Solving Nonconvex Nonsmooth Compound Stochastic Programs with Applications to Risk Measure Minimization

Junyi Liu

Joint work with Jong-Shi Pang and Ying Cui

One World Optimization Seminar

March 8, 2022

Outline

- The setting: Compound stochastic programs
- Applications: Areas of risk measure minimization
- The SMM algorithm: Stochastic majorization minimization algorithm
- Convergence: Almost sure stationarity and probabilistic error bounds for stopping
- Extension: Risk-based robust statistical learning
- Numerical results: OCE-of-deviation optimization and robust statistical learning

The compound stochastic program

$$\boxed{\underset{x \in X}{\text{minimize}} \;\; \Theta(x) \triangleq \psi\left(\mathbf{E}\left[\, \varphi(G(x,\widetilde{\omega}),\, \mathbf{E}\left[\, F(x,\widetilde{\omega}) \,\right]) \,\right] \right)}$$

- a random variable $\widetilde{\omega} : \Xi \to \Omega \subseteq \mathbb{R}^m$, independent of x
- a closed convex set X contained in an open set $Y \subseteq \mathbb{R}^n$
- $G : Y \times \Omega \to \mathbb{R}^{\ell_G}$ and $F : Y \times \Omega \to \mathbb{R}^{\ell_F}$ are continuous, yet potentially nonconvex and nondifferentiable functions
- $\varphi : \mathbb{R}^{\ell_G + \ell_F} \to \mathbb{R}^{\ell_\varphi}$ and $\psi : \mathbb{R}^{\ell_\varphi} \to \mathbb{R}$ are isotone; ψ and $\{\varphi_j\}$ are convex;
- $\ell_F = 0 \Longrightarrow$ composite stochastic program with a single expectation

Application: elementary risk deviations

Deviation from the mean

- Expected **squared** deviation from the mean: $\mathbb{E}\left[Z - \mathbb{E}[Z]\right]^2$ (variance)
- Expected **absolute** deviation from the mean: $\mathbb{E}\left|Z - \mathbb{E}[Z]\right|$
- Expected **semi-**deviation from the mean: $\mathbb{E}\left[(Z - \mathbb{E}[Z])_+\right]$ where $(t)_+ = \max(0, t)$

Backgrounds: risk measures

Given a random variable Z with cumulative distribution function $F_Z(\bullet)$.

- α-Value-at-Risk (VaR) with $\alpha \in (0,1)$: $\mathrm{VaR}_\alpha(Z) \triangleq \min\{z : F_Z(z) \geq \alpha\}$
- τ-Probability Of Exceedance (POE) with $\tau \in \mathbb{R}$: $\mathrm{POE}(Z;\tau) \triangleq \mathbb{P}(Z > \tau) = 1 - F_Z(\tau)$.

Informally, $1-$ POE (distribution function) is the inverse of VaR (quantile function).

- α-Conditional Value-at-Risk[1] (CVaR) with $\alpha \in (0,1)$

$$\mathrm{CVaR}_\alpha(Z) \triangleq \frac{1}{1-\alpha}\int_{z \geq \mathrm{VaR}_\alpha(Z)} z dF_Z(z) = \min_{\eta \in \mathbb{R}} \left\{ \eta + \frac{1}{1-\alpha} \mathbf{E}\left[Z - \eta \right]_+ \right\}$$

- τ-buffered Probability Of Exceedance[2] (bPOE) with $\tau \in \mathbb{R}$ and $\tau \leq \sup(Z)$,

$$\mathrm{bPOE}(Z;\tau) \triangleq 1 - \min\left\{ \alpha \in (0,1) : \mathrm{CVaR}_\alpha(Z) \geq \tau \right\} = \underset{a \geq 0}{\text{minimum}}\ \mathbf{E}\left[a(Z - \tau) + 1 \right]_+$$

Informally, $1-$ bPOE (superdistribution function) is the inverse of CVaR (superquantile function).

[1] Rockafellar RT, Uryasev S (2000) Optimization of conditional value-at-risk. *Journal of Risk 2:2142.*

[2] Rockafellar RT, Royset JO (2010) On buffered failure probability in design and optimization of structures. *Reliability Engrg. System Safety 95(5):499510.*

Backgrounds: risk measures

CVaR is a type of utility-based Optimized Certainty Equivalent[1] (OCE)

Let $u : \mathbb{R} \to [-\infty, \infty)$ be a proper closed concave and nondecreasing utility function with $u(0) = 0$ and $1 \in \partial u(0)$.

$$S_u(Z) \triangleq \sup_{\eta \in \mathbb{R}} \left\{ \eta + \mathbb{E}\left[\, u(Z - \eta) \,\right] \right\} = \max \left\{ \eta + \mathbb{E}\left[\, u(Z - \eta) \,\right] \;\middle|\; \eta \in [z_{\min}, z_{\max}] \right\}$$

where $[\, z_{\min},\, z_{\max} \,]$ is the support interval of Z.

