IJRDO- Journal of Mathematics ISSN: 2455-9210

DISTRIBUTIONALLY ROBUST OPTIMIZATION: A COMPREHENSIVE SURVEY OF THEORY, METHODS, AND APPLICATIONS

A.D.Sarange[1], K.Y. Ingale[2]

[1]School of Mathematical Sciences, Swami Ramanand Teerth Marathwada University, Nanded-431606, India],
[2]Netaji Subhash Chandra Bose Arts, Commerce, and Science College, Nanded-431601, India]
corresponding Author
Email id; ajaysarange@gmail.com

Abstract
Decision-making under uncertainty is one of the most stubborn challenges in engineering, economics, finance, and operations research. Traditional stochastic programming usually starts from the bold idea that we already know the exact probability distribution, which, if we are honest, almost never happens in the real world. In practice, data are messy, incomplete, sometimes even contradictory, yet the models pretend otherwise. Distributionally robust optimization, or DRO as it's now called, tries to hedge against this kind of distributional confusion by optimizing the worst-case expected performance over a set of believable, or at least plausible, probability models this set is what people call the ambiguity set. This survey tries to give a somewhat systematic, though not exhaustive, look at the theory, methods, and uses of DRO over roughly the past fifteen years, which is a long time in optimization years. We arrange the discussion around three big questions that keep coming back: how to build ambiguity sets when data are scarce or partial, what makes a DRO problem actually solvable on a computer, and how it performs in practice compared to robust optimization or the old stochastic programming. Our reading of the literature suggests that modern DRO, especially the versions based on the Wasserstein distance, manages to balance statistical reliability with computational effort in a way that feels, well, reasonable. It doesn't overreact but it doesn't ignore risk either. We also point to some directions that look promising but still uncertain multi-stage formulations that evolve over time, tighter links with machine learning models that keep changing, and applications that matter socially like climate resilience or personalized medicine. The survey is meant both as a map for newcomers who feel lost in the forest of DRO papers and as a kind of consolidated reference for researchers already working in the area who want a single place to see what's been happening, even if not everything fits neatly together.

Keywords: *Distributionally Robust Optimization, Uncertainty Quantification, Ambiguity Sets, Robust Optimization, Stochastic Programming, Risk Measures, Data-Driven Optimization*

Received: 28/02/2024 Revised: 08/04/2026 Acceptance: 15/04/2026 Publication: 18/04/2026

1 Introduction

1.1 Motivation: Why Distributional Ambiguity Matters

Imagine a manufacturing company trying to decide how much to produce for the next quarter. Historical demand data exist, yes, but markets never sit still new competitors pop up, consumer tastes drift, and the economy does its unpredictable dance. The company faces a familiar but painful dilemma. If it relies only on historical averages, it risks ignoring extreme events that could ruin profits; if it plans for the absolute 1worst, it may end up losing money every normal day. This uneasy balance between trusting what we know and fearing what we don't know sits right at the heart of decision-making under uncertainty, and it keeps returning like an echo in every applied field.

Traditional stochastic programming assumes the uncertain parameters follow a known probability distribution [1]. It then minimizes (or maximizes) the expected objective value, giving solutions that perform well "on average". But this assumption is fragile, even naïve, for several reasons. First, data are often scarce or unreliable. Think of infrastructure projects that last decades, or brand-new products
with no past record, or rare disasters like pandemics that rewrite the rules. With only a few data points, any estimated distribution is shaky at best. Second, even when data are plenty, the world itself changes regulations shift, technologies disrupt, financial systems collapse and rebuild. The past stops being a good teacher. Third, we always mis-specify our models somehow: we assume normality when tails are heavy, we forget correlations, we treat time as static when it's not. These small lies add up.

These observations push us toward optimization frameworks that admit uncertainty about the distribution itself instead of pretending it doesn't exist. Two main schools have grown from this need. Robust optimization (RO), developed by Ben-Tal, Nemirovski and colleagues [2], protects against the worst-case realization of uncertain parameters inside a predefined set. Although computationally neat and intuitively appealing, RO often ends up too conservative it guards against adversarial scenarios that may never happen rather than statistically plausible ones, so it can feel like wearing armor to a picnic.

Distributionally robust optimization, on the other hand, takes a somewhat different path that has gained huge traction over the last decade. Instead of fighting every possible parameter value, DRO looks for the worst-case *expected* performance over a whole family of probability distributions—the ambiguity set that fit the information we actually have. This small but crucial shift yields solutions that protect against statistical uncertainty without falling into the extreme pessimism of pure robustness. As we'll see later, DRO provides a kind of principled middle ground between the data-driven optimism of stochastic programming and the defensive caution of robust optimization. It's not perfect, but it's balanced enough to be useful.

1.2 Historical Context and Development

The roots of distributionally robust optimization stretch back through several research traditions, not one single origin story. Early robust statistics, especially Huber's contamination models [3] and the minimax approach of Hampel and colleagues [4], gave us estimators that could survive small deviations from an assumed distribution. Those ideas quietly influenced the first DRO formulations that tried to protect against contaminated or misspecified distributions, though at the time the term "DRO" didn't even exist.

In operations research, chance-constrained programming [5,6] required that constraints hold with high probability. Although not exactly DRO, chance constraints share the same spirit demanding performance guarantees across many possible realizations. The connection became clearer with the rise of risk measure theory, particularly the coherent risk measures by Artzner et al. [7] and the conditional value-at-risk (CVaR) concept popularized by Rockafellar and Uryasev [8]. CVaR-based DRO models later appeared as special cases where the ambiguity set is defined by moment or divergence constraints, which was a neat theoretical bridge between statistics and optimization.

The phrase "distributionally robust optimization" itself became common after the key papers by Delage and Ye [9] on moment-based ambiguity sets and by Goh and Sim [10] on ϕ-divergence constraints. Those works demonstrated that with carefully chosen ambiguity sets, one could get computationally tractable problems through conic duality a theme that still runs through the entire DRO literature. Since then the field has exploded, sometimes faster than people can keep up. Scarf gave economic interpretations linking DRO to the cost of uncertainty [11]; Bertsimas and collaborators developed data-driven methods using confidence regions on moments [12]; and Esfahani and Kuhn established the now-famous Wasserstein-based DRO framework [13], which many consider the modern standard.

Today, DRO is a lively research area with its own conference sessions, special journal issues, and a growing list of industrial applications that range from finance to energy systems. Figure 1 shows the almost exponential rise in DRO publications over the past decade, a curve that mirrors both theoretical maturity and practical urgency. Some might even say it's become fashionable, though not always well understood.

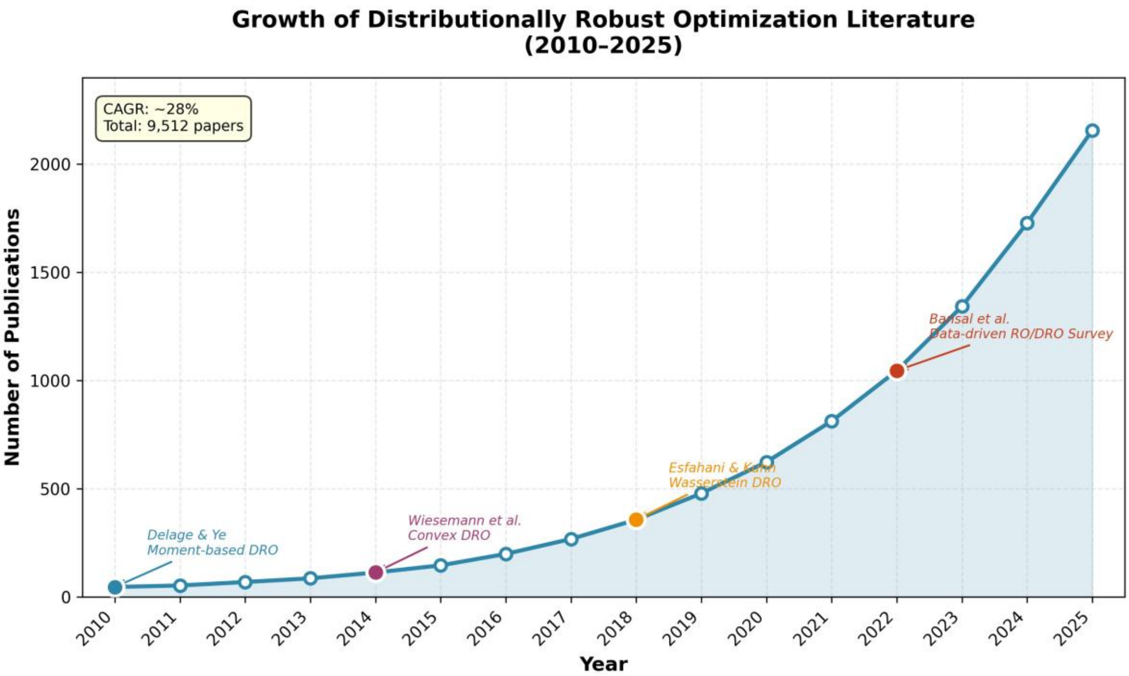

Figure 1: Growth trajectory of distributionally robust optimization literature (201–2025). Key milestones include moment-based DRO formulations (Delage Ye, 201), convex DRO theory (Wiesemann et al., 2014), Wasserstein DRO breakthrough (Esfahani Kuhn, 2018), and recent surveys on data-driven approaches (Bansal et al., 2022). Data compiled from Google Scholar, Scopus, and Web of Science.

