Learning Relations From Data With Conditional Gradients

Elias S. Wirth

Technische Universität Berlin
and
Zuse Institute Berlin

wirth@zib.de

ZIB Meeting · November 19, 2020

Importance of Feature Extraction

Non-Noisy Data [6]

Setup

Let $X = \{x_1, \ldots, x_s\} \subseteq \mathbb{R}^n$ be a data set. Consider the ideal

$$G := \{g \in \mathbb{R}[x_1, \ldots, x_n] \mid g(x) = 0 \text{ for all } x \in X\}.$$

Non-Noisy Data [6]

Setup

Let $X = \{x_1, \ldots, x_s\} \subseteq \mathbb{R}^n$ be a data set. Consider the ideal

$$G := \{g \in \mathbb{R}[x_1, \ldots, x_n] \mid g(x) = 0 \text{ for all } x \in X\}.$$

Multiclass Classification

- Data set $X = \{x_1, \ldots, x_s\} \subseteq \mathbb{R}^n$ and labels $Y = (y_1, \ldots, y_s)^T \subseteq \{1, \ldots, k\}^s$

Non-Noisy Data [6]

Setup

Let $X = \{x_1, \ldots, x_s\} \subseteq \mathbb{R}^n$ be a data set. Consider the ideal

$$G := \{g \in \mathbb{R}[x_1, \ldots, x_n] \mid g(x) = 0 \text{ for all } x \in X\}.$$

Multiclass Classification

- Data set $X = \{x_1, \ldots, x_s\} \subseteq \mathbb{R}^n$ and labels $Y = (y_1, \ldots, y_s)^T \subseteq \{1, \ldots, k\}^s$
- Let $G = \bigcup_{i=1}^k G_i = \{g_1, \ldots, g_t\}$.

Non-Noisy Data [6]

Setup

Let $X = \{x_1, \ldots, x_s\} \subseteq \mathbb{R}^n$ be a data set. Consider the ideal

$$G := \{g \in \mathbb{R}[x_1, \ldots, x_n] \mid g(x) = 0 \text{ for all } x \in X\}.$$

Multiclass Classification

- Data set $X = \{x_1, \ldots, x_s\} \subseteq \mathbb{R}^n$ and labels $Y = (y_1, \ldots, y_s)^T \subseteq \{1, \ldots, k\}^s$
- Let $G = \bigcup_{i=1}^k G_i = \{g_1, \ldots, g_t\}$.

$$G(X) := \begin{bmatrix} - & G(x_1) & - \\ & \vdots & \\ - & G(x_s) & - \end{bmatrix} = \begin{bmatrix} g_1(x_1) & g_2(x_1) & \ldots & g_t(x_1) \\ \vdots & \vdots & & \vdots \\ g_1(x_s) & g_2(x_s) & \ldots & g_t(x_s) \end{bmatrix} \in \mathrm{Mat}_{s,t}(\mathbb{R}).$$

Non-Noisy Data [6]

Setup

Let $X = \{x_1, \ldots, x_s\} \subseteq \mathbb{R}^n$ be a data set. Consider the ideal

$$\mathcal{G} := \{g \in \mathbb{R}[x_1, \ldots, x_n] \mid g(x) = 0 \text{ for all } x \in X\}.$$

Multiclass Classification

- Data set $X = \{x_1, \ldots, x_s\} \subseteq \mathbb{R}^n$ and labels $Y = (y_1, \ldots, y_s)^T \subseteq \{1, \ldots, k\}^s$
- Let $\mathcal{G} = \bigcup_{i=1}^k \mathcal{G}_i = \{g_1, \ldots, g_t\}$.

$$\mathcal{G}(X) := \begin{bmatrix} - & \mathcal{G}(x_1) & - \\ & \vdots & \\ - & \mathcal{G}(x_s) & - \end{bmatrix} = \begin{bmatrix} g_1(x_1) & g_2(x_1) & \ldots & g_t(x_1) \\ \vdots & \vdots & & \vdots \\ g_1(x_s) & g_2(x_s) & \ldots & g_t(x_s) \end{bmatrix} \in \text{Mat}_{s,t}(\mathbb{R}).$$

- Train a classifier on $\mathcal{G}(X)$ and Y.

