

# Learning Models over Relational Data using Sparse Tensors and Functional Dependencies

Mahmoud Abo Khamis  
RelationalAI, Inc.

Hung Q. Ngo  
RelationalAI, Inc.

XuanLong Nguyen  
University of Michigan

Dan Olteanu  
University of Oxford

Maximilian Schleich  
University of Oxford

## Abstract

Integrated solutions for analytics over relational databases are of great practical importance as they avoid the costly repeated loop data scientists have to deal with on a daily basis: select features from data residing in relational databases using feature extraction queries involving joins, projections, and aggregations; export the training dataset defined by such queries; convert this dataset into the format of an external learning tool; and train the desired model using this tool. These integrated solutions are also a fertile ground of theoretically fundamental and challenging problems at the intersection of relational and statistical data models.

This article introduces a unified framework for training and evaluating a class of statistical learning models over relational databases. This class includes ridge linear regression, polynomial regression, factorization machines, and principal component analysis. We show that, by synergizing key tools from database theory such as schema information, query structure, functional dependencies, recent advances in query evaluation algorithms, and from linear algebra such as tensor and matrix operations, one can formulate relational analytics problems and design efficient (query and data) structure-aware algorithms to solve them.

This theoretical development informed the design and implementation of the AC/DC system for structure-aware learning. We benchmark the performance of AC/DC against R, MADlib, libFM, and TensorFlow. For typical retail forecasting and advertisement planning applications, AC/DC can learn polynomial regression models and factorization machines with at least the same accuracy as its competitors and up to three orders of magnitude faster than its competitors whenever they do not run out of memory, exceed 24-hour timeout, or encounter internal design limitations.

## 1 Introduction

Although both disciplines of databases and statistics occupy foundational roles for the emerging field of data science, they are largely seen as complementary. Most fundamental contributions made by statisticians and machine learning researchers are abstracted away from the underlying infrastructure for data management. However, there is undoubtedly clear value in tight integration of statistics and database models and techniques. This is receiving an increasing interest in both academia and industry [2, 42, 61]. This is motivated by the realization that in many practical cases data used for training resides inside relational databases and bringing the analytics closer to the data saves non-trivial time usually spent on data import/export at the interface between database systems and statistical packages [38]. A complementary realization is that large chunks of statistical machine learning code can be expressed as relational queries and computed using database techniques [27, 43, 66, 24].

The problem of solving analytics over databases naturally lends itself to a systematic investigation using the toolbox of concepts and techniques developed by the database theorist, and by synergizing ideas from both relational and statistical data modeling. One can exploit database schema information, functional dependencies, state-of-the-art query evaluation algorithms, and well-understood complexity analysis.

## Contributions

Our *conceptual contribution* is the introduction of a framework for training and evaluating a class of statistical learning models over relational databases. This class, commonly used in retail-planning and forecasting applications [11], includes ridge linear regression, polynomial regression, factorization machines, and principal component analysis. In such applications, the training dataset is the result of a feature extraction query over the database. Typical databases include weekly sales data, promotions, and product descriptions. A retailer would like to compute a parameterized model, which can predict, for instance, the additional demand generated for a given product due to promotion [50]. The feature extraction query is commonly a natural join of the database relations, yet it may join additional relations derived from the input ones using aggregation. The features correspond to database attributes, their categorical values, or aggregates over them. As is prevalent in practical machine learning, the models are trained using a first-order optimization algorithm such as batch or stochastic gradient descent, in part because their convergence rates are dimension-free (for well-behaved objectives). This is a crucial property given the high-dimensionality of our problem as elaborated next.

The main *computational challenge* posed by analytics over databases is the large number of records and of features in the training dataset. There are two types of features: continuous (quantitative) such as price and sales; and categorical (qualitative) such as colors, cities, and countries.<sup>1</sup> While continuous features allow for aggregation over their domains, categorical features cannot be aggregated together. To accommodate the latter, the state-of-the-art approach is to one-hot encode their active domain: each value in the active domain of an attribute is encoded by an indicator vector whose dimension is the size of the domain. For instance, the colors in the domain {red, green, blue} can be represented by indicator vectors [1, 0, 0] for red, [0, 1, 0] for green, and [0, 0, 1] for blue. The one-hot encoding amounts to a relational representation of the training dataset with one new attribute per distinct category of each categorical feature and with wide tuples whose values are mostly 0. This entails huge redundancy due to the presence of the many 0 values. The one-hot encoding also blurs the usual distinction between schema and data, since the schema can become as large as the input database.

Closely related to the computational challenge is a *cultural challenge*: the feasibility of a tight integration of analytics and databases may be called into question. In terms of pure algorithmic performance, why would such an approach be more efficient than the common approach that decouples the computation of the training dataset from the learning task, given the widely available plethora of tools and techniques for the latter?

Our answer to these challenges is that, for a large class of feature extraction queries, it is possible to train a model in time *sub-linear* in the output size of the feature extraction query! This makes our approach competitive *regardless of the learning techniques* used by the mainstream approaches that first materialize the training dataset, including those that use sampling and stochastic gradient descent to only process a subset of the training dataset.

More concretely, our approach entails *three database-centric technical contributions*.

First, we exploit join dependencies and their factorization in the training dataset to asymptotically improve the per-iteration computation time of a gradient descent algorithm.

Second, we exploit functional dependencies present in the database to reduce the dimensionality of the underlying optimization problem by only optimizing for those parameters that functionally determine the others and by subsequently recovering the functionally determined parameters using their dependencies.

Third, we address the shortcomings of one-hot encoding by expressing the sum-product aggregates used to compute the gradient and point evaluation as functional aggregate queries (FAQs) [8]. The aggregates over continuous features are expressed as FAQs without free (i.e., group-by) variables and their computation yields scalar values. In contrast, aggregates over categorical features originating from a set  $S$  of database attributes are expressed as FAQs with free variables  $S$ . The tuples in the result of such FAQs are combinations of categorical values that occur in the training dataset. The ensemble of FAQs defining the gradient form a *sparse tensor representation and computation solution* with lower space and time complexity than solutions based on one-hot encoding. In particular, the complexity of our end-to-end solution can be arbitrarily smaller than that of materializing the result of the feature extraction query.

The above three technical contributions led to the design and implementation of AC/DC, a gradient descent solver for polynomial regression models and factorization machines over databases. To train such models of up to 66K features over the natural join of all relations from a real-world dataset of up to 86M tuples, AC/DC needs

---

<sup>1</sup>Most of the raw features we observed in datasets for retail applications are categorical. In several domains, such as statistical arbitrage [45], it is common to derive many continuous features from categorical features.

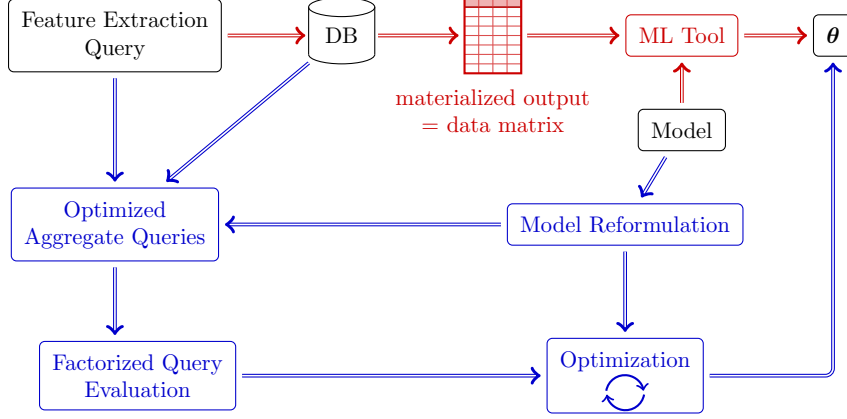


Figure 1: **Structure-aware** vs. **structure-agnostic** learning: High-level diagram.

up to 15 minutes on eight cores of a commodity machine. AC/DC is up to 1,031 times faster than its competitors MadLib [38], libFM [64], and TensorFlow [1] whenever they do not exceed memory limitation, 24-hour timeout, or internal design limitations.

Figure 1 depicts schematically the workflows of our approach and of the mainstream approach for solving optimization problems. The latter materializes the result of the feature extraction query, exports it out of the database and import it as the training dataset in the ML tool, where the desired model is learned. We call it *structure-agnostic* since it does not exploit the relational structure of the underlying training dataset to avoid learning over the full materialization of the training dataset. In contrast, our *structure-aware* approach avoids this materialization and has the following steps: (1) it defines a set of aggregates needed to compute the gradient of the objective function for the desired model; (2) it optimizes these aggregates over the feature extraction query and under dependencies holding in the database and join dependencies defined by the feature extraction query; (3) it computes these aggregates in bulk using factorization techniques and exploiting subexpressions common among them; and (5) it uses a gradient descent solver to compute the model parameters based on the computed aggregates.

This article brings together and extends two lines of our prior work: The theoretical development of model reparameterization under functional dependencies and of factorized learning [5] and a preliminary report on the design and implementation of AC/DC [4]. The extensions concern the treatment of PCA, simplified proof for model reparameterization under functional dependencies, different experiments, and a classification of the existing landscape of structure-aware versus structure-agnostic approaches to analytics.

## Organization

The structure of the paper follows our contributions. Section 2 introduces preliminary notions needed throughout the article. Section 3 describes our unified framework for structure-aware analytics. Section 4 introduces our sparse tensor representation and computation approach for square loss problems (learning polynomial regression models and factorization machines) and principal component analysis together with its complexity analysis. Section 5 shows how to exploit functional dependencies to reduce the dimensionality of learned models. Section 6 discusses the design and implementation of the AC/DC system for learning models over relational databases. Section 7 presents our experimental findings. Section 8 overviews several strands of related work. Finally, Section 9 lists promising directions for future work. Further preliminaries and proofs of some theorems are deferred to the electronic appendix.

## 2 Preliminaries

We use the following notational conventions: bold face letters, e.g.,  $\mathbf{x}$ ,  $\boldsymbol{\theta}$ ,  $\mathbf{x}_i$ ,  $\boldsymbol{\theta}_j$ , denote vectors or matrices, and normal face letters, e.g.,  $x_i$ ,  $\theta_j$ ,  $\theta_i^{(j)}$ , denote scalars. For any positive integer  $n$ ,  $[n]$  denotes the set  $\{1, \dots, n\}$ .

For any set  $S$  and positive integer  $k$ ,  $\binom{S}{k}$  denotes the collection of all  $k$ -subsets of  $S$ . Let  $S$  be a finite set and  $\text{Dom}$  be any domain, then  $\mathbf{a}_S = (a_j)_{j \in S} \in \text{Dom}^{|S|}$  is a *tuple* indexed by  $S$ , whose components are in  $\text{Dom}$ . If  $S$  and  $T$  are disjoint, and given tuples  $\mathbf{a}_S$  and  $\mathbf{a}_T$ , the tuple  $(\mathbf{a}_S, \mathbf{a}_T)$  is interpreted naturally as the tuple  $\mathbf{a}_{S \cup T}$ . The tuple  $\mathbf{0}_S$  is the all-0 tuple indexed by  $S$ . If  $S \subseteq G$ , then the tuple  $\mathbf{1}_{S|G}$  is the characteristic vector of the subset  $S$ , i.e.,  $\mathbf{1}_{S|G}(v) = 1$  if  $v \in S$ , and 0 if  $v \in G - S$ .

## 2.1 Feature Extraction Query

We consider the setting where the training dataset  $D$  used as input to machine learning is the result of a query  $Q$  called *feature extraction query*, over a relational database  $I$ . This query is typically the natural join of the relations in the database. It is also common to join in further relations that are derived from the input relations by aggregating some of their columns. These further relations provide derived features, which add to the raw features readily provided by the input relations.

We use standard notation for query hypergraphs. Let  $\mathcal{H} = (\mathcal{V}, \mathcal{E})$  denote the hypergraph of the join query  $Q$ , where  $\mathcal{V}$  is the set of variables occurring in  $Q$  and  $\mathcal{E}$  is the set of hyperedges with one hyperedge per set of variables in a relation symbol  $R$  in the body of  $Q$ . We denote by  $V \subseteq \mathcal{V}$  the subset of variables selected as features, and let  $n = |V|$ . The features in  $V$  corresponding to qualitative attributes are called *categorical*, while those corresponding to quantitative attributes are *continuous*. Let  $N$  be the size of the largest input relation  $R$  in  $Q$ . Each tuple  $(\mathbf{x}, y) \in D$  contains a scalar response (regressand)  $y$  and a tuple  $\mathbf{x}$  encoding features (regressors).

**Example 1.** Consider the following natural join query  $Q$  that is a simplified version of a feature extraction query:

$$\begin{aligned} Q(\text{sku}, \text{store}, \text{day}, \text{color}, \text{quarter}, \text{city}, \text{country}, \text{unitsSold}, \text{price}, \text{size}) \\ \leftarrow \text{Sales}(\text{sku}, \text{store}, \text{day}, \text{unitsSold}), \text{Items}(\text{sku}, \text{color}, \text{price}), \\ \text{Quarter}(\text{day}, \text{quarter}), \text{Stores}(\text{store}, \text{city}, \text{size}), \text{Country}(\text{city}, \text{country}). \end{aligned}$$

Relation **Sales** records the number of units of a given **sku** (stock keeping unit) sold at a **store** on a particular **day**. The retailer is a global business, so it has stores in different cities and countries. One objective is to predict the number of blue units to be sold next year in the Fall quarter in Berlin. The response is the continuous variable *unitsSold*,  $\mathcal{V}$  is the set of all variables, and  $V = \mathcal{V} - \{\text{unitsSold}, \text{day}\}$ , all of which are categorical except *price* and *size*.  $\square$

## 2.2 Matrix calculus

We introduce basic concepts of matrix calculus and the following operations: the Kronecker/tensor product  $\otimes$ ; the Hadamard product  $\odot$ ; the Khatri-Rao product  $\star$ ; and the Frobenius inner product of two matrices  $\langle \cdot, \cdot \rangle$ , which reduces to the vector inner product when the matrices have one column each. We defer further preliminaries on matrix calculus to Appendix A and connection of tensor computation and the FAQ framework [8] to Appendix B.

### Basics

We list here common identities we often use in the paper; for more details see the Matrix Cookbook [60]. We use *denominator layout* for differentiation, i.e., the gradient is a column vector. Let  $\mathbf{A}$  be a matrix, and  $\mathbf{u}, \mathbf{v}, \mathbf{x}, \mathbf{b}$  be

vectors, where  $\mathbf{A}$  and  $\mathbf{b}$  are independent of  $\mathbf{x}$ , and  $\mathbf{u}$  and  $\mathbf{v}$  are functions of  $\mathbf{x}$  then

$$\frac{\partial \langle \mathbf{b}, \mathbf{x} \rangle}{\partial \mathbf{x}} = \mathbf{b} \quad (1)$$

$$\frac{\partial \mathbf{x}^\top \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\top) \mathbf{x} \quad (2)$$

$$\frac{\partial \|\mathbf{A} \mathbf{x} - \mathbf{b}\|_2^2}{\partial \mathbf{x}} = 2\mathbf{A}^\top (\mathbf{A} \mathbf{x} - \mathbf{b}) \quad (3)$$

$$\frac{\partial \mathbf{u}^\top \mathbf{v}}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}^\top}{\partial \mathbf{x}} \mathbf{v} + \frac{\partial \mathbf{v}^\top}{\partial \mathbf{x}} \mathbf{u} \quad (4)$$

$$\frac{\partial (\mathbf{B} \mathbf{x} + \mathbf{b})^\top \mathbf{C} (\mathbf{D} \mathbf{x} + \mathbf{d})}{\partial \mathbf{x}} = \mathbf{B}^\top \mathbf{C} (\mathbf{D} \mathbf{x} + \mathbf{d}) + \mathbf{D}^\top \mathbf{C}^\top (\mathbf{B} \mathbf{x} + \mathbf{b}). \quad (5)$$

### The Product Cookbook: Tensor product, Kronecker product, and Khatri-Rao product

Next, we discuss some identities regarding tensors. We use  $\otimes$  to denote the *tensor product*. When taking tensor product of two matrices, this is called the *Kronecker product*, which is *not* the same as the outer product for matrices, even though the two are isomorphic maps. If  $\mathbf{A} = (a_{ij})$  is an  $m \times n$  matrix and  $\mathbf{B} = (b_{k\ell})$  is a  $p \times q$  matrix, then the tensor product  $\mathbf{A} \otimes \mathbf{B}$  is an  $mp \times nq$  matrix whose  $((i, k), (j, \ell))$  entry is  $a_{ij}b_{k\ell}$ . In particular, if  $\mathbf{x} = (x_i)_{i \in [m]}$  is an  $m$ -dimensional vector and  $\mathbf{y} = (y_j)_{j \in [p]}$  is an  $p$ -dimensional vector, then  $\mathbf{x} \otimes \mathbf{y}$  is an  $mp$ -dimensional vector whose  $(i, j)$  entry is  $x_i y_j$ ; this is *not* an  $m \times p$  matrix as in the case of the outer product. This layout is the correct layout from the definition of the tensor (Kronecker) product. If  $\mathbf{A}$  is matrix, then  $\mathbf{A}^{\otimes k}$  denote the tensor power  $\underbrace{\mathbf{A} \otimes \cdots \otimes \mathbf{A}}_{k \text{ times}}$ .

**Definition 1** (Tensor product). Let  $\mathbf{A}$  and  $\mathbf{B}$  be tensors of order  $r$  and  $s$  respectively, i.e., functions  $\psi_A(X_1, \dots, X_r)$  and  $\psi_B(Y_1, \dots, Y_s)$ . The tensor product  $\mathbf{A} \otimes \mathbf{B}$  is the multilinear function

$$\psi(X_1, \dots, X_r, Y_1, \dots, Y_s) = \psi_A(X_1, \dots, X_r) \psi_B(Y_1, \dots, Y_s).$$

A matrix is a tensor of order 2.

**Definition 2** (Khatri-Rao product). Let  $\mathbf{A}$  and  $\mathbf{B}$  be two matrices each with  $n$  columns. We use  $\mathbf{A} \star \mathbf{B}$  to denote the matrix with  $n$  columns, where the  $j$ th column of  $\mathbf{A} \star \mathbf{B}$  is the tensor product of the  $j$ th column of  $\mathbf{A}$  with the  $j$ th columns of  $\mathbf{B}$ . The operator  $\star$  is a (special case of) the *Khatri-Rao product* [40], where we partition the input matrices into blocks of one column each. More elaborately, if  $\mathbf{A}$  has columns  $\mathbf{a}_1, \dots, \mathbf{a}_n$ , and  $\mathbf{B}$  has columns  $\mathbf{b}_1, \dots, \mathbf{b}_n$ , then one can visualize the  $\star$  operator as follows:

$$\mathbf{A} \star \mathbf{B} = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & \cdots & | \end{bmatrix} \star \begin{bmatrix} | & | & \cdots & | \\ \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_n \\ | & | & \cdots & | \end{bmatrix} = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{a}_1 \otimes \mathbf{b}_1 & \mathbf{a}_2 \otimes \mathbf{b}_2 & \cdots & \mathbf{a}_n \otimes \mathbf{b}_n \\ | & | & \cdots & | \end{bmatrix}.$$

(Note  $\mathbf{A}$  and  $\mathbf{B}$  do not need to have the same number of rows.)

**Definition 3** (Hadamard product). Let  $\mathbf{A} = (a_{ij})$  and  $\mathbf{B} = (b_{ij})$  be two  $m \times n$  matrices, then the Hadamard product  $\mathbf{A} \circ \mathbf{B}$  is an  $m \times n$  matrix, where each  $i, j$  element is given by  $(\mathbf{A} \circ \mathbf{B})_{ij} = a_{ij}b_{ij}$ .

## 3 Problem formulation

This section introduces a general formulation for a range of machine learning tasks and then lays out a versatile mathematical representation suitable for the in-database treatment of these tasks.

### 3.1 Continuous features

We start with a standard formulation in machine learning, where all model features are numerical.

The training dataset  $D$ , which is defined by a feature extraction query over a relational database, consists of tuples  $(\mathbf{x}, y)$  of a feature vector  $\mathbf{x}$  and a response  $y$ .

In case of continuous features,  $\mathbf{x} \in \mathbb{R}^n$  is the vector of  $n$  raw input features, or equivalently the variables in the feature extraction query. We denote by  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p) \in \mathbb{R}^p$  the vector of  $p$  so-called parameters. Let  $m \geq n$  be an integer. We define feature and parameter maps as follows.

The feature map  $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$  transforms the raw input vector  $\mathbf{x}$  into an  $m$ -vector of “monomial features”  $h(\mathbf{x}) = (h_j(\mathbf{x}))_{j \in [m]}$ . Each component  $h_j$  is a multivariate *monomial* designed to capture the *interactions* among dimensions of input  $\mathbf{x}$ . In particular, we write  $h_j(\mathbf{x}) = \prod_{i \in [n]} x_i^{a_j(i)}$ , where degree  $a_j(i)$  represents the level of participation of input dimension  $i$  in the  $j$ -th monomial feature.

The parameters  $\boldsymbol{\theta}$  produce the coefficients associated with features  $h$  via parameter map  $g : \mathbb{R}^p \rightarrow \mathbb{R}^m$ ,  $g(\boldsymbol{\theta}) = (g_j(\boldsymbol{\theta}))_{j \in [m]}$ . Each component  $g_j$  is a multivariate *polynomial* of  $\boldsymbol{\theta}$ .

A large number of machine learning tasks learn a functional quantity of the form  $\langle g(\boldsymbol{\theta}), h(\mathbf{x}) \rangle$ , where the parameters  $\boldsymbol{\theta}$  are obtained by solving  $\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$  with

$$J(\boldsymbol{\theta}) = \sum_{(\mathbf{x}, y) \in D} \mathcal{L}(\langle g(\boldsymbol{\theta}), h(\mathbf{x}) \rangle, y) + \Omega(\boldsymbol{\theta}). \quad (6)$$

$\mathcal{L}$  is a loss function, e.g., square loss, and  $\Omega$  is a regularizer, e.g.,  $\ell_1$ - or  $\ell_2$ -norm of  $\boldsymbol{\theta}$ . For square loss and  $\ell_2$ -regularization,  $J(\boldsymbol{\theta})$  becomes:

$$J(\boldsymbol{\theta}) = \frac{1}{2|D|} \sum_{(\mathbf{x}, y) \in D} (\langle g(\boldsymbol{\theta}), h(\mathbf{x}) \rangle - y)^2 + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2. \quad (7)$$

**Example 2.** The *ridge linear regression* (LR) model with response  $y$  and regressors  $x_1, \dots, x_n$  has  $p = n + 1$ , parameters  $\boldsymbol{\theta} = (\theta_0, \dots, \theta_n)$ . For convenience, we set  $x_0 = 1$  corresponding to the bias parameter  $\theta_0$ . Then,  $m = n + 1$ ,  $\mathbf{g}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ , and  $h(\mathbf{x}) = \mathbf{x}$ . The inner product becomes  $\langle g(\boldsymbol{\theta}), h(\mathbf{x}) \rangle = \langle \boldsymbol{\theta}, \mathbf{x} \rangle = \sum_{i=0}^n \theta_i x_i$  and Equation (7) becomes  $J(\boldsymbol{\theta}) = \frac{1}{2|D|} \sum_{(\mathbf{x}, y) \in D} (\sum_{i=0}^n \theta_i x_i - y)^2 + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2$ .  $\square$

**Example 3.** The *degree- $d$  polynomial regression* ( $\text{PR}^d$ ) model with response  $y$  and regressors  $x_0 = 1, x_1, \dots, x_n$  has  $p = m = \binom{n+d}{d} = \sum_{i=0}^d \binom{n+i-1}{i}$  parameters  $\boldsymbol{\theta} = (\theta_{\mathbf{a}})$ , where  $\mathbf{a} = (a_1, \dots, a_n)$  is a tuple of non-negative integers such that  $\|\mathbf{a}\|_1 \leq d$ . In this case,  $g(\boldsymbol{\theta}) = \boldsymbol{\theta}$ , while the components of  $h$  are given by  $h_{\mathbf{a}}(\mathbf{x}) = \prod_{i=1}^n x_i^{a_i}$ .  $\square$

**Example 4.** In contrast to polynomial regression models, factorization machines [65] factorize the space of model parameters to better capture data correlations. The *degree-2 rank- $r$  factorization machines* ( $\text{FaMa}_r^2$ ) model with regressors  $x_0 = 1, x_1, \dots, x_n$  and regressand  $y$  has parameters  $\boldsymbol{\theta}$  consisting of  $\theta_i$  for  $i \in \{0, \dots, n\}$  and  $\theta_i^{(\ell)}$  for  $i \in [n]$  and  $\ell \in [r]$ . Training  $\text{FaMa}_r^2$  corresponds to minimizing the following  $J(\boldsymbol{\theta})$ :

$$\frac{1}{2|D|} \sum_{(\mathbf{x}, y) \in D} \left( \sum_{i=0}^n \theta_i x_i + \sum_{\substack{\{i,j\} \in \binom{[n]}{2} \\ \ell \in [r]}} \theta_i^{(\ell)} \theta_j^{(\ell)} x_i x_j - y \right)^2 + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2.$$

This loss function follows Equation (7) with  $p = 1 + n + rn$ ,  $m = 1 + n + \binom{n}{2}$ , and the maps

$$\begin{aligned} h_S(\mathbf{x}) &= \prod_{i \in S} x_i, \text{ for } S \subseteq [n], |S| \leq 2 \\ g_S(\boldsymbol{\theta}) &= \begin{cases} \theta_0 & \text{when } |S| = 0 \\ \theta_i & \text{when } S = \{i\} \\ \sum_{\ell=1}^r \theta_i^{(\ell)} \theta_j^{(\ell)} & \text{when } S = \{i, j\}. \end{cases} \end{aligned}$$

$\square$

**Example 5.** *Classification methods* such as support vector machines (SVM), logistic regression and Adaboost also fall under the same optimization framework, but with different choices of loss  $\mathcal{L}$  and regularizer  $\Omega$ . Typically,  $\Omega(\boldsymbol{\theta}) = \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2$ . Restricting to binary class labels  $y \in \{\pm 1\}$ , the loss function  $\mathcal{L}(\gamma, y)$ , where  $\gamma = \langle g(\boldsymbol{\theta}), h(\mathbf{x}) \rangle$ , takes the form  $\mathcal{L}(\gamma, y) = \max\{1 - y\gamma, 0\}$  for SVM,  $\mathcal{L}(\gamma, y) = \log(1 + e^{-y\gamma})$  for logistic regression and  $\mathcal{L}(\gamma, y) = e^{-y\gamma}$  for Adaboost.  $\square$

**Example 6.** Various unsupervised learning techniques can be expressed as iterative optimization procedures according to which each iteration is reduced to an optimization problem of the generic form given above. For example, the *Principal Component Analysis (PCA)* requires solving the following optimization problem to obtain a principal component direction

$$\max_{\|\boldsymbol{\theta}\|_2=1} \boldsymbol{\theta}^\top \boldsymbol{\Sigma} \boldsymbol{\theta} = \max_{\boldsymbol{\theta} \in \mathbb{R}^p} \min_{\lambda \in \mathbb{R}} \boldsymbol{\theta}^\top \boldsymbol{\Sigma} \boldsymbol{\theta} + \lambda(\|\boldsymbol{\theta}\|_2^2 - 1),$$

where  $\boldsymbol{\Sigma} = \frac{1}{|D|} \sum_{\mathbf{x} \in D} \mathbf{x} \mathbf{x}^\top$  is the (empirical) correlation matrix of the given data. Although there is no response/class label  $y$ , within each iteration of the above iteration, for a fixed  $\lambda$ , there is a loss function  $\mathcal{L}$  acting on feature vector  $h(\mathbf{x})$  and parameter vector  $g(\boldsymbol{\theta})$ , along with a regularizer  $\Omega$ . Specifically, we have  $h(\mathbf{x}) = \boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ ,  $g(\boldsymbol{\theta}) = \boldsymbol{\theta} \otimes \boldsymbol{\theta} \in \mathbb{R}^{p \times p}$ ,  $\mathcal{L} = \langle g(\boldsymbol{\theta}), h(\mathbf{x}) \rangle_F$ , where the Frobenius inner product is now employed. In addition,  $\Omega(\boldsymbol{\theta}) = \lambda(\|\boldsymbol{\theta}\|_2^2 - 1)$ .  $\square$

### 3.2 Categorical features

The active domain of a categorical feature/variable is a set of possible values or categories, e.g., **vietnam**, **england**, and **usa** are possible categories of the categorical feature **country**. Categorical features constitute the vast majority of features we observed in machine learning applications.

It is common practice to one-hot encode categorical variables [36]. Whereas a continuous variable such as **salary** is mapped to a scalar value  $x_{\text{salary}}$ , a categorical variable such as **country** is mapped to an indicator vector  $\mathbf{x}_{\text{country}}$  – a vector of binary values indicating the category that the variable takes on. For example, if the active domain of **country** consists of **vietnam**, **england**, and **usa**, then  $\mathbf{x}_{\text{country}} = [x_{\text{vietnam}}, x_{\text{england}}, x_{\text{usa}}] \in \{0, 1\}^3$ . If a tuple in the training dataset has **country** = “**england**”, then  $\mathbf{x}_{\text{country}} = [0, 1, 0]$  for that tuple.

In general, the feature vector  $\mathbf{x}$  has the form  $\mathbf{x} = (\mathbf{x}_c)_{c \in V}$ , where each component  $\mathbf{x}_c$  is an indicator vector if  $c$  is a categorical variable and a scalar otherwise. Similarly, each component of the parameter vector  $\boldsymbol{\theta}$  becomes a matrix, or a vector if the matrix has one column.

### 3.3 Tensor product representation

We accommodate both continuous and categorical features in our problem formulation (7) by replacing arithmetic product by tensor product in the component functions of the parameter map  $g$  and the feature map  $h$ . Specifically, monomials  $h_j$  now take the form

$$h_j(\mathbf{x}) = \bigotimes_{f \in V} \mathbf{x}_f^{\otimes a_j(f)} \quad (8)$$

with degree vector  $\mathbf{a}_j = (a_j(f))_{f \in V} \in \mathbb{N}^n$ . For each  $j \in [m]$ , the set  $V_j = \{f \in V \mid a_j(f) > 0\}$  consists of features that participate in the interaction captured by the (hyper-) monomial  $h_j$ . Let  $C \subseteq V$  denote the set of categorical variables and  $C_j = C \cap V_j$  the subset of categorical variables in  $V_j$ . For  $f \in C_j$ ,  $h_j$  represents  $\prod_{f \in C_j} |\pi_f(D)|$  many monomials, one for each combination of the categories, where  $\pi_f(D)$  denotes the projection of  $D$  onto variable  $f$ . Due to one-hot encoding, each element in the vector  $\mathbf{x}_f$  for a categorical variable  $f$  is either 0 or 1, and  $\mathbf{x}_f^{a_j(f)} = \mathbf{x}_f$  for  $a_j(f) > 0$ . Hence,  $h_j$  can be simplified as follows:

$$h_j(\mathbf{x}) = \prod_{f \in V_j - C_j} x_f^{a_j(f)} \cdot \bigotimes_{f \in C_j} \mathbf{x}_f. \quad (9)$$

Note that we use  $x_f$  instead of boldface  $\mathbf{x}_f$  since each variable  $f \in V_j - C_j$  is continuous.

**Example 7.** For illustration, consider a query that extracts tuples over schema  $(\text{country}, a, b, c, \text{color})$  from the database, where  $\text{country}$  and  $\text{color}$  are categorical variables, while  $a, b, c$  are continuous variables. Moreover, there are two countries `vietnam` and `england`, and three colors `red`, `green`, and `blue` in the training dataset  $D$ . Consider three of the possible feature functions:

$$h_1(\mathbf{x}) = \mathbf{x}_{\text{country}} \otimes x_a^2 x_c \quad (10)$$

$$h_2(\mathbf{x}) = \mathbf{x}_{\text{country}} \otimes \mathbf{x}_{\text{color}} \otimes x_b \quad (11)$$

$$h_3(\mathbf{x}) = x_b x_c. \quad (12)$$

Under the one-hot encoding, the schema of the tuples becomes:

$$(\text{vietnam}, \text{england}, a, b, c, \text{red}, \text{green}, \text{blue}).$$

Equation (9) says that the functions  $h_1$  and  $h_2$  are actually encoding 8 functions:

$$\begin{aligned} h_{1,\text{vietnam}}(\mathbf{x}) &= x_{\text{vietnam}} x_a^2 x_c \\ h_{1,\text{england}}(\mathbf{x}) &= x_{\text{england}} x_a^2 x_c \\ h_{2,\text{vietnam},\text{red}}(\mathbf{x}) &= x_{\text{vietnam}} x_{\text{red}} x_b \\ h_{2,\text{vietnam},\text{green}}(\mathbf{x}) &= x_{\text{vietnam}} x_{\text{green}} x_b \\ h_{2,\text{vietnam},\text{blue}}(\mathbf{x}) &= x_{\text{vietnam}} x_{\text{blue}} x_b \\ h_{2,\text{england},\text{red}}(\mathbf{x}) &= x_{\text{england}} x_{\text{red}} x_b \\ h_{2,\text{england},\text{green}}(\mathbf{x}) &= x_{\text{england}} x_{\text{green}} x_b \\ h_{2,\text{england},\text{blue}}(\mathbf{x}) &= x_{\text{england}} x_{\text{blue}} x_b. \end{aligned}$$

□

We elaborate the tensor product representation for the considered learning models.

**Example 8.** In linear regression, parameter  $\boldsymbol{\theta}$  is a vector of vectors:  $\boldsymbol{\theta} = [\boldsymbol{\theta}_0, \dots, \boldsymbol{\theta}_n]$ . Since our inner product is Frobenius, when computing  $\langle \boldsymbol{\theta}, \mathbf{x} \rangle$  we should be multiplying, for example,  $\boldsymbol{\theta}_{\text{usa}}$  with  $x_{\text{usa}}$  correspondingly. □

**Example 9.** In polynomial regression, the parameter  $\boldsymbol{\theta}$  is a vector of tensors (i.e., high-dimensional matrices). Consider for instance the second order term  $\theta_{ij} x_i x_j$ . When both  $i$  and  $j$  are continuous,  $\theta_{ij}$  is just a scalar. Now, suppose  $i$  is `country` and  $j$  is `color`. Then, the model has terms  $\theta_{\text{vietnam},\text{red}} x_{\text{vietnam}} x_{\text{red}}$ ,  $\theta_{\text{usa},\text{green}} x_{\text{usa}} x_{\text{green}}$ , and so on. All these terms are captured by the Frobenius inner product  $\langle \boldsymbol{\theta}_{ij}, \mathbf{x}_i \otimes \mathbf{x}_j \rangle$ . The component  $\boldsymbol{\theta}_{ij}$  is a matrix whose number of entries is the number of pairs  $(\text{country}, \text{color})$  that appear together in some tuple in the training dataset. This number can be much less than the product of the numbers of countries and of colors in the input database. □

**Example 10.** Consider the  $\text{FaMa}_r^2$  model from Example (4), but now with categorical variables. From the previous examples, we already know how to interpret the linear part  $\sum_{i=0}^n \theta_i x_i$  of the model when features are categorical. Consider a term in the quadratic part such as  $\sum_{\ell \in [r]} \theta_i^{(\ell)} \theta_j^{(\ell)} x_i x_j$ . When  $i$  and  $j$  are categorical, the term becomes  $\left\langle \sum_{\ell \in [r]} \boldsymbol{\theta}_i^{(\ell)} \otimes \boldsymbol{\theta}_j^{(\ell)}, \mathbf{x}_i \otimes \mathbf{x}_j \right\rangle$ . □

## 4 Database-centric problem reformulation

In this section, we show how we reformulate the square loss optimization problems (learning polynomial regression and factorization machine models) and PCA to encode their data-intensive components as FAQs. The ensemble of these FAQs form a sparse tensor representation and computation solution with lower space and time complexity than solutions based on one-hot encoding.



---

**Algorithm 1:** BGD with Armijo line search.

---

```
 $\theta \leftarrow$  a random point;  
while not converged yet do  
   $\alpha \leftarrow$  next step size;  
   $\mathbf{d} \leftarrow \nabla J(\theta)$ ;  
  while  $J(\theta - \alpha \mathbf{d}) \geq J(\theta) - \frac{\alpha}{2} \|\mathbf{d}\|_2^2$  do  
     $\alpha \leftarrow \alpha/2$  // line search;  
  end  
   $\theta \leftarrow \theta - \alpha \mathbf{d}$ ;  
end
```

---

#### 4.1 Solution for square loss problems

We introduce our approach to learning statistical models for the setting of square loss function  $J(\theta)$  and  $\ell_2$ -norm as in (7). We use a gradient-based optimization algorithm that employs the first-order gradient information to optimize the loss function  $J(\theta)$ . It repeatedly updates the parameters  $\theta$  by some step size  $\alpha$  in the direction of the gradient  $\nabla J(\theta)$  until convergence. To guarantee convergence, it uses backtracking line search to ensure that  $\alpha$  is sufficiently small to decrease the loss for each step. Each update step requires two computations: (1) *Point evaluation*: Given  $\theta$ , compute the scalar  $J(\theta)$ ; and (2) *Gradient computation*: Given  $\theta$ , compute the vector  $\nabla J(\theta)$ . In particular, we use the batch gradient descent (BGD) algorithm with the Armijo line search condition and the Barzilai-Borwein step size adjustment [14, 28], as depicted in Algorithm 1. Quasi-Newton optimization algorithms (e.g., L-BFGS) and other common line search conditions are also applicable in our framework. We refer the reader to the excellent review article [31] for additional details on fast implementations of gradient-descent optimization methods.

##### Continuous features

We first consider the case without categorical features. We rewrite the square-loss function (7) to factor out the data-dependent part of the point evaluation and gradient computation. Recall that, for  $j \in [m]$ ,  $h_j$  denotes the  $j$ th component function of the vector-valued function  $h$ , and  $h_j$  is a multivariate monomial in  $\mathbf{x}$ .