[1] Ben-Tal A, Teboulle M (2007) An old-new concept of convex risk measures: The optimized certainty equivalent. *Math. Finance* 17(3): 449476.

Application: generalized deviation minimization

- OCE based: Given a loss function $f(x,\omega)$,

$$
\begin{aligned}
& \text{minimize}_{x\in X} \quad -S_u\left(f(x,\widetilde{\omega})-\mathbb{E}[f(x,\widetilde{\omega})]\right) \\
\Longleftrightarrow\quad & \text{minimize}_{x\in X,\eta\in\mathbb{R}} \quad -\eta-\mathbb{E}\left[u\left(f(x,\widetilde{\omega})-\mathbb{E}\left[f(x,\widetilde{\omega})+\eta\right]\right)\right]
\end{aligned}
$$

- bPOE based: with the same loss function

$$
\begin{aligned}
& \underset{x\in X}{\text{minimize}} \quad \mathrm{bPOE}\left(f(x,\widetilde{\omega})-\mathbb{E}[f(x,\widetilde{\omega})];\tau\right) \\
\Longleftrightarrow\quad & \underset{x\in X,0\le a\le \bar{A}}{\text{minimize}} \quad \mathbb{E}\left[a\left(f(x,\widetilde{\omega})-\mathbb{E}[f(x,\widetilde{\omega})]-\tau\right)+1\right]_+
\end{aligned}
$$

Both involve compound expectations, an inner composition function φ, but without the outer function ψ.

$$
\boxed{\underset{x\in X}{\text{minimize}}\ \ \Theta(x)\triangleq\psi\left(\mathbf{E}\left[\varphi(G(x,\widetilde{\omega}),\mathbf{E}\left[F(x,\widetilde{\omega})\right])\right]\right)}
$$

Application: cost-based multiclass classification

Example: medical diagnosis classification

Suppose

- 5 levels of disease condition: $\{1, 2, 3, 4, 5\}$, the higher number represents the worse condition.
- groups of errors based on the gap between the true level and the categorized level

true \ classified	1	2	3	4	5
1	(1,1)				
2		(2,2)			
3			(3,3)		
4				(4,4)	
5					(5,5)

Errors in each group can result in similar costs, but the costs of errors among different groups could be significantly different.

Application: cost-based multiclass classification

Suppose

- attribute-class pairs: (X, Y) with $Y \in \{1, \ldots, M\}$
- scoring function: $h(X, \mu^m)$ for $m = 1, \cdots, M$
- **classifier:** an input X is classified into the class j if

$$j \in \operatorname{argmax}\{\, h(X, \mu^m) : m \in [\, M\,]\,\}.$$

- a set of misclassified label pairs: $T \triangleq \{\, (i, j) \in [\, M\,] \times [\, M\,] \mid i \neq j \,\}$.

  a label pair $(i, j) \in T$ means that a true label i is misclassified as $j \neq i$.

- a partition of $M \times (M-1)$ types of classification errors into S groups:

  $T = \bigcup_{k=1}^{S} T_k$, each T_k is associated an individual cost in learning the classification

Application: cost-based multiclass classification

- probability of misclassifying label i into label j with tolerance $\tau_{ij} \geq 0$:

$$\mathbb{P}\left(h(X, \mu^j) - \max_{m \in [M]} h(X, \mu^m) \geq -\tau_{i,j} \mid Y = i \right)$$

the probability of exceedance (POE) can be approximated by buffered probability of exceedance (bPOE) which considers the tail probability distribution.

- buffered cost-based classification problem

$$\underset{\{\mu^j\}_{j=1}^M}{\text{minimize}} \quad \sum_{s=1}^{S} \lambda_s \left\{ \max_{(i,j) \in T_s} \text{bPOE}\left(h(X^i, \mu^j) - \max_{m \in [M]} h(X^i, \mu^m); -\tau_{i,j} \right) \right\}.$$

Application: cost-based multiclass classification

By expressing each bPOE using its minimization formula in terms of an auxiliary variable, we can obtain a compound SP.

$$\operatorname*{minimize}_{\{\mu^j\}_{j=1}^M,\{a_{i,j}\}\geq 0} \quad \sum_{s=1}^{S} \lambda_s \left\{ \max_{(i,j)\in T_s} \mathbf{E} \left[\underbrace{a_{i,j} \left(h(X^i,\mu^j) - \max_{m\in[M]} h(X^i,\mu^m) + \tau_{i,j} \right) + 1}_{F_{i,j}(\{\mu^j\},a_{i,j},X^i)} \right]_+ \right\}.$$

Even when $h(X,\bullet)$ is a linear function, $F_{i,j}(\bullet,\bullet,X)$ is a nonconvex nonsmooth function.

