

---

# Provable Unlearning in Topic Modeling and Downstream Tasks

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

Machine unlearning algorithms are increasingly important as legal concerns arise around the provenance of training data, but verifying the success of unlearning is often difficult. Provable guarantees for unlearning are often limited to supervised learning settings. In this paper, we provide the first theoretical guarantees for unlearning in the pre-training and fine-tuning paradigm by studying topic models, simple bag-of-words language models that can be adapted to solve downstream tasks like retrieval and classification. First, we design a provably effective unlearning algorithm for topic models that incurs a computational overhead independent of the size of the original dataset. Our analysis additionally quantifies the deletion capacity of the model – *i.e.*, the number of examples that can be unlearned without incurring a significant cost in model performance. Finally, we formally extend our analyses to account for adaptation to a given downstream task. In particular, we design an efficient algorithm to perform unlearning after fine-tuning the topic model via a linear head. Notably, we show that it is easier to unlearn pre-training data from models that have been fine-tuned to a particular task, and one can unlearn this data without modifying the base model.

## 1 Introduction

Modern-day machine learning has shifted from single-stage supervised learning on manually constructed datasets to a paradigm in which models are pre-trained and subsequently fine-tuned (Bommasani et al., 2022). In this setting, a model initially learns a good representation of the data using a self-supervised objective on a large unstructured corpus. The resulting pre-trained model is later adapted to solve specific tasks for which it is difficult or costly to curate a large dataset. This blueprint has yielded strong performance in text (*e.g.*, Devlin et al., 2019; Brown et al., 2020), vision (*e.g.*, Oquab et al., 2024; He et al., 2022), and multimodal (*e.g.*, Radford et al., 2021; Zhai et al., 2023) settings. It is well-known that the scale of the pre-training data is strongly correlated with the final performance of the model (Hoffmann et al., 2022), leading to the construction of larger datasets via broad internet scrapes (Gao et al., 2020; Schuhmann et al., 2022; Soldaini et al., 2024; Penedo et al., 2023). Such datasets have been found to often inadvertently include private, sensitive, and unsafe data (Birhane et al., 2021; Longpre et al., 2024; He et al., 2024).

Unsafe data can generally degrade model performance and introduce biases, making the model less useful for various applications (McKenna et al., 2023; Birhane & Prabhu, 2021; Choenni et al., 2021; Naous et al., 2024). Using private and sensitive data, even unknowingly, poses legal risks (Bommasani et al., 2022; Henderson et al., 2023). In particular, recent works have shown that models can memorize and thus permit the extraction of training data (Somepalli et al., 2023; Carlini et al., 2021, 2023). Moreover, one may be requested to remove data in accordance with GDPR’s *right to be forgotten* (European Parliament & Council of the European Union), or as part of a copyright-related lawsuit (*Tremblay v. OpenAI, Inc.*, 2023; *DOE 1 v. GitHub, Inc.*, N.D. Cal. 2022).

Therefore, there is great empirical interest in developing machine unlearning algorithms that can surgically remove portions of the training data from an already learned model without harming performance. The gold standard for machine unlearning is for the model to behave as though it had never been trained on that datapoint (Cao & Yang, 2015). As it is often undesirable to completely retrain models, especially as they grow larger, many works have proposed computationally cheaper heuristics for solving this problem (e.g., Jang et al., 2023; Foster et al., 2024; Kurmanji et al., 2023; Zhang et al., 2024b; Eldan & Russinovich, 2023; Gandikota et al., 2023). In the absence of theoretical guarantees, it is common to use empirics to measure the success of these algorithms. However, recent works have shown that such evaluations often overestimate the success of these unlearning methods (Hayes et al., 2024; Shi et al., 2024; Maini et al., 2024) and thus it has proven difficult to confidently ascertain whether the proposed methods meet the necessary compliance standards. In this context, it is highly desirable to design efficient unlearning algorithms with well-motivated guarantees that are salient to the pre-training and finetuning paradigm (Thudi et al., 2022; Lee et al., 2024).

While there are some instances of such algorithms for linear models (Guo et al., 2020; Izzo et al., 2021; Mahadevan & Mathioudakis, 2023), general convex models (Ullah et al., 2021; Sekhari et al., 2021; Neel et al., 2021), Bayesian models (Nguyen et al., 2020), and GANs (Liu et al., 2024), there are no works on the paradigm of pre-training and fine-tuning algorithms. One of the most classical such algorithms is topic modeling (Hofmann et al., 1999; Blei et al., 2003; Blei & Lafferty, 2006; Li & McCallum, 2006), which can also be thought of as the simplest language model. *In this paper, we present the first provably effective and efficient unlearning algorithms for topic models.*

Topic models are generally pre-trained to extract latent structure (i.e., a small set of underlying topics) from a large corpus of documents. This feature extractor is then used for a variety of downstream applications, including retrieval, classification, and recommendation (Boyd-Graber et al., 2017). Despite their simplicity, topic models can be used to effectively solve many real-world natural language problems — see a survey in Churchill & Singh (2022).

## 1.1 Overview of Results

We focus on the setting in Arora et al. (2012b), because it admits an efficient learning algorithm with provable guarantees (Arora et al., 2012a). The corpus is assumed to contain  $r$  underlying topics, where each topic defines a distribution over words. Let  $\mathcal{D}$  be a distribution over topic distributions. Then, each document  $d$  is generated by sampling a topic distribution  $W_d \sim \mathcal{D}$  over topics, and then sampling words according to  $W_d$ . The dataset of  $m$  documents is a matrix  $M \in \mathbb{R}^{n \times m}$ , where  $M$  permits a non-negative matrix factorization  $M = A^* X$ . Here,  $A^* \in \mathbb{R}^{n \times r}$  is the distribution of words in each of the  $r$  unknown underlying topics, and  $X \in \mathbb{R}^{r \times m}$  is the sampled distribution of topics in each document. In particular,  $A^*, X$  have columns on the probability simplex. We seek to learn the embedding function  $A^*$  and the topic-topic covariance  $R^* = \mathbb{E}_{\mathcal{D}}[XX^\top]$ .

To derive provable guarantees on the success of unlearning, we adapt the notion of  $(\epsilon, \delta)$ -unlearning introduced in Sekhari et al. (2021) to the topic modeling setting. The unlearned model is required to behave indistinguishably from a model that was retrained on the modified dataset. We define a notion of *utility-preserving unlearning* that combines this condition with an analysis on the *deletion capacity* — i.e., the number of datapoints that can be unlearned without performance degradation (Definition 4). We now state our main result on utility-preserving unlearning in topic models.

**Main Result 1** (Informal version of Theorem 2). Suppose we trained a topic model  $A^S, X^S$  on a training set  $S$  containing  $m$  documents. Algorithm 1 can perform utility-preserving unlearning of

$$m_U = \tilde{O}\left(\frac{m}{r^2 \sqrt{nr}}\right)$$

documents from the pre-trained topic model, where  $\tilde{O}(\cdot)$  hides constants depending on the learning and unlearning algorithm.

To adapt a topic model to a downstream topic classification task, we learn a head  $w \in \mathbb{R}^r$  on top of  $A$  to minimize a strongly convex loss function (Definition 2). When  $A$  and  $w$  are both released, one would necessarily have to first unlearn from  $A$ , which makes unlearning just as hard as it was in pre-training (Theorem 3). This setting is rather unrealistic, because there is no obvious case in which one would want to use  $w$  without  $A$  or vice versa. We thus advocate for viewing fine-tuned

model  $B = Aw$  as a whole i.e. it is not allowed to access outputs of  $A$  solely, and we show that it is easier to perform utility-preserving unlearning of pre-training data in this case.

**Main Result 2** (Informal version of Theorem 4). After adapting the model to a downstream task (Definitions 1 and 2), Algorithm 2 can perform utility-preserving unlearning of  $\tilde{\Omega}\left(\frac{mq}{r\sqrt{nr}}\right)$  documents, where  $q \in [1/r, 1]$  is a task-dependent quantity, without modifying the base model  $A$ . Simpler downstream tasks have a larger  $q$ , increasing the separation from the pre-training result.

We demonstrate that our unlearning algorithms run substantially faster than retraining the model (Table 1). Overall, our results imply the following takeaways in the context of topic models. (1) It is possible to effectively and efficiently unlearn datapoints from a pre-trained model without retraining it (Algorithm 1 and Theorem 2). (2) One can effectively unlearn more pre-training data from a model that has been adapted to a downstream task without harming the utility of the base and fine-tuned models (Theorem 4). (3) One can unlearn pre-training data from a fine-tuned model without modifying the base model (Algorithm 2 and Theorem 4).

## 2 Topic Models

As we previously discussed, topic models can be considered as one of the simplest language models that one can pre-train in a self-supervised fashion and later fine-tune for other language-related tasks. This pipeline mirrors the modern-day paradigm of pre-training large language models to build a general understanding of natural language and later fine-tuning them to solve a variety of tasks ranging from classification to code generation.

### 2.1 Problem Description

Topic modeling is a classical, bag-of-words method to discover structure in a corpus of documents (Hofmann et al., 1999). One assumes that each document contains a convex combination of topics, each of which can be described in terms of a distribution over the vocabulary. Different assumptions on the structure of this distribution and the topics have yielded a variety of topic modeling methodologies (Blei & Lafferty, 2006; Li & McCallum, 2006) – perhaps most famous among these is the latent Dirichlet allocation (LDA, Blei et al. (2003)). Many early works established the statistical learnability of topic models under such assumptions, but the learning algorithms generally were not efficient in real-world settings (Arora et al., 2012b; Recht et al., 2012).

Our paper focuses on the setting in Arora et al. (2012b), for which Arora et al. (2012a) provided an empirically efficient learning algorithm. The dataset consists of a set of  $m$  documents  $d_1, \dots, d_m$ , where each document contains  $L$  words from a vocabulary  $\mathcal{V}$  with  $|\mathcal{V}| = n$ .<sup>1</sup> The corpus contains  $r$  different underlying topics, each of which defines a distribution over words. Each word in document  $d$  is generated by: (1) sampling a distribution over topics  $W_d \sim \mathcal{D}$ , and then (2) sampling  $L$  words independently according to  $W_d$ .

