semi-definite programming** (SDP) representation, making it computationally attractive.

Limitations: From a statistical standpoint, $\mathscr{B}$ may contain distributions far from the empirical distribution $\hat{\mathbb{P}}_n$, even for small δ_1, δ_2. As $n \to \infty$, $\hat{\mathbb{P}}_n \overset{w.p.1}{\to} \mathbb{P}^{\star}$ under i.i.d. assumptions, but the set $\mathscr{B}$ may still include distributions that may be far from $\mathbb{P}^{\star}$ (overly conservative). This conservatism may hurt performance, except in cases where the optimal solution only depends on first and second moments.

6.4.2 Metric-Based Uncertainty Sets

A more modern and widely adopted approach defines the uncertainty set via a **probability metric/discrepancy** between distributions. Specifically, given a discrepancy measure $D(\cdot,\cdot)$ on the space of probability measures $\mathscr{P}(\Xi)$, define

$$\mathscr{B}_{\delta}(\hat{\mathbb{P}}_n) = \left\{ \mathbb{Q} \,:\, D(\mathbb{Q}, \hat{\mathbb{P}}_n) \leqslant \delta \right\}.$$

Here, $\delta \geqslant 0$ defines the radius of a **ball** around $\hat{\mathbb{P}}_n$ with respect to the chosen discrepancy measure D.

Advantages: Such sets capture distributions *close* to the empirical distribution $\hat{\mathbb{P}}_n$.

6.5 Distances between distributions

6.5.1 ϕ-Divergence

Absolute continuity

A measure Q is said to be absolutely continuous w.r.t. measure P if for every $A \in \mathscr{F}$,

$$P(A) = 0 \Rightarrow Q(A) = 0$$

OR

$$Q(A) > 0 \Rightarrow P(A) > 0$$

Thus, $Q \ll P$ (read as *P dominates Q*).

Continuous Random Variables:

Let $\mathbb{P}_X(A) = \mathbb{P}(X \in A)$ which forms a probability space with $(\Omega, \mathscr{F})$ where $\Omega = \mathbb{R}$, $X : \Omega \to \mathbb{R}$ and $\mathscr{F}$ is Borel σ-field. If $\mathbb{P}_X(A)$ is absolutely continuous w.r.t. the Lebesgue measure λ, then, by the **Radon-Nikodym theorem**, there exists a $\mathscr{F}$-measurable function $p_X : X \to [0, \infty)$, such that for any $A \in \mathscr{F}$,

$$\mathbb{P}_X(A) = \int_{X^{-1}(A)} dP = \int_A p_X \, d\lambda.$$

The function p_X satisfying the above equality is commonly written as $\frac{d\mathbb{P}_X}{d\lambda}$ and is called **Radon-Nikodym derivative**.
In probability theory, this term is known as probability density function of a random variable. Usually, $p_X = p(X)$ and $A = [a,b] \Rightarrow \mathbb{P}_X([a,b]) = \int_a^b p(x)\, dx$.

Definition: [Of what? Use definition environment - AB] Assume $\phi : \mathbb{R}_+ \to (-\infty, +\infty]$ is a convex function with $\phi(0) = \lim_{t \to 0^+} \phi(t)$, then the ϕ-divergence between $\mathbb{Q}$ and $\hat{\mathbb{P}}_n$ is

$$D_\phi(\mathbb{Q}, \hat{\mathbb{P}}_n) = \begin{cases} \int_{\mathscr{Z}} \phi\left(\frac{d\mathbb{Q}}{d\hat{\mathbb{P}}_n}\right) d\hat{\mathbb{P}}_n(z), & \mathbb{Q} \ll \hat{\mathbb{P}}_n, \\ +\infty, & \text{otherwise.} \end{cases}$$

Where, $\frac{d\mathbb{Q}}{d\hat{\mathbb{P}}_n}$ = likelihood ratio (Radon–Nikodym derivative) and $\mathbb{Q}$ is absolutely continuous w.r.t. $\hat{\mathbb{P}}_n$.
The intuition using this is that the adversary can re-weight the relative importance ($w_i = a_i$) of each sample with a budget constraint (δ). So, the adversary systematically explores how re-weighting can potentially impact the performance of an estimator as measured by a given expected loss [2]. The bigger the reweighting, the larger the divergence.

Commonly used ϕ-divergences:

1. KL divergence: $\phi(t) = t \log t$
2. Reverse KL divergence: $\phi(t) = -\log t$
3. Total Variation distance: $\varphi(t) = \frac{1}{2}|t-1|$
4. Jensen-Shannon divergence: $\varphi(t) = \frac{1}{2}\left(t \log t - (t+1)\log(\frac{t+1}{2})\right)$

KL Divergence Intuition:

$$H(Q,P) - H(Q) = \sum_x Q(x) \log \frac{1}{P(x)} - \sum_x Q(x) \log \frac{1}{Q(x)}$$

$$\implies D_{\mathrm{KL}}(Q \| P) = \sum_x Q(x) \log \frac{Q(x)}{P(x)}$$

Thus, this implies excess entropy from incorrect distribution.

Total Variation Distance:

$$\mathrm{TV}(Q,P) = \int_z \frac{1}{2}\left|\frac{Q(x)}{P(x)} - 1\right| P(x)\,dx = \frac{1}{2}\int_z |Q(x) - P(x)|dx = \sup_{A\in\mathscr{F}} |Q(A) - P(A)|$$

HW: Show the last equality. Is total variation a metric?

