Error Analysis of Spherically Constrained Least Squares Reformulation in Solving the Stackelberg Prediction Game

Xiyuan Li [1] Weiwei Liu [1]

[1]School of Computer Science, Wuhan University

NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

The Stackelberg prediction game (SPG) is a popular model for characterizing strategic interactions between a learner and an adversarial data provider. Although optimization problems in SPGs are often NP-hard, a notable special case involving the least squares loss (SPG-LS) has gained significant research attention recently [1, 2, 3]. The latest state-of-the-art method for solving the SPG-LS problem is the spherically constrained least squares reformulation (SCLS) method proposed in the work of [3]. However, the paper [3] lacks theoretical analysis on the error of the SCLS method, which limits its large-scale applications. In this paper, we investigate the estimation error between the learner obtained by the SCLS method and the actual learner. Specifically, we reframe the estimation error of the SCLS method as a Primary Optimization (**PO**) problem and utilize the Convex Gaussian min-max theorem (CGMT) to transform the **PO** problem into an Auxiliary Optimization (**AO**) problem. Subsequently, we provide a theoretical error analysis for the SCLS method based on this simplified **AO** problem. This analysis not only strengthens the theoretical framework of the SCLS method but also confirms the reliability of the learner produced by it. We further conduct experiments to validate our theorems, and the results are in excellent agreement with our theoretical predictions.

The Stackelberg prediction games

Stackelberg prediction games (SPGs) play prominent roles in various machine learning applications. SPG model is often formulated as a bi-level optimization problem, which is generally NP-hard even in the simplest case with linear constraints and objectives [1].

To overcome the NP-hard nature of SPGs, [1, 2, 3] focus on a commonly used subclass of SPGs, termed as the SPG-LS, whose loss functions for the learner and the data provider are least squares. Specifically, SPG-LS has access to a set of n sample tuples denoted by $S = \{(\boldsymbol{x}_i, y_i, z_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is input data with d features, y_i is the true output label of $\boldsymbol{x}_i$, and z_i is the label that the data provider aims to achieve. The learner of SPG-LS aims to train a linear predictor $\boldsymbol{w} \in \mathbb{R}^d$ to best estimate the true output label y_i of the fake data $\boldsymbol{x}_i^*$ by minimizing the least squares loss:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{w}^\top \boldsymbol{x}_i^* - y_i\|^2.$$

Meanwhile, the data provider of SPG-LS, with full knowledge of the learner's predictive model $\boldsymbol{w}$, selects the following least squares attacking strategy (i.e., modifying the data $\hat{\boldsymbol{x}}_i$) to make the corresponding prediction $\boldsymbol{w}^\top \boldsymbol{x}_i^*$ close to the desired label z_i:

$$\boldsymbol{x}_i^* = \arg\min_{\hat{\boldsymbol{x}}} \|\boldsymbol{w}^\top \hat{\boldsymbol{x}}_i - z_i\|^2 + \gamma \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2,$$

where $\gamma > 0$ is a regularizer to adjust the trade-off between the deviation from the original data $\boldsymbol{x}_i$ and closeness to the target z_i. Thus, the SPG-LS model can be expressed as the following bi-level optimization problem, as described in [1, 2, 3]:

$$\min_{\boldsymbol{w}} \|\boldsymbol{X}^*\boldsymbol{w} - \boldsymbol{y}\|^2, \quad \text{s.t. } \boldsymbol{X}^* = \arg\min_{\hat{\boldsymbol{X}}} \|\hat{\boldsymbol{X}}\boldsymbol{w} - \boldsymbol{z}\|^2 + \gamma\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_F^2, \tag{1}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n\times d}$ is the input sample matrix, $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^\top \in \mathbb{R}^n$ is the vector of true output labels, and $\boldsymbol{z} = (z_1, z_2, \cdots, z_n)^\top \in \mathbb{R}^n$ is the vector of labels that the attacker aims to achieve. Moreover, $\|\cdot\|$ denotes the Euclidean norm (l_2) unless otherwise specified.

