

Decomposition Methods For Solving Distributionally Robust Programs

McCormick

Northwestern Engineering

Sanjay Mehrotra, Professor, Northwestern University

Part I: Stochastic Binary Programs

- *Finite support with Wasserstein & Moment Polytopes*

Joint work with Manish Bansal and Kuo-Ling Huang

Part I: Wasserstein RO – Logistic Regression

- *Wasserstein Ball*

Joint work with Fengqiao Luo

Distributionally Robust Two-Stage Stoch. IP

$$\begin{aligned} \min \quad & c^T x + \max_{P \in \mathfrak{P}} \{ \mathbb{E}_{\xi_P} [Q_\omega(x)] \} \\ \text{s.t.} \quad & Ax \geq b \\ & x \in \{0, 1\}^p \end{aligned} \quad (1.1)$$

$$Q_\omega(x) := \min \quad g_\omega^T y_\omega \quad (1.2a)$$

$$\text{s.t.} \quad W_\omega y_\omega \geq r_\omega - T_\omega x \quad (1.2b)$$

$$y_\omega \in \{0, 1\}^{q_1} \times \mathbb{R}^{q-q_1}. \quad (1.2c)$$

We assume that

1. $X := \{x : Ax \geq b, x \in \{0, 1\}^p\}$ is non-empty,
2. $\mathcal{K}_\omega(x) := \{y_\omega : (1.2b)-(1.2c) \text{ hold}\}$ is non-empty for all $x \in X$ and $\omega \in \Omega$,
3. $Q_\omega(x) < \infty$ for all $x \in X$ and $\omega \in \Omega$ (relatively complete recourse)

Wasserstein Ball

$$\left\{ v \in \mathbb{R}^{|\Omega|} : \sum_{i=1}^{|\Omega|} \sum_{j=1}^{|\Omega|} \|\omega^i - \omega^j\|_1 k_{i,j} \leq \epsilon, \sum_{j=1}^{|\Omega|} k_{i,j} = v_i, \quad i = 1, \dots, |\Omega| \right.$$
$$\left. \begin{aligned} \sum_{i=1}^{|\Omega|} k_{i,j} &= v_j^*, & j &= 1, \dots, |\Omega| & \sum_{i=1}^{|\Omega|} v_i &= 1 \\ v_i &\geq 0, & i &= 1, \dots, |\Omega| \\ k_{i,j} &\geq 0, & i &= 1, \dots, |\Omega|, j = 1, \dots, |\Omega| \end{aligned} \right\}$$

Moment Set

$$\left\{ \underline{u} \leq \sum_{l=1}^{|\Omega|} v_l f(\omega^l) \leq \bar{u}, v \geq 0 \right\}$$

$f(\cdot)$ is some mapping of a sample vector to another vector. e.g., vector of monomials, etc.

L-Shaped Formulation

$$\begin{array}{ll} \min & c^T x + \theta \\ \text{s.t.} & Ax \geq b \\ \max_{P \in \mathfrak{P}} & \{\mathbb{E}_{\xi_P} [Q_\omega(x)]\} \leq \theta \end{array}$$



$$\begin{array}{ll} \min & c^T x + \theta \\ \text{s.t.} & Ax \geq b \\ & \mathbb{E}_{\xi_P} [Q_\omega(x)] \leq \theta, P \in \mathfrak{P} \end{array}$$

We want to stay in the space of x variables as much as we can.

A Distributional Cutting Surface Algorithm (I): DR-TS-SLP

$$\begin{array}{ll} \min & c^T x + \theta \\ \text{s.t.} & Ax \geq b \\ & \mathbb{E}_{\xi_P} [Q_\omega(x)] \leq \theta, P \in \mathfrak{P} \end{array}$$



$$\begin{array}{ll} \min & c^T x + \theta \\ \text{s.t.} & Ax \geq b \\ & \mathbb{E}_{\xi_{P_1}} [Q_\omega(x)] \leq \theta \\ & \mathbb{E}_{\xi_{P_2}} [Q_\omega(x)] \leq \theta \\ & \dots \\ & \mathbb{E}_{\xi_{P_k}} [Q_\omega(x)] \leq \theta \end{array}$$

This algorithm will converge in a finite number of iterations if the distributions used to generate “distributional cuts” are “finite”.