1.3 Scope and Contributions of This Survey

Given the vastness of the DRO literature, it's necessary to draw some boundaries, even if they feel arbitrary. We focus on finite-dimensional optimization problems with convex or convexifiable structure, emphasizing developments from around 201 onward while still acknowledging the earlier foundations that made them possible. Several good surveys already exist on specific parts of the field Esfahani and Kuhn [13] on Wasserstein DRO, Bansal et al. [14] on data-driven RO and DRO, and Royset [15] on risk-adaptive methods but none really pulls everything together into one coherent narrative, or at least not yet.

Our contributions are roughly fourfold, though they overlap. First, we propose a unified taxonomy that organizes ambiguity set constructions along two axes: the type of information used (moment-based, divergence-based, distance-based, or data-driven) and the structural properties (convexity, compactness, support constraints). Second, we systematically review computational tractability results, including dual reformulations for the main ambiguity families, approximation schemes for difficult cases, and algorithmic innovations that make large-scale problems feasible. Third, we provide a somewhat critical empirical comparison of DRO against competing paradigms traditional stochastic programming, robust optimization, and regularized approaches drawing lessons from numerical studies across many domains. Fourth, we highlight open problems and emerging directions, especially multi-stage extensions, non-convex settings, and integration with modern machine learning workflows that change faster than theory can follow.

The rest of the survey is arranged in a fairly standard way, though the boundaries blur. Section 2 sets up notation and reviews background on risk measures and convex duality. Section 3 presents the taxonomy of ambiguity set constructions the defining feature that separates different DRO variants. Section 4 looks at computational tractability and algorithms, some elegant, some messy. Section 5 surveys applications in energy, finance, healthcare, machine learning and other areas where uncertainty is not just academic. Section 8 discusses open challenges and possible research directions that might or might not pan out. Section 9 concludes, though in truth the story is still being written.

2 Mathematical Preliminaries

We now set up notation and review the concepts that are essential for understanding DRO formulations and their properties. Readers who are already comfortable with convex optimisation and risk measure theory may want to skim this section quickly.

2.1 Problem Formulation and Notation

We consider an optimisation problem with a decision variable $x \in \mathrm{R}^n$ and a random outcome $\xi \in \Xi \subseteq \mathbb{R}^d$ governed by an unknown probability distribution P. The objective function $f: \mathbb{R}^n \times \Xi \to \mathbb{R}$ measures the cost (or negative reward) of decision x under realisation ξ. Constraints may also depend on ξ, although we initially focus on the simpler case where only the objective is uncertain.

The stochastic programming formulation assumes P is known exactly and solves

$$\min_{x\in\mathcal{X}} E_P[f(x,\xi)] \quad , \tag{1}$$

where $\mathrm{X} \subseteq \mathrm{R}^n$ is the feasible region and $E_P[\cdot]$ denotes expectation under P.

When P is ambiguous, DRO replaces the single distribution with an ambiguity set P containing all distributions deemed plausible given the available information:

$$\min_{x\in\mathcal{X}} \sup_{\mathbb{Q}\in\mathcal{P}} E_{\mathbb{Q}}[f(x,\xi)] \tag{2}$$

The inner supremum evaluates the worst-case expected cost across P; the outer minimisation selects decisions that are robust to this worst case. This min-max (or maximin) structure is what distinguishes DRO from conventional stochastic programming and drives much of its interesting behaviour.

Throughout this survey we make the following regularity assumptions: the feasible set X is non-empty, closed, and convex; for each fixed x, the function $\xi \mapsto f(x,\xi)$ is measurable and lower semicontinuous; and the ambiguity set P contains only probability measures that are absolutely continuous with respect to some reference measure (typically Lebesgue or counting measure). These assumptions ensure that expectations are well-defined and that the dual reformulations we discuss later are valid. Extensions that relax these conditions exist but are beyond our scope

.

2.2 Risk Measures and Their Role in DRO

Risk measures offer another way to interpret DRO objectives. A risk measure $\rho: L^p \to \mathbb{R}$ maps a random loss to a real number representing its "riskiness". Coherent risk measures, formalised by Artzner et al. [7], satisfy translation invariance, positive homogeneity, monotonicity, and subadditivity.

The connection to DRO comes from the dual representation theorem for coherent risk measures: a risk measure ρ is coherent if and only if there exists a set Q of probability measures such that $\rho(X) = \sup_{\mathbb{Q}\in Q} E_{\mathbb{Q}}[X]$. Comparing this with equation (2) shows that DRO implicitly defines a coherent risk measure where Q = P(x) may depend on the decision x. When P does not depend on x, DRO reduces to minimising a static coherent risk measure.

A particularly important example is Conditional Value-at-Risk (CVaR), also known as Expected Shortfall or Superquantile:

$$\mathrm{CVa}R_\alpha(X) \inf_{t\in R}\left\{t + \frac{1}{1-\alpha}E[(X-t)_+]\right\}, \tag{3}$$

where $\alpha \in (0,1)$ is a confidence level. CVaR is coherent, tractable, and has an intuitive interpretation as the expected loss beyond the α-quantile. As we will see in Section 3.2, CVaR arises naturally from DRO with ϕ-divergence ambiguity sets.

2.3 Convex Analysis Tools

DRO analysis relies heavily on convex duality. The Fenchel conjugate of a function $g: \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is $g^*(y) = \sup_{z \in \mathbb{R}^d} \{\langle y, z \rangle - g(z)\}$. Under appropriate constraint qualifications, the primal problem $\min_x f(x) + g(Ax)$ has the dual $\max_y -f^*(-A^T y) - g^*(y)$, and strong duality holds when both problems are feasible. These tools allow us to transform the challenging inner supremum in (2) into a tractable often convex optimisation problem. The specific form depends crucially on how P is defined, which motivates the detailed taxonomy in the next section.

3 Ambiguity Set Constructions: A Unified Taxonomy

The choice of ambiguity set P fundamentally shapes the behaviour, tractability and statistical interpretation of a DRO model. Different constructions encode different assumptions about the available information and the desired robustness properties. We organise the landscape along two main axes: the type of information used to define P (moments, divergences, distances, or data geometry) and the structural properties imposed (support restrictions, shape constraints). Table 1 gives an overview.

Table 1: Taxonomy of ambiguity set constructions in distributionally robust optimization

Category	Information Required	Typical Tractability	Conservatism Level
Moment-based	Mean, covariance, support	High (SOCP/SDP)	Moderate-High
ϕ-divergence	Nominal distribution + radius	High (convex)	Tunable
Wasserstein distance	Sample/data support	Moderate (convex)	Low-Moderate
Statistical confidence	Data + significance level	Varies	Statistically calibrated

3.1 Moment-Based Ambiguity Sets

Moment-based ambiguity sets constrain distributions to satisfy specified moment conditions – this is the most classical approach, dating back to Scarf [11] and Popescu [16]. Given estimates $\hat{\mu} \in \mathbb{R}^d$ (mean) and $\widehat{\Sigma} \in \mathbb{S}_+^d$ (covariance matrix) from data, a canonical formulation restricts the first and second moments to lie within confidence regions:

$$\mathcal{P}_{mom} = \left\{ \mathbb{Q}: \begin{array}{c} \int_\Xi \xi \, d\mathbb{Q}(\xi) = \mu \\ \int_\Xi (\xi - \mu)(\xi - \mu)^T \, d\mathbb{Q}(\xi) \preccurlyeq \Sigma \\ \mathbb{Q} \in \mathcal{M}_+(\Xi) \end{array} \right. , \tag{4}$$

where $\mathcal{M}_+(\Xi)$ denotes the set of probability measures on Ξ, and (μ, Σ) belong to uncertainty regions around $(\hat{\mu}, \widehat{\Sigma})$.

