不確実性を考慮した最適化手法

統計数理研究所 / 理化学研究所AIP センター
武田朗子

講義の構成

不確実な最適化問題に対する定式化と解法

- 第１部：　ロバスト最適化　(10:30 – 11:20)
- 第２部：　確率計画法　(11:40 – 12:30)

- 第３部：　ロバスト最適化や確率計画法の機械学習問題への適用　(14:00 – 15:00)
- 第４部：　演習
- 第５部：　総括

Mathematical Optimization

It helps to select a best element (with regard to some criteria) from some set of available alternatives.

Mathematical Optimization Problem:

$$\begin{array}{ll} \min\limits_{\boldsymbol{x}} & f(\boldsymbol{x}) \\ \text{s.t.} & g_i(\boldsymbol{x}) \geq 0, \qquad i = 1, \ldots, m \end{array}$$

- $f(\boldsymbol{x}), g_1(\boldsymbol{x}), \ldots, g_m(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$
- If $f(\boldsymbol{x}), g_1(\boldsymbol{x}), \ldots, g_m(\boldsymbol{x})$ are linear in $\boldsymbol{x}$, the problem is called a linear programming problem.

math.

Solving a system of polynomial equations

$$\begin{aligned} xy &= 1 \\ 2i\,xy^2 + y^2 + x &= 1 \end{aligned}$$

(Linear Programming)

power engineering

Scheduling of generators

Optimal size of solar panel

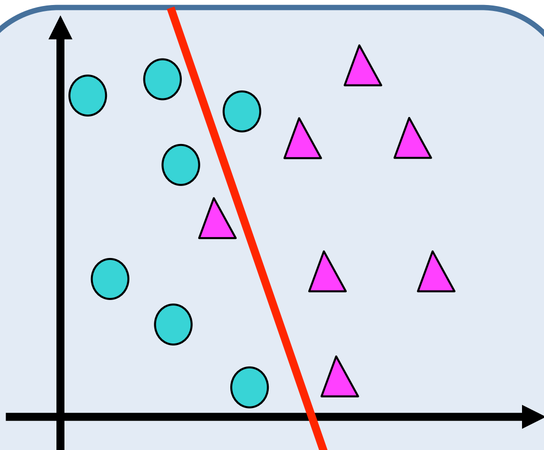
(Robust Optimization)

Optimization Problem

$$\begin{aligned} \textbf{Min:} &\quad f(x) \\ \textbf{subj.to}: &\quad g_1(x) \geq 0 \\ &\quad g_2(x) \geq 0 \\ &\quad \ldots\ldots \end{aligned}$$

machine learning

(Global Optimization)



Support Vector Machine

finance

portfolio allocation

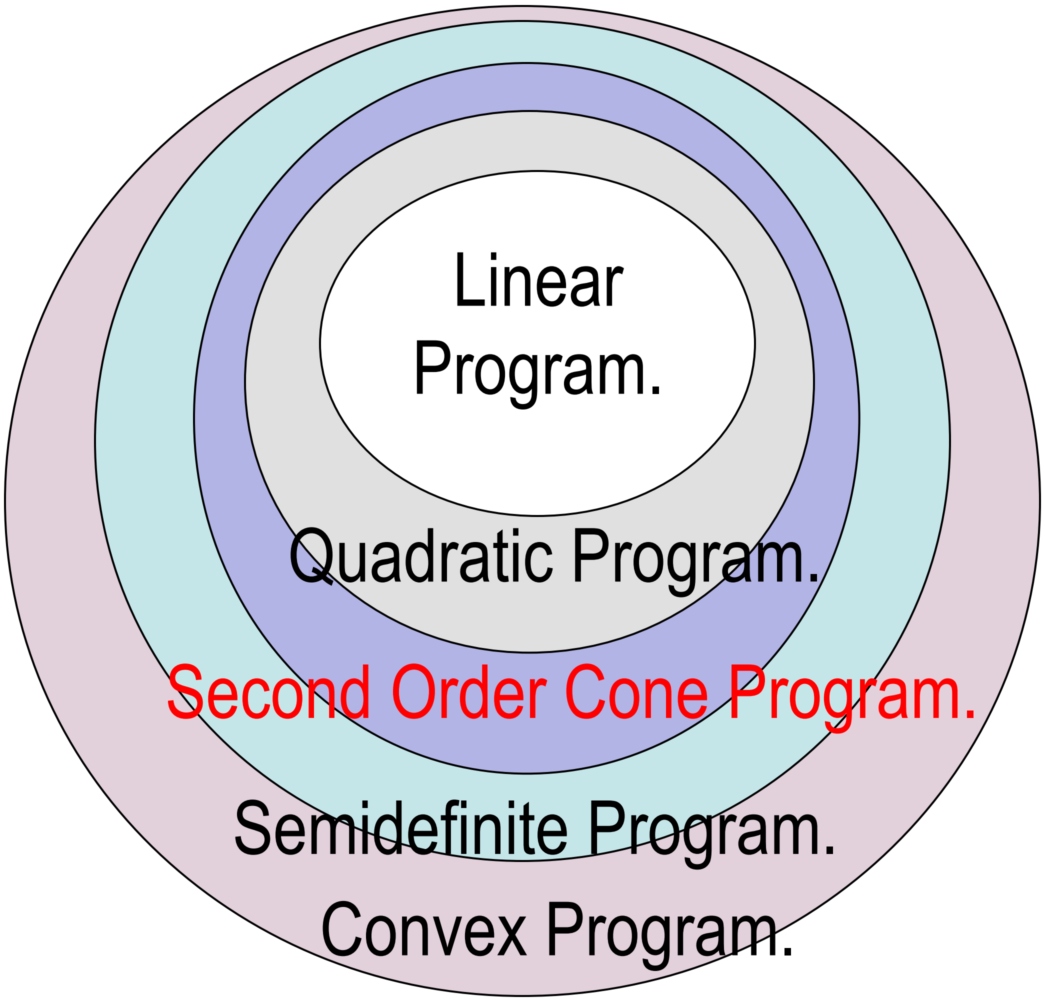
mathematical optimization

Solution Method

- Nonconvex Opt.
- Robust Optimization

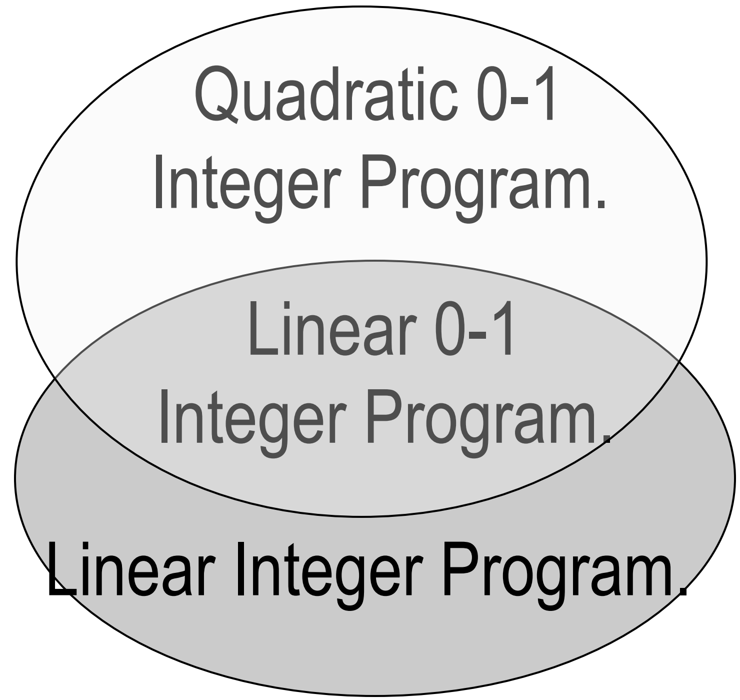
Various Optimization Problems

Continuous Optimization



Nonconvex Quadratic Program.

Discrete Optimization



Problem name based on Application:

Shortest Path Prob., Travelling Salesman Prob. Knapsack Prob.

Second-order cone programming

$$\min_{\boldsymbol{x}} \boldsymbol{f}^\top \boldsymbol{x}$$

$$\text{s.t.} \ \|\boldsymbol{A}_i \boldsymbol{x} + \boldsymbol{b}_i\| \leq \boldsymbol{c}_i^\top \boldsymbol{x} + d_i, \ \ i = 1, \ldots, m$$

Euclidean norm

$$\|\boldsymbol{u}\| = (\boldsymbol{u}^\top \boldsymbol{u})^{1/2}$$

- SOCP can be reformulated as an instance of SDP.
- Convex quadratic programs can also be formulated as SOCPs.
- SOCPs can be solved with great efficiency by interior point methods.

Optimization Method under Uncertainty

- Robust Optimization
  - ✓modeling strategies and solution methods for optimization problems that are defined by uncertain inputs
  - ✓proposed by Ben-Tal & Nemirovski in 1998

- Stochastic Programming
  - ✓classical framework for modeling optimization problems involving uncertainty (studied since the 1950's).
  - ✓assuming that probability distributions are known
  - ✓relation to robust optimization

Example : Power Generation Planning

T. Electric Company has 2 turbines (Fuel : oil, natural gas).
It wants to determine their production outputs to
minimize production costs and satisfy electric demands.

Decision Variable :
x_i : Production Output [MWh]

Unit Cost（Yen/MWh）

$$\min\ \underbrace{135}_{}x_1 + \underbrace{141}_{}x_2$$

$$\text{s.t.}\quad x_1 + x_2 \geq 1000$$

Demand

$$L_o \leq x_1 \leq U_o$$

$$L_g \leq x_2 \leq U_g$$

Linear Programming: LP
（Simplex Method,
Interior Point Method）

Formulation of Robust Optimization

Assump.: uncertain inputs vary within a set (*uncertainty set*).

The best decision is done under the *worst-case* scenario.

uncertainty sets: $\boldsymbol{u}_0 \in \mathcal{U}_0, \ \boldsymbol{u}_i \in \mathcal{U}_i, \forall i$

$$\min_{\boldsymbol{x} \in X} f(\boldsymbol{x}, \boldsymbol{u}_0) \quad \Rightarrow \quad \min_{\boldsymbol{x} \in X} \max_{\boldsymbol{u}_0 \in \mathcal{U}_0} f(\boldsymbol{x}, \boldsymbol{u}_0)$$

$$\text{s.t.} \ \ g_i(\boldsymbol{x}, \boldsymbol{u}_i) \leq 0, \quad \Rightarrow \quad g_i(\boldsymbol{x}, \boldsymbol{u}_i) \leq 0, \ \forall \boldsymbol{u}_i \in \mathcal{U}_i$$

$$i = 1, \ldots, m$$

$$\Leftrightarrow \quad \max_{\boldsymbol{u}_i \in \mathcal{U}_i} g_i(\boldsymbol{x}, \boldsymbol{u}_i) \leq 0$$

Necessity of Robust Solution

Ben-Tal & Nemirovski ['00]

PILOT4 (NETLIB library)

1000 var., 410 const., $\boldsymbol{x}^*$: optimal solution

$$\boldsymbol{a}^\top \boldsymbol{x} \equiv$$
$$-15.79081 x_{826} - 8.598819 x_{827} - 1.88789 x_{828} - 1.362417 \cdots$$
$$-0.031883 x_{849} - 28.725555 x_{850} - 10.792065 x_{851} - 0.190 \cdots$$
$$-12.290832 x_{854} + 717.562256 x_{855} - 0.057865 x_{856} - 3.785$$
$$-122.163055 x_{859} - 6.46609 x_{860} - 0.48371 x_{861} - 0.615264 \cdots$$
$$-84.644257 x_{864} - 122.459045 x_{865} - 43.15593 x_{866} - 1.712 \cdots$$
$$+x_{880} - 0.946049 x_{898} - 0 : 946049 x_{916} \geq \boxed{23.387405} \equiv b$$

Change the coeff. $\boldsymbol{a}$ by its 0.1% $\rightarrow$ $\overline{\boldsymbol{a}}$

e.g., $15.79081 \times 0.001 = 0.0157908$

$\boldsymbol{x}^*$ satisfying $\boldsymbol{a}^\top \boldsymbol{x}^* - b \geq 0$ largely violates the perturbed one:

$$\overline{\boldsymbol{a}}^\top \boldsymbol{x}^* - b < -104.9$$

Applications of Robust Optimization

The obtained solution
•is relatively insensitive to data variations, and
•hedges against catastrophic outcomes.

Ben-Tal & Nemirovski ['97]

Truss topology under the load uncertainties :
constructing a building assuming a typical wind load
→ neglecting the possibility of strong wind
→ causing the building to collapse

Lin, Janak & Floudas ['04]

Robust scheduling of chemical processing :
scheduling of multiproduct and multipurpose batch plants.
→ neglecting variability of process and environmental data.
→ causing fire and explosion

Applications to Radio Therapy

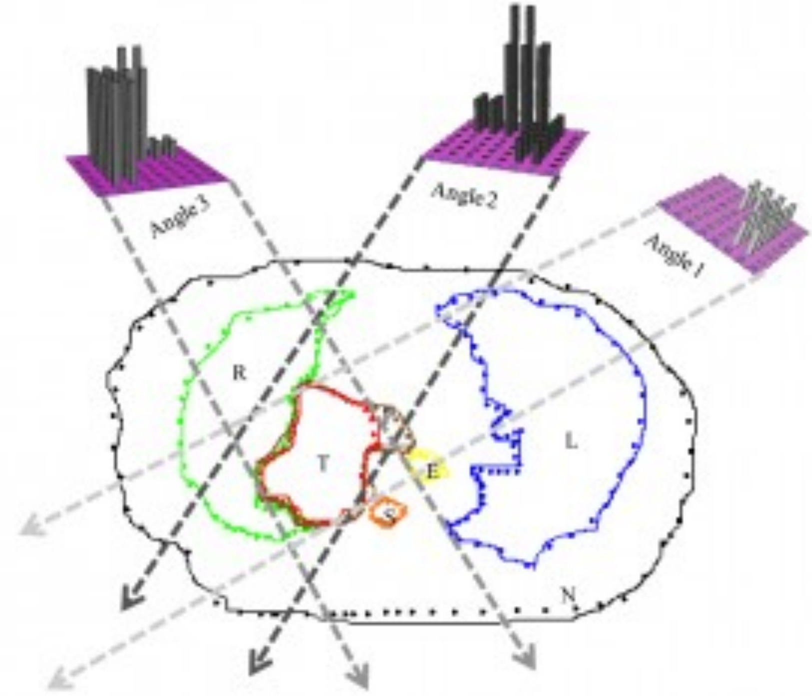
T. C. Y. Chan et al. ['06]

[Radiation Therapy for Cancer Patients]

Beams of radiation are delivered from different angles around a patient, targeting a tumor in their intersection while trying to spare nearby critical organs.

→ Optimization methods determine the angles of the beams and the intensities of the beamlets, etc.

→ Uncertainty in tumor position (e.g., lung tumors move as the patient breathes during treatment)



http://www.newswise.com/articles/improving-radiation-therapy-for-cancer-patients

Applications to Solar Energy System

[Solar Energy System]

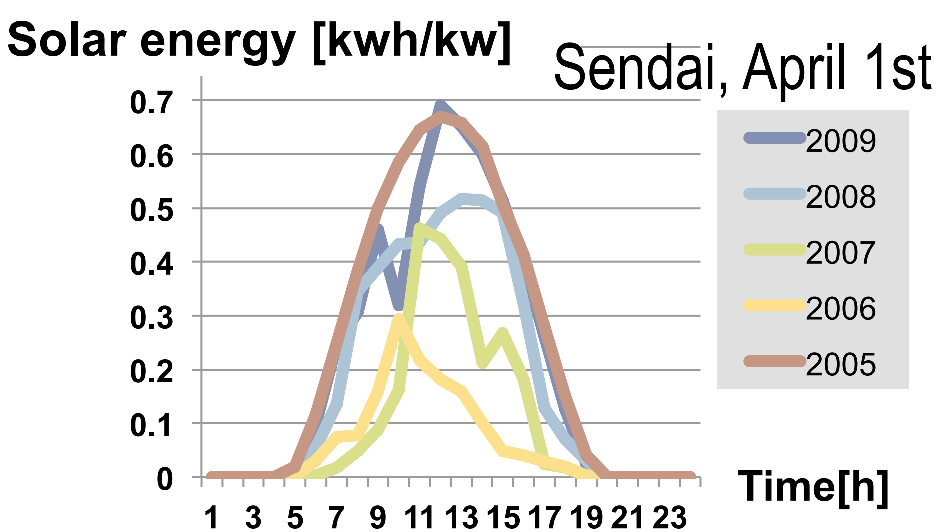
Okido & Takeda ['12]

Determining the optimal size of a residential grid-connected solar system to meet a certain CO2 reduction target at a minimum cost.

[project from Japanese local authority]

→ Useful to determine an amount of subsidy for system owners

→ Taking into consideration uncertainty in the level of solar irradiation (or solar energy) due to weather conditions



What is Robust Optimization?

When the data differs from the assumed nominal values, the generated optimal solution may violate critical constraints and perform poorly.

Want to find a solution immune to data uncertainty.

Robust optimization:
modeling strategies and solution methods for uncertain problems.

It optimizes against the *worst* instance that might arise due to uncertain inputs.