Each sub-problem may be solved using “outer linearization” as in the L-shaped method.

Distribution Separation Problem

we assume that there exists an oracle that provides a probability distribution $P \in \mathfrak{P}$, i.e., $\{p_\omega\}_{\omega \in \Omega}$ where p_ω is the probability of occurrence of scenario $\omega \in \Omega$, by solving the optimization problem:

$$\max_{P \in \mathfrak{P}} \mathbb{E}_{\xi_P} [Q_\omega(x)]$$

for a given $x \in X$.

Moment matching set.

$$\max_{v \in \mathbb{R}^{|\Omega|}} \left\{ \sum_{l=1}^{|\Omega|} v_l Q_{\omega^l}(x) \mid \underline{u} \leq \sum_{l=1}^{|\Omega|} v_l f(\omega^l) \leq \bar{u}, v \geq 0 \right\}$$

Kantorovich set.

$$\max \left\{ \sum_{l=1}^{|\Omega|} v_l Q_{\omega^l}(x) : v \in \mathfrak{P}_K \right\}.$$

The set describing the feasible distributions is a polytope, and an optimum is at its vertex.

L-Shaped Method for DR-TS-SLP

For a given first stage solution (x, θ)

Let $\pi_{\omega,0}^*(x) \in \mathbb{R}^{m_2}$ be the optimal dual corresponding to

$$Q_{\omega}^s(x) := \min g_{\omega}^T y_{\omega}$$

$$s.t. \quad W_{\omega} y_{\omega} \geq r_{\omega} - T_{\omega} x$$

$$y_{\omega} \in \mathbb{R}_+^q.$$

optimality cut

$$\sum_{\omega \in \Omega} p_{\omega} \left\{ \pi_{\omega,0}^*(x)^T (r_{\omega} - T_{\omega} x) \right\} \leq \theta,$$

where $\{p_{\omega}\}_{\omega \in \Omega}$ is obtained by solving the distribution separation problem associated to the ambiguity set \mathfrak{B} .

Note: In this approach, the distributions generating the cuts are not added explicitly.

We do exactly what we do in the L-shaped Method, but solve one additional linear program to determine the weights (probabilities) corresponding to each scenario.

L-Shaped Method for DR-TS-Mixed Binary

First, we define subproblem $\mathcal{S}_\omega(x)$,

$$\begin{aligned} Q_\omega^s(x) &:= \min g_\omega^T y_\omega \\ \text{s.t. } & W_\omega y_\omega \geq r_\omega - T_\omega x \\ & \alpha_\omega^t y_\omega \geq \beta_\omega^t - \psi_\omega^t x, \quad t = 1, \dots, \tau_\omega \\ & y_\omega \in \mathbb{R}_+^q, \end{aligned}$$

where $\alpha_\omega^t \in \mathbb{Q}^q$, $\psi_\omega^t \in \mathbb{Q}^p$, and $\beta_\omega^t \in \mathbb{Q}$ are the coefficients of y_ω , coefficients of x , right hand side, respectively, of a parametric inequality.

optimality cut

For a given

$$\sum_{\omega \in \Omega} p_\omega \left\{ \pi_{\omega,0}^*(x)^T (r_\omega - T_\omega x) + \sum_{t=1}^{\tau_\omega} \pi_{\omega,t}^*(x) (\beta_\omega^t - \psi_\omega^t x) \right\} \leq \theta.$$

Observation: The second stage polytope is (as it gets convexified) does not depend on the probability distribution.