Several studies have solved the SPG-LS (1). Recently, [3] proposes a spherically constrained least squares reformulation (SCLS) method and demonstrates that the SCLS method is currently the state-of-the-art for solving SPG-LS (1), having won the ICML 2022 Outstanding Paper Award.

Spherically Constrained Least Squares Reformulation (SCLS) method

Specifically, expanding upon previous studies by [1, 2], [3] reformulates SPG-LS (1) into the following optimization.

$$\inf_{\boldsymbol{w},\alpha} v(\boldsymbol{w}, \alpha) \triangleq \|\frac{\alpha\boldsymbol{z} + \boldsymbol{X}\boldsymbol{w}}{1+\alpha} - \boldsymbol{y}\|^2, \quad \text{s.t. } \boldsymbol{w}^\top\boldsymbol{w} = \gamma\alpha. \tag{2}$$

Subsequently, [3] makes an assumption on the nonemptiness of the optimal solution set of optimization (2).

Assumption 3.1 [3] *Assume that the optimal solution set of (2) is nonempty.*

Under Assumption 1, [3] employs a nonlinear variable transformation to recast the QFP (2) as a spherical constrained least squares (SCLS) problem:

$$\min_{\tilde{\boldsymbol{w}},\tilde{\alpha}} \tilde{v}(\tilde{\boldsymbol{w}}, \tilde{\alpha}) \triangleq \left\|\frac{\tilde{\alpha}}{2}\boldsymbol{z} + \frac{\sqrt{\gamma}}{2}\boldsymbol{X}\tilde{\boldsymbol{w}} - (\boldsymbol{y} - \frac{\boldsymbol{z}}{2})\right\|^2, \quad \text{s.t. } \tilde{\boldsymbol{w}}^\top\tilde{\boldsymbol{w}} + \tilde{\alpha}^2 = 1, \tag{3}$$

where $\tilde{\boldsymbol{w}}$ and $\tilde{\alpha}$ are defined in Lemmas 3.2 and 3.3.

Lemma 3.2 [3] *Suppose $(\boldsymbol{w}, \alpha)$ is a feasible solution of QFP (2). Then $(\tilde{\boldsymbol{w}}, \tilde{\alpha})$, defined as*

$$\tilde{\boldsymbol{w}} \coloneqq \frac{2}{\sqrt{\gamma}(\alpha+1)}\boldsymbol{w} \quad \text{and} \quad \tilde{\alpha} \coloneqq \frac{\alpha-1}{\alpha+1}, \tag{4}$$

is feasible to SCLS (3) and $v(\boldsymbol{w}, \alpha) = \tilde{v}(\tilde{\boldsymbol{w}}, \tilde{\alpha})$.

Lemma 3.3 [3] *Suppose $(\tilde{\boldsymbol{w}}, \tilde{\alpha})$ is feasible to SCLS (3) with $\tilde{\alpha} \neq 1$. Then $(\boldsymbol{w}, \alpha)$, defined as*

$$\boldsymbol{w} \coloneqq \frac{\sqrt{\gamma}}{1-\tilde{\alpha}}\tilde{\boldsymbol{w}} \quad \text{and} \quad \alpha \coloneqq \frac{1+\tilde{\alpha}}{1-\tilde{\alpha}}, \tag{5}$$

is feasible to QFP (2) and $\tilde{v}(\tilde{\boldsymbol{w}}, \tilde{\alpha}) = v(\boldsymbol{w}, \alpha)$.

Let v^* and $\tilde{v}^*$ represent the optimal values of QFP (2) and SCLS (3), respectively. Subsequently, [3] presents Theorem 3.4 to elucidate the relationship between the solutions of QFP (2) and SCLS (3).