Other Method: Stochastic Programming

Uncertain Optimization Problem:

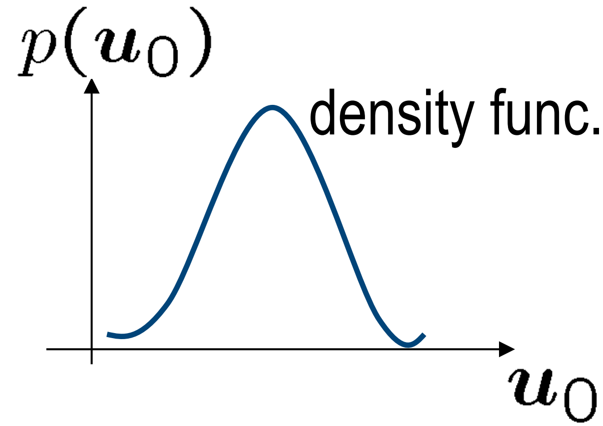
$$\min_{\boldsymbol{x} \in X} f(\boldsymbol{x}, \boldsymbol{u}_0) \text{ s.t. } g(\boldsymbol{x}, \boldsymbol{u}_1) \leq 0$$

$\boldsymbol{u}_0, \boldsymbol{u}_1$: uncertain data

- Stochastic Programming — Dantzig ['55], Beale ['55]

Assump. :

Prob. distributions of $\boldsymbol{u}_0, \boldsymbol{u}_1$ are known.



ex.1)
$$\min_{\boldsymbol{x} \in X} E_{\boldsymbol{u}_0}[f(\boldsymbol{x}, \boldsymbol{u}_0)]$$
$$\text{s.t. } E_{\boldsymbol{u}_1}[g(\boldsymbol{x}, \boldsymbol{u}_1)] \leq 0$$

ex.2) Chance Const.（Probabilistic Const. ） — Charnes & Cooper ['59]

$$\Pr_{\boldsymbol{u}_1}(g(\boldsymbol{x}, \boldsymbol{u}_1) \leq 0) \geq 1 - \epsilon$$

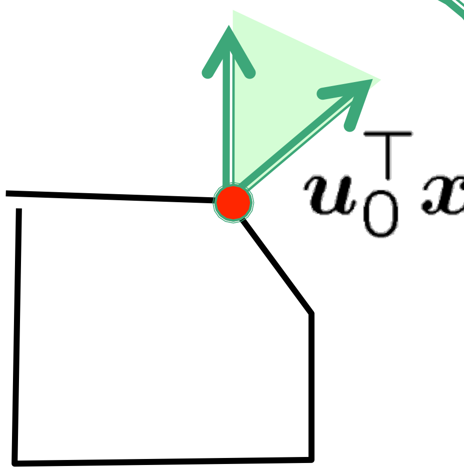
Other Method: Sensitivity Analysis

Uncertain Optimization Problem :

$$\min_{\boldsymbol{x}\in X} f(\boldsymbol{x},\boldsymbol{u}_0) \text{ s.t. } g(\boldsymbol{x},\boldsymbol{u}_1) \le 0$$

$\boldsymbol{u}_0, \boldsymbol{u}_1$: uncertain data

- Post-optimal analysis after obtaining an optimal solution for some $\boldsymbol{u}_0, \boldsymbol{u}_1$.
- It shows whether the optimal solution changes for the data perturbation.



Restrictions: data of objective func. & RHS of LP can be uncertain

History of Robust Optimization

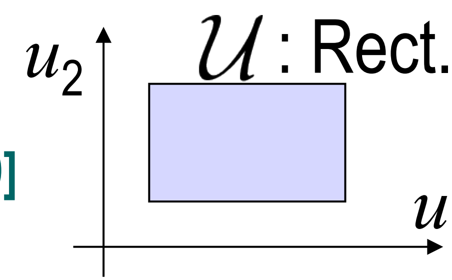
Robust Optimization:

$$\min_{\boldsymbol{x} \in X} \max_{\boldsymbol{u}_0 \in \mathcal{U}_0} f(\boldsymbol{x}, \boldsymbol{u}_0)$$
$$\text{s.t.} \max_{\boldsymbol{u}_i \in \mathcal{U}_i} g_i(\boldsymbol{x}, \boldsymbol{u}_i) \leq 0, \ \forall i$$

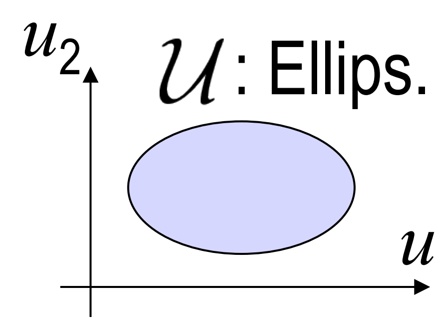
- In 1973, A.L.Soyster proposed “inexact LP” using rectangular $\mathcal{U}$.

Almost no progress (two papers†)

†) **reported by Ben-Tal, El Ghaoui & Nemirovski ['09]**



- In 1998, Ben-Tal & Nemirovski proposed “robust optimization” using ellipsoidal $\mathcal{U}$.



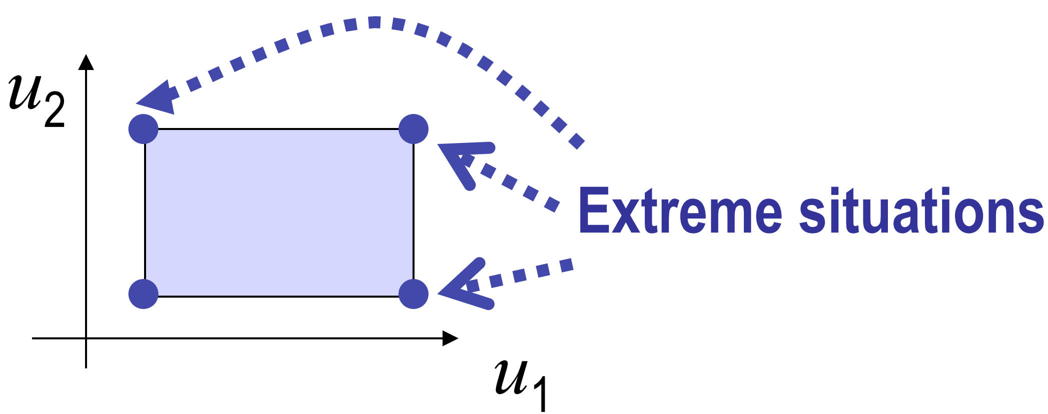
- Studies on robust optimization are going on …

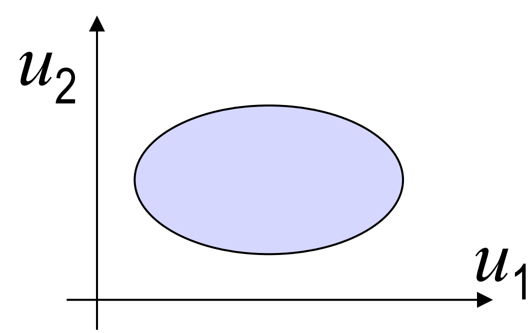
Why robust optimization became popular?

① Inexact LP (=Robust LP with rectangle $\mathcal{U}$) only assumes extreme situations. This drawback was solved by ellipsoidal $\mathcal{U}$.

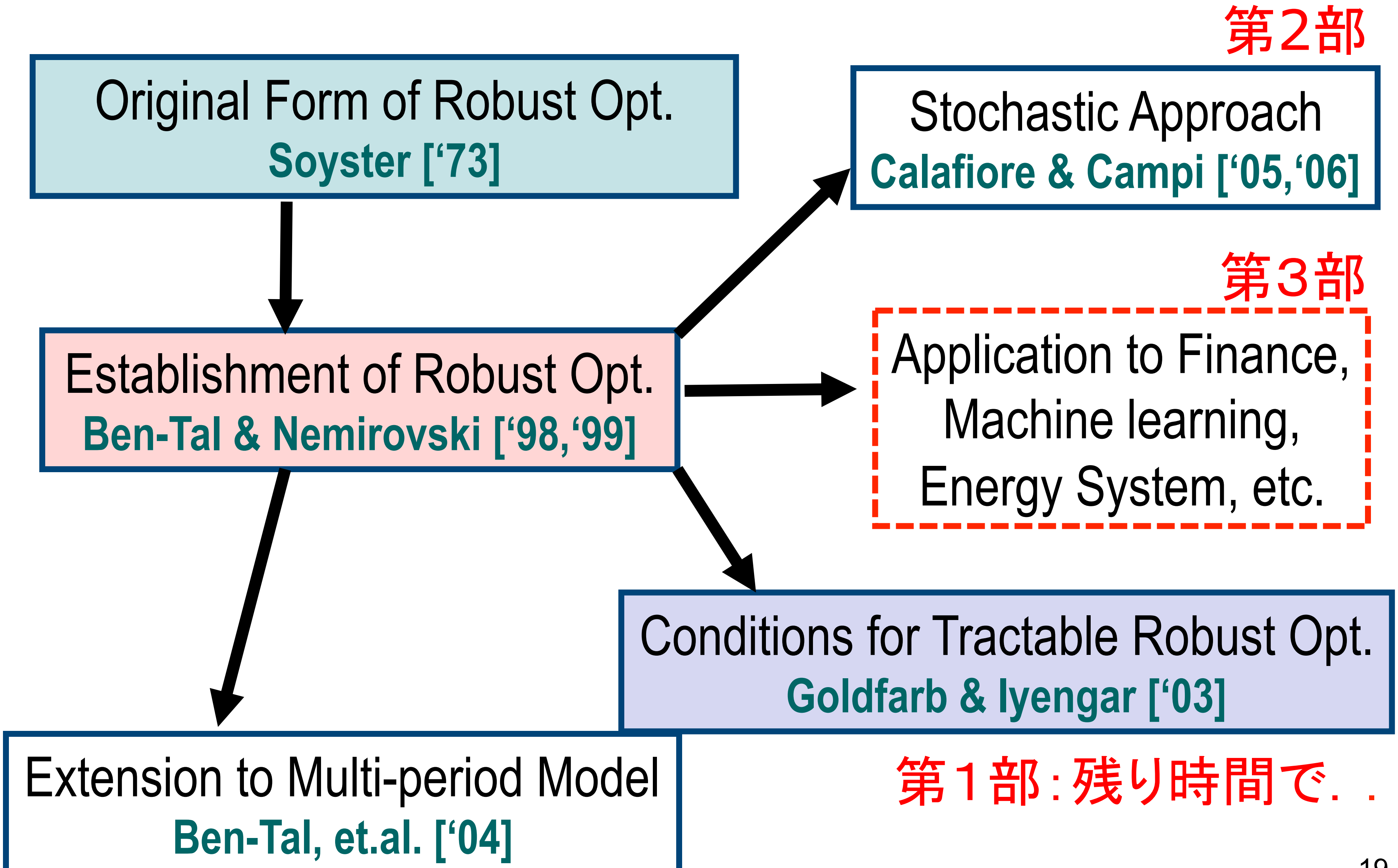
② Resulting in a second-order cone programming (SOCP), semidefinite programming (SDP).

Inexact LP **Soyster ['73]**		Robust LP **Ben-Tal & Nemirovski ['98]**
$\mathcal{U}$ is a rectangle	⟹	$\mathcal{U}$ is an ellipsoid, etc.…





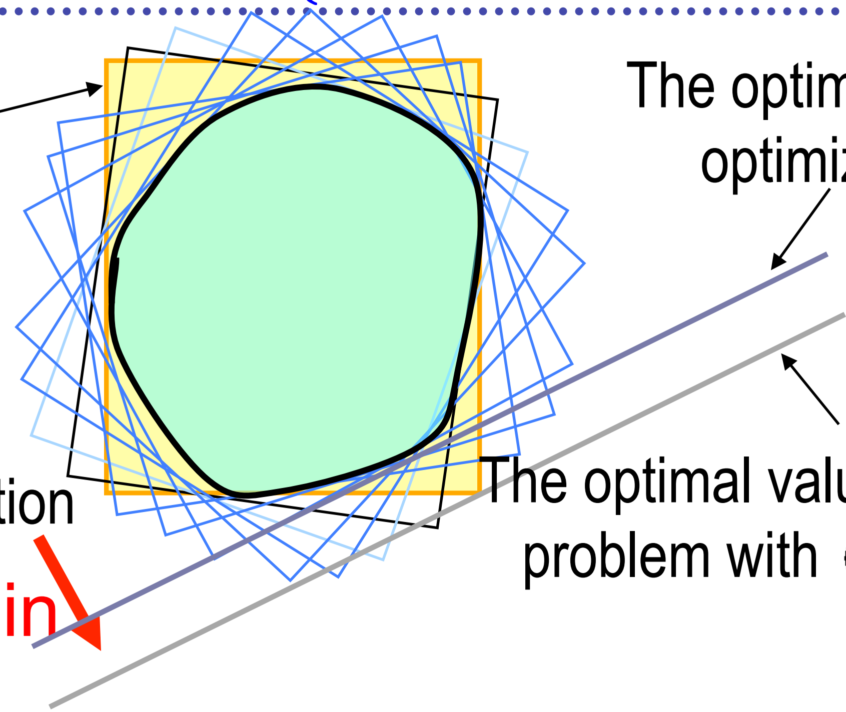
Various Research Directions



Difficult to Be Solved in General

$$\begin{aligned}
\min_{\boldsymbol{x}} \quad & -x_1 + x_2 \\
\text{s.t.} \quad & -1 \le u_1 x_1 + u_2 x_2 \le 1 \\
& -1 \le -u_2 x_1 + u_1 x_2 \le 1 \\
& {}^\forall \boldsymbol{u} \in \mathcal{U} = \left\{ (u_1, u_2) \mid u_1^2 + u_2^2 = 1 \right\}
\end{aligned}$$

infinite number of constraints



One research direction:

Want to define $\mathcal{U}$ so that the RO problem is tractable.

Standard Form for Robust Optimization

$$\min_{\boldsymbol{x} \in X} \ \boldsymbol{c}^\top \boldsymbol{x} \quad \text{s.t.} \quad f_i(\boldsymbol{x}, \boldsymbol{u}_i) \le 0, \quad \forall \boldsymbol{u}_i \in \mathcal{U}_i,$$
$$i = 1, \ldots, m$$

- Constraint-wise uncertainty is assumed.
- $f_i(\boldsymbol{x}, \boldsymbol{u}_i)$: convex in $\boldsymbol{x}$ $\quad (\forall \boldsymbol{u}_i \in \mathcal{U}_i)$
- X : closed convex set, $\quad \mathcal{U}_i$: bounded closed set

- **When the objective function is uncertain**

$$\min_{\boldsymbol{x} \in X} \max_{\boldsymbol{u}_0 \in \mathcal{U}_0} f_0(\boldsymbol{x}, \boldsymbol{u}_0)$$

$$\Longrightarrow \quad \min_{\boldsymbol{x} \in X, t} \ t \quad \text{s.t.} \quad f_0(\boldsymbol{x}, \boldsymbol{u}_0) \le t, \ \forall \boldsymbol{u}_0 \in \mathcal{U}_0$$

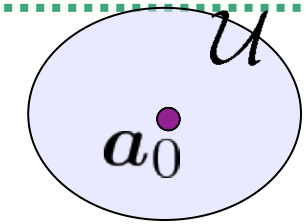
Tractable Robust LP (Ellipsoidal Case)

Ben-Tal & Nemirovski ['99]

$$\min_{\boldsymbol{x}} \ \boldsymbol{c}^\top \boldsymbol{x} \quad \text{s.t.} \quad \boldsymbol{a}(\boldsymbol{u})^\top \boldsymbol{x} \le b, \quad \forall \boldsymbol{u} \in \mathcal{U}$$

Ellipsoidal uncertainty set:

$$\boldsymbol{a}(\boldsymbol{u}) = \boldsymbol{a}_0 + \boldsymbol{A}\boldsymbol{u} \quad \mathcal{U} = \{\ \boldsymbol{u} : \|\boldsymbol{u}\|_2 \le 1\}$$



$$\min_{\boldsymbol{x}} \ \boldsymbol{c}^\top \boldsymbol{x} \quad \text{s.t.} \quad \boldsymbol{a}_0^\top \boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{u} \le b, \quad \|\boldsymbol{u}\|_2 \le 1,$$

$$\boldsymbol{a}_0^\top \boldsymbol{x} + \left(\max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2 \le 1} \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{u} \right) \le b$$

$$\boldsymbol{u}^* = \frac{\boldsymbol{A}^\top \boldsymbol{x}}{\|\boldsymbol{A}^\top \boldsymbol{x}\|_2}$$

$$\min_{\boldsymbol{x}} \ \boldsymbol{c}^\top \boldsymbol{x} \quad \text{s.t.} \quad \boldsymbol{a}_0^\top \boldsymbol{x} + \|\boldsymbol{A}^\top \boldsymbol{x}\|_2 \le b$$

Second order cone programming (SOCP)

Tractable Robust LP (Rectangle Case)

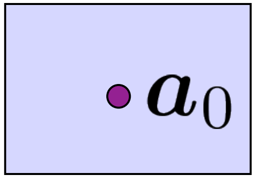
Soyster ['73]

$$\min_{\boldsymbol{x}} \ \boldsymbol{c}^\top \boldsymbol{x} \quad \text{s.t.} \quad \boldsymbol{a}^\top \boldsymbol{x} \le b, \quad \forall \boldsymbol{a} \in \mathcal{U}$$

Rectangle :

$$\mathcal{U} = \{\boldsymbol{u} : \boldsymbol{a}_0 - \bar{\boldsymbol{a}} \le \boldsymbol{u} \le \boldsymbol{a}_0 + \bar{\boldsymbol{a}}\} \subset R^n$$

where $\bar{\boldsymbol{a}} \ge \boldsymbol{0}$



$$\max_{\boldsymbol{a} \in \mathcal{U}} \boldsymbol{a}^\top \boldsymbol{x} = \boldsymbol{a}_0^\top \boldsymbol{x} + \bar{\boldsymbol{a}}^\top |\boldsymbol{x}| \le b$$

A vector constructed by taking absolute values for each element of $\boldsymbol{x}$

Linear Programming Problem

$$\begin{aligned} \min_{\boldsymbol{x},\boldsymbol{y}} \quad & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{a}_0^\top \boldsymbol{x} + \bar{\boldsymbol{a}}^\top \boldsymbol{y} \le b, \quad -\boldsymbol{y} \le \boldsymbol{x} \le \boldsymbol{y}, \quad \boldsymbol{y} \ge \boldsymbol{0} \end{aligned}$$

Conditions for Tractable Robust Optimization

$$\min_{\boldsymbol{x}\in X} \quad \boldsymbol{c}^\top \boldsymbol{x} \quad \text{s.t.} \quad f_i(\boldsymbol{x}, \boldsymbol{u}_i) \le 0, \quad \forall \boldsymbol{u}_i \in \mathcal{U}_i,$$
$$i = 1, \ldots, m$$

→ Want to transform it to a tractable convex prob.

Ben-Tal & Nemirovski ['98], Goldfarb & Iyengar ['03]

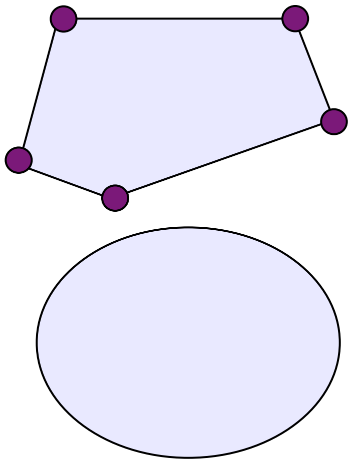
Three Assumptions

(1) $f(\boldsymbol{x}, \boldsymbol{u})$ is convex quadratic in terms of $\boldsymbol{x}$.

$$f(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{x}^\top \boxed{\boldsymbol{Q}(\boldsymbol{u})} \boldsymbol{x} + \boxed{\boldsymbol{q}(\boldsymbol{u})}^\top \boldsymbol{x} + \boxed{\gamma(\boldsymbol{u})}$$

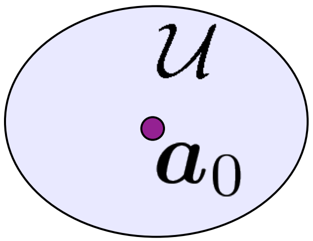
(2) Uncertain data is linear w.r.t $\boldsymbol{u}$.

$$\boldsymbol{Q}(\boldsymbol{u}) = \boldsymbol{Q}_0 + \textstyle\sum_i \boldsymbol{Q}_i u_i$$
$$\boldsymbol{q}(\boldsymbol{u}) = \boldsymbol{q}_0 + \textstyle\sum_i \boldsymbol{q}_i u_i$$



(3) $\mathcal{U}_i$ is a finite set, its convex hull or ellipsoid.

