分布的ロバスト最適化モデリング
—解釈と実用への示唆—

後藤　順哉

本稿では分布的ロバスト最適化の定式化について紹介し，特に最悪ケースの期待値として与えられる目的関数の含意について再考する．またハイパーパラメータである不確実性集合の大きさに対する目的関数値の感度を導入し，正則化付き経験リスク最小化との関係やよく用いられる不確実性集合の違いを論じ，パラメータチューニングに対する示唆を示す．

キーワード：分布的ロバスト最適化，Worst-Case Sensitivity，偏差尺度，正則化

現実に現れる最適化問題では問題を記述するパラメータが推定値であるなど，最適化問題そのものに不確実性が含まれることが多い．一般に，最適化問題は与えられたパラメータに対する「最適」を目指すため，パラメータの想定値が実際値と異なっていれば得られる最適解の現実における最適性は担保されない．実際文献 [1] では，パラメータ値の（丸め誤差のような）わずかなズレが大きな実行不可能性に繋がる例を紹介している．このような，事前には認知が難しい事後での不整合に対し，最悪ケースのズレが起こったとしても最適解の実行可能性やある種の最適性を担保しようとする最適化モデリングが**ロバスト最適化**である．

本稿は著者の最近の論文 [2–4] に基づき，ロバスト最適化の中でも想定する確率分布の不確実性を考慮した**分布的ロバスト最適化**に焦点を絞り，その直感的な含意と，そこで用いられる不確実性の大きさを表すパラメータの選択について示唆を与える．詳細についてはそれら論文をご覧いただきたい．

1. 分布的ロバスト最適化

目的関数が（広義の）コストの期待値で表され，それを最小化する以下の最適化問題を考える．

$$\min_{\mathsf{x}} \ \mathbb{E}_{\mathbb{P}^\star}[f(\mathsf{x},\mathsf{Y})] \tag{1}$$

ここで $\mathbb{E}_{\mathbb{P}}[f]$ は確率分布 $\mathbb{P}$ の元での f の期待値を表し，$\mathbb{P}^\star$ は未知である真の分布を表す．f は決定変数（ベクトル）$\mathsf{x}\in\mathbb{R}^d$ だけでなく，確率変数（ベクトル）Y にも依存する関数である．Y が確率変数のため（各 x に対して）f も確率変数であり，f 自体を最小化することは適切でない．そこで最適化に際しては (1) のようにその期待値を最小化するのが基本である．(1) の形で定式化される問題はオペレーションズ・リサーチをはじめ，統計や機械学習における推定など色々な文脈で現れる．

例) 新聞売り子問題　在庫管理では費用が将来の需要や仕入れ量などに依存する．ここでは在庫管理のエッセンスを取り込んだ古典的な新聞売り子問題を考えよう．今ある新聞（のようにその日に売れないと価値が減じる商品）1 単位を原価 c で仕入れ，売れれば売値 r $(>c)$ だけ獲得し，売れなければ q $(<c)$ で引き取ってもらえるとする．また，売り切れた場合の機会費用を s (≥ 0) としたとき，新聞売り子のコスト（$\equiv(-1)\times$ 利益）は (2) のように表すことができる：

$$f(x,Y) = cx - \underbrace{r\min\{x,Y\}}_{\text{売上}} - \underbrace{q\max\{x-Y,0\}}_{\text{残余価値}} + \underbrace{s\max\{Y-x,0\}}_{\text{機会費用}} \tag{2}$$

ただし x (≥ 0) は仕入れ数を表す決定変数，Y が新聞の需要を表す確率変数である．新聞売り子問題は (2) をコストとした (1) を考える．$0\leq q<c<r$，$s\geq 0$ であれば，(1) は x について凸最小化問題であり，需要 Y の分布関数の逆関数がわかっている場合には解析的に最適解を与えることができる．

例) ポートフォリオ選択問題　手持ちの資金を d 種の株式銘柄に分散投資する問題を考えよう．特に投資家の効用関数が投資収益率 R に対して $U(R)=-\exp(-R)$ で与えられるとする．d 銘柄の株価収益率を並べた収

ごとう　じゅんや
中央大学理工学部
〒 112–8551 東京都文京区春日 1–13–27
jgoto@kc.chuo-u.ac.jp

益率ベクトルを Y，各銘柄への投資比率を並べたポートフォリオベクトルを x とするとおくと，この投資配分比率決定問題は

$$f(\mathsf{x},\mathsf{Y}) := \exp(-\mathsf{Y}^\top \mathsf{x})$$

をコストにした (1) として与えられる．ただし，x は $\mathbf{1}^\top \mathsf{x} = 1$ を満たす必要があるため，制約付き問題となる．

例) ロジスティック回帰　d 個の属性からなる入力ベクトル $\mathsf{a} \in \mathbb{R}^d$ から二値ラベル $b \in \{\pm 1\}$ が決まるとき，n 個の入出力の標本 $(\mathsf{a}_i, b_i) \in \mathbb{R}^d \times \{\pm 1\}$ からその決定規則を記述する線形モデル $b \leftarrow \mathsf{a}^\top \mathsf{x}$ を推定したい．この推定法の一つであるロジスティック回帰はロジスティック分布に対する最尤法として記述される．これは対数尤度を -1 倍した最小化であり，以下の f に対する (3) として記述される：

$$f(\mathsf{x},(\mathsf{a},b)) := \ln\left(1+\exp(-b\mathsf{a}^\top \mathsf{x})\right)$$