DR-TS-Mixed Binary Programs:

DR version of Stochastic Server Location; DR-Multiple Knapsacks

DRSLP and DRMKP Instances									
Instance	Stage I			Stage II					
	#Cons	#BinVar	#ContVar	#Cons	#BinVar	#ContVar	$ \Omega $	RandParam	
DRSLP.5.25.50	1	5	0	30	125	5	50	RHS	
DRSLP.5.25.100	1	5	0	30	125	5	100	RHS	
DRSLP.10.50.50	1	10	0	60	500	10	50	RHS	
DRSLP.10.50.100	1	10	0	60	500	10	100	RHS	
DRSLP.10.50.500	1	10	0	60	500	10	500	RHS	
DRSLP.15.45.5	1	15	0	60	675	15	5	RHS	
DRSLP.15.45.10	1	15	0	60	675	15	10	RHS	
DRSLP.15.45.15	1	15	0	60	675	15	15	RHS	
DRMKP.1	50	240	0	5	120	0	20	OBJ	
DRMKP.2	50	240	0	5	120	0	20	OBJ	
DRMKP.3	50	240	0	5	120	0	20	OBJ	
DRMKP.4	50	240	0	5	120	0	20	OBJ	
DRMKP.5	50	240	0	5	120	0	20	OBJ	
DRMKP.6	50	240	0	5	120	0	20	OBJ	
DRMKP.7	50	240	0	5	120	0	20	OBJ	
DRMKP.8	50	240	0	5	120	0	20	OBJ	
DRMKP.9	50	240	0	5	120	0	20	OBJ	
DRMKP.10	50	240	0	5	120	0	20	OBJ	

DR-TS-Mixed Binary Programs: Full-Distribution Cuts versus L-Shaped on Wasserstein Ball Models

Instance	L-Shaped			Full-Dist. Cuts	
	$\epsilon = 5.0$			$\epsilon = 5.0$	
	z_{opt}	#DCs	$T(s)$	#DCs	$T(s)$
DRSLP.5.25.50	14.0	7	2.6	5	4.1
DRSLP.5.25.100	-40.0	10	7.3	7	21.2
DRSLP.10.50.50	-200.0	5	240.0	3	136.2
DRSLP.10.50.100	-237.0	16	656.1	7	712.9
DRSLP.10.50.500	-159.0	7	1151.6	3	611.9
DRSLP.15.45.5	-252.0	5	288.5	5	182.2
DRSLP.15.45.10	-220.0	7	518.4	5	772.5
DRSLP.15.45.15	-208.0	11	1203.2	4	584.0
DRMKP.1	9686.0	10	285.4	10	243.2
DRMKP.2	9388.0	9	906.0	10	878.4
DRMKP.3	8844.0	10	1462.2	11	1345.1
DRMKP.4	9237.0	23	2695.1	14	10800.0
DRMKP.5	10024.0	9	1656.9	11	3732.5
DRMKP.6	9515.0	9	257.3	10	225.0
DRMKP.7	10003.0	9	434.4	10	386.7
DRMKP.8	9427.0	28	4554.3	18	2910.4
DRMKP.9	10038.0	10	1090.0	18	10800.0
DRMKP.10	9082.2	13	4870.0	10	10800.0

Full-Distribution Cut version fails in 3-hours

DR-TS-Mixed Binary Programs: Full-Distribution Cuts versus L-Shaped on 3-Moment Models

