Recursive Optimization of Mean-Semideviation Risk Measures with Variable Assessment

Dionysios S. Kalogerias
Joint work with Warren B. Powell

June 2018

Department of Operations Research & Financial Engineering (ORFE)

PRINCETON UNIVERSITY

This material is based upon work supported by the U.S. Navy / SPAWAR Systems Center Pacific under Contract No. *N66001-18-C-4031*. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Navy / SPAWAR Systems Center Pacific.

Problem Formulation: Mean-Semideviation Risk Measures

Base Problem

- We consider risk-averse optimization problems of the form

$$\inf_{x \in \mathcal{X}} \underbrace{\mathbb{E}\left\{F\left(x, W\right)\right\}}_{\text{risk-neutral}} + \underbrace{c\left[\mathbb{E}\left\{\left[\mathcal{R}\left(F\left(x, W\right) - \mathbb{E}\left\{F\left(x, W\right)\right\}\right)\right]^p\right\}\right]^{1/p}}_{\text{risk-averse}}, \tag{1}$$

for fixed risk penalty multiplier $c \geq 0$, order $p \in [1, \infty)$, and where

- $F : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}$ is convex in its first argument.
- $\widetilde{F}\left(x, \cdot\right) \equiv F\left(x, W\left(\cdot\right)\right) \in \mathcal{L}_p\left(\Omega, \mathscr{F}, \mathcal{P}; \mathbb{R}\right) \triangleq \mathcal{Z}_p$.
- $\mathcal{R} : \mathbb{R} \to \mathbb{R}$ acts on the central deviation $F\left(\cdot, W\right) - \mathbb{E}\left\{F\left(\cdot, W\right)\right\}$, satisfying:
  - S1 $\mathcal{R}$ is convex.
  - S2 $\mathcal{R}$ is nonnegative.
  - S3 $\mathcal{R}$ is nondecreasing.
  - S4 For every $\alpha \geq 0$, it is true that $\mathcal{R}\left(x + \alpha\right) \leq \mathcal{R}\left(x\right) + \alpha$, for all $x \in \mathbb{R}$. (nonexpansive)
    Any function satisfying conditions S1 − S4 is suggestively called a risk regularizer.
- The feasible set $\mathcal{X} \subseteq \mathbb{R}^N$ is closed and convex.

Base Problem

- We consider risk-averse optimization problems of the form

$$\inf_{x \in \mathcal{X}} \underbrace{\mathbb{E}\left\{F\left(x, W\right)\right\}}_{\text{risk-neutral}} + \underbrace{c\left[\mathbb{E}\left\{\left[\mathcal{R}\left(F\left(x, W\right) - \mathbb{E}\left\{F\left(x, W\right)\right\}\right)\right]^p\right\}\right]^{1/p}}_{\text{risk-averse}}, \tag{1}$$

for fixed risk penalty multiplier $c \geq 0$, order $p \in [1, \infty)$, and where

- $F : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}$ is convex in its first argument.
- $\widetilde{F}\left(x, \cdot\right) \equiv F\left(x, W\left(\cdot\right)\right) \in \mathcal{L}_p\left(\Omega, \mathscr{F}, \mathcal{P}; \mathbb{R}\right) \triangleq \mathcal{Z}_p$.
- $\mathcal{R} : \mathbb{R} \to \mathbb{R}$ acts on the central deviation $F\left(\cdot, W\right) - \mathbb{E}\left\{F\left(\cdot, W\right)\right\}$, satisfying:
  - S1 $\mathcal{R}$ is convex.
  - S2 $\mathcal{R}$ is nonnegative.
  - S3 $\mathcal{R}$ is nondecreasing.
  - S4 For every $\alpha \geq 0$, it is true that $\mathcal{R}\left(x + \alpha\right) \leq \mathcal{R}\left(x\right) + \alpha$, for all $x \in \mathbb{R}$. (nonexpansive)
    Any function satisfying conditions S1 − S4 is suggestively called a risk regularizer.
- The feasible set $\mathcal{X} \subseteq \mathbb{R}^N$ is closed and convex.

Base Problem

- We consider risk-averse optimization problems of the form

$$\inf_{x \in \mathcal{X}} \underbrace{\mathbb{E}\left\{F\left(x, W\right)\right\}}_{\text{risk-neutral}} + \underbrace{c\left[\mathbb{E}\left\{\left[\mathcal{R}\left(F\left(x, W\right) - \mathbb{E}\left\{F\left(x, W\right)\right\}\right)\right]^p\right\}\right]^{1/p}}_{\text{risk-averse}}, \tag{1}$$

for fixed risk penalty multiplier $c \geq 0$, order $p \in [1, \infty)$, and where

- $F : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}$ is convex in its first argument.
- $\widetilde{F}\left(x, \cdot\right) \equiv F\left(x, W\left(\cdot\right)\right) \in \mathcal{L}_p\left(\Omega, \mathscr{F}, \mathcal{P}; \mathbb{R}\right) \triangleq \mathcal{Z}_p$.
- $\mathcal{R} : \mathbb{R} \to \mathbb{R}$ acts on the central deviation $F\left(\cdot, W\right) - \mathbb{E}\left\{F\left(\cdot, W\right)\right\}$, satisfying:
  - S1 $\mathcal{R}$ is convex.
  - S2 $\mathcal{R}$ is nonnegative.
  - S3 $\mathcal{R}$ is nondecreasing.
  - S4 For every $\alpha \geq 0$, it is true that $\mathcal{R}\left(x + \alpha\right) \leq \mathcal{R}\left(x\right) + \alpha$, for all $x \in \mathbb{R}$. (nonexpansive)
    Any function satisfying conditions S1 − S4 is suggestively called a risk regularizer.
- The feasible set $\mathcal{X} \subseteq \mathbb{R}^N$ is closed and convex.

Base Problem

- We consider risk-averse optimization problems of the form

$$\inf_{x \in \mathcal{X}} \underbrace{\mathbb{E}\left\{F\left(x, W\right)\right\}}_{\text{risk-neutral}} + \underbrace{c\left[\mathbb{E}\left\{\left[\mathcal{R}\left(F\left(x, W\right) - \mathbb{E}\left\{F\left(x, W\right)\right\}\right)\right]^p\right\}\right]^{1/p}}_{\text{risk-averse}}, \tag{1}$$

for fixed risk penalty multiplier $c \geq 0$, order $p \in [1, \infty)$, and where

- $F : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}$ is convex in its first argument.
- $\widetilde{F}\left(x, \cdot\right) \equiv F\left(x, W\left(\cdot\right)\right) \in \mathcal{L}_p\left(\Omega, \mathscr{F}, \mathcal{P}; \mathbb{R}\right) \triangleq \mathcal{Z}_p$.
- $\mathcal{R} : \mathbb{R} \to \mathbb{R}$ acts on the central deviation $F\left(\cdot, W\right) - \mathbb{E}\left\{F\left(\cdot, W\right)\right\}$, satisfying:
  - S1 $\mathcal{R}$ is convex.
  - S2 $\mathcal{R}$ is nonnegative.
  - S3 $\mathcal{R}$ is nondecreasing.
  - S4 For every $\alpha \geq 0$, it is true that $\mathcal{R}\left(x + \alpha\right) \leq \mathcal{R}\left(x\right) + \alpha$, for all $x \in \mathbb{R}$. (nonexpansive)
    Any function satisfying conditions S1 − S4 is suggestively called a risk regularizer.
- The feasible set $\mathcal{X} \subseteq \mathbb{R}^N$ is closed and convex.

Base Problem

- We consider risk-averse optimization problems of the form

$$\inf_{x \in \mathcal{X}} \underbrace{\mathbb{E}\left\{F\left(x, W\right)\right\}}_{\text{risk-neutral}} + \underbrace{c\left[\mathbb{E}\left\{\left[\mathcal{R}\left(F\left(x, W\right) - \mathbb{E}\left\{F\left(x, W\right)\right\}\right)\right]^p\right\}\right]^{1/p}}_{\text{risk-averse}}, \tag{1}$$

for fixed risk penalty multiplier $c \geq 0$, order $p \in [1, \infty)$, and where

- $F : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}$ is convex in its first argument.
- $\widetilde{F}\left(x, \cdot\right) \equiv F\left(x, W\left(\cdot\right)\right) \in \mathcal{L}_p\left(\Omega, \mathscr{F}, \mathcal{P}; \mathbb{R}\right) \triangleq \mathcal{Z}_p$.
- $\mathcal{R} : \mathbb{R} \to \mathbb{R}$ acts on the central deviation $F\left(\cdot, W\right) - \mathbb{E}\left\{F\left(\cdot, W\right)\right\}$, satisfying:
  - S1 $\mathcal{R}$ is convex.
  - S2 $\mathcal{R}$ is nonnegative.
  - S3 $\mathcal{R}$ is nondecreasing.
  - S4 For every $\alpha \geq 0$, it is true that $\mathcal{R}\left(x + \alpha\right) \leq \mathcal{R}\left(x\right) + \alpha$, for all $x \in \mathbb{R}$. (nonexpansive)
    Any function satisfying conditions S1 − S4 is suggestively called a risk regularizer.
- The feasible set $\mathcal{X} \subseteq \mathbb{R}^N$ is closed and convex.

Mean-Semideviations

- The objective of (1) evaluates the *risk-measure*

$$\rho(Z) \triangleq \mathbb{E}\{Z\} + c\left(\mathbb{E}\left\{\left(\mathcal{R}\left(Z - \mathbb{E}\{Z\}\right)\right)^p\right\}\right)^{1/p}, \quad Z \in \mathcal{Z}_p, \tag{2}$$

  at $Z \equiv F(\boldsymbol{x}, W) \in \mathcal{Z}_p$, for $\boldsymbol{x} \in \mathbb{R}^N$.

- ρ strictly generalizes the p-th order mean-upper-semideviation.
  - Recovered by choosing $\mathcal{R}(\cdot) \equiv (\cdot)_+ \triangleq \max\{\cdot, 0\}$ in (2), satisfying conditions **S1** – **S4**.
- Problem (1) may be written compactly as

$$\inf_{x \in \mathcal{X}} \left\{\rho\left(F(x, W)\right) \triangleq \phi^{\widetilde{F}}(x)\right\}.$$

Mean-Semideviations

- The objective of (1) evaluates the *risk-measure*

$$\rho(Z) \triangleq \mathbb{E}\{Z\} + c\left(\mathbb{E}\left\{\left(\mathcal{R}\left(Z - \mathbb{E}\{Z\}\right)\right)^p\right\}\right)^{1/p}, \quad Z \in \mathcal{Z}_p, \tag{2}$$

  at $Z \equiv F(\boldsymbol{x}, W) \in \mathcal{Z}_p$, for $\boldsymbol{x} \in \mathbb{R}^N$.

- ρ strictly generalizes the p-th order mean-upper-semideviation.
  - Recovered by choosing $\mathcal{R}(\cdot) \equiv (\cdot)_+ \triangleq \max\{\cdot, 0\}$ in (2), satisfying conditions **S1** – **S4**.
- Problem (1) may be written compactly as

$$\inf_{x \in \mathcal{X}} \left\{\rho\left(F(x, W)\right) \triangleq \phi^{\widetilde{F}}(x)\right\}.$$

Mean-Semideviations

- The objective of (1) evaluates the *risk-measure*

$$\rho(Z) \triangleq \mathbb{E}\{Z\} + c\left(\mathbb{E}\left\{\left(\mathcal{R}\left(Z - \mathbb{E}\{Z\}\right)\right)^p\right\}\right)^{1/p}, \quad Z \in \mathcal{Z}_p, \tag{2}$$

  at $Z \equiv F(x, W) \in \mathcal{Z}_p$, for $x \in \mathbb{R}^N$.