We represent the corpus as a matrix  $M \in \mathbb{R}^{n \times m}$ , where  $M$  permits a non-negative matrix factorization  $M = A^*X$ . Here,  $A^* \in \mathbb{R}^{n \times r}$  is the distribution of words in each of the  $r$  topics,  $X \in \mathbb{R}^{r \times m}$  is the distribution of topics in each document, and hence  $M$  is the distribution of words in each document. While there are several algorithms for learning the feature extractor  $A^*$ , it is well-known that it is hard to recover  $X$  exactly (Arora et al., 2012b). Instead, it is desirable to learn how the topics co-occur together, denoted as  $R^* = \mathbb{E}_{\mathcal{D}}[XX^\top]$ . This quantity is termed the *topic-topic covariance*. Further discussion of this has been included in Appendix A.

The topic modeling setting generally determines  $\mathcal{D}$  (e.g., in LDA,  $\mathcal{D}$  is a Dirichlet distribution). In order to recover  $A^*$  and  $R^*$  efficiently and accurately from an observed corpus  $M \sim \mathcal{D}$ , we need to make the following assumption on the underlying data distribution.

**Assumption 1** ( $p$ -separability, Arora et al. (2012b)). *The topic matrix  $A^*$  is  $p$ -separable for  $p > 0$  if for every topic  $k \in [r]$ , there exists a word  $i \in [n]$  such that  $A_{i,k}^* \geq p$  and  $A_{i,k'}^* = 0$  for all  $k' \neq k$ . Such words are called anchor words.*

Without this separability assumption, maximum likelihood estimation of a topic model is NP-hard (Arora et al., 2012b). Assumption 1 requires that  $A^*$  contains a diagonal matrix, up to row

<sup>1</sup>Without loss of generality, we assume  $L = 2$ .

permutations; intuitively, the appearance of an anchor word in a document perfectly indicates the document has nonzero probability of the corresponding topic. As we will detail in Section 4, this observation inspires a two-phase learning algorithm, whereby one first approximates the anchor words for each topic and then leverages them to identify patterns among the topics.

## 2.2 Downstream Adaptation

Topic models are frequently trained on a general corpus, and the embeddings can be later used to classify documents. The classification problem usually involves only a subset of topics. For example, after training a topic model on a large corpus of news articles with diverse topics (e.g., sports, politics, technology, finance, etc.), one relevant downstream task is to classify the subject of a given news article as sports or politics. We formalize the topic classification task below.

**Definition 1** (Topic Classification Task). A topic classification task  $\mathcal{T} = (\mathbb{T}_{\text{clf}}, \mathbf{w}^*)$  is defined by a subset of topics  $\mathbb{T}_{\text{clf}} \subset [r]$  on which the task is defined and a ground-truth labelling vector  $\mathbf{w}^* \in \mathbb{R}^r$  with bounded norm. Importantly,  $\mathbf{w}^*$  only has non-zero coordinates in the positions corresponding to  $\mathbb{T}_{\text{clf}}$ .

The classification task is defined on the latent features of a given document, so it is necessary to first identify the salient topics as they occur in the text. Fitting a topic model to the corpus yields such a feature extractor  $\mathbf{A}$  that embeds a document into the  $r$ -dimensional topic space. In order to adapt a topic model to a particular classification task, we perform head tuning on the feature extractor  $\mathbf{A}$ .

**Definition 2** (Head Tuning). For a given labelled document classification dataset  $\mathbb{D}_{\text{clf}} = \{(d_i, y_i)\}$  representing a topic classification task  $\mathcal{T}$ , embed each document  $d_i$  as a vector  $\mathbf{x}_i \in \mathbb{R}^n$  containing the word counts in the document. To perform head tuning on a pre-trained topic model  $\mathbf{A}$ , we learn  $\mathbf{w} \in \mathbb{R}^r$  to minimize

$$\ell_{\mathcal{T}}(\mathbf{w}; \mathbf{A}) = \frac{1}{|\mathbb{D}_{\text{clf}}|} \sum_{(\mathbf{x}, y) \in \mathbb{D}_{\text{clf}}} f(\mathbf{x}^\top \mathbf{A} \mathbf{w}, y)$$

where  $f$  is strongly convex in  $\mathbf{w}$ .

One example of  $f$  is the logistic loss with  $\ell_2$  regularization. For ease of exposition, we primarily consider binary classification tasks, but we point out that the definition can extend to multi-class tasks solved via the one-vs-all scheme (Rifkin & Klautau, 2004).

We note that head tuning, also referred to as linear probing, is a simpler adaptation technique than fine-tuning  $\mathbf{A}$  alongside  $\mathbf{w}$ . Nonetheless, recent works on popular language models have demonstrated that head tuning can substantially improve the ability of general pre-trained language models to solve complex classification tasks (Malladi et al., 2023a,b). Head tuning thus serves as a convenient yet effective adaptation method that avoids updating the pre-trained model, which is often desirable. For example, if a single pre-trained model needs to be separately adapted to solve many different tasks, then it is desirable to minimize the number of parameters that are fine-tuned to minimize the memory needed to store all of the adapted models.<sup>2</sup>

## 3 Unlearning

As we mentioned previously, there is increased interest in machine unlearning due to the growing scale of modern datasets and the difficulty of manually inspecting each datapoint. Theoretically, the gold standard for unlearning is that the model should behave identically to one that was trained without the datapoint in its corpus (Cao & Yang, 2015). We first define what it means for two models  $\theta_1, \theta_2 \in \Theta$  to behave *almost* identically, where  $\Theta$  denotes the parameter space of a hypothesis class. Due to randomness in learning,  $\theta_1, \theta_2$  are random variables.

**Definition 3** ( $(\epsilon, \delta)$ -indistinguishable models, Dwork et al. (2014)). Two models denoted by random variables  $\theta_1, \theta_2 \in \Theta$  are  $(\epsilon, \delta)$ -indistinguishable if for all possible subsets of models  $T \subseteq \Theta$ ,

$$\begin{aligned} \Pr(\theta_1 \in T) &\leq e^\epsilon \Pr(\theta_2 \in T) + \delta \\ \Pr(\theta_2 \in T) &\leq e^\epsilon \Pr(\theta_1 \in T) + \delta \end{aligned}$$

<sup>2</sup>This motivation has driven widespread development and adoption of parameter-efficient fine-tuning methods for large language models. Liu et al. (2021) contains a survey of such techniques.

175 We denote this as  $\theta_1 \stackrel{\epsilon, \delta}{\approx} \theta_2$ .

176 We adapt the definitions from Sekhari et al. (2021) to the topic modeling setting. A learning algo-  
 177 rithm  $\mathcal{A}$  takes in a set of  $m$  documents  $S$  and returns a topic model  $\theta = (\mathbf{A}, \mathbf{R})$ . Analogously, an  
 178 unlearning algorithm  $\mathcal{U}$  takes in the learned topic model  $\theta$ , a set of documents to unlearn  $S_f \subseteq S$ ,  
 179 and some statistics on the training set  $T(S)$ , and outputs a model. The set of datapoints to unlearn  $S_f$   
 180 is often referred to as the *forget set*. With this in mind, we now define a notion of utility-preserving  
 181 unlearning, whereby the unlearning algorithm needs to not only effectively simulate retraining the  
 182 model from scratch but also maintain the model’s performance.

183 **Definition 4** (Utility-preserving  $(\epsilon, \delta)$ -Unlearning with Deletion Capacity). Let  $m_0 \in \mathbb{N}$  be a con-  
 184 stant that depends on the topic modeling distribution  $\mathcal{D}$  satisfying Assumption 1. For any training  
 185 dataset  $S \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}$  of size at least  $m_0$ , and  $\epsilon, \delta > 0$ , we say that a pair of learning and unlearning  
 186 algorithms  $(\mathcal{A}, \mathcal{U})$  performs *utility-preserving unlearning with deletion capacity*  $T_{\epsilon, \delta}^{\mathcal{A}, \mathcal{U}}(m)$  if

1. With probability at least 0.9 over draws from  $\mathcal{D}$ , for any forget set  $S_f \subseteq S$  of size at most  $T_{\epsilon, \delta}^{\mathcal{A}, \mathcal{U}}(m)$ , model trained on  $S \setminus S_f$  is indistinguishable from that resulting from unlearning with  $\mathcal{U}$ .

$$\mathcal{U}(S_f, \mathcal{A}(S), T(S)) \stackrel{\epsilon, \delta}{\approx} \mathcal{U}(\emptyset, \mathcal{A}(S \setminus S_f), T(S \setminus S_f))$$

2. Even for an adversarially chosen  $S_f$ , the unlearned model does not suffer a large performance degradation. Formally,

$$\mathbb{E}_{\mathcal{A}, \mathcal{U}} \left[ \max_{|S_f| \leq T_{\epsilon, \delta}^{\mathcal{A}, \mathcal{U}}(m)} h(\mathcal{U}(S_f, \mathcal{A}(S), T(S))) - h^* \right] \leq 0.01$$

187 where  $h : \Theta \rightarrow \mathbb{R}$  is the loss of the topic model, and  $h^* = \min_{w \in \mathcal{W}} h(w)$  is the irreducible loss.

188 The above definition can be applied to both the pre-training and the downstream adaptation stages  
 189 of training a topic model. Of particular notice is that (1) does not guarantee (2), since the former  
 190 only concerns indistinguishability between the unlearned and retrained models, while the latter is  
 191 a statement about utility preservation. Moreover, unless  $T(S)$  contains the entire dataset, we note  
 192 that the unlearning algorithm  $\mathcal{U}$  cannot be as simple as retraining the model. In this paper, we will  
 193 design an unlearning algorithm for topic models that satisfies this definition of provable unlearning,  
 194 and the number of statistics  $T(S)$  will not depend on the initial dataset size  $m$ .

195 To show  $(\epsilon, \delta)$ -indistinguishability, we utilize the Gaussian mechanism, a classic tool from differen-  
 196 tial privacy. Given a particular function, the Gaussian mechanism essentially prescribes how much  
 197 noise one must add to the output in order for the input to be indistinguishable from a similar one.  
 198 The guarantee of the Gaussian mechanism is described in the following lemma.

