ISSN 1342-2804

Research Reports on Mathematical and Computing Sciences

A Robust Approach Based on Conditional Value-at-Risk Measure to Statistical Learning Problems

Akiko Takeda and Takafumi Kanamori

August 2005, B–417

Department of Mathematical and Computing Sciences
Tokyo Institute of Technology

SERIES B: Operations Research

B-417 A Robust Approach Based on Conditional Value-at-Risk Measure to Statistical Learning Problems

Akiko Takeda† and Takafumi Kanamori‡
August 2005 (Revised January 2006)

Abstract. Robust optimization is one of typical approaches to optimize a system with incomplete information and considerable uncertainty. The standard robust optimization problem minimizes maximum cost by focusing on the considerable worst case. In some application field, it is certainly important to consider the worst case among all considerable cases, but this min-max criterion tends to lead an overly conservative decision.

In this paper, we regard statistical learning problems as uncertain problems, and introduce a risk measure known as the conditional value-at-risk (CVaR) in order to dissolve overly conservativeness of robust optimization and depresses influence of outliers or measurement error which may be included in assumed uncertainty set. Monte Carlo sampling is applied to obtain an optimal solution of CVaR robust problem approximately, and convergence property of the solution is proved by using Vapnik and Chervonenkis theory. We point out that in the context of machine learning, CVaR robust problem is identical to ν-support vector classification or ν-support vector regression with apt uncertainty, and show that proposed approach is useful to deal with measurement errors in observations.

Key words.
Robust Optimization, Uncertainty, Conditional Value-at-Risk, Support Vector Machine, Monte Carlo Sampling.

† Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, 2-12-1 Oh-Okayama, Meguro-ku, Tokyo 152-8552 Japan. *takeda@is.titech.ac.jp*

‡ Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, 2-12-1 Oh-Okayama, Meguro-ku, Tokyo 152-8552 Japan. *kanamori@is.titech.ac.jp*

1 Introduction

Uncertainty is an inevitable feature of decision-making environments, and many researches have been developed for optimization under uncertainty. There are two typical approaches to optimize a system with incomplete information and considerable uncertainty: stochastic programming (see, *e.g.* [14]) and robust optimization (see [3, 4, 9, 10, 16]). For an uncertain optimization problem whose objective function includes uncertain data, standard stochastic programming models assume that probability distributions governing the data are known or can be estimated, and maximize the expectation of cost function including random variables. On the other hand, robust optimization approach assumes that the uncertain data are known only within certain bounds, which is called uncertainty set $\mathcal{U}$, and minimizes maximum cost by focusing on the considerable worst case in $\mathcal{U}$.

In some application fields such as typical technological process in chemical industry and topology design of truss structures, it is important to consider the worst case among all considerable ones, but the min-max criterion tends to lead an overly conservative solution. In fact, tradeoff between the uncertainty set size and the level of conservativeness has been thoroughly explored both empirically and theoretically. Ben-Tal & Nemirovski [2, 3, 4] and El-Ghaoui & Lebret [9] proposed less conservative robust models by considering uncertain problems with ellipsoidal uncertainties. Also, Bertsimas & Sim [6] proposed to adjust the level of conservatism of the robust solutions in terms of probabilistic bounds of constraint violations.

In this paper, we regard statistical learning problems as uncertain problems, and utilize a popular risk measure in financial risk management, known as the conditional value-at-risk (CVaR) introduced by [13], for addressing overly conservativeness of robust optimization. The risk measure CVaR indicates the β-tail expectation of the cost function. Minimizing CVaR regards worst $100 \times (1-\beta)\%$ cases in all considerable ones as the worst class, and provides an optimal solution which minimizes the conditional expectation of costs in the worst class. In other words, for any decision, we consider not only the worst case but the worst class including the worst case. The corresponding CVaR robust problem has the following properties:

(i). CVaR robust problem connects stochastic programming problem, which minimizes the expectation of cost function, and robust optimization problem by one parameter $\beta \in (0, 1)$. By adjusting β nicely, CVaR robust problem becomes sufficiently close to the robust optimization and also, stochastic programming problem. Clearly, adjusting β implicitly changes the level of conservatism of the robust solution, or the risk aversion of the decision maker. When the best decision determined by the robust problem is too conservative, CVaR robust problem with proper β is helpful.

(ii). The optimal decision induced from CVaR robust problem possibly depresses influence of outliers or measurement error, and thus, fit statistical learning problems well compared to a decision of robust optimization. CVaR robust problem may provide an appropriate decision

even if the underlying distribution function or uncertainty set $\mathcal{U}$ includes some error, while the decision of robust optimization entirely depends on the choice of uncertainty set $\mathcal{U}$ and is greatly influenced by such error. The resulting CVaR robust problems induced from statistical learning problems are almost identical to ν-support vector classification (ν-SVC) and ν-support vector regression (ν-SVR) [15].

The problem of minimizing CVaR is a kind of stochastic programming and requires probability distributions governing the uncertain data instead of uncertainty set $\mathcal{U}$. However, we also assume that uncertain data are within uncertainty set $\mathcal{U}$ in order to show the relationship between the proposed problem and robust optimization.

CVaR risk measure has been recently used in both the robust optimization and stochastic optimization communities. For instance, Nemirovski & Shapiro [12] use CVaR to evaluate approximate solutions to chance-constrained problems. Recent work by Bertsimas & Brown [5] utilizes CVaR risk measures as a means of constructing uncertainty sets $\mathcal{U}$. In this paper, we introduce CVaR risk measure to statistical learning problems in order to dissolve overly conservativeness of robust optimization and depresses influence of outliers or measurement error which may be included in assumed uncertainty set. Actually, we show empirically that CVaR risk measure fit statistical learning problems well compared to the min-max criterion.

CVaR robust problem, whose CVaR risk measure is defined in the integral form, is difficult to solve exactly when the uncertainty set $\mathcal{U}$ consists of infinite number of cases, while the CVaR problem can be solved easily when the number of all considerable cases is finite, *i.e.*, $\mathcal{U}$ consists of finite scenarios. To deal with infinite $\mathcal{U}$, we generate samples $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$, that is, N realizations of the random vector $\boldsymbol{u} \in \mathcal{U}$ based on Monte Carlo sampling, and solve CVaR robust problem approximately via empirical CVaR robust problem constructed from samples $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$. In this paper, we discuss how large N is necessary to ensure that the gap between the optimal value of CVaR robust problem and that of its empirical problem becomes sufficiently small with high probability, differently from [8] and [18]. They estimated N for a more general uncertain problem which includes uncertainty in constraints, so that the optimal solution of the sampled problem is feasible to the original problem with high probability.

The rest of this paper is organized as follows. In Section 2 we describe uncertain problems and then, introduce the definition of risk measures CVaR. Section 3 discusses CVaR robust problems for two kinds of uncertainty sets: one is finite uncertainty set $\mathcal{U}$ and the other is infinite uncertainty set $\mathcal{U}$. Section 4 provides the proof of the theorem related to convergence property of empirical CVaR robust problems, presented in the previous section. In Section 5, we apply our CVaR robust approach to statistical learning problems: linear classification and linear regression problems. From numerical results, we see that CVaR robust problem pursues two objectives by minimizing the conditional expectation of cost function while avoiding poor performance for any considerable cases. Finally, we conclude the paper by adding some remarks and a possible extension.

2 Preliminaries

2.1 Uncertain Problem

We consider an uncertain optimization problem whose objective function $f(\boldsymbol{x}, \boldsymbol{u})$ includes uncertain data $\boldsymbol{u}$, where $\boldsymbol{x} = (x_1, \ldots, x_n)$ is a decision variable vector in this problem. Feasible set for $\boldsymbol{x}$ is denoted by X. The coefficient vector $\boldsymbol{u} = (u_1, \ldots, u_n)$ of $f(\boldsymbol{x}, \boldsymbol{u})$ is unknown at this present moment, but we know that $\boldsymbol{u}$ of a given uncertainty set $\mathcal{U}$ will be revealed in the future. The usual robust optimization problem focuses on the worst case $\max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x}, \boldsymbol{u})$ and minimizes maximum cost such as

$$\min_{\boldsymbol{x}\in X} \max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x}, \boldsymbol{u}). \tag{1}$$

Throughout this paper, we assume that the regions X of $\boldsymbol{x}$ and $\mathcal{U}$ of $\boldsymbol{u}$ are bounded. Also, suppose that $f(\boldsymbol{x}, \boldsymbol{u})$ is convex in $\boldsymbol{x}$ and the feasible set X is convex.

When given uncertainty set $\mathcal{U}$ consists of finite elements, robust problem (1) can be solved exactly by some existing convex optimization techniques. For an uncertainty set $\mathcal{U}$ generally defined as some region (for example, a polytope defined by hyperplanes), however, the robust problem (1) becomes difficult to be solved. Therefore, for robust optimization problems, several kinds of uncertainty sets $\mathcal{U}$ are proposed in [4, 3, 10] so that the resulting problems (1) are solvable. Here we consider robust optimization problem whose objective function is convex quadratic such as $f(\boldsymbol{x}, \boldsymbol{u}) := f(\boldsymbol{x}, (\boldsymbol{Q}, \boldsymbol{q}, \gamma)) = \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \boldsymbol{q}^\top \boldsymbol{x} + \gamma$. For uncertain data $(\boldsymbol{Q}, \boldsymbol{q}, \gamma)$, [4] assumed an uncertainty set $\widetilde{\mathcal{U}}_q$ shown below and proposed the robust problem called uncertain QCQP problem, which is reduced to a semidefinite programming. Moreover, other kinds of uncertainty sets such as polytopic uncertainty set $\widetilde{\mathcal{U}}_p$ and norm-constrained uncertainty set $\widetilde{\mathcal{U}}_n$ are discussed in [10]. These uncertainty sets $\widetilde{\mathcal{U}}_p$, $\widetilde{\mathcal{U}}_n$ and $\widetilde{\mathcal{U}}_q$ are defined as follows.