- ρ strictly generalizes the p-th order mean-upper-semideviation.
  - Recovered by choosing $\mathcal{R}(\cdot) \equiv (\cdot)_+ \triangleq \max\{\cdot, 0\}$ in (2), satisfying conditions **S1** – **S4**.
- Problem (1) may be written compactly as

$$\inf_{x \in \mathcal{X}} \left\{\rho\left(F(x, W)\right) \triangleq \phi^{\widetilde{F}}(x)\right\}.$$

Examples of Mean-Semideviations - 1

- Mean-Upper-Semideviations (with $\mathcal{R}(\cdot) \equiv (\cdot)_+$):

$$\begin{aligned}\rho(Z) &\equiv \mathbb{E}\{Z\} + c\left(\mathbb{E}\left\{\left((Z-\mathbb{E}\{Z\})_+\right)^p\right\}\right)^{1/p} \\ &\equiv \mathbb{E}\{Z\} + c\left\|(Z-\mathbb{E}\{Z\})_+\right\|_{\mathcal{L}_p}, \quad Z \in \mathcal{Z}_p.\end{aligned}$$

- Entropic Mean-Semideviations (with $\mathcal{R}(\cdot) \equiv t^{-1}\log(1+\exp(t(\cdot)))$):

$$\rho(Z) \equiv \mathbb{E}\{Z\} + \frac{c}{t}\left\|\log\left(1+\exp\left(t\left(Z-\mathbb{E}\{Z\}\right)\right)\right)\right\|_{\mathcal{L}_p}, \quad Z \in \mathcal{Z}_p.$$

- Gaussian Antiderivative (GA) Mean-Semideviations (with $\mathcal{R}(\cdot) \equiv (\cdot)\Phi(\cdot) + \varphi(\cdot)$):

$$\rho(Z) \equiv \mathbb{E}\{Z\} + c\left\|(Z-\mathbb{E}\{Z\})\Phi(Z-\mathbb{E}\{Z\}) + \varphi(Z-\mathbb{E}\{Z\})\right\|_{\mathcal{L}_p}, \quad Z \in \mathcal{Z}_p,$$

where Φ and φ are the (standard) Gaussian cdf and pdf, respectively.

Examples of Mean-Semideviations - 1

- Mean-Upper-Semideviations (with $\mathcal{R}(\cdot) \equiv (\cdot)_+$):

$$\begin{aligned}\rho(Z) &\equiv \mathbb{E}\{Z\} + c\left(\mathbb{E}\left\{\left((Z-\mathbb{E}\{Z\})_+\right)^p\right\}\right)^{1/p} \\ &\equiv \mathbb{E}\{Z\} + c\left\|(Z-\mathbb{E}\{Z\})_+\right\|_{\mathcal{L}_p}, \quad Z \in \mathcal{Z}_p.\end{aligned}$$

- Entropic Mean-Semideviations (with $\mathcal{R}(\cdot) \equiv t^{-1}\log(1+\exp(t(\cdot)))$):

$$\rho(Z) \equiv \mathbb{E}\{Z\} + \frac{c}{t}\left\|\log\left(1+\exp\left(t\left(Z-\mathbb{E}\{Z\}\right)\right)\right)\right\|_{\mathcal{L}_p}, \quad Z \in \mathcal{Z}_p.$$

- Gaussian Antiderivative (GA) Mean-Semideviations (with $\mathcal{R}(\cdot) \equiv (\cdot)\Phi(\cdot) + \varphi(\cdot)$):

$$\rho(Z) \equiv \mathbb{E}\{Z\} + c\left\|(Z-\mathbb{E}\{Z\})\Phi(Z-\mathbb{E}\{Z\}) + \varphi(Z-\mathbb{E}\{Z\})\right\|_{\mathcal{L}_p}, \quad Z \in \mathcal{Z}_p,$$

where Φ and φ are the (standard) Gaussian cdf and pdf, respectively.

Examples of Mean-Semideviations - 1

- Mean-Upper-Semideviations (with $\mathcal{R}(\cdot) \equiv (\cdot)_+$):

$$\begin{aligned}\rho(Z) &\equiv \mathbb{E}\{Z\} + c\left(\mathbb{E}\left\{\left((Z-\mathbb{E}\{Z\})_+\right)^p\right\}\right)^{1/p} \\ &\equiv \mathbb{E}\{Z\} + c\left\|(Z-\mathbb{E}\{Z\})_+\right\|_{\mathcal{L}_p}, \quad Z \in \mathcal{Z}_p.\end{aligned}$$

- Entropic Mean-Semideviations (with $\mathcal{R}(\cdot) \equiv t^{-1}\log(1+\exp(t(\cdot)))$):

$$\rho(Z) \equiv \mathbb{E}\{Z\} + \frac{c}{t}\left\|\log(1+\exp(t(Z-\mathbb{E}\{Z\})))\right\|_{\mathcal{L}_p}, \quad Z \in \mathcal{Z}_p.$$

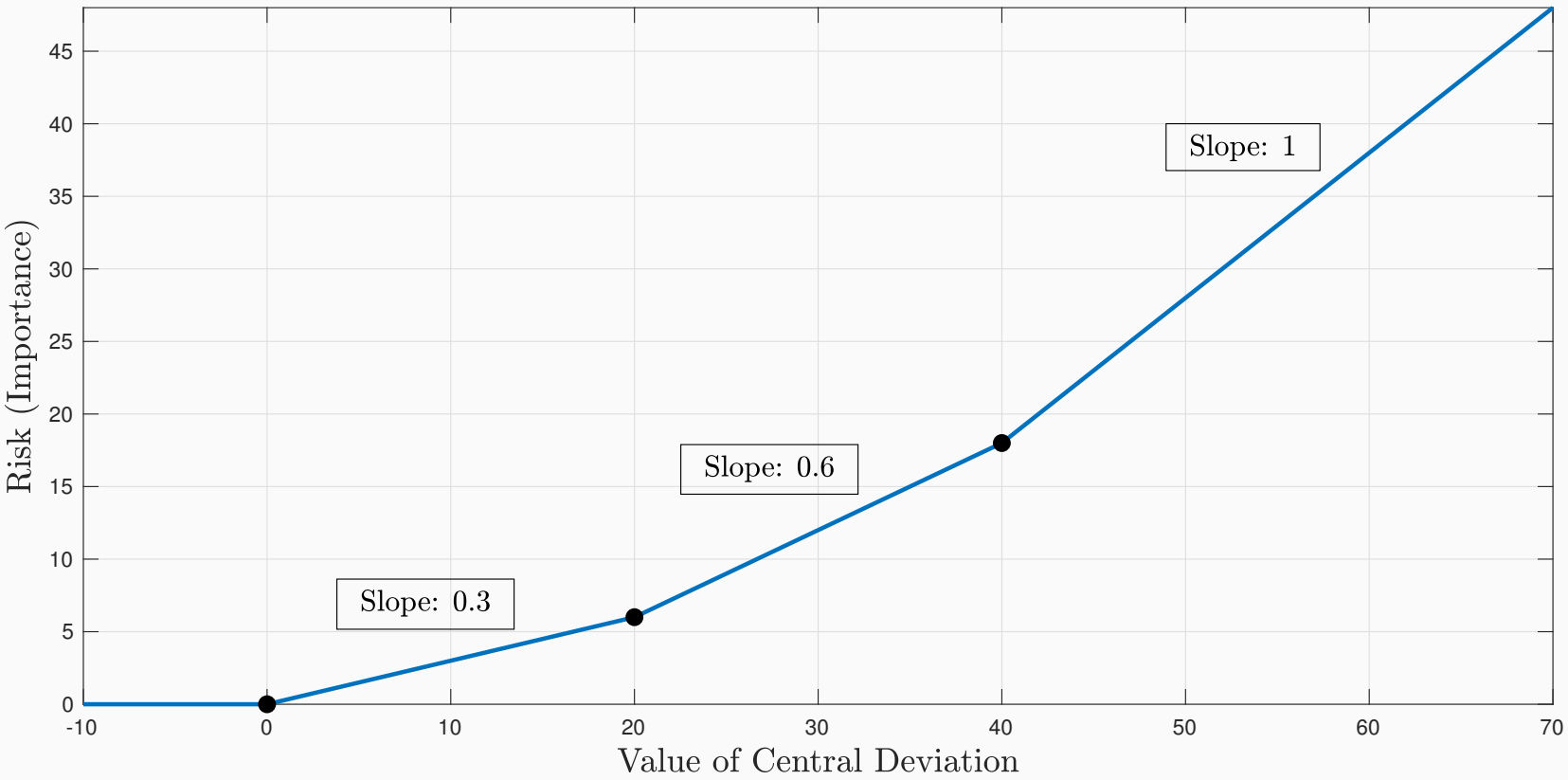
- Gaussian Antiderivative (GA) Mean-Semideviations (with $\mathcal{R}(\cdot) \equiv (\cdot)\Phi(\cdot) + \varphi(\cdot)$):

$$\rho(Z) \equiv \mathbb{E}\{Z\} + c\left\|(Z-\mathbb{E}\{Z\})\Phi(Z-\mathbb{E}\{Z\}) + \varphi(Z-\mathbb{E}\{Z\})\right\|_{\mathcal{L}_p}, \quad Z \in \mathcal{Z}_p,$$

where Φ and φ are the (standard) Gaussian cdf and pdf, respectively.

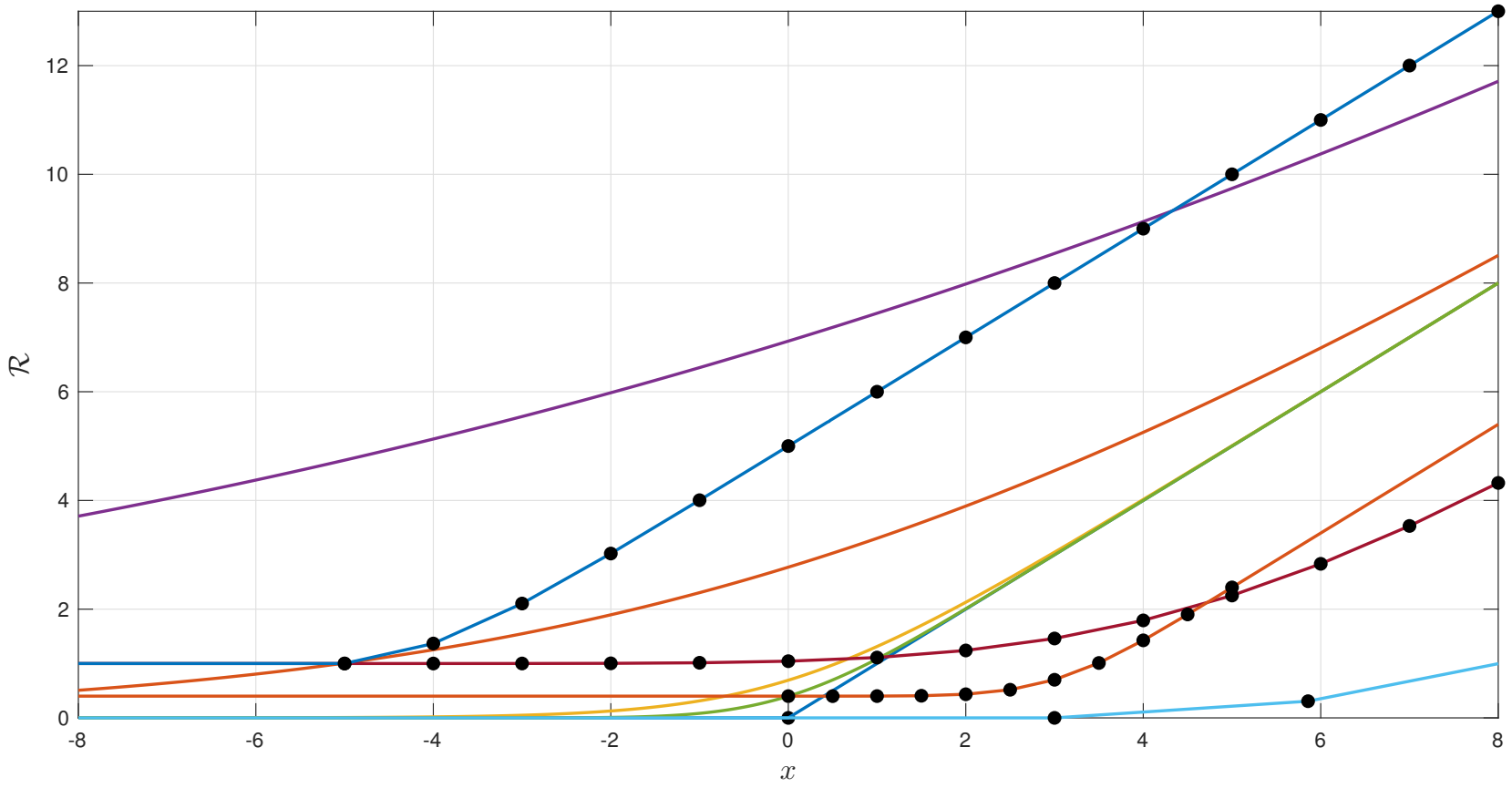
Examples of Mean-Semideviations - 2

- Any convex (S1), nonnegative (S2), increasing (S3), and nonexpansive (S4) **piecewise linear function** (with kinks).