Difficulty of Solving Problems



<u>Assump.</u>

✓ $\mathcal{U}$ is an ellipsoidal uncertainty set

✓ Uncertain data is linear with respect to $\boldsymbol{u} \in \mathcal{U}$

$$\boldsymbol{a}(\boldsymbol{u}) = \boldsymbol{a}_0 + \sum_i \boldsymbol{a}_i u_i, \quad \boldsymbol{F}(\boldsymbol{u}) = \boldsymbol{F}_0 + \sum_i \boldsymbol{F}_i u_i$$

- Robust LP → Second-order Cone Programming（SOCP）
- Robust SOCP → Semidefinite Programming（SDP）
- Robust SDP → ×

  Approximately solved by SDP

Tips on Formulation of Robust Optimization

With robust optimization ... ..

- ✓ How to express uncertainty data is important!
- ✓ There is a great limitation on its expression
  - Uncertainty data is linear w.r.t u.
  - The range for u is an ellipse, etc.

If these conditions are satisfied, a RO problem can be converted to a tractable problem.

In the case where the condition is not satisfied

⇨ stochastic approach by sampling a finite number of constraints among infinitely many constraints

Contents

- Robust Optimization
  - ✓modeling strategies and solution methods for optimization problems that are defined by uncertain inputs
  - ✓proposed by Ben-Tal & Nemirovski in 1998

- Stochastic Programming
  - ✓classical framework for modeling optimization problems involving uncertainty (studied since the 1950’s).
  - ✓assuming that probability distributions are known
  - ✓relation to robust optimization

Stochastic Programming

Uncertain Optimization Problem:

$$\min_{\boldsymbol{x} \in X} f(\boldsymbol{x}, \boldsymbol{u}_0) \text{ s.t. } g(\boldsymbol{x}, \boldsymbol{u}_1) \leq 0$$

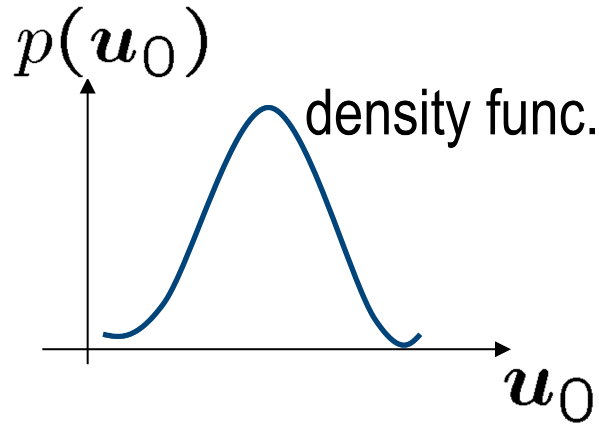
$\boldsymbol{u}_0, \boldsymbol{u}_1$: uncertain data

●Stochastic Programming

Dantzig ['55], Beale ['55]

Assump. :

Prob. distributions of $\boldsymbol{u}_0, \boldsymbol{u}_1$ are known.



ex.1) $$\min_{\boldsymbol{x} \in X} E_{\boldsymbol{u}_0}[f(\boldsymbol{x}, \boldsymbol{u}_0)]$$

$$\text{s.t. } E_{\boldsymbol{u}_1}[g(\boldsymbol{x}, \boldsymbol{u}_1)] \leq 0$$

ex.2) Chance Const.（Probabilistic Const. ）

Charnes & Cooper ['59]

$$\Pr_{\boldsymbol{u}_1}(g(\boldsymbol{x}, \boldsymbol{u}_1) \leq 0) \geq 1 - \epsilon$$

Examples of Another Risk Measure

Rockafellar & Uryasev ['02]

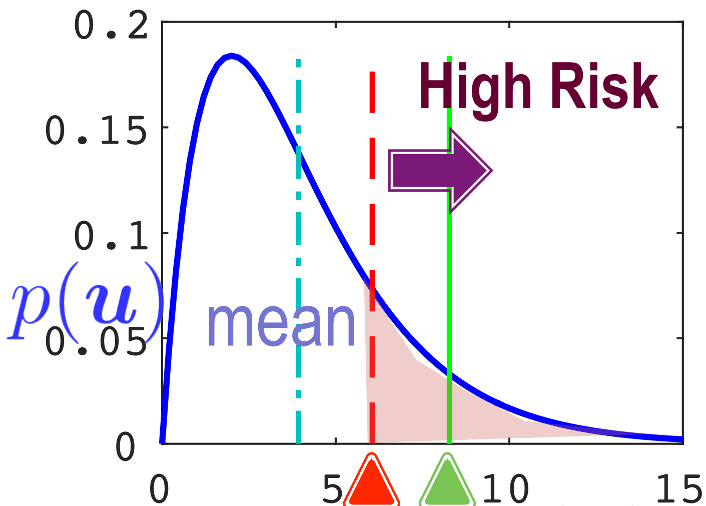
Instead of "Expectation", risk measure "CVaR" is often used.

$$\min_{\boldsymbol{x}\in X} \boxed{E_{\boldsymbol{u}}[f(\boldsymbol{x},\boldsymbol{u})]} \quad \Rightarrow \quad \min_{\boldsymbol{x}\in X} \boxed{\phi_\beta(\boldsymbol{x})} \qquad \beta \in (0,1)$$

CVaR (**C**onditional **V**alue-**at**-**R**isk) : $\phi_\beta(\boldsymbol{x})$

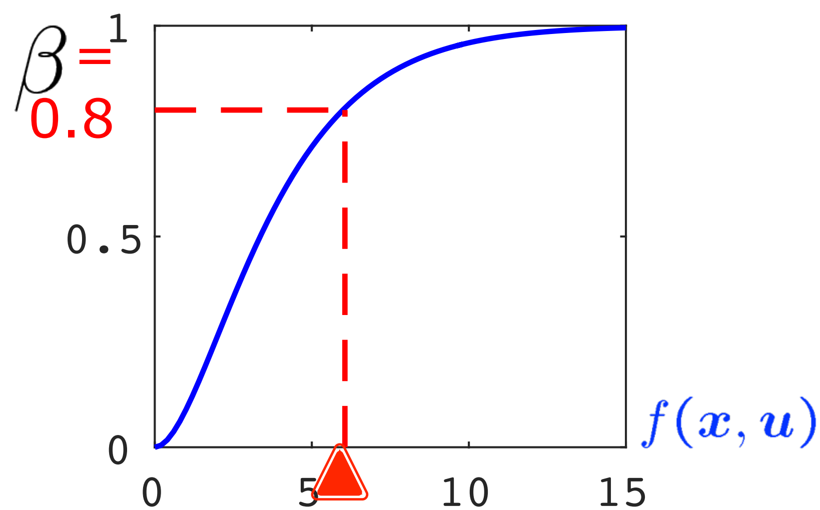
Conditional expectation of $f(\boldsymbol{x},\boldsymbol{u})$ exceeding β-quantile $\alpha_\beta(\boldsymbol{x})$

density function



β-quantile (VaR): $\alpha_\beta(\boldsymbol{x})$ $\phi_\beta(\boldsymbol{x})$

cdf



β-quantile (VaR): $\alpha_\beta(\boldsymbol{x})$

Definition of Conditional Value-at-Risk (CVaR)

Rockafellar & Uryasev ['02]

$\beta \in (0, 1)$

$\alpha_\beta(\boldsymbol{x})$: β-VaR (= β-quantile)

$\phi_\beta(\boldsymbol{x})$: β-CVaR

of the loss $f(\boldsymbol{x}, \boldsymbol{u})$ associated with a decision $\boldsymbol{x}$

random vec.

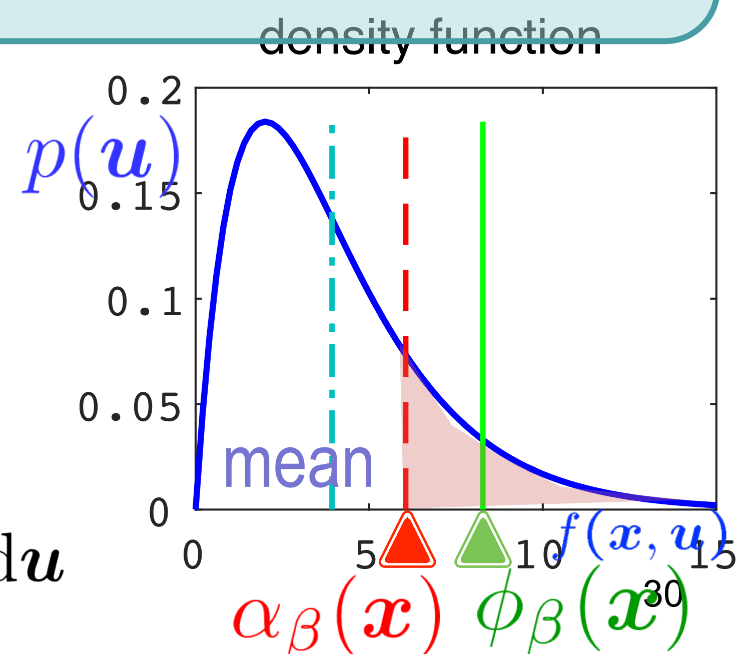
Conditional expectation of $f(\boldsymbol{x}, \boldsymbol{u})$ exceeding β-quantile $\alpha_\beta(\boldsymbol{x})$

$$\phi_\beta(\boldsymbol{x}) = \frac{1}{1-\beta} \int_{f(\boldsymbol{x},\boldsymbol{u}) \geq \alpha_\beta(\boldsymbol{x})} f(\boldsymbol{x}, \boldsymbol{u}) p(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u}$$

$$\alpha_\beta(\boldsymbol{x}) \in \arg\min_{\alpha} F_\beta(\boldsymbol{x}, \alpha)$$

$$\phi_\beta(\boldsymbol{x}) = \min_{\alpha} F_\beta(\boldsymbol{x}, \alpha)$$

$$F_\beta(\boldsymbol{x}, \alpha) := \alpha + \frac{1}{1-\beta} \int_{\boldsymbol{u}} [f(\boldsymbol{x}, \boldsymbol{u}) - \alpha]^+ p(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u}$$



CVaR for Discrete Distribution

When random variables follow a discrete dist. or normal dist., CVaR minimization can be tractable.

Rockafellar & Uryasev ['02]

ex.) For some $\beta \in (0,1)$ and $\boldsymbol{x}$,

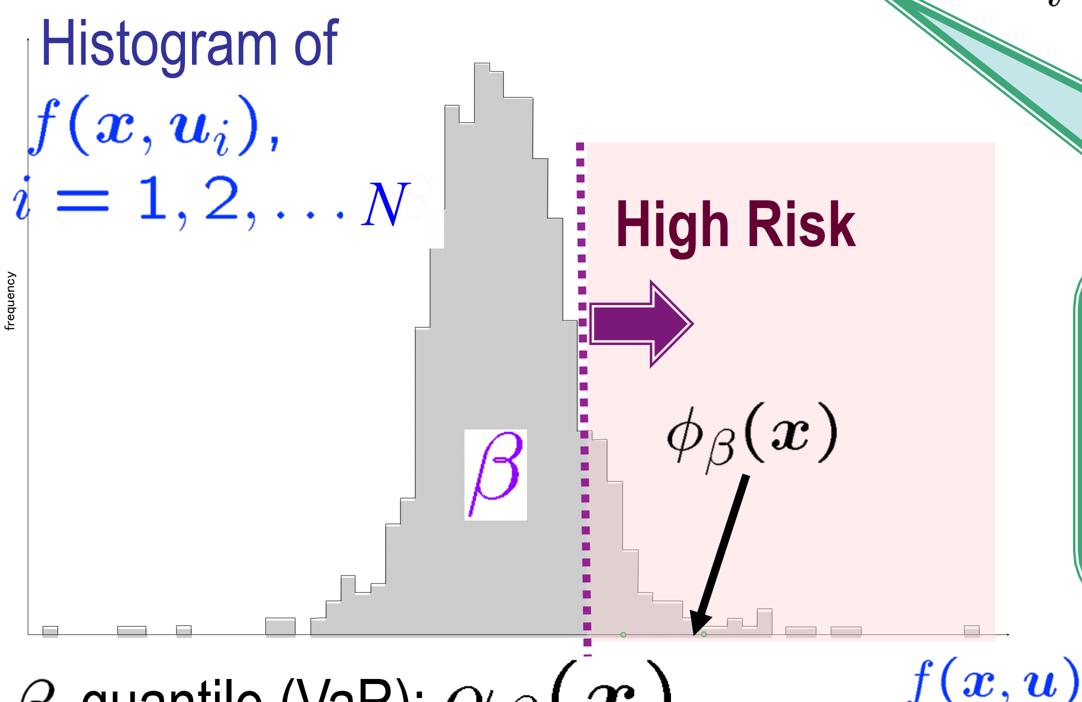
$$\phi_\beta(\boldsymbol{x}) = \min_{\alpha} \ \alpha + \frac{1}{1-\beta} \sum_{i=1}^{N} p_i [f(\boldsymbol{x}, \boldsymbol{u}_i) - \alpha]^+$$

opt.sol: $\alpha^* \approx \alpha_\beta(\boldsymbol{x})$

For the finite support:

$$\mathcal{U} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\}$$

$$\Pr(\boldsymbol{u} = \boldsymbol{u}_i) = p_i$$



β-quantile (VaR): $\alpha_\beta(\boldsymbol{x})$

Tractable Form for CVaR Minimization

Rockafellar & Uryasev ['02]

$$\min_{\boldsymbol{x}\in X} \phi_\beta(\boldsymbol{x})$$

$$\Rightarrow \quad \min_{\boldsymbol{x}\in X,\alpha} \alpha + \frac{1}{1-\beta}\sum_{i=1}^{N} p_i[f(\boldsymbol{x},\boldsymbol{u}_i)-\alpha]^+$$

$$\Rightarrow \quad \min_{\boldsymbol{z},\boldsymbol{x},\alpha} \alpha + \frac{1}{1-\beta}\sum_{i=1}^{N} p_i z_i$$

$$\text{s.t.} \quad f(\boldsymbol{x},\boldsymbol{u}_i)-\alpha-z_i \le 0, \ \forall i$$

$$\boldsymbol{z} \ge \mathbf{0}, \quad \boldsymbol{x} \in X$$

If $f(\boldsymbol{x},\boldsymbol{u}_i)$ is convex in $\boldsymbol{x}$ and X is a convex set, this is a convex optimization prob.

Parameter β of CVaR

$\beta \in (0, 1)$

CVaR (**C**onditional **V**alue-**at**-**R**isk) : $\phi_\beta(\boldsymbol{x})$

Conditional expectation of $f(\boldsymbol{x}, \boldsymbol{u})$ exceeding β-quantile $\alpha_\beta(\boldsymbol{x})$

$$\min_{\boldsymbol{x} \in X} \boxed{\phi_\beta(\boldsymbol{x})}$$

$\beta \to 0$

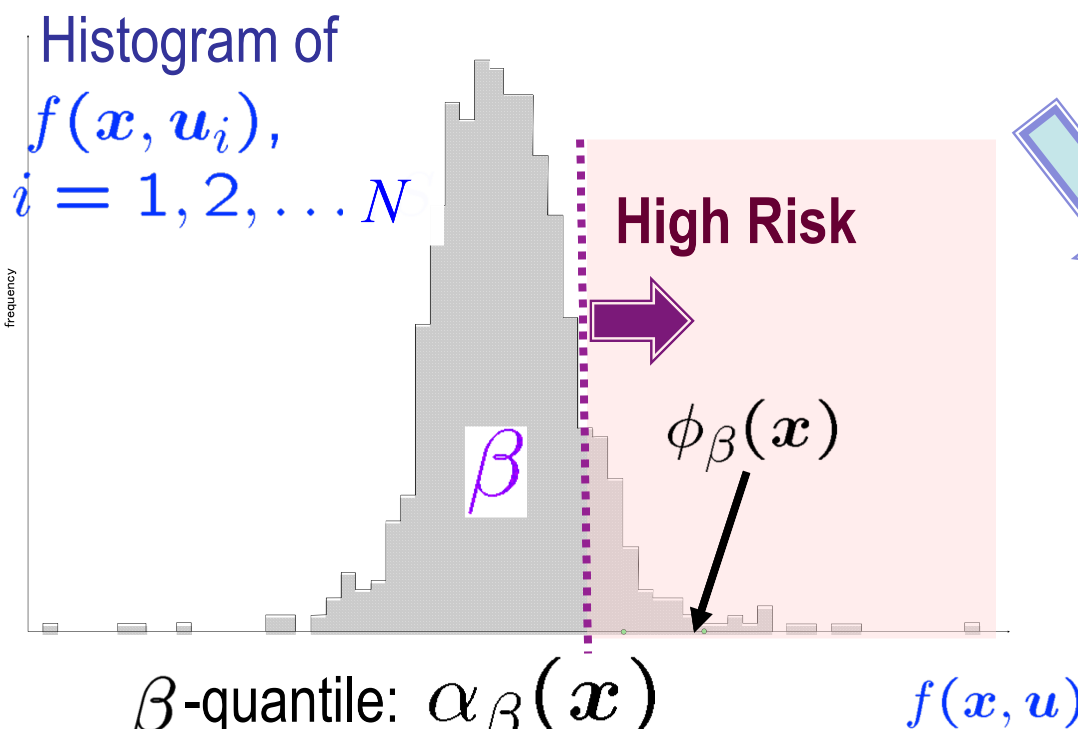
$$\min_{\boldsymbol{x} \in X} E_{\boldsymbol{u}}[f(\boldsymbol{x}, \boldsymbol{u})]$$

traditional stochastic program.

