Tutorial: Robustness and Optimization

John Duchi

UAI 2020
Outline

Part I: (convex) optimization
1. Convex optimization
2. Formulation and “technology”

Part II: robust optimization
1. Formulation of robust optimization problems
2. Data uncertainty and construction

Part III: distributional robustness
1. Ambiguity and confidence
2. Uniform performance and sub-population robustness

Part IV: valid predictions
1. Conformal inference
2. Robustness to the future?
Basic optimization

minimize $f_0(x)$
subject to $f_i(x) \leq 0, \ i = 1, \ldots, m$

▶ $x \in \mathbb{R}^d$ is variable (or decision variable)
▶ $f_0 : \mathbb{R}^d \rightarrow \mathbb{R}$ is objective
▶ $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ are constraints

solution is $x^*$ minimizing $f_0$ subject to constraints
Applications and examples

**Operations research** (1940s on)
- Facility placement: choose location of facility minimize cost of transporting materials
- Portfolio optimization: minimize risk or variance subject to expected returns of investments

**Engineering and control** (1980s on)
- Control: minimize expended energy subject to moving from one location to another (variables are control inputs)
- Device design: (e.g.) minimize power consumption subject to manufacturing limits, timing requirements, size

**Statistics and machine learning** (1990s on)
- minimize prediction error or model mis-fit subject to prior information, sparsity, parameter limits
Convex optimization problems

minimize $f_0(x)$
subject to $f_i(x) = 0, \ i = 1, \ldots, m$

$h_i(x) = b_i, \ i = 1, \ldots, p$

- objective $f_0$ and inequality constraints $f_i$ are convex:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) \quad \text{for} \ 0 \leq \lambda \leq 1$$

- equalities $h_i$ are linear:

$$h_i(x) = a_i^T x$$

this is a technology
Linear programs

objective and constraints are linear

\[
\text{minimize } c^T x \\
\text{subject to } Ax \leq b, \quad Fx = g
\]
Quadratic programs

objective and inequality constraints are quadratic

\[
\text{minimize } x^T A x + b^T x \\
\text{subject to } x^T P_i x + q_i^T x + r_i \leq 0, \ i = 1, \ldots, m \\
Fx = g
\]
Semidefinite programs

variables are matrices $X \in S^n = \{X \in \mathbb{R}^{n \times n} \mid X = X^T\}$,
constraints are in semidefinite order

\[
\begin{align*}
\text{minimize} & \quad \text{tr}(CX) \\
\text{subject to} & \quad \text{tr}(A_iX) = b_i, \quad i = 1, \ldots, m \\
& \quad X \succeq 0
\end{align*}
\]
Example: matrix completion

- partially observed matrix $M \in \mathbb{R}^{m \times n}_{+}$ of movie ratings in locations $(i, j) \in \Omega$
- user $i$ represented by vector $u_i \in \mathbb{R}^r$, movie $j$ by $v_j$, and $M_{ij} = u_i^T v_j$

For $X = UV^T$, $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$,

$$\text{minimize } \text{rank}(X)$$
$$\text{subject to } X_{\Omega} = M_{\Omega}$$

has convex relaxation

$$\text{minimize } \sum_{i=1}^{n} \sigma_i(X) = \|X\|_*$$
$$\text{subject to } X_{\Omega} = M_{\Omega}$$
Nuclear norm minimization

\[
\begin{align*}
\text{minimize} & \quad \sum_{i=1}^{n} \sigma_i(X) = \|X\|_* \\
\text{subject to} & \quad X_\Omega = M_\Omega
\end{align*}
\]

has equivalent semidefinite program

\[
\begin{align*}
\text{minimize} & \quad \text{tr}(Z) + \text{tr}(W) \\
\text{subject to} & \quad X_\Omega = M_\Omega \quad \left[ \begin{array}{cc} Z & -X \\ -X^T & W \end{array} \right] \succeq 0, \quad Z \succeq 0, \quad W \succeq 0
\end{align*}
\]

in variables \( X \in \mathbb{R}^{m \times n}, \ Z \in \mathbb{S}^n, \ W \in \mathbb{S}^m \)
A few important calculus rules

Let \( f_1, f_2 : \mathbb{R}^d \to \mathbb{R} \) be convex functions

- \( f(x) = \alpha f_1(x) + \beta f_2(x) \) is convex for \( \alpha, \beta \geq 0 \)
- maxima of convex functions are convex:
  \[
  f(x) = \max\{f_1(x), f_2(x)\}
  \]
- even for an infinite index set \( \mathcal{A} \),
  \[
  f(x) = \sup_{\alpha \in \mathcal{A}} f_\alpha(x)
  \]
  is convex
A failure of linear programming

\[ c = \begin{bmatrix} 100 \\ 199.9 \\ -5500 \\ -6100 \end{bmatrix}, \quad A = \begin{bmatrix} -0.01 & -0.02 & 0.5 & 0.6 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 90 & 100 \\ 0 & 0 & 40 & 50 \end{bmatrix}, \quad \text{and} \quad b = \begin{bmatrix} 0 \\ 1000 \\ 2000 \\ 800 \\ 100000 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \]