Examples of Mean-Semideviations - 3

- **Countless** other examples!



Motivation for Optimization of Mean-Semideviations

- Relevant whenever **variable risk assessment** relative to the values of the respective central deviation is desirable.
- Typical examples:
  - Part yield maximization in manufacturing facilities.
  - Production planning and related newsvendor-type problems.
- Related areas:
  - Supply chain optimization.
  - Resource allocation (more broadly).
- Risk-aware Learning:
  - machine learning, reinforcement and sequential learning, AI.
- Central research directions:
  - **Design of algorithms** with desirable characteristics:
    - Efficient
    - Data-driven
    - Exact
    - Fast
  - **Convergence analysis** under reasonable assumptions, covering a wide range of problems.

Motivation for Optimization of Mean-Semideviations

- Relevant whenever **variable risk assessment** relative to the values of the respective central deviation is desirable.
- Typical examples:
  - Part yield maximization in manufacturing facilities.
  - Production planning and related newsvendor-type problems.
- Related areas:
  - Supply chain optimization.
  - Resource allocation (more broadly).
- Risk-aware Learning:
  - machine learning, reinforcement and sequential learning, AI.
- Central research directions:
  - **Design of algorithms** with desirable characteristics:
    - Efficient
    - Data-driven
    - Exact
    - Fast
  - **Convergence analysis** under reasonable assumptions, covering a wide range of problems.

Motivation for Optimization of Mean-Semideviations

- Relevant whenever **variable risk assessment** relative to the values of the respective central deviation is desirable.
- Typical examples:
  - Part yield maximization in manufacturing facilities.
  - Production planning and related newsvendor-type problems.
- Related areas:
  - Supply chain optimization.
  - Resource allocation (more broadly).
- Risk-aware Learning:
  - machine learning, reinforcement and sequential learning, AI.
- Central research directions:
  - **Design of algorithms** with desirable characteristics:
    - Efficient
    - Data-driven
    - Exact
    - Fast
  - **Convergence analysis** under reasonable assumptions, covering a wide range of problems.

Motivation for Optimization of Mean-Semideviations

- Relevant whenever **variable risk assessment** relative to the values of the respective central deviation is desirable.
- Typical examples:
  - Part yield maximization in manufacturing facilities.
  - Production planning and related newsvendor-type problems.
- Related areas:
  - Supply chain optimization.
  - Resource allocation (more broadly).
- Risk-aware Learning:
  - machine learning, reinforcement and sequential learning, AI.
- Central research directions:
  - **Design of algorithms** with desirable characteristics:
    - Efficient
    - Data-driven
    - Exact
    - Fast
  - **Convergence analysis** under reasonable assumptions, covering a wide range of problems.

Motivation for Optimization of Mean-Semideviations

- Relevant whenever **variable risk assessment** relative to the values of the respective central deviation is desirable.
- Typical examples:
  - Part yield maximization in manufacturing facilities.
  - Production planning and related newsvendor-type problems.
- Related areas:
  - Supply chain optimization.
  - Resource allocation (more broadly).
- Risk-aware Learning:
  - machine learning, reinforcement and sequential learning, AI.
- **Central research directions:**
  - **Design of algorithms** with desirable characteristics:
    - Efficient
    - Data-driven
    - Exact
    - Fast
  - **Convergence analysis** under reasonable assumptions, covering a wide range of problems.

Main Contributions

- *MESSAGE*^{p} Algorithm[1]:
  - An **efficient**, **data-driven** Stochastic Subgradient Descent (SSD)-type procedure for iteratively solving **convex** mean-semideviation problems **to optimality**.
  - A variation of the **T-Level Stochastic Compositional Gradient Descent (T-SCGD)** algorithm of [Yang, Wang & Fang, 2018][2], analyzed under a **generic framework.**
- **Complete convergence analysis**, under **a new and flexible** structural framework:
  - We establish **pathwise convergence**, in the same strong sense of [Yang et al., 2018].
  - We derive explicit **convergence rates** of the orders of
    - $\mathcal{O}\left(n^{-2/3}\right)$, when $p \equiv 1$ (matching the state of the art).
    - $\mathcal{O}\left(n^{-1/2}\right)$, when $p > 1$ (completely new result).
  - Our problem assumptions reveal a **trade-off** between the **expansiveness** of the random cost and the **smoothness** of the risk regularizer.
  - Our problem assumptions **strictly generalize those of [Yang et al., 2018]**.

[1] *MESSAGE*^{p} = **ME**an-**S**emideviation **S**tochastic composition**A**l sub**G**radient d**E**scent of order p

[2] Yang, Shuoguang, Wang, Mengdi, and Fang, Ethan X., "Multi-Level Stochastic Gradient Methods for Nested Composition Optimization", Arxiv (2018).

Main Contributions

- *MESSAGE*^{p} Algorithm[1]:
  - An **efficient**, **data-driven** Stochastic Subgradient Descent (SSD)-type procedure for iteratively solving **convex** mean-semideviation problems **to optimality**.
  - A variation of the **T-Level Stochastic Compositional Gradient Descent (T-SCGD)** algorithm of [Yang, Wang & Fang, 2018][2], analyzed under a **generic framework.**
- **Complete convergence analysis**, under **a new and flexible** structural framework:
  - We establish **pathwise convergence**, in the same strong sense of [Yang et al., 2018].
  - We derive explicit **convergence rates** of the orders of
    - $\mathcal{O}\left(n^{-2/3}\right)$, when $p \equiv 1$ (matching the state of the art).
    - $\mathcal{O}\left(n^{-1/2}\right)$, when $p > 1$ (completely new result).
  - Our problem assumptions reveal a **trade-off** between the **expansiveness** of the random cost and the **smoothness** of the risk regularizer.
  - Our problem assumptions **strictly generalize those of [Yang et al., 2018]**.

[1] *MESSAGE*^{p} = **ME**an-**S**emideviation **S**tochastic composition**A**l sub**G**radient d**E**scent of order p

[2] Yang, Shuoguang, Wang, Mengdi, and Fang, Ethan X., "Multi-Level Stochastic Gradient Methods for Nested Composition Optimization", Arxiv (2018).

Main Contributions

- *MESSAGE*^{p} Algorithm[1]:
  - An **efficient**, **data-driven** Stochastic Subgradient Descent (SSD)-type procedure for iteratively solving **convex** mean-semideviation problems **to optimality**.
  - A variation of the **T-Level Stochastic Compositional Gradient Descent (T-SCGD)** algorithm of [Yang, Wang & Fang, 2018][2], analyzed under a **generic framework.**
- **Complete convergence analysis**, under **a new and flexible** structural framework:
  - We establish **pathwise convergence**, in the same strong sense of [Yang et al., 2018].
  - We derive explicit **convergence rates** of the orders of
    - $\mathcal{O}\left(n^{-2/3}\right)$, when $p \equiv 1$ (matching the state of the art).
    - $\mathcal{O}\left(n^{-1/2}\right)$, when $p > 1$ (completely new result).
  - Our problem assumptions reveal a **trade-off** between the **expansiveness** of the random cost and the **smoothness** of the risk regularizer.
  - Our problem assumptions **strictly generalize those of [Yang et al., 2018]**.

[1] *MESSAGE*^{p} = **ME**an-**S**emideviation **S**tochastic composition**A**l sub**G**radient d**E**scent of order p

[2] Yang, Shuoguang, Wang, Mengdi, and Fang, Ethan X., "Multi-Level Stochastic Gradient Methods for Nested Composition Optimization", Arxiv (2018).

Main Contributions

- *MESSAGE*^{p} Algorithm[1]:
  - An **efficient**, **data-driven** Stochastic Subgradient Descent (SSD)-type procedure for iteratively solving **convex** mean-semideviation problems **to optimality**.
  - A variation of the **T-Level Stochastic Compositional Gradient Descent (T-SCGD)** algorithm of [Yang, Wang & Fang, 2018][2], analyzed under a **generic framework.**
- **Complete convergence analysis**, under **a new and flexible** structural framework:
  - We establish **pathwise convergence**, in the same strong sense of [Yang et al., 2018].
  - We derive explicit **convergence rates** of the orders of
    - $\mathcal{O}\left(n^{-2/3}\right)$, when $p \equiv 1$ (matching the state of the art).
    - $\mathcal{O}\left(n^{-1/2}\right)$, when $p > 1$ (completely new result).
  - Our problem assumptions reveal a **trade-off** between the **expansiveness** of the random cost and the **smoothness** of the risk regularizer.
  - Our problem assumptions **strictly generalize those of [Yang et al., 2018]**.

[1] *MESSAGE*^{p} = **ME**an-**S**emideviation **S**tochastic composition**A**l sub**G**radient d**E**scent of order p

[2] Yang, Shuoguang, Wang, Mengdi, and Fang, Ethan X., "Multi-Level Stochastic Gradient Methods for Nested Composition Optimization", Arxiv (2018).

Main Contributions

- *MESSAGE*^{p} Algorithm[1]:
  - An **efficient**, **data-driven** Stochastic Subgradient Descent (SSD)-type procedure for iteratively solving **convex** mean-semideviation problems **to optimality**.
  - A variation of the **T-Level Stochastic Compositional Gradient Descent (T-SCGD)** algorithm of [Yang, Wang & Fang, 2018][2], analyzed under a **generic framework.**
- **Complete convergence analysis**, under **a new and flexible** structural framework:
  - We establish **pathwise convergence**, in the same strong sense of [Yang et al., 2018].
  - We derive explicit **convergence rates** of the orders of
    - $\mathcal{O}\left(n^{-2/3}\right)$, when $p \equiv 1$ (matching the state of the art).
    - $\mathcal{O}\left(n^{-1/2}\right)$, when $p > 1$ (completely new result).
  - Our problem assumptions reveal a **trade-off** between the **expansiveness** of the random cost and the **smoothness** of the risk regularizer.
  - Our problem assumptions **strictly generalize those of [Yang et al., 2018]**.

[1] *MESSAGE*^{p} = **ME**an-**S**emideviation **S**tochastic composition**A**l sub**G**radient d**E**scent of order p

[2] Yang, Shuoguang, Wang, Mengdi, and Fang, Ethan X., "Multi-Level Stochastic Gradient Methods for Nested Composition Optimization", Arxiv (2018).

Main Contributions

- *MESSAGE*^{p} Algorithm[1]:
  - An **efficient**, **data-driven** Stochastic Subgradient Descent (SSD)-type procedure for iteratively solving **convex** mean-semideviation problems **to optimality**.
  - A variation of the **T-Level Stochastic Compositional Gradient Descent (T-SCGD)** algorithm of [Yang, Wang & Fang, 2018][2], analyzed under a **generic framework.**
- **Complete convergence analysis**, under **a new and flexible** structural framework:
  - We establish **pathwise convergence**, in the same strong sense of [Yang et al., 2018].
  - We derive explicit **convergence rates** of the orders of
    - $\mathcal{O}\left(n^{-2/3}\right)$, when $p \equiv 1$ (matching the state of the art).
    - $\mathcal{O}\left(n^{-1/2}\right)$, when $p > 1$ (completely new result).
  - Our problem assumptions reveal a **trade-off** between the **expansiveness** of the random cost and the **smoothness** of the risk regularizer.
  - Our problem assumptions **strictly generalize those of [Yang et al., 2018]**.

[1] *MESSAGE*^{p} = **ME**an-**S**emideviation **S**tochastic composition**A**l sub**G**radient d**E**scent of order p

[2] Yang, Shuoguang, Wang, Mengdi, and Fang, Ethan X., "Multi-Level Stochastic Gradient Methods for Nested Composition Optimization", Arxiv (2018).