Compound Stochastic Program

$$\underset{x \in X}{\text{minimize}} \ \Theta(x) \triangleq \psi\left(\mathbf{E}\left[\varphi(G(x,\widetilde{\omega}), \mathbf{E}\left[F(x,\widetilde{\omega})\right])\right]\right)$$

Current literature

• asymptotic and nonasymptotic statistical analysis of sample average approximation (SAA): [Ermoliev & Norkin, 2013], [Dentcheva, Penev, & Ruszczynski, 2015], [Hu, Chen & He, 2020],

• stochastic gradient-based algorithms under the smooth condition for all functions: [Wang, Liu, & Fang, 2016], [Wang, Fang, & Liu, 2017], [Ghadimi, Ruszczynski, & Wang, 2020]

• stochastic generalized subgradient-based algorithm for nonsmooth and nonconvex multi-level composite optimization: [Ruszczyski, 2021]

Challenges:

• computational challenge due to the coupled nonsmooth and nonconvex feature of G and F
• sampling strategies due to the compound structure

Stochastic Majorization-Minimization (SMM) Algorithm

- Surrogation

for every $x' \in X$ and $\omega \in \Omega$, there exists a family $\mathcal{G}(x', \xi)$ consisting of functions $\widehat{G}(\bullet, \xi; x')$ satisfying the following conditions:

(1) $\widehat{G}(x', \omega; x') = G(x', \omega)$;   (2) $\widehat{G}(x, \omega; x') \geq G(x, \omega)$ for any $x \in X$;
(3) each $\widehat{G}_i(\bullet, \omega; x')$ for $i = 1, \cdots, \ell_G$ is convex on X;
(4) uniform outer semicontinuity: a technical assumption for the convergence

- Sampling

incrementally discretize the nested expectations with independent sample sets $\{\xi^t\}_{t=1}^N$, $\{\eta^s\}_{s=1}^N$

The sampling-based surrogate objective

$$\boxed{\widehat{V}_N(x; x') \triangleq \psi\left(\frac{1}{N}\sum_{t=1}^{N}\left[\varphi\left(\widehat{G}^t(x, \xi^t; x'), \frac{1}{N}\sum_{s=1}^{N}[\widehat{F}^s(x, \eta^s; x')]\right)\right]\right).}$$

Example of convex surrogation function

difference-of-convex functions

Suppose $G(x,\omega) = g(x,\omega) - h(x,\omega)$ with $g(\bullet,\omega)$ and $h(\bullet,\omega)$ being convex functions.

For any given $x' \in X$ and $\omega \in \Omega$, we can construct the convex surrogate family $\mathcal{G}(x',\omega)$:

$$\mathcal{G}(x',\omega) = \left\{ \begin{array}{l} \widehat{G}(\bullet,\omega;x') : \widehat{G}(x,\omega;x') \triangleq g(x,\omega) - \underbrace{\Big(h(x',\omega) + a(x',\omega)^\top (x - x')\Big)}_{\text{linearization of } h(\bullet,\omega) \text{ at } x'} \\ \text{with } a(x',\omega) \in \partial_x h(x',\omega) \end{array} \right\}$$

Main iteration in SMM algorithm

For $\nu = 1, 2, \ldots,$ **do**

given the current iterate x^ν, and the current sample sets $\{\xi^t\}_{t=1}^{N_{\nu-1}}$ and $\{\eta^s\}_{s=1}^{N_{\nu-1}}$

1. sample generation

generate i.i.d samples $\left\{\xi^{N_{\nu-1}+t}\right\}_{t=1}^{\Delta_\nu}$ and $\left\{\eta^{N_{\nu-1}+s}\right\}_{s=1}^{\Delta_\nu}$, update $N_\nu \triangleq N_{\nu-1} + \Delta_\nu$,

2. sampling-based convex surrogate function

$$\widehat{V}_{N_\nu}(x; x^\nu) \triangleq \psi\left(\frac{1}{N_\nu}\sum_{t=1}^{N_\nu}\varphi\left(\widehat{G}^{\,t}(x, \xi^t; x^\nu),\ \frac{1}{N_\nu}\sum_{s=1}^{N_\nu}\widehat{F}^{\,s}(x, \eta^s; x^\nu)\right)\right)$$