1.1　標本平均近似から分布的ロバスト最適化へ

このように，(1) のような期待コスト最小化問題は確率変数を内包する関数の最適化として頻繁に現れる．しかしながら多くの場合，そもそも真の分布 $\mathbb{P}^\star$ を知りえないため (1) は厳密な意味では実現できない最適化問題である．そこで，現実には (1) の目的関数を推定したり，確率変数 Y の標本 $\mathsf{Y}_1,...,\mathsf{Y}_n$ を分布の台として，それぞれに $1/n$ の確率を付した経験分布により (1) を直接近似したりすることになる．後者は**標本平均近似**（以降 SAA と略す）などと呼ばれ，以下のように実現できる．

$$\min_{\mathsf{x}} \left\{ \mathbb{E}_{\mathbb{P}}\big[f(\mathsf{x},\mathsf{Y})\big] := \frac{1}{n}\sum_{i=1}^{n} f(\mathsf{x},\mathsf{Y}_i) \right\} \tag{3}$$

ここで $\mathbb{P}$ は経験分布である．もし $\{\mathsf{Y}_1,...,\mathsf{Y}_n\}$ が独立に $\mathbb{P}^\star$ に従う標本であれば，標本数 n が大きくなるにつれ $\mathbb{P}$ は $\mathbb{P}^\star$ に収束していき，(3) の解 $\mathsf{x} = \mathsf{x}_{\mathrm{SAA}}$ は (1) の最適解 $\mathsf{x} = \mathsf{x}^\star$ に近づくことが期待できる．しかし現実には標本数 n は増やせない（今手元にあるもののみ利用可能である）ことが多い．実際，$\mathbb{E}_{\mathbb{P}^\star}[f(\mathsf{x}^\star,\mathsf{Y})]$ に比べ，$\mathbb{E}_{\mathbb{P}^\star}[f(\mathsf{x}_{\mathrm{SAA}},\mathsf{Y})]$ が著しく大きくなること（要するに SSA の解 $\mathsf{x}_{\mathrm{SAA}}$ の真の分布 $\mathbb{P}^\star$ の下での劣化）はよく観察される．このような事後パフォーマンスの劣化は一つには未知である真の分布 $\mathbb{P}^\star$ を，経験分布 $\mathbb{P}$ で置き換えたことに起因していると考えられる．

そこで，経験分布 $\mathbb{P}$ を信じすぎないように分布の不確実性を考慮しつつ，事後パフォーマンスの向上を目指す定式化の変更が考えられている．有名なものとして，機械学習などの分野でよく用いられる正則化付き SAA がある．これは (4) のように，SAA に加え正則化項と呼ばれる項も同時に小さくするように (3) を変更した形をとる：

$$\min_{\mathsf{x}} \ \frac{1}{n}\sum_{i=1}^{n} f(\mathsf{x},\mathsf{Y}_i) + \underbrace{C\mathsf{x}^\top \mathsf{x}}_{\text{正則化項}} \tag{4}$$

ここで $C \geq 0$ は標本データに依存しない正則化項の係数定数であり，これが大きいほど SAA の最小化への依存を小さくすることを意味する．$C = 0$ のとき (4) は (3) に一致することから，うまく C を決めることで，(4) の解は SAA 解 $\mathsf{x}_{\mathrm{SAA}}$ よりも事後パフォーマンスが良くなることが期待できる．

機械学習では標本を部分的に採用する，交差検証法やブートストラップ法といった再標本法を用いて疑似的に分布の不確実性をシミュレートすることで C を決定し，事後パフォーマンスの向上を図ることが多い．また正則化項では (4) の $\mathsf{x}^\top \mathsf{x}$ 以外にも，$\sum_{j=1}^{d} |x_j|$ など，いくつかバリエーションがある．一方で，正則化項は決定変数の自由度を制限することや変数ベクトルの非ゼロ要素数削減，計算の容易さが主眼であり，その関数形自体の意味は二の次といった観がある．

一方で, 本稿で扱う分布的ロバスト最適化（以降 **DRO**）は，経験分布 $\mathbb{P}$ の不確実性を集合で表し，その集合の中で最悪ケースの期待コストの最小化を目指す．具体的には，ありうる分布の集合を $\mathcal{Q}$ で表し，

$$\min_{\mathsf{x}} \max_{\mathbb{Q}\in\mathcal{Q}} \ \mathbb{E}_{\mathbb{Q}}[f(\mathsf{x},\mathsf{Y})] \tag{5}$$

なる最適化問題を考える．一般には x の制約条件に期待値を含む場合など色々拡張が考えられるが，本稿では目的関数にのみ期待値が含まれている状況を考える．

この時点では，(5) を考えることが，果たして $\mathsf{x}^\star$ に近づくのかどうか不明である．結論からいえば，多くの場合についてそれを期待する根拠はほとんどない．

さて，(5) で前提となる分布の集合 $\mathcal{Q}$ を本稿では**不確実性集合**と呼ぶ．$\mathcal{Q}$ は一般性を失うことなく凸集合に限定してよいが，依然大きな自由度がある．本稿では主として分布 $\mathbb{Q}$ の $\mathbb{P}$ からの差を測るダイバージェンス d を用い

$$\mathcal{Q}(\varepsilon) = \{\mathbb{Q} : d(\mathbb{Q}|\mathbb{P}) \leq \varepsilon\} \tag{6}$$

