Distributionally Robust Stochastic Optimization and Regularization

Yinyu Ye

Department of Management Science and Engineering

Stanford University

Stanford, CA 94305, U.S.A.

http://www.stanford.edu/~yyye

General Optimization Problems

Consider mathematical programming (MP) problem in general form

$$(\mathsf{P}) \qquad \begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in X. \end{array}$$

A feasible solution of a given problem is a solution point that satisfies all constraints, that is, in feasible region X; while a global optimal solution is a feasible solution who possesses the lowest objective value, and a local optimal solution is a feasible solution who possesses the lowest objective value among its neighboring feasible solutions.

Question: How does one recognize or certify a (local) optimal solution to a generally constrained and objectived optimization problem? Answer: Lagrangian Theory and Optimality Conditions.

We now review the Lagrangian Duality theory as an alternative to Conic Duality theory. For general nonlinear constraints, the Lagrangian Duality theory is more applicable.

Lagrangian Theory and Review

Let the optimization problem be represented by

$$\text{(GCO)} \qquad \begin{array}{lll} \min & f(\mathbf{x}) & \\ \text{s.t.} & c_i\mathbf{x}) & (\leq, =, \geq) \quad 0,\ i = 1, ..., m, \end{array}$$

For Lagrange Multipliers:

$$Y := \{y_i \quad (\leq, '\text{ free}', \geq) \quad 0,\ i = 1, ..., m\},$$

the Lagrangian Function is given by

$$L(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - \mathbf{y}^T \mathbf{c}(\mathbf{x}) = f(\mathbf{x}) - \sum_{i=1}^{m} y_i c_i(\mathbf{x}),\ \mathbf{y} \in Y.$$

Optimality Conditions: Find both feasible $\mathbf{x}$ and $\mathbf{y}$ such that

$$\nabla_x L(\mathbf{x}, \mathbf{y}) = \mathbf{0} \quad \text{and} \quad y_i c_i(\mathbf{x}) = 0 \ \forall i.$$

Toy Example

$$\begin{aligned} \text{minimize} \quad & (x_1 - 1)^2 + (x_2 - 1)^2 \\ \text{subject to} \quad & x_1 + 2x_2 - 1 \leq 0, \\ & 2x_1 + x_2 - 1 \leq 0. \end{aligned}$$

$$L(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - \mathbf{y}^T \mathbf{c}(\mathbf{x}) = f(\mathbf{x}) - \sum_{i=1}^{2} y_i c_i(\mathbf{x}) =$$

$$= (x_1 - 1)^2 + (x_2 - 1)^2 - y_1(x_1 + 2x_2 - 1) - y_2(2x_1 + x_2 - 1), \ (y_1; y_2) \leq \mathbf{0}$$

where

$$\nabla L_x(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} 2(x_1 - 1) - y_1 - 2y_2 \\ 2(x_2 - 1) - 2y_1 - y_2 \end{pmatrix}$$

Lagrangian Relaxation Problem

For given multipliers $\mathbf{y} \in Y$, consider problem

$$(LRP) \quad \inf \quad L(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - \mathbf{y}^T \mathbf{c}(\mathbf{x})$$
$$\text{s.t.} \quad \mathbf{x} \in R^n.$$

Here, $\mathbf{y}_i$ can be viewed as a penalty parameter to penalize constraint violation $c_i(\mathbf{x}), i = 1, ..., m$.

In the toy example, for given $(y_1; y_2) \leq \mathbf{0}$, the LRP is:

$$\inf \quad (x_1 - 1)^2 + (x_2 - 1)^2 - y_1(x_1 + 2x_2 - 1) - y_2(2x_1 + x_2 - 1)$$
$$\text{s.t.} \quad (x_1; x_2) \in R^2,$$

and it has a close form solution $\mathbf{x}$:

$$x_1 = \frac{y_1 + 2y_2}{2} + 1 \quad \text{and} \quad x_2 = \frac{2y_1 + y_2}{2} + 1$$

with the minimal value function $= -1.25y_1^2 - 1.25y_2^2 - 2y_1y_2 - 2y_1 - 2y_2$.

Inf-Value Function as the Dual Objective

For any $\mathbf{y} \in Y$, the the minimal value function (including unbounded from below or infeasible cases) and the Lagrangian Dual Problem (LDP) are given by:

$$\begin{array}{lllll} \phi(\mathbf{y}) := & \inf_{\mathbf{x}} & L(\mathbf{x}, \mathbf{y}), & \text{s.t.} & \mathbf{x} \in R^n. \\ (LDP) & \sup_{\mathbf{y}} & \phi(\mathbf{y}), & \text{s.t.} & \mathbf{y} \in Y. \end{array}$$

Theorem 1 *The Lagrangian dual objective $\phi(\mathbf{y})$ is a concave function.*

Theorem 2 *(Weak duality theorem) For every $\mathbf{y} \in Y$, the Lagrangian dual function $\phi(\mathbf{y})$ is less or equal to the infimum value of the original GCO problem.*

Proof:

$$\begin{aligned} \phi(\mathbf{y}) &= \inf_{\mathbf{x}} \{f(\mathbf{x}) - \mathbf{y}^T \mathbf{c}(\mathbf{x})\} \\ &\le \inf_{\mathbf{x}} \{f(\mathbf{x}) - \mathbf{y}^T \mathbf{c}(\mathbf{x}) \text{ s.t. } \mathbf{c}(\mathbf{x})(\le, =, \ge)\mathbf{0}\} \\ &\le \inf_{\mathbf{x}} \{f(\mathbf{x}) : \text{ s.t. } \mathbf{c}(\mathbf{x})(\le, =, \ge)\mathbf{0}\}. \end{aligned}$$

The first inequality is from the fact that the unconstrained inf-value is no greater than the constrained one, and the second inequality is from $\mathbf{c}(\mathbf{x})(\le, =, \ge)\mathbf{0}$ and $\mathbf{y}(\le, '\text{free}', \ge)\mathbf{0}$ imply $-\mathbf{y}^T \mathbf{c}(\mathbf{x}) \le 0$.

The Lagrangian Dual Problem for the Toy Example

$$\begin{aligned} \text{minimize} \quad & (x_1 - 1)^2 + (x_2 - 1)^2 \\ \text{subject to} \quad & x_1 + 2x_2 - 1 \leq 0, \\ & 2x_1 + x_2 - 1 \leq 0; \end{aligned}$$

where $\mathbf{x}^* = \left(\frac{1}{3};\ \frac{1}{3}\right)$.

$$\phi(\mathbf{y}) = -1.25y_1^2 - 1.25y_2^2 - 2y_1y_2 - 2y_1 - 2y_2,\ \mathbf{y} \leq \mathbf{0}.$$

$$\begin{aligned} \max \quad & -1.25y_1^2 - 1.25y_2^2 - 2y_1y_2 - 2y_1 - 2y_2 \\ \text{s.t.} \quad & (y_1; y_2) \leq \mathbf{0}. \end{aligned}$$

where $\mathbf{y}^* = \left(\frac{-4}{9};\ \frac{-4}{9}\right)$.

The Lagrangian Dual of LP I

Consider LP problem

$$(LP) \quad \text{minimize} \quad \mathbf{c}^T\mathbf{x}$$
$$\text{subject to} \quad A\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0};$$

and its conic dual problem is given by

$$(LD) \quad \text{maximize} \quad \mathbf{b}^T\mathbf{y}$$
$$\text{subject to} \quad A^T\mathbf{y} + \mathbf{s} = \mathbf{c},\ \mathbf{s} \geq \mathbf{0}.$$

We now derive the Lagrangian Dual of (LP). Let the Lagrangian multipliers be $\mathbf{y}$('free') for equalities and $\mathbf{s} \geq \mathbf{0}$ for constraints $\mathbf{x} \geq \mathbf{0}$. Then the Lagrangian function would be

$$L(\mathbf{x}, \mathbf{y}, \mathbf{s}) = \mathbf{c}^T\mathbf{x} - \mathbf{y}^T(A\mathbf{x} - \mathbf{b}) - \mathbf{s}^T\mathbf{x} = (\mathbf{c} - A^T\mathbf{y} - \mathbf{s})^T\mathbf{x} + \mathbf{b}^T\mathbf{y};$$

where $\mathbf{x}$ is "free".

The Lagrangian Dual of LP II

Now consider the Lagrangian dual objective

$$\phi(\mathbf{y}, \mathbf{s}) = \inf_{\mathbf{x} \in R^n} L(\mathbf{x}, \mathbf{y}, \mathbf{s}) = \inf_{\mathbf{x} \in R^n} \left[(\mathbf{c} - A^T\mathbf{y} - \mathbf{s})^T\mathbf{x} + \mathbf{b}^T\mathbf{y} \right].$$

If $(\mathbf{c} - A^T\mathbf{y} - \mathbf{s}) \neq \mathbf{0}$, then $\phi(\mathbf{y}, \mathbf{s}) = -\infty$. Thus, in order to maximize $\phi(\mathbf{y}, \mathbf{s})$, the dual must choose its variables $(\mathbf{y}, \mathbf{s} \geq \mathbf{0})$ such that $(\mathbf{c} - A^T\mathbf{y} - \mathbf{s}) = \mathbf{0}$.

