The Era of "Non"-Optimization Problems

Jong-Shi Pang

Daniel J. Epstein Department of Industrial and Systems Engineering

University of Southern California

presented at

OneWorld Optimization Seminar Series

3:30–4:30 PM CEST Monday September 14 2020

Based on joint work with Dr. Ying Cui at the University of Minnesota and many collaborators.

Contents of presentation

- Some general thoughts of the state of the field

Contents of presentation

- Some general thoughts of the state of the field
- A readiness test (a light-hearted humor)

Contents of presentation

- Some general thoughts of the state of the field
- A readiness test (a light-hearted humor)
- Some key "non"-features

Contents of presentation

- Some general thoughts of the state of the field
- A readiness test (a light-hearted humor)
- Some key “non”-features
- Some novel “non”-optimization problems

Contents of presentation

- Some general thoughts of the state of the field
- A readiness test (a light-hearted humor)
- Some key “non”-features
- Some novel “non”-optimization problems
- A gentle touch of mathematics

Contents of presentation

- Some general thoughts of the state of the field
- A readiness test (a light-hearted humor)
- Some key "non"-features
- Some novel "non"-optimization problems
- A gentle touch of mathematics
- **The monograph**

Some General Thoughts

addressing the negative impact of the field of machine learning

on the optimization field,

and raising an alarm

- **Convexity and/or differentiability** have been the safe haven for optimization and its applications for decades, particularly popular in the machine learning in recent years

• **Convexity and/or differentiability** have been the safe haven for optimization and its applications for decades, particularly popular in the machine learning in recent years

to the extent that

• modellers are routinely sacrificing realism for convenience,

• algorithm designers will do whatever to stay in this comfort zone, even at the expense of potentially wrong approach,

• practitioners are content with the resulting models, algorithms, and "solutions",

• **Convexity and/or differentiability** have been the safe haven for optimization and its applications for decades, particularly popular in the machine learning in recent years

to the extent that

• modellers are routinely sacrificing realism for convenience,

• algorithm designers will do whatever to stay in this comfort zone, even at the expense of potentially wrong approach,

• practitioners are content with the resulting models, algorithms, and "solutions",

resulting in

• tools being abused,

• advances in science being stalled, and

• future being built on soft ground.

This talk

Benefits, yes or No?

Ready for the “non”-era, yes or no?

This talk

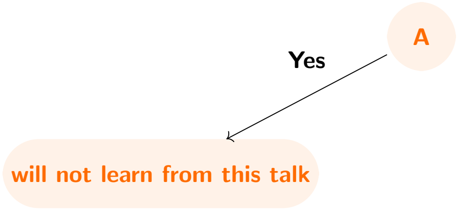
Benefits, yes or No?

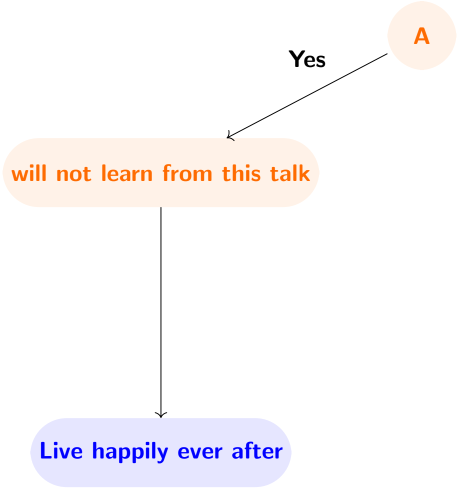
Ready for the “non”-era, yes or no?

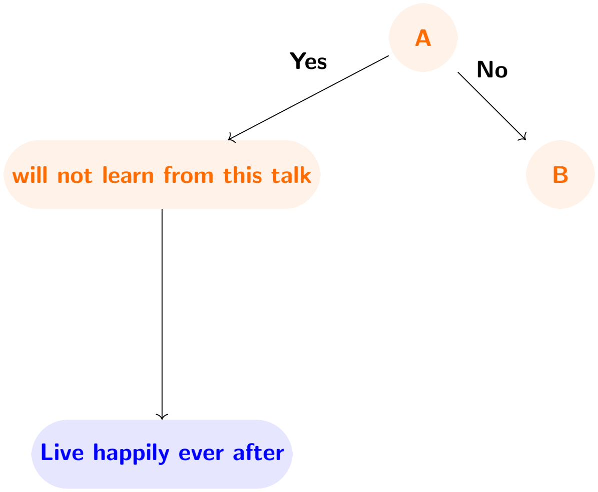
A little humor before the storm

A

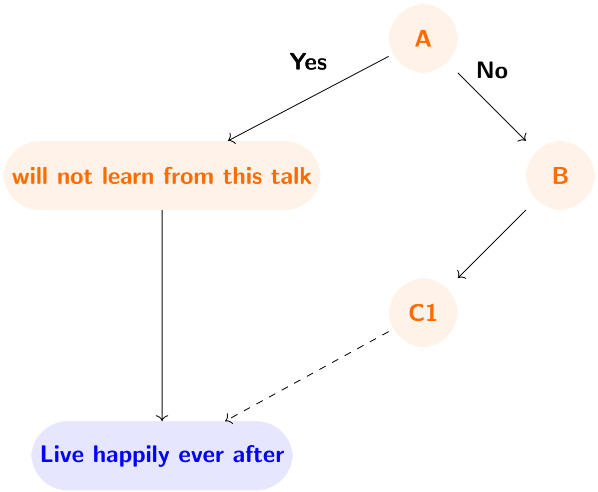
A: Are you satisfied with the comfort zone of convexity and differentiability?