のように与えられるケースを考える．d としては ϕ-ダイバージェンスと Wasserstein 距離を用いるケースに

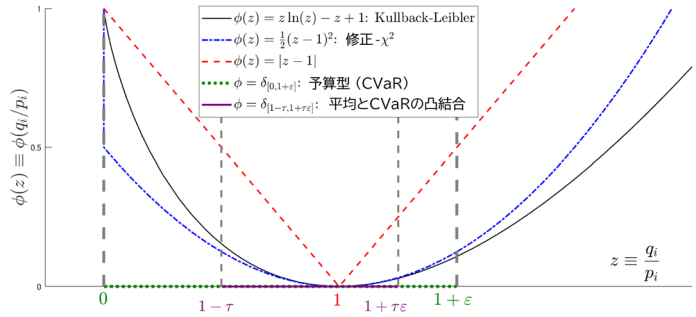

図 1　関数 ϕ の例

焦点を絞る．

ちなみに，(5) の目的関数部分である最悪コストを

$$\rho[f] := \max_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}_{\mathbb{Q}}[f] \tag{7}$$

と置くと，$\rho[f]$ は以下の性質を満たすことが簡単に確かめられる：

- 単調性：$f_1 \le f_2$ ならば $\rho[f_1] \le \rho[f_2]$
- 正 1 次同次性：$A > 0$ に対し $\rho[Af] = A\rho[f]$
- 劣加法性：$\rho[f_1 + f_2] \le \rho[f_1] + \rho[f_2]$
- 平行移動不変性：定数 B に対し $\rho[f+B] = \rho[f]+B$

このような性質は確率変数で表されるコスト f の忌諱度を定量化する関数として，金融工学ではコヒレントリスク尺度 [5] の公理として知られる．逆に，上記の性質を満たすリスク尺度は分布の集合 $\mathcal{Q}$ を用いて (7) のように最悪ケースの期待コストとして表すことができることが知られている．

代表的なコヒレントリスク尺度として知られるのが *CVaR* [6] であり，n 標本上の確率分布 p に対しては

$$\mathcal{Q}_{\mathrm{CVaR}}(\alpha) := \left\{\mathsf{q}\in\mathbb{R}^n \,\middle|\, \mathbf{1}^\top\mathsf{q} = 1,\ 0 \le \mathsf{q} \le \frac{1}{1-\alpha}\mathsf{p}\right\}$$

で与えられる $\mathcal{Q}$ によって特徴づけることができる．この $\mathcal{Q}$ について本稿では $\mathcal{Q}(\varepsilon) = \mathcal{Q}_{\mathrm{CVaR}}(\frac{\varepsilon}{1+\varepsilon})$ とパラメータを置き換え，予算型と呼ぶ．

1.2　ϕ-ダイバージェンス

$z = 1$ で最小値 0 をとる非負の凸関数 $\phi(z)$ に対して以下のように定義されるダイバージェンスを Csiszár の **ϕ-ダイバージェンス**と呼ぶ：

$$\mathcal{H}_\phi(\mathbb{Q}\,|\,\mathbb{P}) := \begin{cases} \sum_i p_i\,\phi\left(\frac{q_i}{p_i}\right), & \mathbf{1}^\top\mathsf{q} = 1, \mathsf{q} \ge 0, \\ +\infty, & \text{その他.} \end{cases}$$

定義より，$\frac{q_i}{p_i} = 1$ すなわち $q_i = p_i$ のとき $\phi(\frac{q_i}{p_i}) = 0$ であり，$q_i \ne p_i$ であれば $\phi(\frac{q_i}{p_i}) \ge 0$ となることから，q が p と異なるときのみ $\mathcal{H}_\phi(\mathbb{Q}\,|\,\mathbb{P})$ は正となりうる．$\phi(z) = z\ln z - z + 1$ のときは $\sum_i p_i\phi\left(\frac{q_i}{p_i}\right) = \sum_i q_i \ln\frac{q_i}{p_i}$ となり Kullback-Leibler ダイバージェンスに一致するなど，ϕ を変えることで色々なダイバージェンスを包含する（図 1）．ちなみに，ϕ は値域として $+\infty$ も含むことを許容することで，区間 I に対する指標関数

$$\delta_I(t) = \begin{cases} 0, & t \in I, \\ \infty, & t \notin I, \end{cases}$$

を用い，$I = [0, \frac{1}{1-\alpha}]$ とすれば，任意の正数 $\eta > 0$ に対して $\mathcal{Q}(\eta) = \mathcal{Q}_{\mathrm{CVaR}}(\alpha)$ となる．

また，$\tau \in [0,1]$ に対して $I = [1-\tau, 1+\tau(\frac{\alpha}{1-\alpha})]$ とすれば，(7) で $\mathcal{Q} = \mathcal{Q}_{\mathrm{c}}(\tau) := (1-\tau)\{\mathsf{p}\} + \tau\mathcal{Q}_{\mathrm{CVaR}}(\alpha)$ とした最悪コストになる．ちなみに $\mathcal{Q} = \mathcal{Q}_{\mathrm{c}}(\tau)$ のとき (7)($\rho(\mathsf{f}) \equiv \rho[f]$) は $\mathbb{P} \equiv \mathsf{p}$ の下での期待コスト $\mathbb{E}_{\mathsf{p}}(\mathsf{f})$ と $\mathrm{CVaR}_{\mathsf{p},\alpha}(\mathsf{f})$ の凸結合に一致する．

1.3　Wasserstein 距離

近年機械学習などで興味をもたれているのが，いわゆる輸送問題に着想を得た分布間の距離尺度である Wasserstein 距離である．本稿では分布の台を標本の値に限定しない連続型分布 $\mathbb{Q} \equiv \gamma(\cdot)$ と，離散型分布である経験分布 $\mathbb{P} \equiv \mathsf{p}$ の間の距離に限定してこれを考える．このとき Wasserstein 距離は以下のように表される：

$$d(\mathbb{Q}|\mathbb{P}) = \left| \begin{array}{ll} \min\limits_{\gamma} & \sum\limits_{i=1}^{n}\int_z \|z - Y_i\|_p\gamma_i(dz) \\ \text{s. t.} & \int_z \gamma_i(dz) = p_i,\ i = 1, ..., n, \\ & \gamma_i(dz) \ge 0 \end{array} \right.$$