$\beta \to 1$

$$\min_{\boldsymbol{x} \in X} \max_{i=1,\ldots,N} f(x, u_i)$$

robust optimization



CVaR for Normal Distribution

$\beta \in (0,1)$

CVaR (**C**onditional **V**alue-**at**-**R**isk) : $\phi_\beta(\boldsymbol{x})$

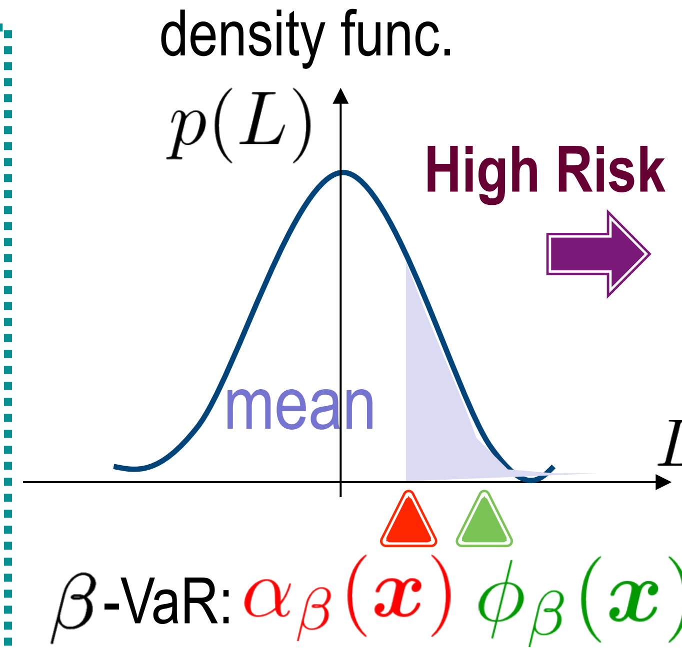
Conditional expectation of $f(\boldsymbol{x}, \boldsymbol{u})$ exceeding β-quantile $\alpha_\beta(\boldsymbol{x})$

Random variable: $\boldsymbol{u} \sim \mathcal{N}_n(\bar{\boldsymbol{u}}, \Sigma)$

$$L = \boldsymbol{u}^\top \boldsymbol{x} \sim \mathcal{N}(\underbrace{\bar{\boldsymbol{u}}^\top \boldsymbol{x}}_{\mu}, \underbrace{\boldsymbol{x}^\top \Sigma \boldsymbol{x}}_{\sigma^2})$$

Probability density of the normal dist. :

$$p(L) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(L-\mu)^2}{2\sigma^2}\right)$$



CVaR is defined as $\phi_\beta(\boldsymbol{x}) = \dfrac{1}{1-\beta} \displaystyle\int_{\alpha_\beta}^{\infty} L \cdot p(L)\, \mathrm{d}L$

$$= \bar{\boldsymbol{u}}^\top \boldsymbol{x} + C\sqrt{\boldsymbol{x}^\top \Sigma \boldsymbol{x}}$$

Probabilistic Constraint

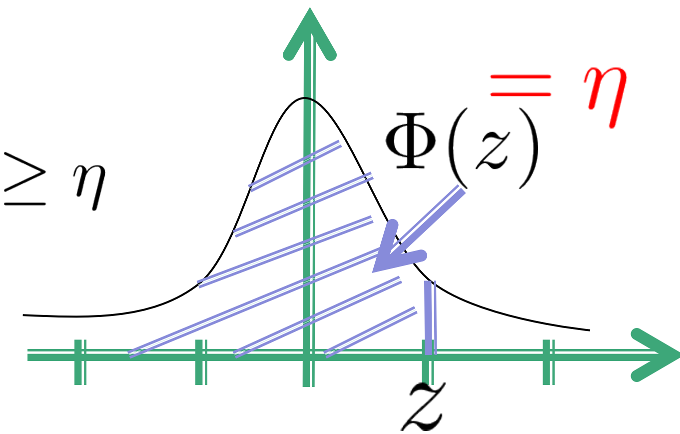
$$\boxed{\Pr_{\boldsymbol{u}}(g(\boldsymbol{x},\boldsymbol{u}) \le 0) \ge 1-\epsilon}$$

ex.)

$$\Pr_{\boldsymbol{u}}(\boldsymbol{u}^\top \boldsymbol{x} \le b) \ge \eta \qquad (ただし, \eta \ge 0.5\ )$$

Under the assump.: $\boldsymbol{u} \sim \mathcal{N}_n(\bar{\boldsymbol{u}}, \Sigma)$

$$\Longleftrightarrow \quad \Pr_{\boldsymbol{u}}\left(\underbrace{\frac{\boldsymbol{u}^\top \boldsymbol{x} - \bar{\boldsymbol{u}}^\top \boldsymbol{x}}{\sqrt{\boldsymbol{x}^\top \Sigma \boldsymbol{x}}}}_{\sim\mathcal{N}(0,1)} \le \frac{b - \bar{\boldsymbol{u}}^\top \boldsymbol{x}}{\sqrt{\boldsymbol{x}^\top \Sigma \boldsymbol{x}}} \right) \ge \eta$$



$$\Longleftrightarrow \quad \frac{b - \bar{\boldsymbol{u}}^\top \boldsymbol{x}}{\sqrt{\boldsymbol{x}^\top \Sigma \boldsymbol{x}}} \ge \Phi^{-1}(\eta)$$

: η-quantile

$\Phi(z)$: cumulative dist. func. (cdf) of $\mathcal{N}(0,1)$

second-order cone constr.

$$\Longleftrightarrow \quad \bar{\boldsymbol{u}}^\top \boldsymbol{x} + \Phi^{-1}(\eta) \|\Sigma^{1/2} \boldsymbol{x}\| \le b$$

Relation to Robust Constraint

Probabilistic Const.

Assump.: $\boldsymbol{u} \sim \mathcal{N}_n(\bar{\boldsymbol{u}}, \Sigma)$

$$\Pr_{\boldsymbol{u}}(\boldsymbol{u}^\top \boldsymbol{x} \le b) \ge \eta \iff \bar{\boldsymbol{u}}^\top \boldsymbol{x} + \Phi^{-1}(\eta)\|\Sigma^{1/2}\boldsymbol{x}\| \le b$$

$$\Updownarrow$$

Robust Const.

Assump.: $\boldsymbol{u} \in \mathcal{U} := \{\bar{\boldsymbol{u}} + \Sigma^{1/2}\boldsymbol{v} : \|\boldsymbol{v}\| \le \Phi^{-1}(\eta)\}$

$$\max_{\boldsymbol{u}\in\mathcal{U}} \boldsymbol{u}^\top \boldsymbol{x} \le b$$

$$\iff \bar{\boldsymbol{u}}^\top \boldsymbol{x} + \max_{\boldsymbol{v}:\|\boldsymbol{v}\|\le\Phi^{-1}(\eta)} \boldsymbol{x}^\top \Sigma^{1/2}\boldsymbol{v} \le b$$

$$= \bar{\boldsymbol{u}}^\top \boldsymbol{x} + \Phi^{-1}(\eta)\|\Sigma^{1/2}\boldsymbol{x}\|$$

Stochastic Interpretation for Uncertainty Set

Assump.: $\boldsymbol{u} \sim \mathcal{N}_n(\bar{\boldsymbol{u}}, \Sigma)$

$$\Pr_{\boldsymbol{u}}(\boldsymbol{u}^\top \boldsymbol{x} \le b) \ge \eta$$

Relation of two Asumptions?

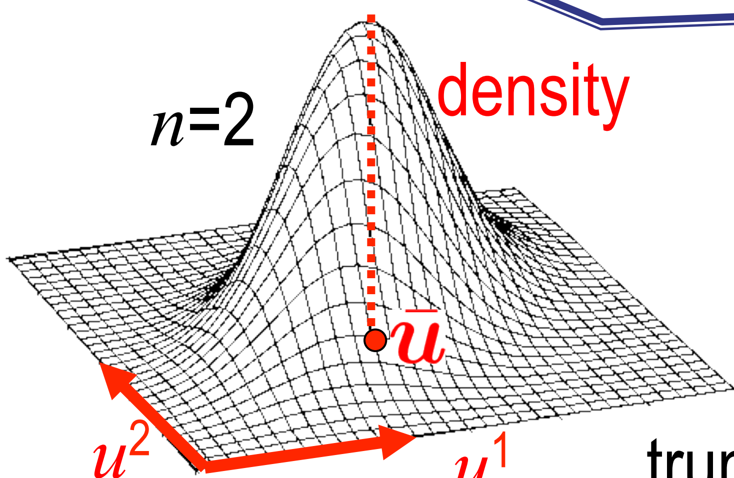
Assump.: $\boldsymbol{u} \in \mathcal{U} := \{\bar{\boldsymbol{u}} + \Sigma^{1/2}\boldsymbol{v} : \|\boldsymbol{v}\| \le \Phi^{-1}(\eta)\}$

$$\max_{\boldsymbol{u} \in \mathcal{U}} \boldsymbol{u}^\top \boldsymbol{x} \le b$$

$$\Pr(\boldsymbol{u} \in \mathcal{U}) = \mathcal{F}_n((\Phi^{-1}(\eta))^2)$$

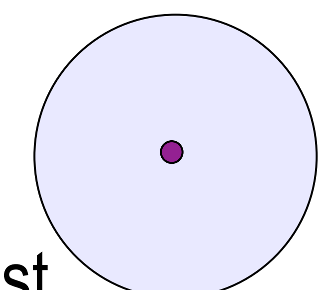
chi-squared distribution with n degrees of freedom

n=2



support for truncated normal dist.

$100\mathcal{F}_n((\Phi^{-1}(\eta))^2)$%
data are covered



Two Optimization Methods under Uncertainty

$$\min_{\boldsymbol{x} \in X} f(\boldsymbol{x}, \boldsymbol{u}_0) \text{ s.t. } g(\boldsymbol{x}, \boldsymbol{u}_1) \leq 0$$

$\boldsymbol{u}_0, \boldsymbol{u}_1$
: uncertain data

Probabilistic Const.

Assump.: $\boldsymbol{u} \sim \mathcal{N}_n(\bar{\boldsymbol{u}}, \Sigma)$

Robust Const.

Assump.: $\boldsymbol{u} \in \mathcal{U} := \{\bar{\boldsymbol{u}} + \Sigma^{1/2}\boldsymbol{v} : \|\boldsymbol{v}\| \leq \Phi^{-1}(\eta)\}$

Boundary between two methods is getting blurred.

Recently, studies on robust optimization using "probability" are increased e.g. for setting the uncertainty set $\mathcal{U}$.

Stochastic Approach for Robust Optimization

Among three assumptions for tractable robust optimization,
(2) Uncertain data is linear w.r.t $\boldsymbol{u}$
(3) $\mathcal{U}$ is a finite set, its convex hull or ellipsoid
can be removed.

$\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$: randomly generated following the distribution on $\mathcal{U}$

Solve a relaxation problem having a finite number of const.

Calafiore & Campi ['05]

Want to estimate the sample size *N* to obtain
a relaxed solution with theoretical guarantee.

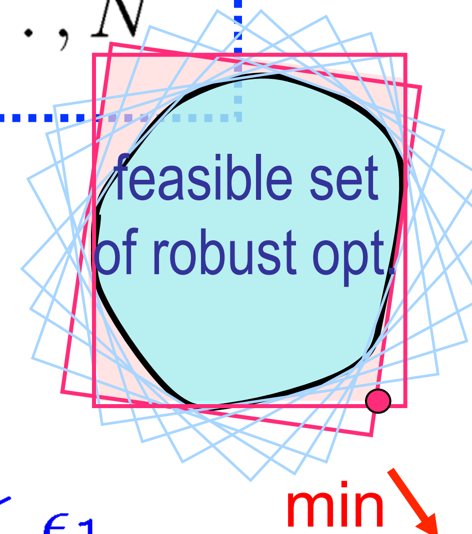
How to determine the sample size *N*

$$\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N \overset{\text{i.i.d.}}{\sim} P$$ (Assume the probability distribution on $\mathcal{U}$)

Randomly generated relaxation problem (SCP_N) :

$$\min_{\boldsymbol{x} \in X} \; \boldsymbol{c}^\top \boldsymbol{x} \quad \text{s.t.} \quad f(\boldsymbol{x}, \boldsymbol{u}_i) \le 0, \; i = 1, \ldots, N$$

Optimal sol. of (SCP_N) : $\widehat{\boldsymbol{x}}_N$



Criteria for deciding $\boldsymbol{N}$:

- Allow $\widehat{\boldsymbol{x}}_N$ to violate some ratio of constraints:

$$V(\widehat{\boldsymbol{x}}_N) = P\{\boldsymbol{u} \in \mathcal{U} : f(\widehat{\boldsymbol{x}}_N, \boldsymbol{u}) > 0\} \le \epsilon_1$$

Calafiore & Campi ['05, '06]

- Allow some amount of constraint violation for $\widehat{\boldsymbol{x}}_N$:

$$\max_{\boldsymbol{u} \in \mathcal{U}} f(\widehat{\boldsymbol{x}}_N, \boldsymbol{u}) \le \epsilon_2$$

Kanamori & Takeda ['12]

Evaluation for Sample Size

$$N(\epsilon,\eta) := \frac{2}{\epsilon}\log\frac{1}{\eta} + 2n + \frac{2n}{\epsilon}\log\frac{2}{\epsilon}$$

Calafiore & Campi ['06]

$$N(\epsilon,\eta) := \min\left\{ N \in \mathbb{N} \mid \sum_{i=0}^{n-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \le \eta \right\}$$

Campi & Garatti ['08]

Theo. (Calafiore & Campi ['05,'06], Campi & Garatti ['08])

Let $\epsilon \in (0,1), \ \eta \in (0,1)$.

The optimal solution $\hat{\boldsymbol{x}}_N \in R^n$ of (SCP_N) generated with $N \ge N(\epsilon,\eta)$ samples satisfies $V(\hat{\boldsymbol{x}}_N) \le \epsilon$ with the probability at least $1-\eta$, that is,

$$P^N\{\ V(\hat{\boldsymbol{x}}_N) \le \epsilon\ \} \ge 1-\eta$$

$\epsilon \to 0, \eta \to 0 \ \Rightarrow \ N(\epsilon,\eta) \to \infty$

Violation probability: $V(\hat{\boldsymbol{x}}_N) = P\{\boldsymbol{u} \in \mathcal{U} : f(\hat{\boldsymbol{x}}_N, \boldsymbol{u}) > 0\}$

A-priori / A-posteriori Evaluations

Takeda, Taguchi & Tanaka ['10]

(A-priori Evaluation)

(construction of function *q is* a key idea)

$$\epsilon \in (0, q(B)),\ \ \eta \in (0,1),\ \ N \geq N(\epsilon, \eta)$$

$\Rightarrow$ Optimal sol. $\hat{\boldsymbol{x}}_N$ of (SCP_N) satisfies

$$P^N\{\ V(\hat{\boldsymbol{x}}_N) \leq \epsilon,\ \max_{\boldsymbol{u}\in\mathcal{U}} f(\hat{\boldsymbol{x}}_N, \boldsymbol{u}) \leq q^{-1}(\epsilon) \} \geq 1-\eta$$

independent from $\hat{\boldsymbol{x}}_N$

(A-posteriori Evaluation)

Optimal sol. $\hat{\boldsymbol{x}}_N$ of (SCP_N), N > 0, satisfies

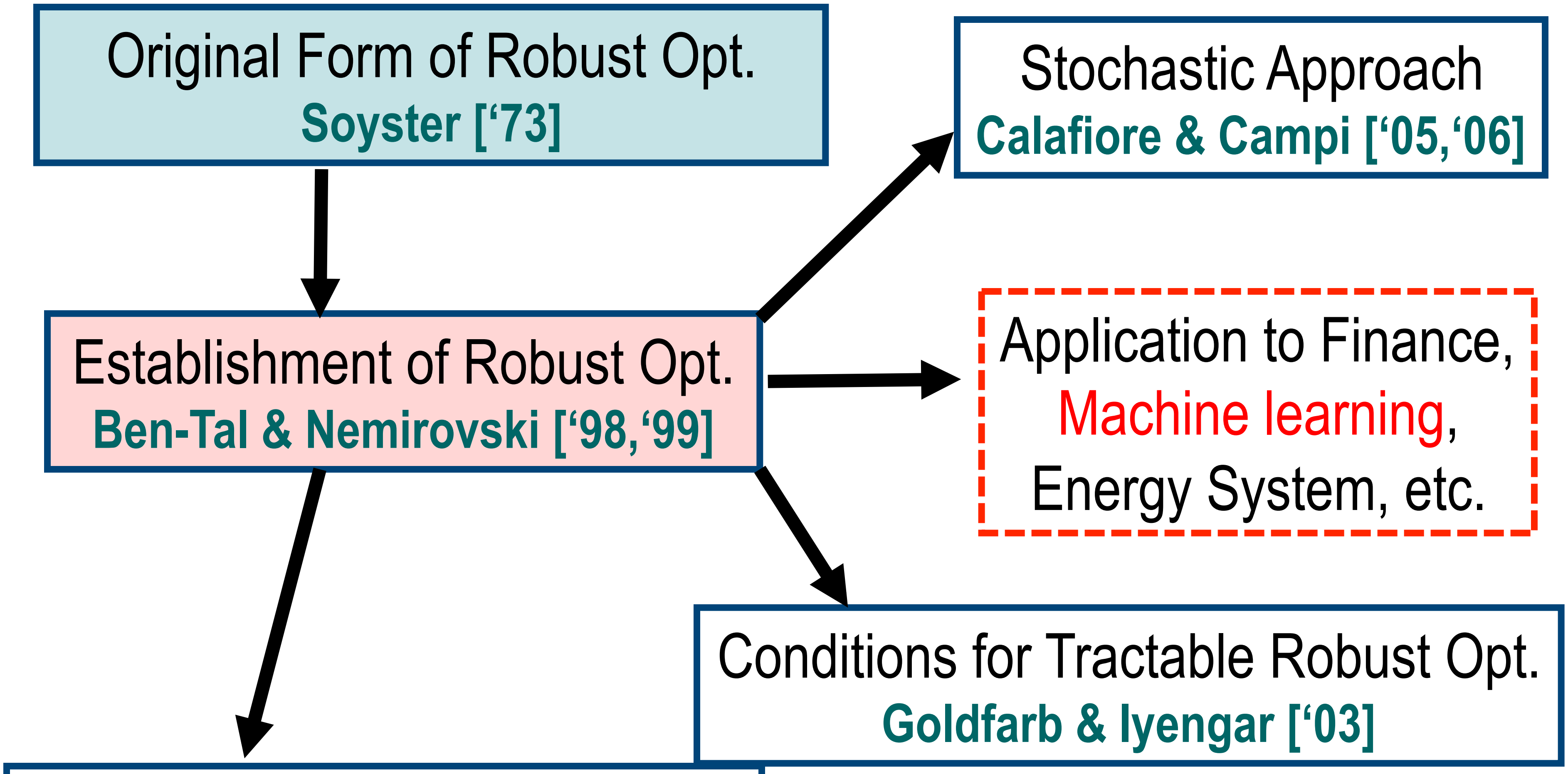
$$\delta \in (0, B],\ \ \eta \in (0,1),\ \ M \geq M(\delta, \eta) := \frac{\ln \eta}{\ln(1-q(\delta))},$$

$$\tilde{\boldsymbol{u}}_1, \ldots, \tilde{\boldsymbol{u}}_M \ \sim\ P,$$

depending on $\hat{\boldsymbol{x}}_N$

$$P^M\{\ \max_{\boldsymbol{u}\in\mathcal{U}} f(\hat{\boldsymbol{x}}_N, \boldsymbol{u}) < \max_{i=1,\ldots,M} f(\hat{\boldsymbol{x}}_N, \tilde{\boldsymbol{u}}_i) + \delta \} \geq 1-\eta$$

Various Research Directions



ロバスト最適化や確率計画法の機械学習問題への適用

統計数理研究所 / 理化学研究所AIP センター
武田朗子

Optimization Techniques in ML

- There are trends in optimization techniques used in ML
  - ✓semidefinite program
  - ✓submodular optimization
  - ✓first-order methods such as APG, ADMM, etc.