This constraint, together with the sign constraint $\mathbf{s} \geq \mathbf{0}$, establish the Lagrangian dual problem:

$$\begin{array}{lll} (LDP) & \text{maximize} & \mathbf{b}^T\mathbf{y} \\ & \text{subject to} & A^T\mathbf{y} + \mathbf{s} = \mathbf{c},\ \mathbf{s} \geq \mathbf{0}. \end{array}$$

which is identical to the conic dual of LP.

Lagrangian Strong Duality Theorem

Theorem 3 *Let (GCO) be a convex minimization problem and the infimum f^* of (GCO) be finite, and the suprermum of (LDP) be ϕ^*. In addition, let (GCO) have an interior-point feasible solution with respect to inequality constraints, that is, there is $\hat{\mathbf{x}}$ such that all inequality constraints are strictly held. Then, $f^* = \phi^*$, and (LDP) admits a maximizer $\mathbf{y}^*$ such that*

$$\phi(\mathbf{y}^*) = f^*.$$

Furthermore, if (GCO) admits a minimizer $\mathbf{x}^*$, then

$$y_i^* c_i(\mathbf{x}^*) = 0, \ \forall i = 1, ..., m.$$

The assumption of "interior-point feasible solution" is called Constraint Qualification condition, which was also needed as a condition to prove the strong duality theorem for general Conic Linear Optimization.

Note that the problem would be a convex minimization problem if all equality constraints are hyperplane or affine functions $c_i(\mathbf{x}) = \mathbf{a}_i\mathbf{x} - b_i$, all other level sets are convex.

More on Lagrangian Duality

Consider the constrained problem with additional constraints

$$\begin{array}{lll} (GCO) & \inf & f(\mathbf{x}) \\ & \text{s.t.} & \mathbf{c}_i(\mathbf{x}) \ (\leq, =, \geq) \ 0, \ i = 1, ..., m, \\ & & \mathbf{x} \in \Omega \subset R^n. \end{array}$$

Typically, Ω has a simple form such as the cone

$$\Omega = R^n_+ = \{\mathbf{x} : \ \mathbf{x} \geq \mathbf{0}\}$$

or the box

$$\Omega := \{\mathbf{x} : \ -\mathbf{e} \leq \mathbf{x} \leq \mathbf{e}.\}$$

Then, when derive the Lagrangian dual, there is not need to introduce multipliers for Ω constraints.

Lagrangian Relaxation Problem

Consider again the (partial) Lagrangian Function:

$$L(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - \mathbf{y}^T \mathbf{c}(\mathbf{x}),\ \mathbf{y} \in Y;$$

and define the dual objective function of $\mathbf{y}$ be

$$\begin{array}{lll} \phi(\mathbf{y}) := & \inf_{\mathbf{x}} & L(\mathbf{x}, \mathbf{y}) \\ & \text{s.t.} & \mathbf{x} \in \Omega. \end{array}$$

Theorem 4 *The Lagrangian dual function $\phi(\mathbf{y})$ is a concave function.*

Theorem 5 *(Weak duality theorem) For every $\mathbf{y} \in Y$, the Lagrangian dual function value $\phi(\mathbf{y})$ is less or equal to the infimum value of the original GCO problem.*

The Lagrangian Dual Problem

$$(LDP) \quad \sup \quad \phi(\mathbf{y})$$
$$\text{s.t.} \quad \mathbf{y} \in Y.$$

would called the Lagrangian dual of the original GCO problem:

Theorem 6 *(Strong duality theorem) Let (GCO) be a convex minimization problem, the infimum f^* of (GCO) be finite, and the suprermum of (LDP) be ϕ^*. In addition, let (GCO) have an interior-point feasible solution with respect to inequality constraints, that is, there is $\hat{\mathbf{x}}$ such that all inequality constraints are strictly held. Then, $f^* = \phi^*$, and (LDP) admits a maximizer $\mathbf{y}^*$ such that*

$$\phi(\mathbf{y}^*) = f^*.$$

Furthermore, if (GCO) admits a minimizer $\mathbf{x}^*$, then

$$y_i^* c_i(\mathbf{x}^*) = 0, \ \forall i = 1, ..., m.$$

The Lagrangian Dual of LP with Bound Constraints

Consider

$$(LP) \quad \text{minimize} \quad \mathbf{c}^T\mathbf{x}$$
$$\text{subject to} \quad A\mathbf{x} = \mathbf{b},\ -\mathbf{e} \le \mathbf{x} \le \mathbf{e}\ (\|\mathbf{x}\|_\infty \le 1);$$

Let the Lagrangian multipliers be $\mathbf{y}$ for equality constraints. Then the Lagrangian dual objective would be

$$\phi(\mathbf{y}) = \inf_{-\mathbf{e}\le\mathbf{x}\le\mathbf{e}} L(\mathbf{x},\mathbf{y}) = \inf_{-\mathbf{e}\le\mathbf{x}\le\mathbf{e}} \left[(\mathbf{c} - A^T\mathbf{y})^T\mathbf{x} + \mathbf{b}^T\mathbf{y}\right];$$

where if $(\mathbf{c} - A^T\mathbf{y})_j \le 0$, $x_j = 1$; and otherwise, $x_j = -1$.

Therefore, the Lagrangian dual is

$$(LDP) \quad \text{maximize} \quad \mathbf{b}^T\mathbf{y} - \|\mathbf{c} - A^T\mathbf{y}\|_1$$
$$\text{subject to} \quad \mathbf{y} \in R^m.$$

Stochastic Optimization

We start from considering a stochastic optimization problem as follows:

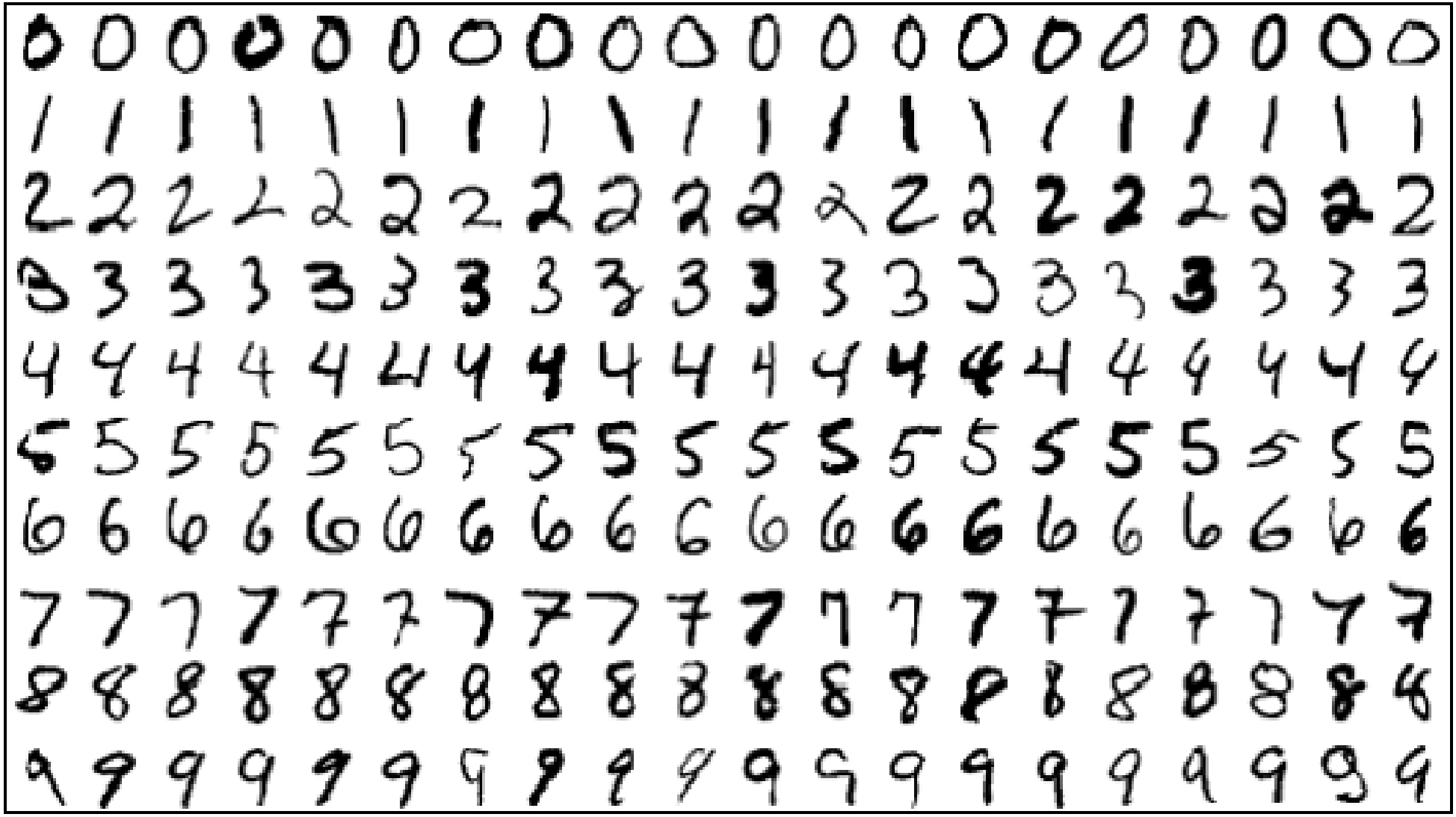
$$\text{minimize}_{\mathbf{x}\in X} \quad \mathbf{E}_{F_\xi}[h(\mathbf{x},\xi)] \tag{1}$$

where $\mathbf{x}$ is the decision variable with feasible region X, ξ is a random variable satisfying probability distribution F_ξ.