A
Yes
will not learn from this talk

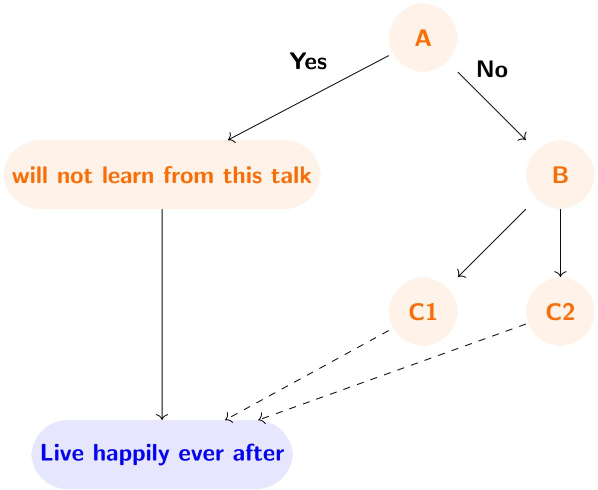

A
Yes
will not learn from this talk
Live happily ever after



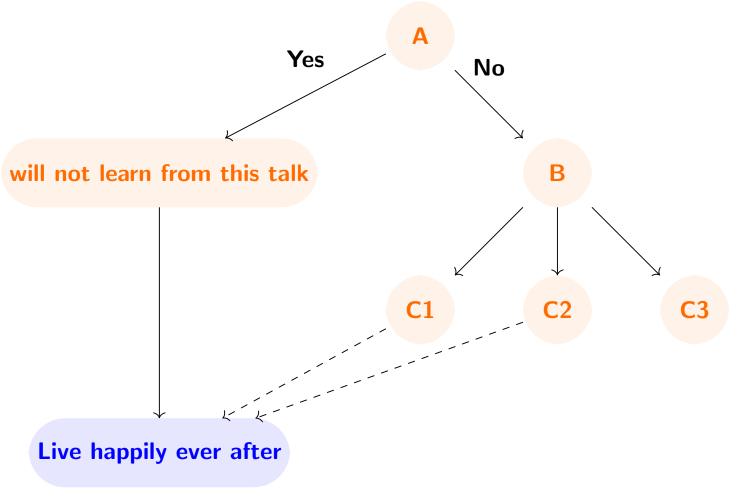
B: How to deal with non-convexity/non-smoothness?



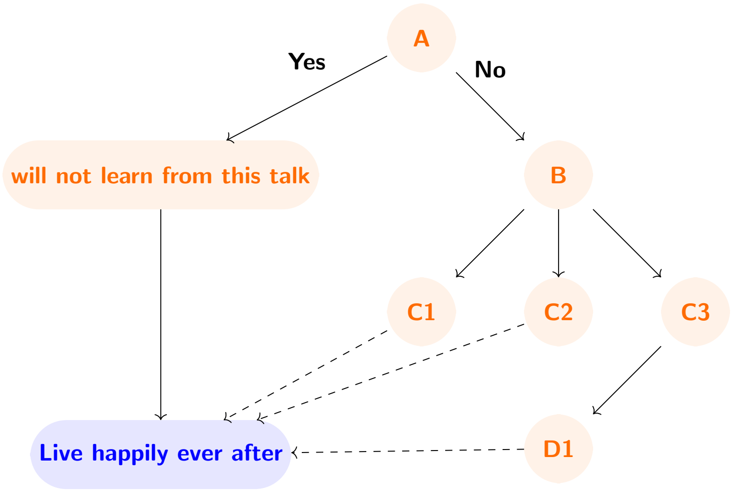
C1: Ignore these properties or give them minimal or even wrong treatment



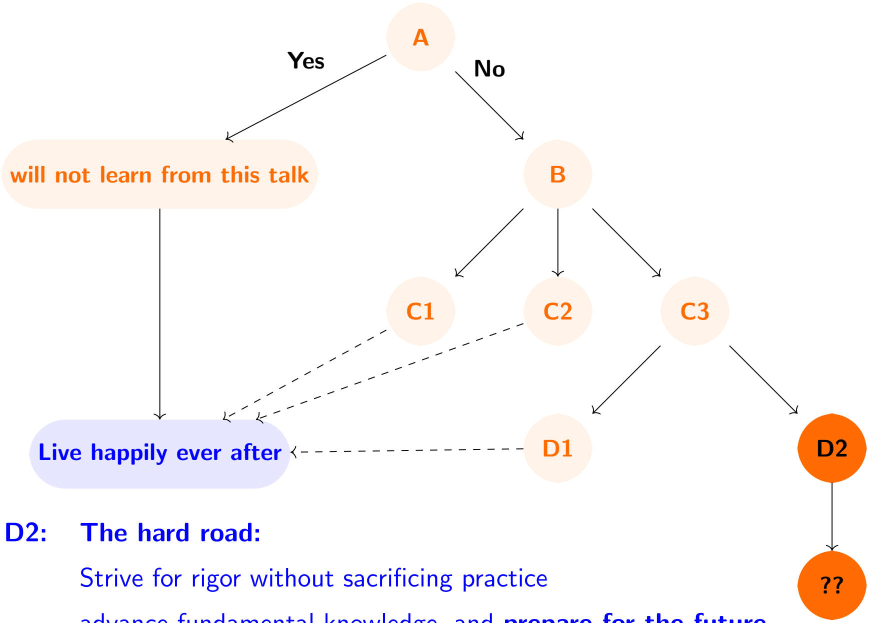
C2: Focus on stylized problems (and there are many interesting ones) and stop at them



C3: Faithfully treat these **non-properties**



D1: Employ heuristics, and potentially, leading to heuristics on heuristics when complex models are being formulated

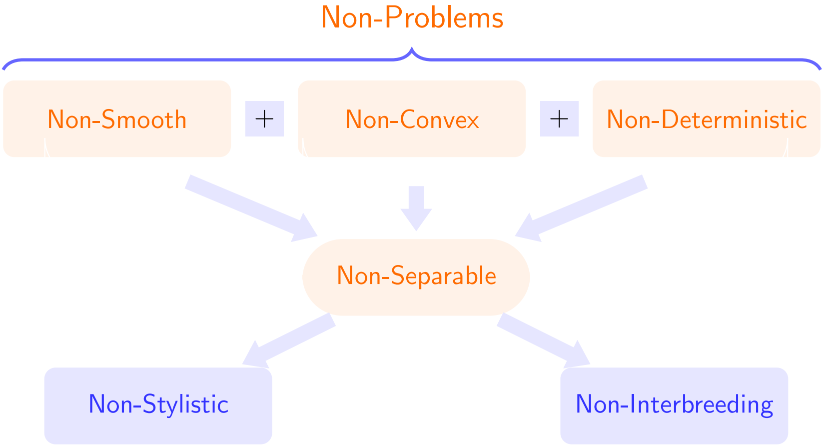


D2: The hard road:

Strive for rigor without sacrificing practice

advance fundamental knowledge, and **prepare for the future**

Features



and many more non-topics $\cdots$

Examples of Non-Problems

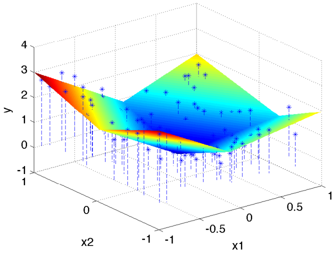
I. Structured Learning

Piecewise Affine Regression

extending classical affine regression

$$\underset{a,\alpha,b,\beta}{\textbf{minimize}} \; \frac{1}{N}\sum_{s=1}^{N}\left(y_s - \left[\underbrace{\max_{1\le i\le k_1}\left\{ (a^i)^\top x^s + \alpha_i \right\} - \max_{1\le i\le k_2}\left\{ (b^i)^\top x^s + \beta_i \right\}}_{\text{capturing all piecewise affine functions}} \right] \right)^2$$

illustrating coupled non-convexity and non-differentiability.