Setting: Noisy Data

Setting: Noisy Data

Let $X = \{x_1, \ldots, x_s\} \subseteq \mathbb{R}^n$ be a data set and $f \in \mathbb{R}[x_1, \ldots, x_n]$.

Setting: Noisy Data

Let $X = \{x_1, \ldots, x_s\} \subseteq \mathbb{R}^n$ be a data set and $f \in \mathbb{R}[x_1, \ldots, x_n]$.

Definition (Evaluation)

The evaluation of f over X is defined as $\mathrm{eval}_X(f) := (f(x_1), \ldots, f(x_s))^T$.

Setting: Noisy Data

Let $X = \{x_1, \ldots, x_s\} \subseteq \mathbb{R}^n$ be a data set and $f \in \mathbb{R}[x_1, \ldots, x_n]$.

Definition (Evaluation)

The evaluation of f over X is defined as $\text{eval}_X(f) := (f(x_1), \ldots, f(x_s))^T$.

Definition (Root Mean Square Error)

Define the root mean square error of f over X as
$\text{rmse}(f, X) := \sqrt{\frac{1}{s}\|\text{eval}_X(f)\|_2^2}$.

Setting: Noisy Data

Let $X = \{x_1, \ldots, x_s\} \subseteq \mathbb{R}^n$ be a data set and $f \in \mathbb{R}[x_1, \ldots, x_n]$.

Definition (Evaluation)

The evaluation of f over X is defined as $\text{eval}_X(f) := (f(x_1), \ldots, f(x_s))^T$.

Definition (Root Mean Square Error)

Define the root mean square error of f over X as $\text{rmse}(f, X) := \sqrt{\frac{1}{s}\| \text{eval}_X(f)\|_2^2}$.

Definition (ψ-Approximate Vanishing Ideal)

An ideal $G \subseteq \mathbb{R}[x_1, \ldots, x_n]$ is ψ-approximately vanishing if G is generated by a set of unitary polynomials $f_1, \ldots, f_k$ of G that satisfy $\text{rmse}(f_i, X) \leq \psi$.

Problem Setting

Problem Setting

Given a data set $X = \{x_1, \ldots, x_s\} \subseteq [-1, 1]^n$ and $\psi > 0$, construct a set of unitary polynomials, G, such that

- G generates a ψ-approximately vanishing ideal,
- any ψ-approximately vanishing polynomial $g \in \mathbb{R}[x_1, \ldots, x_n]$ is contained in $\langle G \rangle_{\mathbb{R}[x_1,\ldots,x_n]}$.

How Do We Construct G?

Definition (Border)

Let $\mathcal{O} \subseteq \mathbb{R}[x_1, \ldots, x_n]$ be a set of monomials. A monomial $t \in \mathbb{R}[x_1, \ldots, x_n] \setminus \mathcal{O}$ is a border term of $\mathcal{O}$ if all divisors of t are in $\mathcal{O}$. The set of all degree d border terms of $\mathcal{O}$ is denoted by $\partial\mathcal{O}^d$.

How Do We Construct G?

Definition (Border)

Let $\mathcal{O} \subseteq \mathbb{R}[x_1, \ldots, x_n]$ be a set of monomials. A monomial $t \in \mathbb{R}[x_1, \ldots, x_n] \setminus \mathcal{O}$ is a border term of $\mathcal{O}$ if all divisors of t are in $\mathcal{O}$. The set of all degree d border terms of $\mathcal{O}$ is denoted by $\partial\mathcal{O}^d$.