Instance	L-Shaped			Full-Dist. Cuts		
	<i>CI</i> = 80%			<i>CI</i> = 80%		
	z_{opt}	#DCs	<i>T</i> (s)	z_{opt}	#DCs	<i>T</i> (s)
DRSLP.5.25.50	-93.39	99	4.2	-93.39	12	24.7
DRSLP.5.25.100	-107.73	108	10.1	-107.73	11	32.8
DRSLP.10.50.50	-332.79	385	162.3	-332.79	7	248.4
DRSLP.10.50.100	-325.03	472	413.6	-325.03	8	466.8
DRSLP.10.50.500	-325.03	499	7234.2	-325.03	23	10191
DRSLP.15.45.5	-255.03	30	272.6	-255.03	6	63.2
DRSLP.15.45.10	-242.70	119	743.7	-242.70	6	175.4
DRSLP.15.45.15	-237.05	347	650.1	-237.05	8	588.7
DRMKP.1	9418.71	13	289.3	9418.71	5	292.5
DRMKP.2	9093.42	15	319.1	9093.42	6	622.5
DRMKP.3	8619.42	19	389.4	8619.42	6	513.6
DRMKP.4	8990.40	31	661.9	8990.40	6	724.0
DRMKP.5	9503.67	15	510.1	9503.67	8	1188.1
DRMKP.6	9204.78	18	486.0	9204.78	7	1129.1
DRMKP.7	9709.79	13	684.0	9709.79	6	1397.0
DRMKP.8	9199.72	40	1555.2	9199.72	6	1610.7
DRMKP.9	9830.45	50	1298.0	9830.45	6	1621.4
DRMKP.10	8864.22	49	4547.7	8864.22	6	3165.9

← Full-Distribution Cut version fails in 3-hours

WRO + Machine Learning (Logistic Regression):

$$\min_{\theta} \max_{P \in \mathcal{P}} \mathbb{E}_P[h(\theta^T \xi)].$$

WR0-Logistic Regression

	$h(\theta, \xi)$	Ξ	Master	Sep	Method
E&K (2015)	convex in θ , concave in ξ	convex compact	convex	convex	conjugate
S-A (2015)	loss function of log. reg.	\mathbb{R}^k	convex	convex	closed form sol.
L&M (2017)	convex in θ and ξ	convex compact	convex SIP	DC	central cutting-surface

E&K: Esfahan and Kuhn (2015) S-A: Shafieezadeh-Abadeh et al. (2015)

L&M: Luo and Mehrotra (2017)

WRO-Primal Conic Reformulation in Joint Probability Space

Theorem

Let Θ and Ξ be *compact* sets. The function $h(\cdot, \cdot)$ is *bounded* on $\Theta \times \Xi$. For every $\theta \in \Theta$, there exists a $C(\theta) > 0$ such that $|h(\theta, s_1) - h(\theta, s_2)| \leq C(\theta)d(s_1, s_2)$, $\forall s_1, s_2 \in \Xi$. Then Wass-DRO can be reformulated as a conic linear program as follows:

$$\min_{\theta \in \Theta} \max_{P \in \mathcal{P}} \mathbb{E}_{\xi \sim P} [h(\theta, \xi)]$$

$$\text{st. } \mathcal{W}(P, P_0) \leq r$$

(WRO)

Equivalent to



$$\min_{\theta \in \Theta} \max_{\mu} \int_{\Xi} h(\theta, s) \mu(ds \times \Xi)$$

$$\text{st. } \mu(\Xi, \{\hat{\xi}_i\}) = 1/m, \quad i \in [m]$$

$$\mu(\Xi \times \Xi^{m+1}) \geq 0$$

$$\sum_{i=1}^m \int_{\Xi} d(s, s^i) \mu(ds \times \{\hat{\xi}_i\}) \leq r$$

$$\mu \succeq 0 \quad (\text{ConicLP})$$

WRO-Dual Reformulation: Decomposes by Scenario

Theorem

Applying conic duality, (ConicLP) can be reformulated as the following semi-infinite program, and the duality gap is zero.

$$\min_{\theta \in \Theta} \max_{\mu} \int_{\Xi} h(\theta, s) \mu(ds \times \Xi)$$

$$\text{st. } \mu(\Xi, \{\hat{\xi}_i\}) = 1/m, \quad i \in [m]$$

$$\mu(\Xi \times \Xi^{m+1}) \geq 0$$

$$\sum_{i=1}^m \int_{\Xi} d(s, s^i) \mu(ds \times \{\hat{\xi}_i\}) \leq r$$

$$\mu \succeq 0 \quad (\text{ConicLP})$$

$$\min_{\theta, v} \frac{1}{m} \sum_{i=1}^m v_i + r \cdot v_{m+1}$$

$$\text{st. } h(\theta, s) - v_i - v_{m+1} \cdot d(s, \hat{\xi}_i) \leq 0, \\ \forall s \in \Xi, i \in [m]$$

$$\theta \in \Theta, v_{m+1} \geq 0$$

(WRO-dual)