Estimate Change Points/Hyperplanes
in Linear Regression

Comments

- This is a convex quadratic composed with a piecewise affine function, thus is piecewise linear-quadratic.

Comments

- This is a convex quadratic composed with a piecewise affine function, thus is piecewise linear-quadratic.
- It has the property that every directional stationary solution is a local minimizer.

Comments

- This is a convex quadratic composed with a piecewise affine function, thus is piecewise linear-quadratic.
- It has the property that every directional stationary solution is a local minimizer.
- It has finitely many stationary values.

Comments

- This is a convex quadratic composed with a piecewise affine function, thus is piecewise linear-quadratic.

- It has the property that every directional stationary solution is a local minimizer.

- It has finitely many stationary values.

- A directional stationary solution can be computed by an iterative convex-programming based algorithm.

Comments

• This is a convex quadratic composed with a piecewise affine function, thus is piecewise linear-quadratic.

• It has the property that every directional stationary solution is a local minimizer.

• It has finitely many stationary values.

• A directional stationary solution can be computed by an iterative convex-programming based algorithm.

• Leads to several classes of nonconvex nonsmooth functions; most general is a difference-of-convex function composed with a difference-of-convex function, an example of which is

$$\varphi_1\left(\max_{1\leq i\leq k_{11}}\psi^1_{1i}(x)-\max_{1\leq i\leq k_{12}}\psi^1_{2i}(x)\right)-\varphi_2\left(\max_{1\leq i\leq k_{21}}\psi^2_{1i}(x)-\max_{1\leq i\leq k_2}\psi^2_{2i}(x)\right)$$

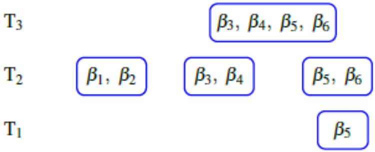
where all functions are convex.

Logical Sparsity Learning

$$\begin{aligned} \underset{\beta \in \mathbb{R}^d}{\textbf{minimize}} \quad & f(\beta) \\ \text{subject to} \quad & \sum_{j=1}^{d} a_{ij} \, | \, \beta_j \, |_0 \, \leq \, b_i, \quad i \, = \, 1, \cdots, m, \end{aligned}$$

where $| \, t \, |_0 \triangleq \begin{cases} 1 & \text{if } t \neq 0 \\ 0 & \text{otherwise.} \end{cases}$

- strong/weak hierarchy: e.g. $| \, \beta_j \, |_0 \leq | \, \beta_k \, |_0$
- cost-based variable selection: coefficients ≥ 0 but not all equal
- hierarchical/group variable selection



Comments

- The ℓ_0 function is discrete in nature; an ASC can be formulated either as complementarity constraints, or by integer variables under boundedness of the original variables (the big M-formulation), when A is nonnegative.

Comments

• The ℓ_0 function is discrete in nature; an ASC can be formulated either as complementarity constraints, or by integer variables under boundedness of the original variables (the big M-formulation), when A is nonnegative.

• Soft formulation of ASCs is possible, leading to an objective of the kind:

$$\underset{\beta \in \mathbb{R}^d}{\textbf{minimize}}\; f(\beta) + \gamma \sum_{i=1}^{m} \left[b_i - \sum_{j=1}^{d} a_{ij} \,|\, \beta_j \,|_0 \right]_+, \quad \text{for a given } \gamma > 0.$$

• When the matrix A has negative entries, the resulting ASC set may not closed; its optimization formulation may not have a lower semi-continuous objective.

• When $|\bullet|_0$ is approximated by a folded concave function, obtain some nondifferentiable, difference-of-convex constraints.

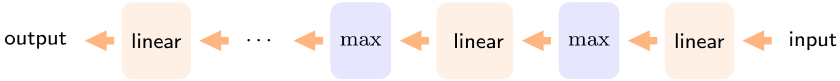
Deep Neural Network with Feed-Forward Structure

leading to multi-composite, nonconvex, nonsmooth optimization problems

$$\underset{\theta=\left\{\left(W^\ell,w^\ell\right)_{\ell=1}^L\right\}}{\textbf{minimize}} \quad \frac{1}{S}\sum_{s=1}^{S}\left[\, y_s - f_\theta(x^s)\,\right]^2$$

where f_θ is the L-fold composite function with the non-smooth max ReLU activation:

$$f_\theta(x) = \underbrace{\left(W^L \bullet + w^L\right)}_{\text{Output Layer}} \circ \cdots \circ \underbrace{\max\left(0, W^2 \bullet + w^2\right)}_{\text{Second Layer}} \circ \underbrace{\max\left(0, W^1 x + w^1\right)}_{\text{Input Layer}}$$



Comments

• Problem can be treated by exact penalization, a theory that addresses stationary solutions rather than minimizers.

Comments

• Problem can be treated by exact penalization, a theory that addresses stationary solutions rather than minimizers.

• Note the ℓ_1 penalization.

$$\boxed{\underset{z \in \mathcal{Z};\, u}{\textbf{minimize}}\ \zeta_\rho(z;u) \triangleq \frac{1}{S}\sum_{s=1}^{S}\left[\varphi_s(u^{s;L}) + \rho\,\underbrace{\sum_{\ell=1}^{L}\sum_{j=1}^{N_\ell}\left|\,u_j^{s;\ell} - \psi_j^\ell(z^\ell, u^{s;\ell-1})\,\right|}_{\text{penalizing the components}}\right]}$$

Comments

• Problem can be treated by exact penalization, a theory that addresses stationary solutions rather than minimizers.