図 2 に示すように，直感的には，$\mathbb{P}$ から $\mathbb{Q}$ へと確率を移動する輸送問題を考え，単位輸送費用が $\|z - Y_i\|_p$ で与えられた場合の最小輸送費用がダイバージェンス

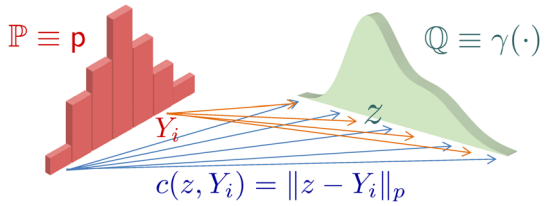

図 2　一般化した輸送問題としての Wasserstein 距離

$d(\mathbb{Q}|\mathbb{P})$ を与える．

ϕ-ダイバージェンスを用いた分布 $\mathbb{Q}$ の分布の台が $\mathbb{P}$ のそれに限定される（すなわち Y は実現値 $\mathsf{Y}_1, ..., \mathsf{Y}_n$ のいずれかであるの）に対し，Wasserstein 距離を用いると $\mathbb{Q}$ の分布の台は限定されない．現実の分布は過去の実現値以外で起こるのが普通であるし，連続分布を扱えることから Wasserstein 距離の方が（少なくとも分布の台について）柔軟性があり，現在盛んに研究がなされている．

たとえば文献 [7] に依れば，ロジスティック回帰に対し (a_i, b_i) から (a, b) への輸送費用を $c((\mathsf{a}_i, b_i), (\mathsf{a}, b)) = \|\mathsf{a}_i - \mathsf{a}\|_2 + \kappa|b_i - b|$ とする（ただし $\kappa \geq 0$ は属性 a に対するラベル b の輸送費用の相対的重み定数）と，ロジスティック回帰の DRO は凸計画問題に帰着できる：

$$\begin{aligned}
&\underset{\lambda, \mathsf{s}, \mathsf{x}}{\text{minimize}}\ \varepsilon\lambda + \tfrac{1}{n}\sum_{i=1}^{n} s_i \\
&\text{subject to} \\
&\qquad s_i \geq \ln\left(1 + \exp(-b_i \mathsf{a}_i^\top \mathsf{x})\right), \qquad i \in [n], \\
&\qquad s_i \geq \ln\left(1 + \exp(b_i \mathsf{a}_i^\top \mathsf{x})\right) - \kappa\lambda,\ i \in [n], \\
&\qquad \lambda \geq \|\mathsf{x}\|_2
\end{aligned}$$

ただし $\|\mathsf{x}\|_2 := \sqrt{\mathsf{x}^\top \mathsf{x}}$．特に $\kappa \to \infty$ のとき，属性の不確実性のみ考慮され，次の正則化付き SAA に一致する：

$$\underset{\mathsf{x}}{\text{minimize}}\ \varepsilon\|\mathsf{x}\|_2 + \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + \exp(-b_i \mathsf{a}_i^\top \mathsf{x})\right) \quad (8)$$

これは（適当な輸送費用の下）Wasserstein 距離を用いた DRO が正則付きロジスティック回帰と一致し，DRO が正則化の解釈を与えることを示している．

2.　最悪ケースの意味

最悪ケースの期待コスト (7) は不確実性集合 $\mathcal{Q}$ に依存し，$\mathcal{Q}$ の代表的な与え方にも色々あることを見てきた．素朴な疑問として $\mathcal{Q}$ の違いが最悪ケースにどのような含意をもたらすのだろうか？ 既に見たように，最悪ケースの期待コスト $\rho[f]$ はコヒレントリスク尺度であるので，ρ の形を比較するのが一つである．しかし DRO では $\mathcal{Q}$ が ε に依存しており，データや応用例ごとに ε のチューニングの必要が生じる．たとえば，$\mathcal{Q} = \mathcal{Q}_{\mathrm{CVaR}}(\alpha)$ とすれば $\rho[f]$ は CVaR であることを紹介したが，金融実務で CVaR でリスク計測をする状況でパラメータ α は 0.95 とか 0.99 といった数値（統計的検定における信頼水準のようなもの）で固定して用いるため，データに合わせるチューニングは通常行われない．その意味で，形式的には同じものを眺めているにも拘らず DRO におけるパラメータの決定は金融工学では対象となり難い問題であることを注意しておこう．

本節では，なぜ「最悪ケース」を考える必要があるのか？ あるいは，そもそも異なる $\mathcal{Q}$ 上での「最悪ケース」を考える含意は何なのかを考える．なお以降では簡単のため，経験分布を念頭に離散型の確率分布 $\mathbb{P} \equiv \mathsf{p}$ を考える．

2.1　滑らかな ϕ-ダイバージェンスの場合

はじめに $\phi''(1) > 0$ を満たすような，$z = 1$ 付近で狭義凸であるような ϕ を用いた ϕ-ダイバージェンスを用いて $\mathcal{Q}$ を定義する場合について考えよう．このとき，十分小さい $\varepsilon > 0$ に対し，以下の関係が成り立つ：

$$\mathbb{E}_{\mathsf{q}(\varepsilon)}(\mathsf{f}) = \mathbb{E}_{\mathsf{p}}(\mathsf{f}) + \sqrt{\varepsilon}\sqrt{\frac{2\mathbb{V}_{\mathsf{p}}(\mathsf{f})}{\phi''(1)}} + o(\sqrt{\varepsilon})$$