- Stochastic Program. and Robust Optimization are not popular in ML
  - ✓but they are implicitly used.

Contents

- Provide a view based on Robust Optimization for various Binary Classification Models including
  - ✓Support Vector Machine (SVM),
    Minimax Probability Machine (MPM) and
    Fisher Discriminant Analysis (FDA), etc.

- Provide a view based on Stochastic Programming
  - ✓ν-SVM & Eν-SVM → Generalization Bound
  - ✓Minimum Margin MPM

Application of Robust Optimization to ML

- ✓ Introducing the work of Xu, Caramanis and Mannor [2009]
- ✓ Showing a unified view for various ML models such as SVM MPM, FDA, logistic regression.

We use robust optimization techniques in a different problem setting

Binary Classification Problem

extendable to nonlinear one using kernel

Find a decision function $f(\boldsymbol{x}) = \widehat{\boldsymbol{w}}^\top \boldsymbol{x} + \widehat{b}$

based on given training samples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)$

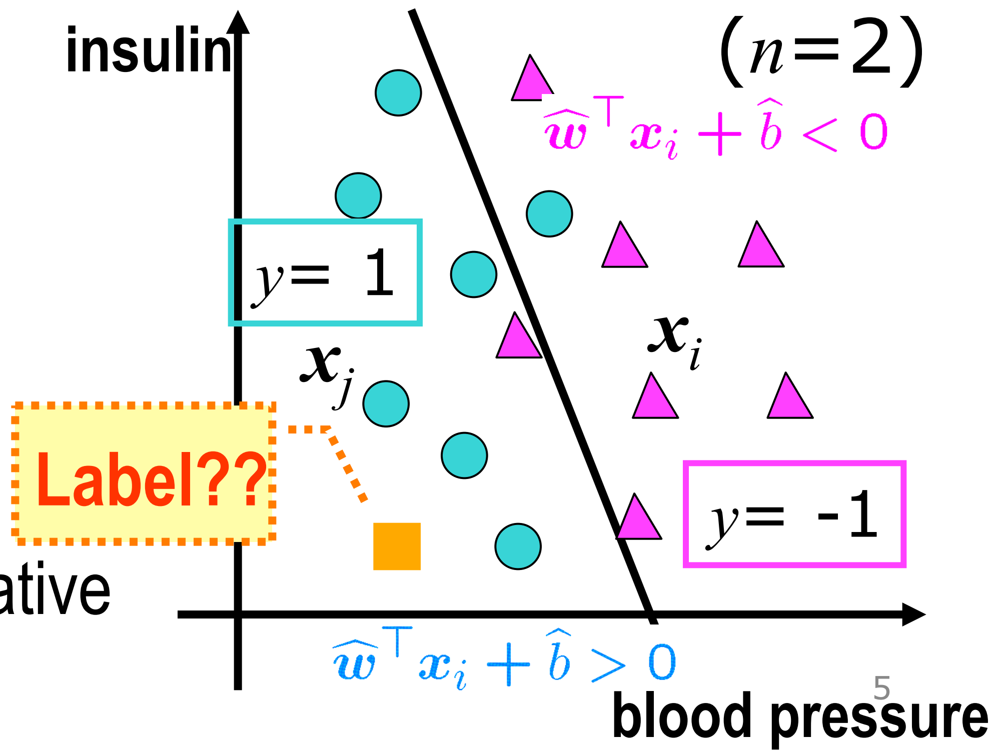
to correctly classify new samples.

EX.) diagnosis of diabetes

$\boldsymbol{x}_i \in R^n$ ← medical examination

$y_i \in \{\pm 1\}$ ← tested positive/negative

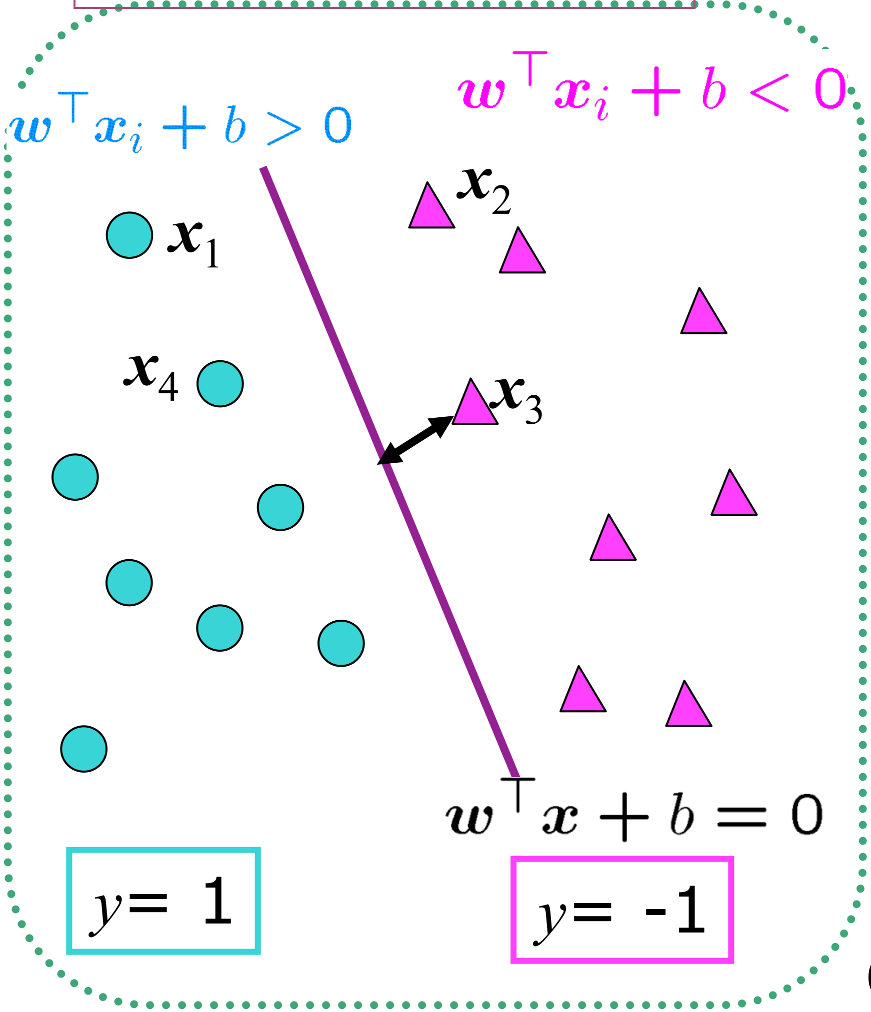
$i \in M := \{1, 2, \ldots, m\}$



Hard margin SVM (support vector machine)

Boser, Guyon & Vapnik ['92]

Linearly Separable



Maximize the minimum distance to the hyperplane

$$\max_{w\neq 0,b} \ \overbrace{\min_{i=1,\ldots,m} \ y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)}^{=1} \Big/ \|\boldsymbol{w}\|$$

regularization penalty

$$\min_{w,b} \ \boxed{\frac{1}{2}\|\boldsymbol{w}\|^2}$$

$$\text{s.t.} \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1$$

$$i = 1, \ldots, m$$

Minimizing a regularization penalty enhances generalization performance (prediction accuracy for test dataset)

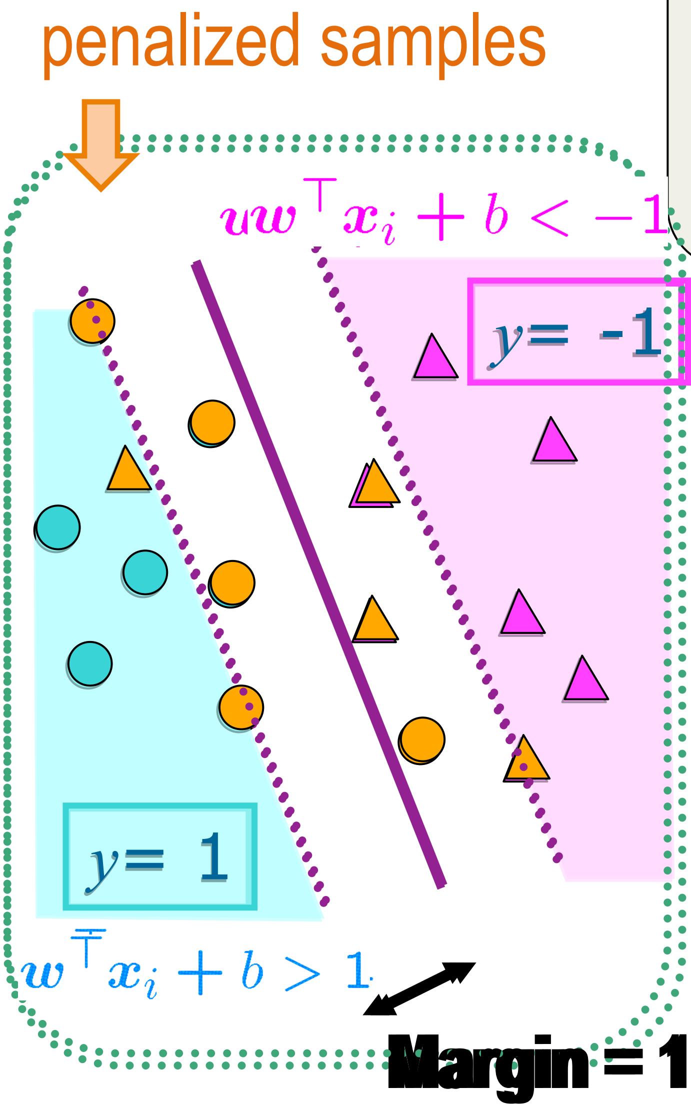
C-SVM

Cortes & Vapnik ['95]

$$\min_{\boldsymbol{w},b,\boldsymbol{z}} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} z_i$$

$$\text{s.t.} \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - z_i, \quad (i \in M)$$

$$\boldsymbol{z} \geq \boldsymbol{0}$$



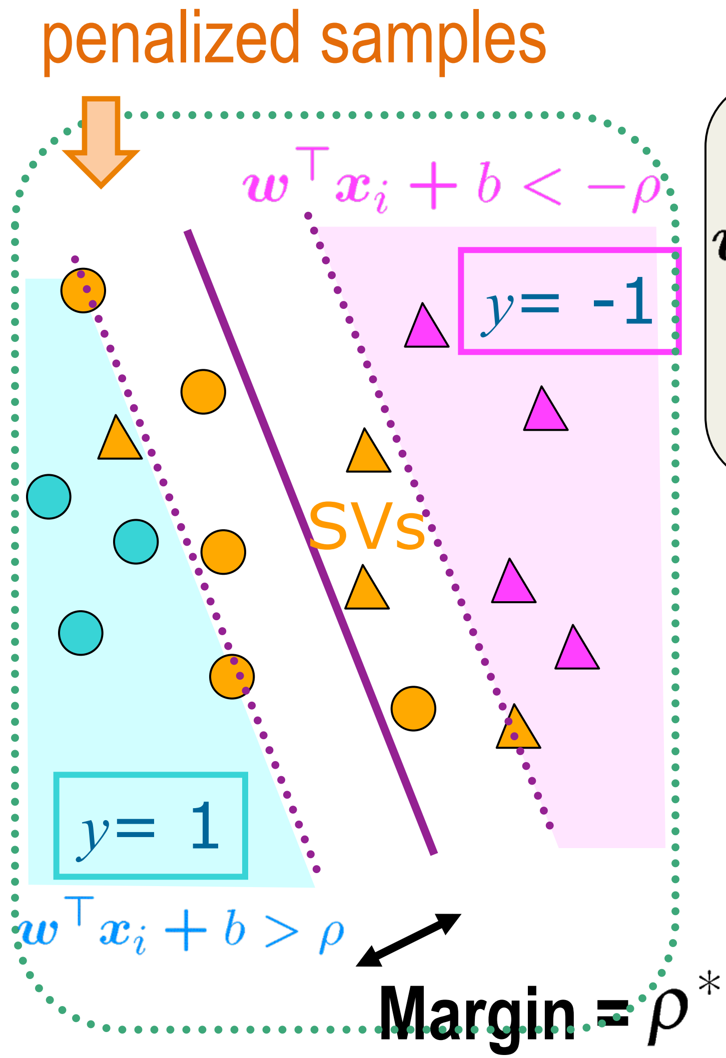
Two conflicting goals

- minimizing training error
- minimizing a regularization penalty

- the trade-off between these goals is controlled by C

ν-SVM

Scholkopf, Smola, Williamson & Bartlett ['00]



C is replaced by an intuitive parameter ν

$$\begin{array}{ll} \min\limits_{w,b,z,\rho} & \frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{m}\sum_{i=1}^{m} z_i \\ \text{s.t.} & y_i(w^\top x_i + b) \geq \rho - z_i \quad (i \in M) \\ & z \geq 0 \end{array}$$

- C-SVM with $C = \frac{1}{m\rho^*}$ $\longleftrightarrow$ ν-SVM
- margin is nonnegative : $\rho^* \geq 0$
- admissible values of ν are limited
  $(\ \nu \in (\ \nu_{\min}, \nu_{\max}\ ] \subseteq (0,1]\ )$
- **0** opt. solution for small ν

Extended ν-SVM (Eν-SVM)

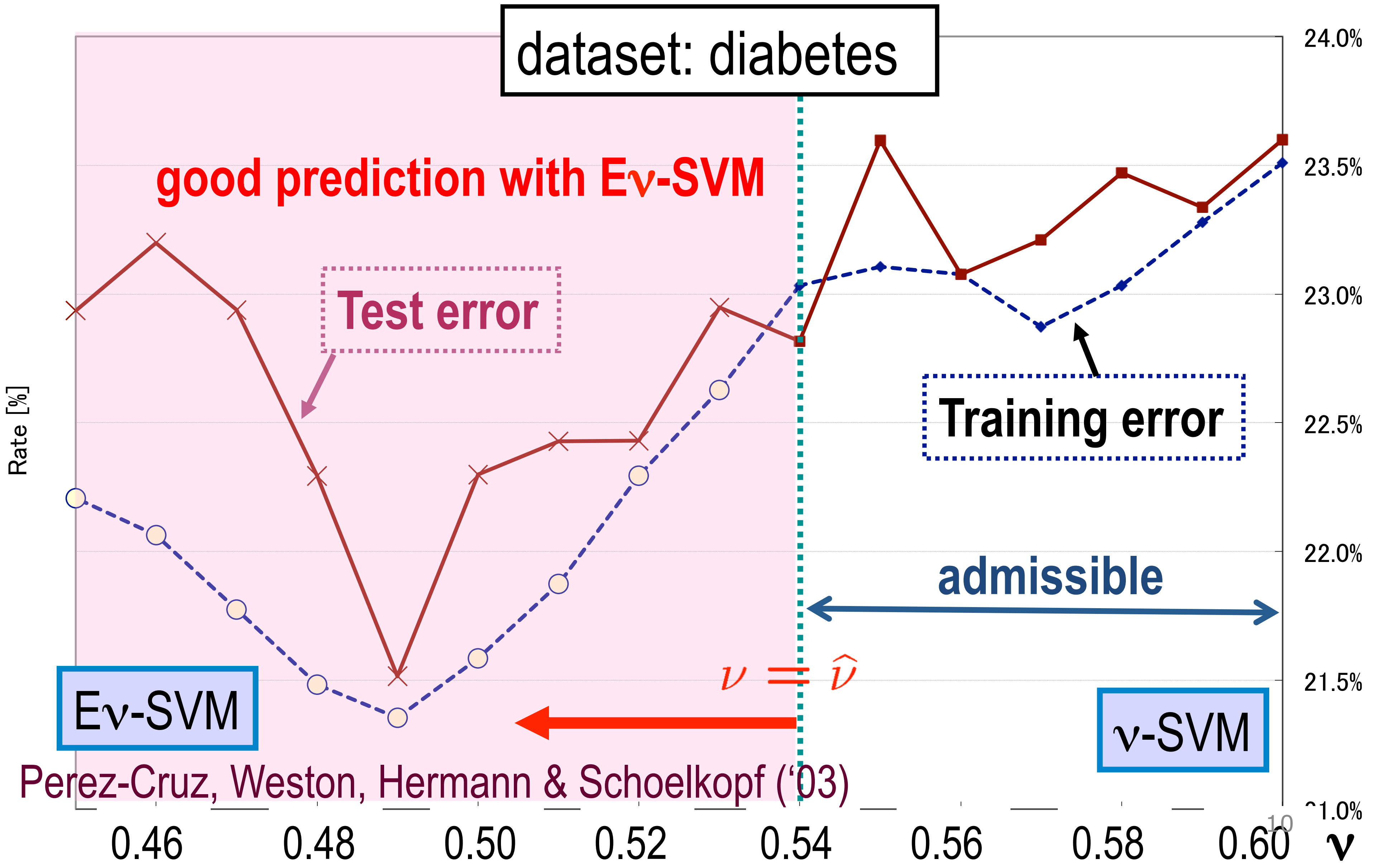
Perez-Cruz, Weston, Hermann & Scholkopf ['03]

$$
\begin{aligned}
\min_{\boldsymbol{w}, b, \boldsymbol{z}, \rho} \quad & -\nu\rho + \frac{1}{m}\sum_{i=1}^{m} z_i \\
\text{s.t.} \quad & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq \rho - z_i, \quad (i \in M) \\
& \boldsymbol{z} \geq \boldsymbol{0}, \quad \boldsymbol{w}^\top \boldsymbol{w} = 1
\end{aligned}
$$

Nonconvex optimization

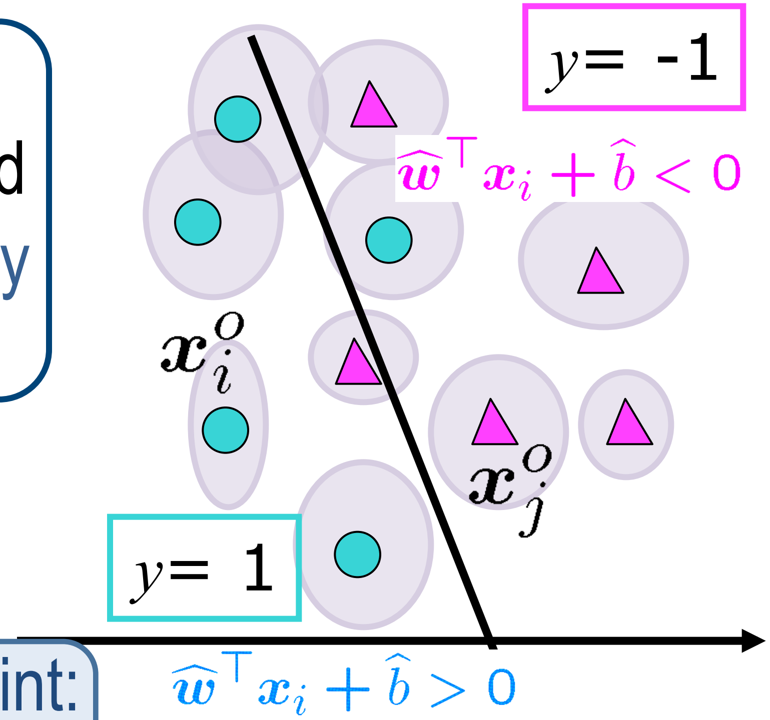
- The margin ρ^* is negative for $\nu \in (0, \ \nu_{\min}]$.
- A non-trivial solution is obtained even for the range.
- The same optimal sol. with ν-SVM for $\nu \in (\ \nu_{\min}, \nu_{\max}\ ]$
- An iterative algorithm was proposed for a local solution.