• Note the ℓ_1 penalization.

$$\boxed{\underset{z \in \mathcal{Z};\, u}{\textbf{minimize}}\ \zeta_\rho(z;u) \triangleq \frac{1}{S}\sum_{s=1}^{S}\left[\varphi_s(u^{s;L}) + \underbrace{\rho\sum_{\ell=1}^{L}\sum_{j=1}^{N_\ell}\left|u_j^{s;\ell} - \psi_j^\ell(z^\ell, u^{s;\ell-1})\right|}_{\text{penalizing the components}}\right]}$$

• Rigorous algorithms with supporting convergence theory for computing a directional stationary point can be developed.

Comments

• Problem can be treated by exact penalization, a theory that addresses stationary solutions rather than minimizers.

• Note the ℓ_1 penalization.

$$\boxed{\underset{z \in \mathcal{Z};\, u}{\textbf{minimize}}\ \zeta_\rho(z; u) \triangleq \frac{1}{S} \sum_{s=1}^{S} \left[\varphi_s(u^{s;L}) + \rho \underbrace{\sum_{\ell=1}^{L} \sum_{j=1}^{N_\ell} \left| u_j^{s;\ell} - \psi_j^\ell(z^\ell, u^{s;\ell-1}) \right|}_{\text{penalizing the components}} \right]}$$

• Rigorous algorithms with supporting convergence theory for computing a directional stationary point can be developed.

• Framework allows the treatment of data perturbation.

Comments

• Problem can be treated by exact penalization, a theory that addresses stationary solutions rather than minimizers.

• Note the ℓ_1 penalization.

$$\boxed{\underset{z \in \mathcal{Z};\, u}{\textbf{minimize}}\ \zeta_\rho(z; u) \triangleq \frac{1}{S} \sum_{s=1}^{S} \left[\varphi_s(u^{s;L}) + \underbrace{\rho \sum_{\ell=1}^{L} \sum_{j=1}^{N_\ell} \left| u_j^{s;\ell} - \psi_j^\ell(z^\ell, u^{s;\ell-1}) \right|}_{\text{penalizing the components}} \right]}$$

• Rigorous algorithms with supporting convergence theory for computing a directional stationary point can be developed.

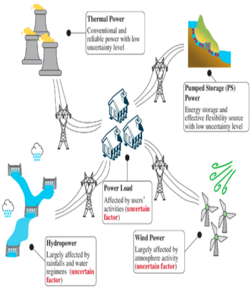
• Framework allows the treatment of data perturbation.

• Leads to an advanced study of exact penalization and error bounds of nondifferentiable optimization problems.

Examples of Non-Problems

II. Smart Planning

Power Systems Planning



ω_1: uncertain renewable energy generation
ω_2: uncertain demand of electricity

x: production from traditional thermal plants

y: production from backup fast-ramping generators

$$\underset{x \in X}{\textbf{minimize}} \quad \varphi(x) + \mathbb{E}_{\omega_1, \omega_2}\left[\psi(x, \omega_1, \omega_2)\right]$$

where the recourse function is

$$\psi(x, \omega_1, \omega_2) \triangleq \underset{y}{\textbf{minimum}}\, y^\top c\left([\omega_2 - x - \omega_1]_+\right)$$
$$\textbf{subject to} \quad A(\omega)x + Dy \geq \xi(\omega)$$

Unit cost of fast-ramping generators depends on

- observed renewable production & demand of electricity
- shortfall due to the first-stage thermal power decisions

Comments

• Leads to the study of the class of two-stage linearly bi-parameterized stochastic program with quadratic recourse:

$$\operatorname*{\mathbf{minimize}}_{x} \quad \zeta(x) \triangleq \varphi(x) + \mathbf{E}_{\tilde{\xi}}\left[\psi(x,\tilde{\xi})\right] \quad \textbf{subject to} \quad x \in X \subseteq \mathbb{R}^{n_1},$$

where the recourse function $\psi(x,\xi)$ is the optimal objective value of the quadratic program (QP): with Q being symmetric positive semidefinite:

$$\begin{array}{rll}\psi(x,\xi) \triangleq & \operatorname*{minimum}_{y} & \left[\, f(\xi) + G(\xi)x \,\right]^\top y + \frac{1}{2}\, y^\top Q y \\ & \text{subject to} & y \in Y(x,\xi) \triangleq \left\{ y \in \mathbb{R}^{n_2} \mid A(\xi)x + Dy \geq b(\xi) \right\}.\end{array}$$

Comments

- Leads to the study of the class of two-stage linearly bi-parameterized stochastic program with quadratic recourse:

$$\underset{x}{\textbf{minimize}} \quad \zeta(x) \triangleq \varphi(x) + \mathbf{E}_{\tilde{\xi}}\left[\psi(x,\tilde{\xi})\right] \quad \textbf{subject to} \quad x \in X \subseteq \mathbb{R}^{n_1},$$

where the recourse function $\psi(x,\xi)$ is the optimal objective value of the quadratic program (QP): with Q being symmetric positive semidefinite:

$$\begin{array}{rll} \psi(x,\xi) \triangleq & \underset{y}{\text{minimum}} & \left[\, f(\xi) + G(\xi)x \,\right]^\top y + \frac{1}{2}\, y^\top Q y \\ & \text{subject to} & y \in Y(x,\xi) \triangleq \left\{ y \in \mathbb{R}^{n_2} \mid A(\xi)x + Dy \geq b(\xi) \right\}. \end{array}$$

- The recourse function $\psi(\bullet,\xi)$ is of the implicit convex-concave kind.

Comments

• Leads to the study of the class of two-stage linearly bi-parameterized stochastic program with quadratic recourse:

$$\underset{x}{\textbf{minimize}} \quad \zeta(x) \triangleq \varphi(x) + \mathbf{E}_{\tilde{\xi}}\left[\psi(x,\tilde{\xi})\right] \quad \textbf{subject to} \quad x \in X \subseteq \mathbb{R}^{n_1},$$

where the recourse function $\psi(x,\xi)$ is the optimal objective value of the quadratic program (QP): with Q being symmetric positive semidefinite:

$$\begin{array}{rll} \psi(x,\xi) \triangleq & \underset{y}{\text{minimum}} & \left[\, f(\xi) + G(\xi)x \,\right]^\top y + \frac{1}{2}\, y^\top Q y \\ & \text{subject to} & y \in Y(x,\xi) \triangleq \left\{ y \in \mathbb{R}^{n_2} \mid A(\xi)x + Dy \geq b(\xi) \right\}. \end{array}$$

• The recourse function $\psi(\bullet,\xi)$ is of the implicit convex-concave kind.