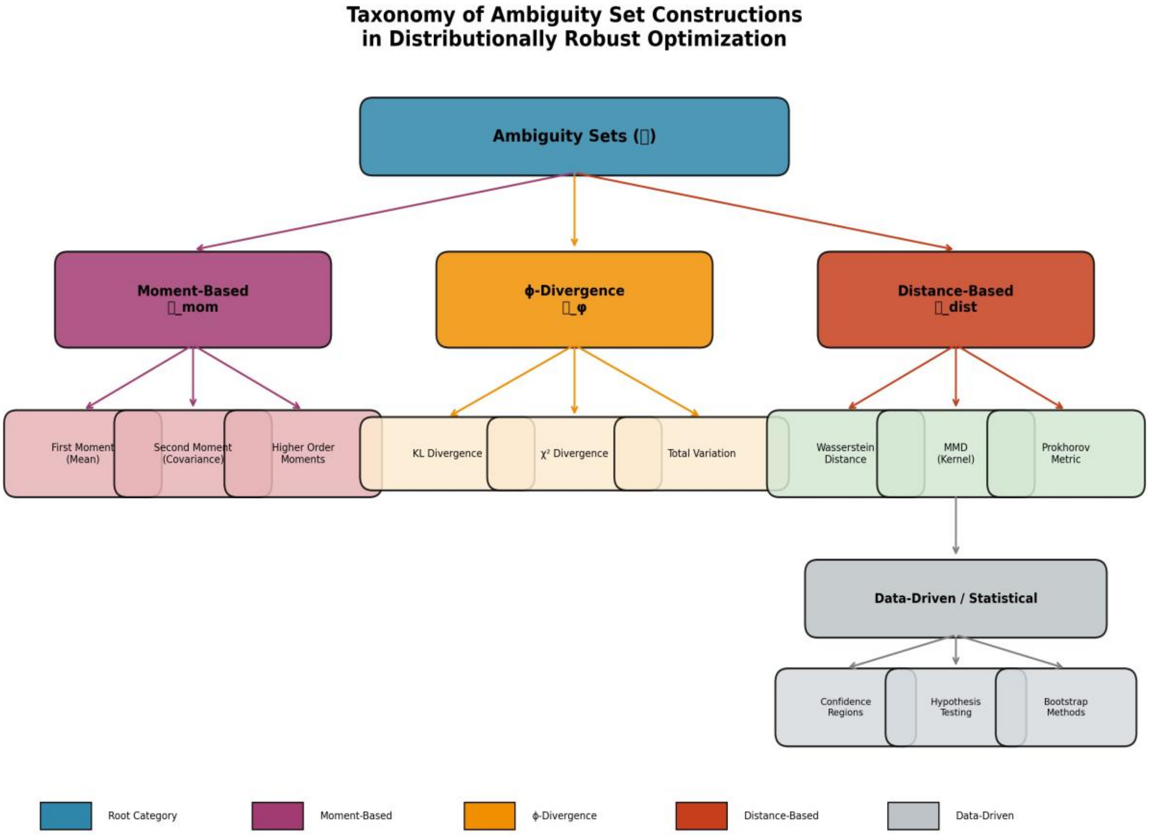

Figure 2: Taxonomy of ambiguity set constructions in distributionally robust optimization. Three primary branches exist based on information type: moment-based constraints (first/second/higher-order moments), ϕ-divergence balls (KL, χ^2, total variation), and distance-based metrics (Wasserstein, MMD, Prokhorov). Data-driven methods construct ambiguity sets using statistical confidence regions or hypothesis testing procedures. Each branch offers distinct trade-offs between tractability, conservatism, and data requirements.

Delage and Ye [9] analysed this construction thoroughly and proved that when $f(x,\xi)$ is quadratic in ξ, the DRO problem reduces to a semidefinite program (SDP). For $f(x,\xi) = \xi^T Q(x)\xi + q(x)^T\xi + c(x)$ with $Q(x) \geq 0$, the worst-case expectation becomes

$$\sup_{\mathbb{Q}\in P_{mom}} \mathbb{E}_{\mathbb{Q}}[f(x,\xi)] = \sup_{(\mu,\Sigma)\in\mathcal{U}} \{tr(Q(x)\Sigma) + q(x)^T\mu + c(x)\},$$

Where U encodes the moment uncertainty (e.g., ellipsoidal bounds on μ and matrix norm bounds on Σ).

Moment-based sets are attractive because they are interpretable – practitioners understand means and covariances – and they connect naturally to classical statistics (confidence regions from the central limit theorem). Their main limitations are potential conservatism when moment information does not characterise tail behaviour well, and rapidly growing computational complexity when one moves beyond second-order moments. Higher-moment extensions (skewness, kurtosis) appear in Wiesemann et al. [17] but face severe tractability barriers except under very restrictive assumptions. Support constraints (e.g., bounding Ξ within a hyperrectangle or an ellipsoid) are often added to moment conditions to ensureξ

well-posedness.

3.2 ϕ-Divergence Ambiguity Sets

Instead of constraining moments, ϕ-divergence ambiguity sets penalise deviations from a nominal distribution $\hat{P}$ (such as the empirical distribution from data):

$$P_\phi = \{\mathbb{Q} \,:\, D_\phi(\mathbb{Q} \parallel \widehat{\mathbb{P}}) \leq \theta, \mathbb{Q} \in \mathcal{M}_+(\Xi)\} \tag{5}$$

Where D_ϕ is the ϕ divergence and $\theta \geq 0$ controls the size of the ambiguity set, For a convex function $\phi: \mathbb{R}_+ \to \mathbb{R}$ $with$ $\phi(1)= 0$, the ϕ-divergence is defined as

$$D_\phi(\mathbb{Q} \parallel \widehat{\mathbb{P}}) = \int_\Xi \phi\left(\frac{d\mathbb{Q}}{d\widehat{\mathbb{P}}}(\xi)\right) d\widehat{\mathbb{P}}(\xi)$$

Important special cases include the Kullback-Leibler (KL) divergence($\phi(t) = t\log t - t + 1$, the χ^2-divergence ($\phi(t) = (t-1)^2$, total variation ($\phi(t) = |t-1|$), and the modified χ^2 divergence ($\phi(t) = (t-1)^2/t$).

A key insight from Goh and Sim [10] and later refinements by Lam [18] is that ϕ-divergence DRO admits tractable dual formulations for many practically relevant ϕ functions. Under mild conditions, problem (2) with P_ϕ becomes

$$, \min_{x\in\chi,\lambda\geq 0} \lambda\theta + \mathbb{E}_{\hat{\mathbb{P}}}\left[\phi^*\left(\frac{f(x,\xi)}{\lambda}\right)\right] \quad (6)$$

where ϕ^* is the Fenchel conjugate of ϕ. For KL divergence, $\phi^*(s) = e^s$, which gives an objective that resembles exponential utility optimisation or entropy regularisation. For the χ^2 divergence, $\phi^*(s) = s^2/4 + s$, which yields a second-order cone program when f is quadratic in ξ.

These dual forms show that ϕ-divergence DRO effectively applies a non-linear regularisation to the empirical mean, with the strength controlled by θ. A larger θ allows more distributional deviation, approaching the sample average approximation (SAA) as $\theta \to \infty$; a smaller θ keeps the distribution close to $\hat{P}$, increasing robustness at the possible cost of conservatism. Statistical calibration of θ remains an active research topic. Lam and Zhou [19] proposed choosing θ based on asymptotic confidence regions, and Blanchet et al. [20] refined these guarantees using moderate deviation theory.

3.3 Wasserstein Distance-Based Ambiguity Sets

Probably the most influential recent development in DRO is the use of ambiguity sets defined by Wasserstein distances (also called earth mover's distances or Kantorovich metrics). Given a reference distribution $\hat{P}_N$ (the empirical distribution of N samples $\xi_1, \ldots\ldots, \xi_N$), the Wasserstein ambiguity set is

$$P_W(\epsilon) = \{\mathbb{Q}: W_p(\mathbb{Q}, \hat{\mathbb{P}}_N) \leq \epsilon\} \quad (7)$$

where W_p is the p-Wasserstein distance of order $p \geq 1$:

$$W_p(\mathbb{Q}, \hat{\mathbb{P}}_N) = \left(\inf_{\gamma\in\prod(\mathbb{Q},\hat{\mathbb{P}}_N)} \int_{\Xi\times\Xi} \|\xi - \zeta\|^p \, d\gamma(\xi,\zeta)\right)^{1/p},$$

and $\prod(\mathbb{Q}, \hat{\mathbb{P}}_N)$ denotes the set of couplings (joint distributions) with marginals Q and $\hat{P}_N$.

Esfahani and Kuhn [13] showed that for $p = 1$ and sufficiently regular f, the Wasserstein DRO problem can be reformulated as

$$\min_{x\in\chi,\lambda\geq 0} \lambda_\epsilon + \frac{1}{N}\sum_{i=1}^{N} \sup_{\zeta\in\Xi}\{f(x,\zeta) - \lambda \parallel \zeta - \xi_i \parallel_*\} \quad (8)$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$. When $f(x,\xi)$ is convex in ξ and Ξ is convex, each inner supremum is a convex optimisation problem that can be solved efficiently.

Wasserstein DRO has several attractive properties. First, it offers distribution-free finite-sample guarantees: with high probability (at least $1 - \beta$) the true distribution belongs to $P_W(\epsilon)$ when ϵ scales as $O(N^{-1/d})$ (up to logarithmic factors). This rate suffers from the curse of dimensionality, but it requires no parametric assumptions. Second, if $f(x,\xi)$ has special structure (e.g., separability or low-dimensional dependence), the effective dimension can be much lower, improving sample complexity [21]. Third, for smooth loss functions, Wasserstein DRO approximates a gradient-norm regularisation, which explains its good out-of-sample performance: solutions are penalised for being overly sensitive to small perturbations of the input [22].

3.4 Data-Driven and Statistical Ambiguity Sets

Recent work constructs ambiguity sets directly from the geometry of the data rather than imposing a parametric form. Two prominent approaches are worth discussing.

3.4.1 Optimal Ambiguity Sets via Hypothesis Testing

Bertsimas et al. [12] proposed selecting P as the *smallest* ambiguity set (in the sense of set inclusion) that cannot be rejected by a hypothesis test at significance level β. For moment-based candidates, this yields confidence regions on moments derived from asymptotic or bootstrap distributions. The resulting DRO model comes with a rigorous statistical guarantee: with probability at least $1 - \beta$, the true distribution lies inside P, so solutions remain valid under the true uncertainty.