\( c \) vector of costs/profits for two drugs, constraints \( Ax \leq b \) on production

▶ what happens if we vary percentages .01, .02 (chemical composition of raw materials) by .5% and 2%, i.e. .01 ± .00005 and .02 ± .0004?
Example failure for linear programming

Frequently lose 15–20% of profits
Example (Truss Design)

**Problem:** Choose thickness of bars to (1) minimize use of material and (2) support desired load.
Example (Truss Design)

**Problem:** Choose thickness of bars to (1) minimize use of material and (2) support desired load.
**Problem:** Choose thickness of bars to (1) minimize use of material and (2) support desired load

![Occasional load displacement diagram](image-url)
Outline

Part I: (convex) optimization
1 Convex optimization
2 Formulation and “technology”

Part II: robust optimization
1 Formulation of robust optimization problems
2 Data uncertainty and construction

Part III: distributional robustness
1 Ambiguity and confidence
2 Uniform performance and sub-population robustness

Part IV: valid predictions
1 Conformal inference
2 Robustness to the future?
Robust optimization

objective $f_0 : \mathbb{R}^n \to \mathbb{R}$, uncertainty set $\mathcal{U}$, $f_i : \mathbb{R}^n \times \mathcal{U} \to \mathbb{R}$, $f_i(x, u)$ convex in $x$ for all $u \in \mathcal{U}$

general form

minimize $f_0(x)$
subject to $f_i(x, u) \leq 0$ for all $u \in \mathcal{U}, i = 1, \ldots, m$.

equivalent to

minimize $f_0(x)$
subject to $\sup_{u \in \mathcal{U}} f_i(x, u) \leq 0, i = 1, \ldots, m$.

▶ Bertsimas, Ben-Tal, El-Ghaoui, Nemirovski (1990s–now)
Setting up robust problem

- can replace objective $f_0$ with $\sup_{u \in \mathcal{U}} f_0(x, u)$, rewrite as

  minimize $t$

  subject to $\sup_{u} f_0(x, u) \leq t$, $\sup_{u} f_i(x, u) \leq 0$, $i = 1, \ldots, m$

- equality constraints make no sense: a robust equality $a^T(x + u) = b$ for all $u \in \mathcal{U}$?

three questions:

- is robust formulation useful?
- is robust formulation computable?
- how should we choose $\mathcal{U}$?
A failure of linear programming

\[ c = \begin{bmatrix} 100 \\ 199.9 \\ -5500 \\ -6100 \end{bmatrix} \quad A = \begin{bmatrix} -.01 & -.02 & .5 & .6 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 90 & 100 \\ 0 & 0 & 40 & 50 \\ 100 & 199.9 & 700 & 800 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 0 \\ 1000 \\ 2000 \\ 800 \\ 100000 \end{bmatrix}. \]

\( c \) vector of costs/profits for two drugs, constraints \( Ax \leq b \) on production

- what happens if we vary percentages .01, .02 (chemical composition of raw materials) by .5% and 2%, i.e. .01 ± .00005 and .02 ± .0004?
Example failure for linear programming

Frequently lose 15–20% of profits
minimize $c^T x$

subject to $(A + \Delta)x \leq b$, all $\Delta \in \mathcal{U}$

where $|\Delta_{11}| \leq .00005$, $|\Delta_{12}| \leq .0004$, $\Delta_{ij} = 0$ otherwise

* solution $x_{\text{robust}}$ has degradation *provably* no worse than 6%
Example (Truss Design)

**Problem:** Choose thickness of bars to (1) minimize use of material and (2) support desired load
**Example (Truss Design)**

**Problem:** Choose thickness of bars to (1) minimize use of material and (2) support desired load.
**Example (Truss Design)**

**Problem:** Choose thickness of bars to (1) minimize use of material and (2) support desired load
Example (Truss Design)

**Problem:** Choose thickness of bars to (1) minimize use of material and (2) support desired load.
Example (Truss Design)

**Problem:** Choose thickness of bars to (1) minimize use of material and (2) support desired load
Example (Truss Design)

**Problem:** Choose thickness of bars to (1) minimize use of material and (2) support desired load.
How to choose uncertainty sets

- uncertainty set $\mathcal{U}$ a modeling choice
- common idea: let $U$ be random variable, want constraints that

$$\Pr(f_i(x, U) \geq 0) \leq \epsilon$$

(1)

- typically hard (non-convex except in special cases)
- find set $\mathcal{U}$ such that $\Pr(U \in \mathcal{U}) \geq 1 - \epsilon$, then sufficient condition for (1)

$$f_i(x, u) \leq 0 \text{ for all } u \in \mathcal{U}$$
Uncertainty set with Gaussian data

\[
\begin{align*}
\text{minimize} & \quad c^T x \\
\text{subject to} & \quad \Pr(a_i^T x > b_i) \leq \epsilon, \quad i = 1, \ldots, m
\end{align*}
\]

