Sparse Optimization with Risk Constraints Relaxation and its Application to Support Vector Machines

Zhicheng Zhang (D3)

Division of Operations Research (Fujisaki Lab)
Department of Information and Physical Sciences
Graduate School of Information Science and Technology
Osaka University

IST Summer School for Mathematical Optimization

Background

Occam's Razor

- Sparse Optimization always appears in compressed sensing (CS), optimal control (OC), machine learning (ML), and statistics, etc.

  e.g., model (variable) reduction, structural sparsity.

  - Methods : greedy algorithms, iterative reweighted least squares (IRLS), mixed-integer program (MIP), and submodular func., etc.

Proverb : *"There is nothing certain, but the uncertain."*

- Risk Assessment is closely related to probability and statistics.
  - Stochastic Programming (SP), e.g., $\mathbb{E}_{q\sim\mathbb{P}}[h(x,q)]$
  - Chance Constrained Optimization (CCP), e.g., SAA
  - Dristributionally Robust Optimization (DRO)

    e.g., worst case $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\{h(x,q)\leq 0\}\leq\epsilon$ (optimal transport)

Chance Constrained Sparse Optimization (Non-Convex Optim.)

Consider an exact sparse optimization with chance constraints as follows

$$(\text{CCSOP}^0_\epsilon) \qquad \min_{x \in \mathcal{X}} \quad \|x\|_0 \qquad \text{s.t.} \quad \mathbb{P}\{q \in \mathbb{Q} : h(x, q) \leq 0\} \geq 1 - \epsilon,$$

where $\mathcal{X} \subseteq \mathbb{R}^n$ be a compact convex set, and $\epsilon \in (0, 1)$ represents the risk (or constraint violation) level for the chance constrained framework.

- Sparsity : ℓ_0 quasi-norm of the vector $x \in \mathbb{R}^n$ depends on its support $\|x\|_0 = \text{supp}\{j : x_j \neq 0\}$ that counts the no. of nonzero elements.
- Probability : Denote $(\mathcal{Q}, \mathfrak{B}(\mathcal{Q}), \mathbb{P})$ be a probability space, where $\mathcal{Q}$ is a metric space w.r.t. Borel σ-algebra $\mathfrak{B}(\mathcal{Q})$.
- Uncertainty : A measurable *uncertain function* $h : \mathcal{X} \times \mathcal{Q} \to \mathbb{R}$, which is “convex” in x for each $q \in \mathcal{Q}$, and “bounded” in q for each $x \in \mathcal{X}$.

Sparse Optimization (II)

Chance Constrained Sparse Optimization (Non-Convex Optim.)

Consider an exact sparse optimization with chance constraints as follows

$$(\mathrm{CCSOP}^0_\epsilon) \qquad \min_{x \in \mathcal{X}} \quad \|x\|_0$$
$$\text{s.t.} \quad \mathbb{P}\{q \in \mathbb{Q} : h(x, q) \leq 0\} \geq 1 - \epsilon,$$

where $\mathcal{X} \subseteq \mathbb{R}^n$ be a compact convex set, and $\epsilon \in (0, 1)$ represents the risk (or constraint violation) level for the chance constrained framework.

Challenge :

- The ℓ_0 cost is "non-convex" and "non-smooth", leading to NP-hard.
- Risk constraint is related to the calculation of of multiple integrals.

Oracle :

- Direct : ℓ_1 norm convex relaxation for objective $\|x\|_1$.
- Data-driven sampling for uncertainty $\{q\}_{i=1}^N$, like Monte-Carlo.

Sparse Optimization with Risk Constraints Relaxation

Consider a sparse optimization with risk constraints relaxation as follows

$$(\mathrm{SSCOP}_N^{\rho}) \qquad \min_{x \in \mathcal{X}, \xi_i \geq 0} \quad \|x\|_1 + \rho \sum_{i=1}^{N} \xi_i$$
$$\text{subject to} \quad h(x, q_i) \leq \xi_i, \quad i = 1, \ldots, N,$$

★ Linear program (LP) reformulation by taking $x \doteq x^+ - x^-$, that is,

$$\|x\|_1 = \sum_{j=1}^{n} |x_j| = \sum_{j=1}^{n} \left(x_j^+ + x_j^- \right),$$
$$x_j^+ = \max\{x_j, 0\}, \quad x_j^- = \max\{-x_j, 0\}, \quad x_j^+, \ x_j^- \geq 0.$$

- Empirical risk : if takes replace weight ρ by $\dfrac{\rho}{N}$

Application : Support Vector Machine (SVM)

- SVM setup : Given a training data $q_i = \{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, N$ of i.i.d. outcomes from some probability $\mathbb{P}$, and true probability is not known.
- here $\mathbf{x}_i \subset \mathcal{X} \in \mathbb{R}^d$ are the feature vectors or samples.
- Two labels $y_i \in \{-1, +1\}$ represent different classifiers.
- Objective : Learn a linear classifier $\hat{y}_i = \text{sign}(\mathbf{x}_i^\top \beta + \beta_0)$, here $\beta_0 \in \mathbb{R}$ is referred to as the offset term.

Recap : Standard L2-SVM

Consider a standard L2 norm support vector machine as follows

$$
\text{(L2-SVM)} \quad \begin{aligned} \min_{\substack{\beta \in \mathbb{R}^p,\ \beta_0 \in \mathbb{R}, \\ \xi_i \geq 0,\ i \in [N]}} \quad & \lambda \|\beta\|_2^2 + \frac{\rho}{N} \sum_{i=1}^{N} \xi_i \\ \text{subject to} \quad & 1 - y_i \left(\mathbf{x}_i^\top \beta + \beta_0 \right) \leq \xi_i, \qquad i \in [N], \\ & \xi_i \geq 0. \end{aligned}
$$

Thinking : L1 norm Support Vector Machine

L1-SVM

To shrink the feature variables, we replace the ℓ_2 norm of the coefficients by using a convex surrogate ℓ_1 norm, resulting in L1-SVM

$$
\text{(L1-SVM)} \quad \begin{aligned}
\min_{\substack{\beta_0 \in \mathbb{R},\ \xi_i \geq 0, \\ \beta^+,\ \beta^- \in \mathbb{R}^p}} \quad & \lambda \sum_{j=1}^{p} \left(\beta_j^+ + \beta_j^-\right) + \rho \sum_{i=1}^{N} \xi_i \\
\text{subject to} \quad & \xi_i + y_i \mathbf{x}_i^\top (\beta^+ - \beta^-) + y_i \beta_0 \geq 1, && i \in [N] \\
& \xi_i \geq 0, \quad \beta_j^+ \geq 0, \quad \beta_j^- \geq 0, && i \in [N],\ j \in [p]
\end{aligned}
$$

- Sparse cost plays an important role in variable reduction, e.g., $\mathcal{J} \subseteq \mathcal{X}$, and $|\mathcal{J}| \leq p$.
- Provide probabilistic robustness guarantees for mis-classification.
- Design fast algorithms to obtain optimal solution (β^*, ξ^*).