Theorem 3.4 [3] *Given Assumption 1, then there exists an optimal solution $(\tilde{\boldsymbol{w}}, \tilde{\alpha})$ to SCLS (3) with $\tilde{\alpha} \neq 1$. Moreover, $(\boldsymbol{w}, \alpha)$, defined by (5), is an optimal solution to (2) and $v^* = v(\boldsymbol{w}, \alpha) = \tilde{v}(\tilde{\boldsymbol{w}}, \tilde{\alpha}) = \tilde{v}^*$.*

The Error Analysis for the SCLS method

We investigate the estimation error between the learner (e.t.$\boldsymbol{w}^*$) estimated by the SCLS method and the true learner (denoted as $\boldsymbol{w}_0$) to validate the reliability of $\boldsymbol{w}^*$. Specifically, we assume the samples $S = \{(\boldsymbol{x}_i, y_i, z_i)\}_{i=1}^n$ are generated by the following black box model:

$$\boldsymbol{X}^* = \arg\min_{\hat{\boldsymbol{X}}} \|\hat{\boldsymbol{X}}\boldsymbol{w}_0 - \boldsymbol{z}\|^2 + \gamma\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_F^2, \quad \boldsymbol{y} = \boldsymbol{X}^*\boldsymbol{w}_0 + \boldsymbol{\epsilon}, \tag{6}$$

where $\boldsymbol{w}_0 \in \mathbb{R}^d$ represents the "true" weight parameter of the real learner, and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^\top \in \mathbb{R}^n$ is the noise vector. Moreover, the entries of $\boldsymbol{X}$ and $\boldsymbol{z}$ are drawn i.i.d. from $\mathcal{N}(0,1)$; the entries of $\boldsymbol{\epsilon}$ are drawn i.i.d. from $\mathcal{N}(0, \sigma^2)$; and we assume $\lim_{n\to\infty}\frac{d}{n} \in (0,1)$.

Given $\boldsymbol{X}$, $\boldsymbol{z}$, and $\boldsymbol{y}$ generated by this model (6), we solve SPG-LS (1) by the SCLS method to obtain $\boldsymbol{w}^*$ that is used to estimate the target vector $\boldsymbol{w}_0$. Our task is to measure the optimal estimation error of the SCLS method, represented by $\|\boldsymbol{w}^* - \boldsymbol{w}_0\|$.

Because the sample $(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{z})$ is generated by black box model (6), we have:

$$\boldsymbol{y} = \boldsymbol{X}^*\boldsymbol{w}_0 + \boldsymbol{\epsilon} = \frac{\alpha_0\boldsymbol{z} + \boldsymbol{X}\boldsymbol{w}_0}{1+\alpha_0} + \boldsymbol{\epsilon}, \tag{7}$$

where $\alpha_0 = \boldsymbol{w}_0^\top\boldsymbol{w}_0/\gamma$. Taking (7) to SCLS problem (3) simplifies to:

$$\min_{\tilde{\boldsymbol{\beta}}} \frac{1}{n}\left\|\boldsymbol{c}^\top\tilde{\boldsymbol{\beta}}\boldsymbol{z} + \boldsymbol{X}\tilde{\boldsymbol{\beta}} - \frac{2\boldsymbol{\epsilon}}{\sqrt{\gamma}}\right\|^2. \tag{8}$$

where $\tilde{\boldsymbol{\beta}} \coloneqq \tilde{\boldsymbol{w}} - \tilde{\boldsymbol{w}}_0$, and $\boldsymbol{c} \coloneqq \boldsymbol{c}(\tilde{\boldsymbol{w}}_0, \gamma)$. The **PO** problem associated with (8) is:

$$\Phi_{\text{SCLS}}(\boldsymbol{X}) = \min_{\tilde{\boldsymbol{\beta}}}\max_{\boldsymbol{u}} \frac{1}{n}\big(\boldsymbol{u}^\top\boldsymbol{X}\tilde{\boldsymbol{\beta}} + \psi(\tilde{\boldsymbol{\beta}}, \boldsymbol{u})\big), \tag{9}$$

where $\psi(\tilde{\boldsymbol{\beta}}, \boldsymbol{u}) \coloneqq \boldsymbol{c}^\top\tilde{\boldsymbol{\beta}}\cdot\boldsymbol{u}^\top\boldsymbol{z} - \frac{2\boldsymbol{u}^\top\boldsymbol{\epsilon}}{\sqrt{\gamma}} - \frac{\|\boldsymbol{u}\|^2}{4}$. The **PO** (9) can be simplified as **AO** using CGMT:

