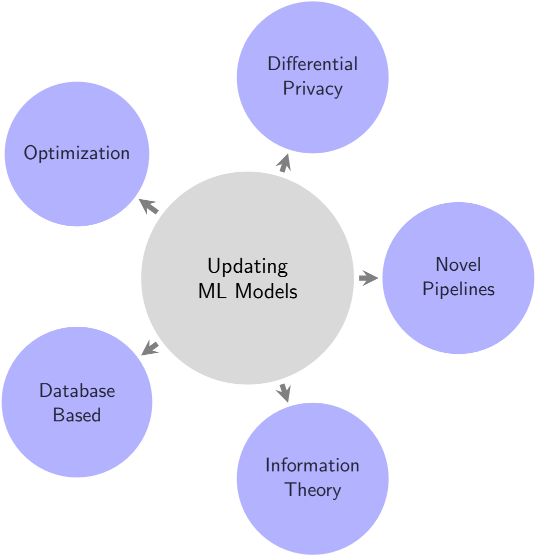
Updating ML Models

Ananth Mahadevan

November 7, 2020

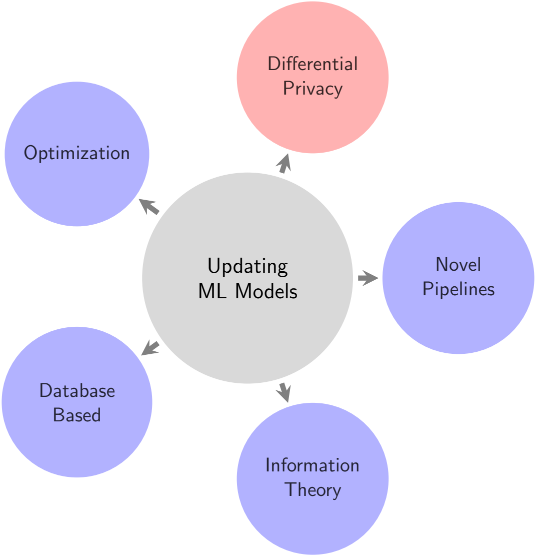
Overview

1. Motivation
2. Problem Overview
3. Approaches
   - Differential Privacy
   - Optimization
   - Database Based
   - Information Theory
   - Novel Pipelines
4. Next Directions


Differential Privacy
Optimization
Updating ML Models
Novel Pipelines
Database Based
Information Theory

Common Terminology

- Fixed training Dataset $\mathcal{D}$
- Learning Algorithm A (can be randomized)
- Datapoints to be remove $\mathcal{D}_{\mathcal{R}}$, where $|\mathcal{D}_{\mathcal{R}}| = r$, remaining dataset $\mathcal{D}' = \mathcal{D} - \mathcal{D}_{\mathcal{R}}$
- Naive approach is retraining from scratch, i.e, $A(\mathcal{D}')$
- Mechanism M which offers an efficient way to update the model


Differential Privacy
Optimization
Updating ML Models
Novel Pipelines
Database Based
Information Theory

Certified Data Removal [Guo et al., 2020]

- A outputs a model in hypothesis space $\mathcal{H}$
- Defines ϵ-*certified removal*, $\forall \mathcal{T} \subseteq \mathcal{H}$

$$e^{-\epsilon} \leq \frac{P(M(A(\mathcal{D}), \mathcal{D}_{\mathcal{R}}) \in \mathcal{T})}{P(A(\mathcal{D}') \in \mathcal{T})} \leq e^{\epsilon}$$

- Insufficiency of Parametric indistinguishability
  - Approximate removal processes leaves a gradient residual
  - Residuals can reveal the prior presence of that training sample