E.g.: For $\mathcal{O} = \{1, x_1, x_2, x_3, x_1x_2, x_2^2, x_3^2\}$, we have $\partial\mathcal{O}^3 = \{x_1x_2^2, x_2^3, x_2^2x_3, x_2x_3^2, x_3^3\}$. Note that $x_1^2x_2 \notin \partial\mathcal{O}^d$.

How Do We Construct G?

Definition (Border)

Let $\mathcal{O} \subseteq \mathbb{R}[x_1, \ldots, x_n]$ be a set of monomials. A monomial $t \in \mathbb{R}[x_1, \ldots, x_n] \setminus \mathcal{O}$ is a border term of $\mathcal{O}$ if all divisors of t are in $\mathcal{O}$. The set of all degree d border terms of $\mathcal{O}$ is denoted by $\partial\mathcal{O}^d$.

E.g.: For $\mathcal{O} = \{1, x_1, x_2, x_3, x_1x_2, x_2^2, x_3^2\}$, we have $\partial\mathcal{O}^3 = \{x_1x_2^2, x_2^3, x_2^2x_3, x_2x_3^2, x_3^3\}$. Note that $x_1^2x_2 \notin \partial\mathcal{O}^d$.

ψ-Approximately Vanishing Polynomial Oracle (AVPO)

Input: A data set $X \subseteq \mathbb{R}^n$ and a set of unitary monomials $\mathcal{O}$.
Output: If a unitary ψ-approximately vanishing polynomial g with terms only in $\mathcal{O}$ exists, returns g. Else, returns any unitary polynomial with terms only in $\mathcal{O}$.

How Do We Construct G?

Definition (Border)

Let $\mathcal{O} \subseteq \mathbb{R}[x_1, \ldots, x_n]$ be a set of monomials. A monomial $t \in \mathbb{R}[x_1, \ldots, x_n] \setminus \mathcal{O}$ is a border term of $\mathcal{O}$ if all divisors of t are in $\mathcal{O}$. The set of all degree d border terms of $\mathcal{O}$ is denoted by $\partial\mathcal{O}^d$.

E.g.: For $\mathcal{O} = \{1, x_1, x_2, x_3, x_1x_2, x_2^2, x_3^2\}$, we have $\partial\mathcal{O}^3 = \{x_1x_2^2, x_2^3, x_2^2x_3, x_2x_3^2, x_3^3\}$. Note that $x_1^2x_2 \notin \partial\mathcal{O}^d$.

ψ-Approximately Vanishing Polynomial Oracle (AVPO)

Input: A data set $X \subseteq \mathbb{R}^n$ and a set of unitary monomials $\mathcal{O}$.
Output: If a unitary ψ-approximately vanishing polynomial g with terms only in $\mathcal{O}$ exists, returns g. Else, returns any unitary polynomial with terms only in $\mathcal{O}$.

A call to this oracle is denoted by $\mathrm{AVPO}(X, \mathcal{O})$.

Algorithm Approximate Vanishing Ideal Algorithm Template [4]

Input: A set $X = \{x_1, \ldots, x_s\} \subseteq [-1, 1]^n$ and $\psi > 0$.

Output: A set of unitary polynomials G that generates a ψ-approximate vanishing ideal of X.

```
d ← 1
O ← {1}
G ← ∅
while ∂O^d ≠ ∅ do
    L ← ∂O^d
    for t ∈ L do
        g ← AVPO(X, O ∪ {t})
        if rmse(g, X) ≤ ψ then
            G ← G ∪ {g}
        else
            O ← O ∪ {t}
        end if
    end for
    d ← d + 1
end while
```

Theoretical Guarantees: AVI

AVPO

- Singular Value Decomposition for AVI.

Theoretical Guarantees: AVI

AVPO

- Singular Value Decomposition for AVI.

Result	AVI
Maximality of G	🟢
Applicable to non-homogeneous relations	🔴
Correct leading term	🔴
Generalization bounds	🔴
Sparse polynomials	🔴

Algorithm Approximate Vanishing Ideal Algorithm[4]

Input: A set $X = \{x_1, \ldots, x_s\} \subseteq [-1, 1]^n$ and $\psi > 0$.