- Polytopic uncertainty set [10] :

$$\widetilde{\mathcal{U}}_p = \left\{ (\boldsymbol{Q}, \boldsymbol{q}, \gamma) : \begin{array}{c} (\boldsymbol{Q}, \boldsymbol{q}, \gamma) = \sum_{j=1}^{\ell} u_j (\boldsymbol{Q}_j, \boldsymbol{q}_j, \gamma_j) \\ \boldsymbol{Q}_j \succeq O, j = 1, \ldots, \ell, \ \ \boldsymbol{u} \in \mathcal{U}_p \end{array} \right\},$$

$$\mathcal{U}_p = \left\{ \boldsymbol{u} : \boldsymbol{u} \geq \boldsymbol{0}, \ \sum_{j=1}^{\ell} u_j = 1 \right\},$$

$$f(\boldsymbol{x}, \boldsymbol{u}) = \textstyle\sum_{j=1}^{\ell} (\boldsymbol{x}^\top \boldsymbol{Q}_j \boldsymbol{x} + \boldsymbol{q}_j^\top \boldsymbol{x} + \gamma_j) u_j, \ \ \boldsymbol{x} \in X, \ \boldsymbol{u} \in \mathcal{U}_p.$$

- Norm-constrained uncertainty set [10] :

$$\widetilde{\mathcal{U}}_n = \left\{ (\boldsymbol{Q}, \boldsymbol{q}, \gamma) : \begin{array}{l} (\boldsymbol{Q}, \boldsymbol{q}, \gamma) = (\boldsymbol{Q}_0, \boldsymbol{q}_0, \gamma_0) + \sum_{j=1}^{\ell} u_j (\boldsymbol{Q}_j, \boldsymbol{q}_j, \gamma_j) \\ \boldsymbol{Q}_j \succeq O, j = 0, 1, \ldots, \ell, \ \ \boldsymbol{u} \in \mathcal{U}_n \end{array} \right\},$$

$$\mathcal{U}_n = \{\boldsymbol{u} : \boldsymbol{u} \geq \boldsymbol{0}, \ \|\boldsymbol{u}\| \leq 1\}$$

$$f(\boldsymbol{x}, \boldsymbol{u}) = \sum_{j=1}^{\ell} (\boldsymbol{x}^\top \boldsymbol{Q}_j \boldsymbol{x} + \boldsymbol{q}_j^\top \boldsymbol{x} + \gamma_j) u_j + (\boldsymbol{x}^\top \boldsymbol{Q}_0 \boldsymbol{x} + \boldsymbol{q}_0^\top \boldsymbol{x} + \gamma_0), \ \ \boldsymbol{x} \in X, \ \boldsymbol{u} \in \mathcal{U}_n.$$

- Quadratic uncertainty set [4] :

$$\widetilde{\mathcal{U}}_q = \left\{ (\boldsymbol{Q}, \boldsymbol{q}, \gamma) : \begin{array}{l} \boldsymbol{Q} = (\boldsymbol{R}_0 + \sum_{j=1}^{\ell} u_j \boldsymbol{R}_j)^\top (\boldsymbol{R}_0 + \sum_{j=1}^{\ell} u_j \boldsymbol{R}_j) \\ (\boldsymbol{q}, \gamma) = (\boldsymbol{q}_0, \gamma_0) + \sum_{j=1}^{\ell} u_j (\boldsymbol{q}_j, \gamma_j) \\ \boldsymbol{u} \in \mathcal{U}_q \end{array} \right\}$$

$$\mathcal{U}_q = \{\boldsymbol{u} : \|\boldsymbol{u}\| \leq 1\}$$

$$f(\boldsymbol{x}, \boldsymbol{u}) = \sum_{i,j=1}^{\ell} (\boldsymbol{x}^\top \boldsymbol{R}_i^\top \boldsymbol{R}_j \boldsymbol{x}) u_i u_j + \sum_{j=1}^{\ell} \left\{ \boldsymbol{x}^\top (\boldsymbol{R}_0^\top \boldsymbol{R}_j + \boldsymbol{R}_j^\top \boldsymbol{R}_0) \boldsymbol{x} + \boldsymbol{q}_j^\top \boldsymbol{x} + \gamma_j \right\} u_j$$
$$+ (\boldsymbol{x}^\top \boldsymbol{R}_0^\top \boldsymbol{R}_0 \boldsymbol{x} + \boldsymbol{q}_0^\top \boldsymbol{x} + \gamma_0), \ \ \boldsymbol{x} \in X, \ \boldsymbol{u} \in \mathcal{U}_q.$$

Our empirical CVaR robust problem, presented in Section 3, requires no particular assumption on uncertainty set $\widetilde{\mathcal{U}}$, but if a suitable uncertainty set such as described above is provided, it is possible to ensure the convergence of the empirical CVaR robust problem to CVaR robust problem.

2.2 A Risk Measure: Conditional Value-at-Risk

For introducing a risk measure to be minimized, we regard $\boldsymbol{u}$ as a random vector, governed by a probability measure on $\mathcal{U}$. The distribution function $\Phi(\cdot\,|\boldsymbol{x})$ of $f(\boldsymbol{x}, \boldsymbol{u})$ and a threshold $\alpha_\beta(\boldsymbol{x})$ with some confidence level $\beta \in (0, 1)$ are defined as follows:

$$\Phi(\alpha\,|\boldsymbol{x}) := \Pr\{\, f(\boldsymbol{x}, \boldsymbol{u}) \leq \alpha \,\},$$
$$\alpha_\beta(\boldsymbol{x}) := \min\{\, \alpha \ : \ \Phi(\alpha\,|\boldsymbol{x}) \geq \beta \,\}.$$

We note that α_β is well-defined because $\Phi(\alpha\,|\boldsymbol{x})$ is right continuous and non-decreasing with respect to α. α_β is known as the *value-at-risk* (VaR) in the context of financial risk management, and it is expected that f exceeds α_β only in $(1 - \beta) \times 100\%$.

Following the discussion of [13], we introduce the β-tail distribution function to focus on the tail part of $\Phi(\alpha\,|\boldsymbol{x})$ as

$$\Phi_\beta(\alpha\,|\,\boldsymbol{x}) := \left\{ \begin{array}{ll} 0 & \text{for } \alpha < \alpha_\beta(\boldsymbol{x}) \\ \dfrac{\Phi(\alpha\,|\,\boldsymbol{x}) - \beta}{1 - \beta} & \text{for } \alpha \geq \alpha_\beta(\boldsymbol{x}) \end{array} \right. .$$

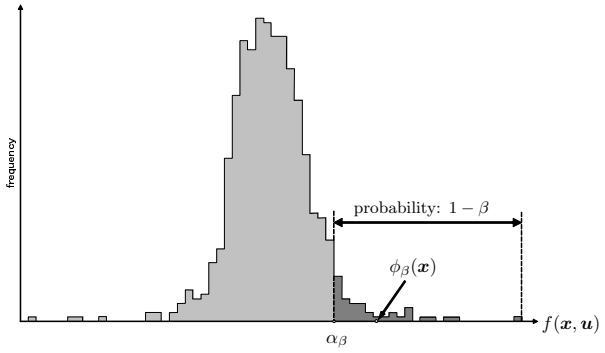

Figure 1: Illustration of the β-tail expectation $\phi(\boldsymbol{x})$ of f for fixed $\boldsymbol{x}$

Using the expectation operator $\boldsymbol{E}_\beta[\cdot]$ under the β-tail distribution Φ_β, let us define the β-tail expectation of f by

$$\phi_\beta(\boldsymbol{x}) := \boldsymbol{E}_\beta\left[\, f(\boldsymbol{x},\boldsymbol{u})\,\right], \tag{2}$$

which is the risk measure known as the *conditional value-at-risk* (CVaR). Denoting the expectation under the original distribution Φ by $\boldsymbol{E}[\cdot]$, the following relation shown in [13]:

$$\alpha_\beta(\boldsymbol{x}) \le \boldsymbol{E}[\, f(\boldsymbol{x},\boldsymbol{u}) \,|\, f(\boldsymbol{x},\boldsymbol{u}) \ge \alpha_\beta(\boldsymbol{x})\,] \;\le\; \phi_\beta(\boldsymbol{x}) \;\le\; \boldsymbol{E}[\, f(\boldsymbol{x},\boldsymbol{u}) \,|\, f(\boldsymbol{x},\boldsymbol{u}) > \alpha_\beta(\boldsymbol{x})\,]$$

implies that ϕ_β is approximately equal to the conditional expectation of f which exceeds the threshold α_β with fixed variables $\boldsymbol{x}$.

To minimize $\phi_\beta(\boldsymbol{x})$, [13] introduces a simpler auxiliary function $F_\beta : R^{n+1} \to R$, defined by

$$F_\beta(\boldsymbol{x},\alpha) := \alpha + \frac{1}{1-\beta}\boldsymbol{E}\left[\,[f(\boldsymbol{x},\boldsymbol{u}) - \alpha]^+\,\right],$$

where $[X]^+ := \max\{X, 0\}$, and confirms the formula

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

This equality provides a shortcut to minimizing $\phi_\beta(\boldsymbol{x})$ as

$$\min_{\boldsymbol{x}\in X} \phi_\beta(\boldsymbol{x}) = \min_{(\boldsymbol{x},\alpha)\in X\times R} F_\beta(\boldsymbol{x},\alpha), \tag{3}$$

that is, the minimal value $\phi_\beta(\boldsymbol{x})$ and an optimal solution can be achieved by minimizing $F_\beta(\boldsymbol{x},\alpha)$ with respect to $\boldsymbol{x} \in X$ and $\alpha \in R$ simultaneously. Furthermore, it is shown in [13] that, with an optimal solution $(\boldsymbol{x}^*, \alpha^*)$ of the right-hand side optimization problem, α^* is almost (or sometimes exactly) equal to $\alpha_\beta(\boldsymbol{x}^*)$.

It should be noted that the optimal value of (3), *i.e.*, $\min_{(\boldsymbol{x},\alpha)\in X\times R} F_\beta(\boldsymbol{x},\alpha)$ is non-decreasing with respect to β. Let its optimal solution be $(\boldsymbol{x}^*_\beta, \alpha^*_\beta)$. For arbitrary $(\bar{\boldsymbol{x}}, \bar{\alpha})$ and $0 < \beta_1 \le \beta_2 < 1$, we have $F_{\beta_1}(\bar{\boldsymbol{x}}, \bar{\alpha}) \le F_{\beta_2}(\bar{\boldsymbol{x}}, \bar{\alpha})$. Therefore, we see that

$$\min_{(\boldsymbol{x},\alpha)\in X\times R} F_{\beta_1}(\boldsymbol{x},\alpha) = F_{\beta_1}(\boldsymbol{x}^*_{\beta_1}, \alpha^*_{\beta_1}) \le F_{\beta_1}(\boldsymbol{x}^*_{\beta_2}, \alpha^*_{\beta_2}) \le F_{\beta_2}(\boldsymbol{x}^*_{\beta_2}, \alpha^*_{\beta_2}) = \min_{(\boldsymbol{x},\alpha)\in X\times R} F_{\beta_2}(\boldsymbol{x},\alpha)$$

which prove the non-decreasingness of $\min_{(\boldsymbol{x},\alpha)\in X\times R} F_\beta(\boldsymbol{x},\alpha)$ with respect to β.

The following proposition, which is a part of Proposition 8 of [13], evaluates the VaR for the extreme case where discreteness of probability distribution rules entirely, as in scenario-based optimization under uncertainty.

Proposition 2.1 (VaR for scenario models) : *Suppose the probability measure is concentrated in finitely many points $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N$ of $\mathcal{U}$. Fixing $\boldsymbol{x}$, let those corresponding values be ordered as $f(\boldsymbol{x},\boldsymbol{u}_1)<\ldots<f(\boldsymbol{x},\boldsymbol{u}_N)$, with the probability of $f(\boldsymbol{x},\boldsymbol{u}_k)$ being $p_k>0$. Let k_β be the unique index such that $\sum_{k=1}^{k_\beta} p_k \geq \beta > \sum_{k=1}^{k_\beta-1} p_k$. Then, the VaR is given by $\alpha_\beta(\boldsymbol{x}) = f(\boldsymbol{x},\boldsymbol{u}_{k_\beta})$.*

3 CVaR Robust Optimization Problem

3.1 Finite Uncertainty Set $\mathcal{U}$

We consider an uncertainty set $\mathcal{U}$ which consists of a finite number of elements such as $\mathcal{U} = \{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N\}$. The elements $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N$ are regarded as scenarios for uncertain data, and assumed to have the same occurrence probability. CVaR robust problem minimizes the β-tail expectation of $f(\boldsymbol{x},\boldsymbol{u})$ over $100\times(1-\beta)\%$ worst cases, that is, $f(\boldsymbol{x},\boldsymbol{u})>\alpha_\beta(\boldsymbol{x})$. This problem is described as

$$\min_{\boldsymbol{x}\in X,\alpha} \alpha + \frac{1}{(1-\beta)N}\sum_{i=1}^{N}[f(\boldsymbol{x},\boldsymbol{u}_i)-\alpha]^+, \tag{4}$$

which can be transformed into the problem:

$$\begin{aligned} \min_{\boldsymbol{x},\alpha,\boldsymbol{z}} \quad & \alpha + \frac{1}{(1-\beta)N}\sum_{i=1}^{N} z_i \\ \text{s.t.} \quad & z_i \geq f(\boldsymbol{x},\boldsymbol{u}_i)-\alpha, \quad i=1,\ldots,N \\ & \boldsymbol{z}\geq \boldsymbol{0}, \quad \boldsymbol{x}\in X. \end{aligned}$$

This problem is solvable since X is convex and $f(\boldsymbol{x},\boldsymbol{u})$ is convex in $\boldsymbol{x}$.