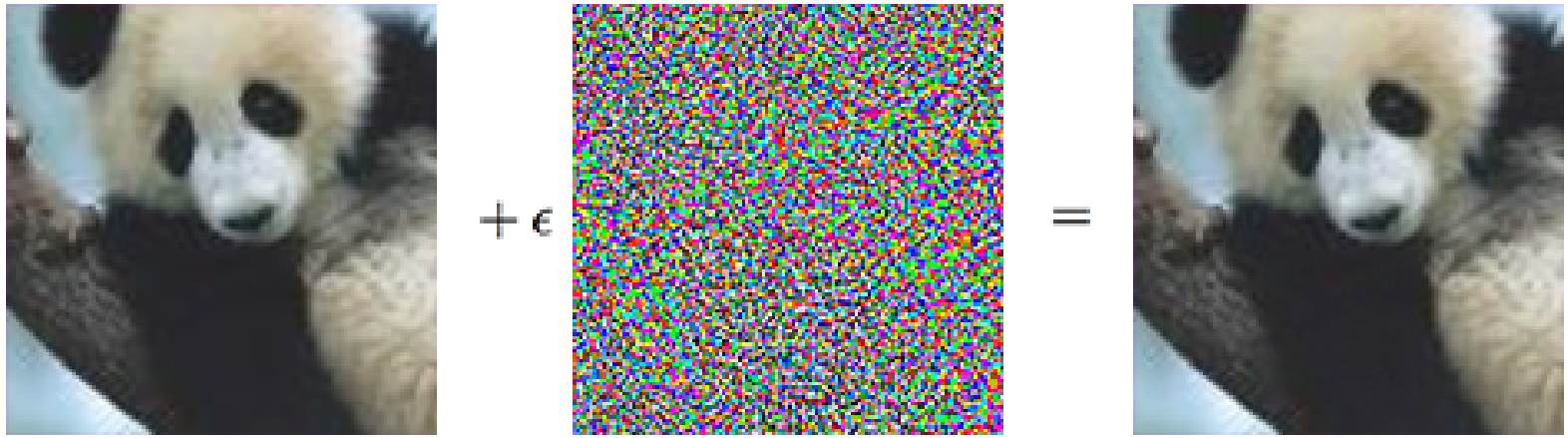
- Pros: In many cases, the expected value is a good measure of performance, and can be solved by sampling.
- Cons: i) Variance or risk is not considered; ii) one has to know the exact distribution of ξ to perform the stochastic optimization; and deviant from the assumed distribution may result in sub-optimal solutions.

Fact: $\mathbf{E}_{F_\xi}[h(x,\xi)]$ is a linear function of probability distribution F_ξ.

Sample-Based Learning



Sample-Based Learning with Noises/Distortions



Goodfellow et al. [2014]

Robust Optimization

In order to overcome the lack of knowledge on the distribution, people proposed the following (static) robust optimization approach:

$$\text{minimize}_{\mathbf{x}\in X} \quad \max_{\xi\in\Xi} h(\mathbf{x},\xi) \tag{2}$$

where Ξ is the support of ξ

- Pros: Robust to any distribution, and only the support of the uncertain parameters are needed.
- Cons: Too conservative – the decision that maximizes the worst-case pay-off may perform badly in usual cases. Also, even for fixed $\mathbf{x}$ the function of ξ may be neither convex nor concave, and Ξ may be difficult to define.

Distributionally Robust Optimization

- In practice, although the exact distribution of the random variables may not be known, people usually know that it falls into a certain range
- People want to choose an intermediate approach between stochastic optimization, which has no robustness in the error of distribution, and the robust optimization, which admits "too many" distributions, even those unrealistic ones (e.g., single point distribution on the boundary of support set)

A solution to the above-mentioned question is to take the following Distributionally Robust Optimization (DRO) approach:

$$\text{minimize}_{\mathbf{x}\in X} \quad \max_{F_\xi \in \Gamma} \mathbf{E}_{F_\xi}[h(\mathbf{x},\xi)] \tag{3}$$

In DRO, we consider a set of distributions Γ and maximizes the worst-case expected value among the distributions in Γ.

Again, $\mathbf{E}_{F_\xi}[h(\mathbf{x},\xi)]$ is a linear function of probability distribution F_ξ so that it is always a convex optimization when Γ is a convex set.

The Choice of the Distribution Set

In DRO, the most important issue is the choice of the distribution set Γ. When choosing Γ, we need to consider the following:

- Tractability
- Statistical Inferences
- Practical performance (the potential risk comparing to the fully robust approach)

In the following, we will address these topics by introducing three pieces of our research related to DRO

Brief History of DRO

- First introduced by Scarf [1958] in the context of inventory control problem with a single random demand variable.
- Distribution set based on moments: Dupacova [1987], Prekopa [1995], Bertsimas and Popescu [2005], etc; HP-hard in general for multi-random variables
- Firs named in Delage and Y [2009,2010], and gave a convex SDP representation
- Distribution set based on Likelihood/Divergences: Nilim and El Ghaoui [2005], Iyanger [2005], Wang, Glynn and Y [2012], etc
- Distribution set based on Wasserstein ambiguity set: Mohajerin Esfahani and Kuhn [2015], Blanchet et al. [2016], Duchi et al. [2016,17], Gao et al. [2017]
- Axiomatic motivation for DRO: Delage et al. [2017]; Ambiguous Joint Chance Constraints Under Mean and Dispersion Information: Hanasusanto et al. [2017]

DRO with Moment Information

We consider the distributionally robust optimization problem as follows:

$$\text{minimize}_{\mathbf{x}\in X} \quad \max_{F_\xi \in \Gamma} \mathbf{E}_{F_\xi}[h(\mathbf{x},\xi)] \tag{4}$$

where

$$\Gamma = \left\{ F_\xi \in \mathcal{M} \;\middle|\; \begin{array}{l} P(\xi \in \mathcal{S}) = 1 \\ (\mathbf{E}[\xi]-\mu_0)^T \Sigma_0^{-1} (\mathbf{E}[\xi]-\mu_0) \le \gamma_1 \\ \mathbf{E}[(\xi-\mu_0)(\xi-\mu_0)^T] \preceq \gamma_2 \Sigma_0 \end{array} \right\}$$

That is, the distribution set is defined based on the support, first and second order moments (confidence) constraints.

Theorem 7 *Under mild conditions, DRO model presented in previous slide can be solved to any precision* ϵ *in time polynomial in* $\log(1/\epsilon)$ *and the sizes of* $\mathbf{x}$ *and* ξ.

Confidence Region for f_ξ

For

$$\Gamma(\gamma_1, \gamma_2) = \left\{ F_\xi \in \mathcal{M} \left| \begin{array}{l} P(\xi \in \mathcal{S}) = 1 \\ (\mathbf{E}[\xi] - \mu_0)^T \Sigma_0^{-1} (\mathbf{E}[\xi] - \mu_0) \leq \gamma_1 \\ \mathbf{E}[(\xi - \mu_0)(\xi - \mu_0)^T] \preceq \gamma_2 \Sigma_0 \end{array} \right. \right\}$$

When vector μ_0 and matrix Σ_0 are point estimates from the empirical data (of size m) and S lies in a ball of radius R such that $||\xi||_2 \leq R$ a.s.. Then

Theorem 8 *For* $\gamma_1 = O(\frac{R^2}{m} \log(4/\delta))$ *and* $\gamma_2 = O(\frac{R^2}{\sqrt{m}} \sqrt{\log(4/\delta)})$, *we have*

$$P(F_\xi \in \Gamma(\gamma_1, \gamma_2)) \geq 1 - \delta.$$

DRO with Likelihood Bounds

Define the distribution set by the constraint on the likelihood ratio.

With observed Data: $\xi_1, \xi_2, ...\xi_N$, we define

$$\Gamma_N = \left\{ F_\xi \;\middle|\; \begin{array}{l} P(\xi \in \Xi) = 1 \\ L(\xi, F_\xi) \geq \gamma \end{array} \right\}$$

where γ adjusts the level of robustness and N represents the sample size.

For example, assume the support of the uncertainty is finite

$$\xi_1, \xi_2, ...\xi_n$$

and we observed m_i samples on ξ_i. Then, F_ξ has a finite discrete distribution $p_1, ..., p_n$ and

$$L(\xi, F_\xi) = \sum_{i=1}^{n} m_i \log p_i.$$

Theory on Likelihood Bounds

The model is a convex optimization problem, and connects to many statistical theories:

- Statistical Divergence theory: provide a bound on KL divergence
- Bayesian Statistics with the threshold γ estimated by samples: confidence level on the true distribution
- Non-parametric Empirical Likelihood theory: inference based on empirical likelihood by Owen
- Asymptotic Theory of the likelihood region
- Possible extensions to deal with Continuous Case

Wang, Glynn and Y [2012,2016]

DRO using Wasserstein Ambiguity Set

By the Kantorovich-Rubinstein theorem, the Wasserstein distance between two distributions can be expressed as the minimum cost of moving one to the other, which is a semi-infinite transportation LP.