**Theorem 4.1.** *Let  $J(\theta)$  be the function in (7). Define the matrix  $\Sigma = (\sigma_{ij})_{i,j \in [m]}$ , the vector  $\mathbf{c} = (c_i)_{i \in [m]}$ , and the scalar  $s_Y$  by*

$$\Sigma = \frac{1}{|D|} \sum_{(\mathbf{x}, y) \in D} h(\mathbf{x}) h(\mathbf{x})^\top \quad (13)$$

$$\mathbf{c} = \frac{1}{|D|} \sum_{(\mathbf{x}, y) \in D} y \cdot h(\mathbf{x}) \quad (14)$$

$$s_Y = \frac{1}{|D|} \sum_{(\mathbf{x}, y) \in D} y^2. \quad (15)$$

Then,

$$J(\theta) = \frac{1}{2} g(\theta)^\top \Sigma g(\theta) - \langle g(\theta), \mathbf{c} \rangle + \frac{s_Y}{2} + \frac{\lambda}{2} \|\theta\|^2 \quad (16)$$

$$\nabla J(\theta) = \frac{\partial g(\theta)^\top}{\partial \theta} \Sigma g(\theta) - \frac{\partial g(\theta)^\top}{\partial \theta} \mathbf{c} + \lambda \theta. \quad (17)$$

Note that  $\frac{\partial g(\theta)^\top}{\partial \theta}$  is a  $p \times m$  matrix, and  $\Sigma$  is an  $m \times m$  matrix. Statistically,  $\Sigma$  is related to the covariance matrix,  $\mathbf{c}$  to the correlation between the response and the regressors, and  $s_Y$  to the empirical second moment of

the response variable. Theorem 4.1 allows us to compute the two key steps of BGD *without* scanning through the data again, because the quantities  $(\Sigma, \mathbf{c}, s_Y)$  can be computed efficiently in a preprocessing step *inside the database* as aggregates over the feature extraction query  $Q$ .

**Example 11.** Consider the query  $Q$  in Example 1, where the set of features is  $\{\text{sku}, \text{store}, \text{day}, \text{color}, \text{quarter}, \text{city}, \text{country}, \text{price}, \text{size}\}$  and  $\text{unitsSold}$  is the response variable. In this query  $n = 9$ , and thus for a  $\text{PR}_2$  model we have  $m = 1 + 9 + \binom{9}{2} = 46$  parameters. Consider two indices  $i$  and  $j$  to the component functions of  $g$  and  $h$ , where  $i = (\text{price})$  and  $j = (\text{size})$ . Then we can compute the entry  $\sigma_{ij} \in \Sigma$  with the following SQL query:

SELECT SUM( $\text{price} * \text{size}$ ) FROM D;

□

When  $g$  is the identity function, i.e., the model is linear, as is the case in PR (and thus LR) model, (16) and (17) become particularly simple:

**Corollary 4.2.** *In a linear model (i.e.,  $g(\theta) = \theta$ ),*

$$J(\theta) = \frac{1}{2} \theta^\top \Sigma \theta - \langle \theta, \mathbf{c} \rangle + \frac{s_Y}{2} + \frac{\lambda}{2} \|\theta\|_2^2 \quad (18)$$

$$\nabla J(\theta) = \Sigma \theta + \lambda \theta - \mathbf{c}. \quad (19)$$

Let  $\mathbf{d} = \nabla J(\theta)$ . Then,

$$\nabla J(\theta - \alpha \mathbf{d}) = (1 - \alpha) \mathbf{d} - \alpha \Sigma \mathbf{d}. \quad (20)$$

The Armijo condition  $J(\theta - \alpha \mathbf{d}) \geq J(\theta) - \frac{\alpha}{2} \|\mathbf{d}\|_2^2$  becomes:

$$\alpha \theta^\top \Sigma \mathbf{d} - \frac{\alpha^2}{2} \mathbf{d}^\top \Sigma \mathbf{d} - \alpha \langle \mathbf{c}, \mathbf{d} \rangle + \lambda \alpha \langle \theta, \mathbf{d} \rangle \leq \frac{\alpha}{2} (\lambda \alpha + 1) \|\mathbf{d}\|_2^2. \quad (21)$$

The significance of (21) is as follows. In a typical iteration of BGD, we have to backtrack a few times (say  $t$  times) for each value of  $\alpha$ . If we were to recompute  $J(\theta - \alpha \mathbf{d})$  using (18) each time, then the runtime of Armijo backtracking search is  $O(tm^2)$ , even after we have already computed  $\mathbf{d}$  and  $J(\theta)$ . Now, using (21), we can compute in advance the following quantities (in this order):  $\mathbf{d}$ ,  $\|\theta\|_2^2$ ,  $\Sigma \mathbf{d}$ ,  $\langle \mathbf{c}, \mathbf{d} \rangle$ ,  $\langle \theta, \mathbf{d} \rangle$ ,  $\mathbf{d}^\top \Sigma \mathbf{d}$ ,  $\theta^\top \Sigma \mathbf{d}$ . Then, each check for inequality (21) can be done in  $O(1)$ -time, for a total of  $O(m^2 + t)$ -times. Once we have determined the step size  $\alpha$ , (20) allows us to compute the next gradient (i.e., the next  $\mathbf{d}$ ) in  $O(m)$ , because we have already computed  $\Sigma \mathbf{d}$  for line search.

To implement BGD, we need to compute four quantities efficiently: the  $\Sigma$  matrix in (13), the vector  $\mathbf{c}$  in (14), point evaluation in (16), and the gradient in (17). The covariance matrix and the correlation vector only have to be computed once in a pre-processing step. The gradient is computed at every iteration, which includes several point evaluations as we perform line search.<sup>2</sup> We do not need to compute the second moment  $s_Y$  because optimizing  $J(\theta)$  is the same as optimizing  $J(\theta) - s_Y$ . Before describing how those four quantities can be computed efficiently, we discuss how we deal with categorical features.

### Categorical features via sparse tensors

The more interesting, more common, and also considerably challenging situation is in the presence of categorical features. We next explain how we accommodate categorical features in the computation of  $\Sigma$  and  $\mathbf{c}$ .

**Example 12.** In Example 7, the matrix  $\Sigma$  is of size  $8 \times 8$  instead of  $3 \times 3$  after one-hot encoding. However, many of those entries are 0, for instance  $(\forall (\mathbf{x}, y) \in D)$ :

$$\begin{aligned} h_{1,\text{vietnam}}(\mathbf{x}) h_{1,\text{england}}(\mathbf{x}) &= 0 \\ h_{1,\text{england}}(\mathbf{x}) h_{2,\text{vietnam,blue}}(\mathbf{x}) &= 0 \\ h_{2,\text{vietnam,blue}}(\mathbf{x}) h_{2,\text{england,blue}}(\mathbf{x}) &= 0 \\ h_{2,\text{vietnam,blue}}(\mathbf{x}) h_{2,\text{vietnam,red}}(\mathbf{x}) &= 0. \end{aligned}$$

<sup>2</sup>In our implementation, each iteration typically involves 1-4 backtracking steps.

The reason is that the indicator variables  $x_{\text{blue}}$  and  $x_{\text{england}}$  act like selection clauses  $x_{\text{color}} = \text{blue}$  and  $x_{\text{country}} = \text{england}$ . More concretely, we can rewrite the entry  $\sigma_{ij}$  as an aggregate over a more selective query. For instance, the entry that corresponds to the product of functions  $h_{1,\text{vietnam}}(\mathbf{x})$  and  $h_{2,\text{vietnam,red}}(\mathbf{x})$  from Example 7 can be rewritten as follows:

$$\sum_{(\mathbf{x},y) \in D} h_{1,\text{vietnam}}(\mathbf{x}) h_{2,\text{vietnam,red}}(\mathbf{x}) = \sum_{\phi} x_a^2 x_c x_b,$$

where  $\phi = ((\mathbf{x}, y) \in D \wedge x_{\text{color}} = \text{red} \wedge x_{\text{country}} = \text{vietnam})$ .  $\square$

Extrapolating straightforwardly, if we were to write  $\Sigma$  down in the one-hot encoded feature space, then the entries  $\sigma_{ij}$  under one-hot encoding got unrolled into many entries. Let  $C_i$  and  $C_j$  be the set of categorical variables for  $h_i$  and  $h_j$  as defined in Section 3.3. Then,  $\sigma_{ij}$  is in fact a tensor  $\sigma_{ij}$  of dimension  $\prod_{f \in C_i} |\pi_f(D)| \times \prod_{f \in C_j} |\pi_f(D)|$ , because

$$\sigma_{ij} = \frac{1}{|D|} \sum_{(\mathbf{x},y) \in D} h_i(\mathbf{x}) h_j(\mathbf{x})^\top. \quad (22)$$

Similarly, each component  $c_j$  of  $\mathbf{c}$  defined in (14) is a tensor  $\mathbf{c}_j$  of dimension  $\prod_{f \in C_j} |\pi_f(D)|$ , because  $h_j(\mathbf{x})$  is a tensor in the categorical case. The following follows immediately.

**Theorem 4.3.** *Theorem 4.1 remains valid even when some features are categorical.*

Note that the outer product in (22) specifies the matrix layout of  $\sigma_{ij}$ , and so  $\Sigma$  is a block matrix, each of whose blocks is  $\sigma_{ij}$ . Furthermore, if we were to layout the tensor  $\sigma_{ij}$  as a vector, we can also write it as

$$\sigma_{ij} = \frac{1}{|D|} \sum_{(\mathbf{x},y) \in D} h_i(\mathbf{x}) \otimes h_j(\mathbf{x}). \quad (23)$$

The previous example demonstrates that the dimensionalities of  $\sigma_{ij}$  and  $\mathbf{c}_j$  can be very large. Fortunately, the tensors are very sparse, and a sparse representation of them can be computed with functional aggregate queries (in the FAQ-framework [8]) as shown in Proposition 4.4 below. We next illustrate the sparsity.

**Example 13.** We extend the Example 11 for entries in Sigma with categorical variables. Consider two indices  $i$  and  $j$  to the component functions of  $g$  and  $h$ , where  $i = (\text{store}, \text{city})$  and  $j = (\text{city})$ . Suppose the query result states that the retailer has  $N_s$  stores in  $N_c$  countries. Then, the full dimensionality of the tensor  $\sigma_{ij}$  is  $N_s \times N_c^2$ , because by definition it was defined to be

$$\sigma_{ij} = \frac{1}{|D|} \sum_{(\mathbf{x},y) \in D} \underbrace{\mathbf{x}_{\text{store}} \otimes \mathbf{x}_{\text{city}}}_{h_i(\mathbf{x})} \otimes \underbrace{\mathbf{x}_{\text{city}}}_{h_j(\mathbf{x})}. \quad (24)$$

Recall that  $\mathbf{x}_{\text{store}}$  and  $\mathbf{x}_{\text{city}}$  are both indicator vectors. The above tensor has the following straightforward interpretation: for every triple  $(\text{store}, \text{city}_1, \text{city}_2)$ , where  $s$  is a store and  $c_1$  and  $c_2$  are cities, this triple entry of the tensor counts the number of data points  $(\mathbf{x}, y) \in D$  for this particular combination of store and cities (divided by  $1/|D|$ ). Most of these  $(s, c_1, c_2)$ -entries are 0. For example, if  $c_1 \neq c_2$  then the count is zero. Thus, we can concentrate on computing entries of the form  $(s, c, c)$ :

SELECT store, city, COUNT(\*) FROM D GROUP BY store, city;

Better yet, since **store** functionally determines **city**, the number of entries in the query output is bounded by  $N_s$ . Using relations to represent sparse tensor results in massive space saving.

We can also succinctly represent entries in  $\Sigma$  that are composed of continuous and categorical variables. Consider the entry that corresponds to dimensions  $i = (\text{store}, \text{city})$  and  $j = (\text{city}, \text{price})$ . We can compute this entry with the following SQL query:

SELECT store, city, SUM(price) FROM D GROUP BY store, city;

$\square$

## 4.2 Solution for Principal Component Analysis

We next consider principal component analysis (PCA) over the training dataset defined by a feature extraction query. We focus on the problem of computing the top- $K$  principal components, which correspond to the eigenvectors of the covariance matrix. Once computed, the principal components are then used to transform the data to a lower dimensional space. We show that the solution to this problem requires similar computations as our solution for square loss problems in Section 4.1.

### Continuous features

We first consider the case with continuous features only. Let  $\boldsymbol{\mu} = \frac{1}{|D|} \sum_{\mathbf{x} \in D} \mathbf{x}$  be the vector of means for each variable in the feature extraction query, and  $\boldsymbol{\Sigma}_1 = \frac{1}{|D|} \sum_{\mathbf{x} \in D} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top$  the centered covariance matrix. The top- $K$  eigenvectors  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$  and the corresponding eigenvalues  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$  can be computed one at a time using the min-max theorem based on the Rayleigh quotient [67]:

$$\max_{\boldsymbol{\theta}_j \in \mathbb{R}^p} \min_{\lambda_j \in \mathbb{R}} \boldsymbol{\theta}_j^\top \boldsymbol{\Sigma}_j \boldsymbol{\theta}_j + \lambda_j (\|\boldsymbol{\theta}_j\|^2 - 1) \quad (25)$$

We compute the optimal solution for the top eigenvector  $\boldsymbol{\theta}_1$  using a gradient-based optimization algorithm, which optimizes the following loss function by alternating between performing gradient ascent with respect to  $\boldsymbol{\theta}_1$  and gradient descent with respect to  $\lambda_1$  until convergence:

$$J(\boldsymbol{\theta}_1, \lambda_1) = \boldsymbol{\theta}_1^\top \boldsymbol{\Sigma}_1 \boldsymbol{\theta}_1 + \lambda_1 (\|\boldsymbol{\theta}_1\|^2 - 1) \quad (26)$$

The gradient optimization steps can then be done with Algorithm 1, where the gradient of  $J(\boldsymbol{\theta}_1, \lambda_1)$  for the two subproblems is given by:

$$\nabla_{\boldsymbol{\theta}_1} J(\boldsymbol{\theta}_1, \lambda_1) = \boldsymbol{\Sigma}_1 \boldsymbol{\theta}_1 - 2\lambda_1 \boldsymbol{\theta}_1 \quad (27)$$

$$\nabla_{\lambda_1} J(\boldsymbol{\theta}_1, \lambda_1) = \|\boldsymbol{\theta}_1\|^2 - 1 \quad (28)$$

The subsequent eigenvectors are computed with the same optimization procedure but over an updated covariance matrix that subtracts all previously computed principal components. The iteration step assumes we already computed the covariance matrix  $\boldsymbol{\Sigma}_l$ , the eigenvector  $\boldsymbol{\theta}_l$ , and the eigenvalue  $\lambda_l$  for  $l \in [K - 1]$ . The step then computes the eigenvector  $\boldsymbol{\theta}_{l+1}$  and the eigenvalue  $\lambda_{l+1}$  over the covariance matrix

$$\boldsymbol{\Sigma}_{l+1} = \boldsymbol{\Sigma}_l - \lambda_l \boldsymbol{\theta}_l \boldsymbol{\theta}_l^\top. \quad (29)$$

Once we computed the top- $K$  eigenvectors  $\boldsymbol{\theta}$ , the projection of a training sample  $\mathbf{x} \in D$  onto the lower  $K$ -dimensional space is given by the inner product  $\mathbf{x}^\top \boldsymbol{\theta}$ .

As for the square-loss problems, we can compute  $\boldsymbol{\Sigma}_1$  once, and then compute the eigenvectors without scanning the data again. If the data is centered in a preprocessing step, then the computation of  $\boldsymbol{\Sigma}_1$  for PCA is identical to (13) for the case of linear regression. If the data is not centered, we can compute the covariance matrix with the following reformulation:

$$\begin{aligned} \boldsymbol{\Sigma}_1 &= \frac{1}{|D|} \sum_{\mathbf{x} \in D} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top = \frac{1}{|D|} \sum_{\mathbf{x} \in D} \mathbf{x} \mathbf{x}^\top - \frac{2\boldsymbol{\mu}}{|D|} \sum_{\mathbf{x} \in D} \mathbf{x}^\top + \frac{1}{|D|} \sum_{\mathbf{x} \in D} \boldsymbol{\mu} \boldsymbol{\mu}^\top \\ &= \frac{1}{|D|} \sum_{\mathbf{x} \in D} \mathbf{x} \mathbf{x}^\top - 2\boldsymbol{\mu} \boldsymbol{\mu}^\top + \boldsymbol{\mu} \boldsymbol{\mu}^\top = \frac{1}{|D|} \sum_{\mathbf{x} \in D} \mathbf{x} \mathbf{x}^\top - \boldsymbol{\mu} \boldsymbol{\mu}^\top \end{aligned} \quad (30)$$

Thus, we can compute the covariance matrix by first computing the matrix from (13) where  $h(\mathbf{x}) = \mathbf{x}$  and then subtracting  $\boldsymbol{\mu} \boldsymbol{\mu}^\top$  to center the data.

The gradient with respect to  $\boldsymbol{\theta}$  for PCA requires the same computation over  $\boldsymbol{\Sigma}_1$  as the gradient for linear regression models (c.f. (19)).

### Categorical features via sparse tensors

PCA is based on the analysis of variance between variables, and therefore it cannot be computed directly over categorical data. It can however be meaningful to compute PCA over one-hot encoded categorical data, which would provide insights into the variance of the frequency of the co-occurrence of categories for different categorical variables. We can compute PCA over one-hot encoded categorical variables efficiently by computing it over the sparse representation of the covariance matrix, which is a variant of the sparse tensor representation of the  $\Sigma$  matrix that we introduced for the case of square-loss problems.

One difference between the square loss problems and PCA is that PCA requires its features to be linearly independent. This property is not satisfied by one-hot encoding, because it is possible to derive the indicator value for one category based on a linear combination of the indicator values for all other categories. For this reason, it is common practice to do one-hot encoding of the categorical variables for all but one category. In our problem formulation, this means that for a categorical variable  $c$ , we encode the corresponding component  $\mathbf{x}_c$  as an indicator vector whose size is the number of its categories minus one, so that we one-hot encode over all but the last category of  $c$ . This encoding is often referred to as *dummy encoding* in many data science tools.

Another difference is the requirement to center the data. For categorical data, it is not desirable to center the data in a preprocessing step, as this would require a one-hot encoding of the input relations. To avoid the materialization of the one-hot encoding, we compute the non-centered matrix first, and then subtract  $\mu\mu^\top$ , as shown in (30). The sparse representation of the covariance matrix is then a block matrix, where each entry  $\sigma_{ij} \in \Sigma_1$  is defined as:

$$\sigma_{ij} = \frac{1}{|D|} \sum_{\mathbf{x} \in D} \mathbf{x}_i \mathbf{x}_j^\top - \mu_i \mu_j^\top \quad (31)$$

The vector of means  $\mu$  has the same dimension as  $\mathbf{x}$ , where for each categorical variable  $c \in V$  the component  $\mu_c \in \mu$  is the vector of frequencies for all but one category in the domain of  $c$ :

$$\mu_c = \frac{1}{|D|} \sum_{\mathbf{x} \in D} \mathbf{x}_c. \quad (32)$$

The vector  $\mu_c$  can be computed efficiently as a SQL count query with group-by variable  $c$ . We drop the group with the lowest count and divide the count for each other group by  $|D|$ .

The resulting matrix  $\Sigma_1$  has the same structure as the sparse tensor that is computed for linear regression problems. In fact, the quantity  $\sigma_{ij}$  in (31) is simply the centered variant of the expression in (22) for the case where  $h(\mathbf{x}) = \mathbf{x}$ . The centering of the  $\Sigma_1$  as well as updating the matrix for subsequent eigenvectors can be expressed as group-by aggregate queries and computed without materializing the quantities  $\mu\mu^\top$  and  $\theta\theta^\top$ .

**Example 14.** Consider the entry  $\sigma_{ij} \in \Sigma_1$  where  $i = (\text{store})$  and  $j = (\text{city})$ . We can compute the centered entry in the covariance matrix based on the non-centered entry  $\sigma_{ij}$  and the frequency vectors for *store* and *city*. The non-centered entry is computed with the SQL query in Example 13,

Let  $\mu_s(\text{store}, \text{val})$  and  $\mu_c(\text{city}, \text{val})$  be the relational encoding of the frequency vectors for *store* and respectively *city*. The relations store tuples that give for each city and respectively store the corresponding frequency that is denoted by *val*. We can then compute the  $(i, j)$  entry in the centered covariance matrix without materializing the product of  $\mu_s$  and  $\mu_c^\top$  with the following SQL query:

```
SELECT store, city, SUM( $\sigma_{ij}.\text{val} - \mu_c.\text{val} * \mu_s.\text{val}$ )
FROM  $\sigma_{ij}, \mu_c, \mu_s$  WHERE  $\sigma_{ij}.\text{city} = \mu_c.\text{city}$  AND  $\sigma_{ij}.\text{store} = \mu_s.\text{store}$ 
GROUP BY store, city;
```

Let  $\theta_s(\text{store}, \text{val})$  and  $\theta_c(\text{city}, \text{val})$  be the relational encodings of the components in  $\theta_1$  that correspond to *store* and respectively *city*. We can compute the updated entry  $\sigma_{ij} \in \Sigma_2$  based on (29) without materializing the product of  $\theta_s$  and  $\theta_c^\top$  with the following query:

```
SELECT store, city, SUM( $\sigma_{ij}.\text{val} - \lambda_1 * \theta_c.\text{val} * \theta_s.\text{val}$ )
FROM  $\sigma_{ij}, \theta_c, \theta_s$  WHERE  $\sigma_{ij}.\text{city} = \theta_c.\text{city}$  AND  $\sigma_{ij}.\text{store} = \theta_s.\text{store}$ ;
GROUP BY store, city;
```

□

The eigenvectors and eigenvalues can be computed on top of the sparse representation of the covariance matrix without touching the input database, and the gradient with respect to the eigenvectors requires similar computation as the gradient for square loss problems. We next show how we can compute the sparse tensor representation.

### 4.3 Efficient computation of the sparse tensor representation

We consider the problem of computing the sparse tensor representation for a given optimization problem. For square-loss problems the sparse tensor captures the quantities  $\Sigma$  and  $\mathbf{c}$ , and for PCA we compute the non-centered covariance matrix (referred to as  $\Sigma$  for uniformity), which is then centered in a subsequent step as shown in Example 14.

An immediate approach to computing this representation is to first materialize the result of the feature extraction query  $Q$  using an efficient query engine, e.g., a worst-case optimal join algorithm, and then compute the entries in the representation as aggregates over the query result. This approach, however, is suboptimal, since the listing representation of the query result is highly redundant and not necessary for the computation of the aggregates.

We employ two orthogonal observations to avoid this redundancy.

First, we use the FAQ [8] and FDB [55] frameworks for factorized computation of aggregates over joins. In a nutshell, factorized aggregate computation unifies three powerful ideas: worst-case optimal join processing, query plans defined by fractional hypertree decompositions of join queries, and pushing aggregates past joins.

Second, we exploit the observation that in the computation of  $\Sigma$  many *distinct* tensors  $\sigma_{ij}$  have *identical* sparse representations. For instance, the tensor  $\sigma_{ij}$  from Example 13 corresponding to  $i = (\text{store}, \text{city})$  and  $j = (\text{city})$  has the same sparse representation as any of the following tensors:  $(i, j) \in \{((\text{city}, \text{city}), \text{store}), ((\text{store}, \text{store}), \text{city}), ((\text{store}, \text{city}), \text{store}), \dots\}$ . This is because **store** and **city** are categorical features and taking any power of the binary values in their indicator vectors does not change these values. Furthermore, any of the two features can be in  $i$  and/or  $j$ .

The time complexity of computing the representation can be lower than that of materializing the result of the feature extraction query  $Q$ . Let  $|\sigma_{ij}|$  denote the *size* (i.e., number of tuples) of the sparse representation of the  $\sigma_{ij}$  tensor. Let  $\text{faqw}(i, j)$  denote the *FAQ-width* of the FAQ-query<sup>3</sup> that expresses the aggregate  $\sigma_{ij}$  over  $Q$ ;  $\text{fhtw}$  be the fractional hypertree width of  $Q$ ; and  $\rho^*$  be the fractional edge cover number<sup>4</sup> of  $Q$ . Let  $I$  be the input database and  $D = Q(I)$ . Let  $N$  be the size of the largest input relation in  $Q$ , which means that  $|D| = O(N^{\rho^*})$ . Recall that  $\mathcal{V}$  is the set of query variables in  $Q$ ,  $\mathcal{E}$  is the set of relations in  $Q$ , and  $m$  is the number of features. The time to compute the sparse tensor representation can be bounded as follows.

**Proposition 4.4.** *The tensors  $\sigma_{ij}$  and  $\mathbf{c}_j$  can be sparsely represented by FAQ-queries with group-by variables  $C_i \cup C_j$  and  $C_j$ , respectively. They can be computed in time*

$$O \left( |\mathcal{V}|^2 \cdot |\mathcal{E}| \cdot \sum_{i,j \in [m]} (N^{\text{faqw}(i,j)} + |\sigma_{ij}|) \cdot \log N \right).$$

In case all features in  $D$  are continuous, i.e.,  $C_j = \emptyset$  for all  $j \in [m]$ , then  $\text{faqw}(i, j) = \text{fhtw}$  [8] and the overall runtime becomes  $O(|\mathcal{V}|^2 \cdot |\mathcal{E}| \cdot m^2 \cdot N^{\text{fhtw}} \cdot \log N)$ . When some features are categorical, we can also bound the width  $\text{faqw}(i, j)$  and tensor size.

**Proposition 4.5.** *Let  $c = \max_{i,j} |C_i \cup C_j|$  be the maximum number of categorical variables for any  $\sigma_{ij}$ . Then,  $\text{faqw}(i, j) \leq \text{fhtw} + c - 1$  and  $|\sigma_{ij}| \leq N^{\min\{\rho^*, c\}}$ ,  $\forall i, j \in [m]$ .*

*For any query  $Q$  with  $\rho^* > \text{fhtw} + c - 1$ , there are infinitely many database instances for which*

$$\lim_{N \rightarrow \infty} \frac{N^{\rho^*}}{\sum_{i,j \in [m]} (N^{\text{faqw}(i,j)} + N^{\min\{\rho^*, c\}}) \log N} = \infty. \quad (33)$$

<sup>3</sup>We show in the proof of Proposition 4.4 in Appendix D how to express  $\sigma_{ij}$  and  $\mathbf{c}_j$  as FAQ-queries.

<sup>4</sup>Due to space limitation, these width notions are defined in Appendix C.

Our precomputation step takes strictly sub-output-size runtime for infinitely many queries and database instances. If we were to compute  $\sigma_{ij}$  on a training dataset with categorical variables one-hot encoded, then the complexity would raise to  $O(|\mathcal{V}|^2 \cdot |\mathcal{E}| \cdot m^2 \cdot N^{\text{ftw}+2d} \log N)$ , where  $d$  is the degree of the polynomial regression model or factorization machine.

#### 4.4 Point evaluation and gradient computation

We introduce two ideas for efficient point evaluation and gradient computation.

First, we employ a sparse representation of tensors in the *parameter space*. We need to evaluate the component functions of  $g$ , which are polynomial. In the **FaMa<sub>r</sub><sup>2</sup>** example, for instance, we evaluate expressions of the form

$$g_{\text{store, city}}(\theta) = \sum_{\ell=1}^r \theta_{\text{store}}^{(\ell)} \otimes \theta_{\text{city}}^{(\ell)}. \quad (34)$$

The result is a 2-way tensor whose CP-decomposition (a sum of rank-1 tensors) is already given by (34)! There is no point in materializing the result of  $g_{\text{store, city}}(\theta)$  and we instead keep it as is. Assuming  $N_c$  distinct cities and  $N_s$  distinct stores in the training dataset  $D$ , if we were to materialize the tensor, then we would end up with an  $\Omega(N_c N_s)$ -sized result for absolutely no gain in computational and space complexity, while the space complexity of the CP-decomposition is only  $O(N_c + N_s)$ . This is a prime example of factorization of the parameter space.

Second, we explain how to evaluate (16) and (17) with our sparse tensor representation. The same techniques can also be applied to evaluate (26) and (27) for PCA. There are two aspects of our solution worth spelling out: (1) how to multiply two tensors, e.g.,  $\sigma_{ij}$  and  $g_j(\theta)$ , and (2) how to exploit that some tensors have the same representation to speed up the point evaluation and gradient computation.

To answer question (1), we need to know the intrinsic dimension of the tensor  $\sigma_{ij}$ . In order to compute  $\Sigma g(\theta)$  in Example 13, we need to multiply  $\sigma_{ij}$  with  $g_j(\theta)$  for  $i = (\text{store, city})$  and  $j = (\text{city})$ . In a linear model,  $g_j(\theta) = \theta_j = \theta_{\text{city}}$ . In this case, when computing  $\sigma_{ij} \theta_{\text{city}}$  we marginalize away one city dimension of the tensor, while keeping the other two dimensions store, city. This is captured by the following query:

```
SELECT store, city, SUM( $\sigma_{i,j}.\text{val} * \theta_j.\text{val}$ )
FROM  $\sigma_{i,j}, \theta_j$  WHERE  $\sigma_{i,j}.\text{city} = \theta_j.\text{city}$ 
GROUP BY store, city;
```

where the tensors  $\sigma_{i,j}$  and  $\theta_j$  map (store, city) and respectively (city) to aggregate values. In words,  $\sigma_{ij} g_j(\theta)$  is computed by a group-by aggregate query where the group-by variables are precisely the variables in  $C_i$ .

For question (2), we use the CP-decomposition of the parameter space as discussed earlier. Suppose now we are looking at the  $\sigma_{ij}$  tensor where  $i = (\text{city})$  and  $j = (\text{store, city})$ . Note that this tensor has the identical representation as the above tensor, but it is a *different* tensor. In a **FaMa<sub>r</sub><sup>2</sup>** model, we would want to multiply this tensor with the component function  $g_j(\theta)$  defined in (34) above. We do so by multiplying it with each of the terms  $\theta_{\text{store}}^{(\ell)} \otimes \theta_{\text{city}}^{(\ell)}$ , one by one for  $\ell = 1, \dots, r$ , and then add up the result. Multiplying the tensor  $\sigma_{ij}$  with the first term  $\theta_{\text{store}}^{(1)} \otimes \theta_{\text{city}}^{(1)}$  corresponds precisely to the following query:

```
SELECT city, SUM( $\sigma_{i,j}.\text{val} * \theta_{\text{store}}^{(1)}.\text{val} * \theta_{\text{city}}^{(1)}.\text{val}$ )
FROM  $\sigma_{i,j}, \theta_{\text{store}}^{(1)}, \theta_{\text{city}}^{(1)}$  WHERE  $\sigma_{i,j}.\text{city} = \theta_{\text{city}}^{(1)}.\text{city}$  AND  $\sigma_{i,j}.\text{store} = \theta_{\text{store}}^{(1)}.\text{store}$ 
GROUP BY city;
```

where the tensors  $\sigma_{i,j}$ ,  $\theta_{\text{city}}^{(1)}$ , and  $\theta_{\text{store}}^{(1)}$  map (store, city), (city), and respectively (store) to aggregate values. Finally, to answer question (2), note that for the same column  $j$  (i.e., the same component function  $g_j(\theta)$ ), there can be multiple tensors  $\sigma_{ij}$  which have identical sparse representations. (This holds especially in models of degree  $> 1$ .)

In such cases, we have queries for point evaluation and gradient computation with identical from-where blocks but different select-group-by clauses, because the tensors have different group-by variables. Nevertheless, all such

queries can share computation as we can compute the from-where clause once for all of them and then scan this result to compute each specific tensor. This analysis gives rise to the following straightforward (and conservative) estimates.

For each  $j \in [m]$ , let  $d_j$  denote the degree and  $t_j$  denote the number of terms in the polynomial  $g_j$  (a component function of  $g$ ). Recall that  $p$  is the number of parameters.

**Proposition 4.6.** *Point evaluation (16) and gradient computation (17) can be computed in time  $O(\sum_{i,j \in [m]} t_i t_j d_i d_j |\sigma_{ij}|)$ , and respectively  $O(p \sum_{i,j \in [m]} t_i t_j d_i d_j |\sigma_{ij}|)$ .*

The times for point evaluation and gradient computation are:  $O(d^2 \sum_{i,j \in [m]} |\sigma_{ij}|)$  and respectively  $O(n^d \sum_{i,j \in [m]} |\sigma_{ij}|)$  for the  $\text{PR}^d$  model;  $O(r^2 d^2 \sum_{i,j \in [m]} |\sigma_{ij}|)$  and respectively  $O(nr^3 d^2 \sum_{i,j \in [m]} |\sigma_{ij}|)$  for the  $\text{FaMa}_r^d$  model; and  $O(\sum_{i,j \in [m]} |\sigma_{ij}|)$  and respectively  $O(n \sum_{i,j \in [m]} |\sigma_{ij}|)$  for PCA. Recall that the case for PCA is similar to that of LR, or equivalently  $\text{PR}_1$ .

Overall, there are a couple of remarkable facts regarding the overall runtime of our approach. Without loss of generality, suppose the number of iterations of BGD is bounded. (This bound is typically dimension-free, dependent on the Lipschitz constant of  $J$ .) Then, from Proposition 4.5, there are infinitely many queries for which the overall runtime of BGD is unboundedly better than the output size. First, our approach is faster than even the data-export step of the mainstream approach that uses an external tool to train the model. Second, it is often well-agreed upon that SGD is faster than BGD. However, a single iteration of SGD requires iterating through all data tuples, which takes time at least the output size. In particular, by training the model using BGD in the factorized form, BGD can be unboundedly faster than a single iteration of SGD.

## 4.5 Diagram of our structure-aware approach revisited

Figure 2 refines Figure 1 and depicts key ideas behind the performance improvements of our structure-aware framework over structure-agnostic learning approaches.

In structure-agnostic learning, a query engine takes the input relations (of size  $\leq N$ ) and joins them into a potentially much larger output relation of size  $|D|$ , which might in turn get blown up even more inside the machine learning tool. In our structure-aware framework, the input query, the input relations, and function  $h$  are first translated into FAQ [8], which is a language that is suitable for aggregate query specification and optimization. In particular, each entry  $\sigma_{i,j}$  (and  $c_j$ ) of our target matrix  $\Sigma$  (and vector  $\mathbf{c}$ ) is expressed as the answer to an FAQ query  $\varphi^{(i,j)}$  (or  $\varphi^{(j)}$ ). All those queries are fed into an FAQ query optimizer. The optimizer factorizes each query  $\varphi^{(i,j)}$  into small sub-queries  $\varphi_a^{(i,j)}, \varphi_b^{(i,j)}, \varphi_c^{(i,j)}, \dots$  and solves them individually. Each sub-query results in a relation of size  $\leq N^{\text{faqw}}$ , which can be much smaller than the size of the data matrix  $D$ . By solving the FAQ queries  $\varphi^{(i,j)}$ , we obtain  $\Sigma$  and  $\mathbf{c}$ , which are all that is needed as input for the convergence step in a batch gradient descent solver.

## 5 FD-Aware Optimization

In this section, we show how to exploit functional dependencies among variables to reduce the dimensionality of the optimization problem by eliminating functionally determined variables and re-parameterizing the model. We compute the quantities  $(\Sigma, \mathbf{c})$  on the subset of features that are not functionally determined, and then solve the lower-dimensional optimization problem. Finally, we recover the parameters in the original space in closed form. Exploiting functional dependencies drastically reduces the computation time for  $(\Sigma, \mathbf{c})$  and the gradient.

### 5.1 Introduction to the main ideas

Consider a query  $Q$  with categorical variables `country` and `city`. For simplicity, assume that there are only two countries “vietnam” and “england”, and 5 cities “saigon”, “hanoi”, “oxford”, “leeds”, and “bristol”. Under one-hot encoding, the corresponding features are encoded as indicators  $x_{\text{vietnam}}, x_{\text{england}}, x_{\text{saigon}}, x_{\text{hanoi}}, x_{\text{oxford}}, x_{\text{leeds}}, x_{\text{bristol}}$ . Since  $\text{city} \rightarrow \text{country}$  is a functional dependency (FD), for a given tuple  $\mathbf{x}$  in the training dataset, the



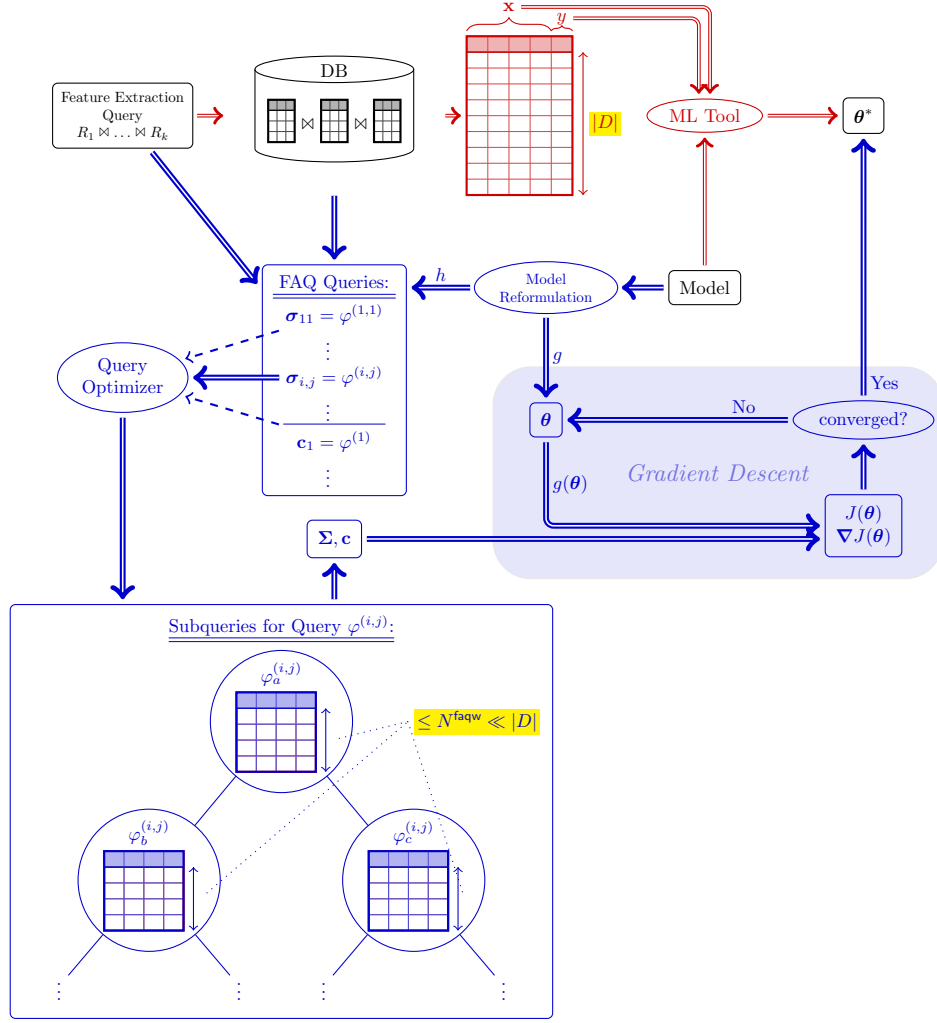


Figure 2: **Structure-aware** vs. **structure-agnostic** learning: Low-level diagram.

following hold:

$$x_{\text{vietnam}} = x_{\text{saigon}} + x_{\text{hanoi}} \quad (35)$$

$$x_{\text{england}} = x_{\text{oxford}} + x_{\text{leeds}} + x_{\text{bristol}}. \quad (36)$$

The first identity states that if a tuple has “vietnam” as the value for **country** ( $x_{\text{vietnam}} = 1$ ), then its value for **city** can only be either “saigon” or “hanoi”, i.e.,  $[x_{\text{saigon}}, x_{\text{hanoi}}]$  is either  $[1, 0]$  or  $[0, 1]$ , respectively. The second identity is explained similarly.