Dualization



WRO-Dual Reformulation as a Semi-infinite Program

Define the following functions:

$$f(x) := \frac{1}{m} \sum_{i=1}^m v_i + r_0 \cdot v_{m+1},$$

$$g_i(x, s) := h(\theta, s) - v_i - v_{m+1} \cdot d(s, \xi^i), \quad i \in [m].$$

The problem (WRO-dual) can be rewritten as:

$$\min_x f(x)$$

$$\text{st. } g_i(x, s) \leq 0, \forall s \in \Xi, i \in [m]$$

$$x \in X$$

(SIP)

$$d(s) \leftarrow v_i + v_{m+1} \cdot d(s, \xi^i),$$

$$g(s) \leftarrow h_\theta(l(\theta, s)) - d(s).$$

$$\text{Logistic Reg.} \quad \log \left(1 + \exp \left[-y(\theta_0 + \theta^T x) \right] \right)$$

Separation Oracle for a Cutting Surface Algorithm

$$\max_{s \in \Xi} g_i(\tilde{x}, s) := h(\tilde{\theta}, s) - \tilde{v}_i - \tilde{v}_{m+1} \cdot d(s, \hat{\xi}_i)$$

- The separation problem is equivalent to the following unconstrained DC optimization:

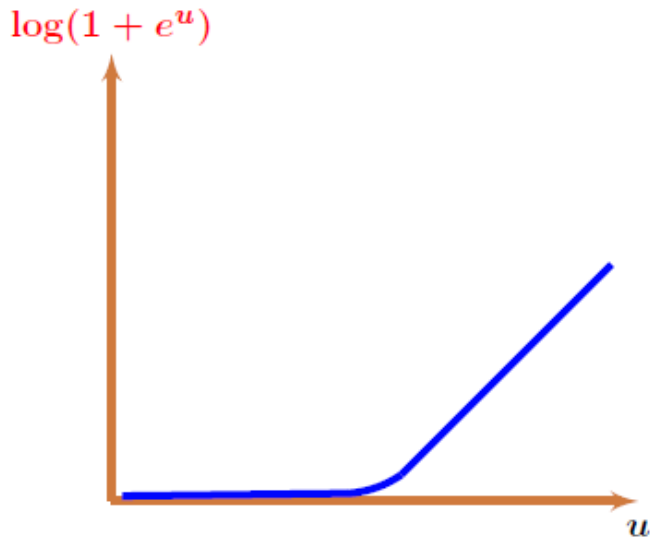
$$\max_u \psi(u) := h_\theta(u) - \phi(u),$$

where

$$\phi(u) = \min_{s \in \Xi} d(s), \quad \text{s.t. } u = l(\tilde{\theta}, s).$$

- **Assumption:** Ξ is a polytope, and the metric $d(s_1, s_2) := \|s_1 - s_2\|_1$ is the 1-norm.
- $\phi(u)$ becomes a univariate piecewise-linear convex function.
- $h_\theta(u)$ is an univariate convex function \implies piecewise-linear approximation.
- Subproblem induced by each linear piece is convex optimization.

WRO-Logistic Regression Separation Problem



$$h_{\theta}(u) = \log(1 + e^u)$$

$$h_{\theta}(u) \rightarrow u, \quad \text{as } u \rightarrow \infty$$

$$h_{\theta}(u) \rightarrow 0, \quad \text{as } u \rightarrow -\infty$$

$$h''_{\theta}(u) = \mathcal{O}(e^{-|u|}).$$

Theorem

For the distributionally robust logistic regression (DRLR) model with (univariate) logistic loss function $h_{\theta}(u) = \log(1 + e^u)$, the separation problem can be solved in at most $O\left(\frac{1}{\sqrt{\varepsilon}} \log \log \frac{L}{\varepsilon}\right)$ iterations, where $L := u_{ub} - u_{lb}$.