Main Contributions

- *MESSAGE*^{p} Algorithm[1]:
  - An **efficient**, **data-driven** Stochastic Subgradient Descent (SSD)-type procedure for iteratively solving **convex** mean-semideviation problems **to optimality**.
  - A variation of the **T-Level Stochastic Compositional Gradient Descent (T-SCGD)** algorithm of [Yang, Wang & Fang, 2018][2], analyzed under a **generic framework.**
- **Complete convergence analysis**, under **a new and flexible** structural framework:
  - We establish **pathwise convergence**, in the same strong sense of [Yang et al., 2018].
  - We derive explicit **convergence rates** of the orders of
    - $\mathcal{O}\left(n^{-2/3}\right)$, when $p \equiv 1$ (matching the state of the art).
    - $\mathcal{O}\left(n^{-1/2}\right)$, when $p > 1$ (completely new result).
  - Our problem assumptions reveal a **trade-off** between the **expansiveness** of the random cost and the **smoothness** of the risk regularizer.
  - Our problem assumptions **strictly generalize those of [Yang et al., 2018]**.

[1] *MESSAGE*^{p} = **ME**an-**S**emideviation **S**tochastic composition**A**l sub**G**radient d**E**scent of order p

[2] Yang, Shuoguang, Wang, Mengdi, and Fang, Ethan X., "Multi-Level Stochastic Gradient Methods for Nested Composition Optimization", Arxiv (2018).

The *MESSAGE*^{p} Algorithm

Basic Properties of the Base Problem

Proposition (When are mean-semideviations convex-monotone?)

Fix $p \in [1, \infty)$ and choose any risk regularizer $\mathcal{R} : \mathbb{R} \to \mathbb{R}$. Then, as long as $c \in [0, 1]$, the risk measure ρ is convex (convex, monotone and translation equivariant).

Corollary (When is (1) Convex?)

Fix $p \in [1, \infty)$ and choose any risk regularizer $\mathcal{R} : \mathbb{R} \to \mathbb{R}$. Then, as long as $c \in [0, 1]$, the composite function $\phi^{\widetilde{F}}(\cdot) \equiv \rho(F(\cdot, W))$ is convex on $\mathbb{R}^N$, and (1) constitutes a convex stochastic program.

Proposition (When are mean-semideviations convex-monotone?)

Fix $p \in [1, \infty)$ and choose any risk regularizer $\mathcal{R} : \mathbb{R} \to \mathbb{R}$. Then, as long as $c \in [0, 1]$, the risk measure ρ is convex (convex, monotone and translation equivariant).

Corollary (When is (1) Convex?)

Fix $p \in [1, \infty)$ and choose any risk regularizer $\mathcal{R} : \mathbb{R} \to \mathbb{R}$. Then, as long as $c \in [0, 1]$, the composite function $\phi^{\widetilde{F}}(\cdot) \equiv \rho(F(\cdot, W))$ is convex on $\mathbb{R}^N$, and (1) constitutes a convex stochastic program.

Mean-Semideviations in Compositional Form

- Define **expectation functions**

$$\varrho(x) \triangleq x^{1/p}, \quad x > 0 \text{ (when } p > 1),$$
$$g^{\widetilde{F}}(x, y) \triangleq \mathbb{E}\left\{\left(\mathcal{R}\left(F(x, W) - y\right)\right)^p\right\} \quad \text{and}$$
$$h^{\widetilde{F}}(x) \triangleq \left[x\, \mathbb{E}\left\{F(x, W)\right\}\right],$$

  for every admissible choice of F and $\mathcal{P}_W$ (the pushforward (distribution) of W).

- Objective may be expressed as

$$\phi^{\widetilde{F}}(x) \equiv \mathbb{E}\left\{F(x, W)\right\} + c\varrho\left(g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\right), \quad \forall x \in \mathcal{X}.$$

- Equivalent formulation of the base problem:

$$\inf_{x \in \mathcal{X}} \phi^{\widetilde{F}}(x) \triangleq \mathbb{E}\left\{F(x, W)\right\} + c\varrho\left(g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\right).$$

Mean-Semideviations in Compositional Form

- Define **expectation functions**

$$\varrho(x) \triangleq x^{1/p}, \quad x > 0 \text{ (when } p > 1),$$
$$g^{\widetilde{F}}(x, y) \triangleq \mathbb{E}\left\{\left(\mathcal{R}\left(F(x, W) - y\right)\right)^p\right\} \quad \text{and}$$
$$h^{\widetilde{F}}(x) \triangleq \left[x\, \mathbb{E}\left\{F(x, W)\right\}\right],$$

  for every admissible choice of F and $\mathcal{P}_W$ (the pushforward (distribution) of W).

- Objective may be expressed as

$$\phi^{\widetilde{F}}(x) \equiv \mathbb{E}\left\{F(x, W)\right\} + c\varrho\left(g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\right), \quad \forall x \in \mathcal{X}.$$

- Equivalent formulation of the base problem:

$$\inf_{x \in \mathcal{X}} \phi^{\widetilde{F}}(x) \triangleq \mathbb{E}\left\{F(x, W)\right\} + c\varrho\left(g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\right).$$

Mean-Semideviations in Compositional Form

- Define **expectation functions**

$$\varrho(x) \triangleq x^{1/p}, \quad x > 0 \text{ (when } p > 1),$$
$$g^{\widetilde{F}}(x, y) \triangleq \mathbb{E}\left\{\left(\mathcal{R}\left(F(x, W) - y\right)\right)^p\right\} \quad \text{and}$$
$$h^{\widetilde{F}}(x) \triangleq \left[x\, \mathbb{E}\left\{F(x, W)\right\}\right],$$

  for every admissible choice of F and $\mathcal{P}_W$ (the pushforward (distribution) of W).

- Objective may be expressed as

$$\phi^{\widetilde{F}}(x) \equiv \mathbb{E}\left\{F(x, W)\right\} + c\varrho\left(g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\right), \quad \forall x \in \mathcal{X}.$$

- Equivalent formulation of the base problem:

$$\inf_{x \in \mathcal{X}} \phi^{\widetilde{F}}(x) \triangleq \mathbb{E}\left\{F(x, W)\right\} + c\varrho\left(g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\right).$$

Differentiability of $\phi^{\widetilde{F}}(\cdot) \equiv \rho(F(\cdot, W))$

Lemma (Differentiability of $\phi^{\widetilde{F}}$)

Under some very mild technical assumptions:

$$\nabla\phi^{\widetilde{F}}(x) \equiv \mathbb{E}\left\{\underline{\nabla}F(x,W)\right\} + c\nabla h^{\widetilde{F}}(x)\,\nabla g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\nabla\varrho\left(g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\right),\ \forall x \in \mathcal{X},$$

where

$$\nabla\varrho(x) \equiv \begin{cases} 1, & \text{if } p \equiv 1 \\ \dfrac{x^{(1-p)/p}}{p}, & \text{if } p \in (1,\infty) \end{cases},\ \nabla h^{\widetilde{F}}(x) \equiv \mathbb{E}\left\{\left[\begin{array}{c|c} I_N & \underline{\nabla}F(x,W) \end{array}\right]\right\} \text{ and}$$

$$\nabla g^{\widetilde{F}}(x,y) \equiv \mathbb{E}\left\{p\left(\mathcal{R}\left(F(x,W)-y\right)\right)^{p-1}\underline{\nabla}\mathcal{R}\left(F(x,W)-y\right)\left[\begin{array}{c} \underline{\nabla}F(x,W) \\ \hline -1 \end{array}\right]\right\},$$

respectively, for every $(x,y) \in \mathrm{Graph}_{\mathcal{X}}\left(\mathbb{E}\left\{F(\cdot,W)\right\}\right)$.

Take a closer look at $\nabla\phi^{\widetilde{F}}$

$$\boxed{\nabla\phi^{\widetilde{F}}(x) \equiv \mathbb{E}\left\{\underline{\nabla} F(x,W)\right\} + c\nabla h^{\widetilde{F}}(x)\,\nabla g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\nabla\varrho\left(g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\right)}$$

$$h^{\widetilde{F}}(x) \equiv \left[x\,\mathbb{E}\left\{F(x,W)\right\}\right]$$

$$g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right) \equiv \mathbb{E}\left\{\left(\mathcal{R}\left(F(x,W) - \mathbb{E}\left\{F(x,W)\right\}\right)\right)^{p}\right\}$$

- If $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ where known, a stochastic gradient could be formed.
- But both $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ are expectation functions *themselves*!
  - Not possible to generate a a stochastic gradient in a single sampling step.
- For fixed x, $\mathbb{E}\left\{F(x,W)\right\}$ can be recursively estimated using **stochastic approximation**.
- Same for $\mathbb{E}\left\{\left(\mathcal{R}\left(F(x,W) - y\right)\right)^{p}\right\}$, for also fixed y.
- $\Rightarrow$ **Simplest Idea**: TRACK $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ using **biased stochastic approximation**!

Take a closer look at $\nabla\phi^{\widetilde{F}}$

$$\boxed{\nabla\phi^{\widetilde{F}}(x) \equiv \mathbb{E}\left\{\underline{\nabla} F(x,W)\right\} + c\nabla h^{\widetilde{F}}(x)\,\nabla g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\nabla\varrho\left(g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\right)}$$

$$h^{\widetilde{F}}(x) \equiv \left[x\,\mathbb{E}\left\{F(x,W)\right\}\right]$$

$$g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right) \equiv \mathbb{E}\left\{\left(\mathcal{R}\left(F(x,W) - \mathbb{E}\left\{F(x,W)\right\}\right)\right)^{p}\right\}$$

- If $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ where known, a stochastic gradient could be formed.
- But both $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ are expectation functions *themselves*!
  - Not possible to generate a a stochastic gradient in a single sampling step.
- For fixed x, $\mathbb{E}\left\{F(x,W)\right\}$ can be recursively estimated using **stochastic approximation**.
- Same for $\mathbb{E}\left\{\left(\mathcal{R}\left(F(x,W) - y\right)\right)^{p}\right\}$, for also fixed y.
- $\Rightarrow$ **Simplest Idea**: TRACK $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ using **biased stochastic approximation**!

Take a closer look at $\nabla\phi^{\widetilde{F}}$

$$\boxed{\nabla\phi^{\widetilde{F}}(x) \equiv \mathbb{E}\left\{\underline{\nabla} F(x,W)\right\} + c\nabla h^{\widetilde{F}}(x)\,\nabla g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\nabla\varrho\left(g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\right)}$$

$$h^{\widetilde{F}}(x) \equiv \left[x\,\mathbb{E}\left\{F(x,W)\right\}\right]$$

$$g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right) \equiv \mathbb{E}\left\{\left(\mathcal{R}\left(F(x,W) - \mathbb{E}\left\{F(x,W)\right\}\right)\right)^{p}\right\}$$

- If $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ where known, a stochastic gradient could be formed.
- But both $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ are expectation functions *themselves*!
  - Not possible to generate a a stochastic gradient in a single sampling step.
- For fixed x, $\mathbb{E}\left\{F(x,W)\right\}$ can be recursively estimated using **stochastic approximation**.
- Same for $\mathbb{E}\left\{\left(\mathcal{R}\left(F(x,W) - y\right)\right)^{p}\right\}$, for also fixed y.
- $\Rightarrow$ **Simplest Idea**: TRACK $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ using **biased stochastic approximation**!

Take a closer look at $\nabla\phi^{\widetilde{F}}$

$$\boxed{\nabla\phi^{\widetilde{F}}(x) \equiv \mathbb{E}\left\{\underline{\nabla} F(x,W)\right\} + c\nabla h^{\widetilde{F}}(x)\,\nabla g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\nabla\varrho\left(g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\right)}$$

$$h^{\widetilde{F}}(x) \equiv \left[x\,\mathbb{E}\left\{F(x,W)\right\}\right]$$

$$g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right) \equiv \mathbb{E}\left\{\left(\mathcal{R}\left(F(x,W) - \mathbb{E}\left\{F(x,W)\right\}\right)\right)^{p}\right\}$$

- If $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ where known, a stochastic gradient could be formed.
- But both $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ are expectation functions *themselves*!
  - Not possible to generate a a stochastic gradient in a single sampling step.
- For fixed x, $\mathbb{E}\left\{F(x,W)\right\}$ can be recursively estimated using **stochastic approximation**.
- Same for $\mathbb{E}\left\{\left(\mathcal{R}\left(F(x,W) - y\right)\right)^{p}\right\}$, for also fixed y.
- $\Rightarrow$ **Simplest Idea**: TRACK $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ using **biased stochastic approximation**!