• Developed a Regularization, Convexification, sampling combined method for computing a *generalized critical point* with almost sure convergence guarantee.

(Simplified) Defender-Attacker Game with Costs

Underlying problem is: $\boxed{\textbf{minimize } c^\top x \text{ subject to } Ax = b \text{ and } x \geq 0}$

Defender's problem: Anticipating the attacker's disruption $\delta \in \Delta$, the defender undertakes the operation by solving the problem:

$$\underset{x \geq 0}{\textbf{minimize}} \quad (c + \delta c)^\top x + \underbrace{\rho \left\| \left[(A + \delta A)x - b - \delta b \right]_+ \right\|_\infty}_{\text{constraint residual}}$$

Attacker's problem: Anticipating the defender's activity x, the attacker aims to disrupt the defender's operations by solving the problem:

$$\underset{\delta \in \Delta}{\textbf{maximum}} \quad \underbrace{\lambda_1 (c + \delta c)^\top x}_{\text{disrupt objective}} + \underbrace{\lambda_2 \left\| \left[(A + \delta A)x - b - \delta b \right]_+ \right\|_\infty}_{\text{disrupt constraints}} - \underbrace{C(\delta)}_{\text{cost}}.$$

- Many variations are possible.
- Offer alternative approaches to the convexity-based robust formulations.
- **Adversarial Programming**

Robustification of nonconvex problems

A general nonlinear program:

$$\begin{array}{ll} \underset{x \in S}{\textbf{minimize}} & \theta(x) + f(x) \\ \textbf{subject to} & H(x) = 0, \quad \text{where} \end{array}$$

• S is a closed convex set in $\mathbb{R}^n$ not subject to alterations,
• θ and f are real-valued functions defined on an open convex set $\mathcal{O}$ in $\mathbb{R}^n$ containing S,
• $H : \mathcal{O} \to \mathbb{R}^m$ is a vector-valued (nonsmooth) function that can be used to model both inequality constraints $g(x) \leq 0$ and equality constraints $h(x) = 0$.

Robustification of nonconvex problems

A general nonlinear program:

$$\begin{array}{ll} \underset{x \in S}{\textbf{minimize}} & \theta(x) + f(x) \\ \textbf{subject to} & H(x) = 0, \quad \text{where} \end{array}$$

• S is a closed convex set in $\mathbb{R}^n$ not subject to alterations,
• θ and f are real-valued functions defined on an open convex set $\mathcal{O}$ in $\mathbb{R}^n$ containing S,
• $H : \mathcal{O} \to \mathbb{R}^m$ is a vector-valued (nonsmooth) function that can be used to model both inequality constraints $g(x) \leq 0$ and equality constraints $h(x) = 0$.

Robust counterpart: with separate perturbations

$$\left.\begin{array}{ll} \underset{x \in S}{\textbf{minimize}} & \theta(x) + \underset{\widehat{\delta} \in \widehat{\Delta}}{\textbf{maximum}}\, f(x, \widehat{\delta}) \\ \textbf{subject to} & \underbrace{H(x, \widetilde{\delta}) = 0, \quad \forall\, \widetilde{\delta} \in \widetilde{\Delta}}_{\text{easily infeasible}} \end{array}\right\} \quad \begin{array}{l} \text{in current way} \\ \text{of modeling} \end{array}$$

New formulation

via a pessimistic value-function minimization: given $\gamma > 0$,

$$\boxed{\underset{x \in S}{\textbf{minimize}} \quad \theta(x) + v_\gamma(x)}$$

where $v_\gamma(x)$ models constraint violation with cost of model attacks:

$$v_\gamma(x) \quad \triangleq \quad \textbf{maximum} \text{ over } \boldsymbol{\delta} \in \boldsymbol{\Delta} \text{ of}$$

$$\phi_\gamma(x, \boldsymbol{\delta}) \quad \triangleq \quad f(x, \boldsymbol{\delta}) + \gamma \left[\underbrace{\| H(x, \boldsymbol{\delta}) \|_\infty}_{\text{constraint residual}} - \underbrace{C(\boldsymbol{\delta})}_{\text{attack cost}} \right]$$

with the novel features:

- combined perturbation $\boldsymbol{\delta}$ in joint uncertainty set $\boldsymbol{\Delta} \subseteq \mathbb{R}^L$;
- robustify constraint satisfaction via penalization

Comments

- Related the VFOP to the two game formulations in terms of directional solutions.

Comments

- Related the VFOP to the two game formulations in terms of directional solutions.

- Showed that the value function admits an explicit form under piecewise structures of $f(\bullet, \boldsymbol{\delta})$ and $H(\bullet, \boldsymbol{\delta})$, a linear structure of $C(\boldsymbol{\delta})$ with a set-up component, and special simplex-type uncertainty set $\boldsymbol{\Delta}$.

Comments

• Related the VFOP to the two game formulations in terms of directional solutions.

• Showed that the value function admits an explicit form under piecewise structures of $f(\bullet, \boldsymbol{\delta})$ and $H(\bullet, \boldsymbol{\delta})$, a linear structure of $C(\boldsymbol{\delta})$ with a set-up component, and special simplex-type uncertainty set $\boldsymbol{\Delta}$.

• Introduced a Majorization-Minimization (MM) algorithm for solving a unified formulation:

$$\underset{x \in S}{\textbf{minimize}}\ \varphi(x) \triangleq \theta(x) + v(x) \quad \text{with}$$

$$v(x) \triangleq \max_{1 \leq j \leq J} \underset{\boldsymbol{\delta} \in \boldsymbol{\Delta} \cap \mathbb{R}^L_+}{\textbf{maximum}} \left[g_j(x, \boldsymbol{\delta}) - h^j(x)^\top \boldsymbol{\delta} \right],$$

Comments

• Related the VFOP to the two game formulations in terms of directional solutions.