Advantage of Extended Range of ν



Uncertainty in Dataset

Bi & Zhang ('04), Shivaswamy et al. ('06), Trafalis & Gilbert ('06), etc. applied robust optimization to handle uncertainty in observations.



$$\boldsymbol{x}_i^o \to \boldsymbol{x}_i^o + \Delta \boldsymbol{x}_i$$

$$\Delta \boldsymbol{x}_i \in \mathcal{U}_i := \{\Delta \boldsymbol{x}_i : \|\Delta \boldsymbol{x}_i\| \leq \delta_i\}$$

Instead of the deterministic constraint:

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i^o + b) \geq 1 - z_i$$

Robust C-SVM model

$$\begin{array}{ll} \displaystyle\min_{\boldsymbol{w},b,\boldsymbol{z}} & \displaystyle\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} z_i \\ \text{s.t.} & \displaystyle\min_{\Delta \boldsymbol{x}_i \in \mathcal{U}_i} y_i(\boldsymbol{w}^\top(\boldsymbol{x}_i^o + \Delta \boldsymbol{x}_i) + b) \geq 1 - z_i, \\ & z_i \geq 0, \quad i = 1, \ldots, m \end{array}$$

$\longrightarrow$ **Second-order cone program**

Regularization = Robustness

Xu-Caramanis-Mannor ['09]

Regularization penalty

$$\min_{\boldsymbol{w},b,\boldsymbol{z}} \quad \delta\|\boldsymbol{w}\| + \sum_{i=1}^{m} z_i$$
$$\text{s.t.} \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i^o + b) \ge 1 - z_i,$$
$$z_i \ge 0, \quad i = 1, \dots, m$$

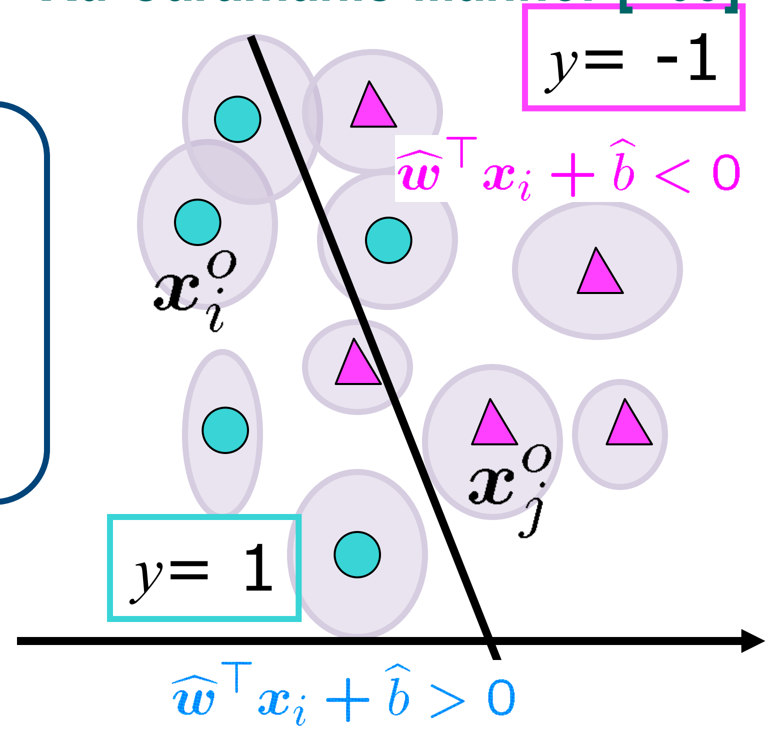
Equivalent

Remove "regularization"

$$\min_{\boldsymbol{w},b} \sum_{i=1}^{m} [1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i^o + b)]^+$$

Consider "robustness"

$$\min_{\boldsymbol{w},b} \max_{(\Delta\boldsymbol{x}_1,\dots,\Delta\boldsymbol{x}_m)\in\mathcal{U}} \sum_{i=1}^{m} [1 - y_i\{\boldsymbol{w}^\top(\boldsymbol{x}_i^o + \Delta\boldsymbol{x}_i) + b\}]^+$$
$$\mathcal{U} = \{(\Delta\boldsymbol{x}_1, \dots, \Delta\boldsymbol{x}_m) : \sum_{i=1}^{m} \|\Delta\boldsymbol{x}_i\| \le \delta\}$$



Robust Classification Model (RCM)

Takeda-Mitsugi-Kanamori ['12]

Max-min form. finds a robust solution with the best worst-case performance.

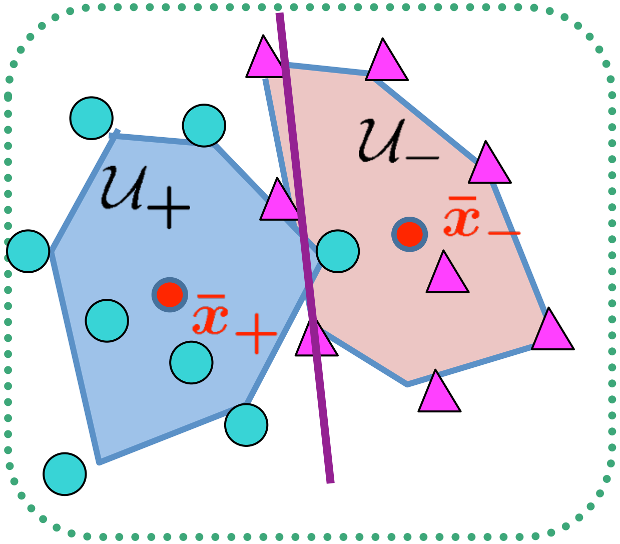
RCM: $$\max_{\|\boldsymbol{w}\|=1} \min_{\boldsymbol{x}_+ \in \mathcal{U}_+, \boldsymbol{x}_- \in \mathcal{U}_-} (\boldsymbol{x}_+ - \boldsymbol{x}_-)^\top \boldsymbol{w}$$

Uncertain Inputs

- ✓ $\boldsymbol{x}_+, \boldsymbol{x}_-$: representative points (or means) of each class.
- ✓ $\mathcal{U}_+$ (resp. $\mathcal{U}_-$) : set of possible points $\boldsymbol{x}_+$ (resp. $\boldsymbol{x}_-$) for each class, called uncertainty set.
- ✓ $\boldsymbol{w}$ is optimized under the worst-case vectors $\boldsymbol{x}_+^*, \boldsymbol{x}_-^*$.
- ✓ b is determined by using $\boldsymbol{x}_+^*$ and $\boldsymbol{x}_-^*$; e.g., so as to go though in the middle of $\boldsymbol{x}_+^*$ and $\boldsymbol{x}_-^*$.

Examples of Uncertainty Sets

$\mathcal{U}_+$ and $\mathcal{U}_-$ are defined with training samples in each class.



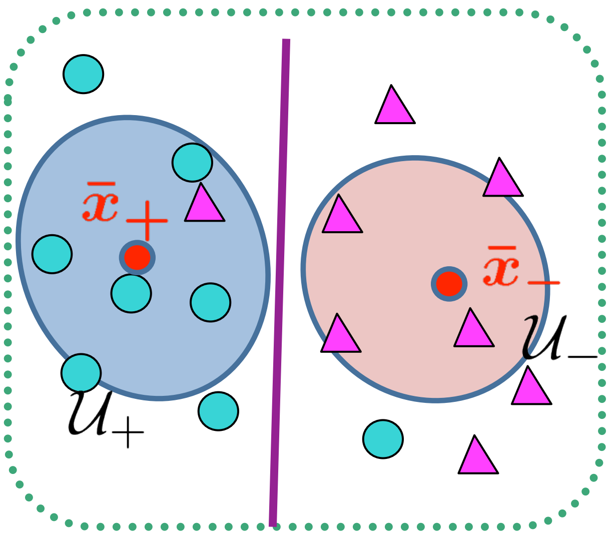
Reduced convex hull (RCH) with param. κ :

$$\kappa \in \left[\frac{1}{m_+}, 1\right]$$

$$\mathcal{U}_+ = \left\{ \sum_{i \in M_+} \lambda_i \boldsymbol{x}_i \ : \ \begin{array}{l} \boldsymbol{e}^\top \boldsymbol{\lambda} = 1, \\ \boldsymbol{0} \le \boldsymbol{\lambda} \le \kappa \boldsymbol{e} \end{array} \right\}$$

a set of discrete distributions

M_+ : index set of samples with label +1



Ellipsoid with param. κ :

$$\mathcal{U}_+ = \left\{ \bar{\boldsymbol{x}}_+ + \Sigma_+^{1/2} \boldsymbol{u} \ : \ \|\boldsymbol{u}\| \le \kappa \right\}$$

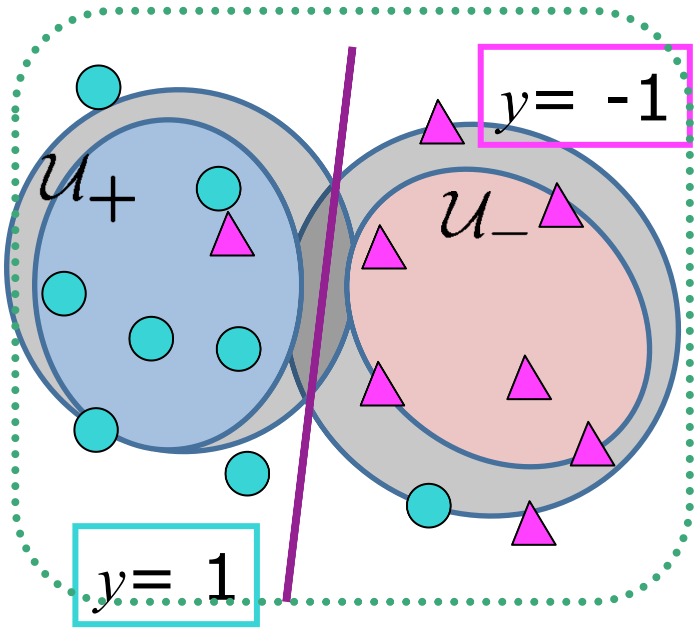
using sample mean : $\bar{\boldsymbol{x}}_+, \ \bar{\boldsymbol{x}}_-$

sample covariance : Σ_+, Σ_-

of samples in each class.

Intersecting or Non-intersecting Uncertainty Set

RCM: $$\max_{\|\boldsymbol{w}\|=1} \min_{\boldsymbol{x}_+\in\mathcal{U}_+,\boldsymbol{x}_-\in\mathcal{U}_-} (\boldsymbol{x}_+ - \boldsymbol{x}_-)^\top \boldsymbol{w}$$



Two uncertainty sets do not intersect.

- $\|\boldsymbol{w}\| = 1$ is replaced by $\|\boldsymbol{w}\| \leq 1$.
- $$\min_{\boldsymbol{x}_+\in\mathcal{U}_+,\boldsymbol{x}_-\in\mathcal{U}_-} \|\boldsymbol{x}_+ - \boldsymbol{x}_-\|$$

  Optimal solution: $\boldsymbol{w} = \boldsymbol{x}_+^* - \boldsymbol{x}_-^*$

→ RCM is a convex problem.

Two uncertainty sets intersect.

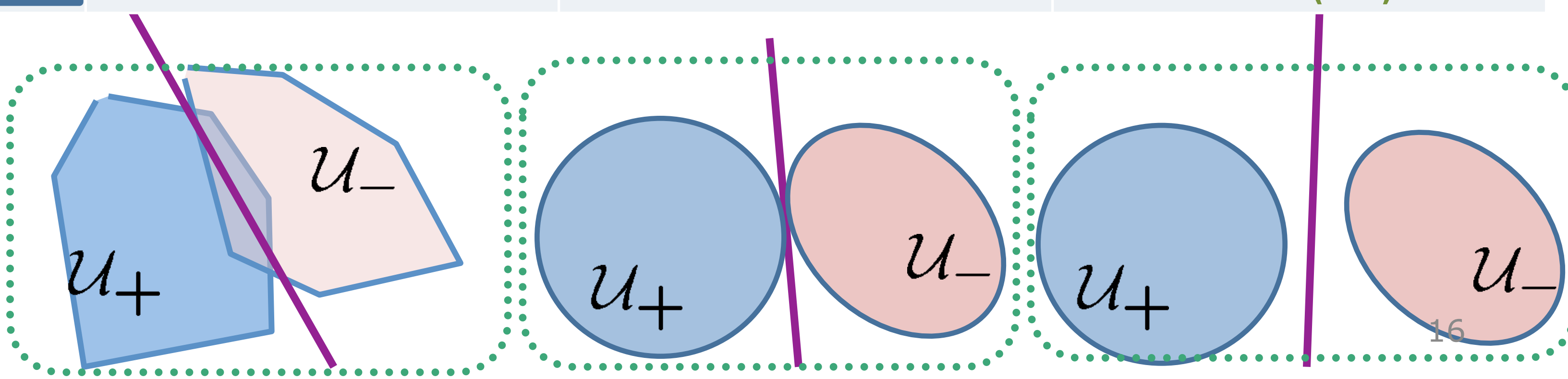
- $\|\boldsymbol{w}\| = 1$ is replaced by $\|\boldsymbol{w}\| \geq 1$.

→ RCM is a non-convex problem.

RCMs with specific sets $\mathcal{U}_\pm$ are reduced to well-known models.

Correspondence to Existing Classifiers

Uncertainty sets	Intersecting	They touch externally	Non-intersecting
Ellipsoid 1 :	No corresponding model	Minimax Probability Machine (MPM) Lanckriet et al. ('02)	Minimum Margin-MPM Nath & Bhattacharyya ('07)
Ellipsoid 2 :	No corresponding model	Fisher Discriminant Analysis (FDA) Fukunaga ('90)	Sparse Feature Selection Bhattacharyya ('04)
Reduced convex hull :	Eν-SVM Perez-Cruz et al. ('03)	$\nu_{\min}$ Crisp & Burges ('00)	ν-SVM (= C-SVM) Scholkopf et al. ('00)
Convex hull : $\nu \to \infty$	----	----	Hard Margin SVM Boser et al. ('92)



What Can We Achieve from Robust-Opt View?

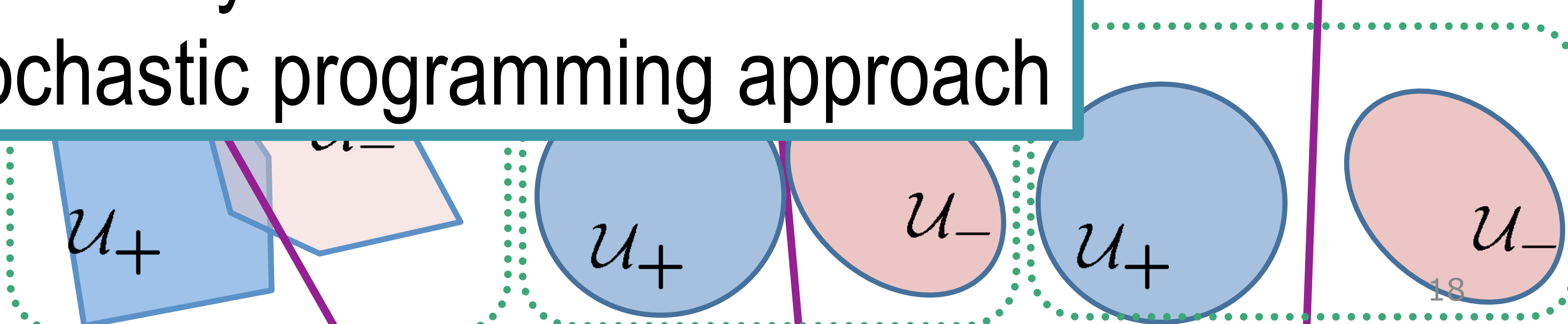
We could give an unified interpretation as robust optimization for some existing classification models.

- ✓ Main difference of those models is **in the definition of their uncertainty sets** for the mean of each class.
- ✓ New models can be available by defining new uncertainty sets.
- ✓ The parameter range can be extended so that the intersection of two sets are allowed.
- ✓ **Unified solution method based on APG** is applicable to convex models (nonintersecting cases).

Correspondence to Existing Classifiers

Uncertainty sets	Intersecting	They touch externally	Non-intersecting
Ellipsoid 1 :	No corresponding model	Minimax Probability Machine (MPM) Lanckriet et al. ('02)	Minimum Margin-MPM Nath & Bhattacharyya ('07)
Ellipsoid 2 :	No corresponding model	Fisher Discriminant Analysis (FDA) Fukunaga ('90)	Sparse Feature Selection Bhattacharyya ('04)
Reduced convex hull :	Eν-SVM Perez-Cruz et al. ('03)	$\nu_{\min}$ Crisp & Burges ('00)	ν-SVM (= C-SVM) Scholkopf et al. ('00)
Convex hull :			Hard Margin SVM Boser et al. ('92)

Analyze these models
by stochastic programming approach



Contents

- Provide a view based on Robust Optimization for various Binary Classification Models including
  - ✓Support Vector Machine (SVM), Minimax Probability Machine (MPM) and Fisher Discriminant Analysis (FDA), etc.