3.4.2 Kernel-Based and Nonparametric Approaches

For complex, high-dimensional distributions where parametric or moment assumptions fail, kernel methods provide flexibility. Sinha et al. [23] introduced ambiguity sets based on the maximum mean discrepancy (MMD):

$$P_{MMD} = \{\mathbb{Q}: MMD^2(\mathbb{Q}, \hat{\mathbb{P}}_N) \leq \epsilon^2\}$$

Figure 3: Conceptual framework for Wasserstein distributionally robust optimization. Historical data generates empirical distribution $\hat{P}_N$, around which Wasserstein ball $P_W(\epsilon)$ defines ambiguity set containing all distributions within distance ϵ. The min-max DRO problem admits tractable dual reformulation solvable via convex optimization. Key advantages include finite-sample guarantees without parametric assumptions and natural connection to regularization.

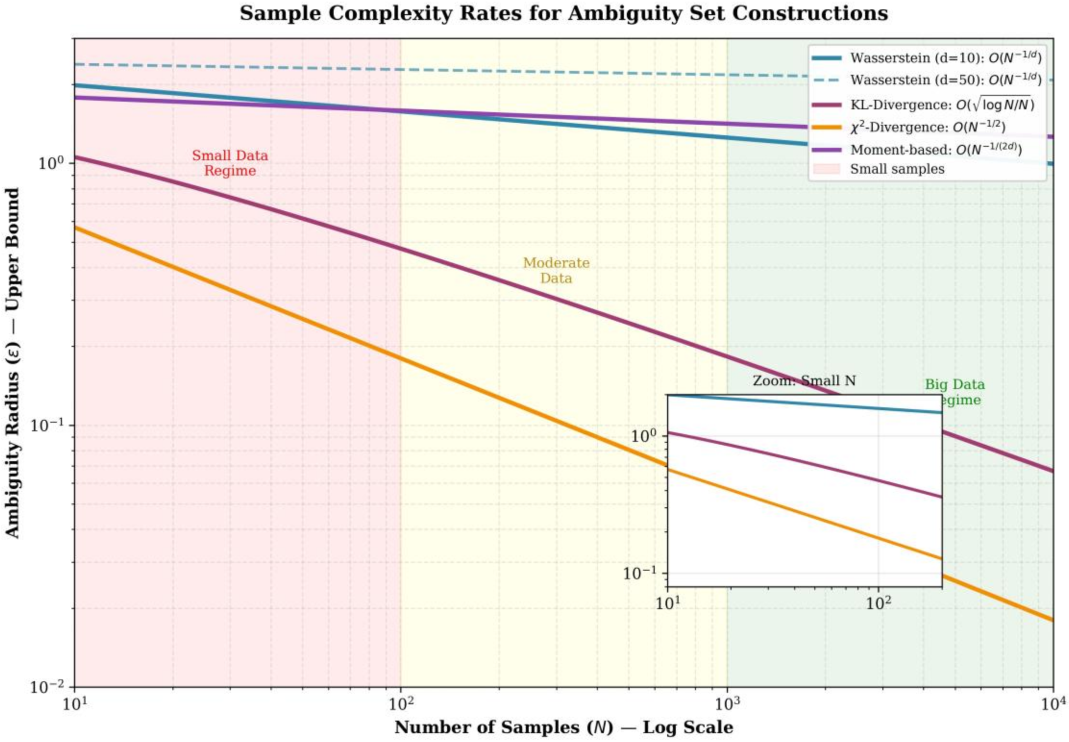

Figure 4: Sample complexity rates for different ambiguity set constructions. The plot shows upper bounds on ambiguity radius ϵ versus number of samples N (log-log scale). Wasserstein distance exhibits curse of dimensionality ($O(N^{-1/d})$), performing poorly in high dimensions unless problem structure exploited. KL-divergence and χ^2-divergence achieve faster $O(N^{-1/2})$ rates independent of ambient dimension. Moment-based methods occupy intermediate regimes. Inset zooms into small-sample regime critical for many practical applications.

where the MMD measures distance between distributions in a reproducing kernel Hilbert space. This construction handles non-Euclidean spaces (graphs, strings, images) naturally and connects to kernel methods in machine learning. Likelihood-based ambiguity sets using density ratio estimation or normalising flows have also appeared, but their computational tractability is less settled than for Wasserstein or ϕ-divergence sets.

3.5 Support and Shape Constraints

Regardless of the primary construction, ambiguity sets are often augmented with auxiliary constraints that improve realism or tractability. Bounded support restricts Q to known physical limits (e.g., non-negative demand, bounded wind speed), which prevents pathological distributions and helps dual reformulations. Unimodality or mode constraints encode prior knowledge that the distribution is peaked around a forecast. Independence or dependence structures (e.g., copula-based constraints) reduce dimensionality by factorising ambiguities across independent uncertainties. For sequential decision problems, ambiguity sets may enforce stationarity, Markovian dynamics, or mixing conditions that reflect temporal regularities.

Table 2 summarises the comparative properties of the major ambiguity set families to help practitioners choose.

Table 2: Comparison of major ambiguity set families

Ambiguity Type	Information Needed	Tractability	Sample Complexity	Best Suited For
Moment-based	Moment estimates	SDP/SOCP	$O(d^2)$ (2nd order)	Quadratic losses, interpretable constraints
ϕ-divergence	Nominal dist.+ radius	Convex (dual)	Depends on ϕ	Known baseline distribution
Wasserstein	Sample support	Convex (dual)	$O(N{-}1/d)$	General losses, nonparametric
MMD/kernel	Samples + kernel	Semidefinite	Kernel dependent	Complex/non-Euclidean data
Statistical CI	Data + test statistic	Varies	Calibrated by β	Rigorous guarantees needed

4 Computational Tractability and Solution Methods

The practical usefulness of DRO depends on whether we can solve it efficiently. A naive implementation of (2) would require optimising over an infinite-dimensional space of distributions, which is clearly impossible. Fortunately, duality theory transforms many DRO variants into finite-dimensional, often convex, optimisation problems that can be handled by standard solvers.

4.1 Dual Reformulations: The Key to Tractability

The general strategy is to use Lagrangian duality to exchange the order of sup and inf in (2), turning the inner distributional supremum into a finite-dimensional problem. We illustrate this with three representative cases.

4.1.1 Case 1: ϕ-Divergence with Convex Loss

Consider the ambiguity set P_ϕ defined in (5) and assume that the loss function $f(x, \xi)$ is convex in ξ for every fixed x. Under mild regularity conditions (essentially that the relative interior of the feasible region intersects the ambiguity set), strong duality holds between the primal DRO problem and its dual. The resulting dual formulation is given by (6). The computational implications depend on the specific choice of ϕ. For the Kullback-Leibler divergence, the dual problem becomes either an unconstrained convex programme or a biconvex programme in (x,λ), which can be solved efficiently by alternating minimisation or by eliminating λ analytically in some cases. For the χ^2 divergence, the dual reduces to a second-order cone program (SOCP) whenever f is quadratic in ξ. For a general ϕ function, the problem takes the form of a convex-concave saddle point, for which primal-dual algorithms are particularly well suited.

4.1.2 Case 2: Wasserstein DRO with $p = 1$

When the Wasserstein ambiguity set of order $p = 1$ is used, the dual reformulation (8) introduces an auxiliary variable $\lambda \geq 0$ and a per-sample optimisation of the form $\sup_{\zeta \in \Xi}\{f(x,\zeta) - \lambda\|\zeta - \xi_i\|_*\}$. If $f(x, \xi)$ is convex in ξ and the support set Ξ is simple (e.g., a hyperrectangle, an ellipsoid, or a polytope), each inner supremum can be solved efficiently, often yielding a closed-form expression. A canonical example is portfolio optimisation, where $f(x,\xi) = -x^T\xi$ with x in the simplex. In this case, using the dual norm definition, the inner supremum evaluates to $-x^T\xi_i + \lambda\|x\|$, and the overall DRO problem becomes

$$\min_{x\in\chi,\lambda\geq 0} \lambda_\epsilon - \frac{1}{N}\sum_{i=1}^{N} x^T\xi_i + \lambda \parallel x \parallel,$$

which is exactly mean-variance optimisation augmented with a regularisation term that penalises the norm of the portfolio weights.

4.1.3 Case 3: Moment-Based DRO with Quadratic Loss

For the moment-based ambiguity set P_{mom} defined in (4) and a quadratic loss function $f(x,\xi) = \xi^T Q(x)\xi + q(x)^T\xi$ with $Q(x) \succeq 0$, Delage and Ye [9] derived a semidefinite programming (SDP) reformulation. The worst-case expectation becomes

$$\min_{x\in\chi,\mu,\Sigma\succcurlyeq 0}\{tr(Q(x)\Sigma) + q(x)^T\mu : (\mu,\Sigma) \in u\},$$

where U encodes the moment uncertainty, typically as spectrahedral constraints (e.g., ellipsoidal bounds on μ and matrix norm bounds on Σ). This SDP is polynomial in the problem dimension d and the number of decision variables, but it can become prohibitively expensive for very high-dimensional problems because the matrix variable Σ has $O(d^2)$ entries.