We explore the relation between CVaR robust problem (4) and the usual robust problem (1) which minimizes the worst case among $\boldsymbol{u}\in\mathcal{U}$. The following theorem implies that the min-max robust optimization problem (1) is a special case of the proposed problem (4) with $\beta\in(1-\frac{1}{N},1)$ and we can say that (4) is a general robust optimization problem with a parameter $\beta\in(0,1)$ which corresponds to confidence level of the conditional value-at-risk measure.

Theorem 3.1. *When β is sufficiently close to 1, concretely, $1-\frac{1}{N}<\beta<1$, (4) is equivalent to (1).*

Proof: Fixing $\boldsymbol{x} \in X$, let an optimal solution of $\max_{i=1,\ldots,N} f(\boldsymbol{x}, \boldsymbol{u}_i)$ be $\boldsymbol{u}_{k^*}$, where k^* may depend on $\boldsymbol{x}$. Proposition 2.1 shows that if the probability of the worst scenario $\boldsymbol{u}_{k^*} \in \mathcal{U}$ is greater than $1-\beta$, then $\phi_\beta(\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{u}_{k^*})$ holds for any $\boldsymbol{x} \in X$. Therefore, under uniform distribution,

$$\min_{\boldsymbol{x}\in X} \phi_\beta(\boldsymbol{x}) = \min_{\boldsymbol{x}\in X} \max_{i=1,\ldots,N} f(\boldsymbol{x}, \boldsymbol{u}_i)$$

holds for $1 - \frac{1}{N} < \beta < 1$. ∎

From the above theorem and the property that the optimal value of (3) is non-decreasing with respect to β, we see that the optimal value of (4) with parameter $\beta < 1$ is less than that of (1), *i.e.*,

$$\min_{\boldsymbol{x}\in X,\alpha} \alpha + \frac{1}{(1-\beta)N} \sum_{i=1}^{N} [f(\boldsymbol{x}, \boldsymbol{u}_i) - \alpha]^+ \leq \min_{\boldsymbol{x}\in X} \max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x}, \boldsymbol{u}).$$

As β decreases from 1 to 0, the optimal value of (4) decreases. This implies that when the best decision determined by the robust problem is too conservative, the conservativeness is eased by CVaR robust problem with appropriate parameter $\beta < 1$.

Theorem 3.2. *When β is sufficiently close to 0, (4) corresponds to a scenario-based one-stage stochastic programming problem,* $\min_{\boldsymbol{x}\in X} \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}, \boldsymbol{u}_i)$.

Proof: From Proposition 2.1, we see that $\alpha_\beta(\boldsymbol{x}) = \min_{i=1,\ldots,N} f(\boldsymbol{x}, \boldsymbol{u}_i)$ holds for β satisfying $0 < \beta \leq \frac{1}{N}$. Therefore, the objective function of (4) is replaced by

$$\min_{i=1,\ldots,N} f(\boldsymbol{x}, \boldsymbol{u}_i) + \frac{1}{(1-\beta)N} \sum_{i=1}^{N} \left(f(\boldsymbol{x}, \boldsymbol{u}_i) - \min_{i=1,\ldots,N} f(\boldsymbol{x}, \boldsymbol{u}_i) \right),$$

and CVaR robust problem (4) converge to $\min_{\boldsymbol{x}\in X} \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}, \boldsymbol{u}_i)$ as $\beta \to 0$. ∎

3.2 Infinite Uncertainty Set $\mathcal{U}$

Supposing that $\mathcal{U}$ is a bounded Lebesgue measurable set with positive volume, we provide a probability density function $g(\boldsymbol{u})$ over $\mathcal{U}$ and consider CVaR robust problem:

$$(P) \left| \begin{array}{rcl} \min\limits_{\boldsymbol{x}\in X,\alpha} F_\beta(\boldsymbol{x}, \alpha) := & \alpha + \dfrac{1}{1-\beta} \boldsymbol{E}\left[[f(\boldsymbol{x}, \boldsymbol{u}) - \alpha]^+ \right]. \\ = & \alpha + \dfrac{1}{1-\beta} \displaystyle\int_{\boldsymbol{u}\in\mathcal{U}} [f(\boldsymbol{x}, \boldsymbol{u}) - \alpha]^+ g(\boldsymbol{u}) d\boldsymbol{u}. \end{array} \right. \tag{5}$$

It is difficult to solve the problem (5) with a generally defined uncertainty set $\mathcal{U}$, and thus, by using a finite set $\mathcal{U}_{(N)} := \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\} \subset \mathcal{U}$ which consists of independently and identically distributed random samples on $\mathcal{U}$ and replacing the integral with empirical mean, one has

$$(P_N) \left| \min_{\boldsymbol{x}\in X,\alpha} \widehat{F}_\beta(\boldsymbol{x}, \alpha) := \alpha + \frac{1}{(1-\beta)N} \sum_{i=1}^{N} [f(\boldsymbol{x}, \boldsymbol{u}_i) - \alpha]^+. \right. \tag{6}$$

Furthermore, the empirical CVaR robust problem (P_N) can be transformed into

$$\begin{aligned} \min_{\boldsymbol{x},\alpha,\boldsymbol{z}} \quad & \alpha + \frac{1}{(1-\beta)N}\sum_{i=1}^{N} z_i \\ \text{s.t.} \quad & z_i \geq f(\boldsymbol{x},\boldsymbol{u}_i) - \alpha, \;\; i = 1,\ldots,N \\ & \boldsymbol{z} \geq \boldsymbol{0}, \;\; \boldsymbol{x} \in X. \end{aligned}$$

Note that (P_N) is solvable since X is convex and $f(\boldsymbol{x},\boldsymbol{u})$ is convex in $\boldsymbol{x}$. When the number of samples is sufficiently large, (P_N) would be a nice approximation for (P) while computational tasks increase.

To elucidate the convergence property of (P_N), we need some results in the field of machine learning [19]. The assertion related to convergence properties of (P_N) is stated below without proof. The proof of Theorem 3.3 is provided in Section 4, together with brief reference to Vapnik-Chervonenkis (VC) dimension h.

Theorem 3.3. *Suppose that $|f(\boldsymbol{x},\boldsymbol{u})| \leq M$ for any $\boldsymbol{x} \in X$ and $\boldsymbol{u} \in \mathcal{U}$, and the VC dimension h of function set*

$$\mathcal{F}_\beta = \left\{ \alpha + \frac{1}{1-\beta}[f(\boldsymbol{x},\cdot) - \alpha]^+ : \mathcal{U} \to R \mid x \in X, \alpha \in [-M, M] \right\} \tag{7}$$

is finite. Then, inequality

$$\left| \min_{\boldsymbol{x}\in X,\alpha} \widehat{F}_\beta(\boldsymbol{x},\alpha) - \min_{\boldsymbol{x}\in X,\alpha} F_\beta(\boldsymbol{x},\alpha) \right| \leq \mathcal{E}_\beta(N,\eta)$$

holds with probability at least $1-\eta$, where

$$\mathcal{E}_\beta(N,\eta) := \frac{2M(3-\beta)}{1-\beta}\sqrt{\frac{h\log 2N + h - h\log h - \log(\eta/4)}{N}} + \frac{1}{N}. \tag{8}$$

Corollary 3.4. *Suppose that the VC dimension of $\mathcal{F}_\beta$ is finite. For any $\epsilon > 0$, the optimal value of (P_N) converges to that of (P) in probability, i.e.,*

$$\lim_{N\to\infty} \Pr\left\{ \left| \min_{\boldsymbol{x}\in X,\alpha} \widehat{F}_\beta(\boldsymbol{x},\alpha) - \min_{\boldsymbol{x}\in X,\alpha} F_\beta(\boldsymbol{x},\alpha) \right| > \epsilon \right\} = 0.$$

Now we show the relation between the usual robust problem (1) and the proposed CVaR robust problem (5) (or the empirical approximation (6)).

Theorem 3.5. *For an arbitrary $\beta \in (0,1)$, the optimal values of (5) and (6) are less than that of (1).*

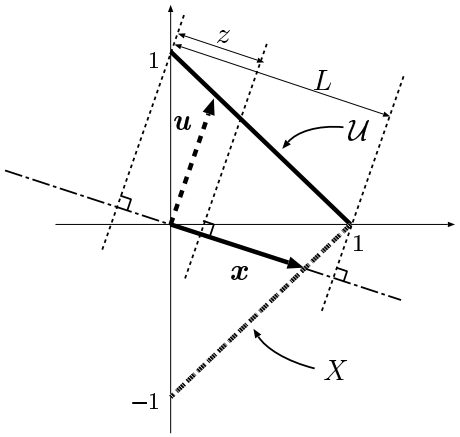

Figure 2: The constraint on $\boldsymbol{x}$ and the uncertainty set $\mathcal{U}$ in Example 3.6.

Proof: The definition of $\phi_\beta(\boldsymbol{x})$, presented in (2), implies $\phi_\beta(\boldsymbol{x}) \leq \max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{u})$ for any $\beta \in (0,1)$. Denoting an optimal solution of $\min_{\boldsymbol{x}\in X} \max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{u})$ as $\boldsymbol{x}^*$, we have

$$\min_{\boldsymbol{x}\in X,\alpha} F_\beta(\boldsymbol{x},\alpha) = \min_{\boldsymbol{x}\in X} \phi_\beta(\boldsymbol{x}) \leq \phi_\beta(\boldsymbol{x}^*) \leq \max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x}^*,\boldsymbol{u}) = \min_{\boldsymbol{x}\in X} \max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{u}).$$

Likewise, we have

$$\min_{\boldsymbol{x}\in X,\alpha} \widehat{F}_\beta(\boldsymbol{x},\alpha) \leq \min_{\alpha} \widehat{F}_\beta(\boldsymbol{x}^*,\alpha) \leq \max_{\boldsymbol{u}\in\mathcal{U}_{(N)}} f(\boldsymbol{x}^*,\boldsymbol{u}) \leq \max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x}^*,\boldsymbol{u}) = \min_{x\in X} \max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{u}).$$

Above inequalities prove this theorem. ∎

It is likely that the optimal value of CVaR robust problem (5) converges to that of the robust problem (1) as β converges to one. We show a simple example in which it is the case.

Example 3.6. Let X and $\mathcal{U}$ be

$$X = \{(x_1,x_2) \in R^2 \mid x_1 - x_2 = 1, x_1 \geq 0, x_2 \leq 0\},$$
$$\mathcal{U} = \{(u_1,u_2) \in R^2 \mid u_1 + u_2 = 1, u_1 \geq 0, u_2 \geq 0\}.$$

The objective function $f(\boldsymbol{x},\boldsymbol{u})$ is defined as $f(\boldsymbol{x},\boldsymbol{u}) = x_1u_1 + x_2u_2$. The constraint X and the uncertainty set $\mathcal{U}$ are depicted in Figure 2.