• Showed that the value function admits an explicit form under piecewise structures of $f(\bullet, \boldsymbol{\delta})$ and $H(\bullet, \boldsymbol{\delta})$, a linear structure of $C(\boldsymbol{\delta})$ with a set-up component, and special simplex-type uncertainty set $\boldsymbol{\Delta}$.

• Introduced a Majorization-Minimization (MM) algorithm for solving a unified formulation:

$$\underset{x \in S}{\textbf{minimize}}\ \varphi(x) \triangleq \theta(x) + v(x) \quad \text{with}$$

$$v(x) \triangleq \max_{1 \le j \le J}\ \underset{\boldsymbol{\delta} \in \boldsymbol{\Delta} \cap \mathbb{R}^L_+}{\textbf{maximum}} \left[g_j(x, \boldsymbol{\delta}) - h^j(x)^\top \boldsymbol{\delta} \right],$$

• Implemented algorithm and reported computation results to demonstrate the viability of the robust paradigm in the "non"-framework.

Examples of Non-Problems

III. Operations Research meets Statistics

Composite CVaR and bPOE minimization

for risk-based statistical estimation to control left and right-tail distributions simultaneously

Interval Conditional Value-at-Risk of a random variable Z, at levels $0 \leq \alpha < \beta \leq 1$, treating two-sided estimation errors,

$$\begin{aligned}\text{In-CVaR}_\alpha^\beta(Z) &\triangleq \frac{1}{\beta - \alpha} \int_{\mathrm{VaR}_\beta(Z) > z \geq \mathrm{VaR}_\alpha(Z)} z \, dF_Z(z) \\ &= \frac{1}{\beta - \alpha} \left[(1-\alpha)\mathrm{CVaR}_\alpha(Z) - (1-\beta)\mathrm{CVaR}_\beta(Z) \right]\end{aligned}$$

Composite CVaR and bPOE minimization

for risk-based statistical estimation to control left and right-tail distributions simultaneously

Interval Conditional Value-at-Risk of a random variable Z, at levels $0 \leq \alpha < \beta \leq 1$, treating two-sided estimation errors,

$$\begin{aligned} \text{In-CVaR}_{\alpha}^{\beta}(Z) \;&\triangleq\; \frac{1}{\beta - \alpha} \int_{\mathrm{VaR}_{\beta}(Z) > z \geq \mathrm{VaR}_{\alpha}(Z)} z \, dF_Z(z) \\ &= \frac{1}{\beta - \alpha} \left[(1-\alpha)\mathrm{CVaR}_{\alpha}(Z) - (1-\beta)\mathrm{CVaR}_{\beta}(Z) \right] \end{aligned}$$

Buffered probability of exceedance/deceedance at level $\tau > 0$

$$\text{bPOE}(Z;\tau) \triangleq 1 - \min \left\{ \alpha \in [0,1] \,:\, \text{CVaR}_{\alpha}(Z) \geq \tau \right\}$$

$$\text{bPOD}(Z;u) \triangleq \min \left\{ \beta \in [0,1] \,:\, \text{In-CVaR}_{0}^{\beta}(Z) \geq u \right\}.$$

Buffered probability of interval $(\tau, u]$

We have the following bound for an interval probability:

$$\mathbb{P}(u \geq Z > \tau) \, \leq \, \mathsf{bPOE}(Z;\tau) + \mathsf{bPOD}(Z;u) - 1 \triangleq \mathsf{bPOI}_{\tau}^{u}(Z).$$

Two risk minimization problems of a composite random functional:

$$\underbrace{\underset{\theta\in\Theta}{\textbf{minimize}} \;\; \text{In-CVaR}_{\alpha}^{\beta}(\mathcal{Z}(\theta))}_{\substack{\text{a difference of two convex} \\ \text{SP value functions}}} \quad \text{and} \quad \underbrace{\underset{\theta\in\Theta}{\textbf{minimize}} \;\; \mathsf{bPOI}_{\tau}^{u}(\mathcal{Z}(\theta))}_{\substack{\text{a stochastic program with} \\ \text{a difference-of-convex objective}}},$$

where

$$\mathcal{Z}(\theta) \triangleq c \circ \left[\underbrace{g(\theta;Z) - h(\theta;Z)}_{g(\bullet;z) - h(\bullet;z) \text{ is difference of convex}} \right],$$

with $c : \mathbb{R} \to \mathbb{R}$ being a univariate piecewise affine function, Z is a m-dimensional random vector, and for each z, $g(\bullet;z)$ and $h(\bullet;z)$ are both convex functions.

Comments

- The In-CVaR minimization leads to the following stochastic program:

$$\textbf{minimize over } \theta \in \Theta \text{ of}$$

$$\underbrace{\underset{\eta_1 \in \Upsilon_1}{\textbf{minimum}} \, \mathbf{E}_{\widetilde{\omega}} \left[\, \varphi_1(\theta, \eta_1; \widetilde{\omega}) \, \right]}_{\text{denoted } v_1(\theta)} - \underbrace{\underset{\eta_2 \in \Upsilon_2}{\textbf{minimum}} \, \mathbf{E}_{\widetilde{\omega}} \left[\, \varphi_2(\theta, \eta_2; \widetilde{\omega}) \, \right]}_{\text{denoted } v_2(\theta)}$$

Both functions v_1 and v_2 are (convex) value functions of a univariate expectation minimization; thus overall objective is a difference-of-convex objective.

Comments

- The In-CVaR minimization leads to the following stochastic program:

$$\textbf{minimize} \text{ over } \theta \in \Theta \text{ of}$$

$$\underbrace{\underset{\eta_1 \in \Upsilon_1}{\textbf{minimum}} \, \mathbf{E}_{\widetilde{\omega}} \left[\varphi_1(\theta, \eta_1; \widetilde{\omega}) \right]}_{\text{denoted } v_1(\theta)} - \underbrace{\underset{\eta_2 \in \Upsilon_2}{\textbf{minimum}} \, \mathbf{E}_{\widetilde{\omega}} \left[\varphi_2(\theta, \eta_2; \widetilde{\omega}) \right]}_{\text{denoted } v_2(\theta)}$$

Both functions v_1 and v_2 are (convex) value functions of a univariate expectation minimization; thus overall objective is a difference-of-convex objective.