199 **Lemma 1** (Gaussian Mechanism, Dwork et al. (2014)). *Let  $f$  be an arbitrary  $d$ -dimensional func-*  
 200 *tion, and define its  $\ell_2$ -sensitivity to be  $\Delta_2 f := \max_{\text{adjacent } x, y} \|f(x) - f(y)\|_2$ . Then, for  $c^2 > 2 \log \frac{1.25}{\delta}$ ,*  
 201 *the Gaussian mechanism with parameter  $\sigma \geq c \Delta_2 f / \epsilon$  is  $(\epsilon, \delta)$ -differentially private.*

202 In our case, we define adjacent inputs (i.e., training datasets) as the case where  $y$  is a superset of  $x$ .

## 203 4 Learning and Unlearning Topic Models

204 In this section, we present the learning and unlearning algorithms and guarantees for topic models.

205 **Notation.** We use  $\mathbf{A}^*$  to refer to the ground-truth topic model,  $\mathbf{A}^S$  to refer to a topic model trained  
 206 on  $S$ , and  $\mathbf{A}^F$  to denote a topic model retrained with the forget set removed  $S \setminus S_f$ . We also use  $\bar{\mathbf{A}}$   
 207 to denote the unlearned topic model before applying the Gaussian mechanism and  $\tilde{\mathbf{A}}$  to denote the  
 208 model after the mechanism is applied. Analogous notations are used for  $\mathbf{R}$ .

### 209 4.1 Learning Algorithm and Guarantees

210 Per Arora et al. (2012a), the learning algorithm  $\mathcal{A}_{\text{base}}$  takes in a corpus of documents  $S =$   
 211  $\{d_1, \dots, d_m\}$  and consists of the following three phases to learn a topic model  $\theta = (\mathbf{A}^S, \mathbf{R}^S)$ .

- 212 1. **Measure the word co-occurrences.** Compute the word co-occurrence matrix  $\mathbf{Q} \in \mathbb{R}^{n \times n}$ ,  
 213 where  $Q_{ij}$  is the number of times word  $i$  appears in the same document as word  $j$ . We also  
 214 compute  $\bar{\mathbf{Q}}$ , which normalizes the rows of  $\mathbf{Q}$  to sum to 1. A detailed discussion of the con-  
 215 struction of  $\bar{\mathbf{Q}}$  and its relationship to the factorization  $\mathbf{M} = \mathbf{A}^* \mathbf{X}$  is included in Appendix A.
- 216 2. **Identify the anchor words  $P$ .** Recall that in order to be able to learn topic models efficiently,  
 217 there must exist a set of anchor words  $P$  with  $|P| = r$ , and each anchor word must appear  
 218 exclusively in a single topic (Assumption 1). This subroutine uses  $\bar{\mathbf{Q}}$  to approximately identify  
 219 the  $r$  anchor words  $P$ .
- 220 3. **Learn the feature extractor  $\mathbf{A}^S$  and the topic-topic covariance  $\mathbf{R}^S$ .** The algorithm uses the  
 221 anchor words  $P$  and the word co-occurrences  $\bar{\mathbf{Q}}$  to learn  $\mathbf{A}^S$  and  $\mathbf{R}^S$ . Each word is expressed  
 222 as a convex combination of anchor words, and thus, topics. With appropriate normalization  
 223 and by cross-referencing information with the co-occurrence matrix, one can recover  $\mathbf{A}^*$ ,  $\mathbf{R}^*$   
 224 in the infinite data limit.

225 We sketch how this algorithm recovers the ground truth  $\mathbf{A}^*$ ,  $\mathbf{R}^*$  when one has infinitely many doc-  
 226 uments in Appendix A. Arora et al. (2012a) gives the following finite-document guarantee.

227 **Theorem 1** (Learning Guarantee). *Running  $\mathcal{A}_{base}$  on a dataset  $S$  of size  $m$ , where  $m$  is at least*

$$\max \left\{ \mathcal{O} \left( \frac{ar^3 \log n}{L(\gamma p)^6 \epsilon_0} \right), \mathcal{O} \left( \frac{a^3 r^3 \log n}{L \epsilon_0^3 (\gamma p)^4} \right), \mathcal{O} \left( \frac{r^2 \log r}{L \epsilon_0^2} \right) \right\}$$

228 *recovers  $\mathbf{A}^S$  and  $\mathbf{R}^S$  with entrywise additive error up to  $\epsilon_0$  from the ground truth  $\mathbf{A}^*$ ,  $\mathbf{R}^*$ , respec-*  
 229 *tively. Here,  $a$  is the topic imbalance parameter, and  $\gamma$  is the condition number of the ground truth*  
 230  *$\mathbf{R}^*$ . Formally, we have  $a = \max_{i,j \in [r]} \Pr_{\mathcal{D}}[z = i] / \Pr_{\mathcal{D}}[z = j]$ .*

231 **Approximating the anchor words.** We defer a precise description of the anchor word identification  
 232 algorithm to Appendix A and instead focus here on the intuitions driving its design and the guar-  
 233 antees we will use throughout the paper. First, we note the relationship between  $\bar{\mathbf{Q}}$  and the set of  
 234 anchor words. If we had infinitely many documents, then the convex hull of the rows in  $\bar{\mathbf{Q}}$  will be a  
 235 simplex with vertices corresponding to the anchor words, because each anchor word corresponds to  
 236 a topic, and each topic prescribes a distribution over words. However, in the finite document setting,  
 237 each row of  $\bar{\mathbf{Q}}$  only approximates their expected value, and so one must approximate the vertices of  
 238 a convex hull when given access to a perturbation of the points that define it.

239 We start by requiring that each topic is distinctly different from any mixture on the other topics.  
 240 Formally, this requires that the simplex is robust, in that each vertex (i.e., anchor word) is sufficiently  
 241 far from any combination of the other topics. Most topic modeling settings define lower bounds on  
 242 the robustness of the simplex. By a result in Arora et al. (2012b), the simplex defined by the  $r$   
 243 anchor word rows of the population  $\bar{\mathbf{Q}}$  is  $\gamma p$ -robust. We can now define exactly the sense in which  
 244 a  $\bar{\mathbf{Q}}$  computed on a finite dataset approximates the population co-occurrence matrix.

245 **Definition 5.** Let  $\{a_i\}_{i=1}^n$  be a set of points whose convex hull  $P$  is a simplex with vertices  $\{v_i\}_{i=1}^r$ .  
 246 We say a set of  $r$  points is  $\epsilon$ -close to the vertex set  $\{v_i\}_{i=1}^r$  if each of the  $r$  points is  $\epsilon$ -close in  $\ell_2$   
 247 distance to a different vertex in  $P$ . Moreover, we say that a simplex  $P$  is  $\beta$ -robust if for every vertex  
 248  $v$  of  $P$ , the  $\ell_2$  distance between  $v$  and the convex hull of the rest of the vertices is at least  $\beta$ .

249 In the context of this definition,  $P$  corresponds to the ground truth convex hull, and the finite sample  
 250  $\bar{\mathbf{Q}}$  can be seen as a perturbation to it. In particular, Arora et al. (2012a) used this to establish a  
 251 guarantee on the accuracy of anchor word recovery.

252 **Lemma 2** (Approximation Guarantee on Anchor Words). *Suppose each row of  $\bar{\mathbf{Q}}$  is at most  $\delta$*   
 253 *distance away from the ground truth  $\gamma p$ -robust simplex  $\bar{\mathbf{Q}}^*$  in  $\ell_2$  norm. If  $20r\delta/(\gamma p)^2 < \gamma p$ , then*  
 254 *the set of anchor words found by the algorithm is  $O(\delta/\gamma p)$ -close to the ground truth anchor words.*

255 We now describe how to use the recovered approximate anchor words to learn the topic model.

256 **Learning the topic model from anchor words.** We are given the set of anchor words  $P$ , the word  
 257 co-occurrence matrix  $\mathbf{Q} \in \mathbb{R}^{n \times n}$ , and the normalized co-occurrence matrix  $\bar{\mathbf{Q}}$ . Our goal is to use  
 258 these quantities to learn  $\mathbf{A} \in \mathbb{R}^{n \times r}$  and  $\mathbf{R} \in \mathbb{R}^{r \times r}$ . We will do so by first expressing each word  
 259  $i \in [n]$  as a convex combination of the anchor words (and thus, the topics). In particular, for each  
 260 word  $i$ , we learn the coefficients  $\mathbf{C}_i \in \Delta_r$  as

$$\mathbf{C}_i = \arg \min_{v \in \Delta_r} \|\bar{\mathbf{Q}}_i - v^\top \bar{\mathbf{Q}}_P\|^2 \quad (1)$$

---

**Algorithm 1** Unlearning algorithm ( $\mathcal{U}_{\text{base}}$ )

---

**Input:** Forget set  $S_f \subseteq S$ , statistics  $T(S)$  which include  $\{C_i^S\}_{i=1}^n, Q^S, P$ , normalization constants  $p^S$

**Output:** Unlearned model  $\tilde{A}, \tilde{R}$

Compute the updated co-occurrence matrix  $Q^F$  by subtracting documents in  $S_f$

Store the updated normalization constants  $p^F = Q^F \mathbf{1}$

**for**  $i$  in  $1, \dots, n$  **do**

    Newton step update on  $C_i$ 's:

$$\bar{C}_i^F \leftarrow C_i^S - H_{C_i^S}^{-1} \nabla \mathcal{L}(C_i^S, S \setminus S_f) \quad (2)$$

$$\bar{C}_i^F \leftarrow \text{proj}_{\Delta_r}(\bar{C}_i^F) \quad (3)$$

    where  $\mathcal{L}(v, S \setminus S_f) := \|\bar{Q}_i^F - v^\top \bar{Q}_P^F\|^2$  and  $H_{C_i^S} = \nabla^2 \mathcal{L}(C_i^S, S \setminus S_f)$

**end for**

$\bar{A}' = \text{diag}(p^F) \bar{C}$

$\bar{A}$  = column normalized  $\bar{A}'$

$\bar{R} = \bar{A}^\dagger Q^F \bar{A}^{\dagger\top}$  where  $\bar{A}^\dagger$  is the pseudoinverse of  $\bar{A}$

Sample  $\nu_A, \nu_R$  from normal distribution defined by Gaussian mechanism guarantee

$\tilde{A}$  = Project each column of  $\bar{A} + \nu_A$  to  $\Delta_n$ .

$\tilde{R}$  = Project  $\bar{R} + \nu_R$  onto the set of PSD matrices.

**return** The unlearned topic model  $\tilde{A}, \tilde{R}$

---

261 where  $\bar{Q}_P$  is the  $P$  rows of  $\bar{Q}$  corresponding to the anchor words. Arora et al. (2012a) showed the  
 262 following approximation guarantee for  $C_i$  compared to the ground-truth coefficients.