Theorem 9 *When using the Wasserstein ambiguity set*

$$\Gamma_N := \{F_\xi \mid P(\xi \in \Xi) = 1 \ \& \ d(F_\xi, \hat{F}_N) \leq \varepsilon_N\},$$

where $d(F_1, F_2)$ *is the Wasserstein distance function and* N *is the sample size, the DRO model satisfies the following properties:*

- *Finite sample guarantee : the correctness probability* $\bar{P}^N$ *is high*
- *Asymptotic guarantee :* $\bar{P}^\infty(\lim_{N\to\infty} \hat{x}_{\varepsilon_N} = x^*) = 1$
- *Tractability : DRO is in the same complexity class as SAA*

Mohajerin Esfahani & Kuhn [15, 17], Blanchet, Kang, Murthy [16], Duchi, Glynn, Namkoong [16]

The Case of Logistic Regression DRO

- Let $\{(\hat{\xi}_i, \hat{\lambda}_i)\}_{i=1}^N$ be a feature-label training set i.i.d. from P, and consider applying logistic regression :

$$\min_{\mathbf{x}} \frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{x}, \hat{\xi}_i, \hat{\lambda}_i) \text{ where } \ell(\mathbf{x}, \xi, \lambda) = \ln(1 + \exp(-\lambda \mathbf{x}^T \xi))$$

- DRO suggests solving

$$\min_{\mathbf{x}} \sup_{F \in \Gamma_N} {}_F[\ell(\mathbf{x}, \xi_i, \lambda_i)]$$

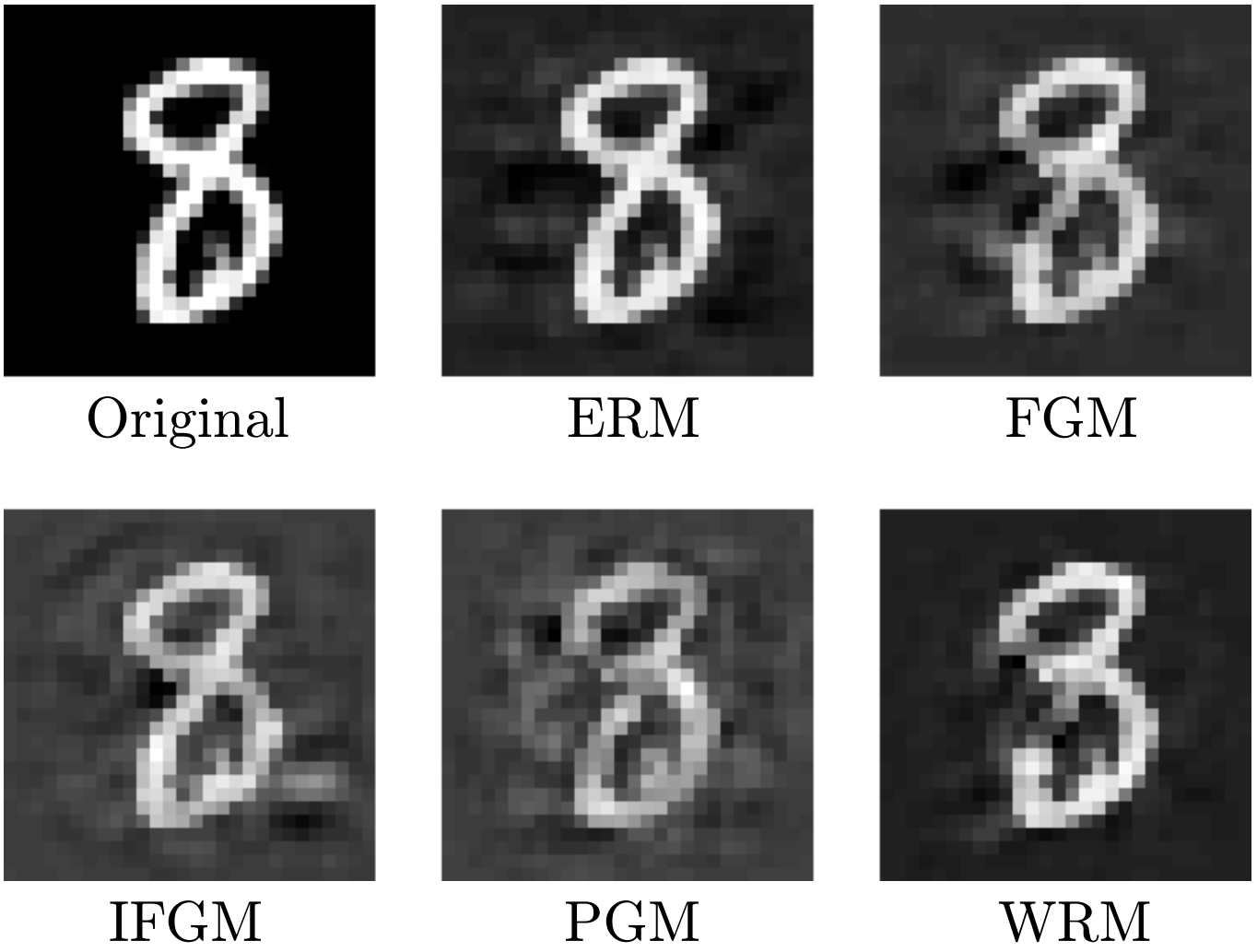
with the Wasserstein ambiguity set.

- When labels are considered to be error free, DRO with Γ_N reduces to regularized logistic regression:

$$\min_{\mathbf{x}} \frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{x}, \hat{\xi}_i, \hat{\lambda}_i) + \varepsilon \|\mathbf{x}\|_*$$

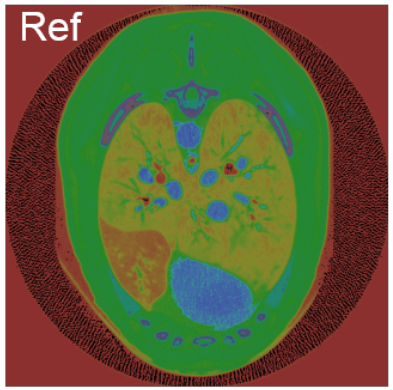
Shafieezadeh Abadeh, Mohajerin Esfahani, & Kuhn, NIPS, [2015]

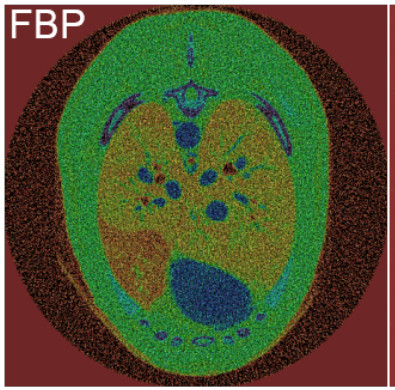
Result of the DRO Learning

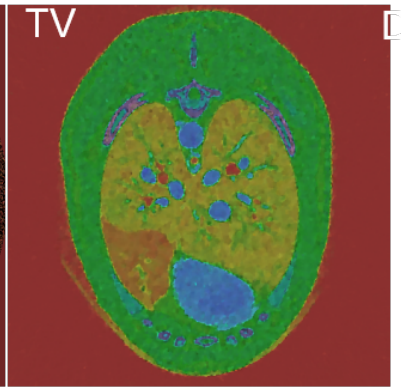


Sinha, Namkoong and Duchi [2017]

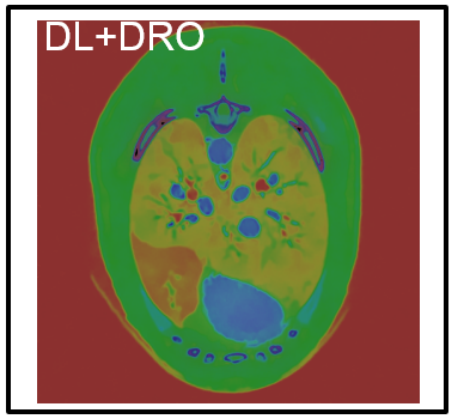
Medical Decision: CT Imaging of Sheep Thorax