Coefficient vectors \(a_i\) i.i.d. \(\mathcal{N}(\bar{a}, \Sigma)\) and failure probability \(\epsilon\)

- marginally \(a_i^T x \sim \mathcal{N}(\bar{a}_i^T x, x^T \Sigma x)\)
- for \(\epsilon = .5\), just LP

\[
\begin{align*}
\text{minimize} & \quad c^T x \\
\text{subject to} & \quad a_i^T x \leq b_i, \quad i = 1, \ldots, m
\end{align*}
\]

- what about \(\epsilon = .1, .9\)?
Gaussian uncertainty sets

\[
\{ x \mid \Pr(a_i^T x > b_i) \leq \epsilon \} = \{ x \mid \bar{a}_i^T x - b_i - \Phi^{-1}(\epsilon)\sqrt{x^T \Sigma x} \leq 0 \}
\]

\(\epsilon = 0.9\) \hspace{2cm} \(\epsilon = 0.5\) \hspace{2cm} \(\epsilon = 0.1\)

(Source: ee364b, Stanford)
Robust problems are convex, so no problem?

not quite...

consider quadratic constraint

\[ \|Ax + Bu\|_2 \leq 1 \quad \text{for all} \quad \|u\|_\infty \leq 1 \]

- convex quadratic *maximization* in \(u\)
- solutions on extreme points \(u \in \{-1, 1\}^n\)
- and NP-hard to maximize (even approximately [Håstad])
  - convex quadratics over hypercube
Important question: when is a robust LP still an LP (robust SOCP an SOCP, robust SDP an SDP)

\[
\begin{align*}
\text{minimize} & \quad c^T x \\
\text{subject to} & \quad (A + U)x \preceq b \quad \text{for} \ U \in \mathcal{U}.
\end{align*}
\]

can always represent formulation constraint-wise, consider only one inequality

\[
(a + u)^T x \leq b \quad \text{for all} \ u \in \mathcal{U}.
\]

- Simple example: \( \mathcal{U} = \{ u \in \mathbb{R}^n \mid \|u\|_\infty \leq \delta \} \), then

\[
a^T x + \delta \|x\|_1 \leq b
\]
When are things tractable?

Duality typically used to get tractability
(but we’re not going to do that)
Portfolio optimization (with robust LPs)

- \( d \) assets \( i = 1, \ldots, d \), random multiplicative return \( R_i \) with
  \[ \mathbb{E}[R_i] = \mu_i \geq 1, \mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \]
- “certain” problem has solution \( x_{\text{nom}} = e_1 \),
  \[ \text{maximize } \mu^T x \text{ subject to } x^T 1 = 1, x \succeq 0 \]
- if asset \( i \) varies in range \( \mu_i \pm u_i \), robust problem
  \[ \text{maximize } \sum_{i=1}^d \inf_{u \in [-u_1, u_i]} (\mu_i + u)x_i \text{ subject to } 1^T x = 1, x \succeq 0 \]
  and equivalent
  \[ \text{maximize } \mu^T x - u^T x \text{ subject to } 1^T x = 1, x \succeq 0 \]
Portfolio optimization (tighter control)

- Returns $R_i \in [\mu_i - u_i, \mu_i + u_i]$ with $\mathbb{E}R_i = \mu_i$
- guarantee return with probability $1 - \epsilon$

\[
\text{maximize}_{x,t} t \\
\text{subject to } \Pr\left(\sum_{i=1}^{n} R_i x_i \geq t\right) \geq 1 - \epsilon, \quad x^T 1 = 1, \quad x \succeq 0
\]

- value at risk is non-convex in $x$, approximate it?
- approximate with high-probability bounds
- less conservative than LP (certain returns) approach
Portfolio optimization: probability approximation

- **Hoeffding’s inequality**

\[
\Pr\left(\sum_{i=1}^{n}(R_i - \mu_i)x_i \leq -t\right) \leq \exp\left(-\frac{t^2}{2 \sum_{i=1}^{n} x_i^2 u_i^2}\right).
\]

- **written differently**

\[
\Pr\left[\sum_{i=1}^{n} R_i x_i \leq \mu^T x - t\left(\sum_{i=1}^{n} u_i^2 x_i^2\right)^{\frac{1}{2}}\right] \leq \exp\left(-\frac{t^2}{2}\right)
\]

- **set** \( t = \sqrt{2 \log(1/\epsilon)} \), **gives robust problem**

\[
\text{maximize } \mu^T x - \sqrt{2 \log \frac{1}{\epsilon}} \|\text{diag}(u)x\|_2 \text{ subject to } 1^T x = 1, \ x \succeq 0.
\]
data $\mu_i = 1.05 + \frac{3(n-i)}{10n}$, uncertainty $|u_i| \leq u_i = .05 + \frac{n-i}{2n}$ and $u_n = 0$

nominal minimizer $x_{\text{nom}} = e_1$

conservative (LP) minimizer $x_{\text{con}} = e_n$ (guaranteed 5% return),

robust (SOCP) minimizer $x_\epsilon$ for value-at risk $\epsilon = 2 \times 10^{-4}$
Portfolio optimization comparison