3. the new iterate

$$x^{\nu+1} \triangleq \underset{x \in X}{\operatorname{argmin}}\left\{\widehat{V}_{N_\nu}(x; x^\nu) + \frac{1}{2\rho}\|x - x^\nu\|^2\right\}$$

Subsequential convergence theorem of SMM

Theorem Under technical conditions and sample sizes $N_\nu = \lceil \nu^{\alpha} \rceil$ for some $\alpha > 1$, for every limit point x^∞ of the sequence $\{x^\nu\}$ produced by the SMM algorithm, there exists $\widehat{G}(\bullet, \omega; x^\infty) \in \mathcal{G}(x^\infty, \omega)$ and $\widehat{F}(\bullet, \omega; x^\infty) \in \mathcal{F}(x^\infty, \omega)$ exist such that with probability 1,

$$x^\infty \in \operatorname*{argmin}_{x \in X} \; \psi\left(\mathbf{E}\left[\varphi\left(\widehat{G}(x, \widetilde{\omega}; x^\infty), \mathbf{E}\left[\widehat{F}(x, \widetilde{\omega}; x^\infty) \right] \right) \right] \right);$$

i.e., x^∞ is a fixed point of the algorithmic map:

$$x' \mapsto \operatorname*{argmin}_{x \in X} \; \psi\left(\mathbf{E}\left[\varphi\left(\widehat{G}(x, \widetilde{\omega}; x'), \mathbf{E}\left[\widehat{F}(x, \widetilde{\omega}; x') \right] \right) \right] \right).$$

Key steps in proof:

- a descent property, with errors, of the sequence of objective values
- finiteness of the accumulated errors through a proper control of the sample sizes

Post-convergence: connections of fixed points to stationarity

- The smooth case. $F_j(\bullet,\omega)$ and $G_i(\bullet,\omega)$ are smooth functions with the Lipschitz gradient modulus κ uniformly for all $\omega \in \Omega$,

$$\begin{aligned} \widehat{F}_j(x,\omega;x') &= F_j(x,\omega) + \frac{\kappa}{2}\,\|\,x - x'\,\|^2 \\ \widehat{G}_i(x,\omega;x') &= G_i(x,\omega) + \frac{\kappa}{2}\,\|\,x - x'\,\|^2. \end{aligned}$$

- The difference-of-convex case.

$$G_i(x,\omega) = g_i^G(x,\omega) - h_i^G(x,\omega), \quad \text{and} \quad F_j(x,\omega) = g_j^F(x,\omega) - h_j^F(x,\omega),$$

with $g_i^G(\bullet,\omega)$, $h_i^G(\bullet,\omega)$, $g_j^F(\bullet,\omega)$ and $h_j^F(\bullet,\omega)$ convex; moreover, $h_i^G(\bullet,\omega)$ and $h_j^F(\bullet,\omega)$ are additionally differentiable with Lipschitz gradient moduli independent of ω.

fixed-point property $\Longrightarrow$ directional stationarity

Error bounds and stopping rules

- Existing approach assessing the solution quality in SP include bounding the optimality gap, or testing the KarushKuhnTuckers conditions ([Higle and Sen, 1991], [Bayraksan and Morton, 2006], [Shapiro et al., 2009])

- For a coupled nonconvex and nondifferentiable SP, we need an error bound of the kind: there exists a constant C, for all test vectors of interest, with high probability,

$$\boxed{\text{distance to stationarity} \leq C \cdot \text{computable residual}.}$$

- with such error bound,

residual is small $\Rightarrow$ distance to stationarity is small, (with high probability)

Stochastic error bound to the compound SP

- Let $\mathcal{M}_{\widehat{V}_N}(\cdot) : \bar{x} \mapsto \underset{x \in X}{\operatorname{argmin}} \widehat{V}_N(x;\bar{x}) + \dfrac{1}{2\rho} \| x - \bar{x} \|^2$ be an algorithmic map
- Let $S^{\rm C}_{X,\,\Theta} \triangleq \{x : 0 \in \partial_C \Theta(\bar{x}) + \mathcal{N}(\bar{x};X)\}$, the set of Clarke stationary points

Theorem. Assume: (1) $S^C_{X,\Theta} \subseteq FIX(\mathcal{M}_{\widehat{V}})$; (2) local error bound; (3) upper bound of the surrogate functions. For any $\varepsilon \in (0,\bar{\varepsilon})$ and $\alpha \in (0,1)$, for all $\widehat{x} \in X$, provided that

$$N \geq \frac{C_1}{\varepsilon^2}\left(n \log\left(\frac{C_2}{\varepsilon}\right) + \log\left(\frac{1}{\alpha}\right)\right),$$

$$\mathbb{P}\left(\underbrace{\text{dist}\left(\widehat{x}; S^C_{X,\Theta}\right)}_{\text{distance to stationarity}} \leq \widehat{\eta} \underbrace{\| \widehat{x} - \mathcal{M}_{\widehat{V}_N}(\widehat{x}) \|}_{\text{sample-based residual}} + \varepsilon \right) \geq 1 - \alpha.$$