How do we express the identities such as (35) and (36) in a formal manner in terms of the input vectors  $\mathbf{x}_{\text{city}}$  and  $\mathbf{x}_{\text{country}}$ ? We can extract in a preprocessing step from the database a relation of the form  $R(\text{city}, \text{country})$  with **city** as primary key. Let  $N_{\text{city}}$  and  $N_{\text{country}}$  be the number of cities and countries, respectively. The predicate  $R(\text{city}, \text{country})$  is the sparse representation of a matrix  $\mathbf{R}$  of size  $N_{\text{country}} \times N_{\text{city}}$ , such that if  $\mathbf{x}_{\text{city}}$  is an indicator vector for **saigon**, then  $\mathbf{R}\mathbf{x}_{\text{city}}$  is an indicator for **vietnam**. In this language, identities (35) and (36) can be written simply as  $\mathbf{x}_{\text{country}} = \mathbf{R}\mathbf{x}_{\text{city}}$ . For example, in the above particular example  $N_{\text{city}} = 5$ ,  $N_{\text{country}} = 2$ , and

$$\mathbf{R} = \begin{array}{ccccc} \text{saigon} & \text{hanoi} & \text{oxford} & \text{leeds} & \text{bristol} \\ \text{vietnam} & 1 & 1 & 0 & 0 & 0 \\ \text{england} & 0 & 0 & 1 & 1 & 1 \end{array} \quad (37)$$

This relationship suggests a natural idea: replace any occurrence of statistics  $\mathbf{x}_{\text{country}}$  by its functionally determining quantity  $\mathbf{x}_{\text{city}}$ . Since these quantities are present only in the loss function  $\mathcal{L}$  via inner products  $\langle g(\mathbf{x}), h(\boldsymbol{\theta}) \rangle$ , such replacements result in a (typically) linear reparameterization of the loss. What happens next is less obvious, due to the presence of the nonlinear penalty function  $\Omega$ . Depending on the specific structure of FDs and the choice of  $\Omega$ , many parameters associated with redundant statistics, which do not affect the loss  $\mathcal{L}$ , can be optimized out directly with respect to the transformed  $\Omega$  penalty.

The remainder of this subsection is a gentle introduction of our idea in the presence of *one* simple FD in the LR model. Consider a query  $Q$  in which **city** and **country** are two of the categorical features and functionally determine one another via a matrix  $\mathbf{R}$  such that  $\mathbf{R}\mathbf{x}_{\text{city}} = \mathbf{x}_{\text{country}}$  for all  $\mathbf{x} = (\cdots, \mathbf{x}_{\text{city}}, \mathbf{x}_{\text{country}}, \cdots) \in D$ . We exploit this fact to “eliminate”  $\mathbf{x}_{\text{country}}$  as follows.

$$\begin{aligned} \langle g(\boldsymbol{\theta}), h(\mathbf{x}) \rangle &= \langle \boldsymbol{\theta}, \mathbf{x} \rangle = \sum_{j \notin \{\text{city}, \text{country}\}} \langle \boldsymbol{\theta}_j, \mathbf{x}_j \rangle + \langle \boldsymbol{\theta}_{\text{city}}, \mathbf{x}_{\text{city}} \rangle + \langle \boldsymbol{\theta}_{\text{country}}, \mathbf{x}_{\text{country}} \rangle \\ &= \sum_{j \notin \{\text{city}, \text{country}\}} \langle \boldsymbol{\theta}_j, \mathbf{x}_j \rangle + \langle \boldsymbol{\theta}_{\text{city}}, \mathbf{x}_{\text{city}} \rangle + \langle \boldsymbol{\theta}_{\text{country}}, \mathbf{R}\mathbf{x}_{\text{city}} \rangle \\ &= \sum_{j \notin \{\text{city}, \text{country}\}} \langle \boldsymbol{\theta}_j, \mathbf{x}_j \rangle + \left\langle \underbrace{\boldsymbol{\theta}_{\text{city}} + \mathbf{R}^\top \boldsymbol{\theta}_{\text{country}}}_{\boldsymbol{\gamma}_{\text{city}}}, \mathbf{x}_{\text{city}} \right\rangle. \end{aligned}$$

Define a new parameter vector  $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_j)_{j \in V - \{\text{country}\}}$  (note that there is no  $\boldsymbol{\gamma}_{\text{country}}$ ), and two functions  $\bar{g} : \mathbb{R}^{n-1} \rightarrow \mathbb{R}^{n-1}$ ,  $\bar{h} : \mathbb{R}^n \rightarrow \mathbb{R}^{n-1}$ :

$$\boldsymbol{\gamma}_j = \begin{cases} \boldsymbol{\theta}_j & j \neq \text{city} \\ \boldsymbol{\theta}_{\text{city}} + \mathbf{R}^\top \boldsymbol{\theta}_{\text{country}} & j = \text{city}. \end{cases} \quad (38)$$

$$\bar{g}(\boldsymbol{\gamma}) = \boldsymbol{\gamma} \quad (39)$$

$$\bar{h}_j(\mathbf{x}) = \mathbf{x}_j, \quad j \neq \text{city}. \quad (40)$$

Then, we can reparameterize  $J(\boldsymbol{\theta})$  in terms of  $\boldsymbol{\gamma}$  by

$$\begin{aligned} J(\boldsymbol{\theta}) &= \frac{1}{2|D|} \sum_{(\mathbf{x}, y) \in D} (\langle g(\boldsymbol{\theta}), h(\mathbf{x}) \rangle - y)^2 + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2 \\ &= \frac{1}{2|D|} \sum_{(\mathbf{x}, y) \in D} (\langle \bar{g}(\boldsymbol{\gamma}), \bar{h}(\mathbf{x}) \rangle - y)^2 + \frac{\lambda}{2} \left( \sum_{j \neq \text{city}} \|\boldsymbol{\gamma}_j\|_2^2 + \|\boldsymbol{\gamma}_{\text{city}} - \mathbf{R}^\top \boldsymbol{\theta}_{\text{country}}\|_2^2 + \|\boldsymbol{\theta}_{\text{country}}\|_2^2 \right). \end{aligned}$$

Note how  $\boldsymbol{\theta}_{\text{country}}$  has disappeared from the loss term, but it still remains in the penalty term. We now “optimize out”  $\boldsymbol{\theta}_{\text{country}}$  by observing that

$$\frac{1}{\lambda} \frac{\partial J}{\partial \boldsymbol{\theta}_{\text{country}}} = \mathbf{R}(\mathbf{R}^\top \boldsymbol{\theta}_{\text{country}} - \boldsymbol{\gamma}_{\text{city}}) + \boldsymbol{\theta}_{\text{country}} \quad (41)$$

By setting (41) to 0 we obtain  $\boldsymbol{\theta}_{\text{country}}$  in terms of  $\boldsymbol{\gamma}_{\text{city}}$ :

$$\boldsymbol{\theta}_{\text{country}} = (\mathbf{I}_{\text{country}} + \mathbf{R}\mathbf{R}^\top)^{-1} \mathbf{R}\boldsymbol{\gamma}_{\text{city}} = \mathbf{R}(\mathbf{I}_{\text{city}} + \mathbf{R}^\top \mathbf{R})^{-1} \boldsymbol{\gamma}_{\text{city}}, \quad (42)$$

where  $\mathbf{I}_{\text{country}}$  is the order- $N_{\text{country}}$  identity matrix and similarly for  $\mathbf{I}_{\text{city}}$ . We can thus express  $J$  and its gradient completely in terms of  $\boldsymbol{\gamma}$ :

$$J(\boldsymbol{\theta}) = \frac{1}{2|D|} \sum_{(\mathbf{x}, y) \in D} (\langle \bar{g}(\boldsymbol{\gamma}), \bar{h}(\mathbf{x}) \rangle - y)^2 + \frac{\lambda}{2} \left( \sum_{j \neq \text{city}} \|\boldsymbol{\gamma}_j\|_2^2 + \langle (\mathbf{I}_{\text{city}} + \mathbf{R}^\top \mathbf{R})^{-1} \boldsymbol{\gamma}_{\text{city}}, \boldsymbol{\gamma}_{\text{city}} \rangle \right) \quad (43)$$

$$\frac{1}{2} \frac{\partial \|\boldsymbol{\theta}\|_2^2}{\partial \boldsymbol{\gamma}_j} = \begin{cases} \boldsymbol{\gamma}_j & j \neq \text{city} \\ (\mathbf{I}_{\text{city}} + \mathbf{R}^\top \mathbf{R})^{-1} \boldsymbol{\gamma}_{\text{city}} & j = \text{city}. \end{cases} \quad (44)$$

(Appendix E.1 contains formal proofs of the above claims.) The gradient of the loss term is computed using the matrix  $\bar{\Sigma}$  and the vector  $\bar{\mathbf{c}}$  with respect to the pair  $(\bar{g}, \bar{h})$  of reduced dimensionality. The matrix  $(\mathbf{I}_{\text{city}} + \mathbf{R}^\top \mathbf{R})$  is a rank- $N_{\text{country}}$  update to the identity matrix  $\mathbf{I}_{\text{city}}$ , strictly positive definite and thus invertible. The inverse can be obtained using database aggregate queries; for numerical stability, one may compute its Cholesky decomposition which can *also* be expressed by aggregate queries. These “linear algebra via aggregate queries” computations are possible because our matrices admit a database interpretation, cf. Section 5.6.

## 5.2 Functional dependencies (FDs)

Composite FDs lead to more complex identities. For instance, the FD  $(\text{guest}, \text{hotel}, \text{date}) \rightarrow \text{room}$  leads to the identity  $x_{\text{room}} = \sum x_{\text{guest}} x_{\text{hotel}} x_{\text{date}}$ . Let  $R$  be a relation on attributes `guest`, `hotel`, `date`, and `room`, encoding this dependency, i.e.,  $R$  has a compound key  $(\text{guest}, \text{hotel}, \text{date})$ . Then, corresponding to  $R$  there is a matrix  $\mathbf{R}$  of dimension  $N_{\text{room}} \times N_{\text{guest}} \cdot N_{\text{hotel}} \cdot N_{\text{date}}$  for which  $\mathbf{x}_{\text{room}} = \mathbf{R}(\mathbf{x}_{\text{guest}} \otimes \mathbf{x}_{\text{hotel}} \otimes \mathbf{x}_{\text{date}})$ . Our results can be extended to the case of composite FDs, yet with a great notational burden; for the sake of clarity, we only state the results for *simple* FDs.

**Definition 4.** An FD is *simple* if its left-hand side is one variable.

Let a query  $Q$  in which there are  $k$  disjoint groups  $G_1, \dots, G_k$  of features, among other features. The  $i$ th group is  $G_i = \{f_i\} \cup S_i$ , where  $f_i$  is a feature,  $S_i$  a set of features, and  $f_i \rightarrow S_i$  is an FD. We shall refer to these as *groups of simple FDs*.

**Example 15.** In a typical feature extraction query for retailer customers, we have  $k = 3$  groups (in addition to other features): the first group contains `week`  $\rightarrow$  `month`  $\rightarrow$  `quarter`  $\rightarrow$  `year`, and thus  $f_1 = \text{week}$  and  $S_1 = \{\text{month}, \text{quarter}, \text{year}\}$ . In the second group,  $f_2 = \text{sku}$  and  $S_2 = \{\text{type}, \text{color}, \text{size}, \dots\}$  (a rather large group). In the third group  $f_3 = \text{store}$  and  $S_3 = \{\text{city}, \text{country}, \text{region}\}$ .  $\square$

For each feature  $c \in S_i$ , let  $\mathbf{R}_c$  denote the matrix for which  $\mathbf{x}_c = \mathbf{R}_c \mathbf{x}_{f_i}$ . For the sake of brevity, we also define a matrix  $\mathbf{R}_{f_i} = \mathbf{I}_{f_i}$  (the identity matrix of dimension equal to the active domain size of attribute  $f_i$ ), so the equality  $\mathbf{R}_c \mathbf{x}_{f_i} = \mathbf{x}_c$  holds for every  $c \in G_i$ .

The linear relationship holds even if the variables are not categorical. For example, consider the FD `sku`  $\rightarrow$  `price` (assuming every stock-keeping unit has a fixed sale-price). The relationship is modeled with a  $1 \times N_{\text{sku}}$  matrix  $\mathbf{R}$ , where the entry corresponding to a `sku` is its `price`. Then,  $\mathbf{R} \mathbf{x}_{\text{sku}} = x_{\text{price}}$  for any indicator vector  $\mathbf{x}_{\text{sku}}$ .

**Definition 5** (FD-reduced pairs of functions). Given a pair of functions  $g$  and  $h$  in our problem setting. Recall that  $C_j$ ’s are defined in Section 3.3, while  $S_k$ ’s are given in Definition 4. Define

$$K = \{j \in [m] \mid C_j \cap (S_1 \cup \dots \cup S_k) \neq \emptyset\}$$

( $K$  is the set of component functions of  $h$  containing at least one functionally determined variable.)

The group of simple FDs induces an *FD-reduced pair* of functions  $\bar{g} : \mathbb{R}^{p-|K|} \rightarrow \mathbb{R}^{m-|K|}$  and  $\bar{h} : \mathbb{R}^n \rightarrow \mathbb{R}^{m-|K|}$  as follows: The component functions of  $\bar{h}$  are obtained from the component functions of  $h$  by removing all component functions  $h_j$  for  $j \in K$ . Similarly,  $\bar{g}$  is obtained from  $g$  by removing all component functions  $g_j$  for which  $j \in K$ . Naturally, define the covariance matrix  $\bar{\Sigma}$  and the correlation vector  $\bar{\mathbf{c}}$  as in (13) and (14), but with respect to  $\bar{h}$ .

We next generalize the above technique to speedup the training of  $\text{PR}^d$  and **FaMa** under an arbitrary collection of simple FDs.

## 5.3 Polynomial regression under FDs

Recall the  $\text{PR}^d$ -model formulated in Example 3. Consider the set  $A_V$  of all tuples  $\mathbf{a}_V = (a_w)_{w \in V} \in \mathbb{N}^V$  of non-negative integers such that  $\|\mathbf{a}_V\|_1 \leq d$ . For any  $(\mathbf{x}, y) \in D$  and  $\mathbf{a} \in A_V$ , define  $\mathbf{x}^{\otimes \mathbf{a}} = \bigotimes_{v \in V} \mathbf{x}_v^{\otimes a_v}$ . In the  $\text{PR}^d$  model we have  $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathbf{a}})_{\|\mathbf{a}\|_1 \leq d}$ ,  $g(\boldsymbol{\theta}) = \boldsymbol{\theta}$ , and  $h_{\mathbf{a}}(\mathbf{x}) = \mathbf{x}^{\otimes \mathbf{a}}$ . If a feature, say  $v \in V$ , is non-categorical, then  $\mathbf{x}_v^{\otimes a_v} = x_v^{a_v}$ . If we knew  $x_v \in \{0, 1\}$ , then  $x_v^{a_v} = x_v$  and thus there is no need to have terms for which  $a_v > 1$ . A similar situation occurs when  $v$  is a categorical variable. To see this, let us consider a simple query

where  $V = \{b, c, w, t\}$ , and all four variables are categorical. Suppose the  $\text{PR}^d$  model has a term corresponding to  $\mathbf{a} = (a_b, a_c, a_w, a_t) = (0, 2, 0, 1)$ . The term of  $\langle \boldsymbol{\theta}, h(\mathbf{x}) \rangle$  indexed by tuple  $\mathbf{a}$  is of the form

$$\langle \boldsymbol{\theta}_{\mathbf{a}}, \mathbf{x}_c^{\otimes 2} \otimes \mathbf{x}_t \rangle = \langle \boldsymbol{\theta}_{\mathbf{a}}, \mathbf{x}_c \otimes \mathbf{x}_c \otimes \mathbf{x}_t \rangle.$$

For the dimensionality to match up,  $\boldsymbol{\theta}_{\mathbf{a}}$  is a 3rd-order tensor, say indexed by  $(i, j, k)$ . The above expression can be simplified as

$$\sum_i \sum_j \sum_k \boldsymbol{\theta}_{\mathbf{a}}(i, j, k) \cdot \mathbf{x}_c(i) \cdot \mathbf{x}_c(j) \cdot \mathbf{x}_t(k) = \sum_j \sum_k \boldsymbol{\theta}_{\mathbf{a}}(j, j, k) \mathbf{x}_c(j) \mathbf{x}_t(k),$$

where the equality holds due to the fact that  $\mathbf{x}_c(j)$  is idempotent. In particular, we only need the entries indexed by  $(j, j, k)$  of  $\boldsymbol{\theta}_{\mathbf{a}}$ . Equivalently, we write:

$$\langle \boldsymbol{\theta}_{\mathbf{a}}, \mathbf{x}_c \otimes \mathbf{x}_c \otimes \mathbf{x}_t \rangle = \langle ((\mathbf{I}_c \star \mathbf{I}_c)^\top \otimes \mathbf{I}_t) \boldsymbol{\theta}_{\mathbf{a}}, \mathbf{x}_c \otimes \mathbf{x}_c \otimes \mathbf{x}_t \rangle.$$

Multiplying on the left by the matrix  $(\mathbf{I}_c \star \mathbf{I}_c)^\top \otimes \mathbf{I}_t$  has precisely the same effect as selecting out only entries  $\boldsymbol{\theta}_{\mathbf{a}}(j, j, k)$  from the tensor  $\boldsymbol{\theta}_{\mathbf{a}}$ . More generally, in the  $\text{PR}^d$  model we can assume that all the indices  $\mathbf{a}_V = (a_v)_{v \in V}$  satisfy the condition that  $a_v \in \{0, 1\}$  whenever  $v$  is categorical. (This is in addition to the degree requirement that  $\|\mathbf{a}_V\|_1 \leq d$ .)

Given  $k$  groups of FDs represented by  $G_1, \dots, G_k$ , let  $G = \bigcup_{i=1}^k G_i$ ,  $S = \bigcup_{i=1}^k S_i$ ,  $\bar{G} = V - G$ ,  $\bar{S} = V - S$ , and  $F = \{f_1, \dots, f_k\}$ . For every non-empty subset  $T \subseteq [k]$ , define  $F_T = \{f_i \mid i \in T\}$ . Given a natural number  $q < d$ , and a non-empty set  $T \subseteq [k]$  with size  $|T| \leq d - q$ , define the collection

$$\mathcal{U}(T, q) = \{U \mid U \subseteq G \wedge U \cap G_i \neq \emptyset, \forall i \in T \wedge U \cap G_i = \emptyset, \forall i \notin T \wedge |U| \leq d - q\}. \quad (45)$$

For every tuple  $\mathbf{a}_{\bar{G}} \in \mathbb{N}^{\bar{G}}$  with  $\|\mathbf{a}_{\bar{G}}\|_1 = q < d$  and every  $U \in \mathcal{U}(T, q)$ , define the following matrices, which play the same role as  $\mathbf{I}_{\text{city}} + \mathbf{R}^\top \mathbf{R}$  in Section 5.1:

$$\mathbf{R}_{\mathbf{a}_{\bar{G}}, U} = \bigotimes_{\substack{w \in \bar{G} \\ a_w > 0}} \mathbf{I}_w \otimes \bigotimes_{i \in T} \bigotimes_{c \in U \cap G_i} \star \mathbf{R}_c. \quad (46)$$

$$\mathbf{B}_{\mathbf{a}_{\bar{G}}, T} = \sum_{W \in \mathcal{U}(T, \|\mathbf{a}_{\bar{G}}\|_1)} \mathbf{R}_{\mathbf{a}_{\bar{G}}, W}^\top \mathbf{R}_{\mathbf{a}_{\bar{G}}, W} \quad (47)$$

Note that the matrices  $\mathbf{B}_{\mathbf{a}_{\bar{G}}, T}$  can be further factorized as each of its terms is a tensor product, but we refrain from doing so here to avoid heavy notational complexity in the proofs and the theorem statement. See (96) and (97) in the appendix for examples of what we mean by factorization of these matrices. The following theorem reparameterizes  $J(\boldsymbol{\theta})$  for  $\text{PR}^d$  ( $d \geq 1$ ) to become  $\bar{J}(\boldsymbol{\gamma})$ . While  $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathbf{a}})$  is a vector indexed by tuples  $\mathbf{a} = \mathbf{a}_V \in \mathbb{N}^V$ , the new parameters  $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{\mathbf{b}})$  are indexed by integer tuples  $\mathbf{b} = \mathbf{b}_{\bar{S}} \in \mathbb{N}^{\bar{S}}$ .

**Theorem 5.1.** *Let the  $\text{PR}^d$ -model with parameters  $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathbf{a}_V})_{\|\mathbf{a}_V\|_1 \leq d}$ , and  $k$  groups of simple FDs  $G_i = \{f_i\} \cup S_i$ ,  $i \in [k]$ . Define the reparameterization:*

$$\boldsymbol{\gamma}_{\mathbf{b}_{\bar{S}}} = \begin{cases} \boldsymbol{\theta}_{(\mathbf{b}_{\bar{S}}, \mathbf{0}_G)} & \|\mathbf{b}_{\bar{S}}\|_1 = 0 \\ \sum_{U \in \mathcal{U}(T, q)} \mathbf{R}_{\mathbf{b}_{\bar{S}}, U}^\top \boldsymbol{\theta}_{(\mathbf{b}_{\bar{S}}, \mathbf{1}_{U \cap G})} & \|\mathbf{b}_{\bar{S}}\|_1 > 0, T := \{j \mid j \in F, b_{f_j} = 1\}, \end{cases}$$

Then, minimizing  $J(\boldsymbol{\theta})$  is equivalent to minimizing the function

$$\bar{J}(\boldsymbol{\gamma}) = \frac{1}{2} \boldsymbol{\gamma}^\top \bar{\boldsymbol{\Sigma}} \boldsymbol{\gamma} - \langle \boldsymbol{\gamma}, \bar{\mathbf{c}} \rangle + \frac{\lambda}{2} \Omega(\boldsymbol{\gamma}), \quad (48)$$

where

$$\Omega(\boldsymbol{\gamma}) = \sum_{\substack{\|\mathbf{b}_{\bar{S}}\|_1 \leq d \\ \|\mathbf{b}_F\|_1 = 0}} \left\| \boldsymbol{\gamma}_{\mathbf{b}_{\bar{S}}} \right\|_2^2 + \sum_{\substack{\|\mathbf{b}_{\bar{S}}\|_1 = q \\ q < d}} \sum_{\substack{T \subseteq [k] \\ 0 < |T| \leq d - q}} \left\langle \mathbf{B}_{\mathbf{b}_{\bar{S}}, T}^{-1} \boldsymbol{\gamma}_{(\mathbf{b}_{\bar{S}}, \mathbf{1}_{F_T \cap F})}, \boldsymbol{\gamma}_{(\mathbf{b}_{\bar{S}}, \mathbf{1}_{F_T \cap F})} \right\rangle. \quad (49)$$

(Recall  $\bar{\boldsymbol{\Sigma}}$  and  $\bar{\mathbf{c}}$  from Definition 5.)

The proof of this theorem (Appendix E.2) is technically involved.  $\bar{J}$  is defined above with respect to the FD-reduced pair of functions  $\bar{g}, \bar{h}$  and a reduced parameter space of  $\gamma$ . Its gradient is simple to compute, since

$$\frac{1}{2} \frac{\partial \Omega(\gamma)}{\partial \gamma_{\mathbf{b}_{\bar{S}}}} = \begin{cases} \gamma_{\mathbf{a}_{\bar{S}}}, & \mathbf{b}_F = \mathbf{0}_F, \\ \mathbf{B}_{\mathbf{b}_{\bar{G}}, T}^{-1} \gamma_{(\mathbf{b}_{\bar{G}}, \mathbf{1}_{F_T|F})}, & T = \{j \mid j \in F, b_j = 1\}. \end{cases} \quad (50)$$

Moreover, once a minimizer  $\gamma$  of  $\bar{J}$  is obtained, we can compute a minimizer  $\theta$  of  $J$  by setting

$$\theta_{\mathbf{a}_V} = \begin{cases} \gamma_{\mathbf{a}_{\bar{S}}}, & \|\mathbf{a}_G\|_1 = 0 \\ \mathbf{R}_{\mathbf{a}_{\bar{G}}, U} \mathbf{B}_{\mathbf{a}_{\bar{G}}, T}^{-1} \gamma_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{F_T|F})}, & \|\mathbf{a}_G\|_1 > 0, T := \{i \mid \exists c \in G_i, a_c > 0\} \end{cases} \quad (51)$$

Theorem 5.1 might be a bit difficult to grasp at first glance due to its generality. To give the reader a sense of how the theorem is applied in specific instances, Appendix E.4 and E.5 present two specializations of the theorem for (ridge) linear regression ( $\text{PR}^1$ ), and degree-2 polynomial regression ( $\text{PR}^2$ ).

## 5.4 Factorization machines under FDs

We now turn our attention to  $\text{FaMa}_r^2$ .

**Theorem 5.2.** *Consider the FaMa model of degree 2, rank  $r$ , parameters  $\theta = (\theta_i, (\theta_i^{(\ell)})_{\ell \in [r]})_{i \in V}$  and  $k$  groups of simple FDs  $G_i = \{f_i\} \cup S_i$ ,  $i \in [k]$ . Let  $G = \cup_{i \in [k]} G_i$ ,*

$$\beta_{f_i} = \sum_{\ell=1}^r \sum_{\{c,t\} \in \binom{G_i}{2}} \mathbf{R}_c^\top \theta_c^{(\ell)} \circ \mathbf{R}_t^\top \theta_t^{(\ell)}, \quad i \in [k] \quad (52)$$

and the following reparameterization:

$$\gamma_w = \begin{cases} \theta_w, & w \notin \bigcup_{i=1}^k G_i \\ \theta_{f_i} + \sum_{c \in S_i} \mathbf{R}_c^\top \theta_c + \beta_{f_i}, & w = f_i, i \in [k]. \end{cases}$$

$$\gamma_w^{(\ell)} = \begin{cases} \theta_w^{(\ell)}, & w \notin F \\ \theta_{f_i}^{(\ell)} + \sum_{c \in S_i} \mathbf{R}_c^\top \theta_c^{(\ell)}, & w = f_i, i \in [k]. \end{cases}$$

Then, minimizing  $J(\theta)$  is equivalent to minimizing the function  $\bar{J}(\gamma) = \frac{1}{2} \bar{g}(\gamma)^\top \bar{\Sigma} \bar{g}(\gamma) - \langle \bar{g}(\gamma), \bar{c} \rangle + \frac{\lambda}{2} \Omega(\gamma)$ , where

$$\Omega(\gamma) = \sum_{w \notin G} \|\gamma_w\|_2^2 + \sum_{i=1}^k \langle \mathbf{B}_i^{-1}(\gamma_{f_i} - \beta_{f_i}), (\gamma_{f_i} - \beta_{f_i}) \rangle + \sum_{\substack{\ell \in [r] \\ w \notin F}} \|\gamma_w^{(\ell)}\|_2^2 + \sum_{\substack{i \in [k] \\ \ell \in [r]}} \left\| \gamma_{f_i}^{(\ell)} - \sum_{c \in S_i} \mathbf{R}_c^\top \gamma_c^{(\ell)} \right\|_2^2. \quad (53)$$

(Recall  $\bar{g}$ ,  $\bar{\Sigma}$  and  $\bar{c}$  from Definition 5.)

In order to optimize  $\bar{J}$  with respect to  $\gamma$ , the following proposition provides a closed form formulae for the relevant gradient.

**Proposition 5.3.** *The gradient of  $\Omega(\gamma)$  defined in (53) can be computed using  $\delta_i^{(\ell)} = \sum_{c \in S_i} \mathbf{R}_c^\top \gamma_c^{(\ell)}$ , and*

$$\beta_{f_i} = \sum_{\ell=1}^r \left[ \left( \gamma_{f_i}^{(\ell)} - \frac{1}{2} \delta_i^{(\ell)} \right) \circ \delta_i^{(\ell)} - \frac{1}{2} \sum_{t \in S_i} \mathbf{R}_t^\top (\gamma_t^{(\ell)} \circ \gamma_t^{(\ell)}) \right]$$

Then,

$$\frac{1}{2} \frac{\partial \Omega(\gamma)}{\partial \gamma_w} = \begin{cases} \gamma_w, & w \notin G \\ \mathbf{B}_i^{-1}(\gamma_{f_i} - \beta_{f_i}), & w = f_i, i \in [k]. \end{cases} \quad (54)$$

$$\frac{1}{2} \frac{\partial \Omega(\gamma)}{\partial \gamma_w^{(\ell)}} = \begin{cases} \gamma_w^{(\ell)}, & w \notin G, \ell \in [r] \\ \gamma_{f_i}^{(\ell)} - \delta_i^{(\ell)} - \frac{1}{2} \delta_i^{(\ell)} \circ \frac{\partial \Omega(\gamma)}{\partial \gamma_{f_i}}, & w = f_i, \ell \in [r] \\ \gamma_w^{(\ell)} - \mathbf{R}_w \left[ \gamma_{f_i}^{(\ell)} \circ \frac{1}{2} \frac{\partial \Omega(\gamma)}{\partial \gamma_{f_i}} + \frac{1}{2} \frac{\partial \Omega(\gamma)}{\partial \gamma_{f_i}^{(\ell)}} \right], & w \in S_i, \ell \in [r]. \end{cases} \quad (55)$$

Suppose that the minimizer  $\gamma$  of  $\bar{J}$  has been obtained, then a minimizer  $\theta$  of  $J$  is available in closed form:

$$\theta_w = \begin{cases} \gamma_w, & w \in V \setminus G \\ \mathbf{R}_t \mathbf{B}_i^{-1}(\gamma_{f_i} - \beta_{f_i}), & \forall t \in G_i, i \in [k]. \end{cases}$$

$$\theta_w^{(\ell)} = \begin{cases} \gamma_w^{(\ell)}, & \forall w \notin F, \ell \in [r]. \\ \gamma_w^{(\ell)} - \delta_i^{(\ell)}, & w = f_i, \ell \in [r]. \end{cases}$$

This section shows that our technique applies to a non-linear model too. It should be obvious that a similar reparameterization works for  $\text{FaMa}_r^d$  for any  $d \geq 1$ . There is some asymmetry in the reparameterization of 1st-order parameters  $\theta_i$  and 2nd-order parameters  $\theta_i^{(\ell)}$  in Theorem 5.2, because we can solve a system of linear equation with matrix inverses, but we don't have closed form solutions for quadratic equations.

## 5.5 Principal Component Analysis under FDs

In this section, we show how to exploit functional dependencies to reduce the number of dimensions of the input to PCA by computing the top- $K$  eigenvectors and eigenvalues over the lower dimensional covariance matrix without the functionally determined features. We show that the eigenvalues of the lower dimensional problem are identical to those of the original problem, while the original eigenvectors can be derived from the solution to the lower dimensional problem.

Recall the functional dependencies of the form  $f_i \rightarrow S_i$ , the sets  $S$  of functionally determined variables and  $\bar{V} = V - S$  of all other variables, as in Section 5.2. Also, recall that each  $\mathbf{x} \in D$  is an  $n$ -dimensional vector, and for each categorical variable  $c$ , the component  $\mathbf{x}_c$  is an indicator vector.

We define  $\bar{\mathbf{x}}$  to be the vector of size  $q = |\bar{V}|$ , which is obtained by removing all components from  $\mathbf{x}$  that correspond to functionally determined variables (i.e., all  $\mathbf{x}_c$  for which  $c \in S$ ). Similar to (30) and (32), we can express the  $q$ -dimensional vector of means and the  $q \times q$  covariance matrix over  $\bar{\mathbf{x}}$ :

$$\bar{\boldsymbol{\mu}} = \frac{1}{|D|} \sum_{\mathbf{x} \in D} \bar{\mathbf{x}}$$

$$\bar{\boldsymbol{\Sigma}}_1 = \frac{1}{|D|} \sum_{\mathbf{x} \in D} \bar{\mathbf{x}} \bar{\mathbf{x}}^\top - \bar{\boldsymbol{\mu}} \bar{\boldsymbol{\mu}}^\top.$$

The covariance matrix  $\bar{\boldsymbol{\Sigma}}_1$  can be computed directly over the input database as in Example 14. The effect of computing a covariance matrix over  $\bar{\mathbf{x}}$  instead of  $\mathbf{x}$  is that its sparse tensor representation does not require the computation of any aggregate over a functionally determined variable.

For each functionally determined variable  $c \in S_i$ , the FD  $f_i \rightarrow c$  induces a mapping from a component in  $\bar{\mathbf{x}}$  to a component in  $\mathbf{x}$ . We define  $\mathbf{U}$  to be the rank- $q$  matrix of all such mappings, so that  $\mathbf{x} = \mathbf{U} \bar{\mathbf{x}}$  and each index  $\mathbf{u}_{kl} \in \mathbf{U}$  maps  $\bar{\mathbf{x}}_l$  to  $\mathbf{x}_k$ . For a variable  $c_k$ , let  $N_k$  be its domain size if  $c_k$  is categorical or one otherwise. Take two such variables  $c_k$  and  $c_l$ . Then, if  $c_k = c_l$  the entry  $\mathbf{u}_{kl}$  is the identity matrix  $I_{N_k}$ . If  $c_k \neq c_l$  and there is no functional dependency between them, then the entry  $\mathbf{u}_{kl}$  is the  $N_k \times N_l$  matrix of zeros. In case there is a functional dependency  $c_l \rightarrow c_k$ , the entry  $\mathbf{u}_{kl}$  is the  $N_k \times N_l$  matrix that encodes this functional dependency. For instance, in case  $l = \text{city}$  and  $k = \text{country}$ , the entry  $\mathbf{u}_{kl}$  is the  $N_{\text{country}} \times N_{\text{city}}$  matrix whose

entries  $(m, n)$  are one if the  $n$ -th city is located in the  $m$ -th country, or zero otherwise (as exemplified in (37) in Section 5.1, and generalized as  $\mathbf{R}_c$  matrices in Section 5.2). We can compute a sparse representation of the matrix  $\mathbf{U}$  as a collection of group-by queries over the input relations, and without materializing the result of the feature extraction query.

The following lemma shows that the eigenvalues of the covariance matrix are preserved under FDs while the eigenvectors are subject to a simple transformation.

**Lemma 5.4.** *For some  $K \leq q$ , let  $\lambda_1, \dots, \lambda_K > 0$  be the top- $K$  (positive-valued) eigenvalues of  $q \times q$  matrix  $\mathbf{U}^\top \mathbf{U} \bar{\Sigma}$  and  $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_K \in \mathbb{R}^q$  be the corresponding eigenvectors. Then  $\lambda_1, \dots, \lambda_K$  are also the top- $K$  eigenvalues of  $\Sigma_1$ . Moreover, the eigenvectors of  $\Sigma_1$  are*

$$\forall j \in [K] : \boldsymbol{\theta}_j = \frac{1}{\lambda_j} \mathbf{U} \bar{\Sigma}_1 \boldsymbol{\eta}_j$$

*Proof.* First note that  $\boldsymbol{\mu} = \frac{1}{|D|} \sum_{\mathbf{x} \in D} \mathbf{x} = \frac{1}{|D|} \sum_{\mathbf{x} \in D} \mathbf{U} \bar{\mathbf{x}} = \mathbf{U} \bar{\boldsymbol{\mu}}$  and  $\Sigma_1 = \frac{1}{|D|} \sum_{\mathbf{x} \in D} \mathbf{x} \mathbf{x}^\top - \boldsymbol{\mu} \boldsymbol{\mu}^\top = \frac{1}{|D|} \sum_{\mathbf{x} \in D} \mathbf{U} \bar{\mathbf{x}} \bar{\mathbf{x}}^\top \mathbf{U}^\top - \mathbf{U} \bar{\boldsymbol{\mu}} \bar{\boldsymbol{\mu}}^\top \mathbf{U}^\top = \mathbf{U} (\frac{1}{|D|} \sum_{\mathbf{x} \in D} \bar{\mathbf{x}} \bar{\mathbf{x}}^\top - \bar{\boldsymbol{\mu}} \bar{\boldsymbol{\mu}}^\top) \mathbf{U}^\top = \mathbf{U} \bar{\Sigma}_1 \mathbf{U}^\top$ . For any eigen-pair  $(\lambda, \boldsymbol{\theta})$  of  $\Sigma_1$ , it holds  $\Sigma_1 \boldsymbol{\theta} = \lambda \boldsymbol{\theta}$  by definition. Thus,  $\mathbf{U} \bar{\Sigma}_1 \mathbf{U}^\top \boldsymbol{\theta} = \lambda \boldsymbol{\theta}$ . Multiplying both sides by  $\mathbf{U}^\top$  to the left, we obtain  $\mathbf{U}^\top \mathbf{U} \bar{\Sigma}_1 \mathbf{U}^\top \boldsymbol{\theta} = \lambda \mathbf{U}^\top \boldsymbol{\theta}$ . Hence,  $(\lambda, \mathbf{U}^\top \boldsymbol{\theta})$  is an eigen-pair of  $\mathbf{U}^\top \mathbf{U} \bar{\Sigma}_1$ . Since  $\mathbf{U}^\top \mathbf{U}$  is full-ranked, the set of positive eigenvalues of  $\Sigma_1$  is identical to that of  $\mathbf{U}^\top \mathbf{U} \bar{\Sigma}_1$ . Moreover, let  $(\lambda, \boldsymbol{\eta})$  be any of the eigen-pairs of  $\mathbf{U}^\top \mathbf{U} \bar{\Sigma}_1$  in which  $\lambda > 0$ , then the corresponding eigenvector of  $\Sigma_1$  may be obtained by the identity:  $\mathbf{U} \bar{\Sigma}_1 \mathbf{U}^\top \boldsymbol{\theta} = \lambda \boldsymbol{\theta}$ , which yields  $\mathbf{U} \bar{\Sigma}_1 \boldsymbol{\eta} = \lambda \boldsymbol{\theta}$ . This gives  $\boldsymbol{\theta} = (1/\lambda) \mathbf{U} \bar{\Sigma}_1 \boldsymbol{\eta}$  to conclude the proof.  $\square$

## 5.6 Linear algebra with database queries

To apply the above results, we need to solve several computational primitives. The first primitive is to compute the matrix inverse  $\mathbf{B}_{T,q}^{-1}$  and its product with another vector. This task can be done by either explicitly computing the inverse, or computing the Cholesky decomposition of the matrix  $\mathbf{B}_{T,q}$ . We next explain how both of these tasks can be done using database queries.