Take a closer look at $\nabla\phi^{\widetilde{F}}$

$$\boxed{\nabla\phi^{\widetilde{F}}(x) \equiv \mathbb{E}\left\{\underline{\nabla} F(x,W)\right\} + c\nabla h^{\widetilde{F}}(x)\,\nabla g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\nabla\varrho\left(g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\right)}$$

$$h^{\widetilde{F}}(x) \equiv \left[x\,\mathbb{E}\left\{F(x,W)\right\}\right]$$

$$g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right) \equiv \mathbb{E}\left\{\left(\mathcal{R}\left(F(x,W) - \mathbb{E}\left\{F(x,W)\right\}\right)\right)^{p}\right\}$$

- If $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ where known, a stochastic gradient could be formed.
- But both $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ are expectation functions *themselves*!
  - Not possible to generate a a stochastic gradient in a single sampling step.
- For fixed x, $\mathbb{E}\left\{F(x,W)\right\}$ can be recursively estimated using **stochastic approximation**.
- Same for $\mathbb{E}\left\{\left(\mathcal{R}\left(F(x,W) - y\right)\right)^{p}\right\}$, for also fixed y.
- $\Rightarrow$ **Simplest Idea**: TRACK $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ using **biased stochastic approximation**!

Take a closer look at $\nabla\phi^{\widetilde{F}}$

$$\boxed{\nabla\phi^{\widetilde{F}}(x) \equiv \mathbb{E}\left\{\underline{\nabla} F(x,W)\right\} + c\nabla h^{\widetilde{F}}(x)\,\nabla g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\nabla\varrho\left(g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right)\right)}$$

$$h^{\widetilde{F}}(x) \equiv \left[x\,\mathbb{E}\left\{F(x,W)\right\}\right]$$

$$g^{\widetilde{F}}\left(h^{\widetilde{F}}(x)\right) \equiv \mathbb{E}\left\{\left(\mathcal{R}\left(F(x,W) - \mathbb{E}\left\{F(x,W)\right\}\right)\right)^{p}\right\}$$

- If $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ where known, a stochastic gradient could be formed.
- But both $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ are expectation functions *themselves*!
  - Not possible to generate a a stochastic gradient in a single sampling step.
- For fixed x, $\mathbb{E}\left\{F(x,W)\right\}$ can be recursively estimated using **stochastic approximation**.
- Same for $\mathbb{E}\left\{\left(\mathcal{R}\left(F(x,W) - y\right)\right)^{p}\right\}$, for also fixed y.
- $\Rightarrow$ **Simplest Idea**: TRACK $h^{\widetilde{F}}(\cdot)$ and $g^{\widetilde{F}}\left(h^{\widetilde{F}}(\cdot)\right)$ using **biased stochastic approximation**!

The *MESSAGE*^{p} Algorithm

Input: Initial points $x^0 \in \mathcal{X}$, $y^0 \in \mathbb{R}$, $z^0 \in \mathbb{R}$, stepsize sequences $\{\alpha_n\}_{n\in\mathbb{N}}$, $\{\beta_n\}_{n\in\mathbb{N}}$, $\{\gamma_n\}_{n\in\mathbb{N}}$, IID sequences $\{W_1^n\}_{n\in\mathbb{N}}$, $\{W_2^n\}_{n\in\mathbb{N}}$ and penalty coefficient $c \in [0,1]$.

Output: Sequence $\{x^n\}_{n\in\mathbb{N}}$.

1: **for** $n = 0, 1, 2, \ldots$ **do**

2: Obtain $F\left(x^n, W_1^{n+1}\right)$ and $\underline{\nabla} F\left(x^n, W_1^{n+1}\right)$ from the $\mathcal{SO}$.

3: Update (First SA Level):

$$\boxed{y^{n+1} = (1-\beta_n)\, y^n + \beta_n F\left(x^n, W_1^{n+1}\right)}$$

4: Obtain $F\left(x^n, W_2^{n+1}\right)$ and $\underline{\nabla} F\left(x^n, W_2^{n+1}\right)$ from the $\mathcal{SO}$.

5: Update (Second SA Level):

$$\boxed{z^{n+1} = \begin{cases} 1, & \text{if } p = 1 \\ (1-\gamma_n)\, z^n + \gamma_n \left(\mathcal{R}\left(F\left(x^n, W_2^{n+1}\right) - y^n\right)\right)^p, & \text{if } p > 1 \end{cases}}$$

6: Define auxiliary variables:

$$\delta = F\left(x^n, W_2^{n+1}\right) - y^n$$

$$\boldsymbol{\delta}^{\underline{\nabla}} = \underline{\nabla} F\left(x^n, W_2^{n+1}\right) - \underline{\nabla} F\left(x^n, W_1^{n+1}\right)$$

$$\Delta = \boldsymbol{\delta}^{\underline{\nabla}} \underline{\nabla} \mathcal{R}(\delta) \left(\mathcal{R}(\delta)\right)^{p-1} \left(z^n\right)^{(1-p)/p}$$

7: Update (Third SA Level):

$$\boxed{x^{n+1} = \Pi_{\mathcal{X}} \left\{x^n - \alpha_n \left(\underline{\nabla} F\left(x^n, W_2^{n+1}\right) + c\Delta\right)\right\}}$$

8: **end for**

Advantages of the *MESSAGE*^{p} Algorithm

- **Effective: If** convergent, solves underlying problem **to optimality**.
  - Compare with Sample Average Approximation (SAA).
- **Efficient:**
  - Can exploit new information available to the decision maker.
  - Suitable when information is available sequentially, and decisions are made adaptively over time.
  - Minimal and fixed time and space complexity per iteration.
  - Suitable for problems with massive amounts of available data.

Good, but:

Does it converge?

Under what assumptions?

How fast?

Advantages of the *MESSAGE*^{p} Algorithm

- **Effective: If** convergent, solves underlying problem **to optimality**.
  - Compare with Sample Average Approximation (SAA).
- Efficient:
  - Can exploit new information available to the decision maker.
  - Suitable when information is available sequentially, and decisions are made adaptively over time.
  - Minimal and fixed time and space complexity per iteration.
  - Suitable for problems with massive amounts of available data.

Good, but:

Does it converge?

Under what assumptions?

How fast?

Advantages of the *MESSAGE*^{p} Algorithm

- **Effective: If** convergent, solves underlying problem **to optimality**.
  - Compare with Sample Average Approximation (SAA).
- **Efficient:**
  - Can exploit new information available to the decision maker.
  - Suitable when information is available sequentially, and decisions are made adaptively over time.
  - Minimal and fixed time and space complexity per iteration.
  - Suitable for problems with massive amounts of available data.

Good, but:

Does it converge?

Under what assumptions?

How fast?

Advantages of the *MESSAGE*^{p} Algorithm

- **Effective: If** convergent, solves underlying problem **to optimality**.
  - Compare with Sample Average Approximation (SAA).
- **Efficient:**
  - Can exploit new information available to the decision maker.
  - Suitable when information is available sequentially, and decisions are made adaptively over time.
  - Minimal and fixed time and space complexity per iteration.
  - Suitable for problems with massive amounts of available data.

Good, but:
Does it converge?
Under what assumptions?
How fast?

Advantages of the *MESSAGE*^{p} Algorithm

- **Effective: If** convergent, solves underlying problem **to optimality**.
  - Compare with Sample Average Approximation (SAA).
- **Efficient:**
  - Can exploit new information available to the decision maker.
  - Suitable when information is available sequentially, and decisions are made adaptively over time.
  - Minimal and fixed time and space complexity per iteration.
  - Suitable for problems with massive amounts of available data.

Good, but:
Does it converge?
Under what assumptions?
How fast?

Advantages of the *MESSAGE*^{p} Algorithm

- **Effective: If** convergent, solves underlying problem **to optimality**.
  - Compare with Sample Average Approximation (SAA).
- **Efficient:**
  - Can exploit new information available to the decision maker.
  - Suitable when information is available sequentially, and decisions are made adaptively over time.
  - Minimal and fixed time and space complexity per iteration.
  - Suitable for problems with massive amounts of available data.

Good, but:
Does it converge?
Under what assumptions?
How fast?

Advantages of the *MESSAGE*^{p} Algorithm

- **Effective: If** convergent, solves underlying problem **to optimality**.
  - Compare with Sample Average Approximation (SAA).
- **Efficient:**
  - Can exploit new information available to the decision maker.
  - Suitable when information is available sequentially, and decisions are made adaptively over time.
  - Minimal and fixed time and space complexity per iteration.
  - Suitable for problems with massive amounts of available data.

Good, but:
Does it converge?
Under what assumptions?
How fast?

Advantages of the *MESSAGE*^{p} Algorithm

- **Effective: If** convergent, solves underlying problem **to optimality**.
  - Compare with Sample Average Approximation (SAA).
- **Efficient:**
  - Can exploit new information available to the decision maker.
  - Suitable when information is available sequentially, and decisions are made adaptively over time.
  - Minimal and fixed time and space complexity per iteration.
  - Suitable for problems with massive amounts of available data.

Good, but:

Does it converge?

Under what assumptions?

How fast?

Analysis of the *MESSAGE*^{p} Algorithm

- For some measurable set $\Omega_E \subseteq \Omega$, such that $\mathcal{P}(\Omega_E) \equiv 1$, let us define

$$m_l \triangleq \inf_{x\in\mathcal{X}} \inf_{\omega\in\Omega_E} F(x, W(\omega)) \quad \text{and} \quad m_h \triangleq \sup_{x\in\mathcal{X}} \sup_{\omega\in\Omega_E} F(x, W(\omega)).$$

For $P \in [2, \infty]$ and $Q \in [P/(P-1), \infty]$, F and $\mathcal{R}$ satisfy the conditions:

C1 For chosen random subgradient $\underline{\nabla} F(\cdot, W)$, there exists a number $G < \infty$, such that

$$\sup_{x\in\mathcal{X}} \left[\mathbb{E}\left\{\|\underline{\nabla} F(x,W)\|_2^P\right\}\right]^{1/P} \triangleq \sup_{x\in\mathcal{X}} \left\|\|\underline{\nabla} F(x,W)\|_2\right\|_{\mathcal{L}_P} \leq G.$$

C2 There exists a number $V < \infty$, such that

$$\sup_{x\in\mathcal{X}} \mathbb{V}\{F(x,W)\} \triangleq \sup_{x\in\mathcal{X}} \left[\mathbb{E}\left\{(F(x,W))^2\right\} - (\mathbb{E}\{F(x,W)\})^2\right] \leq V.$$

C3 For chosen subderivative $\underline{\nabla}\mathcal{R}$, there exists another number $D < \infty$, such that

$$\sup_{x\in\mathcal{X}} \left\|\left|\underline{\nabla}(\mathcal{R}(z))^p\big|_{z\equiv F(x,W)-y_1} - \underline{\nabla}(\mathcal{R}(z))^p\big|_{z\equiv F(x,W)-y_2}\right|\right\|_{\mathcal{L}_Q} \leq D\,|y_1 - y_2|,$$

for all $(y_1, y_2) \in [\mathrm{cl}\{(m_l, m_h)\}]^2$.

C4 <u>Whenever $p > 1$</u>, it is true that $-\infty < m_l \leq m_h < \infty$, and

$$0 < \varepsilon \triangleq \mathcal{R}(m_l - m_h) \leq \mathcal{R}(m_h - m_l) \triangleq \mathcal{E} < \infty.$$

In other words, $\mathcal{R}$ is positively uniformly bounded within $\mathfrak{R}^{\widetilde{F}} \triangleq \mathrm{cl}\{(m_l - m_h, m_h - m_l)\}$.