ここで，$\mathsf{q}(\varepsilon)$ は (7) を達成する最悪ケースの分布である．このことは十分滑らかな ϕ であればどんな ϕ-ダイバージェンスであっても，$\varepsilon > 0$ が十分小さければ，DRO の目的関数値 $\rho[f]$ は SAA と標準偏差の重み和で近似できることを示している．たとえば修正 χ^2-ダイバージェンス $\phi(z) = \frac{1}{2}(z-1)^2$ の場合は十分小さい ε に対し次式が成り立つ：

$$\mathbb{E}_{\mathsf{q}(\varepsilon)}(\mathsf{f}) = \mathbb{E}_{\mathsf{p}}(\mathsf{f}) + \sqrt{\varepsilon}\sqrt{2\mathbb{V}_{\mathsf{p}}(\mathsf{f})}$$

文献 [8] ではこの関係をベースに DRO に基づく推定を $f(\mathsf{x}, \mathsf{Y})$ の標準偏差を正則化項とした正則化付き経験的リスク最小化と見なして，事後的な期待コストに対する理論的な下限を示している．（ロジスティック回帰に Wasserstein 距離を用いた DRO では直接的に正則化項が得られたことも併せると）f の関数形にも依存するが，標準偏差が (4) の正則化項（の平方根）と似たような役割を果たすのではという期待も生じてくる．以降ではその関係が f に強く依存することを見る．

WCS　まず不確実性集合 $\mathcal{Q}$ が (6) のように与えられ

るとき，最悪ケースの目的関数 (7) が ε について単調増加な凹関数であることに注意する．そこで ε の増加に対する増加が最も大きい $\varepsilon = 0$ 付近の傾きを **Worst-Case Sensitivity**（以降 **WCS**）として定義する．具体的には凹で単調増加，かつ，$g(0) = 0$ であるような g に対し，$V(\varepsilon) - V(0) \sim O(g(\varepsilon))$ のとき，WCS を

$$\mathcal{S}_{\mathbb{P}}[f] = \lim_{\varepsilon \downarrow 0} \frac{\max\limits_{\mathbb{Q} \in \mathcal{Q}(\varepsilon)} \mathbb{E}_{\mathbb{Q}}[f] - \mathbb{E}_{\mathbb{P}}[f]}{g(\varepsilon)}$$

で定義する．このとき $\varepsilon > 0$ に対して

$$\max_{\mathbb{Q} \in \mathcal{Q}(\varepsilon)} \mathbb{E}_{\mathbb{Q}}[f] = \mathbb{E}_{\mathbb{P}}[f] + g(\varepsilon)\, \mathcal{S}_{\mathbb{P}}[f] + o(g(\varepsilon))$$

であり，ε が十分小さければ，DRO の目的関数が経験的な期待値と WCS の加重和で近似できることがわかる．

2.2 滑らかでない場合

次に $z = 1$ で微分ができないような例として $\phi(z) = |z - 1|$ であるような，$\mathbb{P} \equiv \mathsf{p}$ に対する ϕ-ダイバージェンスのケースを考えよう．このとき WCS は

$$\mathcal{S}_{\mathsf{p}}(\mathsf{f}) = \frac{1}{2}\big(\max(\mathsf{f}) - \min(\mathsf{f})\big)$$

となり，分布のレンジ（最大値と最小値の差）が WCS に対応している．

ϕ が変わったことで異なる WCS が得られたが，これら（標準偏差とレンジ）に共通しているのはいずれも f の分布のばらつきを定量化した概念であるということである．この概念を公理としてまとめたものが偏差尺度 [9] である：確率変数 f に対し以下を満たす（汎）関数 $\mathbb{D}$ を f の**偏差尺度** (generalized measure of deviation) と呼ぶ：

1. $\mathbb{D}[f] \geq 0$ であり，$\mathbb{D}[f] = 0$ と “f が定数” が等価
2. 任意の非負定数 $A \geq 0$ に対し $\mathbb{D}[Af] = A\mathbb{D}[f]$
3. 任意の定数 $B \in \mathbb{R}$ に対し $\mathbb{D}[f + B] = \mathbb{D}[f]$

標準偏差もレンジもこれらの性質を満たしていることが簡単に確認できる．一般的に WCS は偏差尺度である．

関数 ϕ が指標関数である例についても WCS を求めておく．まず CVaR と関連付けられる $\mathcal{Q} = \mathcal{Q}_{\mathrm{CVaR}}(\alpha)$ については WCS が以下のように計算される[1]:

[1] ちなみに最悪コストである CVaR は α について凹ではないが，$\varepsilon = \frac{\alpha}{1-\alpha}$ とパラメータを置き換える，すなわち $\mathcal{Q}(\varepsilon) = \mathcal{Q}_{\mathrm{CVaR}}(\frac{1}{1+\varepsilon})$ とすると，

$$\mathcal{Q}(\varepsilon) = \Big\{ \mathsf{q} \in \mathbb{R}^n \,\Big|\, \sum_{i=1}^{n} p_i \delta_{[0,1+\varepsilon]}\big(\frac{q_i}{p_i}\big) \leq 1,\ 1^{\top}\mathsf{q} = 1 \Big\}$$

となる．

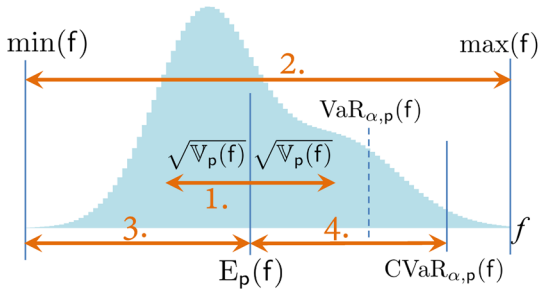

図 3　さまざまな偏差尺度
1. は標準偏差，2. はレンジ，3. は下半レンジ（期待値と最小値の差），4. は CVaR 偏差

$$\mathcal{S}_{\mathsf{p}}(\mathsf{f}) = \mathbb{E}_{\mathsf{p}}(\mathsf{f}) - \min(\mathsf{f})$$