Extension: robust regression

- **robust M-estimator**[1]: min $\sum_{i=1}^{N} \rho(f(\theta, X^i) - Y^i)$ with the robust loss function ρ, such as the ℓ_1 loss, Huber's loss, etc.
- **Trimmed M-estimator**[2]: minimize the average of h smallest losses.

computational challenges:

- a nonconvex and nonsmooth optimization problem, even for the linear regression model
- in heuristic algorithms [3], the size of the subproblems is in the order of the data size and the property of the obtained solution is not guaranteed
- [Aravkin A, Davis D, 2020] proposes a stochastic proximal-gradient algorithm by reformulating the problem as a nonconvex optimization problem with a simplex constraint set, under the smooth condition of the loss function

[1] W. Li, and J. J. Swetits. The linear ℓ_1 estimator and the Huber M-estimator. *SIAM Journal on Optimization*, 1998

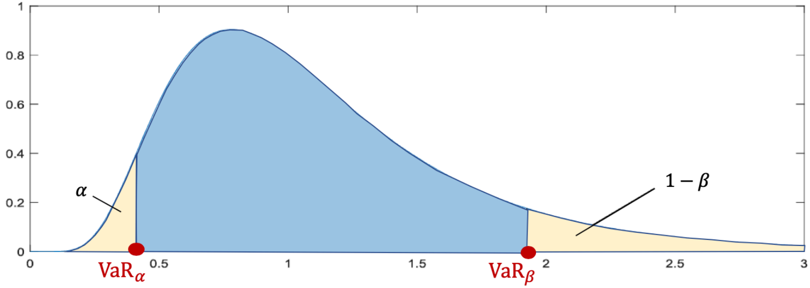
[2] PJ. Rousseeuw. Least median of squares regression. *Journal of the American statistical association.* 1984.

[3] PJ. Rousseeuw, K. Van Driessen. Computing LTS regression for large data sets. *Data mining and knowledge discovery*. 2006 .

Interval CVaR (In-CVaR)[1]

for $0 \leq \alpha < \beta \leq 1$,

$$\text{In-CVaR}_{\alpha}^{\beta}(Z) \triangleq \frac{1}{\beta - \alpha} \int_{\mathrm{VaR}_{\alpha}(Z) > z \geq \mathrm{VaR}_{\beta}(Z)} z dF_Z(z) = \frac{1-\alpha}{\beta-\alpha}\mathrm{CVaR}_{\alpha}(Z) - \frac{1-\beta}{\beta-\alpha}\mathrm{CVaR}_{\beta}(Z)$$



[1] Tsyurmasto P, Uryasev S, Gotoh JY (2013) Support vector classification with positive homogeneous risk functionals. Technical report.

Robust regression with In-CVaR

$$\boxed{\underset{\theta}{\text{minimize}} \quad \ell(\theta) \triangleq \left(\begin{array}{l} \lambda \, \text{In-CVaR}_{\alpha_1}^{\beta_1}(\max\{f(\theta,X)-Y,0\}) \\ +(1-\lambda) \, \text{In-CVaR}_{\alpha_2}^{\beta_2}(\max\{-f(\theta,X)+Y,0\}) \end{array} \right)}$$

Idea:

- excessively large residuals are excluded in model fitting

- excessively small residuals are excluded in model fitting

- over-estimation and under-estimation errors are evaluated with asymmetrical levels (α_1, β_1), (α_2, β_2) and asymmetrical weights $(\lambda, 1-\lambda)$.

Robust regression with In-CVaR

$$\boxed{\underset{\theta}{\text{minimize}} \quad \ell(\theta) \triangleq \left(\begin{array}{l} \lambda \, \text{In-CVaR}_{\alpha_1}^{\beta_1}(\max\{f(\theta,\widetilde{X}) - \widetilde{Y}, 0\}) \\ +(1-\lambda)\, \text{In-CVaR}_{\alpha_2}^{\beta_2}(\max\{-f(\theta,\widetilde{X}) + \widetilde{Y}, 0\}) \end{array} \right)}$$

- **Assumption:**

$f(\cdot, X)$ is a difference-of-convex function, $f(\theta, X) = g(\theta, X) - h(\theta, X)$ where g and h are convex functions.