**Maintaining the matrix inverse with rank-1 updates** Using Sherman-Morrison-Woodbury formula [35], we can incrementally compute the inverse of the matrix  $\mathbf{I} + \sum_{c \in G_i} \mathbf{R}_c^\top \mathbf{R}_c$  as follows. Let  $S \subset G_i$  be some subset and suppose we have already computed the inverse for  $\mathbf{M}_S = \mathbf{I} + \sum_{s \in S} \mathbf{R}_s^\top \mathbf{R}_s$ . We now explain how to compute the inverse for  $\mathbf{M}_{S \cup \{c\}} = \mathbf{I} + \sum_{s \in S \cup \{c\}} \mathbf{R}_s^\top \mathbf{R}_s$ . For concreteness, let the matrix  $\mathbf{R}_c$  map city to country. For each country `country`, let  $\mathbf{e}_{\text{country}}$  denote the 01-vector where there is a 1 for each city the country has. For example,  $\mathbf{e}_{\text{cuba}} = [1 \ 1 \ 0 \ 0 \ 0]^\top$ . Then,  $\mathbf{R}_c^\top \mathbf{R}_c = \sum_{\text{country}} \mathbf{e}_{\text{country}} \mathbf{e}_{\text{country}}^\top$ . And thus, starting with  $\mathbf{M}_S$ , we apply the Sherman-Morrison-Woodbury formula for each country, such as:

$$(\mathbf{M} + \mathbf{e}_{\text{cuba}} \mathbf{e}_{\text{cuba}}^\top)^{-1} = \mathbf{M}^{-1} - \frac{\mathbf{M}^{-1} \mathbf{e}_{\text{cuba}} \mathbf{e}_{\text{cuba}}^\top \mathbf{M}^{-1}}{1 + \mathbf{e}_{\text{cuba}}^\top \mathbf{M}^{-1} \mathbf{e}_{\text{cuba}}}. \quad (56)$$

This update can be done with database aggregate queries, because  $\mathbf{e}_{\text{cuba}}^\top \mathbf{M}^{-1} \mathbf{e}_{\text{cuba}}$  is a sum of entries  $(i, j)$  in  $\mathbf{M}^{-1}$  where both  $i$  and  $j$  are cities in `cuba`;  $\mathbf{v} = \mathbf{M}^{-1} \mathbf{e}_{\text{cuba}}$  is the sum of columns of  $\mathbf{M}^{-1}$  corresponding to `cuba`; and  $\mathbf{M}^{-1} \mathbf{e}_{\text{cuba}} \mathbf{e}_{\text{cuba}}^\top \mathbf{M}^{-1}$  is exactly  $\mathbf{v} \mathbf{v}^\top$ .

Overall, each update (56) can be done in  $O(N_{\text{city}}^2)$ -time, for an overall runtime of  $O(N_{\text{city}}^2 N_{\text{country}})$ . This runtime should be contrasted with Gaussian-elimination-based inverse computation time, which is  $O(N_{\text{city}}^3)$ . When the FDs form a chain, the blocks are nested inside one another, and thus each update is even cheaper as we do not have to access all  $N_{\text{city}}^2$  entries.

**Maintaining a Cholesky decomposition with rank- $k$  update** Maintaining a matrix inverse can be numerically unstable. It would be best to compute a Cholesky decomposition of the matrix, since this strategy is numerically more stable. There are known rank-1 update algorithms [30, 23], using strategies similar to the inverse rank-1 update above. A further common computational primitive is to multiply a tensor product with a vector, such as in  $(\mathbf{B}_i^{-1} \otimes \mathbf{B}_j^{-1}) \boldsymbol{\gamma}_{f_i f_j}$  (also expressible as aggregate queries).

## 5.7 Discussion

### Diagram of our structure-aware learning in the presence of FDs

Figure 3 depicts the enhancements that we introduce to our framework in order to take advantage of FDs in the input database instance and reduce our previous runtime even further (but still compute *the same*  $\theta^*$  as before).

As explained earlier, computing each entry of the matrix  $\Sigma$  and of the vector  $c$  requires solving an FAQ query. However, by utilizing FDs we can filter out many of those entries as unneeded for later stages, thus significantly reducing the number of FAQ queries that we have to solve. After the filtering process,  $\Sigma$  and  $c$  shrink down to  $\bar{\Sigma}$  and  $\bar{c}$ , which we compute and feed to gradient-descent (GD). We also filter the function  $g$  down to  $\bar{g}$  and feed the latter to GD. Now, we run GD in the space of  $\gamma$  (instead of the original higher-dimensional space of  $\theta$ ). During each iteration of GD, in order to compute the objective function  $\bar{J}(\gamma)$  and its gradient  $\nabla \bar{J}(\gamma)$ , we need to use the matrices  $\mathbf{R}$  that represent the functional dependencies. And after GD finishes, we have to convert the resulting optimal solution  $\gamma^*$  back into the original space to get  $\theta^*$ . Such conversion also requires the FD-matrices  $\mathbf{R}$ .

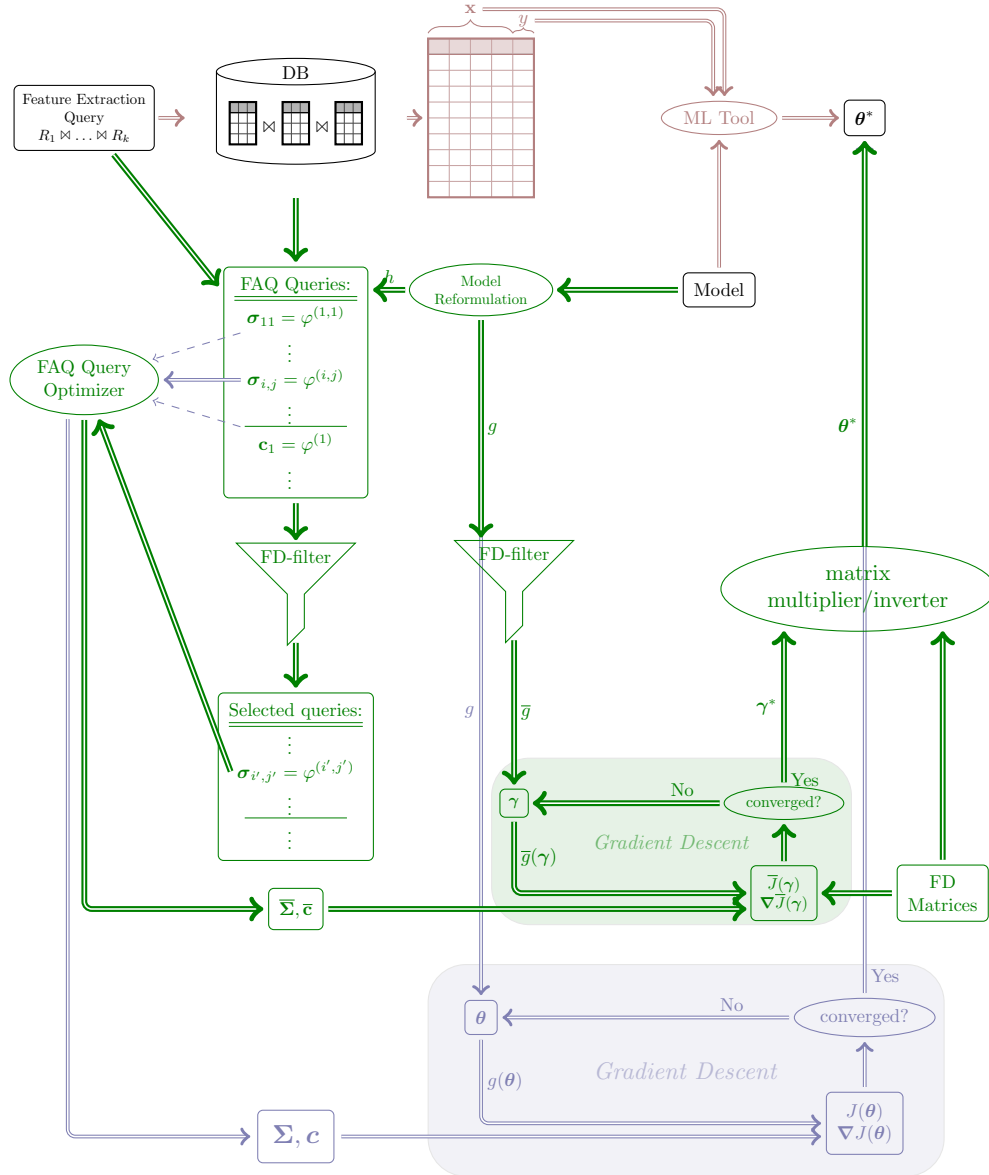


Figure 3: Structure-aware learning: **with FDs** vs. **without FDs**.



## Impact of FDs on model complexity

The prevalence of FDs presents new challenges from both computational and statistical viewpoints. On the one hand, a reasonable and well-worn rule of thumb in statistics dictates that one should always eliminate features that are functionally dependent on others, because this helps reduce both computation and model’s complexity, which in turn leads to reduced generalization error (as also noted in [44]). On the other hand, the statistical effectiveness of such a rule is difficult to gauge when the nature of dependence goes beyond linearity. In such scenarios, it might be desirable to keep some redundant variables, but only if they help construct simpler forms of regression/classification functions, leading to improved approximation ability for the model class.

It is, however, difficult to know a priori which redundant features lead to simple functions. Therefore, the problem of dimensionality reduction cannot be divorced from the model class under consideration. While this remains unsolved in general, in this work we restricted ourselves to specific classes of learning models, the complexity of which may still be varied through regularization via (non-linear) penalties. Within a regularized parametric model class, we introduced dimensionality reduction techniques (variable elimination and re-parameterization) that may not fundamentally change the model’s capacity. The reduction in the number of parameters may still help reduce the variance of parameter estimates, leading to improved generalization error guarantees.

## Impact of FDs on computational complexity

Model reparameterization under FDs does not lower the *data complexity* from Proposition 4.4 for the computation of the sparse tensor representation. Under a simple FD  $A \rightarrow B$ , the number of categories of the functionally determined categorical variable  $B$  cannot exceed that of the functionally determining categorical variable  $A$ . This means that by avoiding the computation of aggregates involving  $B$ , the data complexity for the computation of the sparse tensor representation with both  $A$  and  $B$  is the same as with  $A$  only.

Computing less aggregates means however a reduction in the *query complexity*. In case only  $q < n$  variables functionally determine the entire set of  $n$  variables, the dimensionality of  $\overline{\Sigma}$  for  $\text{PR}^d$  is  $\Theta(q^d) \times \Theta(q^d)$ , which is much smaller than the dimensionality  $\Theta(n^d) \times \Theta(n^d)$  of  $\Sigma$ . This reduction can be significant: In one of our experiments with  $\text{PR}_2$  on the Retailer dataset  $\mathbf{v}_4$ , there is a reduction from 46M to 36M entries in the sparse tensor representation of  $\Sigma$  and  $\mathbf{c}$ . Proposition E.1 in Appendix E.3 provides the corresponding version of Corollary 4.2 with respect to  $\overline{\Sigma}$ .

This reduction in the query complexity comes at a price: The gradient solver has a new data-dependent computation in the regularizer. For instance, under the functional dependency  $\text{city} \rightarrow \text{country}$  used in Section 5.1,  $\theta_{\text{country}} = (\mathbf{I}_{\text{country}} + \mathbf{R}\mathbf{R}^\top)^{-1}\mathbf{R}\gamma_{\text{city}}$  where  $\mathbf{R}$  is a matrix that maps between cities and countries in the input database. Computing this linear algebra expression takes time  $O(N_{\text{city}}^2 N_{\text{country}})$  as explained in Section 5.6, where  $N_{\text{city}}$  and  $N_{\text{country}}$  are the number of cities (categories for the *city* categorical feature) and respectively countries. Assuming these quantities are small, the reduction in the number of aggregates vastly dominates the modest increase in the complexity of the additional linear algebra expression. Figure 7 in Section 7 indeed shows that using one single functional dependency for Retailer leads to a 3.5× performance speedup.

## 6 The Design and Implementation of AC/DC

In this section, we present the design of AC/DC, which is our implementation of the algorithms and optimizations for the end-to-end computation of square loss problems presented in the previous sections. AC/DC computes each entry in the sparse tensor representation of the problem as an aggregate over the feature extraction join query, following the SQL encoding developed in previous sections, e.g., Examples 11 and 13. Two key optimizations used by AC/DC for the computation of these aggregates are: (1) *Factorized computation* of aggregates over the feature extraction query, with low complexity (Section 6.1); and (2) *Massively shared computation* across the aggregates (Section 6.2). AC/DC also exploits functional dependencies to reduce the dimensionality of the problem. By design, AC/DC does not achieve the complexity bound from Proposition 4.4. This is because it would need different query plans for different subsets of the aggregates. Instead, it uses one query plan for all these aggregates. This increases the opportunity to share computation across the aggregates, which proved much more beneficial for the overall performance.

```

aggregates (variable order  $\Delta$ , varMap, relation ranges  $R_1[x_1, y_1], \dots, R_d[x_d, y_d]$ )
 $A = \text{root}(\Delta)$ ; context =  $\pi_{\text{dep}(A)}(\text{varMap})$ ; reset( $\text{aggregates}_A$ ); # $\text{aggregates} = |\text{aggregates}_A|$ ;

if ( $\text{dep}(A) \neq \text{anc}(A)$ ) {  $\text{aggregates}_A = \text{cache}_A[\text{context}]$ ; if ( $\text{aggregates}_A[0] \neq \emptyset$ ) return; }

foreach  $i \in [d]$  do  $R_i[x'_i, y'_i] = R_i[x_i, y_i]$ ;
foreach  $a \in \bigcap_{i \in [d]} \text{such that } A \in \text{vars}(R_i) \pi_A(R_i[x_i, y_i])$  do {
  foreach  $i \in [d]$  such that  $A \in \text{vars}(R_i)$  do
    find range  $R_i[x'_i, y'_i] \subseteq R_i[x_i, y_i]$  such that  $\pi_A(R_i[x'_i, y'_i]) = \{(A : a)\}$ ;
  switch ( $A$ ) :
    continuous feature :  $\lambda_A = [\{() \mapsto 1\}, \{() \mapsto a^1\}, \dots, \{() \mapsto a^{2 \cdot \text{degree}}\}]$ ;
    categorical feature :  $\lambda_A = [\{() \mapsto 1\}, \{a \mapsto 1\}]$ ;
    no feature :  $\lambda_A = [\{() \mapsto 1\}]$ ;
  switch ( $\Delta$ ) :
    leaf node  $A$  :
      foreach  $l \in [\#\text{aggregates}]$  do {  $[i_0] = \mathcal{R}_A[l]$ ;  $\text{aggregates}_A[l] += \lambda_A[i_0]$ ; }
    inner node  $A(\Delta_1, \dots, \Delta_k)$  :
      foreach  $j \in [k]$  do
        aggregates( $\Delta_j$ , varMap  $\times \{(A : a)\}$ , ranges  $R_1[x'_1, y'_1], \dots, R_d[x'_d, y'_d]$ );
        if ( $\forall j \in [k] : \text{aggregates}_{\text{root}(\Delta_j)}[0] \neq \emptyset$ )
          foreach  $l \in [\#\text{aggregates}]$  do {
             $[i_0, i_1, \dots, i_k] = \mathcal{R}_A[l]$ ;
             $\text{aggregates}_A[l] += \lambda_A[i_0] \times \prod_{j \in [k]} \text{aggregates}_{\text{root}(\Delta_j)}[i_j]$ ; }
          }
  }
if ( $\text{dep}(A) \neq \text{anc}(A)$ )  $\text{cache}_A[\text{context}] = \text{aggregates}_A$ ;

```

Figure 4: Algorithm for factorized computation of aggregates. Each aggregate maps tuples over its group-by variables to scalars. The parameters of the initial call are the variable order  $\Delta$  of the feature extraction query, an empty map from variables to values, and the full range of tuples for each input relation  $R_1, \dots, R_d$ .

## 6.1 Factorized aggregate computation

Factorized aggregate computation relies on a variable order for the query  $Q$  to avoid redundant computation. In this paper, we assume that we are given a variable order. Prior work discusses the query optimization problem of finding good orders [13, 8].

*Variable Orders.* State-of-the-art query evaluation uses relation-at-a-time query plans. We use variable-at-a-time query plans, which we call variable orders. These are partial orders on the variables in the query, capture the join dependencies in the query, and dictate the order in which we solve each join variable. For each variable, we join all relations with that variable. Our choice is motivated by the complexity of join evaluation: Relation-at-a-time query plans are provably suboptimal, whereas variable-at-a-time query plans can be chosen to be optimal [53].

For a query  $Q$ , a variable  $X$  *depends* on a variable  $Y$  if both are in the schema of a relation in  $Q$ .

**Definition 1** (adapted from [56]). A variable order  $\Delta$  for a join query  $Q$  is a pair  $(F, \text{dep})$ , where  $F$  is a rooted forest with one node per variable in  $Q$ , and  $\text{dep}$  is a function mapping each variable  $X$  to a set of variables in  $F$ . It satisfies the following constraints:

- For each relation in  $Q$ , its variables lie along the same root-to-leaf path in  $F$ .
- For each variable  $X$ ,  $\text{dep}(X)$  is the subset of its ancestors in  $F$  on which the variables in the subtree rooted at  $X$  depend.

Without loss of generality, we use variables orders that are trees instead of forests. We can convert a forest into a tree by adding to each relation the same dummy join variable that takes a single value. For a variable  $X$

in the variable order  $\Delta$ ,  $anc(X)$  is the set of all ancestor variables of  $X$  in  $\Delta$ . The set of variables in  $\Delta$  (schema of a relation  $R$ ) is denoted by  $vars(\Delta)$  ( $vars(R)$  respectively) and the variable at the root of  $\Delta$  is denoted by  $root(\Delta)$ .

**Example 16.** Figure 6(a) shows a variable order for the natural join of relations  $R(A, B, C)$ ,  $T(B, D)$ , and  $S(A, E)$ . Then,  $anc(D) = \{A, B\}$  and  $dep(D) = \{B\}$ , i.e.,  $D$  has ancestors  $A$  and  $B$ , yet it only depends on  $B$ . Given  $B$ , the variables  $C$  and  $D$  are independent of each other. For queries with group-by variables, we choose a variable order where these variables sit above the other variables [13].  $\square$

Figure 4 presents the AC/DC algorithm for factorized computation of SQL aggregates over the feature extraction query  $Q$ . The backbone of the algorithm without the code in boxes explores the factorized join of the input relations  $R_1, \dots, R_d$  over a variable order  $\Delta$  of  $Q$ . As it traverses  $\Delta$  in depth-first preorder, it assigns values to the query variables. The assignments are kept in `varMap` and used to compute aggregates by the code in the boxes.

The relations are sorted following a depth-first pre-order traversal of  $\Delta$ . Each call takes a range  $[x_i, y_i]$  of tuples in each relation  $R_i$ . Initially, these ranges span the entire relations. Once the root variable  $A$  in  $\Delta$  is assigned a value  $a$  from the intersection of possible  $A$ -values from the input relations, these ranges are narrowed down to those tuples with value  $a$  for  $A$ .

To compute an aggregate over the variable order  $\Delta$  rooted at  $A$ , we first initialize the aggregate to zeros. This is needed since the aggregates might have been used earlier for different assignments of ancestor variables in  $\Delta$ . We next check whether we previously computed the aggregate for the same assignments of variables in  $dep(A)$ , denoted by `context`, and cached it in a map `cacheA`. Caching is useful when  $dep(A)$  is strictly contained in  $anc(A)$ , since this means that the aggregate computed at  $A$  does not need to be recomputed for distinct assignments of variables in  $anc(A) \setminus dep(A)$ . In this case, we probe the cache using as key the assignments in `varMap` of the  $dep(A)$  variables: `cacheA[context]`. If we have already computed the aggregates over that assignment for  $dep(A)$ , then we can just reuse the previously computed aggregates and avoid recomputation.

If  $A$  is a group-by variable, then we compute a map from each  $A$ -value  $a$  to a function of  $a$  and aggregates computed at children of  $A$ , if any. If  $A$  is not a group-by variable, then we compute a map from the empty value  $()$  to such a function; in this latter case, we could have just computed the aggregate instead of the map though we use the map for uniformity. In case there are group-by variables under  $A$ , the computation at  $A$  returns maps whose keys are tuples over all these group-by variables in  $vars(\Delta)$ .

**Example 17.** Consider a feature extraction query  $Q$  with the variable order  $\Delta$  in Figure 6(a). We first compute the assignments for  $A$  as  $Q_A = \pi_A R \bowtie \pi_A T$ . For each assignment  $a \in Q_A$ , we then find assignments for variables under  $A$  within the narrow ranges of tuples that contain  $a$ . The assignments for  $B$  in the context of  $a$  are given by  $Q_B^a = \pi_B(\sigma_{A=a} R) \bowtie \pi_B S$ . For each  $b \in Q_B^a$ , the assignments for  $C$  and  $D$  are given by  $Q_C^{a,b} = \pi_C(\sigma_{A=a \wedge B=b} R)$  and  $Q_D^b = \pi_D(\sigma_{B=b} S)$ . Since  $D$  depends on  $B$  and not on  $A$ , the assignments for  $D$  under a given  $b$  are repeated for every occurrence of  $b$  with assignments for  $A$ . The assignments for  $E$  given  $a \in Q_A$  are computed as  $Q_E^a = \pi_E(\sigma_{A=a} T)$ .

Consider the aggregate  $COUNT(Q)$ . The count at each variable  $X$  is computed as the sum over all value assignments of  $X$  of the product of the counts at the children of  $X$  in  $\Delta$ ; if  $X$  is a leaf in  $\Delta$ , the product at children is considered 1. For our variable order, this computation is captured by the following factorized expression:

$$COUNT = \sum_{a \in Q_A} 1 \cdot \left( \sum_{b \in Q_B^a} 1 \cdot \left( \sum_{c \in Q_C^{a,b}} 1 \cdot V_D(b) \right) \right) \cdot \sum_{e \in Q_E^a} 1 \quad (57)$$

where  $V_D(b) = \sum_{d \in Q_D^b} 1$  is cached the first time we encounter the assignment  $b$  for  $B$  and reused for all subsequent occurrences of this assignment under assignments for  $A$ .

Summing all  $X$ -values in the result of  $Q$  for a variable  $X$  is done similarly, with the difference that at the variable  $X$  in  $\Delta$  we compute the sum of the values of  $X$  weighted by the product of the counts of their children.

For instance, the aggregate  $\text{SUM}(C * E)$  is computed over our variable order by the following factorized expression:

$$\text{SUM}(C * E) = \sum_{a \in Q_A} 1 \cdot \left( \sum_{b \in Q_B^a} 1 \cdot \left( \sum_{c \in Q_C^{a,b}} c \cdot V_D(b) \right) \right) \cdot \sum_{e \in Q_E^a} e \quad (58)$$

To compute the aggregate  $\text{SUM}(C * E)$  **GROUP BY**  $A$ , we compute  $\text{SUM}(C * E)$  for each assignment for  $A$  instead of marginalizing away  $A$ . The result is a map from  $A$ -values to values of  $\text{SUM}(C * E)$ .  $\square$

A good variable order may include variables that are not explicitly used in the optimization problem. This is the case of join variables whose presence in the variable order ensures a good factorization. For instance, if we remove the variable  $B$  from the variable order in Figure 6(a), the variables  $C, D$  are no longer independent and we cannot factorize the computation over  $C$  and  $D$ . AC/DC exploits the conditional independence enabled by  $B$ , but computes no aggregate over  $B$  if this is not required in the problem.

The complexity bound in Proposition 4.4 is achieved by factorizing the computation of each aggregate in  $\Sigma$  over a variable order that has all group-by variables for this aggregate above all other variables. Thus, different aggregates can be computed over different variable orders.

## 6.2 Shared computation of aggregates

Section 6.1 explains how to factorize the computation of one aggregate in  $\Sigma$ ,  $\mathbf{c}$ , and  $s_Y$  over the join of database relations. In this section we show how to share the computation across aggregates.

**Example 18.** We consider the factorized expression of the aggregates  $\text{SUM}(C)$  and  $\text{SUM}(E)$  over  $\Delta$ :

$$\text{SUM}(C) = \sum_{a \in Q_A} 1 \cdot \left( \sum_{b \in Q_B^a} 1 \cdot \left( \sum_{c \in Q_C^{a,b}} c \cdot V_D(b) \right) \right) \cdot \sum_{e \in Q_E^a} 1 \quad (59)$$

$$\text{SUM}(E) = \sum_{a \in Q_A} 1 \cdot \left( \sum_{b \in Q_B^a} 1 \cdot \left( \sum_{c \in Q_C^{a,b}} 1 \cdot V_D(b) \right) \right) \cdot \sum_{e \in Q_E^a} e \quad (60)$$

We can share computation across the expressions (57) to (60) since they are similar. For instance, given an assignment  $b$  for  $B$ , all these aggregates need  $V_D(b)$ . Similarly, for a given assignment  $a$  for  $A$ , the aggregates (58) and (60) can share the computation of the sum aggregate over  $Q_E^a$ . For assignments  $a \in Q_A$  and  $b \in Q_B^a$ , expressions (58) and (59) can share the computation of the sum aggregate over  $Q_C^{a,b}$ .  $\square$

To share as much computation as possible between aggregates, AC/DC computes all aggregates together over a single variable order, which significantly improves the data locality of the aggregate computation. This approach does not follow Proposition 4.4 that assumes that each aggregate is computed over its respective best variable order. AC/DC thus decidedly sacrifices the goal of achieving the lowest-known complexity for individual aggregates for the sake of sharing as much computation as possible across these aggregates.

### Aggregate Decomposition and Registration.

For a model of degree  $degree$  and a set of variables  $\{A_l\}_{l \in [n]}$ , we have aggregates of the form  $\text{SUM}(\prod_{l \in [n]} A_l^{d_l})$ , possibly with a group-by clause, such that  $0 \leq \sum_{l \in [n]} d_l \leq 2 \cdot degree$ ,  $d_l \geq 0$ , and all categorical variables are turned into group-by variables. The reason for  $2 \cdot degree$  is due to the  $\Sigma$  matrix used to compute the gradient of the loss function (17), which pairs any two features of degree up to  $degree$ . Each aggregate is thus defined uniquely by a monomial  $\prod_{l \in [n]} A_l^{d_l}$ ; we may discard the variables with exponent 0. For instance, the monomial for  $\text{SUM}(C * E)$  is CE while for  $\text{SUM}(C * E)$  **GROUP BY**  $A$  is ACE.

Aggregates can be decomposed into shareable components. Consider a variable order  $\Delta = A(\Delta_1, \dots, \Delta_k)$ , with root  $A$  and subtrees  $\Delta_1$  to  $\Delta_k$ . We can decompose any aggregate  $\alpha$  to be computed over  $\Delta$  into  $k + 1$  aggregates such that aggregate 0 is for  $A$  and aggregate  $j \in [k]$  is for  $root(\Delta_j)$ . Then  $\alpha$  is computed as the

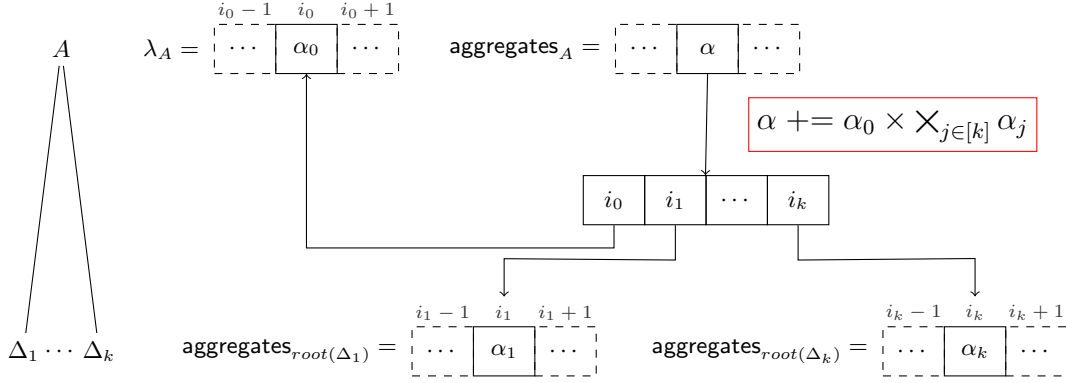


Figure 5: Index structure provided by the aggregate register for a particular aggregate  $\alpha$  that is computed over the variable order  $\Delta = A(\Delta_1, \dots, \Delta_k)$ . The computation of  $\alpha$  is expressed as the sum of the Cartesian products of its aggregate components provided by the indices  $i_0, \dots, i_k$ .

product of its  $k+1$  components. Each of these aggregates is defined by the projection of the monomial of  $\alpha$  onto  $A$  or  $\text{vars}(\Delta_j)$ . The aggregate  $j$  is then pushed down the variable order and computed over the subtree  $\Delta_j$ . If the projection of the monomial is empty, then the aggregate to be pushed down is  $\text{SUM}(1)$ , which computes the size of the join defined by  $\Delta_j$ . If several aggregates push the same aggregate to the subtree  $\Delta_j$ , this is computed only once for all of them.

The decomposed aggregates form a hierarchy whose structure is that of the underlying variable order  $\Delta$ . The aggregates at a variable  $X$  are denoted by  $\text{aggregates}_X$ . All aggregates are to be computed at the root of  $\Delta$ , then fewer are computed at each of its children and so on. This structure is the same regardless of the input data and can be constructed before data processing. We therefore construct at compile time for each variable  $X$  in  $\Delta$  an aggregate register  $\mathcal{R}_X$  that is an array of all aggregates to be computed over the subtree of  $\Delta$  rooted at  $X$ . This register is used as an index structure to facilitate the computation of the actual aggregates. More precisely, an entry for an aggregate  $\alpha$  in the register of  $X$  is labeled by the monomial of  $\alpha$  and holds an array of indices of the components of  $\alpha$  located in the registers at the children of  $X$  in  $\Delta$  and in the local register  $\Lambda_X$  of  $X$ . Figure 5 depicts this construction.

The hierarchy of registers forms an index structure that is used by AC/DC to compute the aggregates. This index structure is stored as one contiguous array in memory, where the entry for an aggregate  $\alpha$  in the register comes with an auxiliary array with the indices of  $\alpha$ 's aggregate components. The aggregates are ordered in the register so that we increase sequential access, and thus cache locality, when updating them.

**Example 19.** Let us compute a regression model of degree 1 over a dataset defined by the join of the relations  $R(A, B, C)$ ,  $S(B, D)$ , and  $T(A, E)$ . We assume that  $B$  and  $E$  are categorical features, and all other variables are continuous. The quantities  $(\Sigma, c, s_Y)$  require the computation of the following aggregates:  $\text{SUM}(1)$ ,  $\text{SUM}(X)$  for each variable  $X$ , and  $\text{SUM}(X * Y)$  for each pair of variables  $X$  and  $Y$ .

Figure 6(a) depicts a variable order  $\Delta$  for the natural join of three relations, and Figure 6(b) illustrates the aggregate register that assigns a list of aggregates to each variable in  $\Delta$ . The aggregates are identified by their respective monomials (the names in the register entries). The categorical variables are shown in bold. Since they are treated as group-by variables, we do not need aggregates whose monomials include categorical variables with exponents higher than 1. Any such aggregate is equivalent to the aggregate whose monomial includes the categorical variable with degree 1 only.

The register  $\mathcal{R}_A$  for the root  $A$  of  $\Delta$  has all aggregates needed to compute the model. The register  $\mathcal{R}_B$  has all aggregates from  $\mathcal{R}_A$  defined over the variables in the subtree of  $\Delta$  rooted at  $B$ . The variables  $C$ ,  $D$ , and  $E$  are leaf nodes in  $\Delta$ , so the monomials for the aggregates in the registers  $\mathcal{R}_C$ ,  $\mathcal{R}_D$ , and  $\mathcal{R}_E$  are the respective variables only. We use two additional registers  $\Lambda_A$  and  $\Lambda_B$ , which hold the aggregates corresponding to projections of the monomials of the aggregates in  $\mathcal{R}_A$ , and respectively  $\mathcal{R}_B$ , onto  $A$ , respectively  $B$ . For a leaf node  $X$ , the registers  $\Lambda_X$  and  $\mathcal{R}_X$  are the same.

A path between two register entries in Figure 6(b) indicates that the aggregate in the register above uses the

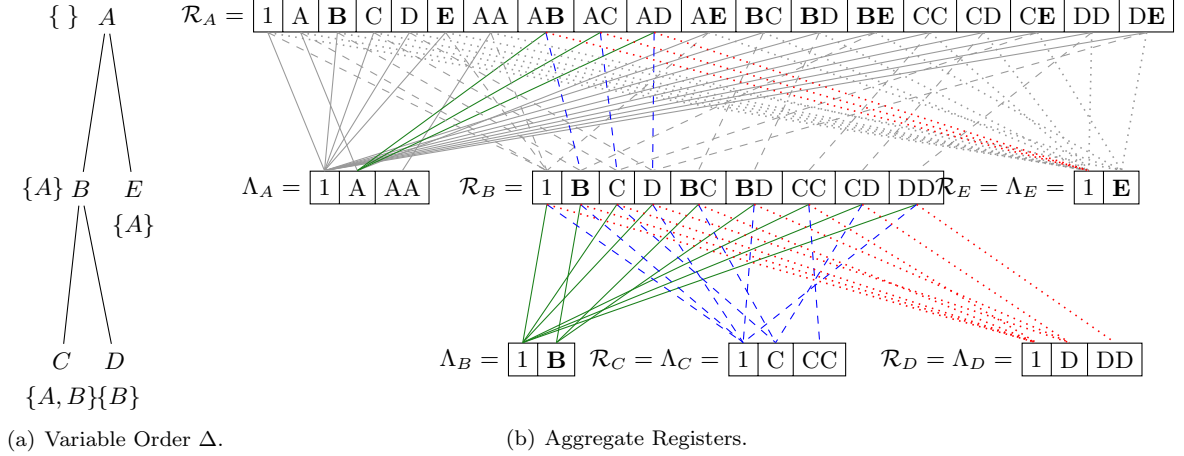


Figure 6: (a) Variable order  $\Delta$  for the natural join of the relations  $R(A,B,C)$ ,  $S(B,D)$ , and  $T(A,E)$ , each variable  $X$  is annotated by the set that  $dep(X)$  maps to; (b) Aggregate registers for the aggregates needed to compute a linear regression model with degree 1 over  $\Delta$ . Categorical variables are shown in bold.

result of the aggregate in the register below. For instance, each aggregate in  $\mathcal{R}_B$  is computed by the product of one aggregate from  $\Lambda_B$ ,  $\mathcal{R}_C$ , and  $\mathcal{R}_D$ . The fan-in of a register entry thus denotes the amount of sharing of its aggregate: All aggregates from registers above with incoming edges to this aggregate share its computation. For instance, the aggregates with monomials  $AB$ ,  $AC$ , and  $AD$  from  $\mathcal{R}_A$  share the computation of the aggregate with monomial  $A$  from  $\Lambda_A$  as well as the count aggregate from  $\mathcal{R}_E$ . Their computation uses a sequential pass over the register  $\mathcal{R}_B$ . This improves performance and access locality as  $\mathcal{R}_B$  can be stored in cache and accessed to compute all these aggregates.  $\square$

### Aggregate Computation.

Once the aggregate registers are in place, we can ingest the input database and compute the aggregates over the join of the database relations following the factorized structure given by a variable order. The algorithm in Figure 4 does precisely this. Section 6.1 explained the factorized computation of a single aggregate over the join. We explain here the case of several aggregates organized into the aggregate registers. This is stated by the pseudocode in the red boxes.

Each aggregate is uniformly stored as a map from tuples over their categorical variables to payloads that represent the sums over the projection of its monomial on all continuous variables. If the aggregate has no categorical variables, the key is the empty tuple.

For each possible  $A$ -value  $a$ , we first compute the array  $\lambda_A$  that consists of the projections of the monomials of the aggregates onto  $A$ . If  $A$  is categorical, then we only need to compute the 0 and 1 powers of  $a$ . If  $A$  is continuous, we need to compute all powers of  $A$  from 0 to  $2 \cdot \text{degree}$ . If  $A$  is not a feature used in the model, then we only compute a trivial count aggregate.