- Developed a convex programming based sampling method for computing a critical point of the objective

Comments

- The In-CVaR minimization leads to the following stochastic program:

$$\boxed{\begin{array}{l} \textbf{minimize} \text{ over } \theta \in \Theta \text{ of} \\[1em] \underbrace{\underset{\eta_1 \in \Upsilon_1}{\textbf{minimum}} \, \mathbf{E}_{\widetilde{\omega}} \left[\varphi_1(\theta, \eta_1; \widetilde{\omega}) \right]}_{\text{denoted } v_1(\theta)} - \underbrace{\underset{\eta_2 \in \Upsilon_2}{\textbf{minimum}} \, \mathbf{E}_{\widetilde{\omega}} \left[\varphi_2(\theta, \eta_2; \widetilde{\omega}) \right]}_{\text{denoted } v_2(\theta)} \end{array}}$$

Both functions v_1 and v_2 are (convex) value functions of a univariate expectation minimization; thus overall objective is a difference-of-convex objective.

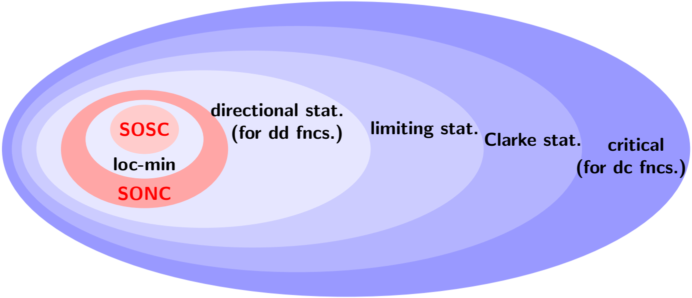
- Developed a convex programming based sampling method for computing a critical point of the objective

- Numerical results demonstrated superior performance in the handling of outliers in regression and classification problems.

A touch of mathematics

bridging computations with theory

First of All, What Can be Computed?



Relationship between the stationary points.

- Stationarity:
  - <u>smooth case:</u> $\nabla V(x) = 0$
  - <u>nonsmooth case:</u> $0 \in \partial V(x)$ [subdifferential: many calculus rules fail]

  e.g. $\partial V_1(x) + \partial V_2(x) \neq \partial(V_1 + V_2)(x)$
  unless one of them is continuously differentiable
- directional stationarity: $V'(x; d) \geq 0$ for all the directions d.

• The sharper the concept, the harder to compute.

Computational Tools

- **Built on a theory of surrogation, supplemented by exact penalization of stationary solutions**
- **A fusion of first-order and second-order algorithms**
  - Dealing with non-convexity: majorization-minimization algorithm
  - Dealing with non-smoothness: nonsmooth Newton algorithm
  - Dealing with pathological constraints: (exact) penalty method
- **An integration of sampling and deterministic methods**
  - Dealing with population optimization: in statistics
  - Dealing with stochastic programming: in operations research
  - Dealing with risk minimization: in financial management and tail control

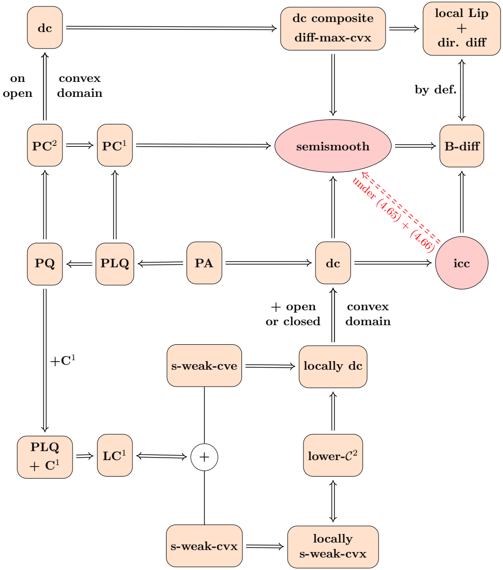
Classes of Nonsmooth Functions

Pervasive in applications

Progressively broader:

- Piecewise smooth*: e.g. piecewise linear-quadratic
- Difference-of convex: facilitates algorithmic design via convex majorization
- Semismooth: enables fast algorithms of the Newton-type
- Bouligand differentiable: locally Lipschitz + directionally differentiable
- Sub-analytic and semi-analytic

$*$ and many more.


dc
dc composite
diff-max-cvx
local Lip
+
dir. diff
on
open
convex
domain
by def.
PC^2
PC^1
semismooth
B-diff
under (4.65) + (4.66)
PQ
PLQ
PA
dc
icc
+ open
or closed
convex
domain
+C^1
s-weak-cve
locally dc
PLQ
+ C^1
LC^1
+
lower-$\mathcal{C}^2$
s-weak-cvx
locally
s-weak-cvx

Completed Monograph:

Modern Nonconvex Nondifferentiable Optimization

Ying Cui and Jong-Shi Pang

submitted to publisher (Friday September 11, 2020)