Returns chosen randomly in $\mu_i \pm u_i$, 10,000 experiments
Tutorial: Robustness and Optimization

John Duchi

UAI 2020
Outline

Part I: (convex) optimization
1 Convex optimization
2 Formulation and “technology”

Part II: robust optimization
1 Formulation of robust optimization problems
2 Data uncertainty and construction

Part III: distributional robustness
1 Ambiguity and confidence
2 Uniform performance and sub-population robustness

Part IV: valid predictions
1 Conformal inference
2 Robustness to the future?
Stochastic optimization

Data $X$ and parameters $\theta$ to learn, with loss

$$
\ell(\theta, X)
$$

**Goal:** Minimize the population risk

$$
\text{minimize } L(\theta) := \mathbb{E}_{P_0}[\ell(\theta, X)] = \int \ell(\theta, x) dP_0(x)
$$

subject to $\theta \in \Theta$

given an i.i.d. sample $X_1, \ldots, X_n \overset{iid}{\sim} P_0$

**Empirical risk minimization:**

$$
\hat{\theta} = \arg\min_{\theta \in \Theta} \mathbb{E}_{\hat{P}_n}[\ell(\theta, X)] = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, X_i)
$$
Liking curly fries on Facebook reveals your high IQ

By PHILIPPA WARR
12 Mar 2013

What you Like on Facebook could reveal your race, age, IQ, sexuality and other personal data, even if you’ve set that information to “private”.

Unlikely to be robust to even small changes in the underlying data
Revisiting uncertainty sets

$$\begin{align*}
\text{minimize} & \quad f_0(x) \\
\text{subject to} & \quad f_i(x, u) \leq 0, \text{ all } u \in \mathcal{U}
\end{align*}$$

the basic idea so far:

- assume uncertainty variable $\mathcal{U}$, choose $\mathcal{U}$ so that
  \[
  \Pr(\mathcal{U} \in \mathcal{U}) \geq 1 - \epsilon
  \]

- use this $\mathcal{U}$ in problem above

When do we actually know $\Pr(\mathcal{U} \in \mathcal{U})$?
Idea: Replace distribution \( P_0 \) with “uncertainty” set \( \mathcal{P} \) of possible distributions around \( P_0 \)

\[
\text{minimize}_{\theta \in \Theta} \ L(\theta) = \mathbb{E}_{P_0}[\ell(\theta, X)]
\]

Big question: How do we choose the set \( \mathcal{P} \)?

(i) Hypothesis testing, covariance, and other moment constraints
(ii) Non-parametric approaches
Distributionally robust optimization

**Idea:** Replace distribution $P_0$ with “uncertainty” set $\mathcal{P}$ of possible distributions around $P_0$

$$\min_{\theta \in \Theta} L(\theta, \mathcal{P}) := \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\theta, X)]$$

**Big question:** How do we choose the set $\mathcal{P}$?

(i) Hypothesis testing, covariance, and other moment constraints

(ii) Non-parametric approaches
A hypothesis testing approach

basic idea in hypothesis testing: for data $X$ drawn from some distribution

- have null hypothesis $H_0 : X \sim P_0$
- have a statistic $T : \mathcal{X} \rightarrow \mathbb{R}$ of observations $X$
- for level $\alpha$, find threshold $\tau_\alpha$ such that

$$P_0(T(X) > \tau_\alpha(P_0)) \leq \alpha$$

- reject null $H_0$ if $T(X) \geq \tau_\alpha$

example

- null is $H_0 : X_i \overset{iid}{\sim} \mathcal{N}(0, 1), i = 1, \ldots, n, T(X_1^n) = |\overline{X}_n|$
- threshold $\tau_\alpha = z_{1-\alpha/2}$
Hypothesis testing/confidence set duality

consider a collection of distributions \( \mathcal{P} \) on space \( \mathcal{X} \)

- let \( T, \tau_{\alpha}(P) \) be a statistic with level \( \alpha \) for distributions \( P \in \mathcal{P} \)
- sample \( X \sim P \), observe \( t^{\text{obs}} = T(X) \)
- confidence set

\[
C(X) := \left\{ P \in \mathcal{P} \mid \Pr_P(T(X) \leq t^{\text{obs}}) > \alpha \right\}
\]

- then

\[
\Pr(P \in C(X)) \geq 1 - \alpha
\]

example

- normal family

\( \mathcal{P} = \{N(\theta, 1) \mid \theta \in \mathbb{R}\} \)

- confidence set (abusing notation) is means

\[
C(X_1^n) = [\bar{X}_n - z_{1-\alpha/2}, \bar{X}_n + z_{1-\alpha/2}]
\]
Asymptotic validity