Substantial noise reduction



Ref: Filtered Back Projection reconstructions of noise-free data

FBP: FBP reconstructions of noisy data

TV: TV-based reconstruction

DL: Dictionary Learning-based reconstruction

DL+DRO: DL+DRO to encourage low-rankness and robustness

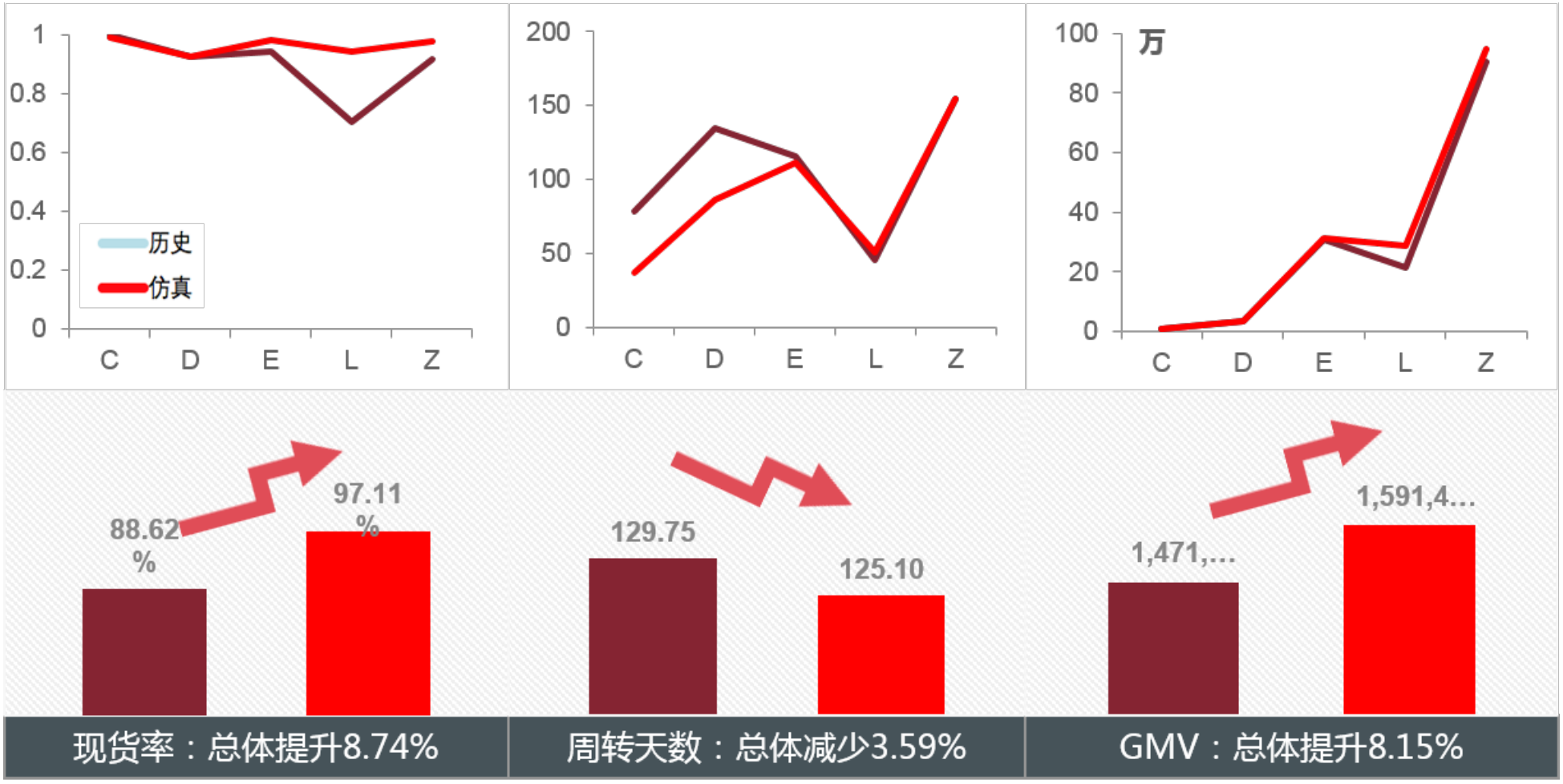
Liu et al. [2017]

Result of the DRO Medical Decision Making

Method	SSIM
FEB	0.37
TV	0.79
DL	0.85
DL+DRO	0.93

Liu et al. [2017]

DRO for Inventory Management (Cardinal Operations)



杉数科技 Cardinal Operations

Summary of DRO under Moment, Likelihood or Wasserstein Ambiguity Set

- Therefore, the DRO models yield a solution with a guaranteed confidence level to the possible distributions. Specifically, the confidence region of the distributions can be constructed upon the historical data and sample distributions.
- The DRO models are tractable, and it sometimes reduces to the same optimization problem with a suitable objective regularization so that it maintains the same computational complexity as the stochastic optimization models with known distribution.
- This approach can be applied to a wide range of problems, including inventory problems (e.g., newsvendor problem), portfolio selection problems, image reconstruction, machine learning, etc., with reported superior numerical results

The key to DRO: $\mathbf{E}_{F_\xi}[h(\mathbf{x}, \xi)]$ is a linear function of probability distribution F_ξ so that it is a convex optimization if Γ is a convex set. Therefore, the inner maximization problem can be replaced by its dual minimization.

Distributionally Robust Learning (DRL) Derivation

The simple sample average minimization

$$\text{mimimize}_{\mathbf{x}\in X} \quad \sum_{k=1}^{N} \hat{p}_k h(\mathbf{x}, \xi_k) \tag{5}$$

where ξ^k represents the kth sample data and $\hat{p}_k$ is its sample/empirical probability.

Suppose we like to "robustfy" the problem by considering

$$\begin{aligned} \text{mimimize}_{\mathbf{x}\in X} \quad & \left[\max_{\mathbf{d}\in\Gamma} \sum_{k=1}^{N} (\hat{p}_k + d_k) h(\mathbf{x}, \xi_k)\right] \\ = \quad & \sum_{k=1}^{N} \hat{p}_k h(\mathbf{x}, \xi_k) + \left[\max_{\mathbf{d}\in\Gamma} \sum_{k=1}^{N} d_k h(\mathbf{x}, \xi_k)\right] \end{aligned} \tag{6}$$

where Γ is given by $\Gamma = \{\mathbf{d} : \sum_{k=1}^{N} d_k = 0, \ \|\mathbf{d}\|_2^2 \le 1/N\}$.

The Inner Optimization Problem

The inner problem is

$$\begin{array}{ll} \max_{\mathbf{d}} & \sum_{k=1}^{N} d_k h(\mathbf{x}, \xi_k) \\ \text{s,t,} & \sum_{k=1}^{N} d_k = 0, \\ & \|\mathbf{d}\|_2^2 \leq 1/N. \end{array}$$

Then, its dual, using the Lagrangian Theory, becomes

$$\min_{y_1} \quad \frac{1}{\sqrt{N}} \|y_1 \mathbf{e} - \mathbf{h}(\mathbf{x}, \xi)\| = \frac{1}{\sqrt{N}} \sqrt{\sum_{k=1}^{N} (y_1 - h(\mathbf{x}, \xi_k))^2}$$

where we implicitly removed the second Lagrange multiplier y_2.

Reformulation of the DRL Problem

$$\text{mimimize}_{\mathbf{x}\in X} \quad \sum_{k=1}^{N} \hat{p}_k h(\mathbf{x},\xi_k) + \quad \frac{1}{\sqrt{N}} \left[\min_{y_1} \sqrt{\sum_{k=1}^{N} (y_1 - h(\mathbf{x},\xi_k))^2} \right]$$

Or

$$\text{mimimize}_{\{\mathbf{x}\in X, y_1\}} \quad \sum_{k=1}^{N} \hat{p}_k h(\mathbf{x},\xi_k) + \frac{1}{\sqrt{N}} \sqrt{\sum_{k=1}^{N} (y_1 - h(\mathbf{x},\xi_k))^2}$$

One can further simplify using

$$y_1 = \frac{1}{N} \sum_{k=1}^{N} h(\mathbf{x},\xi_k)$$

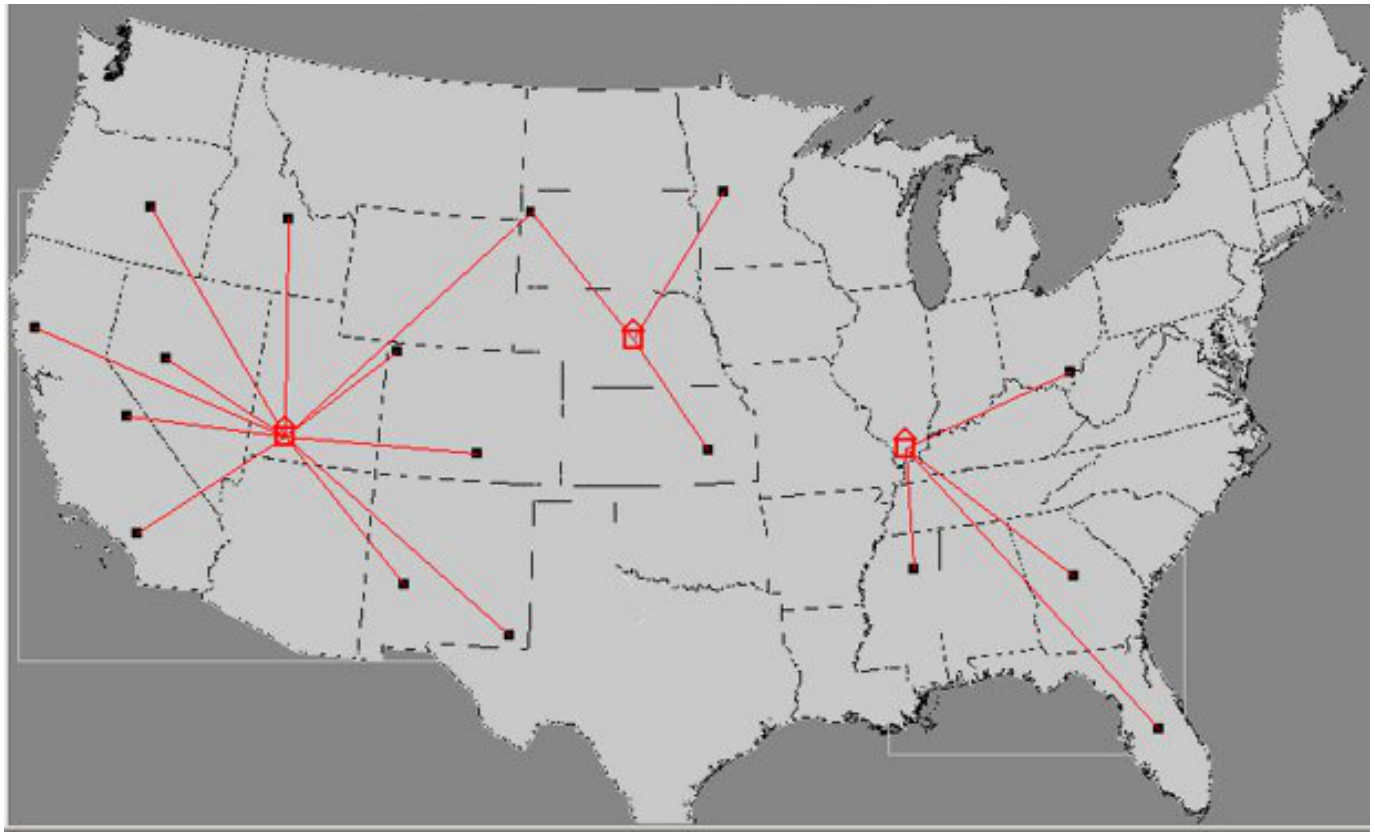
the mean value of $h(\mathbf{x},\xi_k),\ k = 1, ..., N$.