For any $\boldsymbol{x} = (x_1,x_2) \in X$, $x_1 > x_2$ holds, and then, one has $x_2 \leq f(\boldsymbol{x},\boldsymbol{u}) \leq x_1, \forall \boldsymbol{u} \in \mathcal{U}$. Thus the optimal value of robust problem for $f(\boldsymbol{x},\boldsymbol{u})$ is equal to zero because

$$\min_{\boldsymbol{x}\in X} \max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{u}) = \min_{\boldsymbol{x}\in X} x_1 \cdot 1 + x_2 \cdot 0 = 0.$$

We suppose that uniform distribution is defined on $\mathcal{U}$. By referring Figure 2, we find that the distribution function of $f(\boldsymbol{x},\boldsymbol{u})$ for fixed $\boldsymbol{x} \in X$ is given as

$$\Pr\left\{\frac{f(\boldsymbol{x},\boldsymbol{u})}{\|\boldsymbol{x}\|_2} \leq \frac{x_2}{\|\boldsymbol{x}\|_2} + z\right\} = \frac{z}{L} = \frac{z}{(x_1-x_2)/\|\boldsymbol{x}\|_2}, \quad 0 \leq z \leq \frac{x_1-x_2}{\|\boldsymbol{x}\|_2}$$
$$\Longleftrightarrow \Pr\{f(\boldsymbol{x},\boldsymbol{u}) \leq x_2 + \beta(x_1-x_2)\} = \beta, \quad 0 \leq \beta \leq 1.$$

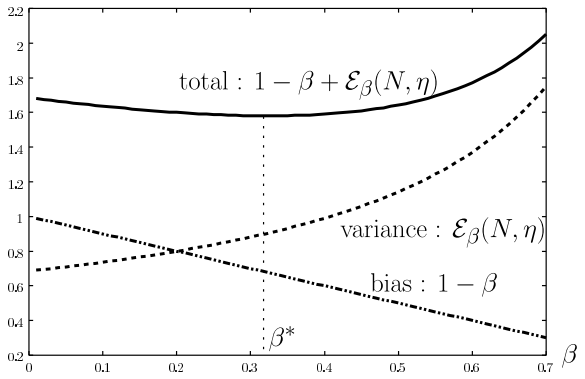

Figure 3: Upper bound of $\min_{\boldsymbol{x}\in X}\max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{u}) - \min_{\boldsymbol{x}\in X,\alpha} F_\beta(\boldsymbol{x},\alpha)$ and minimizer of the bound, β^*, are depicted.

Hence, the value-at-risk of $f(\boldsymbol{x},\boldsymbol{u})$ is equal to $\alpha_\beta(\boldsymbol{x}) = x_2 + \beta(x_1 - x_2) = x_1 + \beta - 1$, and the inequality

$$\beta - 1 \le \alpha_\beta(\boldsymbol{x}) \le \phi_\beta(\boldsymbol{x}) \le \max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{u})$$

holds for an arbitrary $\boldsymbol{x}\in X$, and then, one has

$$\beta - 1 \le \min_{\boldsymbol{x}\in X} \alpha_\beta(\boldsymbol{x}) \le \min_{\boldsymbol{x}\in X} \phi_\beta(\boldsymbol{x}) \le \min_{\boldsymbol{x}\in X}\max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{u}) = 0.$$

Consequently, $\min_{\boldsymbol{x}\in X,\alpha} F_\beta(\boldsymbol{x},\alpha)$ converges to $\min_{\boldsymbol{x}\in X}\max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{u})$ as $\beta\to 1$, that is,

$$\lim_{\beta\to 1}\min_{\boldsymbol{x}\in X,\alpha} F_\beta(\boldsymbol{x},\alpha) = \min_{\boldsymbol{x}\in X}\max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{u}) = 0$$

holds in this example. Moreover, the difference between the optimal value of robust problem and that of empirical CVaR problem is upper bounded in probability at least $1-\eta$ as follows,

$$\begin{aligned}
&\min_{\boldsymbol{x}\in X}\max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{u}) - \min_{x\in X,\alpha} \widehat{F}_\beta(\boldsymbol{x},\alpha)\\
&= \Big\{\min_{\boldsymbol{x}\in X}\max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{u}) - \min_{x\in X,\alpha} F_\beta(\boldsymbol{x},\alpha)\Big\} + \Big\{\min_{x\in X,\alpha} F_\beta(\boldsymbol{x},\alpha) - \min_{x\in X,\alpha} \widehat{F}_\beta(\boldsymbol{x},\alpha)\Big\}\\
&\le 1-\beta+\mathcal{E}_\beta(N,\eta),
\end{aligned}$$

since $\min_{\boldsymbol{x}\in X}\max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{u}) = 0$ and $\beta - 1 \le \min_{x\in X,\alpha} F_\beta(\boldsymbol{x},\alpha) = \min_{x\in X}\phi_\beta(\boldsymbol{x})$. Note that M in $\mathcal{E}_\beta(N,\eta)$ is equal to 1, and VC dimension h is at most 3 (cf. Example 4.2). In the machine learning literature, first term, $1-\beta$, in the upper bound is regarded as bias and second term, $\mathcal{E}_\beta(N,\eta)$, is deemed to be variance. As shown in Figure 3, an optimal solution β^*, which minimize the difference $\min_{\boldsymbol{x}\in X}\max_{\boldsymbol{u}\in\mathcal{U}} f(\boldsymbol{x},\boldsymbol{u}) - \min_{\boldsymbol{x}\in X,\alpha} F_\beta(\boldsymbol{x},\alpha)$, is determined by trade-off between bias and variance. The strategy for selecting β is similar to the structural risk minimization principle [19]. When $\eta = 1/N$, the upper bound $1-\beta+\mathcal{E}_\beta(N,1/N)$ converges to zero by putting $1-\beta = O((\log N/N)^{1/4})$ and the optimal value of $\widehat{F}_\beta(\boldsymbol{x},\boldsymbol{u})$ converges to that of the robust problem in probability.

4 Convergence Property for Empirical Problem

4.1 Proof of Theorem 3.3

To ensure the convergence property of (P_N), we need fundamental results in the field of machine learning [19]. Let us define

$$l_\beta(\boldsymbol{u};\boldsymbol{x},\alpha) := \alpha + \frac{1}{1-\beta}[f(\boldsymbol{x},\boldsymbol{u})-\alpha]^+,$$

then, the auxiliary function F_β and its empirical approximation $\widehat{F}_\beta$ are respectively given as $F_\beta(\boldsymbol{x},\alpha) = \boldsymbol{E}[l_\beta(\boldsymbol{u};\boldsymbol{x},\alpha)]$, and $\widehat{F}_\beta(\boldsymbol{x},\alpha) = \frac{1}{N}\sum_{i=1}^N l_\beta(\boldsymbol{u}_i;\boldsymbol{x},\alpha)$. For each $(\boldsymbol{x},\alpha)$, $\widehat{F}_\beta(\boldsymbol{x},\alpha)$ converges to $F_\beta(\boldsymbol{x},\alpha)$ in probability as $N\to\infty$ by the law of large numbers. However, it is rather nontrivial whether the optimal value of (P_N) converges to that of (P) because generally an optimal solution of (P) is different from that of (P_N). Hence, instead of the law of large numbers, we apply *the uniform law of large numbers* to prove convergence properties of (P_N).

Theorem of the uniform law of large numbers [19] requires the condition that $l_\beta(\boldsymbol{u};\boldsymbol{x},\alpha)$ is bounded for any $\boldsymbol{x}$ and α. When an upper bound of $|f(\boldsymbol{x},\boldsymbol{u})|$ is estimated such as $|f(\boldsymbol{x},\boldsymbol{u})|\le M$ for $\boldsymbol{x}\in X$ and $\boldsymbol{u}\in\mathcal{U}$, then, that of $|l_\beta(\boldsymbol{u};\boldsymbol{x},\alpha)|$ is also estimated with M. Recall that an optimal solution of (P) and that of (P_N) have the form of $(\boldsymbol{x}^*,\alpha_\beta(\boldsymbol{x}^*))$, and clearly $\alpha_\beta(\boldsymbol{x}^*)$ is bounded in the interval $[-M,M]$ regardless of distribution on uncertainty set. Thus, boundedness of $|l_\beta(\boldsymbol{u};\boldsymbol{x},\alpha)|$ is derived as follows:

$$\begin{aligned}
|l_\beta(u;\boldsymbol{x},\alpha)| &\le |\alpha| + \frac{1}{1-\beta}|[f(\boldsymbol{x},\boldsymbol{u})-\alpha]^+| \\
&\le |\alpha| + \frac{1}{1-\beta}|f(\boldsymbol{x},\boldsymbol{u})-\alpha| \\
&\le M + \frac{2M}{1-\beta} \\
&= \frac{M(3-\beta)}{1-\beta}.
\end{aligned}$$

Let $\mathcal{F}_\beta$ be function set defined as (7), that is,

$$\mathcal{F}_\beta = \{l_\beta(\cdot\,;\boldsymbol{x},\alpha) : \mathcal{U}\to R \mid \boldsymbol{x}\in X, \alpha\in[-M,M]\},$$

then, direct application of uniform law of large numbers leads to Theorem 4.1. In the theorem, combinatorial complexity of $\mathcal{F}_\beta$ called VC dimension governs the worst-case convergence property of empirical means.

Theorem 4.1 (uniform law of large numbers [19]) : *Let h be the VC dimension of $\mathcal{F}_\beta$,*

then, one has inequality,

$$\begin{aligned}&\Pr\left\{\sup_{\boldsymbol{x}\in X,\alpha\in[-M,M]}\left|\frac{1}{N}\sum_{i=1}^{N}l_\beta(\boldsymbol{u}_i;\boldsymbol{x},\alpha)-\boldsymbol{E}[l_\beta(\boldsymbol{u};\boldsymbol{x},\alpha)]\right|>\epsilon\right\}\\ \leq&4\exp\left\{\left(\frac{h\log 2N+h-h\log h}{N}-\frac{1}{\left(\frac{2M(3-\beta)}{1-\beta}\right)^2}\left(\epsilon-\frac{1}{N}\right)\right)^2 N\right\},\end{aligned}$$

for arbitrary N such as $2N\geq h$.

The statement of Theorem 4.1 is rephrased as follows: fixing $\eta\in(0,1)$, with probability at least $1-\eta$, inequality

$$\left|\frac{1}{N}\sum_{i=1}^{N}l_\beta(\boldsymbol{u}_i;\boldsymbol{x},\alpha)-\boldsymbol{E}[l_\beta(\boldsymbol{u};\boldsymbol{x},\alpha)]\right|<\mathcal{E}_\beta(N,\eta) \tag{9}$$

holds for any $\boldsymbol{x}\in X$ and $\alpha\in[-M,M]$, where $\mathcal{E}_\beta(N,\eta)$ is defined as (8). Note that the bound $\mathcal{E}_\beta(N,\eta)$ does not depend on probability distribution on $\mathcal{U}$. That is, the bound is valid for any distribution on $\mathcal{U}$.

Uniform law of large numbers assures the uniform convergence of empirical means:

$$\begin{aligned}&\sup_{\boldsymbol{x}\in X,\alpha\in[-M,M]}\left|\frac{1}{N}\sum_{i=1}^{N}l_\beta(\boldsymbol{u}_i;\boldsymbol{x},\alpha)-\boldsymbol{E}[l_\beta(\boldsymbol{u};\boldsymbol{x},\alpha)]\right|\\ =&\sup_{\boldsymbol{x}\in X,\alpha\in[-M,M]}\left|\widehat{F}_\beta(\boldsymbol{x},\alpha)-F_\beta(\boldsymbol{x},\alpha)\right|\longrightarrow 0\quad(N\to\infty),\end{aligned}$$

if the VC dimension of $\mathcal{F}_\beta$ is finite.