これも偏差尺度であるが，コスト分布の左側の偏差を定量化したものといえる（図 3 の 3.）．

平均と CVaR の凸結合の WCS については τ についての感度を考えることで以下を得る：

$$\mathcal{S}_{\mathsf{p}}(\mathsf{f}) = \mathrm{CVaR}_{\mathsf{p},\alpha}(\mathsf{f}) - \mathbb{E}_{\mathsf{p}}(\mathsf{f})$$

ここで右辺は **CVaR 偏差**と呼ばれる偏差尺度で，上位 $100(1-\alpha)\%$ のコストの条件付期待値と（全体の）期待値の差にほぼ等しい（図 3 の 4.）．

2.3 Wasserstein DRO の WCS

Wasserstein DRO については ϕ-ダイバージェンスと異なり輸送費用 $c(\mathsf{z}, \mathsf{Y}_i) = \|\mathsf{z} - \mathsf{Y}_i\|_p$ に依存する．そのため WCS も輸送費用に依存する．またコスト $f(\mathsf{Y}) := f(\mathsf{x}, \mathsf{Y})$ が Y について，ある L が存在して任意の Z, Y に対し $f(\mathsf{Z}) - f(\mathsf{Y}) \leq L\|\mathsf{Z} - \mathsf{Y}\|_p$ とする．このとき WCS は

$$\mathcal{S}_{\mathbb{P}}[f] = \max_{i=1,\cdots,n} \max_{\mathsf{Z}_i} \frac{f(\mathsf{Z}_i) - f(\mathsf{Y}_i)}{\|\mathsf{Z}_i - \mathsf{Y}_i\|_p}.$$

すなわち，Z_i と Y_i の輸送費用あたりのコスト f の変化の最大のものが WCS となる．

1.3 節の最後に紹介したように，ロジスティック回帰に DRO を適用した場合は正則化付き SAA(8) が得られる．また，文献 [10] に依れば，期待収益率に対するポートフォリオ選択でもポートフォリオベクトルに対する正則化が得られることが主張されている．しかしながら，Wasserstein 距離を用いた DRO すべてがそのように単純な正則化に帰着されるわけではない．たとえば新聞売り子問題において $x \in (\min_i Y_i, \max_i Y_i)$ とする．このとき $p = 1$ の Wasserstein 距離を用いた場合の WCS は

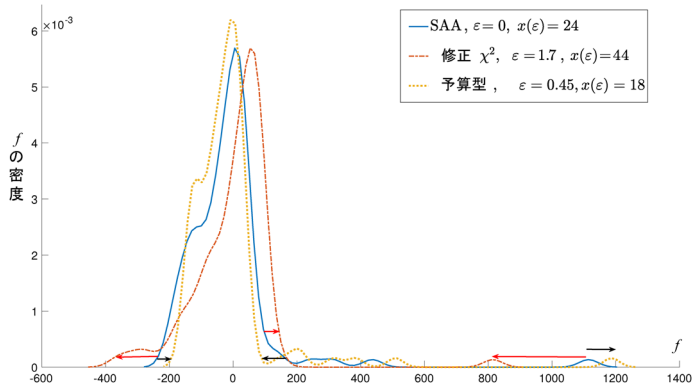

図 4　新聞売り子問題 ($s = 4$) におけるコストの分布

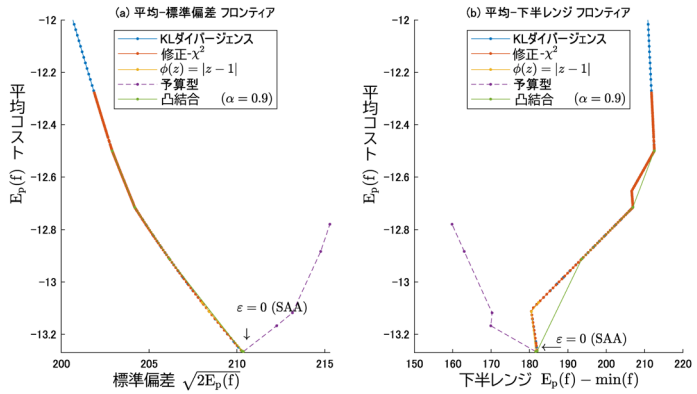

図 5　$s = 4$ の新聞売り子問題に対する二つの平均–感度プロット（左が滑らかな ϕ の感度，右が予算型の感度）

$$\mathcal{S}_{\mathbb{P}}[f(x,Y)] = \max\{r - q, s\}$$

すなわち，WCS が x に依存しない．結果，DRO と SAA の解が一致してしまう．つまり，新聞売り子問題に対して Wasserstein DRO は SAA の結果を「ロバスト」化しない．これらの結果より，Wasserstein DRO が正則化と関連づけられるかどうかや，SAA 解をロバスト化できるかどうかはコスト f の関数の形状といった適用対象の性質に依存することがわかる．

2.4　不確実性集合による差異

ここまでに，異なる $\mathcal{Q}$（あるいは ϕ）が異なる WCS を与えることを見た．ではそれらは明確な違いとなって現れるのだろうか? 図 4 は $s = 4$ である新聞売り子問題に対して SAA と，修正 χ^2 および予算型の ϕ-ダイバージェンスによる DRO で得られたコストの分布（を密度推定した結果）である．