We say a test is *asymptotically of level* $\alpha$ for $H_0 : X_i \overset{iid}{\sim} P$ if

$$\limsup_{n \to \infty} P(T(X^n_1) > \tau_\alpha(P)) \leq \alpha$$

- **asymptotic confidence sets:** for observations $t_{n}^{\text{obs}} = T(X^n_1),$

  $$C(X^n_1) := \left\{ P \in \mathcal{P} \mid \Pr_P(T(X^n_1) \leq t_{n}^{\text{obs}}) > \alpha \right\}$$

- **Then as** $n \to \infty$, get

  $$\liminf_{n \to \infty} \Pr(P \in C(X^n_1)) \geq 1 - \alpha$$
A distributionally robust formulation

**Steps:**
1. choose valid (maybe asymptotically) confidence set $C(X_1^n)$
2. take uncertainty set $\mathcal{P}_n := C(X_1^n)$
3. solve robust problem

$$\minimize_{\theta \in \Theta} L(\theta, \mathcal{P}_n)$$

**Theorem**
Let $L_n^* = \inf_{\theta \in \Theta} L(\theta, \mathcal{P}_n)$ and $\hat{\theta}_n \in \arg\min_{\theta \in \Theta} L(\theta, \mathcal{P}_n)$. Then

$$\limsup_{n \to \infty} \Pr(L(\hat{\theta}_n) \geq L_n^*) \leq \alpha.$$
Example: portfolio optimization

- random returns $R_i \in \mathbb{R}_+^d$ for $d$ assets, periods $i = 1, 2, \ldots$ (assumed i.i.d.), mean returns $\bar{r} = \mathbb{E}[R]$
- goal
  $$\text{maximize } \bar{r}^T \theta \text{ subject to } \theta \succeq 0, \quad 1^T \theta = 1$$
- central limit theorem:
  $$\bar{R}_n = \frac{1}{n} \sum_{i=1}^n R_i \quad \Sigma_n = \frac{1}{n} \sum_{i=1}^n (R_i - \bar{R}_n)(R_i - \bar{R}_n)^T$$
  have
  $$\sqrt{n} \Sigma_n^{-1/2}(\bar{R}_n - \bar{r}) \overset{d}{\to} \mathcal{N}(0, I)$$
- lots of distributional facts about $Z \sim \mathcal{N}(0, I)$ known
Example: portfolio optimization (continued)

▶ choose threshold $\tau_\alpha$ so that

$$\Pr(\|Z\|_2^2 \geq \tau_\alpha) \leq \alpha$$

▶ confidence set

$$\mathcal{P}_n := \left\{ \text{distributions } P \text{ with } \left\| \sqrt{n} \Sigma_n^{-1/2} (\overline{R}_n - \mathbb{E}_P[R]) \right\|_2^2 \leq \tau_\alpha \right\}$$

▶ optimization problem

$$\maximize_{\theta} \inf \left\{ r^T \theta \text{ s.t. } \| \Sigma_n^{-1/2} (\overline{R}_n - r) \|_2^2 \leq \tau_\alpha / n \right\}$$
Example behavior

(Delage and Ye, 2010)
Asymptotic risks

**Challenge:** often very computationally hard to use valid confidence sets (or risk is infinite)
Divergence-based uncertainty sets

The $f$-divergence between distributions $P$ and $Q$ is

$$D_f (P\|Q) := \int f \left( \frac{dP}{dQ} \right) dQ$$

where $f$ is some convex function with $f(1) = 0$. 
Divergence-based uncertainty sets

The $f$-divergence between distributions $P$ and $Q$ is

$$D_f (P\|Q) := \int f \left( \frac{dP}{dQ} \right) dQ$$

where $f$ is some convex function with $f(1) = 0$.

**Familiar examples:**

- $f(t) = -\log t$ gives $D_f (P\|Q) = D_{\text{kl}} (Q\|P)$
- $f(t) = t \log t$ gives $D_f (P\|Q) = D_{\text{kl}} (P\|Q)$
- $f(t) = \frac{1}{2} (t - 1)^2$ gives $D_{\chi^2} (P\|Q)$
- $f(t) = \frac{1}{2} (\sqrt{t} - 1)^2$ gives $d_{\text{Hel}}^2 (P, Q)$
Divergence-based uncertainty sets

The $f$-divergence between distributions $P$ and $Q$ is

$$D_f (P\|Q) := \int f \left( \frac{dP}{dQ} \right) dQ$$

where $f$ is some convex function with $f(1) = 0$. Use uncertainty region

$$\mathcal{P}_\rho := \{ P : D_f (P\|P_0) \leq \rho \}$$
Divergence-based uncertainty sets

The $f$-divergence between distributions $P$ and $Q$ is

$$D_f (P\|Q) := \int f \left( \frac{dP}{dQ} \right) dQ$$

where $f$ is some convex function with $f(1) = 0$.