Thus, the final DRL problem becomes

$$\text{mimimize}_{\mathbf{x}\in X} \quad \sum_{k=1}^{N} \hat{p}_k h(\mathbf{x},\xi_k) + \frac{1}{\sqrt{N}} \sqrt{\sum_{k=1}^{N} \left(\frac{1}{N} \sum_{k=1}^{N} h(\mathbf{x},\xi_k) - h(\mathbf{x},\xi_k) \right)^2}$$

This is the original sample average objective plus the standard deviation of the samples.

Price of Correlation: Planning under High-Dimensional Stochastic Data



$$\min_{\mathbf{x}\in X} \ \mathbf{E}_p[f(\mathbf{x},\xi)]$$

where ξ represents a high-dimensional random vector with joint location/demand high-dimensional probability distribution p.

Curse of dimensionality and Price of Correlation

- Consider stochastic optimization problem

$$\min_{x \in X} \ \mathbf{E}_p[f(x, \xi)]$$

  where ξ is a high-dimensional random vector

- The common solution method is by Sample Average Approximation (SAA).

- However, to sample such a random vector, one suffers from the "curse of dimensionality"

Dimensionality	Required sample size
1	4
2	19
5	786
7	10,700
10	842,000

A Distributionally Robust Approach

We now consider the distributionally robust approach:

$$\min_{\mathbf{x}\in X} \max_{p\in\Gamma} \mathbf{E}_p[f(\mathbf{x},\xi)]$$

where Γ is the set of joint distributions such that the marginal distribution of ξ_i is p_i for each i.

This problem may be too complicated to solve, so that people are tempted to ignore correlations and assume independence among random variables...

However, what is the risk associated with assuming independence? Can we analyze this risk in terms of properties of objective functions?

We precisely quantify this risk as

Price of Correlations (POC)

We now provide tight bounds on POC for various classes of cost functions.

Price of Correlation I

Define

- Let $\hat{\mathbf{x}}$ be the optimal solution of stochastic program with independent distribution $\hat{p}(\xi) = \prod_i p_i(\xi_i)$.

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in X} \mathbf{E}_{\hat{p}}[f(\mathbf{x}, \xi)]$$

- Let $\mathbf{x}^*$ be the optimal solution for the distributionally robust model.

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in X} \max_{p\in\Gamma} \mathbf{E}_p[f(\mathbf{x}, \xi)]$$

Then, Price of Correlations (POC) is the ratio that $\hat{\mathbf{x}}$ achieves for distributionally robust model.

$$\text{POC} = \frac{\max_{p\in\Gamma} \mathbf{E}_p[f(\hat{\mathbf{x}}, \xi)]}{\max_{p\in\Gamma} \mathbf{E}_p[f(\mathbf{x}^*, \xi)]}$$

Price of Correlation II

- This is an approximation of DRO
  - Minimax stochastic program can be replaced by stochastic program with independent distribution to get approximate solution.
  - Often much easy to solve either by sampling or by other algorithmic techniques [e.g., Kleinberg et al. (1997), Möhring et al.(1999)]
- It captures "Value or Price of Information"
  - Small POC means it is not too risky to assume independence.
  - Large POC suggests the importance of investing more on information gathering and learning the correlations in the joint distribution.

Question: What function class has large POC? What function class has small POC?

Our first answer:

- Supermodularity leads to large POC
- Submodularity leads to small POC

Supermodularity leads to large POC

- For any fixed $\mathbf{x}$, function $f(\xi) := f(\mathbf{x}, \xi)$ is supermodular in random variable ξ
- Increasing marginal cost

$$\frac{\partial f(\xi)}{\partial \xi_i \partial \xi_j} \geq 0, \; \forall i \neq j$$

e.g., effects of increase in congestion as demand increases. For example, for convex $u(\cdot)$

$$f(\mathbf{x}, \xi) = u(\mathbf{c}^T \mathbf{x} - \xi^T \mathbf{x}).$$

- In worst case distribution large values of one variable will appear with large values of other variable – positively correlated among uncertain factors
- We can show by example of supermodular set function with POC $= \Omega(2^n)$.

Submodularity leads to small POC

- For any fixed $\mathbf{x}$, function $f(\xi) = f(\mathbf{x}, \xi)$ is submodular in random variable ξ.
- Decreasing marginal cost, economies of scale

$$f(\max\{\xi, \theta\}) + f(\min\{\xi, \theta\}) \leq f(\xi) + f(\theta)$$

- For continuous functions:

$$\frac{\partial f(\xi)}{\partial \xi_i \partial \xi_j} \leq 0, \ \forall i \neq j.$$

Theorem 10 *If $f(\cdot, \xi)$ is monotone and submodular in ξ, then*

$$\textit{POC} \leq e/(e-1).$$

[Calinescu, Chekuri, Pál, Vondrák, 2007] for binary random variables, [Agrawal, Ding, Saberi, Ye, 2010] for general random domains.

Example: Stochastic Bottleneck Matching

$$\min_{\mathbf{x}\in X} \max_{p\in\Gamma} \mathbf{E}_p[\max_i(\xi_i x_i)].$$

reduces to

$$\min_{\mathbf{x}\in X} \mathbf{E}_{\hat{p}}[\max_i(\xi_i x_i)]$$

where expected value is under independent distribution $\hat{p}$.

- This is a monotone submodular function, $e/(e-1) \sim 1.6$ approximation.
- Can be sampled efficiently, Chernoff type concentration bounds hold for monotone submodular functions.
- Reduces to a small convex optimization problem.

More general, we can solve

$$\min_{\mathbf{x}\in X} \mathbf{E}_{\hat{p}}[||\xi.^*\mathbf{x}||_q]$$

where expected value is under independent distribution $\hat{p}$. This a monotone submodular function, and we yield an $e/(e-1) \sim 1.6$ approximation.

Beyond Submodularity?

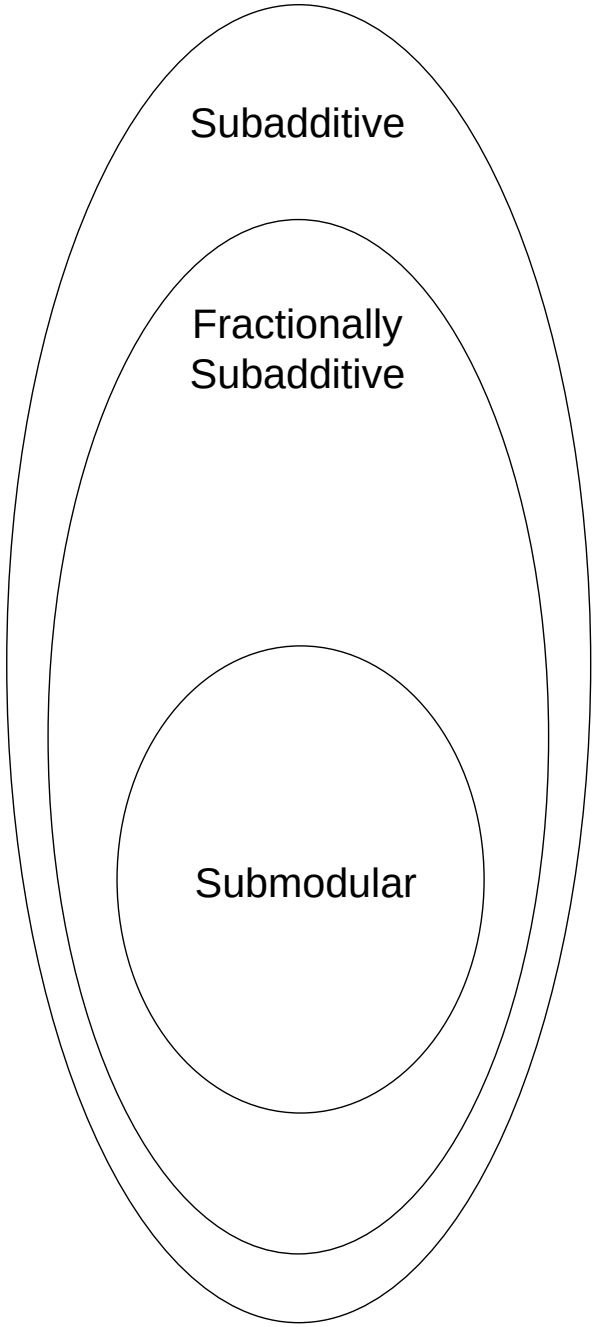
Monotone Subadditive functions?