263 **Lemma 3.** When  $20r\delta/(\gamma p)^2 < \gamma p$ , for every word  $i$ ,  $C_i$  has entrywise error  $O(\delta/(\gamma p)^2)$  from  $C_i^*$ .  
 264

265 We then normalize this  $C_i$  by the total number of co-occurrences that word  $i$  is involved in. Note that  
 266 the  $C_i$  can be assembled into a matrix  $C \in \mathbb{R}^{n \times r}$ . We set  $A$  to be  $C$  after normalizing the columns  
 267 sum to 1, since the columns represent the topic-conditioned distribution over the vocabulary. We  
 268 finally compute  $R = A^\dagger Q A^{\dagger\top}$ , where  $A^\dagger$  denotes the pseudoinverse of  $A$ .

## 269 4.2 Unlearning Algorithm and Guarantees

Learning Phase	Retrain Time	Unlearning Update	Unlearning Time
Co-occurrence matrix computation	$\mathcal{O}(m)$	Updating frequencies	$\mathcal{O}(m_U)$
Identify anchor words	$\mathcal{O}(n^2 + nr/\epsilon_0^2)$	Use learned anchor words	$\mathcal{O}(1)$
Recover topics from anchors	$\mathcal{O}(n^2 r + nr^2/\epsilon_0^2)$	Projected Newton step	$\mathcal{O}(nr^2)$
Head tuning $w$ (Definition 2)	ERM	Newton step	$\mathcal{O}(r^3)$

Table 1: Our unlearning algorithms generally have a runtime shorter than the retraining procedure. ERM denotes empirical risk minimization, and we note the training time relies on the error tolerance.

270 We describe our unlearning algorithm  $\mathcal{U}_{\text{base}}$  to forget a set  $S_f$  from a trained model (Algorithm 1),  
 271 which crucially updates  $C_i$  with a Newton step. We then compute  $\tilde{A}$  from the modified  $C_i$  and  
 272 apply the Gaussian mechanism to ensure indistinguishability. We describe our formal guarantee on  
 273 the unlearning algorithm below, sketching out our utility preserving guarantees with respect to  $A^*$ .  
 274 The arguments for  $R^*$  follow analogously; we defer the discussion to the appendix.

275 **Theorem 2** (Utility-Preserving Unlearning on the Base Model). *Let  $\mathcal{A}_{\text{base}}$  be the learning algo-*  
 276 *rithm described in the prior sections and  $\mathcal{U}_{\text{base}}$  be the unlearning algorithm in Algorithm 1. Then,*  
 277  *$(\mathcal{A}_{\text{base}}, \mathcal{U}_{\text{base}})$  performs utility-preserving unlearning with deletion capacity*

$$T_{\epsilon, \delta}^{\mathcal{A}_{\text{base}}, \mathcal{U}_{\text{base}}}(m) \geq c \cdot \frac{m}{r^2 \sqrt{rn}} \quad (4)$$

where  $m$  is the number of training documents,  $r$  is the number of topics, and  $c$  is a constant dependent on  $\epsilon, \delta$ , and  $\mathcal{D}$ . The loss function  $h$  used in the utility-preserving definition is the maximum entrywise error from the ground truth topic model  $\mathbf{A}^*$ .

**Proof sketch.** The full proof can be found in Appendix B.2. We delete  $m_U \leq \frac{0.001m\epsilon_0(\gamma p)^3}{a^2r^2}$  points. This upper bound ensures that the anchor words are likely unchanged per Lemma 2. Recall that utility-preserving unlearning requires: (1) that the unlearned model is indistinguishable from the retrained model, and (2) that the unlearned model is not too far from the ground-truth model.

*Indistinguishability.* The Gaussian mechanism introduced in Lemma 1 allows us to make two models with a given  $\ell_2$ -sensitivity  $(\epsilon, \delta)$ -indistinguishable from each other. We bound the  $\ell_2$ -sensitivity of the feature extractor  $\mathbf{A}$  by noting that  $\bar{\mathbf{A}}$  is a rescaled version of  $\bar{\mathbf{C}}$ .

**Lemma 4.** For  $\epsilon, \delta > 0$ , the following holds for the  $\bar{\mathbf{C}}$  and the topic matrix  $\bar{\mathbf{A}}$ :

$$\|\bar{\mathbf{C}} - \mathbf{C}^F\|_\infty \leq c \cdot \frac{arm_U}{m\epsilon_0\gamma p} \quad \|\bar{\mathbf{A}} - \mathbf{A}^F\|_\infty \leq (ar) \cdot \|\bar{\mathbf{C}} - \mathbf{C}^F\|_\infty \quad (5)$$

Applying the Gaussian mechanism with noise  $\sigma = \frac{\Delta}{\epsilon} \sqrt{2 \log(1.25/\delta)}$ , where  $\Delta = c\sqrt{nr} \cdot \frac{(ar)^2 m_U}{m\epsilon_0\gamma p}$  and followed by projecting the columns of  $\bar{\mathbf{A}} + \nu_{\mathbf{A}}$  back to  $\Delta_n$  yields the desired result.

*Utility Preservation.* We first apply Lemma 2 to show that, with high probability, the anchor words do not change when unlearning  $m_U$  documents. Then, we use Lemma 8 to bound the distance between the unlearned  $\bar{\mathbf{C}}_i$  and the ground truth  $\mathbf{C}_i^*$ . Accounting for the noise added via the Gaussian mechanism completes the proof.

**Lemma 5.** For  $\epsilon, \delta > 0$ , denote the unlearned model after the Gaussian mechanism described above as  $\tilde{\mathbf{A}}$ . Then,  $\tilde{\mathbf{A}}$  satisfies:

$$\mathbb{E}[\|\tilde{\mathbf{A}} - \mathbf{A}^*\|_\infty] \leq c \cdot \frac{(ar)^2 m_U}{m\epsilon_0\gamma p} \cdot \left( \sqrt{nr} \cdot \sqrt{\log(nr)} \cdot \frac{\sqrt{\log(1/\delta)}}{\epsilon} + 1 \right) \quad (6)$$

Each of the two terms in the above equation yield a constraint on  $m_U$ . In particular,  $m_U \leq \min \left\{ \tilde{\mathcal{O}}\left(\frac{m}{r^2\sqrt{nr}}\right), \mathcal{O}\left(\frac{m}{r^2}\right) \right\}$ , so setting  $m_U \leq \tilde{\mathcal{O}}\left(\frac{m}{r^2\sqrt{nr}}\right)$  completes the proof.

## 5 Unlearning with respect to a Downstream Task

We are interested in unlearning a set of pre-training documents  $S_f \subseteq S$ . A topic classification task is usually defined on a subset of the topics in the dataset — for example, if the pre-training corpus contained diverse news articles, one plausible downstream task is to classify the content of a given document as containing politics or sports. Definition 1 formalizes this: a topic classification task  $\mathcal{T} = (\mathbb{T}_{\text{clf}}, \mathbf{w}^*)$  is defined on a subset of the topics  $\mathbb{T}_{\text{clf}}$  and a  $r$ -length ground-truth labelling vector  $\mathbf{w}^* \in \mathcal{W}_{\text{head}}$ , where  $\mathbf{w}^*$  only has non-zero values in positions corresponding to  $\mathbb{T}_{\text{clf}}$ . We describe two possible settings under which we can show utility-preserving unlearning.

### 5.1 Naive Setting

In the first setting, the learning algorithm  $\mathcal{A}_{\text{head, naive}}$  returns the pre-trained feature extractor  $\mathbf{A}$  and the head  $\mathbf{w}$  separately. So, we must ensure that the forget set  $S_f \subseteq S$  cannot be recovered from either  $\mathbf{A}$  or  $\mathbf{w}$ . As such, we must necessarily perform unlearning on  $\mathbf{A}$  as described in Algorithm 1, which means that unlearning the fine-tuned model is exactly as difficult as unlearning the base model.

**Theorem 3** (Unlearning when releasing  $\mathbf{A}$  and  $\mathbf{w}$ ). For a downstream task  $\mathcal{T}$  with loss function  $\ell_{\mathcal{T}}$ , consider the unlearning algorithm  $\mathcal{U}_{\text{head, naive}}$  that first runs Algorithm 1 to compute  $\tilde{\mathbf{A}} = \mathcal{U}_{\text{base}}(S_f, \mathcal{A}_{\text{base}}(S), T(S))$ , where  $(\mathcal{A}_{\text{base}}, \mathcal{U}_{\text{base}})$  performs utility-preserving unlearning (Theorem 2). Then, it fits a head  $\mathbf{w} = \arg \min_{\mathbf{w} \in \mathcal{W}_{\text{head}}} \ell_{\mathcal{T}}(\mathbf{w}; \tilde{\mathbf{A}})$  and returns  $\tilde{\mathbf{A}}$  and  $\mathbf{w}$ . We assert that  $(\mathcal{A}_{\text{head, naive}}, \mathcal{U}_{\text{head, naive}})$  performs utility-preserving unlearning (Definition 4).

---

**Algorithm 2** Unlearning algorithm for task  $\mathcal{T}$  ( $\mathcal{U}_{head}$ )

---

**Input:** Document deletion requests  $S_f \subseteq S$ , statistics  $T(S)$  which include  $\mathbf{A}^S, \{\mathbf{C}_i^S\}_{i=1}^n, \mathbf{Q}^S, P, \text{diag}(\mathbf{p}^S), \mathbf{w}^S = \arg \min_{\mathbf{w} \in \mathcal{W}_{head}} \ell_{\mathcal{T}}(\mathbf{w}; \mathbf{A}^S)$   
 $\bar{\mathbf{A}}, \bar{\mathbf{R}} = \text{Run Algorithm 1 } (\mathcal{U}_{base})$  up to the Gaussian mechanism  
 $\bar{\mathbf{w}} = \mathbf{w}^S - \mathbf{H}_{\mathbf{w}^S}^{-1} \nabla_{\mathbf{w}} \ell_{\mathcal{T}}(\mathbf{w}^S; \bar{\mathbf{A}})$  where  $\mathbf{H}_{\mathbf{w}^S} = \nabla_{\mathbf{w}}^2 \ell_{\mathcal{T}}(\mathbf{w}^S; \bar{\mathbf{A}})$   
**return**  $(\mathbf{A}^S)^\dagger \bar{\mathbf{A}} \bar{\mathbf{w}} + \xi$ , in accordance with the Gaussian mechanism

---

319 Given the guarantee on  $\tilde{\mathbf{A}}$  from Theorem 2, we show that this result extends to  $\mathbf{w}$  by the well-  
320 known fact: for  $\epsilon, \delta > 0$ , post-processing indistinguishable quantities (Definition 3) preserves  
321  $(\epsilon, \delta)$ -indistinguishability (Dwork et al., 2014). The full proof of utility preservation can be found  
322 in Appendix C, which essentially boils down to a Lipschitz condition. However, there are some  
323 downsides to this algorithm. First, it requires retraining the head  $\mathbf{w}$  for each unlearning request, but  
324 we want to perform unlearning without access to  $\mathbb{D}_{clf}$ . Second, repeatedly noising the base model  
325 via the Gaussian mechanism will erode its utility. We address these issues in the realistic setting.