Pathwise Convergence of the *MESSAGE*^{p} Algorithm

Theorem

Let $\beta_n \in (0,1]$, $\gamma_n \in (0,1]$, for all $n \in \mathbb{N}$. Whenever $p \equiv 1$, suppose that

$$\sum_{n\in\mathbb{N}} \alpha_n \equiv \infty \quad \text{and} \quad \sum_{n\in\mathbb{N}} \alpha_n^2 + \beta_n^2 + \frac{\alpha_n^2}{\beta_n} < \infty,$$

whereas, whenever $p > 1$, choose $y^0 \in [m_l, m_h]$, $z^0 \in [\varepsilon^p, \mathcal{E}^p]$, and suppose additionally that

$$\sum_{n\in\mathbb{N}} \gamma_n^2 + \frac{\alpha_n^2}{\gamma_n} + \frac{\beta_n^2}{\gamma_n} < \infty.$$

Then, as long as $\mathcal{X}^* \equiv \arg\min_{x\in\mathcal{X}} \phi^{\widetilde{F}}(x) \neq \emptyset$, the process $\{x^n\}_{n\in\mathbb{N}}$ generated by the MESSAGE^p algorithm satisfies

$$\mathcal{P}\left(\left\{\omega \in \Omega \,\middle|\, \exists x^*(\omega) \in \mathcal{X}^* \text{ such that } x^n(\omega) \underset{n\to\infty}{\longrightarrow} x^*(\omega)\right\}\right) \equiv 1.$$

In other words, almost everywhere relative to $\mathcal{P}$, the process $\{x^n\}_{n\in\mathbb{N}}$ converges to a random point in the set of optimal solutions of (1).

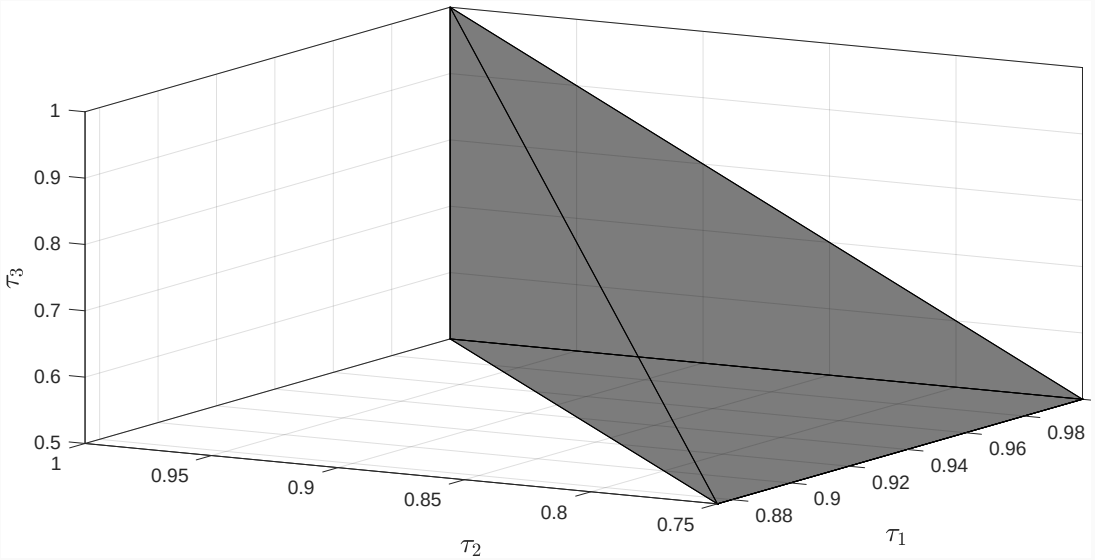
Comments

- And what about actual stepsize selection?
  - Convergence is guaranteed for subharmonic stepsizes (case $p > 1$)

$$\alpha_n \equiv \frac{1}{n^{\tau_1}}, \quad \beta_n \equiv \frac{1}{n^{\tau_2}} \quad \text{and} \quad \gamma_n \equiv \frac{1}{n^{\tau_3}}, \quad n \in \mathbb{N}^+,$$

as long as

$$\frac{7}{8} < \tau_1 \leq 1,$$
$$\frac{3}{4} < \tau_2 < 2\tau_1 - 1 \quad \text{and}$$
$$\frac{1}{2} < \tau_3 < 2\tau_2 - 1.$$



Rates: Convex Case

- It is possible to show that by forming estimates

$$\hat{x}^n \triangleq \frac{1}{\lceil n/2 \rceil} \sum_{i=n-\lceil n/2 \rceil}^{n} x^i, \quad \forall n \in \mathbb{N}^+,$$

the *MESSAGE*^{p} algorithm satisfies

$$\mathbb{E}\left\{\phi^{\widetilde{F}}\left(\hat{x}^n\right) - \phi_*^{\widetilde{F}}\right\} \equiv \begin{cases} \mathcal{O}\left(n^{-1/4}\right), & \text{if } p \equiv 1 \\ \mathcal{O}\left(n^{-1/8}\right) & \text{if } p > 1 \end{cases},$$

with stepsizes chosen as

$$\alpha_n \equiv \frac{1}{n^{3/4}} \quad \text{and} \quad \beta_n \equiv \frac{1}{n^{1/2}}, \quad \text{when } p \equiv 1$$

$$\alpha_n \equiv \frac{1}{n^{7/8}}, \quad \beta_n \equiv \frac{1}{n^{3/4}} \quad \text{and} \quad \gamma_n \equiv \frac{1}{n^{1/2}}, \quad \text{when } p > 1$$

- **Convergence in the general case is slow**, especially when $p > 1$.
- Can we do better?

Rates: Convex Case

- It is possible to show that by forming estimates

$$\hat{x}^n \triangleq \frac{1}{\lceil n/2 \rceil} \sum_{i=n-\lceil n/2 \rceil}^{n} x^i, \quad \forall n \in \mathbb{N}^+,$$

the *MESSAGE*^{p} algorithm satisfies

$$\mathbb{E}\left\{\phi^{\widetilde{F}}\left(\hat{x}^n\right) - \phi_*^{\widetilde{F}}\right\} \equiv \begin{cases} \mathcal{O}\left(n^{-1/4}\right), & \text{if } p \equiv 1 \\ \mathcal{O}\left(n^{-1/8}\right) & \text{if } p > 1 \end{cases},$$

with stepsizes chosen as

$$\alpha_n \equiv \frac{1}{n^{3/4}} \quad \text{and} \quad \beta_n \equiv \frac{1}{n^{1/2}}, \quad \text{when } p \equiv 1$$

$$\alpha_n \equiv \frac{1}{n^{7/8}}, \quad \beta_n \equiv \frac{1}{n^{3/4}} \quad \text{and} \quad \gamma_n \equiv \frac{1}{n^{1/2}}, \quad \text{when } p > 1$$

- **Convergence in the general case is slow**, especially when $p > 1$.
- Can we do better?

Rates: Strongly Convex Case - 1

- Assume that $\phi^{\widetilde{F}}$ is strongly convex in the sense that there exists $\sigma > 0$, such that

$$\phi^{\widetilde{F}}(x) - \phi_*^{\widetilde{F}} \geq \sigma \left\| x - x^* \right\|_2^2, \quad \forall x \in \mathcal{X}.$$

- But how can we guarantee such a condition?

Proposition (When is (1) Strongly Convex?)

Fix $p \in [1, \infty)$ and choose any risk regularizer $\mathcal{R} : \mathbb{R} \to \mathbb{R}$. Then, as long as $c \in [0, 1]$, and if, for $\sigma > 0$ and for every $w \in \mathbb{R}^M$, $F(\cdot, w)$ is σ-strongly convex on $\mathcal{X}$ (such that $F(\cdot, w) - \sigma \|\cdot\|_2^2$ is convex on $\mathcal{X}$), the composite function $\phi^{\widetilde{F}}(\cdot) \equiv \rho(F(\cdot, W))$ is also σ-strongly convex on $\mathcal{X}$.

- Proof is very simple and is due to the risk measure ρ being convex (convex, monotone and translation equivariant):

$$\phi^{\widetilde{F}}(\cdot) - \sigma \|\cdot\|_2^2 = \rho(F(\cdot, W)) - \sigma \|\cdot\|_2^2 \equiv \rho\left(F(\cdot, W) - \sigma \|\cdot\|_2^2\right)$$

Rates: Strongly Convex Case - 1

- Assume that $\phi^{\widetilde{F}}$ is strongly convex in the sense that there exists $\sigma > 0$, such that

$$\phi^{\widetilde{F}}(x) - \phi_*^{\widetilde{F}} \geq \sigma \left\| x - x^* \right\|_2^2, \quad \forall x \in \mathcal{X}.$$

- But how can we guarantee such a condition?

Proposition (When is (1) Strongly Convex?)

Fix $p \in [1, \infty)$ and choose any risk regularizer $\mathcal{R} : \mathbb{R} \to \mathbb{R}$. Then, as long as $c \in [0, 1]$, and if, for $\sigma > 0$ and for every $w \in \mathbb{R}^M$, $F(\cdot, w)$ is σ-strongly convex on $\mathcal{X}$ (such that $F(\cdot, w) - \sigma \|\cdot\|_2^2$ is convex on $\mathcal{X}$), the composite function $\phi^{\widetilde{F}}(\cdot) \equiv \rho(F(\cdot, W))$ is also σ-strongly convex on $\mathcal{X}$.

- Proof is very simple and is due to the risk measure ρ being convex (convex, monotone and translation equivariant):

$$\phi^{\widetilde{F}}(\cdot) - \sigma \|\cdot\|_2^2 = \rho(F(\cdot, W)) - \sigma \|\cdot\|_2^2 \equiv \rho\left(F(\cdot, W) - \sigma \|\cdot\|_2^2\right)$$

Rates: Strongly Convex Case - 1

- Assume that $\phi^{\widetilde{F}}$ is strongly convex in the sense that there exists $\sigma > 0$, such that

$$\phi^{\widetilde{F}}(x) - \phi_*^{\widetilde{F}} \geq \sigma \left\| x - x^* \right\|_2^2, \quad \forall x \in \mathcal{X}.$$

- But how can we guarantee such a condition?

Proposition (When is (1) Strongly Convex?)

Fix $p \in [1, \infty)$ and choose any risk regularizer $\mathcal{R} : \mathbb{R} \to \mathbb{R}$. Then, as long as $c \in [0, 1]$, and if, for $\sigma > 0$ and for every $w \in \mathbb{R}^M$, $F(\cdot, w)$ is σ-strongly convex on $\mathcal{X}$ (such that $F(\cdot, w) - \sigma \|\cdot\|_2^2$ is convex on $\mathcal{X}$), the composite function $\phi^{\widetilde{F}}(\cdot) \equiv \rho(F(\cdot, W))$ is also σ-strongly convex on $\mathcal{X}$.

- Proof is very simple and is due to the risk measure ρ being convex (convex, monotone and translation equivariant):

$$\phi^{\widetilde{F}}(\cdot) - \sigma \|\cdot\|_2^2 = \rho(F(\cdot, W)) - \sigma \|\cdot\|_2^2 \equiv \rho\left(F(\cdot, W) - \sigma \|\cdot\|_2^2\right)$$

Rates: Strongly Convex Case - 2

Theorem (Rate | Strongly Convex Case | Subharmonic Stepsizes | $p > 1$)

Suppose that $\phi^{\widetilde{F}}$ is σ-strongly convex, and that

$$\alpha_n \triangleq \frac{1}{\sigma n}, \quad \beta_n \triangleq \frac{1}{n^{\tau_2}} \quad \text{and} \quad \gamma_n \triangleq \frac{1}{n^{\tau_3}}, \quad \forall n \in \mathbb{N}^+,$$

where $1/2 \leq \tau_3 < \tau_2 < 1$, and with initial values $\alpha_0 \equiv \beta_0 \equiv \gamma_0 \equiv 1$. Define

$$n_o(\tau_2) \equiv \left\lceil \left(1 - \tau_2^{1/(\tau_2+1)}\right)^{-1} \right\rceil \in \mathbb{N}^3 \text{ and } \mathrm{R}(\tau_2, \tau_3) \triangleq \frac{1}{1 - \max\{2 - 2\tau_2, 2\tau_2 - 2\tau_3, \tau_3\}} > 1.$$

Then, for every $n \in \mathbb{N}^{n_o(\tau_2)}$, it is true that

$$\mathbb{E}\left\{\left\|x^{n+1} - x^*\right\|_2^2\right\} \leq \widehat{\Sigma}\frac{n_o(\tau_2)}{n} + \widehat{\Sigma}\frac{\mathrm{R}(\tau_2, \tau_3)}{n^{2\min\{1-\tau_2, \tau_2-\tau_3\}}},$$

for some constant $0 < \widehat{\Sigma} < \infty$. In particular, if, for some $\epsilon \in [0, 1/4)$ and $\delta \in (0, 2)$,

$$\tau_2 \equiv 3/4 + \epsilon \quad \text{and} \quad \tau_3 \equiv 1/2 + \delta\epsilon,$$

then the MESSAGE^p algorithm satisfies

$$\mathcal{O}\left(n^{-(1-4\epsilon)/2}\right) \equiv \mathbb{E}\left\{\left\|x^{n+1} - x^*\right\|_2^2\right\} \leq \frac{\widehat{\Sigma}\left(n_o(\epsilon) + \frac{2}{1-4\epsilon}\right)}{n^{(1-4\epsilon)/2}},$$

for every $n \in \mathbb{N}^{n_o(\epsilon)}$, for each fixed ϵ.