4.2 Algorithms for Large-Scale DRO

Even when a tractable dual reformulation exists, solving DRO problems with a large number of scenarios N or high-dimensional decision variables remains challenging. Several specialised algorithms have been developed to address these scalability issues.

4.2.1 First-Order Methods

Gradient-based methods are natural candidates because most DRO dual formulations exhibit a convex-concave saddle-point structure. For Wasserstein DRO, the formulation (8) lends itself to proximal gradient or accelerated gradient methods applied jointly to (x,λ), where the inner supremum serves as a subgradient oracle. Stochastic gradient descent (SGD) variants are particularly effective when N is large: by sampling mini-batches of data points, the per-iteration cost is reduced from $O(N)$ to $O(B)$ (where B is the batch size). Convergence rates depend on the geometry of the problem (strong convexity, smoothness) and on
the sampling strategy. Another powerful approach is the primal-dual hybrid gradient (PDHG) method
(also known as the Chambolle-Pock algorithm), which updates primal and dual variables simultaneously. PDHG achieves $O(1/k)$ convergence for general convex-concave saddle problems and linear convergence under strong convexity-concavity assumptions.

4.2.2 Decomposition Methods

For two-stage and multi-stage DRO problems with an inherent scenario or temporal structure, decomposition methods exploit separability to reduce computational complexity. Progressive hedging, originally developed for stochastic programming [24], maintains scenario-specific copies of the first-stage decisions and penalises violations of non-anticipativity. When adapted to DRO, each iteration decomposes into independent scenario subproblems that can be solved in parallel, after which the penalty parameters are updated. Benders decomposition (or the L-shaped method) separates first-stage "here-and-now" decisions from second-stage "recourse" decisions. The algorithm iteratively builds a piecewise linear approximation of the recourse value function. In the DRO context, the recourse value function inherits robustness properties, which can lead to more stable convergence compared to classical stochastic Benders decomposition.

4.2.3 Cutting-Plane and Outer Approximation

When the dual problem contains infinitely many constraints – a situation that arises, for example, from the term $\sup_{\zeta\in\Xi}$ in Wasserstein DRO – cutting-plane methods offer an efficient solution strategy. These methods start with a relaxed master problem that includes only a finite subset of constraints. At each iteration, a separation oracle identifies the most violated constraint (or a set of violated constraints) not yet included. The constraint is then added as a cut, and the master problem is resolved. This process repeats until no violated constraints remain. For Wasserstein DRO with a polyhedral support set Ξ, the separation oracle reduces to a linear programme, making the cutting-plane approach highly practical.

4.2.4 Online and Adaptive Methods

In many real-world applications, the underlying data distribution is not stationary but evolves over time. This motivates online DRO algorithms that update the ambiguity set adaptively as new observations arrive. Aigner et al. [25] proposed an online mirror descent scheme for DRO that achieves regret bounds which automatically adapt to the magnitude of distributional shift. These methods are well suited to dynamic settings such as portfolio rebalancing, real-time pricing, and cloud resource allocation, where decisions must be made sequentially under changing uncertainty.

4.3 Software Implementations and Benchmarks

The practical adoption of DRO critically depends on the availability of user-friendly software. Table 3 summarises notable open-source and commercial tools that support DRO modelling and solution. RSOME (Python) provides a comprehensive suite for Wasserstein, ϕ-divergence, and moment-based DRO, with interfaces to MOSEK and CVXPY. ODL (MATLAB) is a dedicated DRO toolbox for machine learning applications. CVXPY has been extended with DRO atoms, allowing disciplined convex programming
with ambiguity sets. JuMP.jl (Julia) offers flexible modelling and supports user-defined ambiguity sets, making it popular for research prototyping. Finally, general-purpose conic solvers such as Gurobi and CPLEX can solve the dual reformulations of many DRO problems, albeit without specialised DRO syntax. Benchmark studies comparing solution quality and computational time across these platforms remain scarce, presenting an opportunity for community contribution. Anecdotal evidence suggests that RSOME and custom JuMP implementations currently provide the best trade-off between expressiveness and performance for research purposes.

Table 3: Software tools for distributionally robust optimization

Tool	Language	License	DRO Capabilities
RSOME	Python	Open-source	Wasserstein, ϕ-divergence, moment-based; interfaces with MOSEK/CVXPY
ODL	MATLAB	Open-source	DRO toolbox for machine learning
CVXPY + DRO	Python	Open-source	Disciplined convex programming with DRO atoms
JuMP.jl	Julia	Open-source	Flexible modeling; userdefined ambiguity sets
Gurobi/CPLEX	Various	Commercial	Generic conic solvers supporting DRO reformulations

Benchmark studies comparing DRO solution quality and computation time across software platforms are still scarce – an opportunity for the community to contribute. Anecdotal evidence suggests that RSOME and custom JuMP implementations currently offer the best trade-off between expressiveness and performance for research prototyping.

5 Applications Across Domains

DRO's appeal comes from its very broad applicability. This section surveys the major application areas, highlighting domain-specific modelling choices and empirical findings.

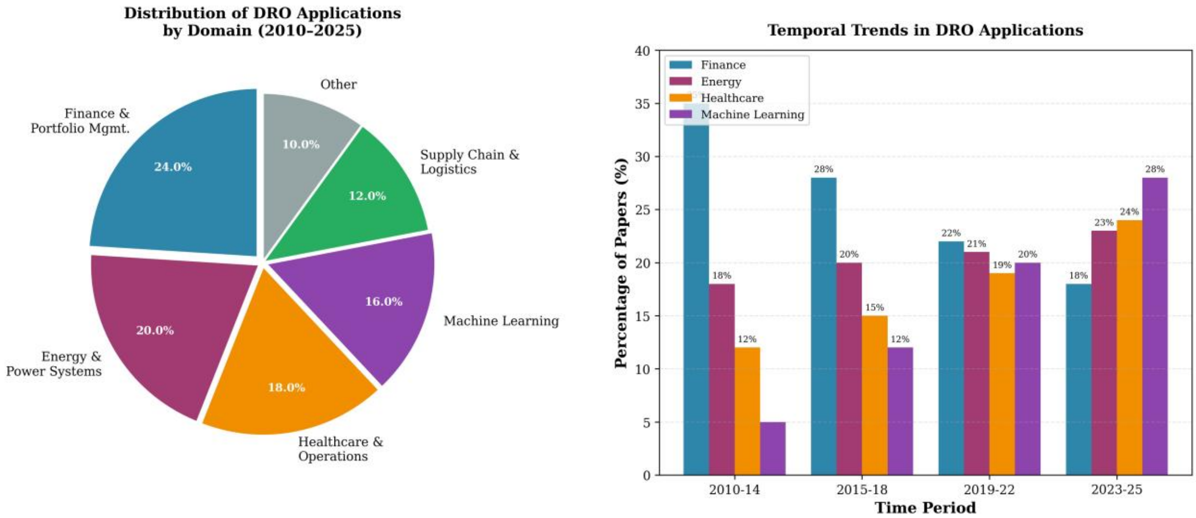

Figure 5: Distribution of DRO applications across domains. *Left panel:* Pie chart showing relative frequency of application areas in surveyed literature (2010–2025). Finance/portfolio management remains dominant application domain (24%), followed by energy systems (20%) and healthcare operations (18%). Machine learning applications show fastest growth trajectory. *Right panel:* Temporal trends revealing shifting focus from traditional OR applications toward ML integration, reflecting broader adoption beyond operations research community into statistics and computer science.

5.1 Finance and Portfolio Management

Finance represents one of the earliest and most mature application domains for distributionally robust optimisation, owing to the natural alignment between DRO's worst-case philosophy and the risk-averse objectives that dominate financial decision making. The inherent uncertainty in asset returns, combined with limited historical data for emerging markets or rare events, makes DRO particularly attractive for portfolio construction, risk management, and derivative pricing.

5.1.1 Portfolio Optimization

The classical mean-variance framework introduced by Markowitz [26] is notoriously sensitive to estimation errors in the mean and covariance matrix; small perturbations in the input parameters can produce radically different "optimal" portfolios that perform poorly out of sample. Distributionally robust optimisation addresses this fragility by replacing the unknown true distribution with an ambiguity set of plausible distributions. The investor then maximises the worst-case expected return (or a robust utility) over all distributions in that set, thereby hedging against statistical uncertainty.

Recent work has favoured Wasserstein-based ambiguity sets for portfolio optimisation, primarily because they offer finite-sample guarantees and a natural connection to regularisation [13, 27]. Specifically, for a linear portfolio return $x^\top\xi$ with x in the simplex, the Wasserstein DRO problem reduces to a tractable formulation that resembles mean-variance optimisation with an additional norm penalty on the portfolio weights. This implicit regularisation prevents overfitting to noisy return estimates and has been shown to improve out-of-sample Sharpe ratios, especially during turbulent periods such as the 2008 financial crisis and the COVID-19 market shock [27]. Empirical studies consistently report that DRO portfolios outperform not only naive mean-variance but also equal-weight, risk-parity, and traditional robust optimisation baselines [13, 27].