$$\phi_{\text{SCLS}}(\boldsymbol{g}, \boldsymbol{h}) = \min_{\tilde{\boldsymbol{\beta}}}\max_{\boldsymbol{u}} \frac{1}{n}\big[(\|\tilde{\boldsymbol{\beta}}\|\boldsymbol{g} + \boldsymbol{c}^\top\tilde{\boldsymbol{\beta}}\boldsymbol{z} - \frac{2\boldsymbol{\epsilon}}{\sqrt{\gamma}})^\top\boldsymbol{u} + \|\boldsymbol{u}\|\boldsymbol{h}^\top\tilde{\boldsymbol{\beta}} - \frac{\|\boldsymbol{u}\|^2}{4}\big], \tag{10}$$

where the entries of $\boldsymbol{g}$ and $\boldsymbol{h}$ are drawn i.i.d. from $\mathcal{N}(0,1)$. As n goes to $+\infty$, the optimal minimizer of **AO** (10) converges to the optimal minimizer of (11) in probability:

$$\min_{\tilde{\boldsymbol{\beta}}} \|\tilde{\boldsymbol{\beta}}\|^2 + (\boldsymbol{c}^\top\tilde{\boldsymbol{\beta}})^2 + \Omega(\tilde{\boldsymbol{\beta}}) + \frac{4\sigma^2}{\gamma}. \tag{11}$$

Here, we successfully reduced the complex **AO** problem (10) to a more manageable deterministic optimization problem (11), effectively focusing only on the estimation error variable $\tilde{\boldsymbol{\beta}}$.

Theorem 4.1 *Suppose $\tilde{\boldsymbol{w}}_0$ is the true weight parameter of the original SCLS problem (3), and $\tilde{\boldsymbol{w}}^*$ is the optimal solution to the objective function of SCLS (3). If $\lim_{n\to\infty}\frac{d}{n} \in (0,1)$, the estimation error of SCLS (3) is given by the following probability limit: $\lim_{n\to\infty}\|\tilde{\boldsymbol{w}}^* - \tilde{\boldsymbol{w}}_0\| \xrightarrow{P} 0$.*

Theorem 4.3 *Suppose $\boldsymbol{w}_0$ is the true weight parameter of the SPG-LS (1), $\tilde{\boldsymbol{w}}^*$ is the optimal solution learned by SCLS (3), and $\boldsymbol{w}^*$ is the optimal solution recovered from $\tilde{\boldsymbol{w}}^*$ by Theorem 1. If $\lim_{n\to\infty}\frac{d}{n} \in (0,1)$, the estimation error of SPG-LS (1) solved by the SCLS (3) is given by the following probability limit:*

$$\lim_{n\to\infty}\|\boldsymbol{w}^* - \boldsymbol{w}_0\| \xrightarrow{P} 0.$$

Experiments

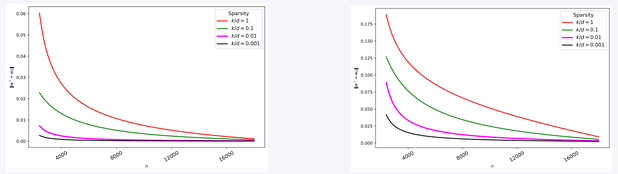

Figure 1. The change of $\|\boldsymbol{w}^* - \boldsymbol{w}_0\|$ with n for SCLS method under different Sparsity k/d.

References

[1] Nick Bishop, Long Tran-Thanh, and Enrico H. Gerding. Optimal learning from verified training data. In *NeurIPS*, 2020.

[2] Jiali Wang, He Chen, Rujun Jiang, Xudong Li, and Zihao Li. Fast algorithms for stackelberg prediction game with least squares loss. In *ICML*, volume 139, pages 10708–10716, 2021.

[3] Jiali Wang, Wen Huang, Rujun Jiang, Xudong Li, and Alex L. Wang. Solving stackelberg prediction game with least squares loss via spherically constrained least squares reformulation. In *ICML*, volume 162, pages 22665–22679, 2022.

Emails: Lee_xiyuan@outlook.com; liuweiwei863@gmail.com