Output: A set of unitary polynomials G that generates a ψ-approximate vanishing ideal of X.

$d \leftarrow 1$
$\mathcal{O} \leftarrow \{1\}$
$G \leftarrow \emptyset$
while $\partial \mathcal{O}^d \neq \emptyset$ **do**
    $L \leftarrow \partial \mathcal{O}^d$
    for $t \in L$ **do**
        $g \leftarrow \text{AVPO}(X, \mathcal{O} \cup \{t\})$
        if $\text{rmse}(g, X) \leq \psi$ **then**
            $G \leftarrow G \cup \{g\}$
        else
            $\mathcal{O} \leftarrow \mathcal{O} \cup \{t\}$
        end if
    end for
    $d \leftarrow d + 1$
end while

Replacing the Singular Value Decomposition

Step of Interest

$$g \leftarrow \text{AVPO}(X, \mathcal{O} \cup \{t\})$$

Replacing the Singular Value Decomposition

Step of Interest

$$g \leftarrow \text{AVPO}(X, \mathcal{O} \cup \{t\})$$

Recall: ψ-Approximately Vanishing Polynomial Oracle (AVPO)

Input: A data set $X \subseteq \mathbb{R}^n$ and a set of unitary monomials $\mathcal{O}$.
Output: If a unitary ψ-approximately vanishing polynomial g with terms only in $\mathcal{O}$ exists, returns g. Else, returns any unitary polynomial with terms only in $\mathcal{O}$.

Replacing the Singular Value Decomposition

Step of Interest

$$g \leftarrow \mathrm{AVPO}(X, \mathcal{O} \cup \{t\})$$

Recall: ψ-Approximately Vanishing Polynomial Oracle (AVPO)

Input: A data set $X \subseteq \mathbb{R}^n$ and a set of unitary monomials $\mathcal{O}$.
Output: If a unitary ψ-approximately vanishing polynomial g with terms only in $\mathcal{O}$ exists, returns g. Else, returns any unitary polynomial with terms only in $\mathcal{O}$.

Notation

- Denote the evaluation matrix of $\mathcal{O}$ by A.
- Let $y := \mathrm{eval}_X(t)$.

Conditional Gradient Approximate Vanishing Ideal Algorithm (CGAVI)

Observation

- A unitary ψ-approximately vanishing polynomial exists iff

$$\min_{x \in \mathbb{R}^{|\mathcal{O}|}} \sqrt{\frac{1}{s}\|Ax - y\|_2^2} \leq \psi.$$

Conditional Gradient Approximate Vanishing Ideal Algorithm (CGAVI)

Observation

- A unitary ψ-approximately vanishing polynomial exists iff

$$\min_{x \in \mathbb{R}^{|\mathcal{O}|}} \sqrt{\frac{1}{s}\|Ax - y\|_2^2} \leq \psi.$$

Adaptation

- Limit size of feasibility region to bound leading term coefficient.

Conditional Gradient Approximate Vanishing Ideal Algorithm (CGAVI)

Observation

- A unitary ψ-approximately vanishing polynomial exists iff

$$\min_{x \in \mathbb{R}^{|\mathcal{O}|}} \sqrt{\frac{1}{s}\|Ax - y\|_2^2} \leq \psi.$$

Adaptation

- Limit size of feasibility region to bound leading term coefficient.
- Least Squares loss (smooth, (strongly) convex).

Conditional Gradient Approximate Vanishing Ideal Algorithm (CGAVI)

Observation

- A unitary ψ-approximately vanishing polynomial exists iff

$$\min_{x \in \mathbb{R}^{|\mathcal{O}|}} \sqrt{\frac{1}{s}\|Ax - y\|_2^2} \leq \psi.$$

Adaptation

- Limit size of feasibility region to bound leading term coefficient.
- Least Squares loss (smooth, (strongly) convex).
- We thus solve

$$\min \frac{1}{s}\|Ax - y\|_2^2,$$
$$\text{such that } \|x\|_1 \leq D.$$

Theoretical Guarantees: AVI vs. CGAVI

AVPO

- Singular Value Decomposition for AVI
- Conditional Gradients a.k.a. Frank-Wolfe [2, 5] for CGAVI.