Now we are ready to prove Theorem 3.3, which evaluates the difference between the optimal value of (P) and that of (P_N).

Proof of Theorem 3.3: We denote an optimal solution of (P_N) by $(\bar{\boldsymbol{x}}_N,\bar{\alpha}_N)$ and that of (P) by $(\boldsymbol{x}^*,\alpha^*)$. Note that $F_\beta(\boldsymbol{x}^*,\alpha^*)\leq F_\beta(\bar{\boldsymbol{x}}_N,\bar{\alpha}_N)$ and $\widehat{F}_\beta(\bar{\boldsymbol{x}}_N,\bar{\alpha}_N)\leq\widehat{F}_\beta(\boldsymbol{x}^*,\alpha^*)$ hold. In addition, (9) ensures that inequalities $-\mathcal{E}_\beta(N,\eta)\leq\widehat{F}_\beta(\bar{\boldsymbol{x}}_N,\bar{\alpha}_N)-F_\beta(\bar{\boldsymbol{x}}_N,\bar{\alpha}_N)$ and $\widehat{F}_\beta(\boldsymbol{x}^*,\alpha^*)-F_\beta(\boldsymbol{x}^*,\alpha^*)\leq\mathcal{E}_\beta(N,\eta)$ hold simultaneously with probability at least $1-\eta$. Hence, one has

$$\begin{aligned}-\mathcal{E}_\beta(N,\eta)&\leq\widehat{F}_\beta(\bar{\boldsymbol{x}}_N,\bar{\alpha}_N)-F_\beta(\bar{\boldsymbol{x}}_N,\bar{\alpha}_N)\\ &\leq\widehat{F}_\beta(\bar{\boldsymbol{x}}_N,\bar{\alpha}_N)-F_\beta(\bar{\boldsymbol{x}}_N,\bar{\alpha}_N)+F_\beta(\bar{\boldsymbol{x}}_N,\bar{\alpha}_N)-F_\beta(\boldsymbol{x}^*,\alpha^*)\\ &=\widehat{F}_\beta(\bar{\boldsymbol{x}}_N,\bar{\alpha}_N)-F_\beta(\boldsymbol{x}^*,\alpha^*)\\ &\leq\widehat{F}_\beta(\bar{\boldsymbol{x}}_N,\bar{\alpha}_N)-F_\beta(\boldsymbol{x}^*,\alpha^*)+\widehat{F}_\beta(\boldsymbol{x}^*,\alpha^*)-\widehat{F}_\beta(\bar{\boldsymbol{x}}_N,\bar{\alpha}_N)\\ &=\widehat{F}_\beta(\boldsymbol{x}^*,\alpha^*)-F_\beta(\boldsymbol{x}^*,\alpha^*)\\ &\leq\mathcal{E}_\beta(N,\eta),\end{aligned}$$

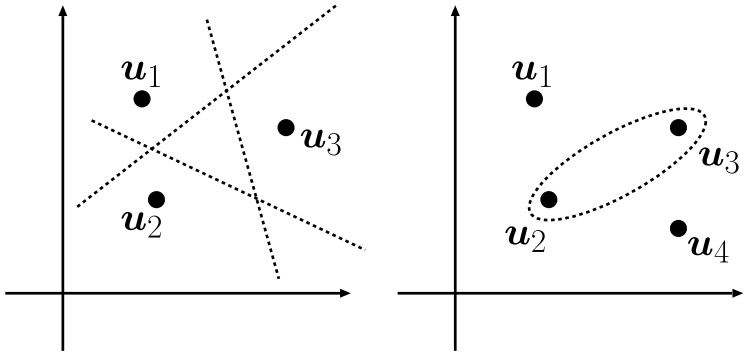

Figure 4: Three points can be separated into two classes in all 2^3 possible ways using affine functions, but not four: The points $\boldsymbol{u}_2, \boldsymbol{u}_3$ cannot be separated by a line from the vectors $\boldsymbol{u}_1, \boldsymbol{u}_4$.

with probability at least $1 - \eta$, and then,

$$\left| \min_{\boldsymbol{x} \in X, \alpha} \widehat{F}_\beta(\boldsymbol{x}, \alpha) - \min_{\boldsymbol{x} \in X, \alpha} F_\beta(\boldsymbol{x}, \alpha) \right| \leq \mathcal{E}_\beta(N, \eta)$$

is proved. ∎

4.2 Estimate of VC dimension and Upper Bound for $|f(\boldsymbol{x}, \boldsymbol{u})|$

Theorem 3.3 requires finite VC dimension h of $\mathcal{F}_\beta$ and an estimate of upper bound M satisfying $|f(\boldsymbol{x}, \boldsymbol{u})| \leq M$ for any $\boldsymbol{x} \in X$ and $\boldsymbol{u} \in \mathcal{U}$. VC dimension is a key concept for the uniform convergence of empirical means. For the detailed definition of VC dimension, one can refer [19]. Here, we show some examples to illustrate the definition of VC dimension.

Example 4.2. (VC dimension for set of affine functions): Let $l(\boldsymbol{u}; \gamma)$ be affine function such as $l(\boldsymbol{u}; \gamma) = \gamma_0 + \gamma_1 u_1 + \gamma_2 u_2$, where $\boldsymbol{u} = (u_1, u_2) \in R^2$, and $\mathcal{F}$ be $\mathcal{F} = \{l(\boldsymbol{u}; \gamma) \mid \gamma = (\gamma_0, \gamma_1, \gamma_2) \in R^3\}$ which consists of all affine functions on R^2. For any $\gamma \in R^3$, $l(\boldsymbol{u}, \gamma) \geq 0$ or $l(\boldsymbol{u}, \gamma) < 0$ holds, and we denote $l(\boldsymbol{u}, \gamma) \geq 0$ by '+' and $l(\boldsymbol{u}, \gamma) < 0$ by '−'. Let $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3\}$ be a set of three points in R^2, then there are eight possible ways for signs of $\{l(\boldsymbol{u}_1, \gamma), l(\boldsymbol{u}_2, \gamma), l(\boldsymbol{u}_3, \gamma)\}$, that is, $\{+,+,+\}, \{-,+,+\}, \ldots, \{-,-,-\}$. We can find three fixed points in R^2 such that all these combinations occur by varying parameter γ. However, there does not exist four points set $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \boldsymbol{u}_4\}$ in R^2 such that all 2^4 possible ways, $\{+,+,+,+\}, \ldots, \{-,-,-,-\}$, occur. Consequently, we find that VC dimension of $\mathcal{F}$ is equal to three. This is the maximum cardinality of subset in R^2 such that all possible combinations of signs occur (see Figure 4).

Example 4.3. (VC dimension for set of polynomial functions [18]): Let $l(\boldsymbol{u}; \gamma)$ be polynomial function on R^m whose degree is at most two and let $\mathcal{F}$ be a subset of polynomials on R^m up to degree two. Then, the VC dimension of $\mathcal{F}$ is less than or equal to $\frac{(m+1)^2}{2}$.

There are various kinds of objective functions $f(\boldsymbol{x}, \boldsymbol{u})$ and uncertainty sets $\widetilde{\mathcal{U}}$ such that a finite upper bound M for $|f(\boldsymbol{x}, \boldsymbol{u})|$ and finite VC dimension h are available. Here, for quadratic objective functions $f(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \boldsymbol{q}^\top \boldsymbol{x} + \gamma$ and the uncertainty sets $\mathcal{U}$ described in Section 2.1, we have actually evaluated upper bounds M and VC dimension h. Let r_x be sufficiently large number so that $\max_{\boldsymbol{x} \in X} \|\boldsymbol{x}\| \leq r_x$ holds, and $\sigma_{max}(\boldsymbol{Q})$ be the maximum eigenvalue of matrix $\boldsymbol{Q}$.

- Polytopic uncertainty set :

$$\begin{aligned} M_p &= \sqrt{\textstyle\sum_{j=1}^{\ell} (\sigma_{max}(\boldsymbol{Q}_j) r_x^2 + \|\boldsymbol{q}_j\| r_x + \gamma_j)^2}, \\ h_p &\leq \ell + 1. \end{aligned}$$

- Norm-constrained uncertainty set :

$$\begin{aligned} M_n &= \sqrt{\textstyle\sum_{j=1}^{\ell} (\sigma_{max}(\boldsymbol{Q}_j) r_x^2 + \|\boldsymbol{q}_j\| r_x + \gamma_j)^2} + \sigma_{max}(\boldsymbol{Q}_0) r_x^2 + \|\boldsymbol{q}_0\| r_x + \gamma_0, \\ h_n &\leq \ell + 1. \end{aligned}$$

- Quadratic uncertainty set :

$$\begin{aligned} M_q &= r_x^2 \sigma_{max}(\textstyle\sum_{i=1}^{\ell} \boldsymbol{R}_i^\top \boldsymbol{R}_i) + \sqrt{\textstyle\sum_{j=1}^{\ell} (\sigma_{max}(\boldsymbol{R}_0^\top \boldsymbol{R}_j + \boldsymbol{R}_j^\top \boldsymbol{R}_0) r_x^2 + \|\boldsymbol{q}_j\| r_x + \gamma_j)^2} \\ &\quad + \sigma_{max}(\boldsymbol{R}_0^\top \boldsymbol{R}_0) r_x^2 + \|\boldsymbol{q}_0\| r_x + \gamma_0, \\ h_q &\leq (m+1)^2/2. \end{aligned}$$

We show the computation of M only for norm-constrained uncertainty set and similar discussion holds for other sets. we have

$$\begin{aligned} & \max_{\boldsymbol{x}, \boldsymbol{u}} \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \boldsymbol{q}^\top \boldsymbol{x} + \gamma \\ \leq\ & \max_{\boldsymbol{x}} \left\| \begin{pmatrix} \boldsymbol{x}^\top \boldsymbol{Q}_1 \boldsymbol{x} + \boldsymbol{q}_1^\top \boldsymbol{x} + \gamma_1 \\ \cdots \\ \boldsymbol{x}^\top \boldsymbol{Q}_\ell \boldsymbol{x} + \boldsymbol{q}_\ell^\top \boldsymbol{x} + \gamma_\ell \end{pmatrix} \right\| + \max_{\boldsymbol{x}} \left\{ \boldsymbol{x}^\top \boldsymbol{Q}_0 \boldsymbol{x} + \boldsymbol{q}_0^\top \boldsymbol{x} + \gamma_0 \right\} \\ \leq\ & \sqrt{\textstyle\sum_{j=1}^{\ell} (\sigma_{max}(\boldsymbol{Q}_j) r_x^2 + \|\boldsymbol{q}_j\| r_x + \gamma_j)^2} + \sigma_{max}(\boldsymbol{Q}_0) r_x^2 + \|\boldsymbol{q}_0\| r_x + \gamma_0 := M_n. \end{aligned}$$

In empirical CVaR robust problem, it is possible to assume any kind of uncertainty set $\widetilde{\mathcal{U}}$. Convergence property of empirical CVaR robust problem is ensured under boundedness of $f(\boldsymbol{x}, \boldsymbol{u})$ and finiteness of VC dimension. These conditions are enough general to deal with practical problems, while uncertainty set of robust problems should be well-structured, such as polytopic, norm-constrained, or quadratic uncertainty set.