Use uncertainty region

$$\mathcal{P}_\rho := \{ P : D_f (P\|P_0) \leq \rho \}$$
Idea: Instead of using empirical distribution $\hat{P}_n$ on sample $X_1, \ldots, X_n$, look at non-parametrically reweighted versions

$$\mathcal{P}_{n, \rho} := \left\{ P : D_f \left( P \parallel \hat{P}_n \right) \leq \frac{\rho}{n} \right\}$$

and minimize

$$L(\theta, \mathcal{P}_{n, \rho}) = \sup_{P \in \mathcal{P}_{n, \rho}} \mathbb{E}_P[\ell(\theta, X)] = \sup_{P \in \mathcal{P}_{n, \rho}} \sum_{i=1}^n p_i \ell(\theta, X_i)$$

$$= \inf_{\lambda \geq 0, \eta} \left\{ \mathbb{E}_{\hat{P}_n} \left[ \lambda f^*(\frac{\ell(\theta, X) - \eta}{\lambda}) \right] + \frac{\rho}{n} \lambda + \eta \right\}$$
Empirical likelihood (Owen 1990)

For data $Z_i \in \mathbb{R}^k$, define confidence ellipse

$$E_n(\rho) := \left\{ \sum_{i=1}^{n} p_i Z_i \mid \sum_{i=1}^{n} (np_i - 1)^2 \leq \rho \right\}$$

then independently of distribution on $Z \in \mathbb{R}^k$

$$\text{Pr}(\mathbb{E}[Z] \in E_n(\rho)) \rightarrow \text{Pr}(\chi_k^2 \leq \rho).$$
Empirical likelihood (Owen 1990)

For data $Z_i \in \mathbb{R}^k$, define confidence ellipse

$$E_n(\rho) := \left\{ \sum_{i=1}^n p_i Z_i \mid \sum_{i=1}^n (np_i - 1)^2 \leq \rho \right\}$$

then independently of distribution on $Z \in \mathbb{R}^k$

$$\Pr(\mathbb{E}[Z] \in E_n(\rho)) \rightarrow \Pr(\chi^2_k \leq \rho).$$
Empirical likelihood (Owen 1990)

For data \( Z_i \in \mathbb{R}^k \), define confidence ellipse

\[
E_n(\rho) := \left\{ \sum_{i=1}^{n} p_i Z_i \mid \sum_{i=1}^{n} (np_i - 1)^2 \leq \rho \right\}
\]

then independently of distribution on \( Z \in \mathbb{R}^k \)

\[
\Pr(\mathbb{E}[Z] \in E_n(\rho)) \to \Pr(\chi^2_k \leq \rho).
\]
Empirical likelihood (Owen 1990)

For data $Z_i \in \mathbb{R}^k$, define confidence ellipse

$$E_n(\rho) := \left\{ \sum_{i=1}^{n} p_i Z_i \mid \sum_{i=1}^{n} (np_i - 1)^2 \leq \rho \right\}$$

then independently of distribution on $Z \in \mathbb{R}^k$

$$\Pr(\mathbb{E}[Z] \in E_n(\rho)) \rightarrow \Pr(\chi^2_k \leq \rho).$$
Empirical likelihood (Owen 1990)

For data $Z_i \in \mathbb{R}^k$, define confidence ellipse

$$E_n(\rho) := \left\{ \sum_{i=1}^{n} p_i Z_i \mid \sum_{i=1}^{n} (np_i - 1)^2 \leq \rho \right\}$$

then independently of distribution on $Z \in \mathbb{R}^k$

$$\Pr(\mathbb{E}[Z] \in E_n(\rho)) \to \Pr(\chi^2_k \leq \rho).$$
On variance expansions

**Confidence ellipse for risk:** Robust risk is

\[
L(\theta, P_{n,\rho}) = \sup_p \left\{ \sum_{i=1}^n p_i \ell(\theta, X_i) \mid \sum_{i=1}^n \frac{1}{n} f \left( \frac{p_i}{1/n} \right) \leq \frac{\rho}{n} \right\}
\]

**Theorem (D., Glynn, Namkoong 20)**

Let \( f \) be convex with \( f'''(1) = 2 \). Then

\[
L(\theta, P_{n,\rho}) = \frac{1}{n} \sum_{i=1}^n \ell(\theta, X_i) + \sqrt{\frac{\rho}{n} \text{Var}_{\hat{P}_n}(\ell(\theta, X))} + O_P(n^{-1})
\]

uniformly in \( \theta \) in compact sets
Problem: Classify documents as a subset of the 4 categories:

\[ \{ \text{Corporate, Economics, Government, Markets} \} \]

- Data: pairs \( x \in \mathbb{R}^d \) represents document, \( y \in \{-1, 1\}^4 \) where \( y_j = 1 \) indicating \( x \) belongs \( j \)-th category.
- Logistic loss, with \( \Theta = \{ \theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1000 \} \)
- \( d = 47,236 \), \( n = 804,414 \). 10-fold cross-validation.
- Use precision and recall to evaluate performance

\[
\text{Precision} = \frac{\# \text{ Correct}}{\# \text{ Guessed Positive}} \quad \text{Recall} = \frac{\# \text{ Correct}}{\# \text{ Actually Positive}}
\]
### Experiment: Reuters Corpus (multi-label)