We update the value of each aggregate  $\alpha$  using the index structure depicted in Figure 5 as we traverse the variable order bottom up. Assume we are at a variable  $A$  in the variable order. In case  $A$  is a leaf, the update is only a specific value in the local register  $\lambda_A$ . In case the variable  $A$  has children in the variable order, the aggregate is updated with the Cartesian product of all its component aggregates, i.e., one value from  $\lambda_A$  and one aggregate for each child of  $A$ . The update value can be expressed in SQL as follows. Assume the aggregate  $\alpha$  has group-by variables  $C$ , which are partitioned across  $A$  and its  $k$  children. Assume also that  $\alpha$ 's components are  $\alpha_0$  and  $(\alpha_j)_{j \in [k]}$ . Recall that all aggregates are maps, which we may represent as relations with columns for keys and one column  $P$  for payload. Then, the update to  $\alpha$  is:

SELECT  $C, (\alpha_0.P * \dots * \alpha_k.P)$  AS  $P$  FROM  $\alpha_0, \dots, \alpha_k$ ;

## Further Considerations.

The auxiliary arrays that provide the precomputed indices of aggregate components within registers speed up the computation of the aggregates. Nevertheless, they still represent one extra level of indirection since each update to an aggregate would first need to fetch the indices and then use them to access the aggregate components in registers that may not be necessarily in the cache. We have been experimenting with an aggressive aggregate compilation approach that resolves all these indices at compile time and generates the specific code for each aggregate update. In experiments with linear regression, this compilation leads to a  $4\times$  performance improvements. However, the downside is that the AC/DC code gets much larger and the C++ compiler needs much more time to compile it. For higher-degree models, it can get into situations where the C++ compiler crashes. We are currently working on a hybrid approach that partially resolves the indices while maintaining a reasonable code size.

## Point Evaluation, Gradient Computation and FD Optimization

For the computation of the point evaluation and gradient computation we use the optimizations we introduced in Section 4.4. Recall that two entries in  $\Sigma$  can have the identical representation, which implies that a single aggregate can be used in distinct products over different components of  $g$ . In order to avoid keep track of which aggregates correspond to which entries in  $\Sigma$ , we construct for each aggregate a list of index pairs  $(i, j)$  for each  $\sigma_{ij} \in \Sigma$  that require this aggregate. AC/DC then uses the index list and the aggregate computed at the root of the variable order to compute the queries for point evaluation and gradient computation that were presented in Section 4.4. Consider the matrix vector product  $\mathbf{p} = \Sigma g(\theta)$ , which is needed for gradient computation. Let  $A$  be the root of the variable order  $\Delta$ . We compute  $\mathbf{p}$  by iterating over all aggregate maps  $\alpha \in \text{aggregates}_A$ , and for each index pair  $(i, j)$  that is assigned to  $\alpha$ , we add to the  $i$ 'th component of  $\mathbf{p}$  the product of  $\alpha$  and  $g_j(\theta)$ . If  $i \neq j$ , we also add to  $j$ 's component of  $\mathbf{p}$  with the product of  $\alpha$  and  $g_i(\theta)$ .

For the FD optimization, it is required to construct the  $\mathbf{R}_c$  matrices that were introduced in Section 5.2. In AC/DC, we represent these matrices as maps that group the values for functionally determining variables by the values that they determine. The maps are sparse representations of  $\mathbf{R}_c$  matrices, and they are populated during the computation of the factorized aggregates over the variable order. We choose this representation because it allows for the efficient computation of the matrix  $I + \mathbf{R}_c^\top \mathbf{R}_c$ , which is the basic building block the matrix  $\mathbf{B}_{T,q}$  from (47). The reparameterization of the regularizer requires the computation of the inverse of  $\mathbf{B}_{T,q}$ . Therefore, we store the matrix  $\mathbf{B}_{T,q}$  as a sparse matrix in the format used by the Eigen linear algebra library [34], and then use Eigen's Sparse Cholesky Decomposition to compute the inverse of  $\mathbf{B}_{T,q}$ .

## 7 Experiments

We report on the performance of learning regression and factorization machine models over three real datasets used in retail and advertisement applications. We benchmark AC/DC against state-of-the-art competitors. AC/DC can compute the models up to  $1,031\times$  faster, while the accuracy of AC/DC's models is always at least as good as the competitor's models. For all experiments, we assume that the model specification is given as input, and all systems compute the same model. We do not consider the orthogonal problem of finding the best model for a given analytics task.

### 7.1 Experimental setup

All experiments were performed on an Intel(R) Core(TM) i7-4770 3.40GHz/64bit/32GB with Ubuntu 18.04, g++7.4, and eight cores. We report wall-clock times by running each system once and then reporting the average of four subsequent runs with warm cache. We do not report the times to load the database into memory for the join. All relations are sorted by their join variables.

#### Competitors

We benchmark AC/DC against six competitors: MADlib [38] 1.16, libFM [64] 1.4.2, TensorFlow [1] 1.13.1, R [63] 3.4.4, scikit-learn [58] 0.20, and Python Statsmodels [68]. We evaluate the performance of learning the models

	Retailer	Favorita	Yelp
Relations	4	5	5
Variables	21	14	26
Categorical Variables	4	10	6
Tuples in Database	87M	125M	8.7M
Size of Database	1.5GB	2.5GB	0.2GB
Tuples in Join Result	86M	125M	360M
Size of Join Result	18GB	7GB	40GB
#values in Listing Representation	2.302G	1.735G	2.835G
#values in Factorized Representation	166M	372M	71.9M
Compression ( Factorized/Listing )	13.91×	4.66×	39.43×

Table 1: Key Characteristics of the three used datasets.

over a training dataset, and then compute the root-mean-squared-error (RMSE) over a separate test dataset to compare the accuracy.

MADlib, R, scikit-learn, and Python StatsModels use *ols* (ordinary least squares) to compute the closed-form solution of regression models, and TensorFlow uses the LinearRegressor estimator with *ftrl* optimization [48], which is based on the conventional SGD optimization algorithm. TensorFlow was compiled from source to enable specialized optimizations that are native to our machine, including AVX optimizations. We use PostgreSQL 10.9 to compute the feature extraction query for libFM and TensorFlow. LibFM supports factorization machines.

AC/DC uses the gradient descent algorithm with the adaptive learning rate from Algorithm 1. For LR and PR models, we run the optimization algorithm until the RMSE over the training dataset changes by less than  $10^{-15}$  in three consecutive iterations. For FaMa models, we use 300 iterations.

The aggregate computation in AC/DC is parallelized on eight cores. Tensorflow also uses multiple threads by default. MADlib and libFM only use a single thread.

The competitors come with various limitations that affect their scalability. MADlib requires the explicit one-hot encoding of the input relations, for which we use its predefined functions.

LibFM requires as input a zero-suppressed encoding of the one-hot encoded join result. Computing this representation is an expensive intermediate step between exporting the query result from the database system and importing the data into libFM. We learn the FaMa models using the MCMC optimization algorithm with a fixed number of runs (300); its SGD implementation requires a fixed learning rate and does not converge.

TensorFlow uses a user-defined iterator interface to load a batch of tuples from the training dataset at a time. This iterator defines a mapping from input tuples to features and is called directly by the learning algorithm. To avoid the explicit one-hot encoding of the features, TensorFlow encodes categorical features using a hash function. Learning over batches requires a random shuffling of the input data, which in TensorFlow amounts to loading the entire dataset into memory. This failed for our experiments due to the large sizes of the datasets. We therefore shuffle the data in PostgreSQL instead and provide the shuffled input to TensorFlow. We benchmark TensorFlow for LR only as it does not provide functionality to create all pairwise interaction terms for PR and FaMa. The optimal batch size for our experiments is 100,000 tuples. Smaller batch sizes require loading too many batches, very large batches cannot fit into memory. Since TensorFlow uses a fixed number of iterations, we report the times to optimize with one epoch over the training dataset. This means that the algorithm learns over each data point in the training dataset once. Our experiments show that it is often necessary to optimize with several epochs to learn a good model.

R, scikit-learn, and Python Statsmodels fail to compute our models due to design limitations. R limits the number of values in their data frames to  $2^{31} - 1$ , which is insufficient to represent out datasets. Scikit-learn and Python Statsmodels both run out of memory for all considered models.

## Datasets

We experimented with three real-world datasets: (1) Retailer [66] is used by a large retailer for forecasting user demand and sales; (2) Favorita [26] is a public dataset used for retail forecasting; and (3) Yelp is based on the Yelp Dataset Challenge [70] and used to predict ratings by a user for a business. The structure and size of these



datasets is common in retail and advertising, where data is easily generated by sales transactions or click streams. The feature extraction query for each dataset is the natural join of the input relations. For each dataset, we consider a subset of the variables. Table 1 presents key characteristics for each dataset. It shows that the join result can be orders of magnitude larger than the input database.

Retailer has a star schema with one fact table **Inventory**, which keeps track of the number of inventory units for products (sku) in a store (locn) at a given date, and three dimension tables: (1) **Location** keeps additional information for each store, including its size in sqft and distances to three competitors; (2) **Items** provides identifiers for the item category, subcategory, category cluster, as well as the price for each sku; and (3) **Weather** keeps information about the weather conditions for each store at a given date (including maximum temperature, and whether it rained). We design two versions of our dataset. The version  $v_1$  includes all variables but sku, date, and locn as features, and has no functional dependencies. Version  $v_2$  extends  $v_1$  with the categorical variable sku, and exploits the functional dependency  $\text{sku} \rightarrow \{\text{category, subcategory, categoryCluster}\}$ . We learned LR,  $\text{PR}_2$ , and  $\text{FaMa}_2^8$  models that predict the amount of inventory units based on all other features. The test data constitutes the inventory in the last month in the dataset (approx. 2.2% of the data). This simulates the realistic usecase where the ML model predicts future inventory.

Favorita has a star schema with one fact table and 4 dimension tables. The fact table **Sales** stores the number of units sold for items for a given date and store, and an indicator whether or not the unit was on promotion at this time. **Items** keeps additional information about the skus, including the item class and whether it is perishable. **Stores** provides additional information about the stores, such as the city and state they are located in, and the type of store. **Transactions** gives the number of transactions at each store on a given date. **Oil** keeps the oil price for each date. Our models predict the number of units sold. We designed two variants for this dataset. Version  $v_1$  learns the model over all variables except sku and date, and version  $v_2$  extends  $v_1$  with sku. We exploit the functional dependency  $\text{store} \rightarrow \{\text{city, state, storetype}\}$  in both variants. The test data constitutes the sales for the last month in the dataset (approx. 1.2% of the data).

Yelp has a star schema with four relations: **Review** stores, for each review, the rating given by user, the review date, and the number of compliments it received (e.g. useful, funny, cool); **Business** keeps the location (city, state, coordinates), the average rating, and the total count of reviews for each business; **User** keeps aggregated statistics for each user, including the number of reviews they wrote, the number of compliments they received, and the average rating they gave to businesses; **Attribute** keeps attributes (e.g. as “open late”) for each business. One user can review many businesses and a business can have many attributes. The result of the feature extraction query is much larger than the input relations. Our models predict the rating that users give to businesses. The models are learned over all variables except the join keys, and exploit the functional dependency  $\text{city} \rightarrow \text{state}$ . The test dataset is a random selection of approximately 2% of the reviews.

## 7.2 Summary of findings

Our findings on the performance comparison between AC/DC and the three competitors are given in Table 2. AC/DC is the fastest system in our experiments. It can compute the models over the input database orders of magnitude faster than its competitors whenever they do not exceed memory limitation, 24-hour timeout, or internal design limitations.

AC/DC learns models with at least as good accuracy as the competitors. In particular, AC/DC’s LR models have comparable accuracy to MADlib’s closed form solution (whenever MADlib does not time out), and consistently better accuracy than the models learned by TensorFlow (trained for one epoch). With additional features, AC/DC can learn more accurate models while the competitors either timeout or fail due to internal design limitations.

The performance gap between competitors and AC/DC is primarily due to the following optimizations supported by AC/DC:

1. It avoids the materialization of the join and the export-import step between database systems and statistical packages, which may take longer than the end-to-end learning of LR models in AC/DC. AC/DC performs the join together with the aggregates using one execution plan;
2. It factorizes the computation of the sparse tensor and the underlying join. The compression factor brought by join factorization is  $13.9\times$  for Retailer,  $4.7\times$  for Favorita, and  $21\times$  for Yelp;

		Retailer $v_1$	Retailer $v_2$	Favorita $v_1$	Favorita $v_2$	Yelp
Join Computation (PSQL)		447.76	447.76	255.16	255.16	195.26
Factorized Computation of Count over Join		19.90	19.90	36.45	36.45	21.07
<b>Linear Regression</b>						
Features (continuous+categorical)		16 + 49	16+3,661	4 + 482	4 + 4,482	21 + 1,068
Number of Entries in Sparse Tensor		1,149	69,777	42,504	455,889	46,401
MADLib (ols)	Encode	0.19	8.46	532.52	545.42	61.79
	Learn	1,124.84	>86,400.00	13,951.44	>86,400.00	44,307.88
TensorFlow (ftrl)	Join+Shuffle+Export	2,266.76	2,266.76	1,417.87	1,417.87	1,110.41
(1 epoch, batch size 100K)	Learn	3,420.91	3,408.34	3,649.73	3,648.52	5,763.88
<b>AC/DC</b>	Aggregate	40.09	118.47	115.02	1,022.44	42.89
	Converge	0.11	285.92	0.94	22.20	0.14
Speedup of <b>AC/DC</b> over	MADlib	27.99 $\times$	> 213.67 $\times$	124.90 $\times$	> 83.23 $\times$	1,031.13 $\times$
	TensorFlow	141.48 $\times$	14.03 $\times$	43.70 $\times$	4.85 $\times$	159.76 $\times$
<b>Polynomial Regression degree 2</b>						
Features (continuous+categorical)		121+980	121+65,996	7+42,016	7+451,403	211+41,559
Number of Entries in Sparse Tensor		72,790	2,517,600	498,641	6,616,551	6,478,164
MADlib (ols)	Encode	0.19	8.46	532.52	545.42	61.79
	Learn	>86,400.00	>86,400.00	>86,400.00	>86,400.00	>86,400.00
<b>AC/DC</b>	Aggregate	122.88	334.35	324.99	7,549.99	3,650.47
	Converge	21.92	621.69	45.34	1,063.88	203.64
Speedup of <b>AC/DC</b> over	MADlib	> 596.69 $\times$	> 90.38 $\times$	> 234.74 $\times$	> 10.09 $\times$	> 22.43 $\times$
<b>Factorization Machine degree 2 rank 8</b>						
Features (continuous+categorical)		107+980	107+65,996	5+42,016	5+451,403	192+41,559
Number of Entries in Sparse Tensor		70,515	2,465,443	497,786	6,607,696	6,454,053
libFM (MCMC)	Join+Ex/Import+Encode	3,368.06	3,368.06	3,214.79	3,214.79	2,719.48
(300 iterations)	Learn	>86,400.00	>86,400.00	>86,400.00	>86,400.00	67,829.59
<b>AC/DC</b>	Aggregate	124.17	324.492	351.68	7,856.88	3,633.93
(300 iterations)	Converge	0.67	42.74	2.60	25.38	265.94
Speedup of <b>AC/DC</b> over	libFM	> 719.06 $\times$	> 244.45 $\times$	> 252.95 $\times$	11.37 $\times$	18.09 $\times$

Table 2: Time performance (seconds) for learning LR, PR, and FaMa models over Retailer, Favorita, and Yelp. The timeout is set to 24 hours (86,400 seconds). R and MADlib do not support FaMa models. TensorFlow does not support PR and FaMa models.

3. It massively shares the computation of many aggregates representing entries in the sparse tensor. For instance, there are up to 2.5M such sum aggregates for PR<sub>2</sub> on Retailer  $v_2$  and they take 150K $\times$  less time than computing the count aggregate 2.5M times, where the count takes 19.9 seconds as reported in Table 2;
4. It decouples the computation of the aggregates on the input data from the parameter convergence step and thus avoids scanning the join result for each iteration;
5. It avoids the upfront one-hot encoding that comes with higher asymptotic complexity and prohibitively large covariance matrices by only computing distinct, non-zero entries in the sparse tensor. For PR<sub>2</sub> on Retailer  $v_2$ , this leads to 868 $\times$  less aggregates to compute;
6. It exploits the functional dependencies in the input data to reduce the number of features of the model, which leads to an improvement factor of up to 2.3 $\times$ .

### 7.3 Further details

#### Categorical features

As we move from Retailer  $v_1$  to  $v_2$ , we increase the number of categorical features by approx. 75 $\times$  for LR (from 49 to 3.7K) and 67 $\times$  for PR<sub>2</sub> and FaMa<sub>2</sub><sup>8</sup> (from 980 to 66K). This translates to a same-order increase in the number of aggregates: 62 $\times$  (56 $\times$ ) more distinct non-zero aggregates in  $v_2$  vs  $v_1$  for LR (resp. PR<sub>2</sub> and FaMa<sub>2</sub><sup>8</sup>). This increase only led to a decrease in performance of AC/DC of 10 $\times$  for LR and 6.6 $\times$  for PR<sub>2</sub>. This sub-linear behavior is partly explained by the ability of AC/DC to process many aggregates much faster in bulk than individually: it takes 19.9 seconds for one count aggregate, but only 334 seconds to compute all 2.5M entries in the sparse tensors for PR<sub>2</sub> on  $v_2$ !

For MADlib, the performance decrease is at least 78 $\times$  for LR when moving from  $v_1$  to  $v_2$  and it times out after 24 hours for  $v_2$ . MADlib also times out for all PR<sub>2</sub> experiments. The performance of TensorFlow is largely invariant to the increase in the number of categorical features, since its internal mapping from tuples in the

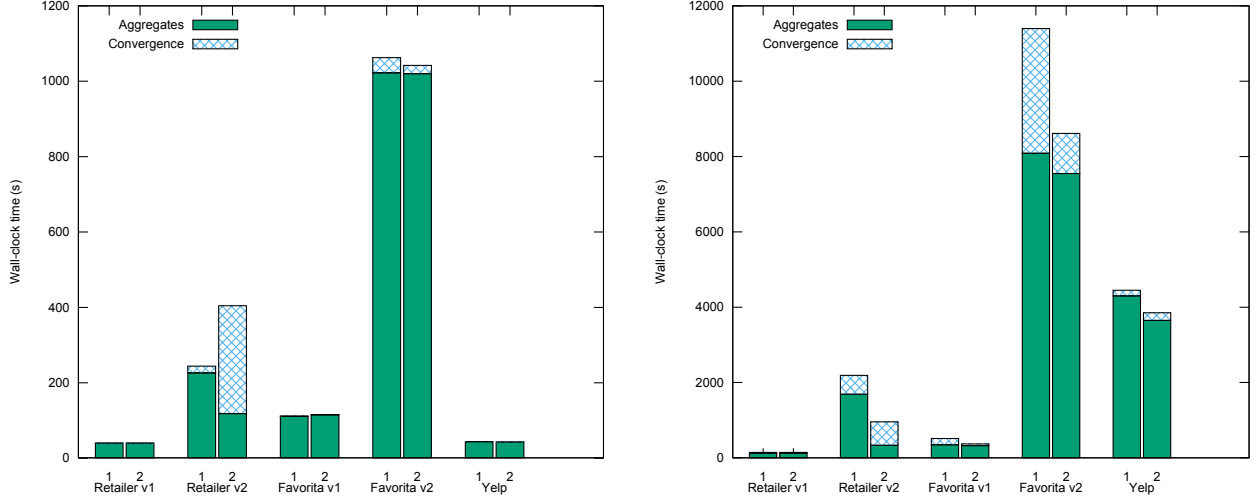


Figure 7: Breakdown of AC/DC performance for LR (left) and PR<sub>2</sub> (right): (1) not using FDs; (2) using FDs.

training dataset to the sparse representation of the features vector remains of similar size. Nevertheless, our system is consistently faster (often by orders of magnitude) than computing only *a single epoch* in TensorFlow. In addition, AC/DC computes the end-to-end models faster than PSQL takes to materialize and shuffle the design matrix that is the input to TensorFlow.

#### One-hot encoding vs. sparse encoding with group-by aggregates.

One-hot encoding categorical features leads to a large number of zero and/or redundant entries in the  $\Sigma$  matrix. For instance, for PR<sub>2</sub> on Retailer v<sub>2</sub>, the number of features is  $m=66,117$ , and then the upper half of  $\Sigma$  would have  $m(m+1)/2 \approx 2.1 \times 10^9$  entries. Most of these are either zero or repeating, as exemplified in Section 4.3. In contrast, AC/DC’s sparse representation only considers 2.5M non-zero and distinct aggregates. The number of aggregates is thus reduced by a factor of 868!

MADlib and libFM require the data be one-hot encoded *before* learning. For MADlib, the static one-hot encoding took up to 545 seconds on Favorita. In addition to the one-hot encoding, libFM requires the input data to be represented in a zero-suppressed sparse format. This data transformation took up to one hour in our experiments. TensorFlow one-hot encodes on the fly using hash functions during the learning phase, which cannot be reported separately.

When running over one-hot encoded input data, AC/DC exceeds the available memory for all models but the linear regression model for Retailer v<sub>1</sub>.

#### Effect of functional dependencies.

Figure 7 shows the performance breakdown for AC/DC with and without exploiting FDs. All other systems do not exploit FDs.

The FDs have a twofold effect on AC/DC: they can reduce the number of features and aggregates, which leads to better performance of the aggregation step; yet it requires a more expensive convergence step due to the more complex regularizer. For LR over Retailer v<sub>2</sub>, the aggregate step becomes 2× faster, while the convergence step increases 15×, offsetting the effect of the faster aggregate computation. The FDs for Favorita and Yelp have a relatively smaller effect on performance for LR (there are only 54 stores in Favorita, so the reduction in the number of aggregates is small).

For PR<sub>2</sub> models, the FD brings an improvement by a factor of 2.3× for Retailer. This is due to a 18% decrease in the number of categorical features, which leads to a 38% decrease in the number of group-by aggregates. For Favorita and Yelp, the performance improvement is 1.3× and respectively 1.2×. For FaMa<sub>2</sub><sup>8</sup> models, the effect of FDs is comparable to that of PR<sub>2</sub> models.

### Accuracy.

The RMSE of the LR models for Retailer  $v_1$  and Favorita  $v_1$  in AC/DC is within 1% of that for the closed form solutions computed in MADlib. By extending the models with the categorical variable `sku` (version  $v_2$  for both datasets), the RMSE decreases by 21% for Retailer and by 6% for Favorita. MADlib fails to learn these more accurate models, because it times out after 24 hours.

For TensorFlow, the models are trained with a single epoch. The resulting models have a consistently higher RMSE than the corresponding model computed in AC/DC. In particular, for Retailer  $v_2$ , the RMSE of the TensorFlow model is 31% higher. TensorFlow requires more epochs to achieve the same accuracy as AC/DC.

Extending the model with pairwise interactions can also improve the model accuracy. For Favorita  $v_2$ , for instance, the RMSE of the  $PR_2$  model is 3% lower than the RMSE of the LR model.

The FaMa models learned by AC/DC do not reach the same accuracy as the  $PR_2$  models, and would therefore require more than 300 iterations to converge to an accurate model. This is due to the non-linear structure of factorization machines.

## 8 Related work

Our work follows closely Chaudhuri’s manifesto on SQL-aware data mining systems from two decades ago [19] in two key aspects. First, the goal of our work is not to invent new machine learning models or data analysis techniques, but identify common data-centric steps across a broad class of learning algorithms and investigate their theoretical and systems challenges. We show that such steps can be encoded as SQL group-by aggregate queries, which are amenable to shared batch computation. Second, our approach performs data analysis not only over materialized relations but more importantly over feature extraction queries, whose results need not be materialized. This enables the interaction between the aggregates encoding the data-centric steps and the underlying queries (this is called ad-hoc mining in Chaudhuri’s terminology).

A reevaluation of Chaudhuri’s manifesto in today’s context brings forth two important technical changes. The first game-changer is represented by the recent development on query processing. This includes a new breed of worst-case optimal join algorithms, which support listing representation [52, 69] and factorized representation [56] of query results, and extensions to aggregate computation [13, 8, 55]. These algorithms exploit developments on (fractional) hypertree decompositions of relational queries [32, 33, 47]. These algorithms overshadow the traditional query plans in both asymptotic complexity [53] and practical performance [13, 54]. The second change is in the workload. Whereas SQL-aware data mining systems were mostly concerned with association rules, decision trees, and clustering, current workloads feature a broader spectrum of increasingly more sophisticated machine learning (ML) models, including polynomial regression models, factorization machines, principal component analysis, generalized linear models, generalized low-rank models, sum-product networks, and convolutional networks. In this article, we introduce a unified approach to learning polynomial regression models, factorization machines, and principal component analysis over non-materialized feature extraction queries with the lowest known complexity and best performance to date. There is also a more profound orthogonal change: There is more data readily available in all aspects of our society and there is more appetite in industry to monetize it by turning it into knowledge.

The current landscape for analytics solutions over multi-relational data can be categorized depending on the degree of integration of the data system, which hosts the data and supports data access via query primitives, with the ML library of models and learning algorithms.

By far the most common solutions provide no integration of the two systems, which are distinct tools on the technology stack: The data system exports the training dataset as one relation, commonly presented as a CSV file, and then the ML system imports it into its own format and learns the desired model. Such solutions are structure-agnostic as they disregard the rich structural information of the materialized training dataset, including the input database schema, database dependencies, and the structure of the feature extraction queries. Prime examples are the pairing of open-source data systems such as Spark [71] or MySQL/PostgreSQL with ML systems such as R [63], Python StatsModels [68], Python Scikit [58], MLpack [22], TensorFlow [1], SystemML [39, 15], MLLib [49], and DeepDist [51]. The advantage of this approach is that the two systems can be developed independently, with virtually any ML model readily available for use.

Two disadvantages of such common solutions are the expensive data export/import at the interface between the two systems and the materialization of the training dataset as a result of a feature extraction query over multi-relational data. The feature extraction query is computed inside the data system, its result exported and imported into the data format of the ML system, where the model is learned. Furthermore, the materialized training dataset may be much larger than the input data (cf. Table 1). This is exacerbated by the stark asymmetry between the two systems: Whereas data systems tend to scale to large datasets, this is not the case for ML libraries. Yet, such solutions expect by design that the ML libraries work on even larger inputs than the data systems! A further disadvantage is that these solutions inherit the limitations of both underlying systems. For instance, the R data frame can host at most  $2^{31}$  values, which makes it impossible to learn models over large datasets, even if data systems can process them. Database systems can only handle up to a few thousand columns per relation, which is usually smaller than the number of features of the model.

The second class of systems features a loose integration, even though they remain structure-agnostic: The ML code migrates inside the space of the data system process, with each ML task being implemented by a distinct user-defined aggregate function (UDAF). Prime examples of this class are MADlib [38] and GLADE PF-OLA [62]. MADlib casts analytics as UDAFs that can be used in SQL queries and executed inside PostgreSQL. GLADE PF-OLA casts analytics as a special form of UDAFs called Generalized Linear Aggregates that can be executed using the GLADE distributed engine [21]. These UDAFs remain black boxes for the underlying query engine, which has to compute the feature extraction query and delegate the UDAF computation on top of the query result to the MADlib’s and GLADE PF-OLA’s specialized code. The advantage of this approach is that the expensive export/import step is avoided. The disadvantage is that each ML task has to be migrated inside the data system space, which comes with design and implementation overhead. A further step towards integration is exemplified by Bismarck [27], which provides a unified programming architecture for many ML tasks instead of one UDAF per task, with possible code reuse across UDAFs.

The third class of systems features a tight integration and are structure-aware: There is one execution strategy for both the feature extraction query and the subsequent learning task, with components of the latter possibly pushed past the joins in the former. Prime examples are Morpheus [43], Hamlet [44], and our prior system F [66] that support generalized linear models, Naïve Bayes classification, and respectively linear regression models with continuous features. This class also contains the recent efforts on in-database linear algebra [20] and on scaling linear algebra using existing distributed database systems [46] and the declarative language BUDS [29], whose compiler can perform deep optimizations of the user’s program. Our approach AC/DC generalizes F to non-linear models, categorical features, and model reparameterization under functional dependencies. A key aspect that sets apart AC/DC and its predecessor F from prior work is that they employ execution strategies for the mixed workload of queries and learning with complexity that may be asymptotically lower than that of query materialization alone. In particular, all machine learning approaches that require as input the materialization of the result of the feature extraction query are asymptotically suboptimal. This complexity gap translates into a performance gap, cf. Section 7.

Figure 1 sums up the difference between the first two classes that fall under structure-agnostic learning and the third class that broadly represents structure-aware learning. The inspiration for our work lies with factorized computation of aggregates over joins [13, 8], which avoids the materialization of joins, and with the LogicBlox system [50, 11], which has a unified system architecture and declarative programming language for hybrid database and optimization workloads.

Beyond the above classification, there are further directions of research looking at ML through database glasses: ML-aware query languages, the effect of dependencies on model training, sparse data representations, and implementations of gradient descent solvers.

Analytical tasks can be expressed to a varying degree within query languages possibly extended with new constructs. Very recent works investigate query languages for matrices [18] and a relational framework for classifier engineering [41]. They follow works on query languages with data mining capabilities [17, 57], also called descriptive or backward-looking analytics, and on in-database data mining solutions, such as frequent itemsets [59] and association rule mining [10]. Our rewriting of ML code into aggregates falls into this line of work as well. The additional fixpoint computation needed on top of the aggregate computation for convergence of the model parameters, which is intrinsic to gradient descent approaches, can be expressed as recursive queries [3].

Functional dependencies (FDs) can be used to avoid key-foreign key joins and reduce the number of features in Naïve Bayes classification and feature selection [44]. In this article we consider the effect of FDs on the

reparameterization of regression models, where a non-trivial development is on the effect of FDs on the model’s non-linear regularization function, cf. Section 5. Our factorized learning approach exploits the join dependencies present in the training dataset, as defined by the feature extraction query. This follows prior work on factorized databases [13, 56].

State-of-the-art machine learning systems use a sparse representation of the input data to avoid redundancy introduced by one-hot encoding [64, 25]. In our setting, however, such systems require an additional data transformation step after the result of the feature extraction query is exported. This additional step is time consuming and makes the use of such systems inefficient in many practical applications. In statistics and machine learning, there is a rich literature on learning with sparse and/or multilinear structures [37]. Such methods complement our framework and it would be of interest to leverage and adapt them to our setting.

Finally, there is a large collection of gradient-based methods proposed in the optimization literature. The description of our approach assumes batch gradient descent (BGD), though our insights are applicable to other methods including Quasi-Newton algorithms. The main rationale for our choice is simplicity and good statistical properties. When combined with backtracking line search (as we do in this article) or second-order gradient estimation (as in Quasi-Newton methods), BGD is guaranteed to converge to a minimum with linear asymptotic convergence rate. A naïve computation of the gradient requires a full pass over the data, which can be inefficient in large-scale analytics. A popular alternative is stochastic gradient descent (SGD), which estimates the gradient with a randomly selected mini-batch of training samples. The convergence of SGD, however, is noisy, requires careful setting of hyperparameters, and does not achieve the linear asymptotic convergence rate of BGD [16]. In our setting, the entire BGD execution can be arbitrarily faster than one SGD iteration over the result of the feature extraction query. The reason is orthogonal to properties of the two gradient descent methods: The complexity of computing the sufficient statistics needed for convergence of model parameters can be asymptotically lower than the complexity of computing the training dataset.

## 9 Open Problems

Our in-database learning framework raises open questions on statistics, algorithm design, and optimization. We next sketch a few representative questions.

One research direction is to further extend the class of statistical models that can be trained efficiently by exploiting the structure of the underlying relational database. Our formulation (6) captures a common class of regression models (such as PR and FaMa), classification models (such as logistic and SVM), and unsupervised learning techniques (such as principal component analysis) which is done by changing the loss function  $\mathcal{L}$ . It remains open how to extend our formulation to capture latent variable models.

The aggregates defining  $\Sigma$ ,  $\mathbf{c}$ , point evaluation, and gradient computation are “multi-output” queries. They deserve a systematic investigation, from formulation to evaluation and complexity analysis. In practice, one often reserves a fragment of the training data for model validation. It is an interesting question to incorporate this data partitioning requirement into our framework.

Understanding how to adapt further optimization algorithms, such as coordinate descent or stochastic gradient, to our structure-aware framework is an important research direction. Furthermore, our FD-aware optimization is specific to the  $\ell_2$ -norm in the penalty term. We would also like to understand the effect of other norms, e.g.,  $\ell_1$ , on model reparameterization under FDs.

Finally, we conjecture that the cost function may be easier to optimize with respect to the reduced set of parameters that are not functionally determined: As redundant variables are eliminated or optimized out, the cost function’s Hessian with respect to reduced parameters becomes less ill-conditioned, resulting in faster convergence behavior for gradient-based optimization techniques. The impact of FD-based dimensionality reduction, from both computational and statistical standpoints, have not been extensively studied for learning (nonlinear) models with categorical variables, which are precisely the kind discussed in our framework.

## Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 682588. XN is supported in part by grants NSF CAREER DMS-1351362, NSF CNS-1409303 and the Margaret and Herman Sokol Faculty Award.

## References

- [1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. A. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A system for large-scale machine learning. In *OSDI*, pages 265–283, 2016.
- [2] S. Abiteboul, M. Arenas, P. Barceló, M. Bienvenu, D. Calvanese, C. David, R. Hull, E. Hüllermeier, B. Kimelfeld, L. Libkin, W. Martens, T. Milo, F. Murlak, F. Neven, M. Ortiz, T. Schwentick, J. Stoyanovich, J. Su, D. Suciu, V. Vianu, and K. Yi. Research directions for principles of data management (dagstuhl perspectives workshop 16151). *Dagstuhl Manifestos*, 7(1):1–29, 2018.
- [3] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.
- [4] M. Abo Khamis, H. Q. Ngo, X. Nguyen, D. Olteanu, and M. Schleich. AC/DC: In-database learning thunderstruck. In *Data Management for End-To-End Machine Learning*, DEEM’18, pages 8:1–8:10, 2018.
- [5] M. Abo Khamis, H. Q. Ngo, X. Nguyen, D. Olteanu, and M. Schleich. In-database learning with sparse tensors. In *PODS*, pages 325–340, 2018.
- [6] M. Abo Khamis, H. Q. Ngo, C. Ré, and A. Rudra. Joins via geometric resolutions: Worst-case and beyond. In *PODS*, pages 213–228, 2015.
- [7] M. Abo Khamis, H. Q. Ngo, and A. Rudra. FAQ: questions asked frequently. *CoRR*, abs/1504.04044, 2015.
- [8] M. Abo Khamis, H. Q. Ngo, and A. Rudra. FAQ: Questions asked frequently. In *PODS*, pages 13–28, 2016.
- [9] I. Adler. *Width functions for hypertree decompositions*. 2006. Ph.D. Dissertation, Albert-Ludwigs-Universität Freiburg. 2006.
- [10] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Advances in knowledge discovery and data mining. chapter Fast Discovery of Association Rules, pages 307–328. 1996.
- [11] M. Aref, B. ten Cate, T. J. Green, B. Kimelfeld, D. Olteanu, E. Pasalic, T. L. Veldhuizen, and G. Washburn. Design and implementation of the LogicBlox system. In *SIGMOD*, pages 1371–1382, 2015.
- [12] A. Atserias, M. Grohe, and D. Marx. Size bounds and query plans for relational joins. In *FOCS*, pages 739–748, 2008.
- [13] N. Bakibayev, T. Kociský, D. Olteanu, and J. Závodný. Aggregation and ordering in factorised databases. *PVLDB*, 6(14):1990–2001, 2013.
- [14] J. Barzilai and J. M. Borwein. Two-point step size gradient methods. *IMA J. Numer. Anal.*, 8(1):141–148, 1988.
- [15] M. Boehm, S. Tatikonda, B. Reinwald, P. Sen, Y. Tian, D. Burdick, and S. Vaithyanathan. Hybrid parallelization strategies for large-scale machine learning in SystemML. *PVLDB*, 7(7):553–564, 2014.
- [16] L. Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade (2nd ed)*, pages 421–436, 2012.
- [17] J.-F. Boulicaut and C. Masson. *Data Mining Query Languages*, pages 715–726. 2005.

- [18] R. Brijder, F. Geerts, J. V. den Bussche, and T. Weerwag. On the expressive power of query languages for matrices. In *ICDT*, pages 10:1–10:17, 2018.
- [19] S. Chaudhuri. Data mining and database systems: Where is the intersection? *IEEE Data Eng. Bull.*, 21(1):4–8, 1998.
- [20] L. Chen, A. Kumar, J. F. Naughton, and J. M. Patel. Towards linear algebra over normalized data. *PVLDB*, 10(11):1214–1225, 2017.
- [21] Y. Cheng, C. Qin, and F. Rusu. Glade: Big data analytics made easy. In *SIGMOD*, pages 697–700, 2012.
- [22] R. R. Curtin, M. Edel, M. Lozhnikov, Y. Mentekidis, S. Ghaisas, and S. Zhang. mlpack 3: a fast, flexible machine learning library. *J. Open Source Software*, 3:726, 2018.
- [23] T. A. Davis and W. W. Hager. Multiple-rank modifications of a sparse Cholesky factorization. *SIAM J. Matrix Anal. Appl.*, 22(4):997–1013, 2001.
- [24] T. Elgamal, S. Luo, M. Boehm, A. V. Evfimievski, S. Tatikonda, B. Reinwald, and P. Sen. SPOOF: sum-product optimization and operator fusion for large-scale machine learning. In *CIDR*, 2017.
- [25] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, 2008.
- [26] C. Favorita. Corporacion Favorita Grocery Sales Forecasting: Can you accurately predict sales for a large grocery chain? <https://www.kaggle.com/c/favorita-grocery-sales-forecasting/>, 2017.
- [27] X. Feng, A. Kumar, B. Recht, and C. Ré. Towards a unified architecture for in-rdbms analytics. In *SIGMOD*, pages 325–336, 2012.
- [28] R. Fletcher. On the Barzilai-Borwein method. In *Optimization and control with applications*, volume 96 of *Appl. Optim.*, pages 235–256. 2005.
- [29] Z. J. Gao, S. Luo, L. L. Perez, and C. Jermaine. The BUDS language for distributed bayesian machine learning. In *SIGMOD*, pages 961–976, 2017.
- [30] P. E. Gill, G. H. Golub, W. Murray, and M. A. Saunders. Methods for modifying matrix factorizations. *Math. Comp.*, 28:505–535, 1974.
- [31] T. Goldstein, C. Studer, and R. G. Baraniuk. A field guide to forward-backward splitting with a FASTA implementation. *CoRR*, abs/1411.3406, 2014.
- [32] G. Gottlob, N. Leone, and F. Scarcello. Hypertree decompositions and tractable queries. In *PODS*, pages 21–32, 1999.
- [33] M. Grohe and D. Marx. Constraint solving via fractional edge covers. In *SODA*, pages 289–298, 2006.
- [34] G. Guennebaud, B. Jacob, et al. Eigen v3. <http://eigen.tuxfamily.org>, 2010.
- [35] W. W. Hager. Updating the inverse of a matrix. *SIAM Rev.*, 31(2):221–239, 1989.
- [36] D. Harris and S. Harris. *Digital Design and Computer Architecture*. 2nd edition, 2012.
- [37] T. Hastie, R. Tibshirani, and M. J. Wainwright. *Statistical Learning with Sparsity: The Lasso and generalizations*. CRC Press, 2015.
- [38] J. M. Hellerstein, C. Ré, F. Schoppmann, D. Z. Wang, E. Fratkin, A. Gorajek, K. S. Ng, C. Welton, X. Feng, K. Li, and A. Kumar. The madlib analytics library or MAD skills, the SQL. *PVLDB*, 5(12):1700–1711, 2012.
- [39] B. Huang, M. Boehm, Y. Tian, B. Reinwald, S. Tatikonda, and F. R. Reiss. Resource elasticity for large-scale machine learning. In *SIGMOD*, pages 137–152, 2015.