5 Applications to Statistical Learning Problems

5.1 Linear Classification

In this section we consider CVaR robust problem with finite uncertainty set for binary classification problem. We suppose that a set of training data $\boldsymbol{x}_i \in R^n$ which are labeled with binary

values $y_i \in \{\pm 1\}$ for $i = 1, \ldots, N$:

$$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N) \in R^n \times \{\pm 1\},$$

is provided. The binary classification problem seeks to find a decision function $g : R^n \to \{\pm 1\}$ using these training data so that g will predict as accurately as possible the labels of new data points, which are generated from the same probability distribution with training data.

If given set of training data is linearly separable, *i.e.*, there exists $(\boldsymbol{w}, b)$ such that $\boldsymbol{w} \neq \boldsymbol{0}$ and $y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) > 0$ hold for $i = 1, \ldots, N$, the hard margin support vector classification (HSVC) [19] provides a most reasonable decision function $g = \mathrm{sign}(\langle \boldsymbol{w}^*, \boldsymbol{x} \rangle + b^*)$, where $(\boldsymbol{w}^*, b^*)$ is an optimal solution of

$$\text{(HSVC)} \quad \left| \max_{\boldsymbol{w} \neq \boldsymbol{0}, b} \min_{i=1,\ldots,N} \frac{y_i \left(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b \right)}{\|\boldsymbol{w}\|}, \right.$$

and $\mathrm{sign}(\xi)$ is a function such that $\mathrm{sign}(\xi) = 1$ if $\xi \geq 0$ and -1, otherwise. (HSVC) is actually transformed to the equivalent problem

$$\min_{\boldsymbol{w}, b} \frac{1}{2} \|\boldsymbol{w}\|^2 \text{ s.t. } y_i \left(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b \right) \geq 1, \quad i = 1, \ldots, N$$

and solved as an convex quadratic program. It should be noted that $\frac{y_i \left(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b \right)}{\|\boldsymbol{w}\|}$ coincides with the Euclidean distance from $\boldsymbol{x}_i$ to the hyperplane $\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0$. When we regard $\mathcal{U} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ as an uncertainty set which consists of finite elements, (HSVC) corresponds to the robust optimization problem (1).

Recently, [11] proposed *conditional geometric score* optimization problem based on CVaR risk measure:

$$\text{(CGS)} \quad \left| \min_{\boldsymbol{w} \neq \boldsymbol{0}, b, \alpha} \alpha + \frac{1}{(1-\beta)N} \sum_{i=1}^{N} \left[-\frac{y_i \left(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b \right)}{\|\boldsymbol{w}\|} - \alpha \right]^+, \right. \tag{10}$$

not only for linearly separable but for non-separable training data set. Its numerical results imply the advantages of (CGS) over other kinds of linear classification such as ν-support vector classification (ν-SVC) [15] and robust linear programming approach [1]. (CGS) corresponds to CVaR robust problem (4) with a finite set $\mathcal{U}$ while (HSVC) does to the usual robust problem (1).

For linearly non-separable training data set, the optimal value of (HSVC) becomes negative and hence, (HSVC) is reduced to a difficult nonconvex problem. Similarly, for the same data set, (CGS) is transformed to a nonconvex optimization problem when the parameter β of (CGS) exceeds some level and its optimal value becomes positive. Hence, from a local optimum solution $(\boldsymbol{w}_h, b_h)$ of (HSVC) and $(\boldsymbol{w}_c, b_c)$ of (CGS), we construct hyperplanes $f_h(\boldsymbol{x}) = \langle \boldsymbol{w}_h, \boldsymbol{x} \rangle + b_h$ and $f_c(\boldsymbol{x}) = \langle \boldsymbol{w}_c, \boldsymbol{x} \rangle + b_c$, respectively, and compare training error and test error rates between these two problems. A local optimum solution $(\boldsymbol{w}_h, b_h)$ of (HSVC) is actually obtained by (CGS) with parameter $\beta > 1 - \frac{1}{N}$ according to Theorem 3.1. For an local optimization algorithm to solve (CGS), see [11].

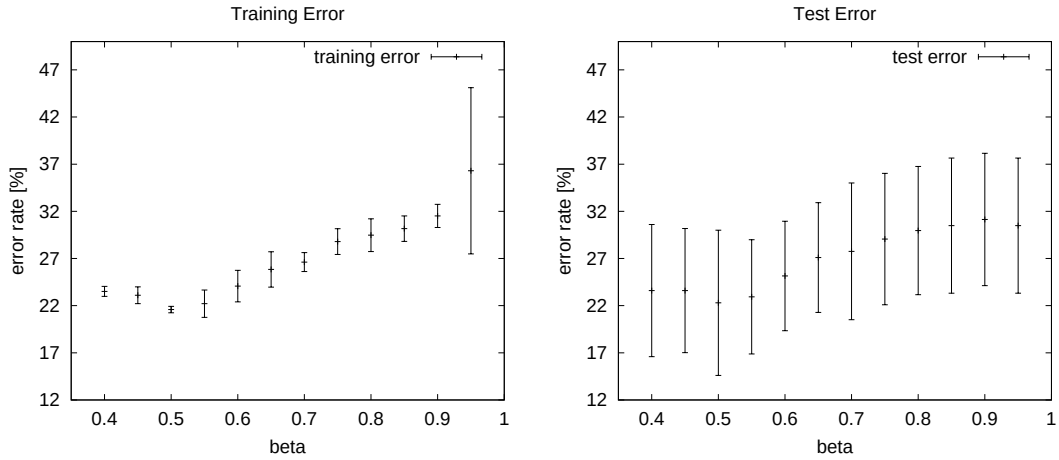

Figure 5: Training error and test error rates over DIABETES dataset.

To show potential superiority of our CVaR robust problem (CGS) over (HSVC), we performed ten-fold cross-validation for DIABETES dataset obtained from the UCI repository of machine learning databases [7]. We separate data into two groups: training data set and test data set. (HSVC) and (CGS) are formulated from training data and their decision function $g = \mathrm{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b)$ are tested over test data set. Figure 5 shows training error and test error rates obtained by (CGS) for $N_1 = 691$ training data set and $N_2 = 77$ test data set, respectively. For $\beta \geq 0.47$, those error rates are measured with a local optimal solution of (CGS), while numerical results of [11] shows that minimum error rate can be achieved at $\beta = 0.51$. Also, we see that (HSVC) corresponds to (CGS) with $\beta \geq 0.999 = 1 - \frac{1}{N_1}$ for training data. Figure 5 (right) implies that the prediction via the optimal solution of the robust problem does not hit right very often, compared with that of CVaR robust problem with $\beta = 0.5$.

5.2 Linear Regression

In linear regression problems, the main issue is to estimate linear function of input vector $\boldsymbol{x}$,

$$f(\boldsymbol{x}; \boldsymbol{w}, b) = \langle \boldsymbol{x}, \boldsymbol{w} \rangle + b, \quad \boldsymbol{x}, \boldsymbol{w} \in R^n, b \in R, \tag{11}$$

that best approximates response values. Selection of desired function is based on a training set of m independent and identically distributed training data, $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)$, drawn according to a probability distribution, where $y_i \in R, i = 1, \ldots, m$ denote response values.

Least square estimator is commonly used to estimate parameters $\boldsymbol{w}$ and b. It is often pointed out that a few number of outliers, which are far away from the bulk of observations, seriously degrade the accuracy of least square estimator. However, even if gross error occurs in response values, L_1 estimator defined as an optimal solution of

$$\min_{\boldsymbol{w}, b} \ \frac{1}{m} \sum_{i=1}^{m} |y_i - f(\boldsymbol{x}_i; \boldsymbol{w}, b)|,$$

depresses influence of outliers. Square regularization term is often added to objective functions such as

$$\min_{\boldsymbol{w},b} \ \frac{C}{m}\sum_{i=1}^{m}|y_i - f(\boldsymbol{x}_i;\boldsymbol{w},b)| + \frac{1}{2}\|\boldsymbol{w}\|^2, \tag{12}$$

where smoothing parameter C is a constant that adjusts effect of regularization. It has been clarified from experimental and theoretical viewpoints that square regularization term raises generalization ability by avoiding over-fitting to training data.

When training data is contaminated and the uncertainty of observation is represented by $\mathcal{U}$, min-max estimator defined as an optimal solution of

$$\min_{\boldsymbol{w},b} \max_{(\boldsymbol{x},y)\in\mathcal{U}} C|y - f(\boldsymbol{x};\boldsymbol{w},b)| + \frac{1}{2}\|\boldsymbol{w}\|^2, \tag{13}$$

is expected to provide conservative estimation results according to robust optimization principle.

In the following, we study CVaR robust problems derived from the min-max estimation (13) with apt uncertainty set. Both finite and infinite uncertainty set are considered. Average error and worst case error are computed to evaluate statistical stability of the estimated function (11).

5.2.1 Finite Uncertainty Set

First, function estimation under finite uncertainty set is studied. Here, uncertainty is defined as a set of training data, $\mathcal{U} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$. Robust optimization with square regularization term is the form of

$$\min_{\boldsymbol{w},b} \max_{i=1,\ldots,m} C|y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i\rangle - b| + \frac{1}{2}\|\boldsymbol{w}\|^2, \tag{14}$$

and then, for uniform distribution on $\mathcal{U}$, CVaR problem is given as

$$\min_{\boldsymbol{w},b,\alpha} C\alpha + \frac{C}{(1-\beta)m}\sum_{i=1}^{m}[|y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i\rangle - b| - \alpha]^+ + \frac{1}{2}\|\boldsymbol{w}\|^2, \tag{15}$$

where maximum residual is replaced by mean value of $100\times(1-\beta)\%$ largest residuals. Note that for $\beta = 0$, CVaR problem is identical to L_1 estimator with square regularization term as shown in Theorem 3.2. That is, CVaR problem connects L_1 estimation problem and robust optimization problem by one parameter β.

CVaR problem (15) is identical to ν-support vector regression (ν-SVR) [15] with $\nu = 1-\beta$ and smoothing parameter $\frac{C}{\nu}$, and Schölkoph, *et al.*, have proved that local movements of outliers do not influence solution of (15), where outlier denotes a sample that has worst $100\times(1-\beta)\%$ error in whole samples. According to Theorem 3.1, when ratio of outliers in $\mathcal{U}$ is more than $1/m$, robust optimization problem (14) would provide an unstable solution, though the optimal solution is stable under less than $1/m$ outlier ratio. To deal with more than $100\times(1-\beta)\%$

outliers in observations, CVaR problem (15) is useful and will provides accurate prediction to typical observations, not to worse-case observations.

In numerical experiments, we study accuracy of estimated function (11) based on observation set $\mathcal{U}$. Worst case error and average error are computed to evaluated the accuracy. Robust optimization will provide an optimal solution that has the smallest worst-case error, but may have large average error. On the other hand, an solution of CVaR problem is expected to have small average error.