Rates: Strongly Convex Case - 3

Theorem (Rate | Strongly Convex Case | Subharmonic Stepsizes | $p \equiv 1$)

Suppose that $\phi^{\widetilde{F}}$ is σ-strongly convex, and that

$$\alpha_n \triangleq \frac{1}{\sigma n}, \quad \text{and} \quad \beta_n \triangleq \frac{1}{n^{\tau_2}} \quad \forall n \in \mathbb{N}^+,$$

where $1/2 < \tau_2 < 1$, and with initial values $\alpha_0 \equiv \beta_0 \equiv 1$. Choose $n_o \equiv n_o(\tau_2)$ as in the previous slide, and define

$$\mathrm{R}(\tau_2) \triangleq \frac{1}{1 - \max\{2 - 2\tau_2, \tau_2\}} > 1.$$

Then, for every $n \in \mathbb{N}^{n_o(\tau_2)}$, it is true that

$$\mathbb{E}\left\{\left\|x^{n+1} - x^*\right\|_2^2\right\} \leq \frac{\overline{\Sigma}\left(n_o(\tau_2) + \mathrm{R}(\tau_2)\right)}{n^{\min\{2-2\tau_2, \tau_2\}}},$$

for some constant $0 < \widehat{\Sigma} < \infty$. In particular, the exponent in the denominator is maximized at $\tau_2^* \equiv 2/3$, yielding a rate of the order of $\mathcal{O}\left(n^{-2/3}\right)$.

Rates: Strongly Convex Case - 4

For $p > 1$:

- If $\epsilon \equiv 0$, the rate attained is of the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - 4x improvement compared to the convex case!
  - However, pathwise convergence **is not guaranteed**.
- If $\epsilon > 0$, the rate attained is **arbitrarily close** to the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - And pathwise convergence **is guaranteed**.

For $p \equiv 1$:

- The rate attained is of the order of $\mathcal{O}\left(n^{-2/3}\right)$.
  - 2.6x improvement compared to the convex case!
  - And pathwise convergence **is guaranteed**.

Rates: Strongly Convex Case - 4

For $p > 1$:

- If $\epsilon \equiv 0$, the rate attained is of the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - 4x improvement compared to the convex case!
  - However, pathwise convergence **is not guaranteed**.
- If $\epsilon > 0$, the rate attained is **arbitrarily close** to the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - And pathwise convergence **is guaranteed**.

For $p \equiv 1$:

- The rate attained is of the order of $\mathcal{O}\left(n^{-2/3}\right)$.
  - 2.6x improvement compared to the convex case!
  - And pathwise convergence **is guaranteed**.

Rates: Strongly Convex Case - 4

For $p > 1$:

- If $\epsilon \equiv 0$, the rate attained is of the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - 4x improvement compared to the convex case!
  - However, pathwise convergence **is not guaranteed**.
- If $\epsilon > 0$, the rate attained is **arbitrarily close** to the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - And pathwise convergence **is guaranteed**.

For $p \equiv 1$:

- The rate attained is of the order of $\mathcal{O}\left(n^{-2/3}\right)$.
  - 2.6x improvement compared to the convex case!
  - And pathwise convergence **is guaranteed**.

Rates: Strongly Convex Case - 4

For $p > 1$:

- If $\epsilon \equiv 0$, the rate attained is of the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - 4x improvement compared to the convex case!
  - However, pathwise convergence **is not guaranteed**.
- If $\epsilon > 0$, the rate attained is **arbitrarily close** to the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - And pathwise convergence **is guaranteed**.

For $p \equiv 1$:

- The rate attained is of the order of $\mathcal{O}\left(n^{-2/3}\right)$.
  - 2.6x improvement compared to the convex case!
  - And pathwise convergence **is guaranteed**.

Rates: Strongly Convex Case - 4

For $p > 1$:

- If $\epsilon \equiv 0$, the rate attained is of the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - 4x **improvement** compared to the convex case!
  - However, pathwise convergence **is not guaranteed**.
- If $\epsilon > 0$, the rate attained is **arbitrarily close** to the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - And pathwise convergence **is guaranteed**.

For $p \equiv 1$:

- The rate attained is of the order of $\mathcal{O}\left(n^{-2/3}\right)$.
  - 2.6x **improvement** compared to the convex case!
  - And pathwise convergence **is guaranteed**.

Rates: Strongly Convex Case - 4

For $p > 1$:

- If $\epsilon \equiv 0$, the rate attained is of the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - 4x **improvement** compared to the convex case!
  - However, pathwise convergence **is not guaranteed**.
- If $\epsilon > 0$, the rate attained is **arbitrarily close** to the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - And pathwise convergence **is guaranteed**.

For $p \equiv 1$:

- The rate attained is of the order of $\mathcal{O}\left(n^{-2/3}\right)$.
  - 2.6x **improvement** compared to the convex case!
  - And pathwise convergence **is guaranteed**.

Rates: Strongly Convex Case - 4

For $p > 1$:

- If $\epsilon \equiv 0$, the rate attained is of the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - 4x **improvement** compared to the convex case!
  - However, pathwise convergence **is not guaranteed**.
- If $\epsilon > 0$, the rate attained is **arbitrarily close** to the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - And pathwise convergence **is guaranteed**.

For $p \equiv 1$:

- The rate attained is of the order of $\mathcal{O}\left(n^{-2/3}\right)$.
  - 2.6x **improvement** compared to the convex case!
  - And pathwise convergence **is guaranteed**.

Rates: Strongly Convex Case - 4

For $p > 1$:

- If $\epsilon \equiv 0$, the rate attained is of the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - 4x **improvement** compared to the convex case!
  - However, pathwise convergence **is not guaranteed**.
- If $\epsilon > 0$, the rate attained is **arbitrarily close** to the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - And pathwise convergence **is guaranteed**.

For $p \equiv 1$:

- The rate attained is of the order of $\mathcal{O}\left(n^{-2/3}\right)$.
  - 2.6x **improvement** compared to the convex case!
  - And pathwise convergence **is guaranteed**.

Rates: Strongly Convex Case - 4

For $p > 1$:

- If $\epsilon \equiv 0$, the rate attained is of the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - 4x **improvement** compared to the convex case!
  - However, pathwise convergence **is not guaranteed**.
- If $\epsilon > 0$, the rate attained is **arbitrarily close** to the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - And pathwise convergence **is guaranteed**.

For $p \equiv 1$:

- The rate attained is of the order of $\mathcal{O}\left(n^{-2/3}\right)$.
  - 2.6x **improvement** compared to the convex case!
  - And pathwise convergence **is guaranteed**.

Rates: Strongly Convex Case - 4

For $p > 1$:

- If $\epsilon \equiv 0$, the rate attained is of the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - 4x improvement compared to the convex case!
  - However, pathwise convergence **is not guaranteed**.
- If $\epsilon > 0$, the rate attained is **arbitrarily close** to the order of $\mathcal{O}\left(n^{-1/2}\right)$.
  - And pathwise convergence **is guaranteed**.

For $p \equiv 1$:

- The rate attained is of the order of $\mathcal{O}\left(n^{-2/3}\right)$.
  - 2.6x improvement compared to the convex case!
  - And pathwise convergence **is guaranteed**.

Comparison with the State of the Art

- Most of the results presented may be proven utilizing the **generic** problem framework of [Yang et al., 2018].
- Question: Are we doing better in terms of the variety of problems our proposed framework supports?
- YES!

Proposition (Structural Comparisons)

Under very mild technical assumptions: The class of mean-semideviation programs supported under the proposed framework contains the respective class supported in [Yang et al., 2018]. Further, the inclusion is strict.

- If fact, under the framework of [Yang et al., 2018]:
  - $(\mathcal{R}(\cdot))^p$ must be differentiable everywhere.
  - $(\mathcal{R}(\cdot))^p$ must be Lipschitz.
  - $F(\cdot, W)$ must be differentiable everywhere on $\mathcal{X}$. If not, $\mathcal{R}$ must be partially constant!
- Our framework is free of all those restrictions.

Comparison with the State of the Art

- Most of the results presented may be proven utilizing the **generic** problem framework of [Yang et al., 2018].
- **Question:** Are we doing better in terms of the variety of problems our proposed framework supports?
- YES!

Proposition (Structural Comparisons)

Under very mild technical assumptions: The class of mean-semideviation programs supported under the proposed framework contains the respective class supported in [Yang et al., 2018]. Further, the inclusion is strict.

- If fact, under the framework of [Yang et al., 2018]:
  - $(\mathcal{R}(\cdot))^p$ must be differentiable everywhere.
  - $(\mathcal{R}(\cdot))^p$ must be Lipschitz.
  - $F(\cdot, W)$ must be differentiable everywhere on $\mathcal{X}$. If not, $\mathcal{R}$ must be partially constant!
- Our framework is free of all those restrictions.

Comparison with the State of the Art

- Most of the results presented may be proven utilizing the **generic** problem framework of [Yang et al., 2018].
- Question: Are we doing better in terms of the variety of problems our proposed framework supports?
- YES!

Proposition (Structural Comparisons)

Under very mild technical assumptions: The class of mean-semideviation programs supported under the proposed framework contains the respective class supported in [Yang et al., 2018]. Further, the inclusion is strict.

- If fact, under the framework of [Yang et al., 2018]:
  - $(\mathcal{R}(\cdot))^p$ must be differentiable everywhere.
  - $(\mathcal{R}(\cdot))^p$ must be Lipschitz.
  - $F(\cdot, W)$ must be differentiable everywhere on $\mathcal{X}$. If not, $\mathcal{R}$ must be partially constant!
- Our framework is free of all those restrictions.

Comparison with the State of the Art

- Most of the results presented may be proven utilizing the **generic** problem framework of [Yang et al., 2018].
- Question: Are we doing better in terms of the variety of problems our proposed framework supports?
- YES!

Proposition (Structural Comparisons)

Under very mild technical assumptions: The class of mean-semideviation programs supported under the proposed framework contains the respective class supported in [Yang et al., 2018]. Further, the inclusion is strict.

- If fact, under the framework of [Yang et al., 2018]:
  - $(\mathcal{R}(\cdot))^p$ must be differentiable everywhere.
  - $(\mathcal{R}(\cdot))^p$ must be Lipschitz.
  - $F(\cdot, W)$ must be differentiable everywhere on $\mathcal{X}$. If not, $\mathcal{R}$ must be partially constant!
- Our framework is free of all those restrictions.

Comparison with the State of the Art

- Most of the results presented may be proven utilizing the **generic** problem framework of [Yang et al., 2018].
- Question: Are we doing better in terms of the variety of problems our proposed framework supports?
- YES!

Proposition (Structural Comparisons)

Under very mild technical assumptions: The class of mean-semideviation programs supported under the proposed framework contains the respective class supported in [Yang et al., 2018]. Further, the inclusion is strict.

- If fact, under the framework of [Yang et al., 2018]:
  - $(\mathcal{R}(\cdot))^p$ must be differentiable everywhere.
  - $(\mathcal{R}(\cdot))^p$ must be Lipschitz.
  - $F(\cdot, W)$ must be differentiable everywhere on $\mathcal{X}$. If not, $\mathcal{R}$ must be partially constant!
- Our framework is free of all those restrictions.

Comparison with the State of the Art

- Most of the results presented may be proven utilizing the **generic** problem framework of [Yang et al., 2018].
- Question: Are we doing better in terms of the variety of problems our proposed framework supports?
- YES!