Result	AVI	CGAVI
Maximality of G	🟢	🟢
Applicable to non-homogeneous relations	🔴	🟢
Correct leading term	🔴	🟢
Generalization bounds	🔴	🟢
Sparse polynomials	🔴	🟢

Conditional Gradient Algorithms

Conditional Gradients Approximate Vanishing Ideal Algorithm (CGAVI)

- Homogeneous problem setup
- Feature extraction for classification

Conditional Gradient Algorithms

Conditional Gradients Approximate Vanishing Ideal Algorithm (CGAVI)

- Homogeneous problem setup
- Feature extraction for classification

Conditional Gradient Learner Algorithm (CGL)

- Non-homogeneous problem setup
- (Feature extraction for) regression

Conditional Gradient Algorithms

Conditional Gradients Approximate Vanishing Ideal Algorithm (CGAVI)

- Homogeneous problem setup
- Feature extraction for classification

Conditional Gradient Learner Algorithm (CGL)

- Non-homogeneous problem setup
- (Feature extraction for) regression

Conditional Gradient Identification Of Equations From Data Algorithm (CGIED)

- Combination of CGAVI + CGL for regression tasks
- Sparse Identification of Nonlinear Dynamics Algorithm (SINDy) [1]

Experiments 1

Table: Results 1. For cancer and fashion, the evaluation metric is test set classification error in percent and for mpg, the evaluation metric is test set root mean square error.

algorithm	cancer	fashion	mpg
CGAVI + SVM	1.678	2.160	–
CGIED	–	–	2.772
AVI + SVM	2.168	2.258	–
CNN	–	1.735	–
DNN	–	–	3.048
SVM	4.649	2.260	–
SVR	–	–	2.577

Experiments 2

Table: Results 2. Results for the noisy Fermi-Pasta-Ulam-Tsingou problem [3].

terms/runs	actual	1	2	3
x_1^3	0.700	0.541	0.566	0.557
$x_1^2x_2$	-2.100	-2.013	-1.957	-1.992
$x_1x_2^2$	2.100	1.997	1.921	2.006
x_2^3	-1.400	-1.074	-1.115	-1.112
$x_2^2x_3$	2.100	2.003	1.936	1.992
$x_2x_3^2$	-2.100	-1.996	-1.918	-1.993
x_3^3	0.700	0.567	0.551	0.549
x_1	1.000	1.105	1.107	1.08
x_2	-2.000	-2.207	-2.215	-2.192
x_3	1.000	1.082	1.11	1.096

References I

S. L. Brunton, J. L. Proctor, and J. N. Kutz.
Discovering governing equations from data by sparse identification of nonlinear dynamical systems.
Proceedings of the national academy of sciences, 113(15):3932–3937, 2016.

M. Frank and P. Wolfe.
An algorithm for quadratic programming.
Naval research logistics quarterly, 3(1-2):95–110, 1956.

P. Gelß, S. Klus, J. Eisert, and C. Schütte.
Multidimensional approximation of nonlinear dynamical systems.
Journal of Computational and Nonlinear Dynamics, 14(6), 2019.

D. Heldt, M. Kreuzer, S. Pokutta, and H. Poulisse.
Approximate computation of zero-dimensional polynomial ideals.
Journal of Symbolic Computation, 44(11):1566–1591, 2009.

References II

E. S. Levitin and B. T. Polyak.
Constrained minimization methods.
USSR Computational mathematics and mathematical physics, 6(5):1–50, 1966.

H. M. Möller and B. Buchberger.
The construction of multivariate polynomials with preassigned zeros.
In European Computer Algebra Conference, pages 24–31. Springer, 1982.