The Data Minimization Principle in Machine Learning

Anonymous Authors[1]

The principle of data minimization requires organizations to *collect, process, and retain only personal data that is adequate, relevant, and limited to what is necessary for specified objectives*. It's grounded in the expectation that not all collected data is essential for the objective and, instead, contributes to a heightened risk of information leakage (Goldsteen et al., 2021; Paul et al., 2021; Sorscher et al., 2022; Shanmugam et al., 2022). However, despite its legal significance and endorsement by global data protection regulations, the data minimization principle lacks a mathematical formalization suitable for real-world ML applications. In particular, the current discourse on data minimization practices often overlooks two crucial aspects: **(1)** the individualized nature of minimization (e.g., information that is unimportant for an individual may be critical for another) and **(2)** its intrinsic link to data privacy.

To overcome these limitations, this paper introduces a formal framework for data minimization in ML while being faithful to its legal notion (which we review in Table 1); adapts and evaluates various optimization algorithms to solve the problem of data minimization; and analyzes their compatibility with real-world privacy. In particular, we seek to answer a critical question: *"Do data minimization requirements in various regulations meet privacy expectations in legal frameworks?"* Our evaluations reveal that the answer is, unfortunately, no. While being an implicit intention, the requirements of data minimization are not necessarily aligned with risk of reconstruction and re-identification and thus may not provide the expected privacy protection.

With this work, we aim to lay down a path for future research for developing ML systems that comply with the legal requirements of data minimization.

Problem settings. Consider a dataset D consisting of n datapoints $(\boldsymbol{x}_i, y_i)$, where $i \in [n]$, each drawn i.i.d. from an unknown distribution. Therein, $\boldsymbol{x}_i \in \mathcal{X}$ is a p-dimensional feature vector and $y_i \in \mathcal{Y}$ is the corresponding output label. As an illustrative example, consider a loan approval task (Ding et al., 2021). Here, $\boldsymbol{x}_i$ could describe an individual's age, income, race, and job, while y_i whether they will repay a loan. The objective is to train a predictor $f_\theta : \mathcal{X} \to \mathcal{Y}$, parametrized by $\theta \in \mathbb{R}^d$, to minimize the empirical risk:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \; J(\theta; \mathbf{X}, \mathbf{Y}) = 1/n \sum_{i=1}^{n} \ell(f_\theta(\boldsymbol{x}_i), y_i),$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a non-negative loss function that evaluates model quality, and $\mathbf{X}$ and $\mathbf{Y}$ represent the matrix of all features and the vector of all labels in D, respectively.

Mapping Regulatory Laws to Formal Principles

We start with a key question: "*How to translate regulatory laws into formal principles?*"

To tackle this challenge, we start by inspecting various global data protection regulations, whose language is summarized in Table 1, which enables us to discern three foundational pillars of the data minimization principle:

1. **Purpose Limitation**: Data should only be collected for a specific, legally justified purpose. In the context of machine learning, this aligns with the goal of achieving a specified task performance via model training, or retaining high performance at inference time.
2. **Data Relevance**: Regulations mandate that collected data be relevant and limited to what is necessary for the stated purpose. In ML tasks, this means striving to minimize data without affecting performance.
3. **Data Privacy**: Data protection laws define personal data as information that can identify an individual. This places an onus on data minimization to prevent any unnecessary usage of such identifiable data.

A formal framework for data minimization should encapsulate these regulatory aspects. Accordingly, the goal of *data minimization* is to reduce the size of D by selectively removing components from the feature vectors $\boldsymbol{x}_i$ (data relevance). This is achieved while also maintaining performance levels comparable to those achieved using the complete dataset (purpose). The privacy goal, in this interpretation, corresponds to retaining only the necessary data. This objective can be stated as a bi-level optimization:

$$\underset{\boldsymbol{B} \in \{\perp, 1\}^{n \times p}}{\text{Minimize}} \quad \|\boldsymbol{B}\|_1 \tag{1a}$$

$$s.t. : J(\hat{\theta}; \mathbf{X}, \mathbf{Y}) - J(\hat{\theta}; \mathbf{X}, \mathbf{Y}) \leq \alpha \tag{1b}$$

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

General Data Protection Regulation (GDPR), Europe	gdpr-info.eu/
Article 4(1): *"personal data" means any information relating to an identifiable natural person ("data subject") [. . .];*	
Article 5(1)(b): *Personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes;*	
Article 5(1)(c): *Personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.*	
California Privacy Rights Act (CPRA), USA	cppa.ca.gov/
Section 1798.100 (a)(1) & (a)(2): *[. . .] A business shall not collect additional categories of (sensitive) personal information or use (sensitive) personal information collected for additional purposes that are incompatible with the disclosed purpose which the (sensitive) personal information was collected without providing the consumer with notice [. . .].*	
Section 1798.140 (v)(1): *"Personal information" means information that identifies, relates to, describes, is reasonably capable of being associated with, or could reasonably be linked with a particular consumer or household.*	
Protection of Personal Information Act (POPIA), South Africa	popia.co.za/
Section 1: *"personal information" means information relating to an identifiable, living, natural person, [. . .]*	
Section 10: *Personal information may only be processed if, given the purpose for which it is processed, it is adequate, relevant and not excessive.*	
Section 13(1): *Personal information must be collected for a specific and lawful purpose related to a function or activity of the responsible party.*	