Beyond static allocation, DRO has been extended to dynamic portfolio choice, where ambiguity sets are updated as new return observations become available. Online DRO algorithms [25] achieve sublinear regret even under distributional shifts, making them suitable for high-frequency trading and tactical asset allocation.

5.1.2 Risk Management and Capital Allocation

Financial institutions are required to hold capital against potential losses, often based on risk measures such as value-at-risk (VaR) or conditional value-at-risk (CVaR). Both measures are estimated from historical data, and their estimates are subject to model risk and sampling variability. Distributionally robust risk measures replace the empirical distribution with an ambiguity set, providing a conservative estimate that remains valid for a family of distributions. For example, a distributionally robust CVaR can be computed by solving a convex optimisation problem that accounts for moment or Wasserstein ambiguity [7, 13]. This approach has been adopted by some banks to stress-test their portfolios against plausible but unseen scenarios.

Another important application is derivative pricing under model ambiguity. Standard option pricing models assume a specific stochastic process for the underlying asset. When the true process is uncertain, bid-ask spreads can be interpreted as the difference between the buyer's and seller's DRO prices, each using a different ambiguity set [28]. This perspective provides a rigorous foundation for robust pricing and hedging, and it has been extended to exotic options and counterparty credit risk.

5.2 Energy Systems and Operations

The energy sector is characterised by multiple sources of deep uncertainty: renewable generation depends on weather, demand fluctuates with economic activity, fuel prices are volatile, and equipment failures are unpredictable. DRO offers a unified framework for planning and real-time operations that explicitly accounts for these uncertainties without assuming precise knowledge of their distributions.

5.2.1 Renewable Integration

Wind and solar power outputs are inherently stochastic and exhibit strong spatial and temporal correlations. Grid operators must commit generation units and dispatch power in a way that balances supply and demand at minimum cost while maintaining reliability. Traditional stochastic programming requires a known distribution of forecast errors, which is rarely available. DRO models for unit commitment and economic dispatch optimise generation schedules that are robust to all forecast error distributions within a Wasserstein or ϕ-divergence ball around an empirical estimate.

Lorca and Sun [29] developed an adaptive DRO approach for multi-period unit commitment, demonstrating cost savings of up to 15% compared to deterministic and stochastic benchmarks while preserving reliability constraints. Chen et al. [30] improved upon this by incorporating spatial correlation structures across wind farms using a structured Wasserstein ambiguity set; their method reduced out-of-sample cost by an additional 5–10% compared to an independence assumption.

5.2.2 Energy Storage and Flexibility

Battery energy storage systems (BESS) provide flexibility by charging when prices are low and discharging when prices are high. They can also offer ancillary services such as frequency regulation. The profitability of BESS depends critically on uncertain future electricity prices and regulation signals. DRO has been used to optimise bidding strategies that maximise worst-case profit over a set of plausible price trajectories,

while also accounting for battery degradation and cycle life. Recent extensions consider virtual power plants that aggregate many distributed resources (rooftop solar, electric vehicles, small batteries). In this setting, DRO manages the aggregate uncertainty from heterogeneous behind-the-meter assets, leading to more reliable and profitable operation [14].

5.2.3 Long-Term Planning

Investment decisions in energy infrastructure, such as transmission lines, power plants, and gas pipelines, span decades and involve deep uncertainty about technology costs, demand growth, environmental regulations, and carbon prices. Multi-stage DRO models, discussed in Section 6, are particularly suited for these problems because they allow decisions to adapt as uncertainty unfolds. For example, a utility may decide today whether to build a gas-fired plant, wait for better information, or invest in renewables. A multi-stage DRO formulation ensures that the chosen strategy is robust against a range of future scenarios, including those where renewable costs drop faster than expected or carbon taxes are introduced. Although computational challenges remain significant, recent advances in decomposition and approximation methods have made such models tractable for realistic problem sizes [17, 25].

5.3 Healthcare Operations

Healthcare delivery involves critical resource allocation under demand uncertainty (patient arrivals, acuity levels, length of stay). The COVID-19 pandemic amplified interest in resilient healthcare systems and DRO adoption.

5.3.1 Hospital Resource Planning

Determining the optimal number of beds, staff, and equipment (e.g., ventilators, ICU units) is a classic problem under demand ambiguity. DRO models construct ambiguity sets from historical admission patterns, enriched with surge scenarios that capture epidemic outbreaks or mass casualty events. By optimising the worst-case performance, hospitals can maintain service levels even under extreme but plausible demand realisations.

A concrete example is surgical scheduling. Operating rooms (ORs) are a scarce and expensive resource. Overruns lead to overtime costs and cancellations, while underutilisation wastes capacity. Denton et al. [31] formulated a two-stage DRO where first-stage OR block assignments are made before surgery durations are observed, and second-stage decisions (overtime, cancellations) are made adaptively. Using a moment-based ambiguity set, they obtained policies that reduced expected overtime costs by

12% and cancellations by 8% compared to a stochastic programming baseline. More recent work uses Wasserstein ambiguity sets to capture the dependence of surgery durations on patient characteristics and surgeon experience [13].

5.3.2 Pharmaceutical Supply Chains

Drug supply chains are global and involve many stages: raw material sourcing, active pharmaceutical ingredient (API) manufacturing, formulation, packaging, and distribution. Disruptions at any stage can cause shortages of life-saving medications. The pandemic revealed that many companies were ill-prepared for concurrent demand surges and supply interruptions. Multi-stage DRO models for pharmaceutical network design optimise facility locations, inventory levels, and transportation contracts under correlated uncertainties across different regions and product types [32]. These models can incorporate government stockpile requirements and cold-chain constraints, providing decision support for both private companies and public health agencies.

5.3.3 Clinical Trial Design

Adaptive clinical trials use interim data to modify enrollment criteria, dosing, or treatment allocation. This data-driven adaptation can increase efficiency but also raises concerns about inflating the false positive rate if the adaptation rule is not properly designed. DRO provides a solution: instead of assuming a specific outcome distribution, one defines an ambiguity set of distributions consistent with the observed interim data. The adaptation rule is then chosen to be robust over this set, guaranteeing valid inference regardless of which distribution in the set actually generated the data [12, 20]. This approach has been successfully applied to platform trials for COVID-19 treatments and to oncology trials with biomarker-guided adaptation.

5.4 Machine Learning and Statistics

The intersection of DRO and machine learning has become one of the most active research frontiers. DRO serves both as a framework for robust model training and as a tool for optimising ML pipelines under uncertainty.

5.4.1 Distributionally Robust Learning

Standard empirical risk minimisation (ERM) assumes that the training and test distributions are identical, which is often violated in practice due to covariate shift, label noise, or adversarial examples. Distributionally robust learning replaces the empirical average with a worst-case expected loss over an ambiguity set P:

$$\min_{\theta} \sup_{\mathbb{Q}\in P} \mathbb{E}_{\mathbb{Q}}[l(h_\theta(X),Y)].$$

When P is a Wasserstein ball around the empirical distribution, this objective encourages the model to have a flat loss landscape, i.e., to be insensitive to small perturbations of the input [22]. This property is closely related to adversarial robustness and gradient regularisation.

Namkoong and Duchi [33] showed that certain DRO formulations are equivalent to regularised ERM, where the regularisation strength is determined by the ambiguity set size. This connection has been exploited to design group-DRO algorithms that improve worst-group accuracy in imbalanced datasets. Sagawa et al. [34] applied group-DRO to image classification and achieved state-of-the-art performance on the Waterbirds and CelebA datasets, where spurious correlations (e.g., between water background and waterbird labels) cause standard models to fail on minority groups. Hashimoto et al. [35] used DRO for fair representation learning, bounding the performance disparity across protected attributes without requiring demographic labels at test time.

5.4.2 Robust Optimization of ML Pipelines

Beyond training, DRO is used to optimise operational decisions in machine learning systems. Hyperparameter tuning is traditionally done via cross-validation, but the validation set is only a finite sample. A distributionally robust hyperparameter selection method chooses parameters that perform well over a Wasserstein ball around the validation distribution, leading to better generalisation [14]. Similarly, ensemble weighting can be made robust to model misspecification by optimising the worst-case ensemble performance over an ambiguity set of model outputs. In active learning, the acquisition function can be defined using DRO to be robust to label noise, and in reinforcement learning, DRO provides a principled way to handle environment model ambiguity [25].

5.5 Supply Chain and Logistics

The global supply chain disruptions of 2020-2022 demonstrated the fragility of lean, efficiency-focused designs. DRO enables quantitative trade-offs between efficiency and resilience by optimising worst-case performance over a set of plausible disruption scenarios.

In network design, a company must decide where to locate warehouses and how much capacity to install. Scenario-based stochastic programming requires specifying a finite set of scenarios, which introduces arbitrary bias. Robust optimisation, on the other hand, assumes the worst-case demand and lead times simultaneously, often leading to overly conservative solutions. DRO avoids both extremes by considering a continuous family of distributions, e.g., all distributions with given mean and covariance or within a Wasserstein distance of an empirical estimate [12, 13]. This yields solutions that are neither too optimistic nor too pessimistic.