- Preserves economy of scale
- Example with POC $= \Omega(\sqrt{n}/\log\log(n))$

Fractionally subadditive?

- POC $\geq \Omega(\sqrt{n}/\log\log(n))$

Cost-sharing to the rescue



$\geq \tilde{\Omega}(\sqrt{n})$

$\geq \tilde{\Omega}(\sqrt{n})$

$\frac{e}{(e-1)}$

Beyond Submodularity?

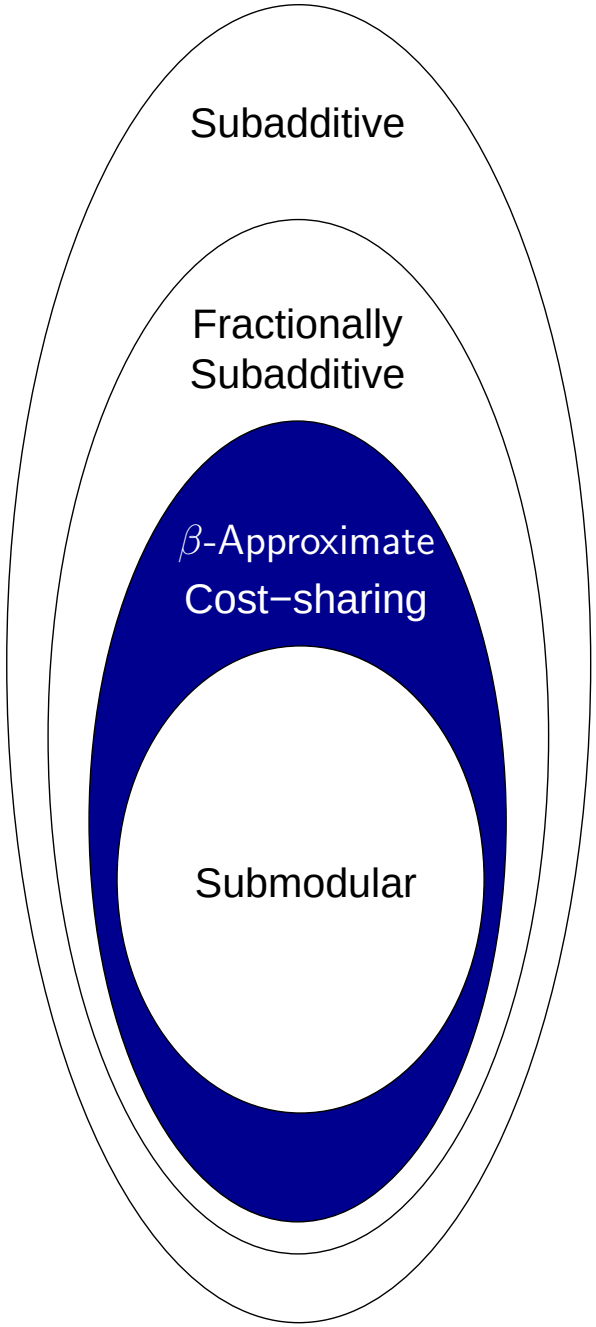
Monotone Subadditive functions?

- Preserves economy of scale
- Example with POC $= \Omega(\sqrt{n}/\log\log(n))$

Fractionally subadditive?

- POC $\geq \Omega(\sqrt{n}/\log\log(n))$

Cost-sharing to the rescue



$\geq \tilde{\Omega}(\sqrt{n})$

$\geq \tilde{\Omega}(\sqrt{n})$

2β Facility Location, Steiner Forest, Spanning tree

$\frac{e}{(e-1)}$

Cross-Monotone Cost-Sharing

A cooperative game theory concept

- Can cost $f(\xi_1, \ldots, \xi_n)$ be charged to participants $1, \ldots, n$ so that the share charged to participant i decreases as the demands of other participants increase?

  [introduced by Thomson (1983, 1995) in context of bargaining]

- For submodular functions – charge marginal costs.

- β-approximate cost-sharing scheme: total cost charged is within β of the original (expected) function value

Approximate cost-sharing schemes exist for non-submodular functions

- 3-approximate cost-sharing for facility location cost function [Pál, Tardos 2003]

- 2-approximate cost-sharing for Steiner forest cost function [Könemann, Leonardi, Schäfer 2005]

Bounding POC via Cost-Sharing and the Tightness

Theorem 11 *If objective function $f(\cdot, \xi)$ is monotone in ξ with β-cost-sharing scheme, POC $\leq 2\beta$.*

- POC ≤ 6 for two-stage stochastic facility location
- POC ≤ 4 for two-stage stochastic Steiner forest network design problem.

Theorem 12 *If correlation gap for function f is less than β, there exists a cross-monotone cost-sharing scheme with expected β-budget balance.*

- Monotone submodular function with POC $\geq \frac{e}{e-1}$.
- Facility location with POC ≥ 3.
- Steiner tree network design with POC ≥ 2.

[Agrawal, Ding, Saberi, Ye, 2010]

Sample Average Approximation with Quasi-Norm Regularization

To solve stochastic optimization $\min_{\mathbf{x}} \; E_{F_\xi} f(\mathbf{x}, \xi)$, we

- Solve instead an approximation problem

$$\min_{\mathbf{x}} \; F_N(\mathbf{x}) := \frac{1}{N} \sum_{j=1}^{N} f(\mathbf{x}, \xi_j)$$

  where $\{\xi_1, \xi_2, ..., \xi_j, ..., \xi_N\}$ is a sequence of samples of ξ, and it is simple to implement often tractably computable

- The DRO may be reduced to

$$\min_{\mathbf{x}} \; F_N(\mathbf{x}) := \frac{1}{N} \sum_{j=1}^{N} f(\mathbf{x}, \xi_j) + \lambda \|\mathbf{x}\|_*$$

  where $*$ is a suitable norm.

It has been show that the quasi-norm $\|\mathbf{x}\|_p, \; 0 < p < 1$ can be effective when $\mathbf{x}$ is sparse and the problem may be over-fitting.

Example: Sparse-Least-Squares with Quasi-Norm Regularization

Consider the problem:

$$\text{minimize}_x \quad f_p(\mathbf{x}) := \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_p^p \tag{7}$$

where data $A \in R^{m \times n}$, $\mathbf{b} \in R^m$, and parameter $0 \leq p < 1$.

$\|\mathbf{x}\|_p$ with $0 < p < 1$ is called quasi-norm of vector $\mathbf{x}$. When $p = 0$:

$$\|\mathbf{x}\|_0^0 := \|\mathbf{x}\|_0 := |\{j : \ x_j \neq 0\}|$$

that is, the number of nonzero entries in $\mathbf{x}$.

Related Problems: Constrained Quasi-Norm Minimization

$$\begin{array}{ll} \text{minimize} & \|\mathbf{x}\|_p^p = \sum_{1 \leq j \leq n} |x_j|^p \\ \text{subject to} & A\mathbf{x} = \mathbf{b}. \end{array} \quad \text{or} \quad \begin{array}{ll} \text{minimize} & p(\mathbf{x}) = \sum_{1 \leq j \leq n} x_j^p \\ \text{subject to} & A\mathbf{x} = \mathbf{b}, \\ & \mathbf{x} \geq \mathbf{0}, \end{array} \tag{8}$$

Theory of L_p Regularized Model I

Theorem 13 *(The first order bound) Let* $\mathbf{x}^*$ *be any local minimizer of (7) and*

$$\ell_j = \left(\frac{\lambda p}{2\|\mathbf{a}_j\|\sqrt{f_p(\mathbf{x}^*)}} \right)^{\frac{1}{1-p}},$$

where $\mathbf{a}_j$ *is the* j*th column of* A*. Then, the following property holds:*

$$\text{for each } j, \qquad x_j^* \in (-\ell_j, \ell_j) \quad \Rightarrow \quad x_j^* = 0.$$

Moreover, the number of nonzero entries in $\mathbf{x}^*$ *is bounded by*

$$\|\mathbf{x}^*\|_0 \leq \min\left(m, \frac{f_p(\mathbf{x}^*)}{\lambda \ell^p} \right);$$

where $\ell = \min\{\ell_j\}$.

[Chen et al, 2010]

Sketch of Proof

Let $\mathbf{x}^*$ be a local minimizer. Then it remains a minimizer after eliminating those variables whose values are zeros. For the nonzero-value variables, they must still satisfy the first-order KKT conditions:

$$2\mathbf{a}_j^T(A\mathbf{x}^* - \mathbf{b}) + \lambda p(|x_j^*|^{p-1} \cdot \text{sign}(x_j^*)) = 0.$$

Thus,

$$|x_j^*|^{1-p} \geq \frac{\lambda p}{2\|\mathbf{a}_j\|\|A\mathbf{x}^* - \mathbf{b}\|} \geq \frac{\lambda p}{2\|\mathbf{a}_j\|\sqrt{f_p(\mathbf{x}^*)}}.$$

Now we show the second part of the theorem. Again,

$$\lambda\|\mathbf{x}^*\|_p^p \leq \|A\mathbf{x}^* - \mathbf{b}\| + \lambda\|\mathbf{x}^*\|_p^p = f_p(\mathbf{x}^*).$$

From the first part of this theorem, any nonzero entry of $\mathbf{x}^*$ is bounded from below by ℓ so that we have the desired result.