Table 1: Excerpts from various data regulation laws on the principle of data minimization, highlighting language on purpose limitation, data relevance, and references to the expectations of data privacy.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(f_\theta(\boldsymbol{x}_i \odot \boldsymbol{B}_i), y_i\right) \tag{1c}$$

Therein, $\boldsymbol{B}$ is an $n \times p$ binary matrix, which we call the *minimization matrix*, taking values in the set $\{\perp, 1\}$, and the ℓ_1-norm of $\boldsymbol{B}$, i.e., $\|\boldsymbol{B}\|_1$, is simply the sum of 1s in the minimization matrix. Here, the symbol $\perp$ represents the concealment or removal of redundant values in the dataset, i.e., $\forall a \in \mathbb{R} : a \times \perp = \perp$, and $\alpha \geq 0$ is an input parameter which thresholds the permitted drop in model quality due to data minimization. The minimized input feature set $\mathbf{X}'$ is defined as the element-wise product of the original feature set $\mathbf{X}$ and the minimization matrix $\boldsymbol{B}$, i.e., $\mathbf{X}' = \mathbf{X} \odot \boldsymbol{B}$.

The optimization problem above defines an operational method to remove entries from the feature set $\mathbf{X}$ in a *personalized* manner (1a), while adhering to pre-specified accuracy requirements on the original dataset (1b), for the final model trained on the minimized dataset (1c). While this formulation captures the original goals expressed in the legal formulation of data minimization, it is however intractable to solve in practice.

We thus implement a gradient-based approach for approximating a batched bi-level program (Eq. (1)). Figure 1 compares three different data modalities—tabular (a), image (b), and text (c)—and demonstrates how this approach (red curves) outperforms baselines based on feature selection[1] or (personalized) subsampling (blue curves) for various degrees of sparsity (x-axis).

In the figure, various threshold checks indicated by α levels provide insight into the associated drop in accuracy and can be used to meet specified performance constraints, as articulated in Eq. (1b). These results highlight a strength of data minimization at its core: even when the dataset size is trimmed down to just 15% of its original volume, the model's utility remains largely unaffected across most data modalities, experiencing less than a 2% dip in performance.

[1] Feature selection is generally applicable in text corpus, thus Fig 1(right) does not report it.

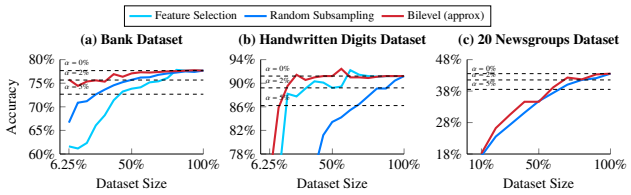

Fig. 1: Task utility across various algorithms, datasets, modalities, and changing the amount of the minimized data (x-axis).

Does Data Minimization meet Privacy Expectations?

While we saw the effectiveness of data minimization with a primary emphasis on utility, there is an expectation of privacy associated with it. The central question of this task is: *"Do current data minimization requirements in various regulations genuinely meet privacy expectations?"* In short, we show that minimization algorithms based purely on translating the legislative language into actionable algorithms fall short in providing acceptable privacy levels. Our analysis (not reported due to space constraints) focuses on assessing reidentification risks, probing the privacy limitations intrinsic to data minimization. These observation necessitate an deep understanding of the underlying causes. This aspect requires developing both a formal analysis and devise mitigation strategies e.g., integrating more formal privacy guarantees into the data minimization process.

This study marks a step in aligning the legal requirements with practical, technical solutions for data minimization in ML. We hope it could set the stage for future work aimed at developing comprehensive, efficient, and ethically sound methodologies for minimization.

References

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Goldsteen, A., Ezov, G., Shmelkin, R., Moffie, M., and Farkash, A. Data minimization for gdpr compliance in machine learning models. *AI and Ethics*, pp. 1–15, 2021.

Paul, M., Ganguli, S., and Dziugaite, G. K. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34: 20596–20607, 2021.

Shanmugam, D., Diaz, F., Shabanian, S., Finck, M., and Biega, A. Learning to limit data collection via scaling laws: A computational interpretation for the legal principle of data minimization. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 839–849, 2022.

Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., and Morcos, A. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.