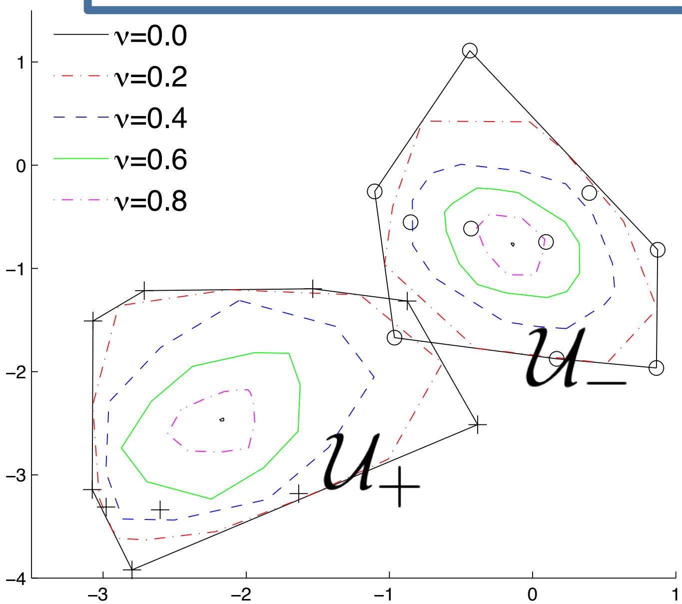
- Provide a view based on Stochastic Programming
  - ✓ν-SVM & Eν-SVM → Generalization Bound
  - ✓Minimum Margin MPM

ν-SVM & Eν-SVM (dual form.)

Robust Classification Model

$$\max_{\|\boldsymbol{w}\|=1} \min_{\boldsymbol{x}_+\in\mathcal{U}_+,\boldsymbol{x}_-\in\mathcal{U}_-} (\boldsymbol{x}_+ - \boldsymbol{x}_-)^\top \boldsymbol{w}$$

- If two RCHs do not intersect (with large ν) → **ν-SVM**
- If two RCHs intersect (with small ν) → **Eν-SVM**



$$\nu \in (0, 2\frac{\min(m_+, m_-)}{m}]$$

Reduced convex hull (RCH) with param. ν :

$$\mathcal{U}_+ = \left\{ \sum_{i\in M_+} \lambda_i \boldsymbol{x}_i : \begin{array}{l} \boldsymbol{e}^\top \boldsymbol{\lambda} = 1, \\ \boldsymbol{0} \le \boldsymbol{\lambda} \le \frac{2}{\nu m}\boldsymbol{e} \end{array} \right\}$$

κ

Shrunk polytopes toward the centers by increasing ν.

ν-SVM & Eν-SVM (primal form.)

Robust Classification Model

$$\max_{\|\boldsymbol{w}\|=1} \min_{\boldsymbol{x}_+ \in \mathcal{U}_+, \boldsymbol{x}_- \in \mathcal{U}_-} (\boldsymbol{x}_+ - \boldsymbol{x}_-)^\top \boldsymbol{w}$$

$\|\boldsymbol{w}\| \ge 1$ — Two RCHs intersect. — Nonconvex Program

$\|\boldsymbol{w}\| \le 1$ — Two RCHs do not intersect. — Convex Program

$\nu = 0$ $\qquad = 2\frac{\min(m_+, m_-)}{m}$

CVaR Minimization??

(Eν-SVM) Perez-Cruz, Weston, Hermann & Schoelkopf ('03)

$$\begin{aligned} \min_{\boldsymbol{w}, b, \boldsymbol{z}, \rho} \quad & -\nu\rho + \frac{1}{m}\sum_{i=1}^{m} z_i \\ \text{s.t.} \quad & z_i + y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - \rho \ge 0, \\ & i = 1, \ldots, m, \\ & \boldsymbol{z} \ge \boldsymbol{0}, \quad \boldsymbol{w}^\top \boldsymbol{w} = 1 \end{aligned}$$

(ν-SVM) Schoelkopf, Smola, Williamson & Bartlett ('00)

$$\begin{aligned} \min_{\boldsymbol{w}, b, \boldsymbol{z}, \rho} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 - \nu\rho + \frac{1}{m}\sum_{i=1}^{m} z_i \\ \text{s.t.} \quad & z_i + y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - \rho \ge 0, \\ & i = 1, \ldots, m, \\ & \boldsymbol{z} \ge \boldsymbol{0}, \quad \cancel{\rho \ge 0} \end{aligned}$$

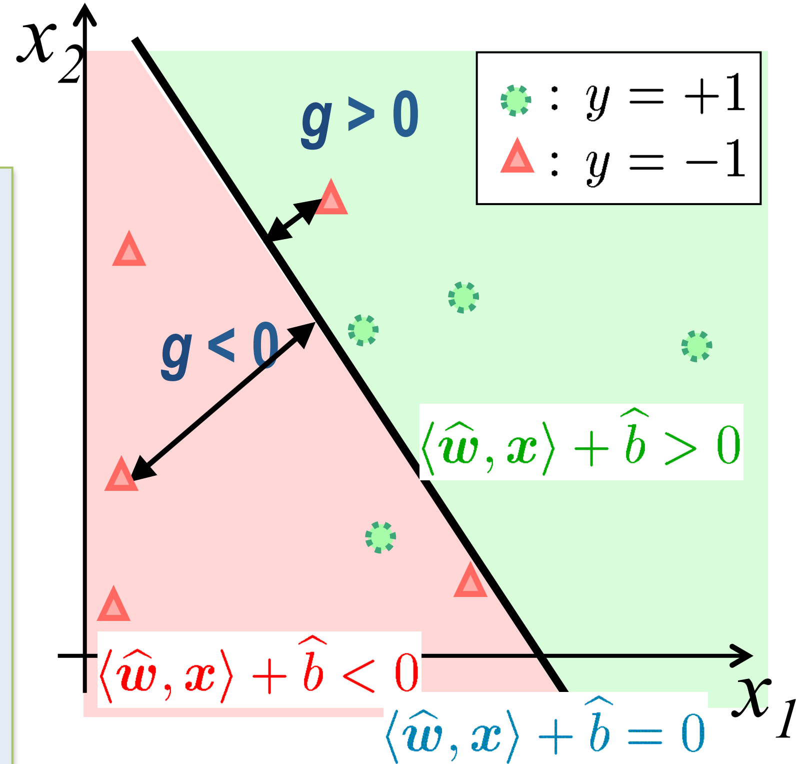
CVaR of Distance

For a hyperplane:

$$\boldsymbol{w}^\top \boldsymbol{x} + b = 0$$

compute the **signed distance (score)** from a point $\boldsymbol{x}_i$ to the hyperplane for all training samples by

$$g(\boldsymbol{w}, b; \boldsymbol{x}_i, y_i) = -\frac{y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)}{\|\boldsymbol{w}\|}$$



$g < 0$ correctly classified, $g > 0$ misclassified

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) > 0$$

Minimize CVaR $\phi_\beta(\boldsymbol{w}, b)$ with $\beta = 1 - \nu$ using $g(\boldsymbol{w}, b; \boldsymbol{x}_i, y_i)$, $i = 1, \ldots, m$

hyperplane of **(E)ν-SVM**

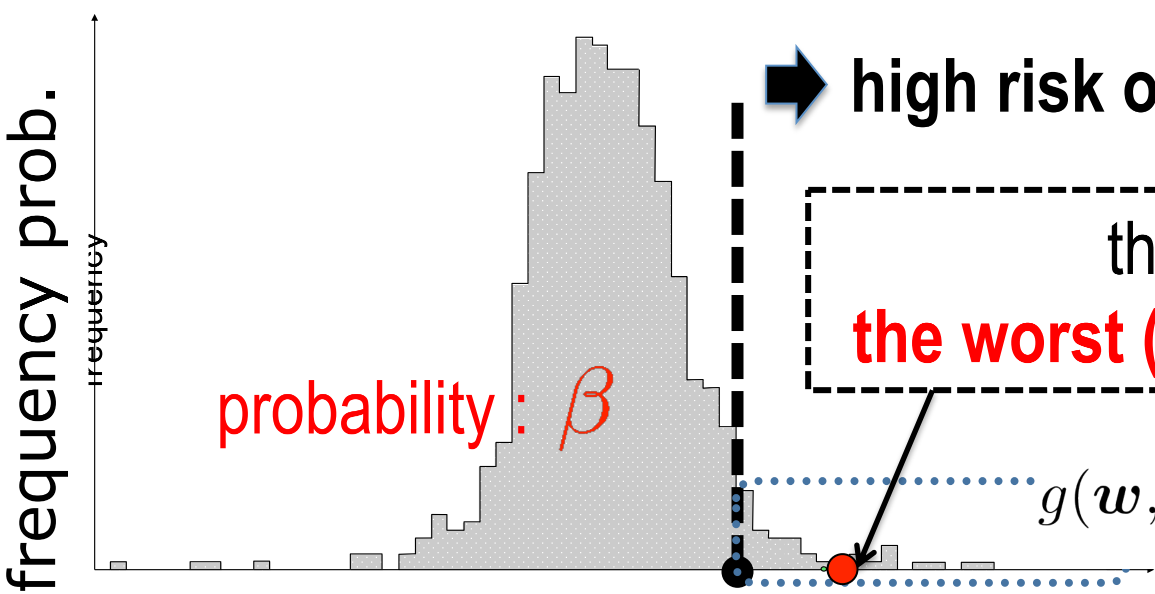
CVaR Minimization for Classification

✓Minimize CVaR $\phi_\beta(\boldsymbol{w}, b)$ with $\beta = 1 - \nu$

and $\Pr((\boldsymbol{x}, y) = (\boldsymbol{x}_i, y_i)) = \dfrac{1}{m}$

by

$$\min_{\boldsymbol{x}, b, \alpha} \alpha + \frac{1}{m\nu} \sum_{i=1}^{m} [g(\boldsymbol{w}, b; \boldsymbol{x}_i, y_i) - \alpha]^+$$



➡ **high risk of misclassification**

the mean of
the worst (1-β)×100% scores

$$g(\boldsymbol{w}, b; \boldsymbol{x}_i, y_i) = -\frac{y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)}{\|\boldsymbol{w}\|}$$

β –quantile: $\alpha_\beta(\boldsymbol{w}, b)$

CVaR: $\phi_\beta(\boldsymbol{w}, b)$

➔ **Minimize**

New interpretation for Eν-SVC

$$\min_{\boldsymbol{x},b,\alpha} \alpha + \frac{1}{m\nu}\sum_{i=1}^{m}\left[-\frac{y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)}{\|\boldsymbol{w}\|} - \alpha\right]^{+}$$

If $\phi_{1-\nu} > 0$

variable: $\rho = -\alpha$

If $\phi_{1-\nu} \leqq 0$

(Eν-SVM) Perez-Cruz, Weston, Hermann & Schoelkopf ('03)

$$\begin{aligned}
\min_{\boldsymbol{w},b,\boldsymbol{z},\rho} \quad & -\nu\rho + \frac{1}{m}\sum_{i=1}^{m} z_i \\
\text{s.t.} \quad & z_i + y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - \rho \ge 0, \\
& i = 1, \ldots, m, \\
& \boldsymbol{z} \ge \boldsymbol{0}, \quad \boldsymbol{w}^\top \boldsymbol{w} = 1
\end{aligned}$$

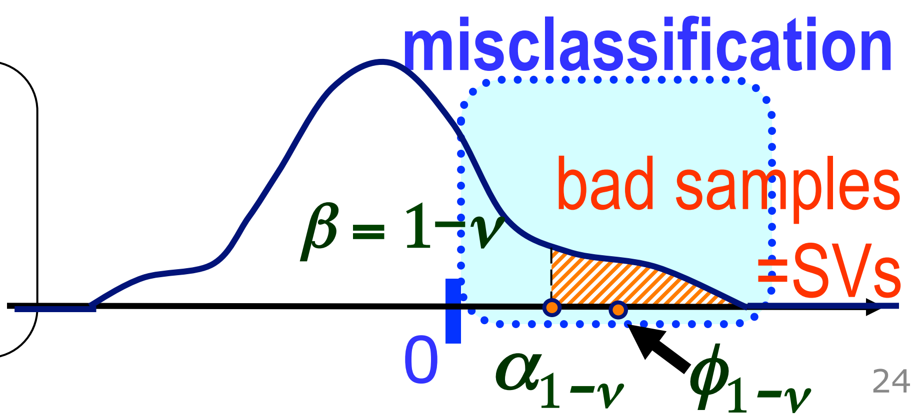
(ν-SVM) Schoelkopf, Smola, Williamson & Bartlett ('00)

$$\begin{aligned}
\min_{\boldsymbol{w},b,\boldsymbol{z},\rho} \quad & \cancel{\frac{1}{2}\|\boldsymbol{w}\|^2} - \nu\rho + \frac{1}{m}\sum_{i=1}^{m} z_i \\
\text{s.t.} \quad & z_i + y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - \rho \ge 0 \\
& i = 1, \ldots, m, \\
& \boldsymbol{z} \ge \boldsymbol{0}, \quad \cancel{\rho \ge 0} \quad \boldsymbol{w}^\top \boldsymbol{w} \le 1
\end{aligned}$$

$$\rho^* = -\alpha^* \approx -\alpha_{1-\nu}$$

: margin of Eν-SVC

negative margin ↔ $\alpha_{1-\nu} > 0$



Three Cases depending on ν

If $\phi_{1-\nu} > 0$

(Eν-SVM) Perez-Cruz, Weston, Hermann & Schoelkopf ('03)

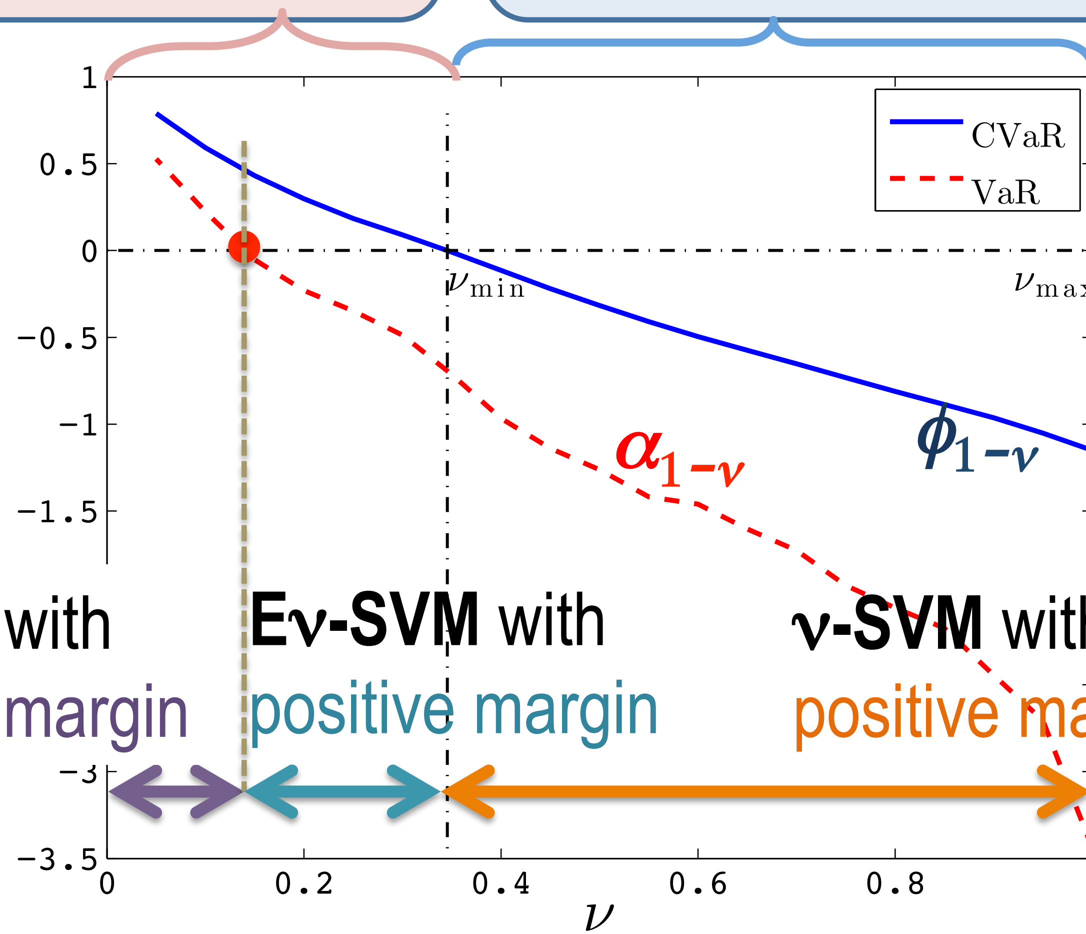
Nonconvex Problem

If $\phi_{1-\nu} \leqq 0$

(ν-SVM) Schoelkopf, Smola, Williamson & Bartlett ('00)