#### Table: Reuters Number of Examples

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corporate</td>
<td>381,327</td>
</tr>
<tr>
<td>Economics</td>
<td>119,920</td>
</tr>
<tr>
<td>Government</td>
<td>239,267</td>
</tr>
<tr>
<td>Markets</td>
<td>204,820</td>
</tr>
</tbody>
</table>
Experiment: Reuters Corpus (multi-label)

Figure: Recall on common category (Corporate)
Experiment: Reuters Corpus (multi-label)

Figure: Recall on rare category (Economics)
Do well **almost all** the time instead of just on average.
Moving beyond “certificates”

**New challenge:** doing well on sub-populations within data

- ML models increasingly used in high-stakes decisions
- Disease diagnosis, hiring decisions, driving vehicles
- Models often underperform on minority, other subpopulations
  - As of 2015, only 1.9 percent of all studies of respiratory disease included minority subjects despite African Americans more likely to suffer respiratory ailments
  - Only 2 percent of more than 10,000 cancer clinical trials funded by the National Cancer Institute focused on a racial or ethnic minority
Approaches: group-based or pure robustness

Given groups $g \in \mathcal{G}$ with populations $P_g$, minimize

$$\max_{g \in \mathcal{G}} \mathbb{E}_{P_g}[\ell(\theta; X)]$$

[Meinshausen & Bühlmann 15; Kearns et al. 19; Sagawa, Koh et al. 19–20]

- requires pre-defined groups
- may be computationally challenging (if large numbers of potentially intersecting groups)

**alternative idea:** pick worst-performing sub-population, optimize that
for random variable $Z \in \mathbb{R}$, $Z \sim P_0$, and $q_{1-\alpha}(Z) = 1 - \alpha$ quantile of $Z$,

\[
\text{CVaR}_\alpha(Z) = \mathbb{E}[Z \mid Z \geq q_{1-\alpha}(Z)]
\]

\[
= \inf_{\eta} \left\{ \alpha^{-1} \mathbb{E}[(Z - \eta)_+] + \eta \right\}
\]

\[
= \sup \left\{ \mathbb{E}_P[Z] \mid \frac{p(z)}{p_0(z)} \leq \frac{1}{\alpha} \right\}
\]

\[
= \sup \left\{ \mathbb{E}_P[Z] \mid \text{there exists } Q, \beta \leq \alpha \text{ s.t. } P_0 = \beta P + (1 - \beta)Q \right\}
\]

intuition: choose worst sub-population of size at least $\alpha$
Generalized conditional value at risk

Theorem (Kusuoka)

For any collection $\mathcal{P}$ of distributions, there is a collection of distributions $\mathcal{M}$ on $[0, 1]$ such that

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[Z] = \sup_{\mu \in \mathcal{M}} \int_0^1 \text{CVaR}_\alpha(Z) \mu(d\alpha).$$

Interpretation: all distributionally robust formulations are mixtures of conditional value at risk
Robustness sets from $f$-divergences

**Proposition (D. & Namkoong 20)**

*For any $f$ of the form $f(t) = t^k - 1$, we have*

$$
\sup_{P : D_f(P \| P_0) \leq \rho} \mathbb{E}_P[Z] = \inf_{\eta} \left\{ (1 + c(\rho)) \mathbb{E} \left[ [Z - \eta]_+^{k_*} \right]^{1/k_*} + \eta \right\}
$$

*where $k_* = \frac{k}{k-1}$*

Consider minimizing robust losses of the form

$$
L(\theta, \{P : D_f(P \| P_0) \leq \rho \}) = \sup_{P : D_f(P \| P_0) \leq \rho} \mathbb{E}_P[\ell(\theta; X)]
$$
Typical results (MNIST classification experiment)

- have dataset of MNIST handwritten digits (60,000 images of digits 0–9)
- smaller dataset of typewritten digits
- training data is mixture of MNIST and typewritten digits
Error on MNIST handwritten digits
Error on all typewritten digits

![Graph showing the error on typewritten digits for different values of ρ.](image)
Error on easy typewritten digit (3)
Error on hard typewritten digit (9)
A few parting thoughts

- Have not talked about statistical consequences
- Still sometimes challenging to solve these at scale
- Hybrids between knowing groups and not knowing groups
- Connections with causality?
Tutorial: Robustness and Optimization

John Duchi

UAI 2020
Outline

Part I: (convex) optimization
1 Convex optimization
2 Formulation and “technology”

Part II: robust optimization
1 Formulation of robust optimization problems
2 Data uncertainty and construction

Part III: distributional robustness
1 Ambiguity and confidence
2 Uniform performance and sub-population robustness

Part IV: valid predictions
1 Conformal inference
2 Robustness to the future?
The actual robustness challenge

Robustness to future data
CIFAR Generalization

are an effective way to improve image classification models. Adaptivity is therefore an unlikely explanation for the accuracy drops. Instead, we propose an alternative explanation based on the relative difficulty of the original and new test sets. We demonstrate that it is possible to recover the original ImageNet accuracies almost exactly if we only include the easiest images from our candidate pool. This suggests that the accuracy scores of even the best image classifiers are still highly sensitive to minutiae of the data cleaning process. This brittleness puts claims about human-level performance into context [20, 31, 48]. It also shows that current classifiers still do not generalize reliably even in the benign environment of a carefully controlled reproducibility experiment.