- [40] C. G. Khatri and C. R. Rao. Solutions to some functional equations and their applications to characterization of probability distributions. *Sankhyā Ser. A*, 30:167–180, 1968.
- [41] B. Kimelfeld and C. Ré. A relational framework for classifier engineering. In *PODS*, pages 5–20, 2017.
- [42] A. Kumar, M. Boehm, and J. Yang. Data management in machine learning: Challenges, techniques, and systems. In *SIGMOD*, pages 1717–1722, 2017.
- [43] A. Kumar, J. F. Naughton, and J. M. Patel. Learning generalized linear models over normalized data. In *SIGMOD*, pages 1969–1984, 2015.
- [44] A. Kumar, J. F. Naughton, J. M. Patel, and X. Zhu. To join or not to join?: Thinking twice about joins before feature selection. In *SIGMOD*, pages 19–34, 2016.
- [45] Y. Liu, S. Zhou, and D. Niu. Implementation of statistical arbitrage using pca factorization approach, 2016.
- [46] S. Luo, Z. J. Gao, M. N. Gubanov, L. L. Perez, and C. M. Jermaine. Scalable linear algebra on a relational database system. *SIGMOD Rec.*, 47(1):24–31, 2018.
- [47] D. Marx. Approximating fractional hypertree width. *ACM Trans. Algorithms*, 6(2):29:1–29:17, Apr. 2010.
- [48] H. B. McMahan and et al. Ad click prediction: A view from the trenches. In *KDD*, pages 1222–1230, 2013.
- [49] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, D. Xin, R. Xin, M. J. Franklin, R. Zadeh, M. Zaharia, and A. Talwalkar. Mllib: Machine learning in apache spark. *J. Mach. Learn. Res.*, 17(1):1235–1241, 2016.
- [50] R. Menich and N. Vasiloglou. The future of LogicBlox machine learning. LogicBlox User Days, 2013.
- [51] D. Neumann. Lightning-fast deep learning on Spark via parallel stochastic gradient updates, [www.deepest.com](http://www.deepest.com), 2015.
- [52] H. Q. Ngo, E. Porat, C. Ré, and A. Rudra. Worst-case optimal join algorithms. In *PODS*, pages 37–48, 2012.
- [53] H. Q. Ngo, C. Ré, and A. Rudra. Skew strikes back: New developments in the theory of join algorithms. In *SIGMOD Rec.*, pages 5–16, 2013.
- [54] D. T. Nguyen, M. Aref, M. Bravenboer, G. Kollias, H. Q. Ngo, C. Ré, and A. Rudra. Join processing for graph patterns: An old dog with new tricks. *CoRR*, abs/1503.04169, 2015. Short version in GRADES@SIGMOD 2015.
- [55] D. Olteanu and M. Schleich. Factorized databases. *SIGMOD Rec.*, 45(2):5–16, 2016.
- [56] D. Olteanu and J. Závodný. Size bounds for factorised representations of query results. *TODS*, 40(1):2, 2015.
- [57] C. Ordonez, Y. Zhang, and W. Cabrera. The gamma matrix to summarize dense and sparse data sets for big data analytics. *IEEE Trans. Knowl. Data Eng.*, 28(7):1905–1918, 2016.
- [58] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *J. Machine Learning Research*, 12:2825–2830, 2011.
- [59] J. Pei, J. Han, and L. V. Lakshmanan. Mining frequent itemsets with convertible constraints. In *ICDE*, pages 433–442, 2001.
- [60] K. B. Petersen and M. S. Pedersen. The matrix cookbook, nov 2012. Version 20121115.
- [61] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich. Data management challenges in production machine learning. In *SIGMOD*, pages 1723–1726, 2017.

- [62] C. Qin and F. Rusu. Speculative approximations for terascale distributed gradient descent optimization. In *DanaC*, pages 1:1–1:10, 2015.
- [63] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, [www.r-project.org](http://www.r-project.org), 2013.
- [64] S. Rendle. Factorization machines with libFM. *ACM Trans. Intell. Syst. Technol.*, 3(3):57:1–57:22, 2012.
- [65] S. Rendle. Scaling factorization machines to relational data. *PVLDB*, 6(5):337–348, 2013.
- [66] M. Schleich, D. Olteanu, and R. Ciucanu. Learning linear regression models over factorized joins. In *SIGMOD*, pages 3–18, 2016.
- [67] G. Strang. *Linear Algebra and Its Applications*. Thomson, Brooks/Cole, 2006.
- [68] The StatsModels development team. StatsModels: Statistics in Python, <http://statsmodels.sourceforge.net>, 2012.
- [69] T. L. Veldhuizen. Triejoin: A simple, worst-case optimal join algorithm. In *ICDT*, pages 96–106, 2014.
- [70] Yelp. Yelp dataset challenge, <https://www.yelp.com/dataset/challenge/>, 2017.
- [71] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *NSDI*, pages 2–2, 2012.

## A Matrix Calculus

We introduce matrix inversion formulas and identities regarding tensors and the various products introduced in Section 2.

We use the following matrix inversion formulas [35].

**Proposition A.1.** *We have*

$$(\mathbf{B} + \mathbf{UCV})^{-1} = \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{VB}^{-1}\mathbf{U})^{-1}\mathbf{VB}^{-1}. \quad (61)$$

whenever all dimensions match up and inverses on the right hand side exist. In particular, the following holds when  $\mathbf{C} = (1)$ ,  $\mathbf{U} = \mathbf{1}$ ,  $\mathbf{V} = \mathbf{1}^\top$ , and  $\mathbf{J}$  is the all-1 matrix:

$$(\mathbf{B} + \mathbf{J})^{-1} = \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{1}(\mathbf{1} + \mathbf{1}^\top\mathbf{B}^{-1}\mathbf{1})^{-1}\mathbf{1}^\top\mathbf{B}^{-1}. \quad (62)$$

Another special case is

$$(\mathbf{A} + \mathbf{U}^\top\mathbf{U})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}^\top(\mathbf{I} + \mathbf{UA}^{-1}\mathbf{U}^\top)^{-1}\mathbf{UA}^{-1}. \quad (63)$$

An even more special case is the Sherman-Morrison formula, where  $\mathbf{U}^\top$  is just a vector  $\mathbf{u}$ . The matrix  $\mathbf{A} + \mathbf{uu}^\top$  is typically called a rank-1 update of  $\mathbf{A}$ :

$$(\mathbf{A} + \mathbf{uu}^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{uu}^\top\mathbf{A}^{-1}}{1 + \mathbf{u}^\top\mathbf{A}^{-1}\mathbf{u}}. \quad (64)$$

Next, we discuss some identities involving tensor, Khatri-Rao, and Hadamard products.

**Proposition A.2.** *We have (if the dimensionalities match up correctly):*

$$(\mathbf{AB} \otimes \mathbf{CD}) = (\mathbf{A} \otimes \mathbf{C})(\mathbf{B} \otimes \mathbf{D}) \quad (65)$$

$$(\mathbf{A} \otimes \mathbf{B})^\top = (\mathbf{A}^\top \otimes \mathbf{B}^\top) \quad (66)$$

$$\langle \mathbf{x}, \mathbf{By} \rangle = \langle \mathbf{B}^\top \mathbf{x}, \mathbf{y} \rangle \quad (67)$$

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = (\mathbf{A}^{-1} \otimes \mathbf{B}^{-1}) \quad \text{if both are square matrices} \quad (68)$$

$$\langle \mathbf{A} \otimes \mathbf{B}, \mathbf{RX} \otimes \mathbf{SY} \rangle = \langle \mathbf{R}^\top \mathbf{A} \otimes \mathbf{S}^\top \mathbf{B}, \mathbf{X} \otimes \mathbf{Y} \rangle. \quad (69)$$

If  $\mathbf{x}$  is a standard  $n$ -dimensional unit vector,  $\mathbf{A}$  and  $\mathbf{B}$  are two matrices with  $n$  columns each, and  $\mathbf{a}$  and  $\mathbf{b}$  are two  $n$ -dimensional vectors, then

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{x} \otimes \mathbf{x}) = (\mathbf{A} \star \mathbf{B})\mathbf{x} \quad (70)$$

$$\langle \mathbf{a} \otimes \mathbf{b}, \mathbf{x} \otimes \mathbf{x} \rangle = \langle \mathbf{a} \circ \mathbf{b}, \mathbf{x} \rangle. \quad (71)$$

Let  $\mathbf{x}$  be a standard  $n$ -dimensional unit vector,  $\mathbf{A}_1, \dots, \mathbf{A}_k$  be  $k$  matrices with  $n$  columns each. Then,

$$\left(\bigotimes_{i=1}^k \mathbf{A}_i\right)(\mathbf{x}^{\otimes k}) = \left(\bigstar_{i=1}^k \mathbf{A}_i\right)\mathbf{x}. \quad (72)$$

We note in passing that the first five identities are very useful in our dimension reduction techniques by exploiting functional dependencies, while (70), (71), and (72) are instrumental in achieving computational reduction in our handling of categorical features.

*Proof.* The identities (65), (66), (67), and (68) can be found in the Matrix Cookbook [60]. Identity (69) follows from (65) and (66). To see (70), note that

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{x} \otimes \mathbf{x}) = \mathbf{Ax} \otimes \mathbf{Bx} = (\mathbf{A} \star \mathbf{B})\mathbf{x},$$

where the last equality follows due to the following reasoning. Suppose  $x_j = 1$  for some  $j$ , then  $\mathbf{Ax} = \mathbf{a}_j$  and  $\mathbf{Bx} = \mathbf{b}_j$ , where  $\mathbf{a}_j$  and  $\mathbf{b}_j$  are the  $j$ th columns of  $\mathbf{A}$  and  $\mathbf{B}$ , respectively. Thus,

$$\mathbf{Ax} \otimes \mathbf{Bx} = \mathbf{a}_j \otimes \mathbf{b}_j = (\mathbf{A} \star \mathbf{B})_j = (\mathbf{A} \star \mathbf{B})\mathbf{x}.$$

Identities (71) and (72) are proved similarly, where (72) is a trivial generalization of (70).  $\square$

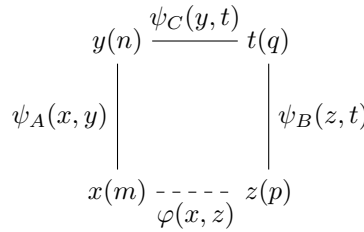
## B Tensor computation and FAQ queries

Quite often we need to compute a product of the form  $(\mathbf{A} \otimes \mathbf{B})\mathbf{C}$ , where  $\mathbf{A}, \mathbf{B}$ , and  $\mathbf{C}$  are tensors, provided that their dimensionalities match up. For example, suppose  $\mathbf{A}$  is an  $m \times n$  matrix,  $\mathbf{B}$  a  $p \times q$  matrix, and  $\mathbf{C}$  a  $nq \times 1$  matrix (i.e. a vector). The result is a  $mp \times 1$  tensor. The brute-force way of computing  $(\mathbf{A} \otimes \mathbf{B})\mathbf{C}$  is to compute  $\mathbf{A} \otimes \mathbf{B}$  first, taking  $\Theta(mnpq)$ -time, and then multiply the result with  $\mathbf{C}$ , for an overall runtime of  $\Theta(mnpq)$ . The brute-force algorithm is a horribly inefficient algorithm.

The better way to compute  $(\mathbf{A} \otimes \mathbf{B})\mathbf{C}$  is to view this as an FAQ-expression [8] (a sum-product form): we think of  $\mathbf{A}$  as a function  $\psi_A(x, y)$ ,  $\mathbf{B}$  as a function  $\psi_B(z, t)$ , and  $\mathbf{C}$  as a function  $\psi_C(y, t)$ . What we want to compute is the function

$$\varphi(x, z) = \sum_y \sum_t \psi_A(x, y) \psi_B(z, t) \psi_C(y, t). \quad (73)$$

This is a 4-cycle FAQ query:



We can pick between the following two evaluation strategies:

- Eliminate  $t$  first, i.e., compute  $\varphi_1(y, z) := \sum_t \psi_B(z, t) \psi_C(y, t)$  in time  $O(npq)$ ; then, eliminate  $y$ , i.e., compute  $\varphi(x, z) = \sum_y \varphi_1(y, z) \psi_A(x, y)$  in time  $O(mnp)$ . The overall runtime is thus  $O(np(m + q))$ .
- Eliminate  $y$  first and then  $t$ . The overall runtime is  $O(mq(n + p))$ .

This is not surprising, since the problem is just matrix chain multiplication. In the language of FAQ evaluation, we want to pick the best tree decomposition and then compute a variable elimination order out of it [8]. We shall see later that a special case of the above that occurs often is when  $\mathbf{B} = \mathbf{I}$ , the identity matrix. In that case,  $\psi_B(z, t)$  is the same as the atom  $z = t$ , and thus it serves as a change of variables:

$$\varphi(x, z) = \sum_y \sum_t \psi_A(x, y) \psi_B(z, t) \psi_C(y, t) = \sum_y \psi_A(x, y) \psi_C(y, z).$$

In other words, we only have to marginalize out one variable instead of two. This situation arises, for example, in (50) and (51).

Appendix C overviews the InsideOut algorithm for FAQ queries and its complexity analysis.

## C Widths for FAQ Queries and the InsideOut Algorithm

### Background: Fractional edge cover number and output size bounds

In what follows, we consider a conjunctive query  $Q$  over a relational database instance  $I$ . We use  $N$  to denote the size of the largest input relation in  $Q$ . We also use  $Q(I)$  to denote the output and  $|Q(I)|$  to denote its size. We use the query  $Q$  and its hypergraph  $\mathcal{H}$  interchangeably.

**Definition 6** (Fractional edge cover number  $\rho^*$ ). Let  $\mathcal{H} = (\mathcal{V}, \mathcal{E})$  be a hypergraph (of some query  $Q$ ). Let  $B \subseteq \mathcal{V}$  be any subset of vertices. A *fractional edge cover* of  $B$  using edges in  $\mathcal{H}$  is a feasible solution  $\lambda = (\lambda_S)_{S \in \mathcal{E}}$  to the

following linear program:

$$\begin{aligned}
\min \quad & \sum_{S \in \mathcal{E}} \lambda_S \\
\text{s.t.} \quad & \sum_{S: v \in S} \lambda_S \geq 1, \quad \forall v \in B \\
& \lambda_S \geq 0, \quad \forall S \in \mathcal{E}.
\end{aligned}$$

The optimal objective value of the above linear program is called the *fractional edge cover number* of  $B$  in  $\mathcal{H}$  and is denoted by  $\rho_{\mathcal{H}}^*(B)$ . When  $\mathcal{H}$  is clear from the context, we drop the subscript  $\mathcal{H}$  and use  $\rho^*(B)$ .

Given a conjunctive query  $Q$ , the fractional edge cover number of  $Q$  is  $\rho_{\mathcal{H}}^*(\mathcal{V})$  where  $\mathcal{H} = (\mathcal{V}, \mathcal{E})$  is the hypergraph of  $Q$ .

**Theorem C.1** (AGM-bound [12, 33]). *Given a full conjunctive query  $Q$  over a relational database instance  $I$ , the output size is bounded by*

$$|Q(I)| \leq N^{\rho^*},$$

where  $\rho^*$  is the fractional edge cover number of  $Q$ .

**Theorem C.2** (AGM-bound is tight [12, 33]). *Given a full conjunctive query  $Q$  and a non-negative number  $N$ , there exists a database instance  $I$  whose relation sizes are upper-bounded by  $N$  and satisfies*

$$|Q(I)| = \Theta(N^{\rho^*}).$$

Worst-case optimal join algorithms [69, 52, 53, 6] can be used to answer any full conjunctive query  $Q$  in time

$$O(|\mathcal{V}| \cdot |\mathcal{E}| \cdot N^{\rho^*} \cdot \log N). \quad (74)$$

## Background: Tree decompositions, acyclicity, and width parameters

**Definition 7** (Tree decomposition). Let  $\mathcal{H} = (\mathcal{V}, \mathcal{E})$  be a hypergraph. A *tree decomposition* of  $\mathcal{H}$  is a pair  $(T, \chi)$  where  $T = (V(T), E(T))$  is a tree and  $\chi : V(T) \rightarrow 2^{\mathcal{V}}$  assigns to each node of the tree  $T$  a subset of vertices of  $\mathcal{H}$ . The sets  $\chi(t)$ ,  $t \in V(T)$ , are called the *bags* of the tree decomposition. There are two properties the bags must satisfy

- (a) For any hyperedge  $F \in \mathcal{E}$ , there is a bag  $\chi(t)$ ,  $t \in V(T)$ , such that  $F \subseteq \chi(t)$ .
- (b) For any vertex  $v \in \mathcal{V}$ , the set  $\{t \mid t \in V(T), v \in \chi(t)\}$  is not empty and forms a connected subtree of  $T$ .

**Definition 8** (acyclicity). A hypergraph  $\mathcal{H} = (\mathcal{V}, \mathcal{E})$  is *acyclic* iff there exists a tree decomposition  $(T, \chi)$  in which every bag  $\chi(t)$  is a hyperedge of  $\mathcal{H}$ .

When  $\mathcal{H}$  represents a join query, the tree  $T$  in the above definition is also called the *join tree* of the query. A query is acyclic if and only if its hypergraph is acyclic.

For non-acyclic queries, we often need a measure of how “close” a query is to being acyclic. To that end, we use *width* notions of a query.

**Definition 9** ( $g$ -width of a hypergraph: a generic width notion [9]). Let  $\mathcal{H} = (\mathcal{V}, \mathcal{E})$  be a hypergraph, and  $g : 2^{\mathcal{V}} \rightarrow \mathbb{R}^+$  be a function that assigns a non-negative real number to each subset of  $\mathcal{V}$ . The  $g$ -width of a tree decomposition  $(T, \chi)$  of  $\mathcal{H}$  is  $\max_{t \in V(T)} g(\chi(t))$ . The  $g$ -width of  $\mathcal{H}$  is the *minimum*  $g$ -width over all tree decompositions of  $\mathcal{H}$ . (Note that the  $g$ -width of a hypergraph is a *Minimax* function.)

**Definition 10** (*Treewidth* and *fractional hypertree width* are special cases of  $g$ -width). Let  $s$  be the following function:  $s(B) = |B| - 1$ ,  $\forall B \subseteq \mathcal{V}$ . Then the *treewidth* of a hypergraph  $\mathcal{H}$ , denoted by  $\text{tw}(\mathcal{H})$ , is exactly its  $s$ -width, and the *fractional hypertree width* of a hypergraph  $\mathcal{H}$ , denoted by  $\text{fhtw}(\mathcal{H})$ , is the  $\rho^*$ -width of  $\mathcal{H}$ .

From the above definitions,  $\text{fhtw}(\mathcal{H}) \geq 1$  for any hypergraph  $\mathcal{H}$ . Moreover,  $\text{fhtw}(\mathcal{H}) = 1$  if and only if  $\mathcal{H}$  is acyclic.

## Background: Vertex/variable orderings and their equivalence to tree decompositions

Besides tree decompositions, there is another way to define acyclicity and width notions of a hypergraph, which is *orderings* of the hypergraph vertices. And just like we refer to queries and hypergraphs interchangeably, we also refer to query variables and hypergraph vertices interchangeably.

In what follows, we use  $n$  to denote the number of vertices of the given hypergraph  $\mathcal{H}$ .

**Definition 11** (Vertex ordering of a hypergraph). A *vertex ordering* of a hypergraph  $\mathcal{H} = (\mathcal{V}, \mathcal{E})$  is simply a listing  $\sigma = (v_1, \dots, v_n)$  of all vertices in  $\mathcal{V}$ .

**Definition 12** (Elimination sets  $U_j^\sigma$  of a vertex ordering  $\sigma$ ). Given a hypergraph  $\mathcal{H} = (\mathcal{V}, \mathcal{E})$  and a vertex ordering  $\sigma = (v_1, \dots, v_n)$ , we define sets  $U_1^\sigma, \dots, U_n^\sigma \subseteq \mathcal{V}$ , called the *elimination sets* of  $\sigma$ , as follows: Let  $\partial(v_n)$  be the set of hyperedges of  $\mathcal{H}$  that contain  $v_n$ . We define  $U_n^\sigma$  to be the union of all hyperedges in  $\partial(v_n)$ :

$$U_n^\sigma = \bigcup_{S \in \partial(v_n)} S.$$

If  $n = 1$ , then we are done. Otherwise, we remove vertex  $v_n$  and all hyperedges in  $\partial(v_n)$  from  $\mathcal{H}$  and add back to  $\mathcal{H}$  a new hyperedge  $U_n^\sigma - \{v_n\}$ , thus turning  $\mathcal{H}$  into a hypergraph with  $n - 1$  vertices:

$$\begin{aligned} \mathcal{V} &\leftarrow \mathcal{V} - \{v_n\}, \\ \mathcal{E} &\leftarrow (\mathcal{E} - \partial(v_n)) \cup \{U_n^\sigma - \{v_n\}\}. \end{aligned}$$

The remaining elimination sets  $U_1^\sigma, \dots, U_{n-1}^\sigma$  are defined inductively to be the elimination sets of the resulting hypergraph (whose vertices are now  $\{v_1, \dots, v_{n-1}\}$ ).

When  $\sigma$  is clear from the context, we drop the superscript  $\sigma$  and use  $U_1, \dots, U_n$ .

**Proposition C.3** (Every vertex ordering has an “equivalent” tree decomposition [7]). *Given a hypergraph  $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ , for every vertex ordering  $\sigma$ , there is a tree decomposition  $(T, \chi)$  whose bags  $\chi(t)$  are the elimination sets  $U_j^\sigma$  of  $\sigma$ .*

By applying the GYO elimination procedure [3] on the bags of any given tree decomposition, we can obtain an “equivalent” vertex ordering:

**Proposition C.4** (Every tree decomposition has an “equivalent” vertex ordering [7]). *Given a hypergraph  $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ , for every tree decomposition  $(T, \chi)$ , there is a vertex ordering  $\sigma$  such that every elimination set  $U_j^\sigma$  of  $\sigma$  is contained in some bag  $\chi(t)$  of the tree decomposition  $(T, \chi)$ .*

## FAQ-width of an FAQ query

Just like a conjunctive query, an FAQ query has a query hypergraph  $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ . But unlike conjunctive queries, an FAQ query also specifies an order of its variables, which is the order in which we aggregate over those variables in the given FAQ-expression. (For example, in expression (73), we sum over  $t$  first, then over  $y$ , and we keep  $z$  and  $x$  as free variables. Hence, the FAQ query in (73) specifies the variable order  $\sigma = (x, z, y, t)$ .) Such a variable order for the query can also be interpreted as a vertex order  $\sigma$  for the query’s hypergraph.

The InsideOut algorithm for answering FAQ queries is based on *variable elimination*. To eliminate variable/vertex  $v_n$ , we have to solve a sub-problem consisting of a smaller FAQ query over the variables in the elimination set  $U_n^\sigma$ . This smaller query can be solved by an algorithm that is based on worst-case optimal join algorithms [69, 52, 53, 6]. From (74), this takes time <sup>5</sup>

$$O(|\mathcal{V}| \cdot |\mathcal{E}| \cdot N^{\rho_{\mathcal{H}}^*(U_n^\sigma)} \cdot \log N). \quad (75)$$

After eliminating  $v_n$ , the remaining variables  $v_{n-1}, v_{n-2}, \dots, v_1$  can be eliminated similarly. This variable elimination algorithm motivates the following width notion.

<sup>5</sup>To achieve this runtime, we need some additional ideas that are beyond the scope of this very brief introduction to FAQ. See [8] for more details.

**Definition 13** (FAQ-width of a given variable ordering  $\sigma$ ). Given an FAQ query  $\varphi$  with a variable ordering  $\sigma$ , we define the FAQ-width of  $\sigma$ , denoted by  $\text{faqw}(\sigma)$ , to be

$$\text{faqw}(\sigma) = \max_{j \in [n]} \{\rho_{\mathcal{H}}^*(U_j^\sigma)\}. \quad (76)$$

By the above definition, the FAQ-width of a variable ordering  $\sigma$  is the same as the fractional hypertree width of the “equivalent” tree decomposition that is referred to in Proposition C.3.

**Theorem C.5** (Runtime of *InsideOut* [8]). *Given an FAQ-query  $\varphi$  with a variable ordering  $\sigma$ , the InsideOut algorithm answers  $\varphi$  in time*

$$O\left(|\mathcal{V}|^2 \cdot |\mathcal{E}| \cdot \left(N^{\text{faqw}(\sigma)} + |\varphi|\right) \cdot \log N\right), \quad (77)$$

where  $|\varphi|$  is the output size in the listing representation.

Let  $\varphi$  be an FAQ query with variable ordering  $\sigma$ . In many cases, there might be a different variable ordering  $\sigma'$  such that if we were to permute the aggregates of  $\varphi$  in the order of  $\sigma'$  instead of  $\sigma$ , we would obtain an FAQ-query  $\varphi'$  that is “semantically-equivalent” to  $\varphi$  (i.e. that always returns the same answer as  $\varphi$  no matter what the input is). If this is the case, then we can run *InsideOut* on  $\varphi$  using the ordering  $\sigma'$  instead of  $\sigma$ , which can lead to a better runtime if  $\text{faqw}(\sigma')$  happens to be smaller than  $\text{faqw}(\sigma)$ . We use  $\text{EVO}(\varphi)$  to denote the set of all such “equivalent” orderings  $\sigma'$ . (For a formal definition, see [8].) Therefore, it is best to consider all orderings  $\sigma'$  in  $\text{EVO}(\varphi)$ , pick the one with the smallest  $\text{faqw}(\sigma')$ , and use it in *InsideOut* algorithm. This motivates the following definition.

**Definition 14** (FAQ-width of an FAQ query). The FAQ-width of an FAQ query  $\varphi$ , denoted by  $\text{faqw}(\varphi)$ , is the minimum one over all orderings  $\sigma'$  in  $\text{EVO}(\varphi)$ , i.e.

$$\text{faqw}(\varphi) = \min \{\text{faqw}(\sigma') \mid \sigma' \in \text{EVO}(\varphi)\}. \quad (78)$$

Characterizing  $\text{EVO}(\varphi)$  for an arbitrary given FAQ-query  $\varphi$  is a technically involved problem (see [8] for hardness background and a general solution). However, the FAQ queries that we need for our machine learning tasks are of a special form that makes the problem easier.: The aggregate operator that we use in such queries is the summation operator  $\sum$ . We refer to those restricted FAQ queries as FAQ-SS queries (see [8]). Our FAQ-SS queries in this work have only two types of variables:

- Variables that we are summing over, e.g. variables  $y$  and  $t$  in (73).
- Free variables (i.e. Group-by variables), e.g. variables  $x$  and  $z$ .

Given an FAQ-SS query  $\varphi$ ,  $\text{EVO}(\varphi)$  contains every ordering  $\sigma'$  that lists all free variables *before* the non-free variables. For example, for the FAQ-SS query  $\varphi(x, z)$  in (73),  $\text{EVO}(\varphi(x, z))$  contains all permutations of  $\{x, y, z, t\}$  where  $\{x, z\}$  come before  $\{y, t\}$ .

**Proposition C.6.** *For any FAQ-SS query  $\varphi$  without free variables, we have  $\text{faqw}(\varphi) = \text{fhtw}(\mathcal{H})$ , where  $\mathcal{H}$  is the hypergraph of  $\mathcal{H}$ .*

*Proof.* In this case,  $\text{EVO}(\varphi)$  contains all  $n!$  possible orderings. By Proposition C.4, for every tree decomposition  $(T, \chi)$ , there is an ordering  $\sigma'$  such that  $\text{faqw}(\sigma') \leq \text{fhtw}((T, \chi))$ . By Proposition C.3, for every ordering  $\sigma'$ , there is a tree decomposition  $(T, \chi)$  such that  $\text{fhtw}((T, \chi)) = \text{faqw}(\sigma')$ . Therefore, we have

$$\min_{\sigma' \in \text{EVO}(\varphi)} \text{faqw}(\sigma') = \min_{(T, \chi)} \text{fhtw}((T, \chi)).$$

□

**Proposition C.7.** *For any FAQ-SS query  $\varphi$  with  $f \geq 1$  free variables, we have  $\text{faqw}(\varphi) \leq \text{fhtw}(\mathcal{H}) + f - 1$ , where  $\mathcal{H}$  is the hypergraph of  $\mathcal{H}$ .*

*Proof.* Find a tree decomposition  $(T, \chi)$  of  $\mathcal{H}$  with minimal  $\text{fhtw}$ , i.e. where  $\text{fhtw}((T, \chi)) = \text{fhtw}(\mathcal{H})$ . WLOG let the  $f$  free variables be  $v_1, \dots, v_f$ . Construct another tree decomposition  $(T, \bar{\chi})$  by extending all bags  $\chi(t)$  of  $(T, \chi)$  with the variables  $\{v_2, \dots, v_f\}$ , i.e. by defining  $\bar{\chi}(t) = \chi(t) \cup \{v_2, \dots, v_f\}$  for all  $t$ . By Definition 7,  $(T, \bar{\chi})$  is indeed a tree decomposition. And because  $\rho^*(\chi(t) \cup \{v_2, \dots, v_f\}) \leq \rho^*(\chi(t)) + f - 1$ , we have

$$\text{fhtw}((T, \bar{\chi})) \leq \text{fhtw}((T, \chi)) + f - 1.$$

Moreover, since  $(T, \chi)$  must have a bag  $\chi(t^*)$  that contains  $v_1$ , the corresponding bag  $\bar{\chi}(t^*)$  of  $(T, \bar{\chi})$  contains all the free variables  $\{v_1, \dots, v_f\}$ . We designate  $t^*$  as the root of  $T$ , and then we run GYO elimination procedure [3] on the bags  $\bar{\chi}(t)$  of  $(T, \bar{\chi})$  to construct a vertex ordering  $\sigma'$  with  $\text{faqw}(\sigma') \leq \text{fhtw}((T, \bar{\chi}))$ . Moreover, if we choose to eliminate the vertices of the root  $t^*$  at the end of GYO elimination (after all other vertices have already been eliminated), we can make the free variables  $\{v_1, \dots, v_f\}$  appear before all other variables in  $\sigma'$ , thus making sure that  $\sigma'$  is indeed in  $\text{EVO}(\varphi)$  and completing the proof. In particular, we apply GYO elimination as follows:

- If the tree  $T$  contains only one node  $t^*$ :
  - We eliminate vertices in  $\bar{\chi}(t^*) - \{v_1, \dots, v_f\}$  before eliminating  $\{v_1, \dots, v_f\}$ .
  - We remove  $t^*$  from  $T$ , thus making  $T$  an empty tree.
- Otherwise, we pick a *leaf* node  $t$  of  $T$  (other than the root  $t^*$ ). Let  $t'$  be the parent of  $t$  in  $T$ :
  - If  $\bar{\chi}(t) \subseteq \bar{\chi}(t')$ , then we remove node  $t$  from  $T$  along with the associated bag  $\bar{\chi}(t)$ .
  - Otherwise,  $\bar{\chi}(t)$  must have a vertex  $u$  that is not in  $\bar{\chi}(t')$ . (Hence, by property (b) of Definition 7,  $u$  is not in  $\bar{\chi}(t'')$  for all  $t''$  in  $T$  other than  $t$ .)
    - \* If  $u$  is the only vertex in  $\bar{\chi}(t)$ , then we remove node  $t$  from  $T$  along with the associated bag  $\bar{\chi}(t)$ .
    - \* Otherwise, we remove  $u$  from  $\bar{\chi}(t)$ .
- We repeat the above steps until  $T$  becomes an empty tree.

□

## D Missing details from Section 4

*Proof of Theorem 4.1.* We start with point evaluation:

$$\begin{aligned}
 \frac{1}{2|D|} \sum_{(\mathbf{x}, y) \in D} (\langle g(\boldsymbol{\theta}), h(\mathbf{x}) \rangle - y)^2 &= \frac{1}{2|D|} \sum_{(\mathbf{x}, y) \in D} (\langle g(\boldsymbol{\theta}), h(\mathbf{x}) \rangle^2 - 2y \langle g(\boldsymbol{\theta}), h(\mathbf{x}) \rangle + y^2) \\
 &= \frac{1}{2|D|} \sum_{(\mathbf{x}, y) \in D} g(\boldsymbol{\theta})^\top (h(\mathbf{x})h(\mathbf{x})^\top) g(\boldsymbol{\theta}) - \left\langle g(\boldsymbol{\theta}), \frac{1}{|D|} \sum_{(\mathbf{x}, y) \in D} yh(\mathbf{x}) \right\rangle \\
 &\quad + \frac{1}{2|D|} \sum_{(\mathbf{x}, y) \in D} y^2 \\
 &= \frac{1}{2} g(\boldsymbol{\theta})^\top \left( \frac{1}{|D|} \sum_{(\mathbf{x}, y) \in D} h(\mathbf{x})h(\mathbf{x})^\top \right) g(\boldsymbol{\theta}) - \langle g(\boldsymbol{\theta}), \mathbf{c} \rangle + \frac{s_Y}{2} \\
 &= \frac{1}{2} g(\boldsymbol{\theta})^\top \Sigma g(\boldsymbol{\theta}) - \langle g(\boldsymbol{\theta}), \mathbf{c} \rangle + \frac{s_Y}{2}.
 \end{aligned}$$

The gradient formula follows straightforwardly from (16) and the chain rule.

□



*Proof of Corollary 4.2.* From (18) we have

$$\begin{aligned}
J(\boldsymbol{\theta}) - J(\boldsymbol{\theta} - \alpha \mathbf{d}) &= \frac{1}{2} \boldsymbol{\theta}^\top \Sigma \boldsymbol{\theta} - \frac{1}{2} (\boldsymbol{\theta} - \alpha \mathbf{d})^\top \Sigma (\boldsymbol{\theta} - \alpha \mathbf{d}) - \langle \boldsymbol{\theta}, \mathbf{c} \rangle + \langle \boldsymbol{\theta} - \alpha \mathbf{d}, \mathbf{c} \rangle + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2 - \frac{\lambda}{2} \|\boldsymbol{\theta} - \alpha \mathbf{d}\|_2^2 \\
&= \frac{1}{2} \boldsymbol{\theta}^\top \Sigma \boldsymbol{\theta} - \frac{1}{2} \left( \boldsymbol{\theta}^\top \Sigma \boldsymbol{\theta} - 2\alpha \boldsymbol{\theta}^\top \Sigma \mathbf{d} + \alpha^2 \mathbf{d}^\top \Sigma \mathbf{d} \right) - \alpha \langle \mathbf{d}, \mathbf{c} \rangle + \lambda \alpha \langle \boldsymbol{\theta}, \mathbf{d} \rangle - \frac{\lambda \alpha^2}{2} \|\mathbf{d}\|_2^2 \\
&= \alpha \boldsymbol{\theta}^\top \Sigma \mathbf{d} - \frac{\alpha^2}{2} \mathbf{d}^\top \Sigma \mathbf{d} - \alpha \langle \mathbf{d}, \mathbf{c} \rangle + \lambda \alpha \langle \boldsymbol{\theta}, \mathbf{d} \rangle - \frac{\lambda \alpha^2}{2} \|\mathbf{d}\|_2^2.
\end{aligned}$$

□

*Proof of Proposition 4.4.* For any event  $E$ , let  $\delta_E$  denote the Kronecker delta, i.e.  $\delta_E = 1$  if  $E$  holds, and  $\delta_E = 0$  otherwise. Recall that the input query  $Q$  has hypergraph  $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ , and there is an input relation  $R_F$  for every hyperedge  $F \in \mathcal{E}$ . Recall that we can write  $\sigma_{ij}$  in the tensor form as shown in Eq. (23). Plugging in the definition of  $h_i$  and  $h_j$  from (9); and, let  $C_{ij} = C_i \cup C_j$  and  $V_{ij} = V_i \cup V_j$ , we have

$$\sigma_{ij} = \frac{1}{|D|} \sum_{(\mathbf{x}, y) \in D} \prod_{f \in V_{ij} - C_{ij}} x_f^{a_i(f) + a_j(f)} \cdot \bigotimes_{f_i \in C_i} \mathbf{x}_{f_i} \otimes \bigotimes_{f_j \in C_j} \mathbf{x}_{f_j}.$$

As illustrated in Example 13, the tensor  $\bigotimes_{f \in C_i} \mathbf{x}_f \otimes \bigotimes_{f \in C_j} \mathbf{x}_f$  is very sparse. For a fixed tuple  $\mathbf{x}$ , in fact, the tensor has only *one* 1 entry, corresponding to the combination of values of the attributes in  $C_{ij}$ . Hence,  $\sigma_{ij}$  is a function of the variables  $C_{ij}$ . In the FAQ-framework, the query representing  $\sigma_{ij}$  can be expressed as a Sum-Product queries with free (i.e., group-by) variables  $C_{ij}$ , defined by:

$$\varphi(C_{ij}) = \frac{1}{|D|} \sum_{x_{f'}: f' \in \mathcal{V} - C_{ij}} \prod_{f \in V_{ij} - C_{ij}} x_f^{a_i(f) + a_j(f)} \cdot \prod_{F \in \mathcal{E}} \delta_{\pi_F(\mathbf{x}) \in R_F}. \quad (79)$$

Similarly, the tensor  $\mathbf{c}_j$  can be sparsely represented by an aggregate query with group-by attributes  $C_j$ , which is expressed as the Sum-Product query

$$\varphi(C_j) = \frac{1}{|D|} \sum_{x_{f'}: f' \in \mathcal{V} - C_j} y \cdot \prod_{f \in V_j - C_j} x_f^{a_j(f)} \cdot \prod_{F \in \mathcal{E}} \delta_{\pi_F(\mathbf{x}) \in R_F}. \quad (80)$$

The overall runtimes for computing the above FAQ-queries follow from applying the InsideOut algorithm and Theorem C.5 [8]. □

*Proof of Proposition 4.5.* The fact that  $\text{faqw}(i, j) \leq \text{fhtw} + c - 1$  follows from Proposition C.7. Since  $\sigma_{ij}$  is a tensor of order at most  $c$ , and each attribute's active domain has size at most  $N$ , it follows that  $|\sigma_{ij}| \leq N^c$ . And,  $|\sigma_{ij}| \leq |D|$  because the support of the tensor  $\sigma_{ij}$  cannot be more than the output size.