修正 χ^2DRO がコスト 0 付近の密度が高い部分において SAA よりもばらつきを大きくしつつ，右裾にあるコブを左にシフトさせる，すなわち最大コストを下げる結果を導いているのに対し，予算型 DRO では右裾の最大コストを微増させつつも高密度部分のばらつきを小さくしている．これは前者の WCS である標準偏差が全体的なばらつきを捉え，全体の広がりを抑えようとするのに対し，後者の WCS がコスト分布の下側のばらつきを抑えようとしている違いから生じていると考えられる．

図 5 は同じの新聞売り子問題において，各 DRO の不確実性集合のサイズ ε を変えながらコストの平均値（縦軸）と 4 種類の WCS（横軸）をプロットしたものである．予算型 DRO のみ ε を大きくしていった際に，ほかの DRO と異なる方向を示していることがわかる．

一方，図 6 は $s = 0$ と変更した（その他のパラメータは不変の）場合についての同様のプロットである．図 5 と異なり，いずれの DRO でもほぼ同一の効率的フロンティアが描かれており，不確実性集合による大きな差異は見られない．

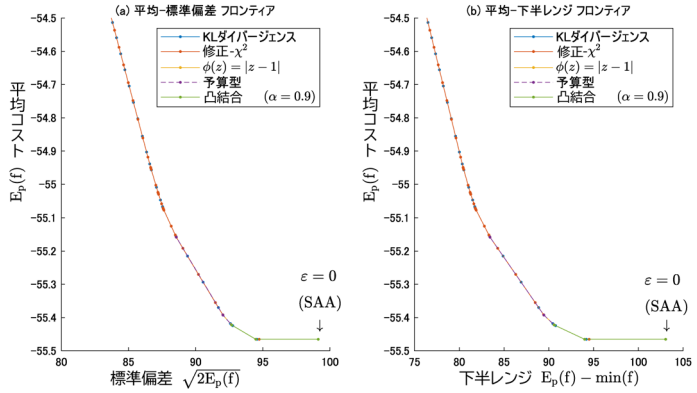

図 6　$s=0$ の新聞売り子問題に対する二つの平均–感度プロット（左が滑らかな ϕ の感度，右が予算型の感度）

このように同じ問題であってもパラメータが変わることでコスト f の関数形（x や Y に対する依存の仕方）が変わるために，DRO の振る舞いに差異が現れたり消滅したりすることがわかる．これらの観察からも，適用例（におけるコスト f）に応じて適切に不確実性集合（あるいは DRO の種類）を選ぶ必要があることが示唆される．その際に WCS は該当する DRO が考慮するロバスト性がどのようなものであるかを明示的に与えてくれるので一定のヒントになるかもしれない．

3.　不確実性パラメータの選択

仮に不確実性集合を選択したとしても，不確実性集合の大きさを表すハイパーパラメータ ε をどのように設定すべきか考える必要がある．以下では経験分布の標本によるばらつきを考慮した ε の決定法に関する理論的結果を紹介する．

分析の簡単化のため，ϕ が十分滑らか ($\phi''(1)>0$) なケースを考え，かつ，(5) の代わりに罰則型の DRO で，x について制約がない場合を考える：

$$\min_{\mathsf{x}} \max_{\mathbb{Q}} \left\{ \mathbb{E}_{\mathbb{Q}}[f(\mathsf{x},\mathsf{Y})] - \frac{1}{\delta}\mathcal{H}_\phi(\mathbb{Q}|\mathbb{P}) \right\} \tag{9}$$

ここで $\delta>0$ は ε に代わる罰則パラメータであり，ε と同様に，δ が大きいほど不確実性集合が大きいことを意味することに注意しておく．

$\mathsf{Y}_1,\dots,\mathsf{Y}_n$ を真の分布 $\mathbb{P}^\star$ からの i.i.d. 標本とし，この経験分布に基づく DRO の解を $\mathsf{x}(\delta)$ と記す．なお $\mathsf{x}(0)$ は SAA (3) の解に一致する．このとき，$\mathbb{P}^\star$ から新たに生成される Y_{n+1} に対する $\mathsf{x}(\delta)$ の振る舞いが興味の対象である．

DRO (9) の WCS は f の分散になり，δ が小さければ目的関数はコストの期待値と分散の 2 目的最適化で近似できることを示すことができる [3]．そこで，

- $\mu_n(\delta) := \mathbb{E}_{\mathbb{P}^\star}[f(\mathsf{x}(\delta),\mathsf{Y}_{n+1})]$　事後の期待コスト
- $v_n(\delta) := \mathbb{V}_{\mathbb{P}^\star}[f(\mathsf{x}(\delta),\mathsf{Y}_{n+1})]$　事後の WCS

とおくと，

$$\frac{\mathrm{d}\mu_n}{\mathrm{d}\delta}(0) = -\frac{\rho}{2n}, \tag{10}$$

$$\frac{\mathrm{d}v_n}{\mathrm{d}\delta}(0) = -\frac{2}{\phi''(1)}\beta^\top H(0)^{-1}\beta + \frac{\theta}{2n}, \tag{11}$$

が得られる．ここで ρ,θ は定数であり，

$$H(0) := \mathbb{E}_{\mathbb{P}^\star}[\nabla_{\mathsf{x}}^2 f(\mathsf{x}^\star(0),\mathsf{Y}_{n+1})],$$
$$\beta := \mathrm{Cov}_{\mathbb{P}^\star}[\nabla_{\mathsf{x}} f(\mathsf{x}^\star(0),\mathsf{Y}_{n+1}), f(\mathsf{x}^\star(0),\mathsf{Y}_{n+1})]$$