Theory of L_p Regularized Model II

Theorem 14 *(The second order bound) Let* $\mathbf{x}^*$ *be any local minimizer of (7), and* $\kappa_j = \left(\frac{\lambda p(1-p)}{2\|\mathbf{a}_j\|^2}\right)^{\frac{1}{2-p}}, j \in \mathcal{N}$. *Then the following property holds:*

$$\text{for each } j, \qquad x_j^* \in (-\kappa_j, \kappa_j) \quad \Rightarrow \quad x_j^* = 0.$$

Again, we remove zero-value variables from $\mathbf{x}^*$ and the remain variables must still satisfy the second-order KKT condition for a local minimizer of (7):

$$\nabla^2 f_p(\mathbf{x}) = 2A^T A - \lambda p(1-p)\text{Diag}(|x_j^*|^{p-2}) \succeq \mathbf{0}.$$

Then all diagonal entries of the Hessian must be nonnegative, which gives the proof.

[Chen et al, 2010]

Theory of L_p Regularized Model III

- The first-order theorem indicates that the lower the objective value, the sparser the solution cardinality bound. Also, for λ sufficiently large but finite, the number of nonzero entries in any local minimizer reduces to 0.
- The result of the second-order theorem depends only on λ and p. In practice, one would typically choose $p = 1/2$.
- The two theorems establish relations between model parameters $p,\ \lambda$ and the desired degree of sparsity of the solution. In particular, it gives a guidance on how to choose the combination of λ and p.
- We can show that a second-order KKT solution of (7) would be relatively easy to compute, either in theory or practice.

Sample-Size Efficacy of SAA

- Assumptions: i.i.d., subgaussian and Lipschitz-like conditions.
- Number of samples N required to achieve ϵ accuracy with probability $1-\alpha$ in solving an d-dimensional problem:

$$P[f(\mathbf{x}^{SAA}) - f(\mathbf{x}^*) \leq \epsilon] \geq 1 - \alpha$$

if N is large enough to satisfy (Shapiro [2003] Stochastic Programming, Handbook in OR & MS and Shapiro et al. [2009] Lectures on Stochastic Programming Modeling and Theory)

$$N > \frac{d}{\epsilon^2} \ln \frac{1}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\alpha}$$

Sample size N grows polynomially when number of dimensions d increases.

Sample-Size Efficacy of SAA with Quasi-Norm-Type Regularization I

We made the following Assumptions:

(a) $f(\mathbf{x}, \xi)$ is subgaussian for any $\mathbf{x} \in [0,\ R]^d$

(b) $f(\cdot, \xi)$ is twice differentiable for almost every ξ

(c) Lipschitz-like condition

(d) The true solution is sparse, i.e., $\|\mathbf{x}^*\|_0 \leq s$

Assumptions (a) to (c) are standard.

Sample-Size Efficacy of SAA with Quasi-Norm-Type Regularization II

	$N >$	Solution-Type	$f(\cdot, \xi)$ convex	$E[f]$ strongly convex & differentiable	$\min_{i \in \mathcal{S}} x_i^{\min} \geq$ threshold
SAA	$\frac{d}{\epsilon^2} \ln \frac{1}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\alpha}$	Global	Not	Not	Not
RSAA	$\frac{s}{\epsilon^3} \left(\ln \frac{d}{\epsilon} \right)^{1.5} + \frac{1}{\epsilon^2} \ln \frac{1}{\alpha}$	Global	Not	Not	Not
RSAA	$\frac{s}{\epsilon^2} \ln \frac{d}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\alpha}$	Local	Required	Required	Required

- *In general:* poly-logarithmic dependence in d at the compromise on increased dependence in ϵ
- *A special case:* poly-logarithmic dependence in p at NO compromise on ϵ

[Liu et al, 2018]

Sparse Portfolio Selection: Quasi-Norm Regularization

Recall the modern portfolio selection problem:

$$\begin{aligned} \text{minimize} \quad & \mathbf{x}^T V \mathbf{x} \\ \text{subject to} \quad & \mathbf{r}^T \mathbf{x} \geq \mu, \\ & \mathbf{e}^T \mathbf{x} = 1, \ \mathbf{x} \geq \mathbf{0}, \end{aligned}$$

where expect-value vector $\mathbf{r}$ and co-variance matrix V are given, and $\mathbf{e}$ is the vector of all ones.

In shorting-allowed models, constraint $\mathbf{x} \geq \mathbf{0}$ is dropped; and it is replaced by $\|\mathbf{x}\|_1 \leq 1 + \delta$ for some $\delta > 0$, where δ controls the leverage of the portfolio.

But the final solution of the model are typically dense ...

A Quasi-Norm Regularized Model

We now consider

$$\begin{array}{rl} \text{minimize} & \mathbf{x}^T V \mathbf{x} + \mathbf{c}^T \mathbf{x} + \lambda \|\mathbf{x}\|_p \\ \text{subject to} & \mathbf{e}^T \mathbf{x} = 1,\ \mathbf{x} \geq \mathbf{0}, \end{array}$$

where we removed the linear expectation constraint for simplicity. Also for simplicity, we fix $p = 1/2$ in the analysis.

One may consider more complicated regularization model:

$$\begin{array}{rl} \text{minimize} & \mathbf{x}^T V \mathbf{x} + \mathbf{c}^T \mathbf{x} + \lambda \|\mathbf{x}\|_p \\ \text{subject to} & \mathbf{e}^T \mathbf{x} = 1,\ \|\mathbf{x}\| \leq 1 + \delta. \end{array}$$

Theory of the Quasi-Norm Regularized Model

Theorem 15 *(The second order theorem) Let* $\mathbf{x}^*$ *be any second-order KKT solution (after removing zero-value entries),* P^* *be the support of* $\mathbf{x}^*$ *and* $K = |P^*|$*, and* V^* *be the corresponding covariance sub-matrix. Furthermore, let*

$$\kappa_j = V^*_{jj} - \frac{2}{K}(V^*\mathbf{e})_j + \frac{1}{K^2}(\mathbf{e}^T V^* \mathbf{e}),\ j \in P^*,$$

which are the diagonal entries of matrix $\left(1 - \frac{1}{K}\mathbf{e}\mathbf{e}^T\right) V^* \left(1 - \frac{1}{K}\mathbf{e}\mathbf{e}^T\right)$*. Then the following properties hold:*

- $(K-1)K^{3/2} \le \frac{4}{\lambda}\sum_{j\in P^*}\kappa_j$.
- *If there is* $\kappa_j = 0$*, then* $K = 1$ *and* $x^*_j = 1$*; otherwise,*

$$x^*_j \ge \left(\frac{\lambda(1-\frac{1}{K})^2}{4\kappa_j}\right)^{2/3}.$$

We sketch the proof [Chen et al, 2013] next...

We only consider variables $j \in P^*$. The second-order condition requires that the Hessian of the Lagrangian function

$$V^* - \frac{\lambda}{4}\text{Diag}\left[(x_j^*)^{-3/2}\right]$$

must be positive semidefinite in the null space of $\mathbf{e} \in R^K$. Or, the projected Hessian matrix

$$\left(I - \frac{1}{K}\mathbf{e}\mathbf{e}^T\right)\left(V^* - \frac{\lambda}{4}\text{Diag}\left[(x_j^*)^{-3/2}\right]\right)\left(I - \frac{1}{K}ee^T\right) \succeq \mathbf{0},$$

must be positive semidefinite.

Thus, the jth diagonal entry of the projected Hessian matrix

$$\kappa_j - \frac{\lambda}{4}\left((x_j^*)^{-3/2}\left(1 - \frac{2}{K}\right) + \frac{\sum_k (x_k^*)^{-3/2}}{K^2}\right) \geq 0, \tag{9}$$

and the trace of projected Hessian matrix

$$\sum_k \kappa_k - \frac{\lambda}{4}\frac{K-1}{K}\sum_k (x_k^*)^{-3/2} \geq 0.$$

The quantity $\sum_k (x_k^*)^{-3/2}$, with $\sum_k x_k^* = 1,\ x_k^* \geq 0$ achieves its minimum at $x_k^* = 1/K$ for all k

with the minimum value $K \cdot K^{3/2}$. Thus,

$$\frac{\lambda}{4}(K-1)K^{3/2} \leq \sum_k \kappa_k, \quad \text{or} \quad (K-1)K^{3/2} \leq \frac{4\sum_k \kappa_k}{\lambda},$$

which complete the proof of the first item.

Again, from (9) we have

$$\frac{\lambda}{4}\left((x_j^*)^{-3/2}\left(1-\frac{2}{K}\right)+\frac{\sum_k (x_k^*)^{-3/2}}{K^2}\right) \leq \kappa_j.$$

Or

$$\frac{\lambda}{4}\left((x_j^*)^{-3/2}\left(1-\frac{1}{K}\right)^2+\frac{\sum_{k,k\neq j} (x_k^*)^{-3/2}}{K^2}\right) \leq \kappa_j,$$

which implies

$$\frac{\lambda}{4}(x_j^*)^{-3/2}\left(1-\frac{1}{K}\right)^2 \leq \kappa_j.$$

Hence, if any $\kappa_j = 0$, we must have $K = 1$ and x_j^* is the only non-zero entry in $\mathbf{x}^*$ so that $x_j^* = 1$. Otherwise, we have the desired second statement in the Theorem.