Removal Mechanism for Linear Classfiers

- A empirical risk $L(\mathbf{w}; \mathcal{D})$ with a convex loss function $\ell(\mathbf{w}^T\mathbf{x}, y)$
- $\mathbf{w}^* = A(\mathcal{D}) = \text{argmin}_w L(\mathbf{w}; \mathcal{D})$
- To remove a single point $\mathcal{D}_{\mathcal{R}} = \{(\mathbf{x}_n, y_n)\}$
- Newton Update Step: $\mathbf{w}^- = M(\mathbf{w}^*, (\mathbf{x}_n, y_n)) = \mathbf{w}^* - H_{\mathbf{w}^*}^{-1}\nabla$
- Where $H_{\mathbf{w}^*} = \nabla^2 L(\mathbf{w}^*, \mathcal{D}')$ and $\nabla = \lambda\mathbf{w}^* + \nabla\ell((\mathbf{w}^*)^T\mathbf{x}_n, y_n)$
- $H_{\mathbf{w}^*}^{-1}\nabla$ is from *influence function* literature

Influence Function

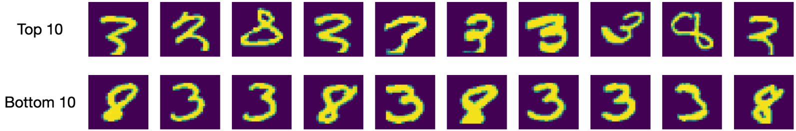


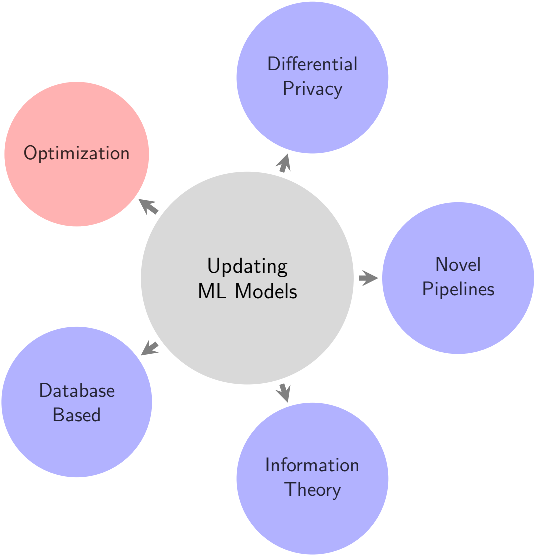
Figure 3. **MNIST training digits sorted by norm of the removal update $\|\mathbf{H}_{\mathbf{w}^*}^{-1}\boldsymbol{\Delta}\|_2$.** The samples with the highest norm (**top**) appear to be atypical, making it harder to undo their effect on the model. The samples with the lowest norm (**bottom**) are prototypical 3s and 8s, and hence are much easier to remove.

Certifing Removal

- $\mathbf{w}^-$ is approximate close to minimizer of $L(\mathbf{w}; \mathcal{D}')$
- $\nabla L(\mathbf{w}^-; \mathcal{D}')$ is gradient residual and if non-zero, reveals Information
- Even a small $\|\nabla L(\mathbf{w}^-; \mathcal{D}')\|_2$ doesn't guarantee certifiable removal
- Therefore, perturb loss at training time

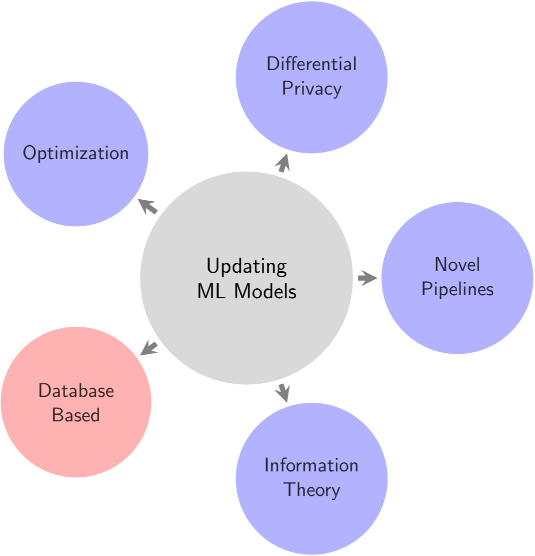
$$L_b(\mathbf{w}; \mathcal{D}) = \sum_{i=1}^{n} \ell(\mathbf{w}^T \mathbf{x}_i, y_i) + \frac{\lambda n}{2} \|\mathbf{w}\|_2^2 + \mathbf{b}^T \mathbf{w}$$