This class of regression functions includes

- linear regression: $f(\theta, X) = \theta^\top X$
- piecewise affine regression: $f(\theta, X) = \max\{\theta_{1,i}^\top X + \theta_{0,i} : i \in \mathcal{I}\} - \max\{\theta_{2,j}^\top X + \theta_{0,j} : j \in \mathcal{J}\}$
- 2-layer neural network with ReLu: $f((A, a, b, \beta), X) = \max\{b^\top \max\{Ax + a, 0\} + \beta, 0\}$

Robust classification with In-CVaR

In binary classification, the attribute $X \in \mathbb{R}^n$, a binary response $Y \in \{1, -1\}$, a discriminant function $f(\theta, X)$.

$$\boxed{\underset{\theta}{\text{minimize}} \quad \text{In-CVaR}_{\alpha}^{\beta}(r(Y \cdot f(\theta, X))) + R(\theta)}$$

- The loss function $r(\cdot)$ could be:
  the hinge loss function $r_{\text{hinge}}(u) = \max\{1-u, 0\}$;
  the logistic loss function $r_{\text{logistic}}(u) = \log(1+\exp(-u))$.
- The discriminant function $f(\cdot, X)$ could be any difference-of-convex function.

The parameter estimation problem: In-CVaR based robust regression

Reformulation of the loss function $\ell(\theta)$

Since $\text{In-CVaR}_{\alpha}^{\beta}(Z) = \dfrac{1-\alpha}{\beta-\alpha}\,\mathrm{CVaR}_{\alpha}(Z) - \dfrac{1-\beta}{\beta-\alpha}\,\mathrm{CVaR}_{\beta}(Z)$, we reformulate the loss function

$$
\begin{aligned}
\ell(\theta) = \quad & \kappa_1\,\mathrm{CVaR}_{\alpha_1}(\max\{f(\theta,\widetilde{X})-\widetilde{Y},0\}) - \kappa_2\,\mathrm{CVaR}_{\beta_1}(\max\{f(\theta,\widetilde{X})-\widetilde{Y},0\}) \\
& +\kappa_3\,\mathrm{CVaR}_{\alpha_2}(\max\{-f(\theta,\widetilde{X})+\widetilde{Y},0\} - \kappa_4\,\mathrm{CVaR}_{\beta_2}(\max\{-f(\theta,\widetilde{X})+\widetilde{Y},0\})
\end{aligned}
$$

where $\kappa_1 = \dfrac{\lambda(1-\alpha_1)}{\beta_1-\alpha_1}, \kappa_2 = \dfrac{\lambda(1-\beta_1)}{\beta_1-\alpha_1}, \kappa_3 = \dfrac{\lambda(1-\alpha_2)}{\beta_2-\alpha_2}, \kappa_4 = \dfrac{\lambda(1-\beta_2)}{\beta_2-\alpha_2}$.

The parameter estimation problem: In-CVaR based robust regression

Reformulation of the loss function $\ell(\theta)$

Since $\mathrm{CVaR}_\beta(Z) = \min_{\eta\in\mathbb{R}} \left\{ \eta + \frac{1}{1-\beta}\,\mathbf{E}\left[\, Z - \eta \,\right]_+ \right\}$, and $f(\theta, X) = g(\theta, X) - h(\theta, X)$,

with some algebraic operations, we can derive that

$$\ell(\theta) = \left(\min_{\eta_1,\eta_2} \mathbf{E}\Big[\underbrace{\varphi_1(\theta,\eta_1,\eta_2;X,Y)}_{\text{jointly convex function}} \Big] \right) - \left(\min_{\eta_3,\eta_4} \mathbf{E}\Big[\underbrace{\varphi_2(\theta,\eta_3,\eta_4;X,Y)}_{\text{jointly convex function}} \Big] \right)$$

$$\begin{aligned}
\varphi_1(\theta,\eta_1,\eta_2;X,Y) \triangleq\ & \kappa_1\,\eta_1 + \frac{\kappa_1}{1-\alpha_1}\max\{g(\theta,X) - Y - \eta_1, h(\theta,X) - \eta_1, h(\theta,X)\} \\
& + \kappa_2\,\eta_2 + \frac{\kappa_2}{1-\alpha_2}\max\{h(\theta,X) + Y - \eta_2, g(\theta,X) - \eta_2, g(\theta,X)\}, \\
\varphi_2(\theta,\eta_3,\eta_4;X,Y) \triangleq\ & \kappa_3\,\eta_3 + \frac{\kappa_3}{1-\beta_1}\max\{g(\theta,X) - Y - \eta_3, h(\theta,X) - \eta_3, h(\theta,X)\} \\
& + \kappa_4\,\eta_4 + \frac{\kappa_4}{1-\beta_2}\max\{h(\theta,X) + Y - \eta_4, g(\theta,X) - \eta_4, g(\theta,X)\}.
\end{aligned}$$