In the following numerical experiments, input vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ in observations are independent and identically distributed from uniform distribution on $[-5, 5]^n$, where vector dimension is $n = 3$ and sample size is $m = 20$. Response values are prepared as $y_i = \langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle + b^* + \epsilon_i$ for $\boldsymbol{w}^* = (1, 1, 1)$ and $b^* = 1$, where noise terms $\epsilon_i, i = 1, \ldots, m$ are independently generated from nominal distribution with mean 0. Almost all noise terms have small variance, 1.0, and a sample $(\boldsymbol{x}, y) = (\boldsymbol{0}, 5)$ is mixed with set of observations as an outlier. Objective functions in robust problem and CVaR problem have smoothing parameter C, and in experiments, C is fixed to $10 \times m$. Estimation accuracy of solution $(\boldsymbol{w}, b)$ is evaluated in two ways: one is the worst-case error, $\max_i |y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b|$, and the other is average error, $\frac{1}{m} \sum_{i=1}^m |y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b|$. Here, the observation set, $\mathcal{U}$, is also used to evaluate estimation accuracy. 100 sets of observations are generated with different random seed, and mean values of average errors and worst case errors are computed over these 100 sets.

For each β, mean values of worst case errors and average errors are plotted with error bars in Figure 6. Note that robust optimization problem corresponds to CVaR problem with β such as $1 - \frac{1}{m} < \beta < 1$, and L_1 estimation does to CVaR problem with $\beta = 0$. Robust solutions minimize worst case errors as expected, while in average error, CVaR solutions are fairly better than robust solutions. That is, an outlier, $(\boldsymbol{x}, y) = (\boldsymbol{0}, 5)$, significantly affects robust solutions and degrades estimation accuracy to the other typical observations. On the other hand, L_1 estimation minimizes average error, but takes relatively large worst case error in comparison with CVaR problem with $\beta > 0$.

5.2.2 Infinite Uncertainty Set

Infinite uncertainty set is useful to deal with measurement error. L_1 estimator depresses influence of outliers in response y as stated before. In addition to outliers in response values, measurement error may mislead conclusions for inference. Typical form of measurement error is given as

$$\begin{pmatrix} \boldsymbol{x} \\ y \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_i^\circ \\ y_i^\circ \end{pmatrix} + \tilde{\boldsymbol{u}}, \tag{16}$$

where measurement error, $\tilde{\boldsymbol{u}}$, is $n + 1$ dimensional vector and $(\boldsymbol{x}_i^\circ, y_i^\circ)$ is i-th nominal data. If there is not measurement error, observation $(\boldsymbol{x}, y)$ is identical to nominal data $(\boldsymbol{x}_i^\circ, y_i^\circ)$. Suppose

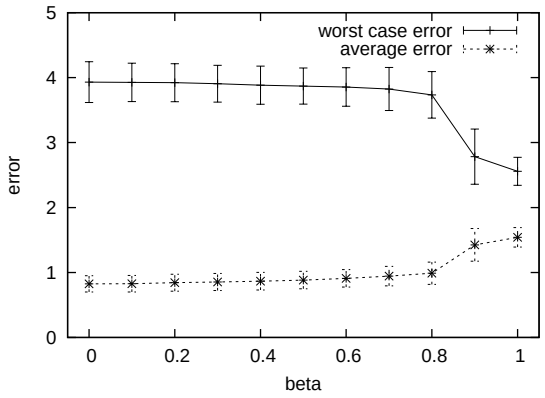

Figure 6: Worst case error and average error with error bars under finite uncertainty set.

that $(\boldsymbol{x}, y) \in \widetilde{\mathcal{U}}$, where $\widetilde{\mathcal{U}}$ is a compact set in R^{n+1} defined as

$$\widetilde{\mathcal{U}} = \cup_{i=1}^{m} \widetilde{\mathcal{U}}_i,$$
$$\widetilde{\mathcal{U}}_i = \left\{ (\boldsymbol{x}, y) \;\middle|\; \begin{pmatrix} \boldsymbol{x} \\ y \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_i^\circ \\ y_i^\circ \end{pmatrix} + D_i \boldsymbol{u}, \;\; \|\boldsymbol{u}\| \leq 1, \;\; \boldsymbol{u} \in \boldsymbol{R}^\ell \right\}.$$

Matrices $D_i \in R^{(n+1)\times \ell}$ are given as $D_i = [\boldsymbol{d}_{i1}, \ldots, \boldsymbol{d}_{i\ell}]$, and specify directions of measurement errors. Thus, robust problem under infinite uncertainty set $\widetilde{\mathcal{U}}$ is constructed as

$$\min_{\boldsymbol{w},b} \max_{(\boldsymbol{x},y)\in\widetilde{\mathcal{U}}} C|y - \langle \boldsymbol{w}, \boldsymbol{x} \rangle - b| + \frac{1}{2}\|\boldsymbol{w}\|^2,$$

and one can reformulate it as a second-order cone programming problem using the technique of [3].

For CVaR problem, let us define probability density p on $\widetilde{\mathcal{U}}$ as

$$p(\boldsymbol{x}, y) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\mathrm{Vol}(\mathcal{U}^i)} \mathbf{1}((\boldsymbol{x}, y) \in \widetilde{\mathcal{U}}_i),$$

where $\mathbf{1}(\cdot)$ is indicator function, and $\mathrm{Vol}(\mathcal{U})$ indicates the volume of a set $\mathcal{U}$. That is, nominal data $(\boldsymbol{x}_i^\circ, y_i^\circ)$ is uniformly chosen from $\{(\boldsymbol{x}_1^\circ, y_1^\circ), \ldots, (\boldsymbol{x}_m^\circ, y_m^\circ)\}$, and then, observation $(\boldsymbol{x}, y)$ is generated uniformly from $\widetilde{\mathcal{U}}_i$. CVaR problem under uncertainty set $\widetilde{\mathcal{U}}$ is given as

$$\min_{\boldsymbol{w},b,\alpha} C\alpha + \frac{C}{1-\beta} \boldsymbol{E}\left\{ [|y - \langle \boldsymbol{w}, \boldsymbol{x} \rangle - b| - \alpha]^+ \right\} + \frac{1}{2}\|\boldsymbol{w}\|^2, \tag{17}$$

and empirical approximation of expectation is constructed as

$$\min_{\boldsymbol{w},b,\alpha} C\alpha + \frac{C}{1-\beta} \frac{1}{N} \sum_{i=1}^{N} [|y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b| - \alpha]^+ + \frac{1}{2}\|\boldsymbol{w}\|^2, \tag{18}$$

where $(\boldsymbol{x}_i, y_i), i = 1, \ldots, N$ are independently and identically distributed from the probability $p(\boldsymbol{x}, y)$.

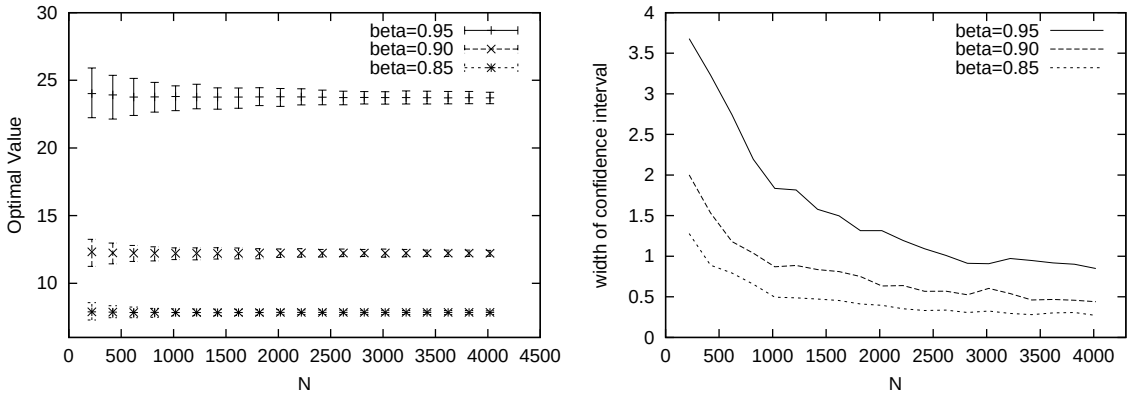

Figure 7: Left figure: Optimal value of empirical approximation of CVaR problem with 90% confidence interval. Right figure: width of 90% confidence interval for optimal value.

In Theorem 3.3, we have proved that optimal value of empirical approximation (18) converges to that of CVaR problem (17). Convergence in probability is confirmed by a simple numerical experiments as follows. Input vectors of nominal data $(\boldsymbol{x}_i^\circ, y_i^\circ)$ are randomly chosen from $[-5, 5]^n$, where $n = 3$ and $m = 20$, and response values are prepared as

$$y_i^\circ = \langle \boldsymbol{w}^*, \boldsymbol{x}_i^\circ \rangle + b^* + \epsilon_i, \quad i = 1, \ldots, 20, \tag{19}$$

for $\boldsymbol{w}^* = (1, 1, 1)$ and $b^* = 1$. The noise term ϵ_i is according to nominal distribution with mean 0 and variance 0.5^2, *i.e.* $N(0, 0.5^2)$. All components of perturbation matrices, D_i, are extracted uniformly from $[-1, 1]$. To compute the distribution of optimal value (18), 100 sets of samples on $\widetilde{\mathcal{U}}$ are generated by different random seeds.

For each β, optimal values of (18) with $C = 1 \times m = 20$ are plotted in Figure 7. In Figure 7 (left), optimal values seem to distribute around a constant value for each β. On the other hand, width of confidence interval for empirical CVaR problem decreases as N increases, as shown in Figure 7(right). Consequently, numerical experiments indicate that optimal value of (18) converges to a constant value. Moreover, optimal value increases as $\beta \to 1$. In Section 2, we have proved that optimal value of CVaR problem is non-decreasing function with respect to β, and this is the case even for empirical approximation of CVaR problem. In our experiments, this property is confirmed.

We have also provided the convergence rate of optimal value, $\mathcal{E}_\beta(N, \eta)$, which is increasing function with respect to β. Applying statistical curve fitting technique to our numerical results, we find that the width of confidence interval approximately decreases in the order of $O(\frac{1}{\sqrt{N}})$ which is almost same as the order of $\mathcal{E}_\beta(N, \eta)$. We also find that width of confidence interval increases as β increases. This is also similar to the property of $\mathcal{E}_\beta(N, \eta)$, which is increasing function with respect to β.

Next experiments illustrate that CVaR robust problems provide accurate estimation results in the sense of average error, compared with robust problems under infinite uncertainty set $\widetilde{\mathcal{U}}$.

Input vectors of nominal data are uniformly distributed on $[-5,5]^3 \in R^3$ and corresponding response values $y_i^\circ, i = 1,\ldots,m$ are prepared by (19), where $m = 20$ and $\epsilon_i, i = 1,\ldots,m$ are independently distributed according to $N(0,2^2)$. Matrices D_i for uncertainty set $\widetilde{\mathcal{U}}$ are all identical and proportional to identity matrix I.

Integration is involved in exact CVaR problem (17), and is approximated by empirical mean such as (18). In our experiments, 100 samples are uniformly distributed on $\widetilde{\mathcal{U}}_i$ for each $i = 1,\ldots,m$, and then, total number of samples, N, is equal to $100 \times m = 2000$. Solution $(\boldsymbol{w}_r, b_r)$ of robust problem and $(\boldsymbol{w}_c, b_c)$ of empirical CVaR problem are computed with smoothing parameter $C = 1 \times m = 20$, and estimation accuracy of parameter $(\boldsymbol{w}, b)$ under uncertainty is evaluated by worst case error,

$$\max_{(\boldsymbol{x},y)\in\widetilde{\mathcal{U}}} |y - \langle \boldsymbol{w}, \boldsymbol{x}\rangle - b|,$$

and approximated average error,

$$\begin{aligned}\boldsymbol{E}[|y - \langle \boldsymbol{w}, \boldsymbol{x}\rangle - b|] &= \frac{1}{m}\sum_{i=1}^{m} \frac{1}{\mathrm{Vol}(\mathcal{U}^i)} \int_{\widetilde{\mathcal{U}}_i} |y - \langle \boldsymbol{w}, \boldsymbol{x}\rangle - b| d\boldsymbol{x} dy \\ &\cong \frac{1}{m}\sum_{i=1}^{m} \frac{1}{1000} \sum_{k=1}^{1000} |\tilde{y}_{ik} - \langle \boldsymbol{w}, \tilde{\boldsymbol{x}}_{ik}\rangle - b|,\end{aligned}$$

where samples, $(\tilde{\boldsymbol{x}}_{ik}, \tilde{y}_{ik}), k = 1,\ldots,1000$, are uniformly distributed on $\widetilde{\mathcal{U}}_i$. For fixed uncertainty set $\mathcal{U}$, we repeated experiments 100 times with different random seeds for scenario sampling in empirical CVaR problem (18).