## 326 5.2 Realistic Setting

327 There is little reason to release  $\mathbf{A}$  and  $\mathbf{w}$  separately after fine-tuning the model, because it is unclear  
328 why one would want to use  $\mathbf{A}$  without  $\mathbf{w}$  or vice versa. One can obtain  $\mathbf{A}$  directly after pre-training  
329 instead of relying on a fine-tuned model, and there is little use for  $\mathbf{w}$  alone, because it is highly  
330 sensitive to the specific topics extracted by  $\mathbf{A}$  and their ordering. As such, we argue for releasing  
331 the fine-tuned model as a single matrix<sup>3</sup>  $\mathbf{B} = \mathbf{A}\mathbf{w}$ , where  $\mathbf{B} \in \mathbb{R}^{n \times 1}$ .

332 **Theorem 4** (Utility-Preserving Unlearning on the Downstream Task). *Suppose that the downstream*  
333 *task  $\mathcal{T}$  only depends on a subset of topics  $\mathbb{T}_{clf} \subseteq [r]$ ; that is,  $\mathbf{w}^* = \arg \min_{\mathbf{v} \in \mathcal{W}_{base}} \ell_{\mathcal{T}}(\mathbf{v}; \mathbf{A}^*)$  has*  
334 *non-zero entries only in the index set  $\mathbb{T}_{clf}$ . Denote  $q := \min_{k \in \mathbb{T}_{clf}} \Pr_{\mathcal{D}}[z = k]$ , and let  $\mathcal{A}_{head}$  be*  
335 *the head tuning algorithm (Definition 2) and  $\mathcal{U}_{head}$  be Algorithm 2. Then,  $(\mathcal{A}_{head}, \mathcal{U}_{head})$  performs*  
336 *utility-preserving unlearning with deletion capacity*

$$T_{\epsilon, \delta}^{\mathcal{A}_{head}, \mathcal{U}_{head}}(m) \geq c' \cdot \frac{mq}{r\sqrt{nr}} \quad (7)$$

337 where  $c'$  is a constant dependent on  $\epsilon, \delta, \mathcal{D}$ , and  $\mathcal{T}$ .

338 The full proof is in Appendix C, including the worst case of  $\mathbb{T}_{clf} = [r]$ . When the task relies heavily  
339 on every single topic (i.e.,  $q = 1/ar$ ), the above guarantee is equivalent to the one in the pre-training  
340 phase. However, in most realistic settings, the downstream task will only depend on a subset of  
341 the latent topics in the corpus. In this case,  $q > 1/ar$ , and we can unlearn more points without  
342 degrading the utility of the model. Intuitively this makes sense too; the more reliance  $\mathcal{T}$  has on a  
343 rare topic, the less adversarial deletion it can tolerate.

344 **Proof sketch.** We again assume that we are deleting  $m_U \leq \frac{0.001m\epsilon_0(\gamma p)^3}{a^2 r^2}$  points. For any mod-  
345 ification made to  $\mathbf{A}$ , there is an equivalent modification that can be made to  $\mathbf{w}$  instead such that  
346  $\mathbf{B} = \mathbf{A}\mathbf{w}$  is preserved, so we do not need to update  $\mathbf{A}$ . We look for  $\mathbf{v} \in \mathcal{W}_{head}$  such that  
347  $\mathbf{A}^S \mathbf{v} = \mathbf{A}^F \mathbf{w}^F$ , where  $\mathbf{w}^F$  is the head learned on  $\mathbf{A}^F$ . It can be shown that  $\bar{\mathbf{A}}^S$  has a unique  
348 pseudoinverse since it is full rank; naturally, we set  $\mathbf{v} = \bar{\mathbf{A}}^{S\dagger} \mathbf{A}^F \mathbf{w}^F$ , thereby ensuring privacy even  
349 if one recovers a part of  $\mathbf{A}$  from  $\mathbf{B} = \mathbf{A}\mathbf{w}$ . We furthermore define  $\bar{\mathbf{v}}$  that is fit to the unlearned  
350 model before the Gaussian mechanism,  $\bar{\mathbf{v}} = \bar{\mathbf{A}}^{S\dagger} \bar{\mathbf{A}} \bar{\mathbf{w}}$ . We now need to show  $\mathbf{v}$  and  $\bar{\mathbf{v}}$  satisfy both  
351 the indistinguishability and utility preservation conditions in Definition 4.

352 **Indistinguishability.** Let  $\bar{\mathbf{w}}^* = \arg \min_{\mathbf{v} \in \mathcal{W}_{head}} \ell_{\mathcal{T}}(\mathbf{v}; \bar{\mathbf{A}})$  denote the result of head tuning  $\bar{\mathbf{A}}$ , and let  
353  $\bar{\mathbf{w}}$  be the result of taking a Newton step on  $\mathbf{w}$  (see Algorithm 2). Then by triangle inequality,

$$\|\bar{\mathbf{A}}\bar{\mathbf{w}} - \mathbf{A}^F \mathbf{w}^F\|_2 \leq \|\bar{\mathbf{A}}\bar{\mathbf{w}} - \bar{\mathbf{A}}\bar{\mathbf{w}}^*\|_2 + \|\bar{\mathbf{A}}\bar{\mathbf{w}}^* - \mathbf{A}^F \bar{\mathbf{w}}^*\|_2 + \|\mathbf{A}^F \bar{\mathbf{w}}^* - \mathbf{A}^F \mathbf{w}^F\|_2 \quad (8)$$

354 Informally, the first term is controlled by the error in the Newton step approximation, and the third  
355 term is bounded by the error to the retrained  $\mathbf{w}^F$ . The remaining term can be rewritten as  $\|(\bar{\mathbf{A}} -$

---

<sup>3</sup>One can generalize this to the case where the downstream task is a  $C$ -way classification, in which case  $\mathbf{B} \in \mathbb{R}^{n \times C}$ .

356  $\mathbf{A}^F)(\bar{\mathbf{w}}^* - \mathbf{w}^*) + (\bar{\mathbf{A}} - \mathbf{A}^F)\mathbf{w}^*\|$ , where the first term can be bounded using the same technique  
 357 use to prove Lemmas 4 and 5. The second term can be bounded by noting that  $\mathbf{w}^*$  is sparse, which  
 358 yields the below lemma that plays a crucial role in establishing the improved deletion capacity.

**Lemma 6** (Modification of Lemma 4 for downstream task). *For  $\epsilon, \delta > 0$ ,*

$$\|\bar{\mathbf{A}} - \mathbf{A}^F\|_\infty \leq \frac{1}{q} \cdot \|\bar{\mathbf{C}} - \mathbf{C}^F\|_\infty = c \cdot \frac{1}{q} \cdot \frac{arm_U}{m\epsilon_0\gamma p}$$

359 As in the pre-training case, we can now set the noise scale in the Gaussian mechanism and complete  
 360 the proof. In the worst case, when the downstream task depends on *every* topic, then  $q = 1/ar$ , and  
 361 we recover Lemma 4; however, this is unlikely to happen in practice.

362 *Utility Preservation.* We compare the value of  $\mathbf{v}$  after the Gaussian mechanism  $\tilde{\mathbf{v}} = \bar{\mathbf{v}} + \nu_{\bar{\mathbf{v}}}$  to what  
 363 it would be for the ground-truth model  $\mathbf{v}^* = (\mathbf{A}^S)^\dagger \mathbf{A}^* \mathbf{w}^*$ . We again rely the sparsity of  $\mathbf{w}^*$  and  
 364 bound  $\mathbb{E}[\|\tilde{\mathbf{v}} - \mathbf{v}^*\|_\infty]$  in Lemma 31.

## 365 6 Related Works

366 **Provable unlearning.** One ideally wants the unlearned model to behave identically to one that  
 367 was retrained from scratch with the forget set removed from the training data (Cao & Yang, 2015;  
 368 Bourtoutle et al., 2021; Gupta et al., 2021). This is difficult to achieve in many settings, so there are  
 369 several notions of approximate unlearning (Ginart et al., 2019; Guo et al., 2020; Neel et al., 2021)  
 370 reminiscent of differential privacy (Dwork et al., 2014). Most relevant to our work is the notion  
 371 of  $(\epsilon, \delta)$ -unlearning introduced in Sekhari et al. (2021), which we adapt to construct Definition 4.  
 372 Our work focuses on deriving unlearning guarantees in the pre-training and fine-tuning pipeline.  
 373 Golatkar et al. (2020) is closest to our work. They show considerably weaker guarantees on un-  
 374 learning information with respect to probes fit to the weights. In contrast, our work is focused on  
 375 realistic topic classification tasks and demonstrates strong guarantees (Definition 4). Recent works  
 376 have extended notions of certified unlearning to nonconvex settings. Zhang et al. (2024a); Mu &  
 377 Klabjan (2024); Chien et al. (2024) provide unlearning algorithms without deletion capacity guar-  
 378 antees. Qiao et al. (2024) also proposes an unlearning method for non-convex settings but analyzes  
 379 its deletion capacity in a convex setting. Our work extends beyond the convex setting to provide  
 380 provable unlearning methods and corresponding deletion capacity analysis for non-convex models.

381 **Theoretical analysis of pre-training and fine-tuning.** Our downstream task definition (Sec-  
 382 tion 2.2) is inspired by works on transfer learning in language models (Saunshi et al., 2021; Wei  
 383 et al., 2021; Wu et al., 2023; Kumar et al., 2022), contrastive learning (Lee et al., 2021; HaoChen &  
 384 Ma, 2023), and meta-learning (Chua et al., 2021; Collins et al., 2022; Yüksel et al., 2024).