The parameter estimation problem: In-CVaR based robust regression

Lemma

Let $v(\bullet) : \Theta \subseteq \mathbb{R}^d \to \mathbb{R}$ *with* $v(\theta) \triangleq \min_{\eta \in \Gamma} \psi(\theta, \eta)$ *where* Γ *is a convex set,* $\psi(\theta, \eta)$ *is a jointly convex function on* $\Theta \times \Gamma$*. Then* $v(\bullet)$ *is a convex function. Furthermore, for any* $\theta \in \Theta$*, we have conv* $\left\{ \partial_\theta \psi(\theta, \eta) : \eta \in \underset{\eta \in \Gamma}{\operatorname{argmin}}\, \psi(\theta, \eta) \right\} \subseteq \partial\, v(\theta)$*.*

To compute the parameter θ, we aim to solve a **difference-of-convex (DC) program**

$$\underset{\theta \in \Theta}{\text{minimize}} \quad \ell(\theta) = \underbrace{\left(\min_{\eta_1, \eta_2} \mathbf{E}\left[\varphi_1(\theta, \eta_1, \eta_2; X, Y) \right] \right)}_{u(\theta):\ \text{a convex function of } \theta} - \underbrace{\left(\min_{\eta_3, \eta_4} \mathbf{E}\left[\varphi_2(\theta, \eta_3, \eta_4; X, Y) \right] \right)}_{v(\theta):\ \text{a convex function of } \theta}$$

In principle we could solve such problem by the classical **difference-of-convex algorithm** [1]

[1] S. Fujiwara, A. Takeda, T. Kanamori, DC Algorithm for Extended Robust Support Vector Machine, *Neural Computation*, 29(5), 2017

Stochastic difference-of-convex algorithm (SDCA)

$$\underset{\theta\in\Theta}{\text{minimize}}\ \ell(\theta) = \underbrace{\left(\min_{\eta_1,\eta_2}\ \mathbf{E}\left[\varphi_1(\theta,\eta_1,\eta_2;X,Y)\right]\right)}_{u(\theta)\,:\ \text{a convex function of } \theta} - \underbrace{\left(\min_{\eta_3,\eta_4}\ \mathbf{E}\left[\varphi_2(\theta,\eta_3,\eta_4;X,Y)\right]\right)}_{v(\theta)\,:\ \text{a convex function of } \theta}$$

Approximations of u and v:

- u is approximated by $u_\nu(\theta)$, utilizing a random subset of samples of size N_ν

$$\boxed{u_\nu(\theta) = \min_{\eta_1,\eta_2} \frac{1}{N_\nu}\sum_{s=1}^{N_\nu} \varphi_1(\theta,\eta_1,\eta_2;X^s,Y^s)}$$

- v is approximated by $\hat{v}_\nu(\theta;\theta^\nu)$, the sampling-based linear approximation function

$$\eta_3^\nu,\eta_4^\nu \in \text{argmin}\ \frac{1}{N_\nu}\sum_{s=1}^{N_\nu}\varphi_2(\theta^\nu,\eta_3,\eta_4;X^s,Y^s), \quad \text{and} \quad a_{\nu,s}\in\partial\,\varphi_2(\theta^\nu,\eta_3^\nu,\eta_4^\nu;X^s,Y^s)$$

$$\boxed{\hat{v}_\nu(\theta;\theta^\nu) = \frac{1}{N_\nu}\sum_{s=1}^{N_\nu}\varphi_2(\theta^\nu,\eta_3^\nu,\eta_4^\nu;X^s,Y^s) + \langle \frac{1}{N_\nu}\sum_{s=1}^{N_\nu} a_{\nu,s},\theta-\theta^\nu\rangle}$$

- accumulating sampling strategy with the appropriate control of the incremental sample sizes

Stochastic Difference-of-Convex Algorithm (SDCA)

For $\nu = 1, 2, \ldots,$ **do**

1. random sample generation

generate i.i.d. samples $\{(X^{N_{\nu-1}+s}, Y^{N_{\nu-1}+s})\}_{s=1}^{\Delta_\nu}$, set $N_\nu = N_{\nu-1} + \Delta_\nu$.