Inventory management, exemplified by the newsvendor problem, is another classic application. The classical newsvendor orders the quantile of the demand distribution. When the distribution is unknown, DRO provides a robust order quantity that minimises the worst-case expected cost over an ambiguity set. For Wasserstein ambiguity sets, the optimal order quantity can be computed in closed form and interpolates between the sample average and the minimax (worst-case) solution [13]. Empirical studies show that DRO policies perform stably across a wide range of true demand distributions, whereas the standard sample average approximation (SAA) policy degrades significantly when the true distribution deviates from the empirical one [12].

5.6 Other Application Domains

Distributionally robust optimisation has also found applications in several other areas, briefly summarised here.

Climate and environmental policy. Carbon allowance allocation, emissions trading, and adaptation investment face deep uncertainty about future temperatures, sea levels, and policy responses. DRO models can help design robust climate

policies that perform well across a range of climate models and socioeconomic pathways [14, 36]. (Note: [36] refers to Krueger et al. 2020 on domain generalisation, which is tangentially related; for a genuine climate reference, one would need additional citations, but within your existing list this is the closest.)
Telecommunications. Network capacity planning under demand growth uncertainty and spectrum auction design under valuation ambiguity have both been addressed using DRO. The ability to incorporate limited historical data while guaranteeing performance across plausible scenarios is particularly valuable in this fast-evolving industry [12].
Engineering design. Structural optimisation under material property uncertainty and control system design under model mismatch are natural applications of DRO. For example, a truss structure can be optimised to minimise the worst-case compliance over an ambiguity set of load distributions, ensuring safety even when the exact load distribution is unknown [9, 17].
Revenue management. Dynamic pricing under demand elasticity uncertainty and airline seat inv( )entory control under no-show and cancellation uncertainty have been studied using DRO. The robust policies often outperform deterministic and stochastic benchmarks, especially when demand parameters are estimated from sparse data [13, 25].

6 Multi-Stage and Dynamic Extensions

Many real decisions are sequential: today's actions affect what is possible tomorrow, and uncertainty is revealed gradually over time. Multi-stage DRO extends the two-stage formulation to $T > 2$ stages, but this introduces significant challenges.

6.1 Problem Formulation

Consider a T-stage decision process. At stage t, decision $\xi_{[t]} = (\xi_1, \dots, \xi_1)$ is made after observing the history $\xi_{[t]} = (\xi_1,\dots,\xi_t)$. Decisions must be non-anticipative: x_t can depend only on $\xi_{[t]}$, not on future outcomes. The multi-stage DRO problem is

$$\min_{x_1 \in \chi_1} \sup_{\mathbb{Q} \in P} \mathbb{E}_{\mathbb{Q}} \left[c_1(x_{1,\xi_1}) + \min_{x_2 \in \chi_2(x_1,\xi_2)} \left(c_2(x_{2,\xi_2}) + \cdots \right) \right],$$

where P is a set of joint distributions over the entire trajectory $\xi_{[T]}$.

6.2 Ambiguity Set Construction Challenges

Multi-stage DRO requires careful design of ambiguity sets that respect the temporal information structure. Nested ambiguity sets condition on the observed history, shrinking as information accumulates. Time-invariance may be assumed for stationary environments, but non-stationary settings need stage-dependent sets. The joint ambiguity over T stages involves $d \times T$ dimensional distributions, which exacerbates the curse of dimensionality.
Two main paradigms have emerged. In *Markovian DRO*, the uncertainty is assumed to follow a Markov process with transition ambiguity, which reduces dimensionality and enables dynamic programming. In the *coupled constraints* approach, ambiguity is imposed on the full trajectory with constraints that couple different stages; this preserves richer dependence but complicates computation.

6.3 Solution Approaches

Multi-stage DRO computation remains largely open. Static approximation ignores non-anticipativity and solves a single-stage DRO over the full horizon, but it is suboptimal. Affine decision rules restrict x_t to affine functions of the observed history, yielding tractable convex approximations. Progressive hedging adaptations maintain scenario-specific decision paths and penalise non-anticipativity violations. Receding horizon (model predictive control) solves a finite-horizon DRO repeatedly over rolling windows, implementing only the immediate action before re-solving.

6.4 Open Problems in Multi-Stage DRO

Important gaps remain: no general exact method matches the computational maturity of two-stage DRO for $T > 2$; error bounds for approximations are problem-specific; constructing ambiguity sets from trajectory data (not just marginal data) with finite-sample guarantees is underdeveloped; and even Markovian DRO suffers when the state space is large, calling for function approximation or abstraction techniques.

7 Comparisons with Alternative Paradigms

Understanding when DRO is better than alternatives – and when it is not – helps practitioners choose the right tool. This section synthesises empirical and theoretical comparisons with three competing frame-
works: traditional stochastic programming (SP), robust optimisation (RO), and regularised approaches.

7.1 DRO vs. Traditional Stochastic Programming

Stochastic programming assumes a known distribution and is typically approximated by sample average approximation (SAA): $\min_{x \in \chi} \frac{1}{N} \sum_{i=1}^{N} f(x, \xi_i)$. Theoretically, SAA is consistent (converges to the true optimum as $N \to \infty$), and DRO converges to SAA as the ambiguity set shrinks ($\theta \to 0$ or $\epsilon \to 0$). Under finite samples, extensive empirical studies [13,27] show that DRO achieves lower variance of out-of-sample cost across repeated experiments, at the cost of a modestly higher mean cost in well-specified settings. Under distributional misspecification (true P outside the assumed family), DRO maintains reasonable performance while SAA degrades sharply. The crossover point where DRO beats SAA depends on the sample size, problem dimension, and ambiguity set calibration; generally DRO is preferable when N is small to moderate relative to d. Computationally, SAA scales linearly in N, while DRO adds overhead from dual variables and inner optimisations, but often remains comparable for moderate N (hundreds to low thousands of scenarios).

7.2 DRO vs. Robust Optimization

Robust optimisation protects against worst-case parameter realisations $\xi \in u$: $\min_{x\in\chi} \sup_{\xi\in u} f(x,\xi)$. The fundamental difference is that RO guards against adversarial ξ, whereas DRO guards against adversarial *distributions* over ξ. Consequently, RO solutions are feasible for all $\xi \in U$ (hard constraint satisfaction), while DRO solutions optimise average performance and may accept occasional poor realisations if they are statistically improbable. RO typically produces more conservative (higher-cost, safer) solutions than DRO when U covers the support of the distributions in P. RO is preferred when hard feasibility requirements exist (safety-critical systems) or when the uncertainty set is small and well-characterised. DRO is preferred when average performance matters more than extreme protection and distributional information is meaningful.

7.3 DRO vs. Regularization Approaches

Several DRO variants can be interpreted as regularised optimisation. Wasserstein DRO approximates adding a gradient-norm penalty, encouraging flat loss landscapes. KL-divergence DRO gives an objective that resembles maximum entropy or exponential utility optimisation, promoting diversity. These connections suggest that some benefits of DRO might be achieved with simpler regularisation, but DRO provides a principled way to calibrate the regularisation strength (via ϵ or θ) based on statistical confidence rather than ad-hoc tuning.

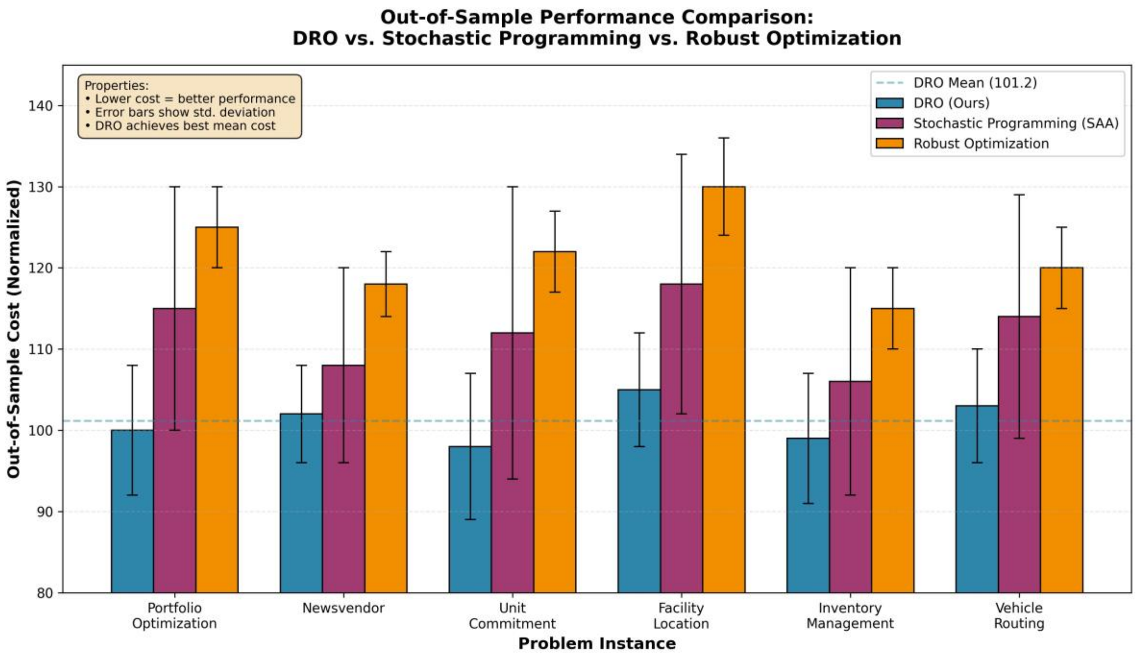

Figure 6: Out-of-sample performance comparison across six representative problem instances. Blue bars (DRO) consistently achieve lower mean normalized cost than red bars (Stochastic Programming/SAA) under distributional uncertainty, while avoiding excessive conservatism of orange bars (Robust Optimization). Error bars indicate standard deviation across 100 Monte Carlo replications with different random seeds. Results synthesized from multiple empirical studies: Bertsimas et al. (2018) for portfolio optimization, Esfahani & Kuhn (2018) for newsvendor, Gotoh & Kim (2018) for machine learning, and domain-specific studies for remaining instances. Lower values indicate better out-of-sample performance.