Where $\mathbf{b} \in \mathbb{R}^d$ drawn randomly from some distribution


Differential Privacy
Optimization
Updating ML Models
Novel Pipelines
Database Based
Information Theory

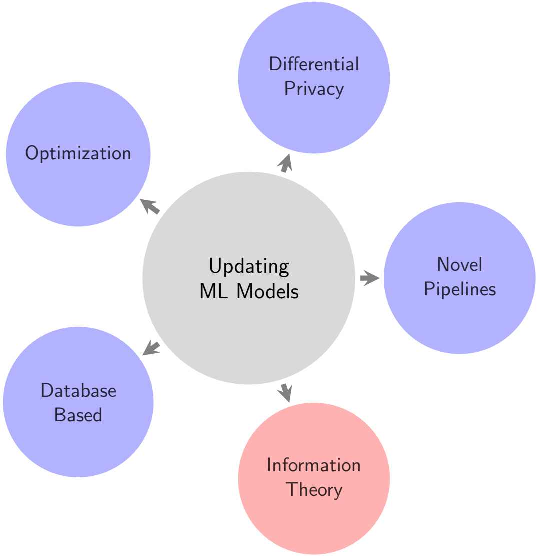
DeltaGrad [Wu et al., 2020a]

-


Differential Privacy
Optimization
Updating ML Models
Novel Pipelines
Database Based
Information Theory

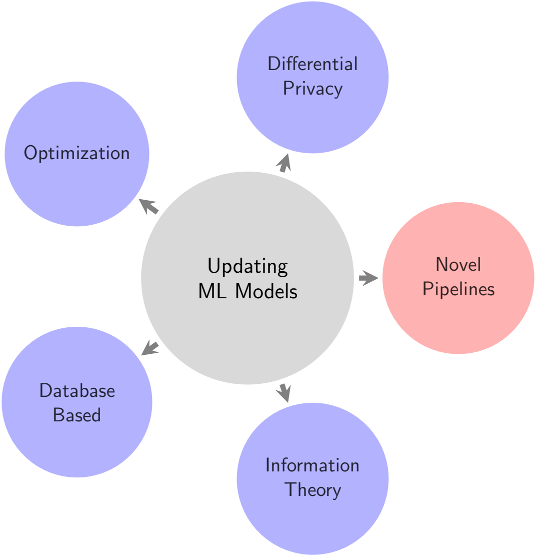
PrIU [Wu et al., 2020b]

-


Differential Privacy
Optimization
Updating ML Models
Novel Pipelines
Database Based
Information Theory

Eternal Sunshine of the Spotless Net [Golatkar et al., 2020]

-


Differential Privacy
Optimization
Updating ML Models
Novel Pipelines
Database Based
Information Theory

Machine Unlearning: SISA [Bourtoule et al., 2020]

-

References I

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. (2020).
Machine Unlearning.
arXiv:1912.03817 [cs].

Golatkar, A., Achille, A., and Soatto, S. (2020).
Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks.
arXiv:1911.04933 [cs, stat].

References II

Guo, C., Goldstein, T., Hannun, A., and van der Maaten, L. (2020).
Certified Data Removal from Machine Learning Models.
arXiv:1911.03030 [cs, stat].

Wu, Y., Dobriban, E., and Davidson, S. B. (2020a).
DeltaGrad: Rapid retraining of machine learning models.
arXiv:2006.14755 [cs, stat].

References III

Wu, Y., Tannen, V., and Davidson, S. B. (2020b).
PrIU: A Provenance-Based Approach for Incrementally Updating Regression Models.
In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 447–462, Portland OR USA. ACM.