Performance Results: Implementation

- All algorithms are Implemented in C++.
- **Master Problem**: twice-differentiable convex program → Interior Point Method (Ipopt: Wächter and Biegler, 2006).
- **Separation Problem**: DC optimization → Sequence of Parametric Linear Programs (Cplex).

Data: UCI Repository

Data sets for numerical study

- Select **11** data sets from UCI machine learning repository.
- Training sample size: **50, 75, 100, 150**.
- Candidate Wasserstein radius $r = 0, 0.01, 0.05, 0.1, 0.5, 1$.
- Each experiment is repeated **100** times.

Data set	Area	No. Attrib.	No. Observ.
BA	Finance	4	1372
VC	Health care	6	310
PID	Health care	8	768
BCW	Health care	9	699
ST-H	Health care	13	270
EES	Health care	14	14980
SPT-H	Health care	22	267
ION	Aerospace	34	351
SPTF-H	Health care	44	267
SPAM	Computer	57	4601
CB	Aerospace	60	208

Performance: Out of Sample Predictability

Compare the mean AUC value between WRLR and LR in 44 cases. With $\alpha = 0.05$, WRLR is better in 55% cases; LR is better in 16% cases; No significant difference in the remaining 29% cases.

Dataset	m	LR AUC	WRLR AUC	Rel. Diff	p-value
BCW	50	.9716	.9916	.7040	.0000
	75	.9773	.9886	.4954	.0000
	100	.9790	.9940	.7122	.0000
	150	.9889	.9945	.5049	.0000
ST-H	50	.8317	.8808	.2914	.0000
	75	.8504	.8903	.2664	.0000
	100	.8945	.9064	.1133	.0000
	150	.8986	.8990	.0042	.4319
ION	50	.8429	.8708	.1775	.0000
	75	.8582	.8919	.2381	.0000
	100	.8606	.8967	.2584	.0000
	150	.8715	.9006	.2264	.0000

Performance: Out of Sample Loss Function

data set	Non-regularized		l_1 -regularized	
	LR mean loss	WRLR mean loss	LR mean loss	WRLR mean loss
BA	0.0454	0.0534	0.0689	0.0786
VC	0.3079	0.3091	0.3150	0.3286
PID	0.5288	0.5153	0.5184	0.5153
BCW	0.2775	0.1110	0.1008	0.0994
ST-H	0.4162	0.3791	0.3970	0.3855
EES	0.6863	0.6685	0.6747	0.6650
SPT-H	0.8240	0.4162	0.3831	0.3821
ION	2.7404	0.4239	0.3656	0.3346
SPTF-H	0.9275	0.3787	0.4044	0.3965
SPAM	3.5443	0.7297	0.3679	0.3346
CB	5.4601	0.8451	0.5058	0.4413

Performance: Computational

Computational performance of solving WR-LogReg

- Number of calls to the master problem: $4 \sim 40$.
- Approximately $2m \sim 20m$ cutting surfaces are added. (m : number of training samples)

Dataset	m	Iters.	Cuts	CPU [sec]	Master (%)	Sep. (%)
BA	50	3.8	66.9	1.21	13.74	86.26
	75	4.3	90.8	0.86	17.96	82.04
	100	3.9	116.7	1.83	13.89	86.11
	150	4.6	157.5	2.30	14.42	85.58
BCW	50	8.6	251.8	6.19	31.44	68.56
	75	9.4	284	7.88	27.48	72.52
	100	8.9	501.4	12.84	34.28	65.72
	150	9.6	786.1	27.31	41.79	58.21
SPT-H	50	21.5	938.8	38.87	88.91	11.09
	75	24.5	1031.2	53.43	83.40	16.60
	100	19.4	1122.5	63.08	83.44	16.56
	150	13.4	1384	49.70	77.91	22.09