である．

ρ は正にも負にもなりえる．正であれば，（δ が小さいとき）DRO の方が SAA よりも低い期待コストを達成する，すなわち，DRO が SAA の結果を改善することが (10) より示唆される．

ただし，標本数 n で除されているため，標本が増えるほどその改善効果は消滅に向かう．一方，$f(\mathsf{x},\mathsf{Y})$ が x について狭義凸であれば，$H(0)$ は正定値であるため，$\beta^\top H(0)^{-1}\beta>0$ であり，$\phi''(1)>0$ と併せると (11) の第 1 項は負である．一方第 2 項は標本が増えれば消滅に向かう．したがって，標本数がある程度あれば，DRO は SAA よりも分散を小さくする一方，平均を改善する効果は限定的であることがわかる．これらのことから，DRO (9) において SAA に対応する $\delta=0$ から不確実性パラメータを増加させていった際，事後的な期待値の減少よりも，WCS である分散の減少の方が支配的になる．その意味で，DRO を採用するということの事後的なご利益は，期待コストを下げる効果

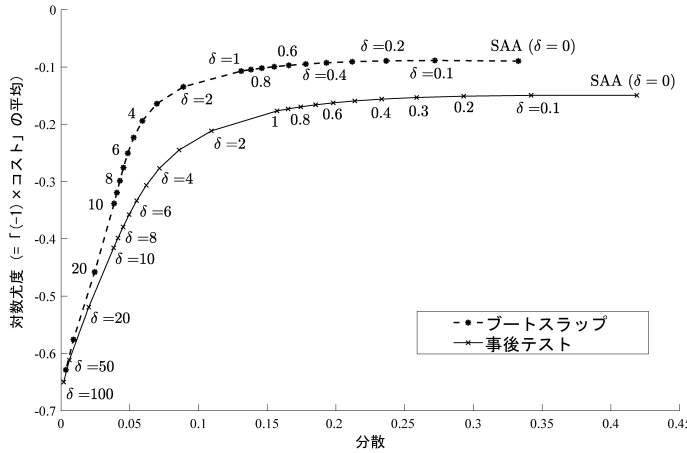

図 7　Kullback-Leibler ダイバージェンスによるロジスティック回帰の DRO に対するブートストラップ・シミュレーションと事後テストに対する平均・分散プロット
WDBC データセットの 30 ある属性のうち三つ（2, 24, 25 番目）の半分の標本に対してブートストラップ法を 20 回適用し，平均の平均と分散の平均のプロットと，得られた解に対して残りの半分の標本に基づく平均と分散のプロット

ではなく，ばらつきを抑える効果に見出せることがわかる．

図 7 はロジスティック回帰に対し DRO (9) を適用した場合にあるデータセットの半数の標本を用いたブートストラップ法によるコストの平均・分散プロットと，経験的な DRO の解を残りの半数に適用して計算される平均・分散プロットを示した例である．なおここでは尤度最大化の観点から縦軸は対数尤度である $(-1) \times f$ であり，上に行くほど尤度が大きく好ましいことを意味する．通常の対数尤度最大化である SAA ($\delta = 0$) を出発点とし δ を増加させたときに，対数尤度（垂直方向）がほとんど変化しない一方で，WCS である分散（水平方向）は大きく減少していること，そして，テスト標本においても同様の傾向が見て取れることがわかる．上記の理論は特に前者の現象が一般的に成り立つことを示している．

このように，DRO の定式化を吟味すると，パラメータ δ（あるいは ε）を選択するにあたっては，（正則化付き SAA (4) でよく採用されるような）コストの平均が最小になるような選択ではなく，ばらつきや WCS の減少を考慮に入れることが定式化に則していることがわかる．

謝辞　研究 [2, 3] は科研費 19H02379，19H00808，20H00285 の助成を部分的に受けて実施されています．

参考文献

[1] A. Ben-Tal and A. Nemirovski, “Robust solutions of linear programming problems contaminated with uncertain data,” *Mathematical Programming*, **88**, pp. 411–424, 2000.
[2] J. Gotoh, M. J. Kim and A. E. B. Lim, “Worst-case sensitivity,” *arXiv*（投稿中）.
[3] J. Gotoh, M. J. Kim and A. E. B. Lim, “Calibration of distributionally robust empirical optimization models,” *Operations Research*（掲載予定）.
[4] J. Gotoh, M. J. Kim and A. E. B. Lim, “Robust empirical optimization is almost the same as mean–variance optimization,” *Operations Research Letters*, **46**, pp. 448–452, 2018.
[5] P. Artzner, F. Delbaen, J. M. Eber and D. Heath, “Coherent measures of risk,”. *Mathematical Finance*, **9**, pp. 203–228, 1999.
[6] R. T. Rockafellar and S. Uryasev, “Optimization of conditional value-at-risk,” *Journal of Risk*, **2**, pp. 21–42, 2000.
[7] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani and D. Kuhn, “Distributionally robust logistic regression,” In *Advances in Neural Information Processing Systems*, pp. 1576–1584, 2015.
[8] J. C. Duchi and H. Namkoong, “Variance-based regularization with convex objectives,” *Journal of Machine Learning Research*, **20**, pp. 1–55, 2019.
[9] R. T. Rockafellar, S. Uryasev and M. Zabarankin, “Generalized deviations in risk analysis,” *Finance and Stochastics*, **10**, pp. 51–74, 2006.
[10] J. Blanchet, L. Chen and X. Y. Zhou, “Distributionally robust mean-variance portfolio selection with Wasserstein distances,” *arXiv preprint*, arXiv:1802.04885, 2018.