Fix a query  $Q$  with  $\rho^* > \text{fhtw} + c - 1 \geq c$ . Consider a database instance  $I$  for which  $|D|$  (the output size of  $Q$ ) is  $\Theta(N^{\rho^*})$ . (The existence of such database instances is guaranteed by Theorem C.2.) From this (33) follows trivially. □

*Proof of Proposition 4.6.* We first analyze the time it takes to compute expression (16), which is dominated by the quadratic form  $g(\boldsymbol{\theta})^\top \Sigma g(\boldsymbol{\theta})$ . To compute this quadratic form, for every pair  $i, j \in [m]$  we need to compute  $g_i(\boldsymbol{\theta})^\top \sigma_{ij} g_j(\boldsymbol{\theta})$ . This product is broken up into a sum of  $t_i t_j$  terms when we expand  $g_i$  and  $g_j$  out. Each of those terms is computed in time  $O(d_i d_j |\sigma_{ij}|)$ . The runtime for computing (17) is analyzed similarly. □

## E Missing details from Section 5

In the proofs below, for each feature  $w \in V$ ,  $\mathbf{I}_w$  denote the identity matrix whose dimension is the size of the effective domain of  $w$ . This is not to be confused with the notation  $\mathbf{I}_n$  which is an order- $n$  identity matrix.

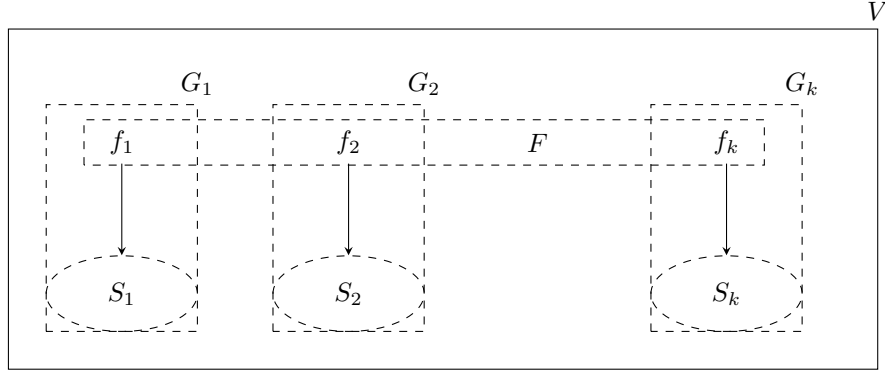


Figure 8: Groups of simple FDs.  $G = G_1 \cup \dots \cup G_k$ .

### E.1 Missing rewriting steps in the example in Section 5.1

We first prove identity (42) formally. By setting the gradient in (41) to 0, we have  $\theta_{\text{country}} = (\mathbf{I}_{\text{country}} + \mathbf{R}\mathbf{R}^\top)^{-1}\mathbf{R}\gamma_{\text{city}}$ . Hence, it remains to show the identity

$$(\mathbf{I}_{\text{country}} + \mathbf{R}\mathbf{R}^\top)^{-1}\mathbf{R} = \mathbf{R}(\mathbf{I}_{\text{city}} + \mathbf{R}^\top\mathbf{R})^{-1}.$$

To see this, we apply the Sherman-Morrison-Woodbury identity (63) with  $\mathbf{A} = \mathbf{I}_{\text{country}}$  and  $\mathbf{U} = \mathbf{R}^\top$  to get

$$(\mathbf{I}_{\text{country}} + \mathbf{R}\mathbf{R}^\top)^{-1} = \mathbf{I}_{\text{country}} - \mathbf{R}(\mathbf{I}_{\text{city}} + \mathbf{R}^\top\mathbf{R})^{-1}\mathbf{R}^\top \quad (81)$$

multiply both sides of (81) on the right by  $\mathbf{R}$ , we obtain

$$\begin{aligned} (\mathbf{I}_{\text{country}} + \mathbf{R}\mathbf{R}^\top)^{-1}\mathbf{R} &= \mathbf{R} - \mathbf{R}(\mathbf{I}_{\text{city}} + \mathbf{R}^\top\mathbf{R})^{-1}\mathbf{R}^\top\mathbf{R} \\ &= \mathbf{R}[\mathbf{I}_{\text{city}} - (\mathbf{I}_{\text{city}} + \mathbf{R}^\top\mathbf{R})^{-1}\mathbf{R}^\top\mathbf{R}] \\ &= \mathbf{R}[\mathbf{I}_{\text{city}} - (\mathbf{I}_{\text{city}} + \mathbf{R}^\top\mathbf{R})^{-1}(\mathbf{I}_{\text{city}} + \mathbf{R}^\top\mathbf{R} - \mathbf{I}_{\text{city}})] \\ &= \mathbf{R}[\mathbf{I}_{\text{city}} - \mathbf{I}_{\text{city}} + (\mathbf{I}_{\text{city}} + \mathbf{R}^\top\mathbf{R})^{-1}] \\ &= \mathbf{R}(\mathbf{I}_{\text{city}} + \mathbf{R}^\top\mathbf{R})^{-1}. \end{aligned}$$

We next show (43); to do so, it is sufficient to verify that

$$\|\gamma_{\text{city}} - \mathbf{R}^\top\theta_{\text{country}}\|_2^2 + \|\theta_{\text{country}}\|_2^2 = \langle (\mathbf{I}_{\text{city}} + \mathbf{R}^\top\mathbf{R})^{-1}\gamma_{\text{city}}, \gamma_{\text{city}} \rangle \quad (82)$$

For the sake of brevity, define  $\mathbf{B} = \mathbf{I}_{\text{city}} + \mathbf{R}^\top\mathbf{R}$  so that  $\theta_{\text{country}} = \mathbf{R}\mathbf{B}^{-1}\gamma_{\text{city}}$ . We compute each term on the left hand side separately:

$$\begin{aligned} \gamma_{\text{city}} - \mathbf{R}^\top\theta_{\text{country}} &= \gamma_{\text{city}} - \mathbf{R}^\top(\mathbf{I}_{\text{country}} + \mathbf{R}\mathbf{R}^\top)^{-1}\mathbf{R}\gamma_{\text{city}} \\ &= [\mathbf{I}_{\text{city}} - \mathbf{R}^\top(\mathbf{I}_{\text{country}} + \mathbf{R}\mathbf{R}^\top)^{-1}\mathbf{R}]\gamma_{\text{city}} \\ (\text{follows from (63)}) &= \mathbf{B}^{-1}\gamma_{\text{city}}, \end{aligned}$$

From here, we derive (82) by

$$\begin{aligned} \|\gamma_{\text{city}} - \mathbf{R}^\top\theta_{\text{country}}\|_2^2 + \|\theta_{\text{country}}\|_2^2 &= \|\mathbf{B}^{-1}\gamma_{\text{city}}\|_2^2 + \|\mathbf{R}\mathbf{B}^{-1}\gamma_{\text{city}}\|_2^2 \\ &= \langle \mathbf{B}^{-1}\gamma_{\text{city}}, \mathbf{B}^{-1}\gamma_{\text{city}} \rangle + \langle \mathbf{R}\mathbf{B}^{-1}\gamma_{\text{city}}, \mathbf{R}\mathbf{B}^{-1}\gamma_{\text{city}} \rangle \\ &= \langle \mathbf{B}^{-1}\gamma_{\text{city}}, \mathbf{B}^{-1}\gamma_{\text{city}} \rangle + \langle \mathbf{B}^{-1}\gamma_{\text{city}}, \mathbf{R}^\top\mathbf{R}\mathbf{B}^{-1}\gamma_{\text{city}} \rangle \\ &= \langle \mathbf{B}^{-1}\gamma_{\text{city}}, (\mathbf{I}_{\text{city}} + \mathbf{R}^\top\mathbf{R})\mathbf{B}^{-1}\gamma_{\text{city}} \rangle \\ &= \langle \mathbf{B}^{-1}\gamma_{\text{city}}, \gamma_{\text{city}} \rangle. \end{aligned}$$

## E.2 Proof of Theorem 5.1

*Proof.* We start by breaking the loss term into two parts

$$\langle \theta, h(\mathbf{x}) \rangle = \sum_{\|\mathbf{a}_V\|_1 \leq d} \langle \theta_{\mathbf{a}}, h_{\mathbf{a}}(\mathbf{x}) \rangle = \sum_{\substack{\|\mathbf{a}_V\|_1 \leq d \\ \|\mathbf{a}_G\|_1 = 0}} \langle \theta_{\mathbf{a}}, \mathbf{x}^{\otimes \mathbf{a}} \rangle + \sum_{\substack{\|\mathbf{a}_V\|_1 \leq d \\ \|\mathbf{a}_G\|_1 > 0}} \langle \theta_{\mathbf{a}}, \mathbf{x}^{\otimes \mathbf{a}} \rangle$$

and rewrite the second part:

$$\sum_{\substack{\|\mathbf{a}_V\|_1 \leq d \\ \|\mathbf{a}_G\|_1 > 0}} \langle \theta_{\mathbf{a}}, \mathbf{x}^{\otimes \mathbf{a}} \rangle = \sum_{\substack{\|\mathbf{a}_V\|_1 \leq d \\ \|\mathbf{a}_G\|_1 > 0}} \left\langle \theta_{\mathbf{a}}, \mathbf{x}_{\overline{G}}^{\otimes \mathbf{a}_{\overline{G}}} \otimes \mathbf{x}_G^{\otimes \mathbf{a}_G} \right\rangle = \sum_{\substack{\|\mathbf{a}_V\|_1 \leq d \\ \|\mathbf{a}_G\|_1 > 0}} \left\langle \theta_{\mathbf{a}}, \mathbf{x}_{\overline{G}}^{\otimes \mathbf{a}_{\overline{G}}} \otimes \bigotimes_{\substack{i \in [k] \\ c \in G_i \\ a_c > 0}} \mathbf{x}_c \right\rangle \quad (83)$$

$$= \sum_{\substack{\|\mathbf{a}_V\|_1 \leq d \\ \|\mathbf{a}_G\|_1 > 0}} \left\langle \theta_{\mathbf{a}}, \mathbf{x}_{\overline{G}}^{\otimes \mathbf{a}_{\overline{G}}} \otimes \bigotimes_{\substack{i \in [k] \\ c \in G_i \\ a_c > 0}} \mathbf{R}_c \mathbf{x}_{f_i} \right\rangle \quad (84)$$

$$= \sum_{\substack{\|\mathbf{a}_V\|_1 \leq d \\ \|\mathbf{a}_G\|_1 > 0}} \left\langle \theta_{\mathbf{a}}, \mathbf{x}_{\overline{G}}^{\otimes \mathbf{a}_{\overline{G}}} \otimes \bigotimes_{\substack{i \in [k] \\ c \in G_i \\ a_c > 0}} \left( \star \mathbf{R}_c \right) \mathbf{x}_{f_i} \right\rangle \quad (85)$$

$$= \sum_{\substack{\|\mathbf{a}_V\|_1 \leq d \\ \|\mathbf{a}_G\|_1 > 0}} \left\langle \theta_{\mathbf{a}}, \left( \bigotimes_{\substack{w \in \overline{G} \\ a_w > 0}} \mathbf{I}_w \otimes \bigotimes_{\substack{i \in [k] \\ c \in G_i \\ a_c > 0}} \star \mathbf{R}_c \right) \left( \mathbf{x}_{\overline{G}}^{\otimes \mathbf{a}_{\overline{G}}} \otimes \bigotimes_{\substack{i \in [k] \\ c \in G_i \\ a_c > 0}} \mathbf{x}_{f_i} \right) \right\rangle \quad (86)$$

$$= \sum_{\substack{\|\mathbf{a}_V\|_1 \leq d \\ \|\mathbf{a}_G\|_1 > 0}} \left\langle \left( \bigotimes_{\substack{w \in \overline{G} \\ a_w > 0}} \mathbf{I}_w \otimes \bigotimes_{\substack{i \in [k] \\ c \in G_i \\ a_c > 0}} \star \mathbf{R}_c \right)^{\top} \theta_{\mathbf{a}}, \mathbf{x}_{\overline{G}}^{\otimes \mathbf{a}_{\overline{G}}} \otimes \bigotimes_{\substack{i \in [k] \\ c \in G_i \\ a_c > 0}} \mathbf{x}_{f_i} \right\rangle \quad (87)$$

$$= \sum_{\substack{\|\mathbf{a}_{\overline{G}}\|_1 = q \\ q < d}} \sum_{\substack{T \subseteq [k] \\ 0 < |T| \leq d-q}} \sum_{U \in \mathcal{U}(T, q)} \left\langle \underbrace{\left( \bigotimes_{\substack{w \in \overline{G} \\ a_w > 0}} \mathbf{I}_w \otimes \bigotimes_{i \in T} \bigotimes_{c \in U \cap G_i} \star \mathbf{R}_c \right)^{\top} \theta_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{U|G})}}_{\mathbf{R}_{\mathbf{a}_{\overline{G}}, U} \text{ defined in (46)}}, \mathbf{x}_{\overline{G}}^{\otimes \mathbf{a}_{\overline{G}}} \otimes \bigotimes_{i \in T} \mathbf{x}_{f_i} \right\rangle \quad (88)$$

$$= \sum_{\substack{\|\mathbf{a}_{\overline{G}}\|_1 = q \\ q < d}} \sum_{\substack{T \subseteq [k] \\ 0 < |T| \leq d-q}} \left\langle \underbrace{\sum_{U \in \mathcal{U}(T, q)} \mathbf{R}_{\mathbf{a}_{\overline{G}}, U}^{\top} \theta_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{U|G})}}_{\gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})}}, \mathbf{x}_{\overline{G}}^{\otimes \mathbf{a}_{\overline{G}}} \otimes \bigotimes_{i \in T} \mathbf{x}_{f_i} \right\rangle \quad (89)$$

$$= \sum_{\substack{\|\mathbf{a}_{\overline{G}}\|_1 = q \\ q < d}} \sum_{\substack{T \subseteq [k] \\ 0 < |T| \leq d-q}} \left\langle \gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})}, \mathbf{x}_{\overline{G}}^{\otimes \mathbf{a}_{\overline{G}}} \otimes \bigotimes_{i \in T} \mathbf{x}_{f_i} \right\rangle \quad (90)$$

$$= \sum_{\|\mathbf{b}_{\overline{S}}\|_1 \leq d} \left\langle \gamma_{\mathbf{b}_{\overline{S}}}, \mathbf{x}_{\overline{S}}^{\otimes \mathbf{b}_{\overline{S}}} \right\rangle. \quad (91)$$

Equality (85) follows from (72). Equality (86) follows from (65). Equality at (88) is a bit loaded. What goes on there is that we broke the sum over  $\mathbf{a}_V$  for which  $\|\mathbf{a}_V\|_1 \leq d$  and  $\|\mathbf{a}_G\|_1 > 0$  into a nested triple sum. First of all, in order for  $\|\mathbf{a}_G\|_1 > 0$ , obviously  $\|\mathbf{a}_{\overline{G}}\|_1 < d$  must hold, so we group by those tuples first. The remaining mass  $\|\mathbf{a}_G\|_1$  can only be at most  $d - \|\mathbf{a}_{\overline{G}}\|_1 = d - q$ . Since all features in  $G$  are categorical, from the above analysis we have  $\mathbf{a}_G = (a_g)_{g \in G} \in \{0, 1\}^G$ , i.e.,  $\mathbf{a}_G$  is a characteristic vector of a subset  $U \subseteq G$ . Let  $T = \{i \mid U_i \neq \emptyset\} \subseteq [k]$ .

Then, in the second summation we group  $U$  by  $T$ . The third summation ranges over all choices of  $U \cap G_i, i \in T$ , for which the total mass is at most  $d - q$ . (Recall the definition of  $\mathcal{U}(T, q)$  in (45).)

Next, in (89) we perform the reparameterization. Recall that  $\mathbf{1}_{F_T|F}$  is the characteristic vector of the set  $F_T = \{f_i\}_{i \in T}$  in the collection  $F = \{f_1, \dots, f_k\}$ . The new parameter  $\gamma_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{F_T|F})}$  is indexed by the tuple  $(\mathbf{a}_{\bar{G}}, \mathbf{1}_{F_T|F})$  whose support is  $\bar{G} \cup F = \bar{S}$ , i.e., the set of all features except for the ones functionally determined by features in  $F$ . After the reparameterization, the loss term is identical to the loss term of a  $\text{PR}^d$  model whose features are  $\bar{S}$ . This explains the collapsed pair  $(\bar{g}, \bar{h})$  used in the theorem.

Next, we explore the new parameter and how it affects the penalty term. Consider a fixed pair  $\mathbf{a}_{\bar{G}}$  and  $T \subseteq [k]$  such that  $T \neq \emptyset$  and  $\|\mathbf{a}_{\bar{G}}\|_1 + |T| \leq d$ . The last condition is implicit for the set  $U$  to exist for which  $U \cap G_i \neq \emptyset$  and  $\|\mathbf{a}_{\bar{G}}\|_1 + |U| \leq d$ . Among all choices of  $U$ , we single out  $U = F_T$  and write

$$\gamma_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{F_T|F})} = \sum_{\substack{U \subseteq G \\ U \cap G_i \neq \emptyset, \forall i \in T \\ \|\mathbf{a}_{\bar{G}}\|_1 + |U| \leq d}} \mathbf{R}_{\mathbf{a}_{\bar{G}}, U}^\top \boldsymbol{\theta}_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{U|G})} = \boldsymbol{\theta}_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{F_T|G})} + \sum_{\substack{F_T \neq U \subseteq G \\ U \cap G_i \neq \emptyset, \forall i \in T \\ \|\mathbf{a}_{\bar{G}}\|_1 + |U| \leq d}} \mathbf{R}_{\mathbf{a}_{\bar{G}}, U}^\top \boldsymbol{\theta}_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{U|G})}.$$

Now we are ready to write the penalty term  $\|\boldsymbol{\theta}\|_2^2$  in terms of the new parameter  $\gamma$  and some “left-over” components of  $\boldsymbol{\theta}$ .

$$\begin{aligned} \|\boldsymbol{\theta}\|_2^2 &= \sum_{\|\mathbf{a}_V\|_1 \leq d} \|\boldsymbol{\theta}_{\mathbf{a}}\|_2^2 \\ &= \sum_{\substack{\|\mathbf{a}_V\|_1 \leq d \\ \|\mathbf{a}_G\|_1 = 0}} \|\boldsymbol{\theta}_{\mathbf{a}_V}\|_2^2 + \sum_{\substack{\|\mathbf{a}_V\|_1 \leq d \\ \|\mathbf{a}_G\|_1 > 0}} \|\boldsymbol{\theta}_{\mathbf{a}_V}\|_2^2 \\ &= \sum_{\substack{\|\mathbf{a}_V\|_1 \leq d \\ \|\mathbf{a}_G\|_1 = 0}} \|\boldsymbol{\theta}_{\mathbf{a}_V}\|_2^2 + \sum_{\substack{\|\mathbf{a}_{\bar{G}}\|_1 = q \\ q < d}} \sum_{\substack{T \subseteq [k] \\ 0 < |T| \leq d - q}} \sum_{U \in \mathcal{U}(T, q)} \left\| \boldsymbol{\theta}_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{U|G})} \right\|_2^2 \\ &= \sum_{\substack{\|\mathbf{b}_{\bar{S}}\|_1 \leq d \\ \|\mathbf{b}_F\|_1 = 0}} \left\| \gamma_{\mathbf{b}_{\bar{S}}} \right\|_2^2 + \sum_{\substack{\|\mathbf{a}_{\bar{G}}\|_1 = q \\ q < d}} \sum_{\substack{T \subseteq [k] \\ 0 < |T| \leq d - q}} \left( \left\| \boldsymbol{\theta}_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{F_T|G})} \right\|_2^2 + \sum_{\substack{W \in \mathcal{U}(T, q) \\ W \neq F_T}} \left\| \boldsymbol{\theta}_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{W|G})} \right\|_2^2 \right) \\ &= \sum_{\substack{\|\mathbf{b}_{\bar{S}}\|_1 \leq d \\ \|\mathbf{b}_F\|_1 = 0}} \left\| \gamma_{\mathbf{b}_{\bar{S}}} \right\|_2^2 + \sum_{\substack{\|\mathbf{a}_{\bar{G}}\|_1 = q \\ q < d}} \sum_{\substack{T \subseteq [k] \\ 0 < |T| \leq d - q}} \left\| \gamma_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{F_T|F})} - \sum_{\substack{U \in \mathcal{U}(T, q) \\ U \neq F_T}} \mathbf{R}_{\mathbf{a}_{\bar{G}}, U}^\top \boldsymbol{\theta}_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{U|G})} \right\|_2^2 \\ &\quad + \sum_{\substack{\|\mathbf{a}_{\bar{G}}\|_1 = q \\ q < d}} \sum_{\substack{T \subseteq [k] \\ 0 < |T| \leq d - q}} \sum_{\substack{W \in \mathcal{U}(T, q) \\ W \neq F_T}} \left\| \boldsymbol{\theta}_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{W|G})} \right\|_2^2. \end{aligned}$$

Next, for every  $W \in \mathcal{U}(T, q) - \{F_T\}$ , we optimize out the parameter  $\boldsymbol{\theta}_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{W|G})}$  by noting that the new loss term does not depend on these parameters. To optimize them out, we compute

$$\begin{aligned} \frac{1}{2} \frac{\partial J}{\partial \boldsymbol{\theta}_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{W|G})}} &= \boldsymbol{\theta}_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{W|G})} - \mathbf{R}_{\mathbf{a}_{\bar{G}}, W} \left( \gamma_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{F_T|F})} - \sum_{\substack{U \in \mathcal{U}(T, \|\mathbf{a}_{\bar{G}}\|_1) \\ U \neq F_T}} \mathbf{R}_{\mathbf{a}_{\bar{G}}, U}^\top \boldsymbol{\theta}_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{U|G})} \right) \\ &= \boldsymbol{\theta}_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{W|G})} - \mathbf{R}_{\mathbf{a}_{\bar{G}}, W} \boldsymbol{\theta}_{(\mathbf{a}_{\bar{G}}, \mathbf{1}_{F_T|G})}. \end{aligned}$$

Setting this partial derivative to 0, we obtain  $\theta_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{W|G})} = \mathbf{R}_{\mathbf{a}_{\overline{G}}, W} \theta_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|G})}$ , which leads to

$$\begin{aligned} \theta_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|G})} &= \gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})} - \sum_{\substack{U \in \mathcal{U}(T, \|\mathbf{a}_{\overline{G}}\|_1) \\ U \neq F_T}} \mathbf{R}_{\mathbf{a}_{\overline{G}}, U}^\top \theta_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{U|G})} \\ &= \gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})} - \sum_{\substack{U \in \mathcal{U}(T, \|\mathbf{a}_{\overline{G}}\|_1) \\ U \neq F_T}} \mathbf{R}_{\mathbf{a}_{\overline{G}}, U}^\top \mathbf{R}_{\mathbf{a}_{\overline{G}}, U} \theta_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|G})}. \end{aligned}$$

Moving and grouping, we obtain

$$\left( \bigotimes_{\substack{g \in \overline{G} \\ a_g > 0}} \mathbf{I}_g \otimes \bigotimes_{i \in T} \mathbf{I}_{f_i} + \sum_{\substack{U \in \mathcal{U}(T, \|\mathbf{a}_{\overline{G}}\|_1) \\ U \neq F_T}} \mathbf{R}_{\mathbf{a}_{\overline{G}}, U}^\top \mathbf{R}_{\mathbf{a}_{\overline{G}}, U} \right) \theta_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|G})} = \gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})}.$$

Or, more compactly,

$$\underbrace{\left( \sum_{W \in \mathcal{U}(T, \|\mathbf{a}_{\overline{G}}\|_1)} \mathbf{R}_{\mathbf{a}_{\overline{G}}, W}^\top \mathbf{R}_{\mathbf{a}_{\overline{G}}, W} \right)}_{\mathbf{B}_{\mathbf{a}_{\overline{G}}, T} \text{ as defined in (47)}} \theta_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|G})} = \gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})}. \quad (92)$$

Consequently, we can completely optimize out the remaining  $\theta$ -components, solving for them in terms of the components of  $\gamma$ :

$$\begin{aligned} \theta_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|G})} &= \mathbf{B}_{\mathbf{a}_{\overline{G}}, T}^{-1} \gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})} \\ \theta_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{U|G})} &= \mathbf{R}_{\mathbf{a}_{\overline{G}}, U} \theta_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|G})} \\ &= \mathbf{R}_{\mathbf{a}_{\overline{G}}, U} \mathbf{B}_{\mathbf{a}_{\overline{G}}, T}^{-1} \gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})} \end{aligned}$$

Thus, for a fixed  $T$  and  $\overline{G}$ , we can simplify the total squared normed involved:

$$\begin{aligned} \sum_{U \in \mathcal{U}(T, \|\mathbf{a}_{\overline{G}}\|_1)} \left\| \theta_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{U|G})} \right\|_2^2 &= \sum_{U \in \mathcal{U}(T, \|\mathbf{a}_{\overline{G}}\|_1)} \left\langle \mathbf{R}_{\mathbf{a}_{\overline{G}}, U} \mathbf{B}_{\mathbf{a}_{\overline{G}}, T}^{-1} \gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})}, \mathbf{R}_{\mathbf{a}_{\overline{G}}, U} \mathbf{B}_{\mathbf{a}_{\overline{G}}, T}^{-1} \gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})} \right\rangle \\ &= \sum_{U \in \mathcal{U}(T, \|\mathbf{a}_{\overline{G}}\|_1)} \left\langle \mathbf{R}_{\mathbf{a}_{\overline{G}}, U}^\top \mathbf{R}_{\mathbf{a}_{\overline{G}}, U} \mathbf{B}_{\mathbf{a}_{\overline{G}}, T}^{-1} \gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})}, \mathbf{B}_{\mathbf{a}_{\overline{G}}, T}^{-1} \gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})} \right\rangle \\ &= \left\langle \left( \sum_{U \in \mathcal{U}(T, \|\mathbf{a}_{\overline{G}}\|_1)} \mathbf{R}_{\mathbf{a}_{\overline{G}}, U}^\top \mathbf{R}_{\mathbf{a}_{\overline{G}}, U} \right) \mathbf{B}_{\mathbf{a}_{\overline{G}}, T}^{-1} \gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})}, \mathbf{B}_{\mathbf{a}_{\overline{G}}, T}^{-1} \gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})} \right\rangle \\ &= \left\langle \gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})}, \mathbf{B}_{\mathbf{a}_{\overline{G}}, T}^{-1} \gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})} \right\rangle. \end{aligned}$$

Finally, we write  $\|\theta\|_2^2$  in terms of the new parameter  $\gamma$  to prove (49):

$$\begin{aligned} \|\theta\|_2^2 &= \sum_{\|\mathbf{a}_V\|_1 \leq d} \|\theta_{\mathbf{a}}\|_2^2 \\ &= \sum_{\substack{\|\mathbf{a}_V\|_1 \leq d \\ \|\mathbf{a}_G\|_1 = 0}} \|\theta_{\mathbf{a}_V}\|_2^2 + \sum_{\substack{\|\mathbf{a}_V\|_1 \leq d \\ \|\mathbf{a}_G\|_1 > 0}} \|\theta_{\mathbf{a}_V}\|_2^2 \end{aligned}$$

$$\begin{aligned}
&= \sum_{\substack{\|\mathbf{a}_V\|_1 \leq d \\ \|\mathbf{a}_G\|_1 = 0}} \|\boldsymbol{\theta}_{\mathbf{a}_V}\|_2^2 + \sum_{\substack{\|\mathbf{a}_{\overline{G}}\|_1 = q \\ q < d}} \sum_{\substack{T \subseteq [k] \\ 0 < |T| \leq d-q}} \sum_{U \in \mathcal{U}(T, q)} \left\| \boldsymbol{\theta}_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{U|G})} \right\|_2^2 \\
&= \sum_{\substack{\|\mathbf{b}_{\overline{S}}\|_1 \leq d \\ \|\mathbf{b}_F\|_1 = 0}} \left\| \gamma_{\mathbf{b}_{\overline{S}}} \right\|_2^2 + \sum_{\substack{\|\mathbf{a}_{\overline{G}}\|_1 = q \\ q < d}} \sum_{\substack{T \subseteq [k] \\ 0 < |T| \leq d-q}} \left\langle \mathbf{B}_{\mathbf{a}_{\overline{G}}, T}^{-1} \gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})}, \gamma_{(\mathbf{a}_{\overline{G}}, \mathbf{1}_{F_T|F})} \right\rangle.
\end{aligned}$$

□

### E.3 Alternative to Corollary 4.2

One big advantage of a linear model in terms of BGD is Corollary 4.2, where we do not have to redo point-evaluation for every backtracking step. After the reparameterization exploiting FD-based dimensionality reduction, Corollary 4.2 does not work as is, because we have changed the penalty terms. However, it is easy to work out a similar result in terms of the new parameter space; see The point of the following proposition is that we only need to compute intermediate results involving the covariance matrix  $\overline{\Sigma}$  once while backtracking. For each new value of  $\alpha$ , we will need to recompute the penalty's objective  $\overline{\Omega}(\gamma - \alpha \mathbf{d})$ , which is an inexpensive operation. If  $\lambda = 0$ , we can even solve for  $\alpha$  directly.

**Proposition E.1.** *With respect to the new parameters (and new objective  $\overline{J}$  defined in (48)), the Armijo condition  $\overline{J}(\gamma) - \overline{J}(\gamma - \alpha \mathbf{d}) \leq \frac{\alpha}{2} \|\mathbf{d}\|_2^2$  is equivalent to*

$$\alpha \left( 2\gamma^\top \overline{\Sigma} \mathbf{d} - \alpha \mathbf{d}^\top \overline{\Sigma} \mathbf{d} - 2 \langle \mathbf{d}, \overline{\mathbf{c}} \rangle - \|\mathbf{d}\|_2^2 \right) + \lambda \overline{\Omega}(\gamma) \leq \lambda \overline{\Omega}(\gamma - \alpha \mathbf{d}),$$

where  $\mathbf{d} = \nabla \overline{J}(\gamma)$ . Furthermore, the next gradient of  $\overline{J}$  is also readily available:

$$\frac{\partial \overline{J}(\gamma - \alpha \mathbf{d})}{\partial \gamma} = \mathbf{d} - \alpha \overline{\Sigma} \mathbf{d} + \frac{\lambda}{2} \left( \frac{\partial \Omega(\gamma - \alpha \mathbf{d})}{\partial \gamma} - \frac{\partial \Omega(\gamma)}{\partial \gamma} \right).$$

*Proof.* Let  $\mathbf{d} = \nabla \overline{J}(\gamma)$ . Then,

$$\begin{aligned}
\overline{J}(\gamma) - \overline{J}(\gamma - \alpha \mathbf{d}) &= \frac{1}{2} \gamma^\top \overline{\Sigma} \gamma - \frac{1}{2} (\gamma - \alpha \mathbf{d})^\top \overline{\Sigma} (\gamma - \alpha \mathbf{d}) + \langle \gamma - \alpha \mathbf{d}, \overline{\mathbf{c}} \rangle + \frac{\lambda}{2} (\overline{\Omega}(\gamma) - \overline{\Omega}(\gamma - \alpha \mathbf{d})) \\
&= \alpha \gamma^\top \overline{\Sigma} \mathbf{d} - \frac{\alpha^2}{2} \mathbf{d}^\top \overline{\Sigma} \mathbf{d} - \alpha \langle \mathbf{d}, \overline{\mathbf{c}} \rangle + \frac{\lambda}{2} (\overline{\Omega}(\gamma) - \overline{\Omega}(\gamma - \alpha \mathbf{d})).
\end{aligned}$$

□

### E.4 Specializing Theorem 5.1 to the LR model

This section specializes Theorem 5.1 to the LR-model. Let us first specialize expressions (45), (46). and (47). We start with (45). Since  $d = 1$ , the only valid choice of  $q$  is 0, and  $|T| = 1$ . If  $T = \{j\}$ , then  $U \in \mathcal{U}(T, q)$  iff  $U = \{c\}$  for some  $c \in G_j$ . In other words, we can replace  $\mathcal{U}(T, q)$  by  $G_j$  itself. Next, consider (46): there is only one valid choice of  $\mathbf{a}_{\overline{G}}$  – the all 0 vector – and  $U = \{c\}$  for some  $c \in G_j$ , the matrix  $\mathbf{R}_{\mathbf{a}_{\overline{G}}, U}$  is *exactly*  $\mathbf{R}_c$ . Lastly, when  $T = \{j\}$  the sum (47) becomes  $\sum_{c \in G_j} \mathbf{R}_c^\top \mathbf{R}_c$ . We have the following corollary:

**Corollary E.2.** *Consider a LR model with parameters  $\boldsymbol{\theta} = (\boldsymbol{\theta}_w)_{w \in V}$  and  $k$  groups of simple FDs  $G_i = \{f_i\} \cup S_i$ ,  $i \in [k]$ . Define the following reparameterization:*

$$\gamma_w = \begin{cases} \boldsymbol{\theta}_w & w \in V - G, \\ \sum_{c \in G_i} \mathbf{R}_c^\top \boldsymbol{\theta}_c & w \in F. \end{cases}$$

Then, minimizing  $J(\boldsymbol{\theta})$  is equivalent to minimizing the function  $\bar{J}(\boldsymbol{\gamma}) = \frac{1}{2}\boldsymbol{\gamma}^\top \bar{\boldsymbol{\Sigma}}\boldsymbol{\gamma} - \langle \boldsymbol{\gamma}, \bar{\mathbf{c}} \rangle + \frac{\lambda}{2}\Omega(\boldsymbol{\gamma})$ , where  $\Omega(\boldsymbol{\gamma}) = \sum_{w \in V \setminus G} \|\boldsymbol{\gamma}_w\|_2^2 + \sum_{i=1}^k \langle \mathbf{B}_i^{-1}\boldsymbol{\gamma}_{f_i}, \boldsymbol{\gamma}_{f_i} \rangle$ , and matrix  $\mathbf{B}_i$  for each  $i \in [k]$  is given by

$$\mathbf{B}_i = \sum_{c \in G_i} \mathbf{R}_c^\top \mathbf{R}_c. \quad (93)$$

$\bar{J}$  is defined with respect to the FD-reduced pair of functions  $\bar{g}, \bar{h}$  and a reduced parameter space of  $\boldsymbol{\gamma}$ . Its gradient is very simple to compute, where we specialize (50):

$$\frac{1}{2} \frac{\partial \Omega(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_w} = \begin{cases} \boldsymbol{\gamma}_w & w \in V - G, \\ \mathbf{B}_i^{-1}\boldsymbol{\gamma}_{f_i} & w \in F. \end{cases} \quad (94)$$

Moreover, once a minimizer  $\boldsymbol{\gamma}$  of  $\bar{J}$  is obtained, following (51), we can compute a minimizer  $\boldsymbol{\theta}$  of  $J$  by setting

$$\boldsymbol{\theta}_w = \begin{cases} \boldsymbol{\gamma}_w & w \in V \setminus G, \\ \mathbf{R}_w \mathbf{B}_i^{-1} \boldsymbol{\gamma}_{f_i} & w \in G_i, i \in [k]. \end{cases}$$

## E.5 Specializing Theorem 5.1 to the PR<sup>2</sup> model

In this section we explore Theorem 5.1 for the special case of degree-2 polynomial regression. This case is significant for three reasons. First, due to the explosion in the number of parameters, in practice one rarely runs polynomial regression of degree higher than 2. In fact, PR<sup>2</sup> may be a sufficiently rich nonlinear regression model for many real-world applications. Second, this is technically already a highly non-trivial application of our general theorem. Third, this case shares some commonality with FaMa<sub>r</sub><sup>2</sup> model to be described in the next section.

As before, we first specialize expressions (45), (46), and (47). To do so, we slightly change the indexing scheme of the model to simplify the presentation. In the general model, we use tuples  $\mathbf{a}$  with  $\|\mathbf{a}\|_1 \leq d$  to index parameters. When the model is of degree 2, we explicitly write down the two types of indices: we use  $\boldsymbol{\theta}_w$ ,  $w \in V$  instead of  $\boldsymbol{\theta}_{\mathbf{a}}$  with  $\|\mathbf{a}\|_1 = 1$ , and we use  $\boldsymbol{\theta}_{cw}$  with  $c, w \in V$  instead of  $\boldsymbol{\theta}_{\mathbf{a}}$  when  $\|\mathbf{a}\|_1 = 2$ .

We start with (45). Since  $d = 2$ , two valid choices of  $q$  are 0 and 1.

- when  $q = 1$ ,  $|T| = \{i\}$  for some  $i \in [k]$ . The set  $\mathcal{U}(\{i\}, 1)$  is the collection of singleton subsets of  $G_i$ . Hence, this is similar to the linear regression situation.
- when  $q = 0$ ,  $|T|$  is either  $\{i\}$  or  $\{i, j\}$ . The set  $\mathcal{U}(\{i, j\}, 0)$  consists of all 2-subsets  $U$  of  $G$  for which  $U$  contains one element from  $G_i$  and one from  $G_j$ . The set  $\mathcal{U}(\{i\}, 0)$  contains all singletons and 2-subsets of  $G_i$ .