Convex Problem

Margin:
$\rho^* = -\alpha_{1-\nu}$



Eν-SVM with negative margin

Eν-SVM with positive margin

ν-SVM with positive margin

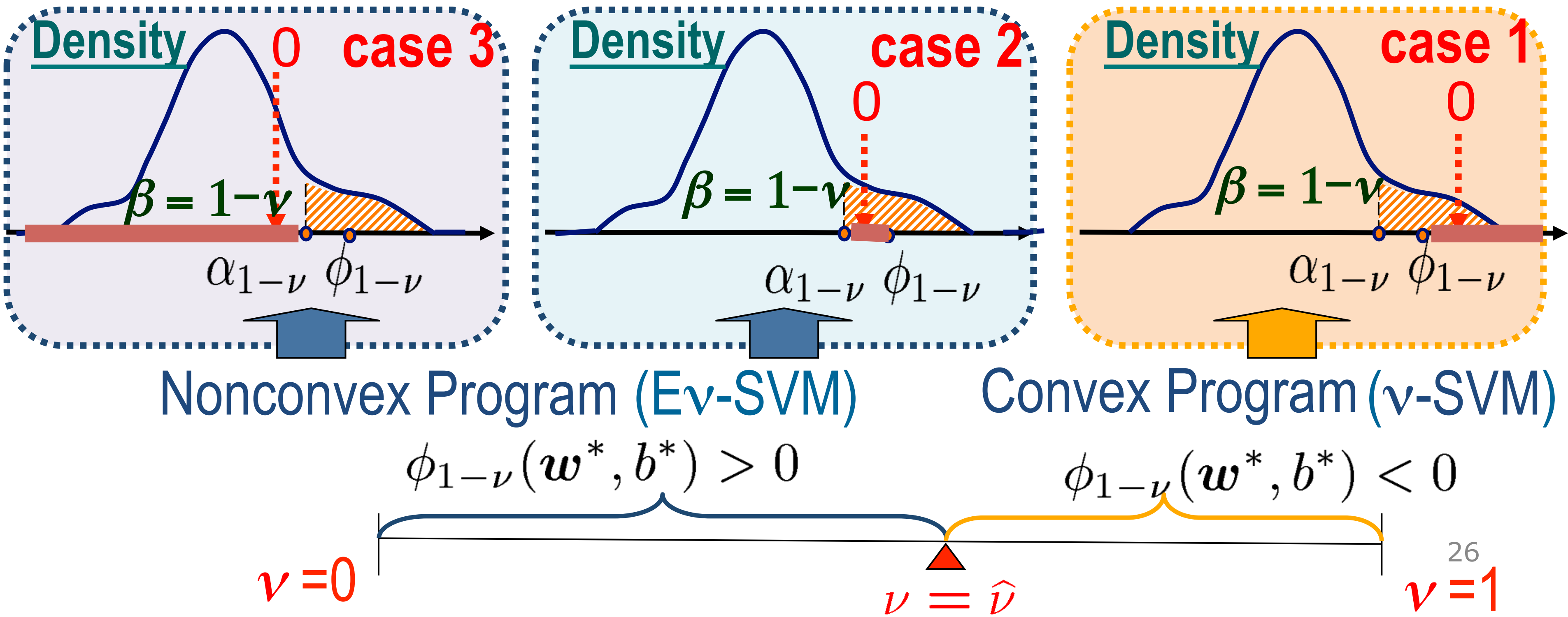
Generalization Error Bounds

New generalization error bounds of Eν-SVM include the CVaR risk measure

error rates for test (new) samples

➔ Minimizing the CVaR lowers the bound

➔ It justifies the use of Eν-SVM & ν-SVM

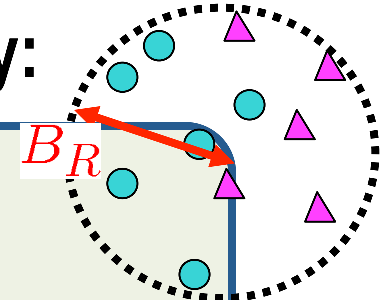


Generalization Error Bound (case 1)

Theorem : (case 1)

Takeda-Sugiyama ['08]

For a feasible sol. $(\boldsymbol{w}, b)$ **of (ν-SVM), the inequality:**



(generalization error with $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$)

$$\leqq \ \nu + G(\boxed{\alpha_{1-\nu}(\boldsymbol{w}, b)}) \ \leqq \ \nu + G(\boxed{\phi_{1-\nu}(\boldsymbol{w}, b)})$$

< 0

$$G(\gamma) := \sqrt{\frac{2}{m}\left(\frac{4c^2(1+B_R^2)^2}{\gamma^2}\log_2(2m) - 1 + \log\left(\frac{2}{\delta}\right)\right)}$$

$G(\rho^*)$ is used for **(ν-SVM)** in **Schoelkopf, Smola, Williamson & Bartlett ('00)**

holds with probability at least $1 - \delta$

➔ CVaR min. gives an opt. solution which minimizes the bound.

➔ ν-SVM is reasonable.

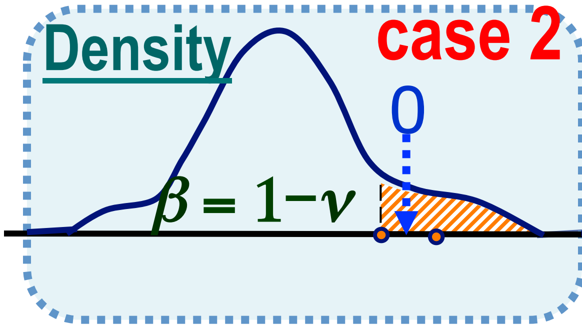
Generalization Error Bound (cases 2&3)

For a feasible sol. $(\boldsymbol{w}, b)$ of (Eν-SVM)

(generalization error with $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$)

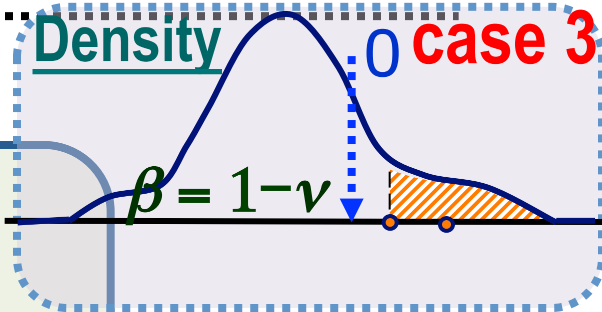
$$\leqq \quad \nu + G(\boxed{\alpha_{1-\nu}(\boldsymbol{w}, b)})$$

holds with probability at least $1 - \delta$



$$G(\gamma) := \sqrt{\frac{2}{m}\left(\frac{4c^2(1+B_R^2)^2}{\gamma^2}\log_2(2m) - 1 + \log\left(\frac{2}{\delta}\right)\right)}$$

For a feasible sol. $(\boldsymbol{w}, b)$ of (Eν-SVM)



(generalization error with $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$)

$$\geqq \quad \nu - G(\boxed{\alpha_{1-\nu}(\boldsymbol{w}, b)})$$

This bound is upper -bounded as

$$\nu - G(\alpha_{1-\nu}(\boldsymbol{w}, b)) \leq \nu - G(\boxed{\phi_{1-\nu}(\boldsymbol{w}, b)})$$

(E)ν-SVM (classification method)

CVaR Min.:
$$\min_{\boldsymbol{w},b,\rho} -\rho + \frac{1}{\nu m}\sum_{i\in M}[g(\boldsymbol{w},b;\boldsymbol{x}_i,y_i)+\rho]^+$$

Stochastic Programming

$$g(\boldsymbol{w},b;\boldsymbol{x}_i,y_i) = -\frac{y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)}{\|\boldsymbol{w}\|}$$

(E)ν-SVM:
$$\begin{array}{lll}\min_{\boldsymbol{w},b,\boldsymbol{z},\rho} & -\nu\rho + \frac{1}{m}\sum_{i=1}^{m} z_i & \\ \text{s.t.} & z_i + y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)-\rho \geq 0, & i\in M, \\ & \boldsymbol{z}\geq \boldsymbol{0},\ \ \boldsymbol{w}^\top\boldsymbol{w}=1 & \text{(or } \boldsymbol{w}^\top\boldsymbol{w}\leq 1\text{)}\end{array}$$

Perez-Cruz, Weston, Hermann & Schoelkopf ('03)

Robust Optimization

by taking dual w.r.t. $b, \boldsymbol{z}, \rho$

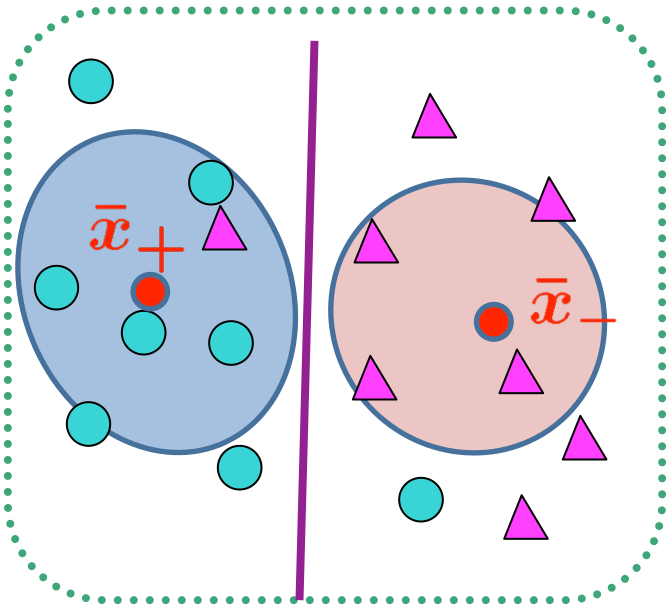
RCM:
$$\max_{\|\boldsymbol{w}\|=1}\ \min_{\boldsymbol{x}_+\in\mathcal{U}_+,\boldsymbol{x}_-\in\mathcal{U}_-} (\boldsymbol{x}_+-\boldsymbol{x}_-)^\top\boldsymbol{w}$$

$$\mathcal{U}_\pm = \left\{\sum_{i\in M_\pm}\lambda_i\boldsymbol{x}_i : \boldsymbol{e}^\top\boldsymbol{\lambda}=1,\ \boldsymbol{0}\leq\boldsymbol{\lambda}\leq\frac{2}{\nu m}\boldsymbol{e}\right\}$$

Ellipsoidal Uncertainty Sets

Robust Classification Model

$$\max_{\|\boldsymbol{w}\|=1} \min_{\boldsymbol{x}_+\in\mathcal{U}_+,\boldsymbol{x}_-\in\mathcal{U}_-} (\boldsymbol{x}_+ - \boldsymbol{x}_-)^\top \boldsymbol{w}$$



Using sample mean : $\bar{\boldsymbol{x}}_+, \ \bar{\boldsymbol{x}}_-$

sample covariance : Σ_+, Σ_-

of samples in each class, let

$$\mathcal{U}_+ = \left\{ \bar{\boldsymbol{x}}_+ + \Sigma_+^{1/2}\boldsymbol{u} : \|\boldsymbol{u}\| \le \kappa \right\}$$

$$\mathcal{U}_- = \left\{ \bar{\boldsymbol{x}}_- + \Sigma_-^{1/2}\boldsymbol{v} : \|\boldsymbol{v}\| \le \kappa \right\}.$$

$$\min_{\|\boldsymbol{w}\|=1} \kappa\|\Sigma_+^{1/2}\boldsymbol{w}\| + \kappa\|\Sigma_-^{1/2}\boldsymbol{w}\| - (\bar{\boldsymbol{x}}_+ - \bar{\boldsymbol{x}}_-)^\top \boldsymbol{w}$$

$\|\boldsymbol{w}\| = 1$ can be replaced by $\|\boldsymbol{w}\| \le 1$ when $\mathcal{U}_+ \cap \mathcal{U}_- = \emptyset$.

Equivalence to Maximum-Margin MPM

Robust Classification Model (non-intersecting case)

$$\min_{\|\boldsymbol{w}\|\leq 1} \; \kappa\|\Sigma_+^{1/2}\boldsymbol{w}\| + \kappa\|\Sigma_-^{1/2}\boldsymbol{w}\| - (\bar{\boldsymbol{x}}_+ - \bar{\boldsymbol{x}}_-)^\top \boldsymbol{w}$$

$$\kappa = \sqrt{\frac{1-\eta}{\eta}}$$

Maximum-Margin MPM

Nath & Bhattacharyya ('07)

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$

Worst-case misclassified probabilities

$$\text{s.t.} \quad \sup_{\boldsymbol{x}_+\sim(\bar{\boldsymbol{x}}_+,\Sigma_+)} \Pr\{\boldsymbol{x}_+^\top \boldsymbol{w} + b < 1\} \leq \eta$$

Using generalized Chebyshev-Cantelli inequality,

$$\bar{\boldsymbol{x}}_+^\top \boldsymbol{w} + b \geq 1 + \sqrt{\frac{1-\eta}{\eta}}\|\Sigma_+^{1/2}\boldsymbol{w}\|$$

$\boldsymbol{x}_+$, $\boldsymbol{x}_-$: random vectors from each of two classes with means and covariance matrices given by $(\bar{\boldsymbol{x}}_+, \Sigma_+)$ and $(\bar{\boldsymbol{x}}_-, \Sigma_-)$.

Stochastic Problem under Normal Distribution

Robust Classification Model with $\mathcal{U}_{\pm} = \{\bar{\boldsymbol{x}}_{\pm} + \Sigma_{\pm}^{1/2}\boldsymbol{u} : \|\boldsymbol{u}\| \le \kappa\}$

$$\max_{\|\boldsymbol{w}\| \le 1} \min_{\boldsymbol{x}_+ \in \mathcal{U}_+, \boldsymbol{x}_- \in \mathcal{U}_-} (\boldsymbol{x}_+ - \boldsymbol{x}_-)^\top \boldsymbol{w}$$

$$\kappa = \sqrt{\frac{1-\eta}{\eta}}$$

$$\min_{\boldsymbol{w}, b} \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$\text{s.t.} \quad \sup_{\boldsymbol{x}_+ \sim (\bar{\boldsymbol{x}}_+, \Sigma_+)} \Pr\{\boldsymbol{x}_+^\top \boldsymbol{w} + b < 1\} \le \eta$$

$$\sup_{\boldsymbol{x}_- \sim (\bar{\boldsymbol{x}}_-, \Sigma_-)} \Pr\{\boldsymbol{x}_-^\top \boldsymbol{w} + b > -1\} \le \eta$$

$$\bar{\boldsymbol{x}}_+^\top \boldsymbol{w} + b \ge 1 + \sqrt{\frac{1-\eta}{\eta}} \|\Sigma_+^{1/2}\boldsymbol{w}\|$$

$$\kappa = \Phi^{-1}(1-\eta)$$

The worst-case prob. distribution is considered in **Nath & Bhattacharyya ('07)**

Under the assump: $\boldsymbol{x}_+ \sim \mathcal{N}_{m_+}(\bar{\boldsymbol{x}}_+, \Sigma_+)$

$$\Pr\{\boldsymbol{x}_+^\top \boldsymbol{w} + b < 0\} \le \eta$$

$$\Longleftrightarrow$$

$$\bar{\boldsymbol{x}}_+^\top \boldsymbol{w} + b \ge 1 + \Phi^{-1}(1-\eta) \|\Sigma_+^{1/2}\boldsymbol{w}\|$$

$\Phi(z)$: cumulative dist. func. (cdf) of $\mathcal{N}(0,1)$

32

Conclusions

- We provided new views based on Robust Optimization / Stochastic Programming for existing machine learning classification models (SVM, MPM, FDA and their variants).

- We could evaluate generalization bounds from the viewpoint of SP and propose an efficient algorithm from the viewpoint of RO.

Summary

- The first textbook on Robust Optimization appears in 2009.

  Ben-Tal, El Ghaoui & Nemirovski ['09]

- Robust optimization techniques are used in various research areas.

  ✓The preface of the book briefly mentions the relation to Robust Control （H_∞ Control), Robust Statistics, Machine learning（SVM）, etc.

- Recently, studies on robust optimization using "probability" are increased. The robust optimization research is still developing.

参考文献 -1-

- E. M. L. Beale. On minimizing a convex function subject to linear inequalities. J. Roy. Statist. Soc. Ser. B. 17 (1955), 173–184.
- A. Ben-Tal, L. El Ghaoui and A. Nemirovski. Robust optimization. Princeton University Press, 2009.
- A. Ben-Tal, A. Goryashko, E. Guslitzer and A. Nemirovski. Adjustable robust solutions of uncertain linear programs. *Math. Progr.* 99 (2004), 351-376.
- A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Math. of Oper. Res.* 23:4 (1998), 769-805.
- A. Ben-Tal and A. Nemirovski. Robust solutions of uncertain linear programs. *OR Letters* 25 (1999), 1-13.
- Ben-Tal and A. Nemirovski. Robust solutions of linear programming problems contaminated with uncertain data. *Math. Progr.* 88 (2000), 411-424.
- B. E. Boser, I. M. Guyon and V. N. Vapnik. A training algorithm for optimal margin classiers. COLT (pp. 144-152). ACM Press, 1992.
- G. Calafiore and M.C. Campi. Uncertain convex programs: Randomized solutions and confidence levels. *Math. Progr. 102:1 (2005), 25–46.*
- G. Calafiore and M.C. Campi. The scenario approach to robust control design. IEEE Transactions on Automatic Control, 51:5 (2006), 742–753.
- T.C.Y. Chan, T. Bortfeld and J.N. Tsitsiklis. A robust approach to IMRT optimization. Phys. Med. Biol. 51 (2006), 2567–2583.

参考文献 -2-

- A. Charnes and W. W. Cooper. Uncertain convex programs: Randomize solutions and confidence level. *Management Sci.* 6 (1959), 73–79.
- C. Cortes and V. Vapnik. Support-vector networks. Machine Learning, 20 (1995), 273-297.
- G. B. Dantzig. Linear programming under uncertainty. Management Sci., 1 (1955), 197-206.
- L. El Ghaoui and H. Lebret. Robust solution to least-squares problems with uncertain data. *SIAM J. of Matrix Anal. Appl.* 18 (1997), 1035-1064.
- D. Goldfarb and G. Iyengar. Robust convex quadratically constrained programs. *Math. Progr.* 97 (2003) , 495-515.
- J. Gotoh and A. Takeda. A linear Classification Model Based on Conditional Geometric Score. Pacific Journal of Optimization 1 (2005), 277-296.
- P. Kouvelis and G. Yu. Robust discrete optimization and its applications. Kluwer Academic, Dordrecht (1997).
- X. Lin, S. L. Janak and C. A. Floudas. A new robust optimization approach for scheduling under uncertainty: I. Bounded uncertainty. Computers and Chemical Engineering 28 (2004), 1069–1085.
- F. Perez-Cruz , J. Weston, D.J.L. Hermann, B. Schölkopf. Extension of the *ν*-SVM range for classification. Advances in Learning Theory: Methods, Models and Applications 190 (2003), 179–196.

参考文献 -3-

- R.T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. J. Bank. Finance 26 (2002) , 1443–1471.
- B. Schoelkopf, A. J. Smola, R. C. Williamson and P. L. Bartlett. New support vector algorithms. Neural Computation 12 (2000), 1207–1245.
- A.L. Soyster. Convex programming with set-inclusive constraints and applications to inexact linear programming. *Oper. Res. 21* (1973), 1154-1157.
- A. Takeda and M. Sugiyama. *v*-support vector machine as conditional value-at-risk minimization. ICML 2008, 2008.
- A. Takeda, H. Mitsugi and T. Kanamori. A unified robust classification model. ICML2012, 2012.
- H. Xu, C. Caramanis and S. Mannor. Robustness and regularization of support vector machines. Journal of Machine Learning Research, 10 (2009), 1485-1510.
- G. Yu and J. Yang. On the robust shortest path problem. Comput. Oper. Res. 25 (1998), 457–468.

- 寒野 善博, ロバスト性を考慮した設計. 2008年度日本建築学会大会(中国)・構造部門(応用力学)パネルディスカッション資料『建築構造設計における冗長性と頑強性の役割—リダンダンシーとロバスト性とは—』, 14-23.
- 土谷隆、笹川卓， 二次錐計画問題による磁気シールドのロバスト最適化、統計数理53：2（2005）, 297-315.