## 385 7 Conclusion

386 This work uses topic models to develop the first provable guarantees on unlearning in the modern-  
 387 day pre-training and fine-tuning paradigm. We propose two unlearning algorithms that can effec-  
 388 tively and efficiently unlearn from both the pre-trained model (Algorithm 1 and Theorem 2) and  
 389 the fine-tuned model (Algorithm 2 and Theorem 4). Notably, we find that it is easier, in terms of  
 390 the deletion capacity (Definition 4), to unlearn pre-training data from the fine-tuned model, and we  
 391 can do so without modifying the pre-trained base model. Our findings suggest that task-specific un-  
 392 learning is easier than full model unlearning, providing a promising path forward to design efficient  
 393 algorithms for large-scale models.

394 The most notable limitation of our work is that our usage of topic models, which permit a tractable  
 395 analysis but cannot capture interesting features of modern-day language models (e.g., their autore-  
 396 gressive nature). Moreover, with the growing popularity of foundation models, there is scholarly  
 397 discussion around meaningful definitions of unlearning and how they can be measured (Thudi et al.,  
 398 2022; Lee et al., 2024). Our work focuses on traditional notions of unlearning centered on differen-  
 399 tial privacy (see Definition 4), but we hope to extend these definitions to capture additional features  
 400 of generative models that are salient to their real-world uses.

## References

- Sanjeev Arora, Rong Ge, Yoni Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees, 2012a. URL <https://arxiv.org/abs/1212.4777>.
- Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models - going beyond svd, 2012b.
- Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1536–1546. IEEE, 2021.
- Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.
- David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*, pp. 113–120, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143859. URL <https://doi.org/10.1145/1143844.1143859>.
- David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avnika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. URL <https://arxiv.org/abs/2108.07258>.
- Lucas Bourtole, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.
- Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296, 2017.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf).

453 Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015*  
454 *IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

455 Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine  
456 Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data  
457 from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp.  
458 2633–2650, 2021.

459 Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwal, Florian Tramer, Borja  
460 Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd*  
461 *USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

462 Eli Chien, Haoyu Wang, Ziang Chen, and Pan Li. Langevin unlearning: A new perspective of noisy  
463 gradient descent for machine unlearning, 2024. URL [https://arxiv.org/abs/2401.](https://arxiv.org/abs/2401.10371)  
464 10371.

465 Rochelle Choenni, Ekaterina Shutova, and Robert van Rooij. Stepmothers are mean and aca-  
466 demics are pretentious: What do pretrained language models learn about you? In Marie-  
467 Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the*  
468 *2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1477–1491, Online  
469 and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguis-  
470 tics. doi: 10.18653/v1/2021.emnlp-main.111. URL [https://aclanthology.org/2021.](https://aclanthology.org/2021.emnlp-main.111)  
471 emnlp-main.111.

472 Kurtland Chua, Qi Lei, and Jason D Lee. How fine-tuning allows for effective meta-learning. *Ad-*  
473 *vances in Neural Information Processing Systems*, 34:8871–8884, 2021.

474 Rob Churchill and Lisa Singh. The evolution of topic modeling. *ACM Comput. Surv.*, 2022.

475 Liam Collins, Aryan Mokhtari, Sewoong Oh, and Sanjay Shakkottai. Maml and anil provably learn  
476 representations. In *International Conference on Machine Learning*, pp. 4238–4310. PMLR, 2022.

477 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of  
478 deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and  
479 Tamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of*  
480 *the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long*  
481 *and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Com-  
482 putational Linguistics. doi: 10.18653/v1/N19-1423. URL [https://aclanthology.org/](https://aclanthology.org/N19-1423)  
483 N19-1423.

484 Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations*  
485 *and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

486 Ronen Eldan and Mark Russinovich. Who’s harry potter? approximate unlearning in llms, 2023.  
487 URL <https://arxiv.org/abs/2310.02238>.

488 *DOE 1 v. GitHub, Inc.* 4:22-cv-06823, N.D. Cal. 2022.

489 *Tremblay v. OpenAI, Inc.,* 23-cv-03416-AMO, (N.D. Cal.), 2023.

490 European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the Euro-  
491 pean Parliament and of the Council. URL [https://data.europa.eu/eli/reg/2016/](https://data.europa.eu/eli/reg/2016/679/oj)  
492 679/oj.

493 Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining  
494 through selective synaptic dampening. In *Proceedings of the AAAI Conference on Artificial Intel-*  
495 *ligence*, volume 38, pp. 12043–12051, 2024.

496 Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts  
497 from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer*  
498 *Vision*, pp. 2426–2436, 2023.

499 Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason  
500 Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile:  
501 An 800gb dataset of diverse text for language modeling, 2020. URL [https://arxiv.org/](https://arxiv.org/abs/2101.00027)  
502 [abs/2101.00027](https://arxiv.org/abs/2101.00027).

503 Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data dele-  
504 tion in machine learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox,  
505 and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Cur-  
506 ran Associates, Inc., 2019. URL [https://proceedings.neurips.cc/paper\\_files/](https://proceedings.neurips.cc/paper_files/paper/2019/file/cb79f8fa58b91d3af6c9c991f63962d3-Paper.pdf)  
507 [paper/2019/file/cb79f8fa58b91d3af6c9c991f63962d3-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/cb79f8fa58b91d3af6c9c991f63962d3-Paper.pdf).

508 Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net:  
509 Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer*  
510 *Vision and Pattern Recognition*, pp. 9304–9312, 2020.

511 Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal  
512 from machine learning models. In *International Conference on Machine Learning*, pp. 3832–  
513 3842. PMLR, 2020.

514 Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites.  
515 Adaptive machine unlearning. *Advances in Neural Information Processing Systems*, 34:16319–  
516 16330, 2021.

517 Jeff Z. HaoChen and Tengyu Ma. A theoretical study of inductive biases in contrastive learning.  
518 In *The Eleventh International Conference on Learning Representations*, 2023. URL [https://](https://openreview.net/forum?id=AuEgN1EAmed)  
519 [openreview.net/forum?id=AuEgN1EAmed](https://openreview.net/forum?id=AuEgN1EAmed).

520 Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. Inex-  
521 act unlearning needs more careful evaluations to avoid a false sense of privacy, 2024. URL  
522 <https://arxiv.org/abs/2403.01218>.

523 Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked au-  
524 toencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer*  
525 *vision and pattern recognition*, pp. 16000–16009, 2022.

526 Luxi He, Yangsibo Huang, Weijia Shi, Tinghao Xie, Haotian Liu, Yue Wang, Luke Zettlemoyer,  
527 Chiyuan Zhang, Danqi Chen, and Peter Henderson. Fantastic copyrighted beasts and how (not)  
528 to generate them. *arXiv preprint arXiv:2406.14526*, 2024.

529 Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A Lemley, and Percy  
530 Liang. Foundation models and fair use. *Journal of Machine Learning Research*, 24(400):1–79,  
531 2023.

532 Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza  
533 Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Train-  
534 ing compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

535 Thomas Hofmann et al. Probabilistic latent semantic analysis. In *UAI*, volume 99, pp. 289–296,  
536 1999.

537 Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion  
538 from machine learning models, 2021.

539 Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and  
540 Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models, 2023. URL  
541 <https://openreview.net/forum?id=zAxuIJLb38>.

542 Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-  
543 tuning can distort pretrained features and underperform out-of-distribution. In *International Con-*  
544 *ference on Learning Representations*, 2022. URL [https://openreview.net/forum?](https://openreview.net/forum?id=UYneFzXSJWh)  
545 [id=UYneFzXSJWh](https://openreview.net/forum?id=UYneFzXSJWh).

546 Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded  
547 machine unlearning. In *Thirty-seventh Conference on Neural Information Processing Systems*,  
548 2023. URL <https://openreview.net/forum?id=OveBaTtUAT>.

549 Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps:  
550 Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–  
551 323, 2021.

552 Katherine Lee, A. Cooper, Christopher Choquette-Choo, Ken Liu, Matthew Jagielski, Niloofar  
553 Mireshghallah, Lama Ahmed, James Grimmelmann, David Bau, Christopher De Sa, Fernando  
554 Delgado, Vitaly Shmatikov, Katja Filippova, Seth Neel, Miranda Bogen, Amy Cyphert, Mark  
555 Lemley, and Nicolas Papernot. Extended abstract: Machine unlearning doesn’t do what you  
556 think, 04 2024.

557 Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic corre-  
558 lations. In *Proceedings of the 23rd international conference on Machine learning*, pp. 577–584,  
559 2006.

560 Jiaqi Liu, Jian Lou, Zhan Qin, and Kui Ren. Certified minimax unlearning with generalization rates  
561 and deletion capacity. *Advances in Neural Information Processing Systems*, 36, 2024.

562 Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-  
563 train, prompt, and predict: A systematic survey of prompting methods in natural language pro-  
564 cessing, 2021. URL <https://arxiv.org/abs/2107.13586>.

565 Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William  
566 Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. A large-  
567 scale audit of dataset licensing and attribution in ai. *Nature Machine Intelligence*, 6(8):975–987,  
568 2024.

569 Ananth Mahadevan and Michael Mathioudakis. Cost-effective retraining of machine learning mod-  
570 els. *arXiv preprint arXiv:2310.04216*, 2023.

571 Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. Tofu: A task  
572 of fictitious unlearning for llms, 2024. URL <https://arxiv.org/abs/2401.06121>.

573 Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and San-  
574 jeev Arora. Fine-tuning language models with just forward passes. In *Thirty-seventh Confer-  
575 ence on Neural Information Processing Systems*, 2023a. URL [https://openreview.net/  
576 forum?id=Vota6rFhBQ](https://openreview.net/forum?id=Vota6rFhBQ).

577 Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based  
578 view of language model fine-tuning. In *International Conference on Machine Learning*, pp.  
579 23610–23641. PMLR, 2023b.

580 Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark  
581 Steedman. Sources of hallucination by large language models on inference tasks. *arXiv preprint  
582 arXiv:2305.14552*, 2023.

583 Siqiao Mu and Diego Klabjan. Rewind-to-delete: Certified machine unlearning for nonconvex func-  
584 tions, 2024. URL <https://arxiv.org/abs/2409.09778>.

585 Tarek Naous, Michael Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring  
586 cultural bias in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar  
587 (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Lin-  
588 guistics (Volume 1: Long Papers)*, pp. 16366–16393, Bangkok, Thailand, August 2024. As-  
589 sociation for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.862. URL [https:  
590 //aclanthology.org/2024.acl-long.862](https://aclanthology.org/2024.acl-long.862).