Figures 8, 9, and 10 show average error and worst case error with error bar. For each figure, uncertainty set is defined by $D_i = 2I$, $D_i = 3I$, and $D_i = 5I$, respectively. Estimation error of "nominal" defined as an optimal solution of

$$\min_{\boldsymbol{w},b} \frac{C}{m}\sum_{i=1}^{m} |y_i^\circ - \langle \boldsymbol{w}, \boldsymbol{x}_i^\circ\rangle - b| + \frac{1}{2}\|\boldsymbol{w}\|^2,$$

is also depicted in these figures. Note that nominal solution does not take into account measurement error represented by the uncertainty set $\widetilde{\mathcal{U}}$.

Solutions of robust problem minimize worst case error in any experiments for a certainty. We also find that when β comes close to 1, both average error and worst case error for empirical CVaR problem tend to converge to those for robust problem, as illustrated in Example 3.6. For $D_i = 2I$ (Figure 8), nominal problem provides smaller average error than empirical CVaR problem with $\beta \geq 0.7$ and robust problem because relatively small measurement error does not causes significant loss of information. As measurement error becomes larger, the estimation via empirical CVaR problem outperforms the other competitors in the sense of average error. Our experimental results indicate that empirical CVaR problem provides fairly good estimation results in comparison with robust problem on average, especially when measurement error, $\boldsymbol{u}$, causes information loss for estimated function more than noise term ϵ.

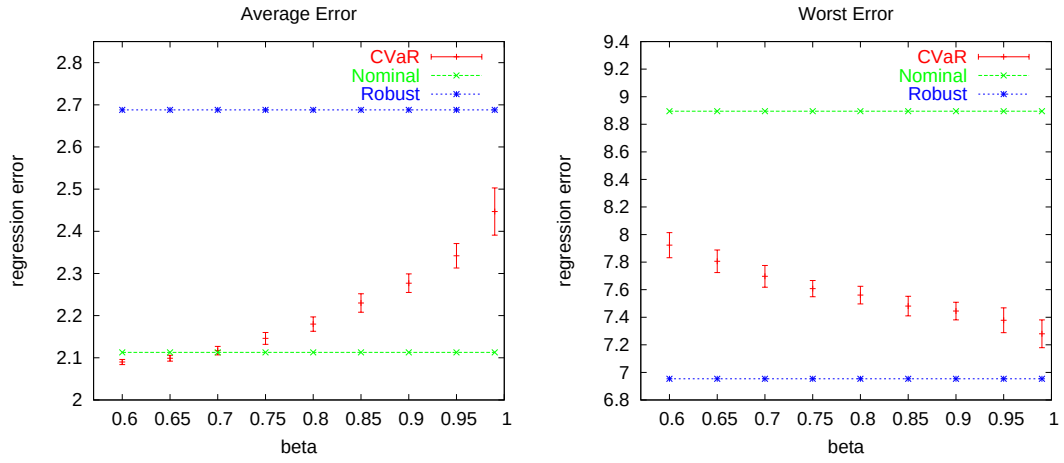

Figure 8: Estimation error plots of optimal solutions for robust problem, nominal problem, and empirical CVaR problem with $D_i = 2I, i = 1, \ldots, m$. Left figure: average error. Right figure: worst case error.

6 Conclusions

We applied a robust optimization approach based on conditional value-at-risk measure to statistical learning problems whose objective functions include uncertain data. The CVaR robust problem includes one parameter $\beta \in (0, 1)$, and minimizes expected value of costs in the worst class defined by β-quantile point (VaR) while the usual robust optimization minimizes cost in the worst case. When β is close to 1, CVaR robust problem is almost same as usual robust problem, and as β is far from 1, CVaR robust problem becomes more interested in the objective of minimizing average cost than the objective of minimizing maximum cost. Therefore, when the best decision determined by robust problem is too conservative, the conservativeness is eased by CVaR robust problem with appropriate parameter $\beta < 1$.

Throughout numerical experiments, we confirmed that CVaR robust problem dissolves overly conservativeness of robust optimization and depresses influence of outliers or measurement error which may be included in assumed uncertainty set. Furthermore, it is shown that CVaR robust approach is effective for linear classification problem and linear regression problem, which are studied frequently in the field of machine learning. CVaR robust problems are closely related to statistical learning models: ν-SVC and ν-SVR.

One of interesting research directions is to develop applications of CVaR robust optimization technique to more wide range of statistical learning. Also, it might be possible to solve CVaR robust problem parametrically by changing parameter β or provide some appropriate strategy for the selection of β.

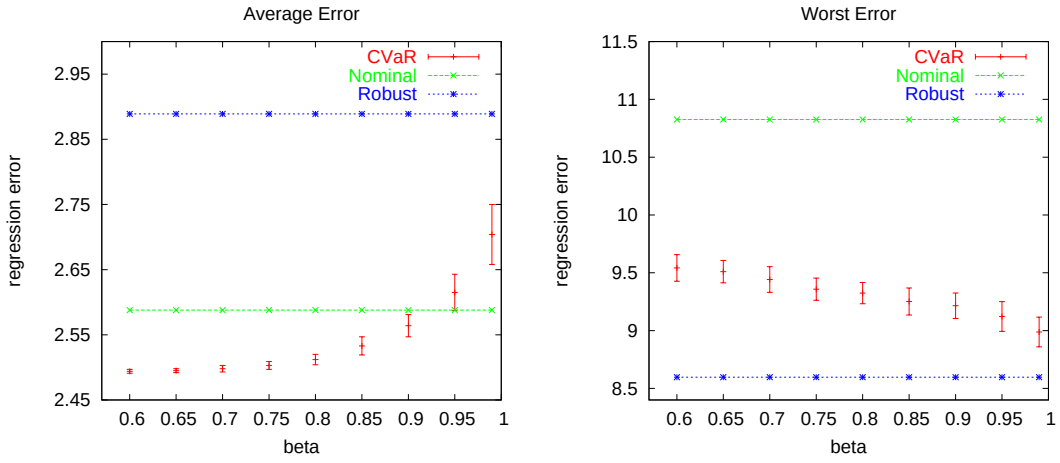

Figure 9: Estimation error plots with $D_i = 3I, i = 1, \ldots, m$.

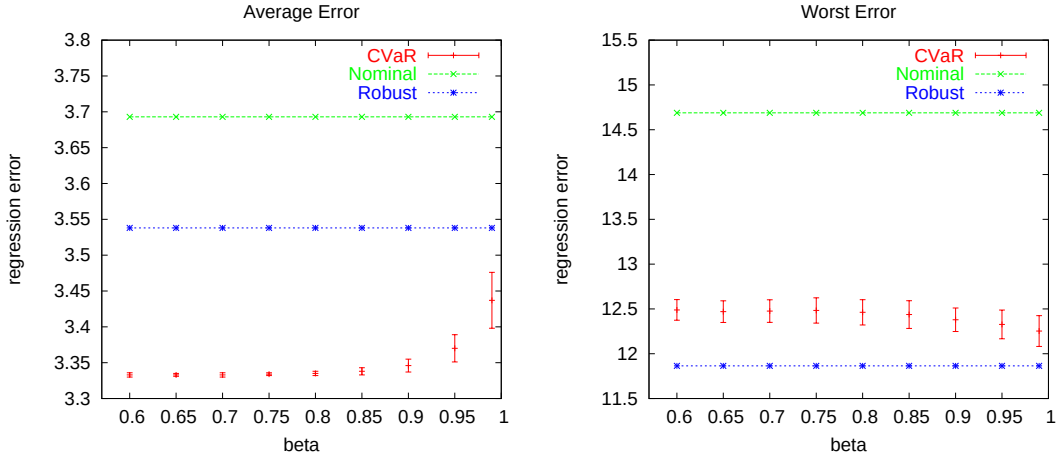

Figure 10: Estimation error plots with $D_i = 5I, i = 1, \ldots, m$.

References

[1] K.P. Bennett and O.L. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets," *Optimization Methods and Software*, **1** (1992) 23-34.

[2] A. Ben-Tal and A. Nemirovski, "Robust solutions of Linear Programming problems contaminated with uncertain data," *Mathematical Programming*, **88** (2000) 411-424.

[3] A. Ben-Tal and A. Nemirovski, "Robust solutions to uncertain linear programs," *Operations Research Letters,* **25** (1999) 1-13.

[4] A. Ben-Tal and A. Nemirovski, "Robust convex optimization," *Mathematics of Operations Research*, **23** (1998) 769-805.

[5] D. Bertsimas and D.B. Brown, "Robust linear optimization and coherent risk measures," working paper, http://web.mit.edu/dbbrown/www/papers/lids_2659.pdf, 2005.

[6] D. Bertsimas and M. Sim, "The price of robustness," *Operations Research*, **52** (2004) 35-53.

[7] C.L. Blake and C.J. Merz, UCI Repository of machine learning databases [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.

[8] G. Calafiore and M.C. Campi, "Uncertain convex programs: randomized solutions and confidence levels," *Mathematical Programming*, **102** (2005) 25-46.

[9] L. El-Ghaoui and H. Lebret, "Robust solutions to least-squares problems with uncertain data," *SIAM Journal on Matrix Analysis and Applications*, **18** (1997) 1035-1064.

[10] D. Goldfarb and G. Iyengar, "Robust convex quadratically constrained programs," *Mathematical Programming*, **97** (2003) 495-515.

[11] J. Gotoh and A. Takeda, "A Linear Classification Model Based on Conditional Geometric Score," *Pacific Journal of Optimization*, **1** (2005) 277-296.

[12] A. Nemirovski and A. Shapiro, "Convex approximations of chance constrained programs," *E-print:* http://www.optimization-online.org, 2004.

[13] R.T. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," *Journal of Banking and Finance*, **26** (2002) 1443-1471.

[14] A. Ruszczyński and A. Shapiro, *Stochastic Programming*, Elsevier Science B.V., (2003).

[15] B. Schölkopf, A. J. Smola, R. C. Williamson and P. L. Bartlett, "New support vector algorithms," *Neural Computation*, **12** (2000) 1207-1245.

[16] A.L. Soyster, "Convex programming with set-inclusive constraints and applications to inexact linear programming," *Operations Research*, **21** (1973) 1154-1157.

[17] J. F. Sturm: "Using SeDuMi 1.02, a Matlab Toolbox for Optimization over Symmetric Cones," *Optimization Methods and Software*, **11-12** (1999) 625-653.

[18] R. Tempo, G. Calafiore, and F. Dabbene: *Randomized Algorithms for Analysis and Control of Uncertain Systems*, Springer-Verlag London Limited, (2005).

[19] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1996.