Proposition (Structural Comparisons)

Under very mild technical assumptions: The class of mean-semideviation programs supported under the proposed framework contains the respective class supported in [Yang et al., 2018]. Further, the inclusion is strict.

- If fact, under the framework of [Yang et al., 2018]:
  - $(\mathcal{R}(\cdot))^p$ must be differentiable everywhere.
  - $(\mathcal{R}(\cdot))^p$ must be Lipschitz.
  - $F(\cdot, W)$ must be differentiable everywhere on $\mathcal{X}$. If not, $\mathcal{R}$ must be partially constant!
- Our framework is free of all those restrictions.

Comparison with the State of the Art

- Most of the results presented may be proven utilizing the **generic** problem framework of [Yang et al., 2018].
- Question: Are we doing better in terms of the variety of problems our proposed framework supports?
- YES!

Proposition (Structural Comparisons)

Under very mild technical assumptions: The class of mean-semideviation programs supported under the proposed framework contains the respective class supported in [Yang et al., 2018]. Further, the inclusion is strict.

- If fact, under the framework of [Yang et al., 2018]:
  - $(\mathcal{R}(\cdot))^p$ must be differentiable everywhere.
  - $(\mathcal{R}(\cdot))^p$ must be Lipschitz.
  - $F(\cdot, W)$ must be differentiable everywhere on $\mathcal{X}$. If not, $\mathcal{R}$ must be partially constant!
- Our framework is free of all those restrictions.

Comparison with the State of the Art

- Most of the results presented may be proven utilizing the **generic** problem framework of [Yang et al., 2018].
- Question: Are we doing better in terms of the variety of problems our proposed framework supports?
- YES!

Proposition (Structural Comparisons)

Under very mild technical assumptions: The class of mean-semideviation programs supported under the proposed framework contains the respective class supported in [Yang et al., 2018]. Further, the inclusion is strict.

- If fact, under the framework of [Yang et al., 2018]:
  - $(\mathcal{R}(\cdot))^p$ must be differentiable everywhere.
  - $(\mathcal{R}(\cdot))^p$ must be Lipschitz.
  - $F(\cdot, W)$ must be differentiable everywhere on $\mathcal{X}$. If not, $\mathcal{R}$ must be partially constant!
- Our framework is free of all those restrictions.

Conclusions

- We introduced and analyzed the ***MESSAGE^p algorithm***, an efficient, data-driven procedure for iteratively solving convex mean-semideviation problems to optimality.
- We proposed a **new framework** of structural assumptions, under which:
  - We established pathwise convergence of the *MESSAGE^p* algorithm, in the strong sense of [Yang et al., 2018].
  - We derived explicit **convergence rates** of the *MESSAGE^p* algorithm, matching and improving the state of the art.
- Our framework:
  - **provably strictly generalizes that** of [Yang et al., 2018].
  - reveals a well-defined structural trade-off between the random cost and the mean-semideviation risk measure considered.
- Many directions for future research:
  - **Adaptive** risk-averse optimization in **nonstationary settings**.
  - Risk-averse **reinforcement learning** and **approximate dynamic programming**.
  - **Acceleration** for problems with more favorable structure (second-order differentiability, etc.)
  - Applications: **Supply chain management**, **multi-agent network control** and **resource allocation**.

Conclusions

- We introduced and analyzed the ***MESSAGE^p algorithm***, an efficient, data-driven procedure for iteratively solving convex mean-semideviation problems to optimality.
- We proposed a **new framework** of structural assumptions, under which:
  - We established pathwise convergence of the *MESSAGE^p* algorithm, in the strong sense of [Yang et al., 2018].
  - We derived explicit **convergence rates** of the *MESSAGE^p* algorithm, matching and improving the state of the art.
- Our framework:
  - **provably strictly generalizes that** of [Yang et al., 2018].
  - reveals a well-defined structural trade-off between the random cost and the mean-semideviation risk measure considered.
- Many directions for future research:
  - **Adaptive** risk-averse optimization in **nonstationary settings**.
  - Risk-averse **reinforcement learning** and **approximate dynamic programming**.
  - **Acceleration** for problems with more favorable structure (second-order differentiability, etc.)
  - Applications: **Supply chain management**, **multi-agent network control** and **resource allocation**.

Conclusions

- We introduced and analyzed the ***MESSAGE^p algorithm***, an efficient, data-driven procedure for iteratively solving convex mean-semideviation problems to optimality.
- We proposed a **new framework** of structural assumptions, under which:
  - We established pathwise convergence of the *MESSAGE^p* algorithm, in the strong sense of [Yang et al., 2018].
  - We derived explicit **convergence rates** of the *MESSAGE^p* algorithm, matching and improving the state of the art.
- Our framework:
  - **provably strictly generalizes that** of [Yang et al., 2018].
  - reveals a well-defined structural trade-off between the random cost and the mean-semideviation risk measure considered.
- Many directions for **future research**:
  - **Adaptive** risk-averse optimization in **nonstationary settings**.
  - Risk-averse **reinforcement learning** and **approximate dynamic programming**.
  - **Acceleration** for problems with more favorable structure (second-order differentiability, etc.)
  - Applications: **Supply chain management**, **multi-agent network control** and **resource allocation**.

Conclusions

- We introduced and analyzed the ***MESSAGE^p algorithm***, an efficient, data-driven procedure for iteratively solving convex mean-semideviation problems to optimality.
- We proposed a **new framework** of structural assumptions, under which:
  - We established pathwise convergence of the *MESSAGE^p* algorithm, in the strong sense of [Yang et al., 2018].
  - We derived explicit **convergence rates** of the *MESSAGE^p* algorithm, matching and improving the state of the art.
- Our framework:
  - **provably strictly generalizes that** of [Yang et al., 2018].
  - reveals a well-defined structural trade-off between the random cost and the mean-semideviation risk measure considered.
- Many directions for **future research**:
  - **Adaptive** risk-averse optimization in **nonstationary settings**.
  - Risk-averse **reinforcement learning** and **approximate dynamic programming**.
  - **Acceleration** for problems with more favorable structure (second-order differentiability, etc.)
  - Applications: **Supply chain management**, **multi-agent network control** and **resource allocation**.

Conclusions

- We introduced and analyzed the ***MESSAGE^p algorithm***, an efficient, data-driven procedure for iteratively solving convex mean-semideviation problems to optimality.
- We proposed a **new framework** of structural assumptions, under which:
  - We established pathwise convergence of the *MESSAGE^p* algorithm, in the strong sense of [Yang et al., 2018].
  - We derived explicit **convergence rates** of the *MESSAGE^p* algorithm, matching and improving the state of the art.
- Our framework:
  - **provably strictly generalizes that** of [Yang et al., 2018].
  - reveals a well-defined structural trade-off between the random cost and the mean-semideviation risk measure considered.
- Many directions for **future research**:
  - **Adaptive** risk-averse optimization in **nonstationary settings**.
  - Risk-averse **reinforcement learning** and **approximate dynamic programming**.
  - **Acceleration** for problems with more favorable structure (second-order differentiability, etc.)
  - Applications: **Supply chain management**, **multi-agent network control** and **resource allocation**.

Conclusions

- We introduced and analyzed the ***MESSAGE^p algorithm***, an efficient, data-driven procedure for iteratively solving convex mean-semideviation problems to optimality.
- We proposed a **new framework** of structural assumptions, under which:
  - We established pathwise convergence of the *MESSAGE^p* algorithm, in the strong sense of [Yang et al., 2018].
  - We derived explicit **convergence rates** of the *MESSAGE^p* algorithm, matching and improving the state of the art.
- Our framework:
  - **provably strictly generalizes that** of [Yang et al., 2018].
  - reveals a well-defined structural trade-off between the random cost and the mean-semideviation risk measure considered.
- Many directions for **future research**:
  - **Adaptive** risk-averse optimization in **nonstationary settings**.
  - Risk-averse **reinforcement learning** and **approximate dynamic programming**.
  - **Acceleration** for problems with more favorable structure (second-order differentiability, etc.)
  - Applications: **Supply chain management**, **multi-agent network control** and **resource allocation**.

Conclusions

- We introduced and analyzed the ***MESSAGE^p algorithm***, an efficient, data-driven procedure for iteratively solving convex mean-semideviation problems to optimality.
- We proposed a **new framework** of structural assumptions, under which:
  - We established pathwise convergence of the *MESSAGE^p* algorithm, in the strong sense of [Yang et al., 2018].
  - We derived explicit **convergence rates** of the *MESSAGE^p* algorithm, matching and improving the state of the art.
- Our framework:
  - **provably strictly generalizes that** of [Yang et al., 2018].
  - reveals a well-defined structural trade-off between the random cost and the mean-semideviation risk measure considered.
- Many directions for **future research**:
  - **Adaptive** risk-averse optimization in **nonstationary settings**.
  - Risk-averse **reinforcement learning** and **approximate dynamic programming**.
  - **Acceleration** for problems with more favorable structure (second-order differentiability, etc.)
  - Applications: **Supply chain management**, **multi-agent network control** and **resource allocation**.

Thank You!

Questions?

BackUp!

How "hard" condition C3 really is?

Proposition (Ensuring Validity of C3)

Assume that, whenever $p > 1$, condition **C4** *is satisfied. Then, the following are true:*

1. Suppose that $(\mathcal{R}(\cdot))^p$ is differentiable on $\mathfrak{R}^{\widetilde{F}}$, and there is $D_{\mathcal{R},p} < \infty$, such that

$$\left|\nabla(\mathcal{R}(y_1))^p - \nabla(\mathcal{R}(y_1))^p\right| \leq D_{\mathcal{R},p}\left|y_1 - y_2\right|, \quad \forall (y_1, y_2) \in \left[\mathfrak{R}^{\widetilde{F}}\right]^2.$$

Then, condition **C3** *is satisfied for every $Q \in [P/(P-1), \infty]$, for every $P \in [2, \infty]$.*

2. Choose $\underline{\nabla}\mathcal{R} \equiv \mathcal{R}'_+$, and if $F_W^{(\cdot)} : \mathbb{R} \to [0,1]$ denotes the cdf of $F(\cdot, W)$, suppose that there exists $D_{\widetilde{F}} < \infty$, such that

$$\sup_{x \in \mathcal{X}} \left|F_W^x(y_1) - F_W^x(y_2)\right| \leq D_{\widetilde{F}}\left|y_1 - y_2\right|, \quad \forall (y_1, y_2) \in \left[\mathrm{cl}\left\{(m_l, m_h)\right\}\right]^2.$$

Then, condition **C3** *is satisfied for $Q \equiv 1$ (implying that $P \equiv \infty$), for every $p \in [1, \infty)$.*

- In fact, more is true!

  But we need more notation...

How "hard" condition C3 really is?

Proposition (Ensuring Validity of C3)

Assume that, whenever $p > 1$, condition **C4** *is satisfied. Then, the following are true:*

1. Suppose that $(\mathcal{R}(\cdot))^p$ is differentiable on $\mathfrak{R}^{\widetilde{F}}$, and there is $D_{\mathcal{R},p} < \infty$, such that

$$\left|\nabla(\mathcal{R}(y_1))^p - \nabla(\mathcal{R}(y_1))^p\right| \leq D_{\mathcal{R},p}\left|y_1 - y_2\right|, \quad \forall (y_1, y_2) \in \left[\mathfrak{R}^{\widetilde{F}}\right]^2.$$

Then, condition **C3** *is satisfied for every $Q \in [P/(P-1), \infty]$, for every $P \in [2, \infty]$.*

2. Choose $\underline{\nabla}\mathcal{R} \equiv \mathcal{R}'_+$, and if $F_W^{(\cdot)} : \mathbb{R} \to [0,1]$ denotes the cdf of $F(\cdot, W)$, suppose that there exists $D_{\widetilde{F}} < \infty$, such that

$$\sup_{x \in \mathcal{X}} \left|F_W^x(y_1) - F_W^x(y_2)\right| \leq D_{\widetilde{F}}\left|y_1 - y_2\right|, \quad \forall (y_1, y_2) \in \left[\mathrm{cl}\left\{(m_l, m_h)\right\}\right]^2.$$

Then, condition **C3** *is satisfied for $Q \equiv 1$ (implying that $P \equiv \infty$), for every $p \in [1, \infty)$.*

- In fact, more is true!

  But we need more notation...