Next, consider (46): there are two valid choices for the pair  $(\mathbf{a}_{\bar{G}}, U)$ :

- when  $\|\mathbf{a}_{\bar{G}}\|_1 = 0$ ,  $U \in \mathcal{U}(\{i, j\}, 0)$  or  $U \in \mathcal{U}(\{i\}, 0)$ . In that case, we have

$$\begin{aligned} \mathbf{R}_{\emptyset, \{c, t\}} &= \mathbf{R}_c \otimes \mathbf{R}_t & (c, t) &\in G_i \times G_j \\ \mathbf{R}_{\emptyset, \{c\}} &= \mathbf{R}_c & c &\in G_i \\ \mathbf{R}_{\emptyset, \{c, t\}} &= \mathbf{R}_c \star \mathbf{R}_t & \{c, t\} &\in \binom{G_i}{2}. \end{aligned}$$

- when  $\|\mathbf{a}_{\bar{G}}\|_1 = 1$ ,  $U \in \mathcal{U}(\{i\}, 1)$  for some  $i \in [k]$ ; and in this case we use  $w \in \bar{G}$  to represent  $\mathbf{a}_{\bar{G}}$ :

$$\mathbf{R}_{w, \{c\}} = \mathbf{I}_w \otimes \mathbf{R}_c.$$

Finally, we write down (47) explicitly; the valid indices for  $\mathbf{B}$  are  $\mathbf{B}_{\emptyset, \{i\}}$ ,  $\mathbf{B}_{\emptyset, \{i, j\}}$ , and  $\mathbf{B}_{w, \{i\}}$  (also recall the definition of  $\mathbf{B}_i$  in (93)):

$$\mathbf{B}_{\emptyset, \{i\}} = \sum_{c \in G_i} \mathbf{R}_c^\top \mathbf{R}_c + \sum_{\{c, t\} \in \binom{G_i}{2}} (\mathbf{R}_c \star \mathbf{R}_t)^\top (\mathbf{R}_c \star \mathbf{R}_t) \quad (95)$$

$$\begin{aligned} \mathbf{B}_{\emptyset, \{i, j\}} &= \sum_{c \in G_i, t \in G_j} (\mathbf{R}_c \otimes \mathbf{R}_t)^\top (\mathbf{R}_c \otimes \mathbf{R}_t) \\ &= \sum_{c \in G_i, t \in G_j} (\mathbf{R}_c^\top \mathbf{R}_c \otimes \mathbf{R}_t^\top \mathbf{R}_t) \\ &= \mathbf{B}_i \otimes \mathbf{B}_j \end{aligned} \quad (96)$$

$$\mathbf{B}_{w, \{i\}} = \sum_{c \in G_i} (\mathbf{I}_w \otimes \mathbf{R}_c)^\top (\mathbf{I}_w \otimes \mathbf{R}_c) = \mathbf{I}_w \otimes \mathbf{B}_i \quad (97)$$

**Corollary E.3.** Consider the  $\text{PR}^2$  model with  $k$  groups of simple FDs  $G_i = \{f_i\} \cup S_i$ ,  $i \in [k]$ . Let

$$\boldsymbol{\theta} = ((\boldsymbol{\theta}_w)_{w \in V}, (\boldsymbol{\theta}_{cw})_{c, w \in v})$$

be the original parameters, and  $G = \cup_{i \in [k]} G_i$ . Define the following reparameterization:

$$\gamma_w = \begin{cases} \boldsymbol{\theta}_w & w \in V \setminus G \\ \sum_{c \in G_i} \mathbf{R}_c^\top \boldsymbol{\theta}_c + \sum_{\{c, t\} \in \binom{G_i}{2}} (\mathbf{R}_c \star \mathbf{R}_t)^\top \boldsymbol{\theta}_{ct} & w = f_i, i \in [k]. \end{cases} \quad (98)$$

$$\gamma_{tw} = \begin{cases} \boldsymbol{\theta}_{tw}, & \{t, w\} \subseteq V \setminus G \\ \sum_{c \in G_i} (\mathbf{I}_w \otimes \mathbf{R}_c^\top) \boldsymbol{\theta}_{wc} & t = f_i, w \notin G \\ \sum_{(c, c') \in G_i \times G_j} (\mathbf{R}_c \star \mathbf{R}_{c'})^\top \boldsymbol{\theta}_{cc'}, & \{t, w\} = \{f_i, f_j\}, \{i, j\} \in \binom{[k]}{2}. \end{cases} \quad (99)$$

Then, minimizing  $J(\boldsymbol{\theta})$  is equivalent to minimizing the function  $\bar{J}(\boldsymbol{\gamma}) = \frac{1}{2} \boldsymbol{\gamma}^\top \bar{\boldsymbol{\Sigma}} \boldsymbol{\gamma} - \langle \boldsymbol{\gamma}, \bar{\mathbf{c}} \rangle + \frac{\lambda}{2} \Omega(\boldsymbol{\gamma})$ , where

$$\begin{aligned} \Omega(\boldsymbol{\gamma}) &= \sum_{w \notin G} \|\gamma_w\|_2^2 + \sum_{\substack{c \notin G \\ t \notin G}} \|\gamma_{ct}\|_2^2 + \sum_{i=1}^k \left\langle \mathbf{B}_{\emptyset, \{i\}}^{-1} \gamma_{f_i}, \gamma_{f_i} \right\rangle \\ &+ \sum_{\substack{i \in [k] \\ w \notin G}} \left\langle (\mathbf{I}_w \otimes \mathbf{B}_i^{-1}) \gamma_{wf_i}, \gamma_{wf_i} \right\rangle + \sum_{ij \in \binom{[k]}{2}} \left\langle \mathbf{B}_i^{-1} \otimes \mathbf{B}_j^{-1} \gamma_{f_i f_j}, \gamma_{f_i f_j} \right\rangle. \end{aligned}$$

The gradient of  $\bar{J}$  is very simple to compute, by noticing that  $\bar{J}$  is defined with respect to the FD-reduced pair of functions  $\bar{g}, \bar{h}$  and a reduced parameter space of  $\boldsymbol{\gamma}$ . Its gradient can be computed by specializing (50):

$$\frac{1}{2} \frac{\partial \Omega(\boldsymbol{\gamma})}{\partial \gamma_w} = \begin{cases} \gamma_w & w \notin G \\ \mathbf{B}_{\emptyset, \{i\}}^{-1} \gamma_{f_i} & w = f_i \end{cases} \quad (100)$$

$$\frac{1}{2} \frac{\partial \Omega(\boldsymbol{\gamma})}{\partial \gamma_{tw}} = \begin{cases} \gamma_{tw} & \{t, w\} \cap \{f_i\}_{i=1}^k = \emptyset \\ (\mathbf{I}_w \otimes \mathbf{B}_i^{-1}) \gamma_{wf_i} & t = f_i, w \notin G \\ (\mathbf{B}_i^{-1} \otimes \mathbf{B}_j^{-1}) \gamma_{f_i f_j} & \{t, w\} = \{f_i, f_j\}. \end{cases} \quad (101)$$



Moreover, once a minimizer  $\gamma$  of  $\bar{J}$  is obtained, following (51), we can compute a minimizer  $\theta$  of  $J$  by setting

$$\begin{aligned}\theta_w &= \begin{cases} \gamma_w & w \in V \setminus G \\ \mathbf{R}_w \mathbf{B}_{\emptyset, \{i\}}^{-1} \gamma_{f_i}, & w \in G_i, i \in [k] \end{cases} \\ \theta_{ct} &= (\mathbf{R}_c \star \mathbf{R}_t) \mathbf{B}_{\emptyset, \{i\}}^{-1} \gamma_{f_i}, \forall \{c, t\} \in \binom{G_i}{2} \\ \theta_{cw} &= \begin{cases} \gamma_{cw}, & w \in V \setminus G \\ (\mathbf{I}_w \otimes \mathbf{R}_c \mathbf{B}_i^{-1}) \gamma_{wf_i}, & c \in G_i, w \notin G, i \in [k] \end{cases} \\ \theta_{ct} &= (\mathbf{R}_c \mathbf{B}_i^{-1} \otimes \mathbf{R}_t \mathbf{B}_j^{-1}) \gamma_{f_i f_j}, (c, t) \in G_i \times G_j.\end{aligned}$$

## E.6 Proofs of results in Section 5.4

*Proof of Theorem 5.2.* We begin with a similar derivation, where “relevant terms” of  $\langle g(\theta), h(\mathbf{x}) \rangle$  are the terms where  $h$  contains a feature  $c \in F_i$  for some  $i \in [k]$ :

$$\begin{aligned}& \text{relevant terms of } \langle g(\theta), h(\mathbf{x}) \rangle \\&= \sum_{\substack{c \in F_i \\ i \in [k]}} \langle \theta_c, \mathbf{x}_c \rangle + \sum_{\substack{\{c, t\} \in \binom{F_i}{2} \\ i \in [k] \\ \ell \in [r]}} \langle \theta_c^{(\ell)} \otimes \theta_t^{(\ell)}, \mathbf{x}_c \otimes \mathbf{x}_t \rangle + \sum_{ij \in \binom{[k]}{2}} \sum_{\substack{c \in F_i \\ t \in F_j \\ \ell \in [r]}} \langle \theta_c^{(\ell)} \otimes \theta_t^{(\ell)}, \mathbf{x}_c \otimes \mathbf{x}_t \rangle \\&+ \sum_{\substack{c \in F \\ w \notin F \\ \ell \in [r]}} \langle \theta_c^{(\ell)} \otimes \theta_w^{(\ell)}, \mathbf{x}_c \otimes \mathbf{x}_w \rangle \\&= \sum_{\substack{c \in F_i \\ i \in [k]}} \langle \theta_c, \mathbf{R}_c \mathbf{x}_{f_i} \rangle + \sum_{\substack{\{c, t\} \in \binom{F_i}{2} \\ i \in [k] \\ \ell \in [r]}} \langle \theta_c^{(\ell)} \otimes \theta_t^{(\ell)}, \mathbf{R}_c \mathbf{x}_{f_i} \otimes \mathbf{R}_t \mathbf{x}_{f_i} \rangle \\&+ \sum_{ij \in \binom{[k]}{2}} \sum_{\substack{c \in F_i \\ t \in F_j \\ \ell \in [r]}} \langle \theta_c^{(\ell)} \otimes \theta_t^{(\ell)}, \mathbf{R}_c \mathbf{x}_{f_i} \otimes \mathbf{R}_t \mathbf{x}_{f_j} \rangle + \sum_{\substack{i \in [k] \\ c \in F_i \\ w \notin F \\ \ell \in [r]}} \langle \theta_c^{(\ell)} \otimes \theta_w^{(\ell)}, \mathbf{R}_c \mathbf{x}_{f_i} \otimes \mathbf{x}_w \rangle \\&= \sum_{\substack{c \in F_i \\ i \in [k]}} \langle \mathbf{R}_c^\top \theta_c, \mathbf{x}_{f_i} \rangle + \sum_{\substack{\{c, t\} \in \binom{F_i}{2} \\ i \in [k] \\ \ell \in [r]}} \langle \mathbf{R}_c^\top \theta_c^{(\ell)} \otimes \mathbf{R}_t^\top \theta_t^{(\ell)}, \mathbf{x}_{f_i} \otimes \mathbf{x}_{f_i} \rangle \\&+ \sum_{ij \in \binom{[k]}{2}} \sum_{\substack{c \in F_i \\ t \in F_j \\ \ell \in [r]}} \langle \mathbf{R}_c^\top \theta_c^{(\ell)} \otimes \mathbf{R}_t^\top \theta_t^{(\ell)}, \mathbf{x}_{f_i} \otimes \mathbf{x}_{f_j} \rangle + \sum_{\substack{i \in [k] \\ c \in F_i \\ w \notin F \\ \ell \in [r]}} \langle \mathbf{R}_c^\top \theta_c^{(\ell)} \otimes \theta_w^{(\ell)}, \mathbf{x}_{f_i} \otimes \mathbf{x}_w \rangle \\&= \sum_{\substack{c \in F_i \\ i \in [k]}} \langle \mathbf{R}_c^\top \theta_c, \mathbf{x}_{f_i} \rangle + \sum_{\substack{\{c, t\} \in \binom{F_i}{2} \\ i \in [k] \\ \ell \in [r]}} \langle \mathbf{R}_c^\top \theta_c^{(\ell)} \circ \mathbf{R}_t^\top \theta_t^{(\ell)}, \mathbf{x}_{f_i} \rangle \\&+ \sum_{ij \in \binom{[k]}{2}} \left\langle \sum_{c \in F_i} \mathbf{R}_c^\top \theta_c^{(\ell)} \otimes \sum_{t \in F_j} \mathbf{R}_t^\top \theta_t^{(\ell)}, \mathbf{x}_{f_i} \otimes \mathbf{x}_{f_j} \right\rangle + \sum_{\substack{i \in [k] \\ c \in F_i \\ w \notin F \\ \ell \in [r]}} \left\langle \sum_{c \in F_i} \mathbf{R}_c^\top \theta_c^{(\ell)} \otimes \theta_w^{(\ell)}, \mathbf{x}_{f_i} \otimes \mathbf{x}_w \right\rangle\end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^k \left\langle \underbrace{\sum_{c \in F_i} \mathbf{R}_c^\top \boldsymbol{\theta}_c + \sum_{\ell=1}^r \sum_{\{c,t\} \in \binom{F_i}{2}} \mathbf{R}_c^\top \boldsymbol{\theta}_c^{(\ell)} \circ \mathbf{R}_t^\top \boldsymbol{\theta}_t^{(\ell)}}_{\boldsymbol{\gamma}_{f_i}}, \mathbf{x}_{f_i} \right\rangle \\
&+ \sum_{\substack{ij \in \binom{[k]}{2} \\ \ell \in [r]}} \left\langle \underbrace{\sum_{c \in F_i} \mathbf{R}_c^\top \boldsymbol{\theta}_c^{(\ell)}}_{\boldsymbol{\gamma}_{f_i}^{(\ell)}} \otimes \underbrace{\sum_{t \in F_j} \mathbf{R}_t^\top \boldsymbol{\theta}_t^{(\ell)}}_{\boldsymbol{\gamma}_{f_j}^{(\ell)}}, \mathbf{x}_{f_i} \otimes \mathbf{x}_{f_j} \right\rangle + \sum_{\substack{i \in [k] \\ w \notin F \\ \ell \in [r]}} \left\langle \underbrace{\sum_{c \in F_i} \mathbf{R}_c^\top \boldsymbol{\theta}_c^{(\ell)}}_{\boldsymbol{\gamma}_{f_i}^{(\ell)}} \otimes \boldsymbol{\theta}_w^{(\ell)}, \mathbf{x}_{f_i} \otimes \mathbf{x}_w \right\rangle \\
&= \sum_{i=1}^k \langle \boldsymbol{\gamma}_{f_i}, \mathbf{x}_{f_i} \rangle + \sum_{\substack{ij \in \binom{[k]}{2} \\ \ell \in [r]}} \langle \boldsymbol{\gamma}_{f_i}^{(\ell)} \otimes \boldsymbol{\gamma}_{f_j}^{(\ell)}, \mathbf{x}_{f_i} \otimes \mathbf{x}_{f_j} \rangle + \sum_{\substack{i \in [k] \\ w \notin F \\ \ell \in [r]}} \langle \boldsymbol{\gamma}_{f_i}^{(\ell)} \otimes \boldsymbol{\theta}_w^{(\ell)}, \mathbf{x}_{f_i} \otimes \mathbf{x}_w \rangle.
\end{aligned}$$

The above derivation immediately yields the reparameterization given in the statement of the theorem, which we reproduce here for the sake of clarity:

$$\begin{aligned}
\boldsymbol{\gamma}_w &= \begin{cases} \boldsymbol{\theta}_w & w \notin F \\ \boldsymbol{\theta}_{f_i} + \sum_{c \in S_i} \mathbf{R}_c^\top \boldsymbol{\theta}_c + \boldsymbol{\beta}_{f_i} & w = f_i, i \in [k]. \end{cases} \\
\boldsymbol{\gamma}_w^{(\ell)} &= \begin{cases} \boldsymbol{\theta}_w^{(\ell)} & w \notin \{f_1, \dots, f_k\} \\ \boldsymbol{\theta}_{f_i}^{(\ell)} + \sum_{c \in S_i} \mathbf{R}_c^\top \boldsymbol{\theta}_c^{(\ell)} & w = f_i, i \in [k]. \end{cases}
\end{aligned}$$

Note that we did not define  $\boldsymbol{\gamma}_w$  for  $w \in S_i, i \in [k]$ . The reason we can do so, is because we can optimize out  $\boldsymbol{\theta}_c$  due to the following trick we have been using (as in the proof of Theorem 5.1). First, we rewrite all the terms in  $\|\boldsymbol{\theta}\|_2^2$  in terms of  $\boldsymbol{\gamma}$  and  $\boldsymbol{\theta}_c, c \in S_i, i \in [k]$ :

$$\begin{aligned}
\|\boldsymbol{\theta}\|_2^2 &= \sum_{w \notin F} \|\boldsymbol{\theta}_w\|_2^2 + \sum_{i=1}^k \sum_{t \in F_i} \|\boldsymbol{\theta}_t\|_2^2 + \sum_{\ell=1}^r \sum_{w \notin \{f_1, \dots, f_k\}} \|\boldsymbol{\theta}_w^{(\ell)}\|_2^2 + \sum_{\ell=1}^r \sum_{i=1}^k \|\boldsymbol{\theta}_{f_i}^{(\ell)}\|_2^2 \\
&= \sum_{w \notin F} \|\boldsymbol{\gamma}_w\|_2^2 + \sum_{i=1}^k \sum_{t \in F_i} \|\boldsymbol{\theta}_t\|_2^2 + \sum_{\ell=1}^r \sum_{w \notin \{f_1, \dots, f_k\}} \|\boldsymbol{\gamma}_w^{(\ell)}\|_2^2 + \sum_{\ell=1}^r \sum_{i=1}^k \|\boldsymbol{\theta}_{f_i}^{(\ell)}\|_2^2 \\
&= \sum_{w \notin F} \|\boldsymbol{\gamma}_w\|_2^2 + \sum_{i=1}^k \left\| \boldsymbol{\gamma}_{f_i} - \sum_{c \in S_i} \mathbf{R}_c^\top \boldsymbol{\theta}_c - \boldsymbol{\beta}_{f_i} \right\|_2^2 + \sum_{i=1}^k \sum_{t \in S_i} \|\boldsymbol{\theta}_t\|_2^2 + \sum_{\ell=1}^r \sum_{w \notin \{f_1, \dots, f_k\}} \|\boldsymbol{\gamma}_w^{(\ell)}\|_2^2 \\
&\quad + \sum_{\ell=1}^r \sum_{i=1}^k \left\| \boldsymbol{\gamma}_{f_i}^{(\ell)} - \sum_{c \in S_i} \mathbf{R}_c^\top \boldsymbol{\gamma}_c^{(\ell)} \right\|_2^2
\end{aligned}$$

Since  $\boldsymbol{\theta}_t, t \in S_i$ , does not depend on the loss term, we have

$$\frac{1}{2} \frac{\partial J}{\partial \boldsymbol{\theta}_t} = \boldsymbol{\theta}_t - \mathbf{R}_t \left( \underbrace{\boldsymbol{\gamma}_{f_i} - \sum_{c \in S_i} \mathbf{R}_c^\top \boldsymbol{\theta}_c - \boldsymbol{\beta}_{f_i}}_{\boldsymbol{\theta}_{f_i}} \right) \quad w \in S_i, i \in [k]. \quad (102)$$

By setting (102) to 0, we have  $\boldsymbol{\theta}_t = \mathbf{R}_t \boldsymbol{\theta}_{f_i}$  for all  $t \in F_i$ , and thus

$$\boldsymbol{\theta}_{f_i} = \boldsymbol{\gamma}_{f_i} - \sum_{c \in S_i} \mathbf{R}_c^\top \boldsymbol{\theta}_c - \boldsymbol{\beta}_{f_i} = \boldsymbol{\gamma}_{f_i} - \sum_{c \in S_i} \mathbf{R}_c^\top \mathbf{R}_c \boldsymbol{\theta}_{f_i} - \boldsymbol{\beta}_{f_i},$$

which implies  $\boldsymbol{\theta}_{f_i} = \mathbf{B}_i^{-1}(\boldsymbol{\gamma}_{f_i} - \boldsymbol{\beta}_{f_i})$ . Hence, the following always holds:

$$\boldsymbol{\theta}_t = \mathbf{R}_t \mathbf{B}_i^{-1}(\boldsymbol{\gamma}_{f_i} - \boldsymbol{\beta}_{f_i}), \quad \forall t \in F_i, i \in [k].$$

Note also that,

$$\begin{aligned} \sum_{t \in F_i} \|\boldsymbol{\theta}_t\|_2^2 &= \sum_{t \in F_i} \|\mathbf{R}_t \mathbf{B}_i^{-1}(\boldsymbol{\gamma}_{f_i} - \boldsymbol{\beta}_{f_i})\|_2^2 \\ &= \sum_{t \in F_i} \langle \mathbf{R}_t^\top \mathbf{R}_t \mathbf{B}_i^{-1}(\boldsymbol{\gamma}_{f_i} - \boldsymbol{\beta}_{f_i}), \mathbf{B}_i^{-1}(\boldsymbol{\gamma}_{f_i} - \boldsymbol{\beta}_{f_i}) \rangle \\ &= \left\langle \left( \sum_{t \in F_i} \mathbf{R}_t^\top \mathbf{R}_t \right) \mathbf{B}_i^{-1}(\boldsymbol{\gamma}_{f_i} - \boldsymbol{\beta}_{f_i}), \mathbf{B}_i^{-1}(\boldsymbol{\gamma}_{f_i} - \boldsymbol{\beta}_{f_i}) \right\rangle \\ &= \langle \mathbf{B}_i \mathbf{B}_i^{-1}(\boldsymbol{\gamma}_{f_i} - \boldsymbol{\beta}_{f_i}), \mathbf{B}_i^{-1}(\boldsymbol{\gamma}_{f_i} - \boldsymbol{\beta}_{f_i}) \rangle \\ &= \langle (\boldsymbol{\gamma}_{f_i} - \boldsymbol{\beta}_{f_i}), \mathbf{B}_i^{-1}(\boldsymbol{\gamma}_{f_i} - \boldsymbol{\beta}_{f_i}) \rangle. \end{aligned}$$

Due to the fact that  $\boldsymbol{\theta}_{f_i}^{(\ell)} = \boldsymbol{\gamma}_{f_i}^{(\ell)} - \sum_{c \in S_i} \mathbf{R}_c^\top \boldsymbol{\gamma}_c^{(\ell)}$ , we can now write the penalty term in terms of the new parameter  $\boldsymbol{\gamma}$ :

$$\begin{aligned} \|\boldsymbol{\theta}\|_2^2 &= \sum_{w \notin F} \|\boldsymbol{\gamma}_w\|_2^2 + \sum_{i=1}^k \sum_{t \in F_i} \|\boldsymbol{\theta}_t\|_2^2 + \sum_{\ell=1}^r \sum_{w \notin \{f_1, \dots, f_k\}} \|\boldsymbol{\gamma}_w^{(\ell)}\|_2^2 + \sum_{\ell=1}^r \sum_{i=1}^k \|\boldsymbol{\theta}_{f_i}^{(\ell)}\|_2^2 \\ &= \sum_{w \notin F} \|\boldsymbol{\gamma}_w\|_2^2 + \sum_{i=1}^k \langle (\boldsymbol{\gamma}_{f_i} - \boldsymbol{\beta}_{f_i}), \mathbf{B}_i^{-1}(\boldsymbol{\gamma}_{f_i} - \boldsymbol{\beta}_{f_i}) \rangle + \sum_{\ell=1}^r \sum_{w \notin \{f_1, \dots, f_k\}} \|\boldsymbol{\gamma}_w^{(\ell)}\|_2^2 \\ &\quad + \sum_{\ell=1}^r \sum_{i=1}^k \left\| \boldsymbol{\gamma}_{f_i}^{(\ell)} - \sum_{c \in S_i} \mathbf{R}_c^\top \boldsymbol{\gamma}_c^{(\ell)} \right\|_2^2. \end{aligned}$$

□

*Proof of Proposition 5.3.* The goal is to derive the gradient of  $\Omega(\boldsymbol{\gamma})$  w.r.t the parameters  $\boldsymbol{\gamma}$ . Since  $\boldsymbol{\beta}_{f_i}$  is a function of  $\boldsymbol{\gamma}_c^{(\ell)}$ ,  $\ell \in [r]$ ,  $c \in F_i$ , the following is immediate:

$$\begin{aligned} \frac{1}{2} \frac{\partial \|\boldsymbol{\theta}\|_2^2}{\partial \boldsymbol{\gamma}_w} &= \begin{cases} \boldsymbol{\gamma}_w, & w \notin F \\ \mathbf{B}_i^{-1}(\boldsymbol{\gamma}_{f_i} - \boldsymbol{\beta}_{f_i}) & w = f_i, i \in [k]. \end{cases} \\ \frac{1}{2} \frac{\partial \|\boldsymbol{\theta}\|_2^2}{\partial \boldsymbol{\gamma}_w^{(\ell)}} &= \boldsymbol{\gamma}_w, \quad w \notin F, \ell \in [r]. \end{aligned}$$

Next, we have to simplify  $\boldsymbol{\beta}_{f_i}$  to facilitate fast computation:

$$\begin{aligned} \boldsymbol{\beta}_{f_i} &= \sum_{\ell=1}^r \sum_{\{c,t\} \in \binom{F_i}{2}} \mathbf{R}_c^\top \boldsymbol{\theta}_c^{(\ell)} \circ \mathbf{R}_t^\top \boldsymbol{\theta}_t^{(\ell)} \\ &= \sum_{\ell=1}^r \left[ \mathbf{R}_{f_i}^\top \boldsymbol{\theta}_{f_i}^{(\ell)} \circ \sum_{c \in S_i} \mathbf{R}_c^\top \boldsymbol{\theta}_c^{(\ell)} + \sum_{\{c,t\} \in \binom{S_i}{2}} \mathbf{R}_c^\top \boldsymbol{\theta}_c^{(\ell)} \circ \mathbf{R}_t^\top \boldsymbol{\theta}_t^{(\ell)} \right] \\ &= \sum_{\ell=1}^r \left[ \boldsymbol{\theta}_{f_i}^{(\ell)} \circ \sum_{c \in S_i} \mathbf{R}_c^\top \boldsymbol{\theta}_c^{(\ell)} + \sum_{\{c,t\} \in \binom{S_i}{2}} \mathbf{R}_c^\top \boldsymbol{\theta}_c^{(\ell)} \circ \mathbf{R}_t^\top \boldsymbol{\theta}_t^{(\ell)} \right] \end{aligned}$$

$$\begin{aligned}
&= \sum_{\ell=1}^r \left[ \left( \gamma_{f_i}^{(\ell)} - \sum_{t \in S_i} \mathbf{R}_t^\top \theta_t^{(\ell)} \right) \circ \sum_{c \in S_i} \mathbf{R}_c^\top \theta_c^{(\ell)} + \sum_{\{c,t\} \in \binom{S_i}{2}} \mathbf{R}_c^\top \theta_c^{(\ell)} \circ \mathbf{R}_t^\top \theta_t^{(\ell)} \right] \\
&= \sum_{\ell=1}^r \left[ \gamma_{f_i}^{(\ell)} \circ \sum_{c \in S_i} \mathbf{R}_c^\top \theta_c^{(\ell)} - \sum_{t \in S_i} \sum_{c \in S_i} \mathbf{R}_t^\top \theta_t^{(\ell)} \circ \mathbf{R}_c^\top \theta_c^{(\ell)} + \sum_{\{c,t\} \in \binom{S_i}{2}} \mathbf{R}_c^\top \theta_c^{(\ell)} \circ \mathbf{R}_t^\top \theta_t^{(\ell)} \right] \\
&= \sum_{\ell=1}^r \left[ \gamma_{f_i}^{(\ell)} \circ \sum_{c \in S_i} \mathbf{R}_c^\top \theta_c^{(\ell)} - \sum_{t \in S_i} \mathbf{R}_t^\top \theta_t^{(\ell)} \circ \mathbf{R}_t^\top \theta_t^{(\ell)} - \sum_{\{c,t\} \in \binom{S_i}{2}} \mathbf{R}_c^\top \theta_c^{(\ell)} \circ \mathbf{R}_t^\top \theta_t^{(\ell)} \right] \\
&= \sum_{\ell=1}^r \left[ \gamma_{f_i}^{(\ell)} \circ \sum_{c \in S_i} \mathbf{R}_c^\top \theta_c^{(\ell)} - \sum_{t \in S_i} \mathbf{R}_t^\top (\theta_t^{(\ell)} \circ \theta_t^{(\ell)}) - \sum_{\{c,t\} \in \binom{S_i}{2}} \mathbf{R}_c^\top \theta_c^{(\ell)} \circ \mathbf{R}_t^\top \theta_t^{(\ell)} \right] \\
&= \sum_{\ell=1}^r \left[ \gamma_{f_i}^{(\ell)} \circ \sum_{c \in S_i} \mathbf{R}_c^\top \gamma_c^{(\ell)} - \sum_{t \in S_i} \mathbf{R}_t^\top (\gamma_t^{(\ell)} \circ \gamma_t^{(\ell)}) - \sum_{\{c,t\} \in \binom{S_i}{2}} \mathbf{R}_c^\top \gamma_c^{(\ell)} \circ \mathbf{R}_t^\top \gamma_t^{(\ell)} \right].
\end{aligned}$$

Next, we derive the partial derivative w.r.t.  $\gamma_{f_i}^{(\ell)}$  for a fixed  $i \in [k]$ ,  $\ell \in [r]$ ; in this computation we make use of (5) above:

$$\begin{aligned}
\frac{1}{2} \frac{\partial \|\theta\|_2^2}{\partial \gamma_{f_i}^{(\ell)}} &= \frac{1}{2} \frac{\partial \langle (\gamma_{f_i} - \beta_{f_i}), \mathbf{B}_i^{-1}(\gamma_{f_i} - \beta_{f_i}) \rangle}{\partial \gamma_{f_i}^{(\ell)}} + \frac{1}{2} \frac{\partial \left\| \gamma_{f_i}^{(\ell)} - \sum_{c \in S_i} \mathbf{R}_c^\top \gamma_c^{(\ell)} \right\|_2^2}{\partial \gamma_{f_i}^{(\ell)}} \\
&= \left( \sum_{c \in S_i} \text{DIAG}(\mathbf{R}_c^\top \gamma_c^{(\ell)}) \right) \mathbf{B}_i^{-1}(\beta_{f_i} - \gamma_{f_i}) + \gamma_{f_i}^{(\ell)} - \sum_{c \in S_i} \mathbf{R}_c^\top \gamma_c^{(\ell)} \\
&= \underbrace{\gamma_{f_i}^{(\ell)} - \sum_{c \in S_i} \mathbf{R}_c^\top \gamma_c^{(\ell)}}_{\delta_i^{(\ell)}} - \underbrace{\left( \sum_{c \in S_i} \mathbf{R}_c^\top \gamma_c^{(\ell)} \right) \circ \mathbf{B}_i^{-1}(\gamma_{f_i} - \beta_{f_i})}_{\delta_i^{(\ell)}} \\
&= \gamma_{f_i}^{(\ell)} - \delta_i^{(\ell)} - \delta_i^{(\ell)} \circ \left( \frac{1}{2} \frac{\partial \|\theta\|_2^2}{\partial \gamma_{f_i}^{(\ell)}} \right)
\end{aligned}$$

Lastly, we move on to the partial derivative w.r.t.  $\gamma_w^{(\ell)}$  for a fixed  $i \in [k]$ ,  $w \in S_i$ ,  $\ell \in [r]$ :

$$\begin{aligned}
\frac{1}{2} \frac{\partial \|\theta\|_2^2}{\partial \gamma_w^{(\ell)}} &= \frac{1}{2} \frac{\partial \left\| \gamma_w^{(\ell)} \right\|_2^2}{\partial \gamma_w^{(\ell)}} + \frac{1}{2} \frac{\partial \langle (\gamma_{f_i} - \beta_{f_i}), \mathbf{B}_i^{-1}(\gamma_{f_i} - \beta_{f_i}) \rangle}{\partial \gamma_w^{(\ell)}} + \frac{1}{2} \frac{\partial \left\| \gamma_{f_i}^{(\ell)} - \sum_{c \in S_i} \mathbf{R}_c^\top \gamma_c^{(\ell)} \right\|_2^2}{\partial \gamma_w^{(\ell)}} \\
&= \gamma_w^{(\ell)} + \mathbf{R}_w \left( \sum_{c \in F_i} \text{DIAG}(\mathbf{R}_c^\top \gamma_c^{(\ell)}) \right) \mathbf{B}_i^{-1}(\beta_{f_i} - \gamma_{f_i}) + \mathbf{R}_w \left( \sum_{c \in S_i} \mathbf{R}_c^\top \gamma_c^{(\ell)} - \gamma_{f_i}^{(\ell)} \right) \\
&= \gamma_w^{(\ell)} + \mathbf{R}_w \left( \gamma_{f_i}^{(\ell)} + \delta_i^{(\ell)} \right) \circ \left( \frac{1}{2} \frac{\partial \|\theta\|_2^2}{\partial \gamma_{f_i}^{(\ell)}} \right) + \mathbf{R}_w \left( \delta_i^{(\ell)} - \gamma_{f_i}^{(\ell)} \right) \\
&= \gamma_w^{(\ell)} + \mathbf{R}_w \left[ \gamma_{f_i}^{(\ell)} \circ \left( \frac{1}{2} \frac{\partial \|\theta\|_2^2}{\partial \gamma_{f_i}^{(\ell)}} \right) + \left( \delta_i^{(\ell)} \circ \left( \frac{1}{2} \frac{\partial \|\theta\|_2^2}{\partial \gamma_{f_i}^{(\ell)}} \right) + \delta_i^{(\ell)} - \gamma_{f_i}^{(\ell)} \right) \right] \\
&= \gamma_w^{(\ell)} + \mathbf{R}_w \left[ \gamma_{f_i}^{(\ell)} \circ \left( \frac{1}{2} \frac{\partial \|\theta\|_2^2}{\partial \gamma_{f_i}^{(\ell)}} \right) - \left( \frac{1}{2} \frac{\partial \|\theta\|_2^2}{\partial \gamma_{f_i}^{(\ell)}} \right) \right].
\end{aligned}$$

In particular, we were able to reuse the computation of  $\frac{1}{2} \frac{\partial \|\boldsymbol{\theta}\|_2^2}{\partial \boldsymbol{\gamma}_{f_i}^{(\ell)}}$  and  $\frac{1}{2} \frac{\partial \|\boldsymbol{\theta}\|_2^2}{\partial \boldsymbol{\gamma}_{f_i}^{(\ell)}}$  to compute  $\frac{1}{2} \frac{\partial \|\boldsymbol{\theta}\|_2^2}{\partial \boldsymbol{\gamma}_w^{(\ell)}}$ . There is, however, still one complicated term  $\beta_{f_i}$  left to compute. We simplify  $\beta_{f_i}$  to make its evaluation faster as follows.

$$\begin{aligned}
\beta_{f_i} &= \sum_{\ell=1}^r \left[ \gamma_{f_i}^{(\ell)} \circ \sum_{c \in S_i} \mathbf{R}_c^\top \gamma_c^{(\ell)} - \sum_{t \in S_i} \mathbf{R}_t^\top (\gamma_t^{(\ell)} \circ \gamma_t^{(\ell)}) - \sum_{\{c,t\} \in \binom{S_i}{2}} \mathbf{R}_c^\top \gamma_c^{(\ell)} \circ \mathbf{R}_t^\top \gamma_t^{(\ell)} \right] \\
&= \sum_{\ell=1}^r \left[ \gamma_{f_i}^{(\ell)} \circ \sum_{c \in S_i} \mathbf{R}_c^\top \gamma_c^{(\ell)} - \frac{1}{2} \sum_{t \in S_i} \mathbf{R}_t^\top (\gamma_t^{(\ell)} \circ \gamma_t^{(\ell)}) - \frac{1}{2} \sum_{c \in S_i} \sum_{t \in S_i} \mathbf{R}_c^\top \gamma_c^{(\ell)} \circ \mathbf{R}_t^\top \gamma_t^{(\ell)} \right] \\
&= \sum_{\ell=1}^r \left[ \gamma_{f_i}^{(\ell)} \circ \underbrace{\sum_{c \in S_i} \mathbf{R}_c^\top \gamma_c^{(\ell)}}_{\boldsymbol{\delta}_i^{(\ell)}} - \frac{1}{2} \sum_{t \in S_i} \mathbf{R}_t^\top (\gamma_t^{(\ell)} \circ \gamma_t^{(\ell)}) - \frac{1}{2} \sum_{c \in S_i} \mathbf{R}_c^\top \gamma_c^{(\ell)} \circ \sum_{t \in S_i} \mathbf{R}_t^\top \gamma_t^{(\ell)} \right] \\
&= \sum_{\ell=1}^r \left[ \gamma_{f_i}^{(\ell)} \circ \boldsymbol{\delta}_i^{(\ell)} - \frac{1}{2} \sum_{t \in S_i} \mathbf{R}_t^\top (\gamma_t^{(\ell)} \circ \gamma_t^{(\ell)}) - \frac{1}{2} \boldsymbol{\delta}_i^{(\ell)} \circ \boldsymbol{\delta}_i^{(\ell)} \right] \\
&= \sum_{\ell=1}^r \left[ \left( \gamma_{f_i}^{(\ell)} - \frac{1}{2} \boldsymbol{\delta}_i^{(\ell)} \right) \circ \boldsymbol{\delta}_i^{(\ell)} - \frac{1}{2} \sum_{t \in S_i} \mathbf{R}_t^\top (\gamma_t^{(\ell)} \circ \gamma_t^{(\ell)}) \right].
\end{aligned}$$

This completes the proof. □