Jensen-Shannon Divergence:

This is a symmetrized version of the KL-divergence.

$$\mathrm{JS}(Q,P) = \tfrac{1}{2}D_{\mathrm{KL}}\left(Q\middle\|\tfrac{Q+P}{2}\right) + \tfrac{1}{2}D_{\mathrm{KL}}\left(P\middle\|\tfrac{Q+P}{2}\right)$$

Notes

1. D_{KL} is not a metric ($\because$ not symmetric, does not satisfies triangle inequality).
2. Square root of JS is a metric.
3. In ϕ-divergence, $\mathbb{Q}$ is only supported where $\hat{\mathbb{P}}_n$ is supported.

Properties of ϕ-divergence

1. *Non-negativity*: For any two probability distributions Q and P,

$$D_\phi(Q,P) \geqslant 0.$$

Proof: From the definition,

$$D_\phi(Q,P) = \int_{\mathscr{Z}} \phi\left(\frac{dQ}{dP}\right) dP,$$

where $\phi : \mathbb{R}_+ \to (-\infty, +\infty]$ is convex with $\phi(1) = 0$.
By Jensen's inequality,

$$\int_{\mathscr{Z}} \phi\left(\frac{dQ}{dP}\right) dP \;\geqslant\; \phi\left(\int_{\mathscr{Z}} \frac{dQ}{dP} dP\right).$$

Since $\int_{\mathscr{Z}} \frac{dQ}{dP} dP = \int_{\mathscr{Z}} dQ = 1$, the RHS equals $\phi(1) = 0$. Thus, $D_\phi(Q,P) \geqslant 0$. □

2. *Identity of indiscernibles*:

$$D_\phi(Q,P) = 0 \quad \Longleftrightarrow \quad Q \overset{a.s.}{=} P \quad \text{(up to measure 0 set)}.$$

Proof: If $Q = P$, then $\frac{dQ}{dP} = 1$ everywhere and hence

$$D_\phi(Q,P) = \int_\Xi \phi(1) dP = 0.$$

Conversely, if $D_\phi(Q,P) = 0$, then by strict convexity of ϕ (minimum at $t = 1$), we must have

$$\frac{dQ}{dP} = 1 \quad P\text{-almost surely.}$$

This implies $Q = P$. □

Bibliography

[1] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

[2] John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[3] P.D. Lax. *Functional Analysis*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2014.

[4] Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

[5] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014.

[6] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Appendix A

Lebesgue integral with Dirac Measure

Let $(X, \mathscr{F})$ be a measurable space and fix $x \in X$. The Dirac measure at x, denoted δ_x, is defined by

$$\delta_x(A) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A, \end{cases} \quad \forall A \in \mathscr{F}.$$

We want to show that for any measurable function $f : X \to [-\infty, \infty]$,

$$\int_x f(y)\, d\delta_x(y) = f(x),$$

with the usual convention that the integral may take values $\pm\infty$ if f is not integrable. When f is integrable (i.e. $\int |f|\, d\delta_x < \infty$), the value is a finite real number equal to $f(x)$. [Add citation to where the proof is from - AB]

Proof

Step 1: Indicator functions

Let $A \in \mathscr{F}$. Then, by definition of the Lebesgue integral for indicator functions,

$$\int_x \mathbf{1}_A(y)\, d\delta_x(y) = \delta_x(A) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

But note that $\mathbf{1}_A(x)$ equals the same right-hand side. Hence

$$\int_x \mathbf{1}_A\, d\delta_x = \mathbf{1}_A(x).$$

Step 2: Simple functions

Let $s = \sum_{k=1}^n a_k \mathbf{1}_{A_k}$ be a (nonnegative) simple function with $a_k \geqslant 0$ and $A_k \in \mathscr{F}$. Using linearity of the integral and Step 1,

$$\int s\, d\delta_x = \sum_{k=1}^n a_k \int \mathbf{1}_{A_k}\, d\delta_x = \sum_{k=1}^n a_k\, \mathbf{1}_{A_k}(x) = s(x).$$

Step 3: Nonnegative measurable functions

Let $f \geqslant 0$ be measurable. By construction of the Lebesgue integral, there exists an increasing sequence of simple functions (s_n) with $0 \leqslant s_n \uparrow f$ pointwise. By the Monotone Convergence Theorem,

$$\int f \, d\delta_x = \lim_{n\to\infty} \int s_n \, d\delta_x = \lim_{n\to\infty} s_n(x) = f(x).$$

Step 4: General measurable functions

For a general measurable function f, write

$$f = f^+ - f^-, \quad \text{where } f^+ = \max\{f, 0\},\ f^- = \max\{-f, 0\}.$$

If at least one of $\int f^{\pm} \, d\delta_x$ is $+\infty$, Step 3 implies

$$\int f \, d\delta_x = f(x)$$

in the extended sense. If f is integrable (i.e. $\int |f| \, d\delta_x < \infty$), then applying Step 3 to $f^{\pm}$ and subtracting gives

$$\int_x f \, d\delta_x = \int f^+ \, d\delta_x - \int f^- \, d\delta_x = f^+(x) - f^-(x) = f(x).$$