2. solve the second inner convex subproblem

compute $\eta_3^\nu, \eta_4^\nu \in \operatorname{argmin} \dfrac{1}{N_\nu} \displaystyle\sum_{s=1}^{N_\nu} \varphi_2(\theta^\nu, \eta_3, \eta_4; X^s, Y^s)$

select a subgradient $a_{\nu,s} \in \partial_\theta\, \varphi_2(\theta^\nu, \eta_3^\nu, \eta_4^\nu; X^s, Y^s)$ for $s = 1, \ldots, N_\nu$

3. solve outer convex subproblem

$$\theta^{\nu+1} = \operatorname*{argmin}_\theta \left\{ u_\nu(\theta) - \hat{v}_\nu(\theta; \theta^\nu) + \frac{1}{2\rho}\|\theta - \theta^\nu\|^2 \right\}$$

Convergence of SDCA

Theorem

Under some technical assumptions and $\{N_\nu\}$ satisfying $N_\nu = \lceil \nu \rceil^\gamma$ for $\gamma > 1$, every accumulation point of the sequence $\{\theta^\nu\}$ is **a critical point** *almost surely.*

SDCA is a convergent algorithm with a proper control of increasing rate of the sample size, which is stronger than the sample size requirement $N_\nu = C\nu$ in [Le Thi et al. (2020)] because of the value-function structure in the ln-CVaR based robust regression model.

Numerical experiment: OCE-of-Deviation optimization

With the exponential utility function,

$$\underset{x\in[0,8]}{\text{minimize}}\ \mathbb{E}\Big[\exp\Big\{-(x-\tilde{\xi}\,)^2+\mathbb{E}[\,(x-\tilde{\xi}\,)^2\,]\Big\}\Big].$$

Table: Comparisons between the SMM algorithm and NASA algorithm[1]

algorithm	initialization range	iteration number	sample size	mean	std	running time
SMM	$[0,\ 8]$	5	16	1.0847	0.0932	0.3613
		10	31	1.0635	0.0210	0.8009
		15	54	1.0545	0.0083	1.5682
NASA	$[0,\ 8]$	50	50	1.6355	0.8684	0.0012
		500	500	1.3131	0.2439	0.0120
		5000	5000	1.2771	0.1140	0.1410
NASA	$[3,\ 5]$	50	50	1.1522	0.0960	0.0013
		500	500	1.0960	0.0556	0.0157
		5000	5000	1.0890	0.0412	0.1142

[1] Ghadimi S, Ruszczynski A, Wang M (2020) A single timescale stochastic approximation method for nested stochastic optimization. *SIAM J. Optim. 30(1):960979*.

Numerical experiment: robust regression

ground truth model: $\phi(x_1, x_2) = \max\{x_1 - 2x_2, -2x_1 + x_2 + 1\} - \max\{3x_1 + 2x_2, 2x_1\}$

Table: Performance of OLS, Huber and In-CVaR based estimators

$(p_0, \underline{\varepsilon}, \overline{\varepsilon})$	Δ_ν	iteration number	sample size	model	MAE (%)	MOE (%)	MUE (%)	running time (s)
(0.1, 3, 5)	10	30	325	**In-CVaR**	**8.91**	**9.47**	**8.20**	12.77
				Huber	11.7	13.3	4.69	8.06
				OLS	33.6	35.4	3.33	9.72
	10	50	525	**In-CVaR**	**7.32**	**9.51**	**4.82**	18.03
				Huber	12.1	12.7	8.00	15.53
				OLS	33.7	34.1	2.58	11.53
	$\lceil \nu^{0.5} \rceil$	50	285	**In-CVaR**	**5.95**	**5.34**	**6.39**	14.57
				Huber	7.26	5.64	9.04	13.4
				OLS	9.30	11.3	5.75	9.95
	k	30	490	**In-CVaR**	**4.38**	**4.66**	**4.07**	7.81
				Huber	7.59	8.36	4.81	9.38
				OLS	19.1	19.6	4.92	4.79

Summary

- Nonconvex and nonsmooth compound stochastic programs have many applications: generalized deviation optimization, cost-based multi-class classfication, robust statistical learning, etc. Extensions include multi-layer compound SP and conditional SP.
- We develop the stochastic majorization minimization (SMM) algorithm based on the surrogate family of functions for obtaining the fixed point of the algorithmic map.
- We establish a stochastic error bound, which theoretically justifies the probabilistic stopping rule for the SMM algorithm.

This talk is based on the following work:

[1] Liu J, Pang J-S, Cui Y (2022). Solving Nonsmooth Nonconvex Compound Stochastic Programs with Applications to Risk Measure Minimization. *Mathematics of Operations Research.*
[2] Liu J, Pang J-S (2022) *Risk-based robust statistical learning by stochastic difference-of-convex value-function optimization. Operations Research.*

Thank you!