Figure 1 shows the main result of our experiment. Before we describe our methodology in Section 3, the next section provides relevant background. To enable future research, we release both our new test sets and the corresponding code.

Potential Causes of Accuracy Drops

We adopt the standard classification setup and posit the existence of a “true” underlying data distribution $D$ over labeled examples $(x, y)$. The overall goal in classification is to find a model $\hat{f}$

1. https://github.com/modestyachts/CIFAR-10

(Recht, Roelofs, Schmidt, Shankar 2019)
ImageNet Generalization

are an effective way to improve image classification models. Adaptivity is therefore an unlikely explanation for the accuracy drops. Instead, we propose an alternative explanation based on the relative difficulty of the original and new test sets. We demonstrate that it is possible to recover the original ImageNet accuracies almost exactly if we only include the easiest images from our candidate pool. This suggests that the accuracy scores of even the best image classifiers are still highly sensitive to minutiae of the data cleaning process. This brittleness puts claims about human-level performance into context (Recht, Roelofs, Schmidt, Shankar 2019). It also shows that current classifiers still do not generalize reliably even in the benign environment of a carefully controlled reproducibility experiment.

Figure 1 shows the main result of our experiment. Before we describe our methodology in Section 3, the next section provides relevant background. To enable future research, we release both our new test sets and the corresponding code.

(Recht, Roelofs, Schmidt, Shankar 2019)
An alternative idea

let’s build valid confidence into systems

**Goal:** get confidence regions \( C(x) \) such that for given level \( \alpha \)

\[
\Pr(Y \in C(X)) \geq 1 - \alpha
\]

Conformal inference (Vovk and colleagues): we can do this for *any* model
Scoring functions

- Prediction or score $s(x, y)$

- Confidence sets of the form

$$C(x) = \{ y \mid s(x, y) \leq \tau \}$$
Split conformal inference

Define scores $S_i = s(X_i, Y_i)$, $i = 1, \ldots, n$, and threshold

$$
\tau_n := \frac{n + 1}{n} (1 - \alpha)\text{-quantile of } \{S_1, \ldots, S_n\}
$$

and confidence set

$$
C(x) := \{y \mid s(x, y) \leq \tau_n\}
$$

Theorem

*If data are i.i.d., then*

$$
\Pr(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha.
$$
Is this enough?
Banking data
California housing data
Delta ailerons data
Kinematics data
Puma data
Problem: Find confidence sets $C(x)$ such that if $s(X_{n+1}, Y_{n+1}) \sim P$ and $s(X_i, Y_i) \overset{iid}{\sim} P_0$ where

$$D_f (P \| P_0) \leq \rho$$

then

$$P(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$$
Robust quantiles and validity under shift

Define

\[ g_{f, \rho}(\beta) := \inf \left\{ z \in [0, 1] : \beta f \left( \frac{z}{\beta} \right) + (1 - \beta) f \left( \frac{1 - z}{1 - \beta} \right) \leq \rho \right\} \]

\[ g_{f, \rho}^{-1}(\tau) = \sup \left\{ \beta \in [\tau, 1] : \beta f \left( \frac{\tau}{\beta} \right) + (1 - \beta) f \left( \frac{1 - \tau}{1 - \beta} \right) \leq \rho \right\} \]

Proposition

We have

\[ \sup_{P : D_{f}(P \| P_0) \leq \rho} \text{Quantile}(\alpha, P) = \text{Quantile}(g_{f, \rho}^{-1}(\alpha), P) \]
A coverage guarantee

Define

\[ C_\rho(x) := \left\{ y \mid s(x, y) \leq \text{Quantile}(g_{f,\rho}^{-1}(1 - \alpha), \hat{P}_n) \right\} \]

**Theorem**

If \( s(X_i, Y_i) \sim P_0 \) for \( i = 1, \ldots, n \), and \( s(X_{n+1}, Y_{n+1}) \sim P \), then for \( \rho \geq D_f(P\|P_0) \)

\[ \Pr(Y_{n+1} \in C_\rho(X_{n+1})) \geq 1 - \alpha - \frac{O(1)}{n}. \]
One experimental result

The diagram illustrates the coverage of different Chi-squared tests: sampling, regression, and classification. The coverage values range from 0.80 to 1.00. The standard method shows the lowest coverage, while the Chi-squared methods for regression and classification have higher coverage, with the regression method having the highest among the three.
A few parting thoughts