Table of Contents

Preface **i**

Acronyms **i**

Glossary of Notation **i**

Index **i**

1 Primer of Mathematical Prerequisites **1**

1.1 Smooth Calculus 1

1.1.1 Differentiability of functions of several variables 7

1.1.2 Mean-value theorems for vector-valued functions 12

1.1.3 Fixed-point theorems and contraction 16

1.1.4 Implicit-function theorems 18

1.2 Matrices and Linear Equations 23

1.2.1 Matrix factorizations, eigenvalues, and singular values 29

1.2.2 Selected matrix classes 36

1.3 Matrix Splitting Methods for Linear Equations 47

1.4 The Conjugate Gradient Method. 50

1.5 Elements of Set-Valued Analysis 54

2 Optimization Background **59**

2.1 Polyhedral Theory . . . 59

2.1.1 The Fourier-Motzkin elimination in linear inequalities . . . 60

2.1.2 Properties of polyhedra . . . 63

2.2 Convex Sets and Functions . . . 67

2.2.1 Euclidean projection and proximality . . . 71

2.3 Convex Quadratic and Nonlinear Programs . . . 79

2.3.1 The KKT conditions and CQs . . . 81

2.4 The Minmax Theory of von Neumann . . . 85

2.5 Basic Descent Methods for Smooth Problems . . . 86

2.5.1 Unconstrained problems: Armijo line search . . . 87

2.5.2 The PGM for convex programs . . . 91

2.5.3 The proximal Newton method . . . 95

2.6 Nonlinear Programming Methods: A Survey . . . 103

3 Structured Learning via Statistics and Optimization **109**

3.1 A Composite View of Statistical Estimation . . . 111

3.1.1 The optimization problems . . . 113

3.1.2 Learning classes . . . 115

3.1.3 Loss functions . . . 122

3.1.4 Regularizers for sparsity . . . 123

3.2 Low Rankness in Matrix Optimization . . . 128

3.2.1 Rank constraints: Hard and soft formulations . . . 128

3.2.2 Shape constrained index models . . . 131

3.3 Affine Sparsity Constraints . . . 137

4 Nonsmooth Analysis **143**

4.1 B(ouligand)-Differentiable Functions . . . 144

4.1.1 Generalized directional derivatives . . . 153

4.2 Second-order Directional Derivatives 156
4.3 Subdifferentials and Regularity 162
4.4 Classes of Nonsmooth Functions 176
4.4.1 Piecewise affine functions 177
4.4.2 Piecewise linear-quadratic functions 190
4.4.3 Weakly convex functions 205
4.4.4 Difference-of-convex functions. 214
4.4.5 Semismooth functions 221
4.4.6 Implicit convex-concave functions 229
4.4.7 A summary 240

5 Value Functions **243**
5.1 Nonlinear Programming based Value Functions 244
5.1.1 Interdiction versus enhancement. 247
5.2 Value Functions of Quadratic Programs 251
5.3 Value Functions Associated with Convex Functions 258
5.4 Robustification of Optimization Problems 261
5.4.1 Attacks without set-up costs 264
5.4.2 Attacks with set-up costs 267
5.5 The Danskin Property 271
5.6 Gap Functions 278
5.7 Some Nonconvex Stochastic Value Functions 282
5.8 Understanding Robust Optimization 289
5.8.1 Uncertainty in data realizations 290
5.8.2 Uncertainty in the probabilities of realizations 304

6 Stationarity Theory **311**
6.1 First-order Theory 313
6.1.1 The convex constrained case 314
6.1.2 Composite ε-strong stationarity 320

6.1.3 Subdifferential based stationarity . . . 324
6.2 Directional Saddle-Point Theory . . . 332
6.3 Difference-of-Convex Constraints . . . 337
6.3.1 Extension to piecewise dd-convex programs . . . 345
6.4 Second-order Theory . . . 348
6.5 Piecewise Linear-Quadratic Programs . . . 350
6.5.1 Review of nonconvex QPs . . . 350
6.5.2 Return to PLQ programs . . . 352
6.5.3 Finiteness of stationary values . . . 362
6.5.4 Testing copositivity: One negative eigenvalue . . . 367

7 Computational Algorithms by Surrogation 371
7.1 Basic Surrogation . . . 372
7.1.1 A broad family of upper surrogation functions . . . 379
7.1.2 Examples of surrogation families . . . 381
7.1.3 Surrogation algorithms for several problem classes . . . 384
7.2 Refinements and Extensions . . . 392
7.2.1 Moreau surrogation . . . 392
7.2.2 Combined BCD and ADMM for coupled constraints . . . 400
7.2.3 Surrogation with line searches . . . 418
7.2.4 Dinkelbach's algorithm with surrogation . . . 424
7.2.5 Randomized choice of subproblems . . . 432
7.2.6 Inexact surrogation . . . 435

8 Theory of Error Bounds 441
8.1 An Introduction to Error Bounds . . . 442
8.1.1 The stationarity inclusions . . . 444
8.2 The Piecewise Polyhedral Inclusion . . . 449
8.3 Error Bounds for Subdifferential-Based Stationarity . . . 454

8.4 Sequential Convergence Under Sufficient Descent 462
8.4.1 Using the local error bound (8.19) 462
8.4.2 The Kurdyka-Łojaziewicz theory 467
8.4.3 Connections between Theorems 8.4.1 and 8.4.4 475
8.5 Error Bounds and Linear Regularity 476
8.5.1 Piecewise systems 483
8.5.2 Stationarity for convex composite dc optimization 488
8.5.3 dd-convex systems 491
8.5.4 Linear regularity and directional Slater 497

9 Theory of Exact Penalization **501**
9.1 Exact Penalization of Optimal Solutions. 503
9.2 Exact Penalization of Stationary Solutions 506
9.3 Pulled-out Penalization of Composite Functions. 511
9.4 Two Applications of Pull-out Penalization 517
9.4.1 B-stationarity of dc composite diff-max programs 518
9.4.2 Local optimality of composite twice semidiff. programs. 521
9.5 Exact Penalization of Affine Sparsity Constraints 527
9.6 A Special Class of Cardinality Constrained Problems. 531
9.7 Penalized Surrogation for Composite Programs 533
9.7.1 Solving the convex penalized subproblems 539
9.7.2 Semismooth Newton methods. 543

10 Nonconvex Stochastic Programs **547**
10.1 Expectation Functions 548
10.1.1 Sample average approximations 555
10.2 Basic Stochastic Surrogation 560
10.2.1 Expectation constraints. 567
10.3 Two-stage SPs with Quadratic Recourse 568
10.3.1 The positive definite case 569

10.3.2 The positive semidefinite case . . . 571
10.4 Probabilisitic Error Bounds . . . 588
10.5 Consistency of Directional Stationary Solutions . . . 591
10.5.1 Convergence of stationary solutions . . . 592
10.5.2 Asymptotic distribution of the stationary values . . . 605
10.5.3 Convergence rate of the stationary points . . . 609
10.6 Noisy Amplitude-based Phase Retrieval . . . 613

11 Nonconvex Game Problems 625
11.1 Definitions and Preliminaries . . . 627
11.1.1 Applications . . . 631
11.1.2 Related models . . . 637
11.2 Existence of a QNE . . . 640
11.3 Computational Algorithms. . . . 648
11.3.1 A Gauss-Seidel best-response algorithm . . . 649
11.4 The Role of Potential . . . 658
11.4.1 Existence of a potential. . . . 665
11.5 Double-Loop Algorithms . . . 669
11.6 The Nikaido-Isoda Optimization Formulations . . . 678
11.6.1 Surrogation applied to the NI function ϕ_c . . . 681
11.6.2 Difference-of-convexity and surrogation of $\widehat{\phi}_c$ . . . 683

Bibliography 691

Index 721