591 Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods  
592 for machine unlearning. In *Algorithmic Learning Theory*, pp. 931–962. PMLR, 2021.

593 Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. Variational bayesian unlearning.  
594 *Advances in Neural Information Processing Systems*, 33:16025–16036, 2020.

595 Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khali-  
596 dov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran,  
597 Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra,  
598 Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick  
599 Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features with-  
600 out supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL  
601 <https://openreview.net/forum?id=a68SUt6zFt>.

602 Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli,  
603 Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb  
604 dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv*  
605 *preprint arXiv:2306.01116*, 2023.

606 Xinbao Qiao, Meng Zhang, Ming Tang, and Ermin Wei. Efficient and generalizable certified un-  
607 learning: A hessian-free recollection approach, 2024. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2404.01712)  
608 [2404.01712](https://arxiv.org/abs/2404.01712).

609 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,  
610 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual  
611 models from natural language supervision. In *International conference on machine learning*, pp.  
612 8748–8763. PMLR, 2021.

613 Ben Recht, Christopher Re, Joel Tropp, and Victor Bittorf. Factoring nonnegative matrices with  
614 linear programs. *Advances in neural information processing systems*, 25, 2012.

615 Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *The Journal of Machine*  
616 *Learning Research*, 5:101–141, 2004.

617 Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. A mathematical exploration of why language  
618 models help solve downstream tasks. In *International Conference on Learning Representations*,  
619 2021. URL <https://openreview.net/forum?id=vVjIW3sEcls>.

620 Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi  
621 Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An  
622 open large-scale dataset for training next generation image-text models. *Advances in Neural*  
623 *Information Processing Systems*, 35:25278–25294, 2022.

624 Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what  
625 you want to forget: Algorithms for machine unlearning, 2021. URL [https://arxiv.org/](https://arxiv.org/abs/2103.03279)  
626 [abs/2103.03279](https://arxiv.org/abs/2103.03279).

627 Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao  
628 Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way  
629 evaluation for language models, 2024. URL <https://arxiv.org/abs/2407.06460>.

630 Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur,  
631 Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha,  
632 Sachin Kumar, Li Lucy, Xinxu Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas  
633 Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle  
634 Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh,  
635 Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge,  
636 and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining re-  
637 search. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd*  
638 *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,  
639 pp. 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguis-  
640 tics. doi: 10.18653/v1/2024.acl-long.840. URL [https://aclanthology.org/2024.](https://aclanthology.org/2024.acl-long.840)  
641 [acl-long.840](https://aclanthology.org/2024.acl-long.840).

642 Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion  
643 art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the*  
644 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048–6058, 2023.

- 645 Anvith Thudi, Hengrui Jia, Ilia Shumailov, and Nicolas Papernot. On the necessity of auditable  
646 algorithmic definitions for machine unlearning. In *31st USENIX Security Symposium (USENIX  
647 Security 22)*, pp. 4007–4022, Boston, MA, August 2022. USENIX Association. ISBN 978-1-  
648 939133-31-1. URL [https://www.usenix.org/conference/usenixsecurity22/  
649 presentation/thudi](https://www.usenix.org/conference/usenixsecurity22/presentation/thudi).
- 650 Enayat Ullah, Tung Mai, Anup Rao, Ryan A Rossi, and Raman Arora. Machine unlearning via  
651 algorithmic stability. In *Conference on Learning Theory*, pp. 4126–4142. PMLR, 2021.
- 652 Colin Wei, Sang Michael Xie, and Tengyu Ma. Why do pretrained language models help in down-  
653 stream tasks? an analysis of head and prompt tuning. *Advances in Neural Information Processing  
654 Systems*, 34:16158–16170, 2021.
- 655 Chenwei Wu, Holden Lee, and Rong Ge. Connecting pre-trained language model and downstream  
656 task via properties of representation. In *Thirty-seventh Conference on Neural Information Pro-  
657 cessing Systems*, 2023. URL <https://openreview.net/forum?id=YLOJ4aKAka>.
- 658 Oğuz Kaan Yüksel, Etienne Boursier, and Nicolas Flammarion. First-order ANIL provably learns  
659 representations despite overparametrisation. In *The Twelfth International Conference on Learning  
660 Representations*, 2024. URL <https://openreview.net/forum?id=if2vRbS8Ew>.
- 661 Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language  
662 image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer  
663 Vision*, pp. 11975–11986, 2023.
- 664 Binchi Zhang, Yushun Dong, Tianhao Wang, and Jundong Li. Towards certified unlearning for deep  
665 neural networks, 2024a. URL <https://arxiv.org/abs/2408.00920>.
- 666 Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catas-  
667 trophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024b. URL  
668 <https://openreview.net/forum?id=MXLBXjQkmb>.

## 669 A Precise Description of $\mathcal{A}_{\text{base}}$

### 670 A.1 Complete Description

---

**Algorithm 3** High level learning algorithm ( $\mathcal{A}$ )

---

**Input:** document corpus  $S = \{d_i\}_{i=1}^m$ , anchor word tolerance  $\epsilon_0$   
**Output:** matrices  $\mathbf{A}, \mathbf{R}$   
 $\mathbf{Q}$  = word co-occurrences  
 $\bar{\mathbf{Q}}$  = row-normalized  $\mathbf{Q}$   
 $P = \text{RecoverAnchors}(\{\bar{\mathbf{Q}}_1, \dots, \bar{\mathbf{Q}}_n\})$   
 $\mathbf{A}, \mathbf{R} = \text{RecoverTopics}(\mathbf{Q}, S)$   
**return**  $\mathbf{A}, \mathbf{R}$

---



---

**Algorithm 4** RecoverAnchors, same as Arora et al. (2012a)

---

**Input:** Row-normalized co-occurrence matrix  $\bar{\mathbf{Q}}$  and  $\epsilon_0$  tolerance parameter  
**Output:**  $r$  points of this perturbed simplex close to the vertices of the actual simplex  
 Project the rows to a randomly chosen  $4 \log n / \epsilon_0^2$  dimensional subspace  
 $S \leftarrow \{\bar{\mathbf{Q}}_i\}$  where  $\bar{\mathbf{Q}}_i$  is the furthest point from the origin  
**for**  $i$  in  $1, \dots, r-1$  **do**  
     Let  $\bar{\mathbf{Q}}_j$  be the row of  $\bar{\mathbf{Q}}$  with largest distance to  $\text{span}(S)$   
      $S \leftarrow S \cup \{\bar{\mathbf{Q}}_j\}$   
**end for**  $S = \{\bar{\mathbf{Q}}_{s_1}, \dots, \bar{\mathbf{Q}}_{s_r}\}$   
**for**  $i$  in  $1, \dots, r$  **do**  
     Let  $\bar{\mathbf{Q}}_j$  be the point that has largest distance to  $\text{span}(S \setminus \{\bar{\mathbf{Q}}_{s_i}\})$   
     Remove  $\bar{\mathbf{Q}}_{s_i}$  from  $S$  and insert  $\bar{\mathbf{Q}}_j$  into  $S$   
**end for**  
**return**  $S$

---



---

**Algorithm 5** Recover Topics, from Arora et al. (2012a)

---

**Input:** Co-occurrence matrix  $\mathbf{Q}$ , anchor words  $P = \{s_1, \dots, s_k\}$ , tolerance parameter  $\epsilon_0$   
**Output:** Matrices  $\mathbf{A}, \mathbf{R}$   
 $\bar{\mathbf{Q}}$  = row normalized  $\mathbf{Q}$   
 Store the normalization constants  $\mathbf{p} = \mathbf{Q}\mathbf{1}$   
**for**  $i$  in  $1, \dots, n$  **do**  
     Solve  $C_i = \arg \min_{v \in \Delta_r} \|\bar{\mathbf{Q}}_i - v^\top \bar{\mathbf{Q}}_P\|^2$   
     up to  $\epsilon_0$  accuracy  
**end for**  
 $\mathbf{A}' = \text{diag}(\mathbf{p})\mathbf{C}$   
 $\mathbf{A}$  = column-sum-one normalized  $\mathbf{A}'$   
 $\mathbf{R} = \mathbf{A}^\dagger \mathbf{Q} \mathbf{A}^{\dagger\top}$  where  $\mathbf{A}^\dagger$  is the pseudoinverse of  $\mathbf{A}$   
**return**  $\mathbf{A}, \mathbf{R}$

---

671 More formally, the co-occurrence matrix is constructed as follows. For each document, let  $H_d \in \mathbb{R}^n$   
 672 be the frequency vector of each word in the document; the sum of its entries should be  $L$ . Then, for  
 673 a document  $d$ , consider the matrix

$$\mathbf{G}_d := \tilde{\mathbf{H}}_d \tilde{\mathbf{H}}_d^\top - \hat{\mathbf{H}}_d \quad (9)$$

674 where

$$\tilde{\mathbf{H}}_d := \frac{\mathbf{H}_d}{\sqrt{L(L-1)}} \quad (10)$$

$$\hat{\mathbf{H}}_d := \frac{\text{diag}(\mathbf{H}_d)}{L(L-1)} \quad (11)$$

675 In particular, the denominator term  $L(L - 1)$  is precisely the number of co-occurrences in each  
 676 document, by simple combinatorics, and it can be seen that the sum of the entries of  $\mathbf{G}_d$  is always  
 677 1. Our co-occurrence matrix  $\mathbf{Q}$  is defined to be

$$\mathbf{Q} := \frac{1}{m} \sum_{i=1}^m \mathbf{G}_d \quad (12)$$

678 so that  $\mathbf{Q}$  also has entries that sum to 1. By linearity of expectation, we have

$$\mathbb{E}[\mathbf{Q}] = \mathbb{E}[\mathbf{G}_d] = \mathbf{A}^* \mathbb{E}[\mathbf{X}_d \mathbf{X}_d^\top] \mathbf{A}^{*\top} \quad (13)$$

679 which implies that as the number of documents increases,  $\mathbf{Q}$  concentrates around  $\mathbf{A} \mathbb{E}[\mathbf{X} \mathbf{X}^\top] \mathbf{A}^\top =$   
 680  $\mathbb{E}[\mathbf{M} \mathbf{M}^\top]$ . Therefore, we should expect  $\mathbf{A}^\dagger \mathbf{Q} \mathbf{A}^{\dagger\top}$  to concentrate around  $\mathbb{E}[\mathbf{X} \mathbf{X}^\top] = \mathbf{R}^*$ .