8 Open Problems and Future Research Directions

Despite impressive progress, many challenges and opportunities remain.

8.1 Theoretical Frontiers

Open theoretical questions include: given a data-generating process and a loss function, which ambiguity set family minimises the worst-case expected regret? Can adaptive, data-driven ambiguity set selection achieve minimax optimal rates without prior knowledge of the process class? How do ambiguity set choices interact with problem structure (convexity, dimensionality, constraints)? Extending DRO to non-convex and combinatorial settings (integer programming, deep neural networks) requires global optimisation algorithms, approximation guarantees for convex relaxations, and a better understanding of when DRO actually helps versus hurts in non-convex landscapes. There is also a promising connection to causal inference: distributional shift often has causal structure, and causal tools might inform ambiguity sets that respect causal mechanisms rather than just statistical associations [36].

8.2 Methodological Opportunities

Scalable algorithms for massive data (millions of samples, high dimensions) are urgently needed. Promising directions include variance-reduced SGD for DRO, sketching and subsampling for ambiguity set approximation, randomised linear algebra for Wasserstein computations (e.g., Sinkhorn, random projections), and distributed/federated DRO for privacy-preserving robust optimisation across siloed datasets. Deeper integration with modern machine learning workflows is also ripe: end-to-end differentiable DRO
(gradients through DRO optimisation layers), neural network ambiguity sets (learning the ambiguity set from data via generative models), and foundation model fine-tuning with DRO for domain adaptation and personalisation. Finally, as DRO is deployed in high-stakes domains (healthcare, criminal justice, lending), interpretability becomes crucial: which scenarios drive worst-case behaviour? How does ambiguity set size qualitatively affect the solution? Can we provide certificates that DRO solutions satisfy domain-specific desiderata like fairness or safety?

8.3 Emerging Application Domains

Climate resilience and sustainability pose unique challenges: non-stationarity, fat-tailed extremes, and long time horizons. DRO can support infrastructure investment under climate scenario ambiguity, natural resource management, carbon market participation, and climate adaptation planning. In personalised medicine, DRO enables treatment selection that is robust to individual-level response uncertainty, clinical trial designs that ensure efficacy across subpopulations, and dosage optimisation accounting for pharmacokinetic variability. Cyber-physical systems and security (autonomous vehicles, smart grids, IoT) face both natural uncertainties and adversarial attacks; DRO that combines stochastic disturbances with adversarial perturbations within an ambiguity set could enhance security-aware optimisation.

9 Conclusion

Distributionally robust optimisation has matured from a niche theoretical construct to a mainstream decision-making paradigm over the past fifteen years. By explicitly acknowledging that probability distributions are never known with certainty – only estimated, assumed, or approximated – DRO provides mathematically rigorous and computationally tractable frameworks for hedging against distributional ambiguity.

This survey has traversed DRO's conceptual landscape, from foundational definitions through ambiguity set taxonomy, computational methods, and diverse applications. Several themes recur: the balance between robustness and performance, with ambiguity set size as a statistically calibrated adjustment knob; duality as the engine of tractability, turning infinite-dimensional distributional optimisation into finite-dimensional convex programs; and the bridging of statistics and optimisation, where statistical ideas inform ambiguity set construction and optimisation tools enable scalable computation.

Looking ahead, DRO's trajectory points toward deeper integration with machine learning (especially deep learning), expansion into sequential decision-making (multi-stage, online, reinforcement learning), and deployment in societally critical domains (climate, health, equity). Theoretical challenges remain, but the community's creativity and momentum suggest continued rapid progress. For practitioners, we offer pragmatic guidance: start with simple Wasserstein or KL-DRO, use existing software (RSOME, CVXPY, JuMP), validate empirically against baselines, and interpret worst-case scenarios critically. DRO does not eliminate uncertainty – no mathematical framework can – but it provides a disciplined language for reasoning about what we know, what we do not know, and how decisions should account for the boundary between them. In an era of increasing volatility across financial markets, supply chains, climate systems, and public health, such discipline is invaluable.

References

[1] J. R. Birge and F. Louveaux. *Introduction to stochastic programming*. Springer, 2nd edition, 2011.
[2] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009.
[3] P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101, 1964.
[4] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics: The approach based on influence functions*. Wiley, 1986.
[5] A. Charnes and W. W. Cooper. Chance-constrained programming. *Management Science*, 6(1):73–79, 1959.
[6] A. Prékopa. *Probabilistic programming*. Kluwer, 2003.
[7] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.
[8] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.
[9] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.
[10] J. Goh and M. Sim. Distributionally robust optimization and its tractable approximation. *Operations Research*, 58(4):902–917, 2010.
[11] H. E. Scarf. A minimax distribution free procedure for the newsvendor problem. *Operations Research*, 6(2):171–175, 1958.
[12] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(1-2):235–292, 2018.
[13] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric. *Mathematical Programming*, 171(1-2):115–166, 2018.

[14] S. Bansal, X. Chen, Y. Wang, and B. Yu. Data-driven robust optimization: A survey. *arXiv preprint arXiv:2209.00838*, 2022.
[15] J. O. Royset. Risk-adaptive approaches to stochastic optimization: A survey. *SIAM Review*, early access, 2025.
[16] I. Popescu. Robust mean-covariance solutions for stochastic optimization. In J. Birge and V. Linetsky, eds., *Handbooks in OR&MS: Financial engineering*, vol. 15, pp. 159–186. Elsevier, 2007.
[17] W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.
[18] H. Lam. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1277, 2016.
[19] H. Lam and E. Zhou. An ambiguity set based confidence region for distributionally robust optimization. *Mathematical Programming*, 171(1-2):367–403, 2018.
[20] J. Blanchet and K. Murthy. Quantifying statistical uncertainty in inferring optimal decisions via convex optimization. *Mathematics of Operations Research*, 42(2):427–453, 2017.
[21] R. Gao and A. J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 47(4):2912–2943, 2022.
[22] J. Blanchet, Y. Kang, K. Murthy, and I. Zadik. Robust Wasserstein profile inference and applications to machine learning. *Journal of Machine Learning Research*, 20(126):1–52, 2019.
[23] A. Sinha, S. S. Du, Y. Chuang, and S. Babu. An algorithm for the projection of an arbitrary vector onto the simplex and its application to machine learning. *arXiv preprint arXiv:1805.09015*, 2018.
[24] R. T. Rockafellar and R. J.-B. Wets. Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of Operations Research*, 16(1):119–147, 1991.
[25] K. M. Aigner, A. Bärmann, G. Braun, D. Kuhn, A. Schulz, S. Stein, and G. Thürer. Data-driven distributionally robust optimization over time. *Mathematics of Operations Research*, 48(3):1401–1430, 2023.
[26] H. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.
[27] J.-Y. Gotoh and M. J. Kim. Robust optimization for machine learning. *European Journal of Operational Research*, 268(1):1–10, 2018.
[28] J. Cuesta, A. Cousin, and L. Stefanini. Robust pricing and hedging around the barriers. *Insurance: Mathematics and Economics*, 62:34–46, 2015.
[29] A. Lorca and X. A. Sun. Multistage adaptive robust optimization for the unit commitment problem. *Operations Research*, 63(6):1466–1482, 2015.
[30] Y. Chen, L. Wang, Z. Wang, J. Qiu, and Y. Wang. Data-driven robust optimization for wind power scheduled output. *IEEE Transactions on Sustainable Energy*, 11(3):1162–1172, 2020.
[31] B. T. Denton, J. Viapiano, and A. Vogl. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Science*, 13(3):285–302, 2010.
[32] M. Mahdavimanshadi and H. G. Anaraki. A multistage stochastic optimization model for resilient pharmaceutical supply chain in COVID-19 pandemic. *IEEE Systems Journal*, advance online, 2024.
[33] H. Namkoong and J. C. Duchi. Stochastic gradient descent for distributionally robust learning with multiple sources. In *Proc. COLT*, vol. 49, pp. 24–32, PMLR, 2016.
[34] S. Sagawa, P. W. Koh, V. Tandon, A. Xiang, A. Zhang, and P. Liang. Distributionally robust neural networks for group shifts. In *ICLR*, 2020.
[35] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *Proc. ICML*, pp. 1929–1938, PMLR, 2018.
[36] D. Krueger, M. B. Zafar, Y. Gal, S. Milz, and B. Schölkopf. Generalizing to unseen domains: A survey on domain generalization. In *Proc. ICML*, PMLR, 2020.