## 681 A.2 Sketch: Population Analysis

682 To understand this algorithm, consider the setting where we have infinitely many documents. Specif-  
 683 ically, consider two words  $w_1, w_2$  in a document and their respective topics  $z_1, z_2$ . Then, this  
 684 population co-occurrence matrix  $\mathbf{Q}$  will have elements  $Q_{i,j} = \Pr[w_1 = i, w_2 = j]$ , and the row-  
 685 normalized co-occurrence matrix  $\bar{\mathbf{Q}}$  will have entries  $\bar{Q}_{i,j} = \Pr[w_2 = j | w_1 = i]$ . Moreover, we  
 686 have that  $A_{i,k} = \Pr[w_1 = i | z_1 = k] = \Pr[w_2 = i | z_2 = k]$ .

687 Consider the set of anchor words  $P = \{s_1, \dots, s_r\} \subseteq [n]$ , where  $s_k$  is the anchor word for topic  $k$ .  
 688 Then, observe that for an anchor word row  $s_k$  of  $\mathbf{Q}$ , it holds that

$$\bar{Q}_{s_k,j} = \Pr[w_2 = j | w_1 = s_k] = \sum_{k'} \Pr[z_1 = k' | w_1 = s_k] \Pr[w_2 = j | w_1 = s_k, z_1 = k'] \quad (14)$$

$$= \Pr[w_2 = j | w_1 = s_k, z_1 = k] \quad (15)$$

$$= \Pr[w_2 = j | z_1 = k] \quad (16)$$

689 where the second line follows from only  $\Pr[z_1 = k | w_1 = s_k] = 1$  in the summation, and the last  
 690 line follows from  $w_2, w_1$  are conditionally independent given  $z_1$ . Furthermore, for non-anchor word  
 691 rows  $i$  of  $\bar{\mathbf{Q}}$ , it holds that

$$\bar{Q}_{i,j} = \sum_k \Pr[z_1 = k | w_1 = i] \Pr[w_2 = j | z_1 = k] \quad (17)$$

692 where again we use that  $w_2, w_1$  are conditionally independent  $z_1$ . For a word  $i$ , let  $\mathbf{C}_i \in \mathbb{R}^r$  be  
 693 the vector such that  $\mathbf{C}_{i,k} := \Pr[z_1 = k | w_1 = i]$ . Then, it holds that  $\bar{\mathbf{Q}}_i = \mathbf{C}_i^\top \bar{\mathbf{Q}}_S$ , where  $\bar{\mathbf{Q}}_S$   
 694 is the submatrix of  $\bar{\mathbf{Q}}$  constrained to the anchor word rows. In other words, for every word  $i$ ,  $\bar{\mathbf{Q}}_i$  is a  
 695 convex combination of rows of  $\bar{\mathbf{Q}}_S$ .

696 In the algorithm, one can see that  $\mathbf{A}'_{i,k} = \mathbf{C}_{i,k} \mathbf{p}_i$ . Normalizing this along each column, we obtain

$$\mathbf{A}_{i,k} = \frac{\mathbf{C}_{i,k} \mathbf{p}_i}{\sum_{i'} \mathbf{C}_{i',k} \mathbf{p}_{i'}} = \frac{\Pr[z_1 = k | w_1 = i] \Pr[w_1 = i]}{\sum_{i'} \Pr[z_1 = k | w_1 = i'] \Pr[w_1 = i']} = \Pr[w_1 = i | z_1 = k] \quad (18)$$

697 Hence, in the infinite document limit, this algorithm recovers the ground truth  $\mathbf{A}^*, \mathbf{R}^*$ .

## 698 B From properties of the learning algorithm to the proof of Theorem 2

699 We first give the formal statement of Theorem 2.

700 **Theorem 5** (Formal statement of Theorem 2). *Let  $\mathcal{A}_{base}$  be the learning algorithm described in the*  
 701 *prior sections and  $\mathcal{U}_{base}$  be the unlearning algorithm in Algorithm 1. Then,  $(\mathcal{A}_{base}, \mathcal{U}_{base})$  performs*  
 702 *utility-preserving unlearning with deletion capacity*

$$T_{\epsilon, \delta}^{\mathcal{A}_{base}, \mathcal{U}_{base}}(m) \geq c \cdot \min \left\{ \frac{m\epsilon}{r^2 \sqrt{rn \log 1/\delta}}, \frac{0.001m}{r^2} \right\} \quad (19)$$

703 where  $m$  is the number of training documents,  $r$  is the number of topics, and  $c$  is a constant depen-  
 704 dent on  $\mathcal{D}$ . The loss function  $h$  used in the utility-preserving definition is the maximum entrywise  
 705 error from the ground truth topic model  $\mathbf{A}^*$ .

## B.1 Preliminaries

When the norm is not specified, we assume that it is the Euclidean norm  $\|\cdot\|_2$ . We now start off with a technical assumption on the precision of the learning algorithm.

**Assumption 2.**  $\epsilon_0 \leq O(1/\sqrt{nr})$ .

**Assumption 3.** Every word appears with probability  $\epsilon_0/4ar$  without loss of generality; see discussion in Arora et al. (2012b). Essentially, less probable words can be combined in a sense to form a single category of "rare" words.

We recall the definitions from Arora et al. (2012a).

**Definition 6** ( $\beta$ -robust simplex). A simplex  $P$  is  $\beta$ -robust if for every vertex  $v$  of  $P$ , the  $\ell_2$  distance between  $v$  and the convex hull of the rest of the vertices is at least  $\beta$ .

**Definition 7.** Let  $\{a_i\}_{i=1}^n$  be a set of points whose convex hull is a simplex with vertices  $\{v_i\}_{i=1}^r$ . We say a set of  $r$  points is  $\epsilon$ -close to the vertex set  $\{v_i\}_{i=1}^r$  if each of the  $r$  points is  $\epsilon$ -close in  $\ell_2$  distance to a different vertex in this vertex set.

The following result will be used throughout our proof.

**Proposition 1** (Arora et al. (2012b)).  $\bar{Q}_P^*$  in population is  $\gamma p$ -robust.

We now list the high probability events we condition on throughout our proof. These follow from previous results in Arora et al. (2012a); they concern the properties of the output of the learning algorithm.

**Proposition 2.** With high probability, in our regime of  $m$ , the following hold:

- The correct anchor words are selected.
- Each word appears at least  $O(\frac{m\epsilon_0}{4ar})$  times.
- The error in the empirical matrix  $\hat{Q}$  is entrywise at most  $\tilde{O}(1/\sqrt{m})$  from the population  $Q^*$ .

We also utilize the following two key lemmas from Arora et al. (2012a) that we touched upon in the main paper.

**Lemma 7** (Approximation Guarantee on Anchor Words). Suppose each row of  $\bar{Q}$  is at most  $\delta$  distance away from the ground truth  $\gamma p$ -robust simplex  $Q^*$  in  $\ell_2$  norm. If  $20r\delta/(\gamma p)^2 < \gamma p$ , then the set of anchor words found by the algorithm is  $O(\delta/\gamma p)$ -close to the ground truth anchor words.

**Lemma 8.** When  $20r\delta/(\gamma p)^2 < \gamma p$ , it holds for every word  $i$  that  $C_i$  has entrywise error  $O(\delta/(\gamma p)^2)$  from  $C_i^*$ .

## B.2 Proof of Theorem 2

The following are lemmas bounding the relation between  $\bar{Q}_i^S, \bar{Q}_i^F, \bar{Q}_i^*$ .

**Lemma 9.** After training, the error of each row of  $\bar{Q}^S$  is at most  $\delta_2 := O(\sqrt{\frac{4ar}{m\epsilon_0}})$ . That is,  $\|\bar{Q}_i^S - \bar{Q}_i^*\| \leq \delta_2$  for all words  $i$ .

Importantly, note that

$$20r\delta_2/(\gamma p)^2 < \gamma p \quad (20)$$

This implies that the anchor words of  $\bar{Q}_i^S$  are  $O(\delta_2/(\gamma p))$  close to the anchor words of  $\bar{Q}_i^*$ .

Consequently, it holds that

$$\|C^S - C^*\|_\infty \leq O(\delta_2/(\gamma p)^2) \quad (21)$$

*Proof.* The first part follows directly from the fact that if the number of documents  $m = \tilde{\Omega}(1/\epsilon_Q^2)$ , then  $\|\bar{Q}_i^S - \bar{Q}_i^*\| \leq \delta_2$  for each row  $i$ . To show that

$$20r\delta_2/(\gamma p)^2 < \gamma p \quad (22)$$

746 we note that by the sample complexity guarantee,

$$m\epsilon_0 \geq \tilde{O}\left(\frac{ar^3}{(\gamma p)^6}\right) \quad (23)$$

747 which implies that

$$\delta_2 \leq \tilde{O}\left(\frac{(\gamma p)^3}{r}\right) \quad (24)$$

748 as desired.  $\square$

749 **Lemma 10.** *When we delete  $m_U \leq \frac{0.001m\epsilon_0(\gamma p)^3}{a^2r^2}$ , it holds that*

$$\|\bar{\mathbf{Q}}_i^F - \bar{\mathbf{Q}}_i^S\| \leq \frac{m_U}{m\epsilon_0/4ar} = \frac{4arm_U}{m\epsilon_0} \quad (25)$$

750 *In particular, this is smaller than*

$$\frac{0.001m\epsilon_0(\gamma p)^3}{a^2r^2} \cdot \frac{1}{m\epsilon_0/4ar} = \frac{0.004(\gamma p)^3}{ar} \quad (26)$$

751 *Proof.* For a word  $i$ , consider the change in  $\bar{\mathbf{Q}}_i$  after deletion requests. Let  $F$  be the initial sum of  
752 the the  $i$ th row of  $\mathbf{Q}$ . Each coordinate  $j \in [n]$  will change as follows:

$$\delta_j = \frac{f_j - t_j}{F - m_U} - \frac{f_j}{F} = \frac{m_U f_j - F t_j}{F(F - m_U)} \quad (27)$$

753 where  $f_j$  is the initial number of cooccurrences of words  $i, j$  and  $t_j$  is the number of documents  
754 removed that have this cooccurrence. Moreover,  $F$  is the number of initial occurrences of word  $i$ , and  
755  $T$  is the number of deletions of the word  $i$ . From the previous lemma, it holds that  $F \geq m\epsilon_0/4ar$ ,  
756 and that  $m_U \geq \sum_{j=1}^n t_j$ . Hence, it follows that the squared Euclidean norm of the change is:

$$\sum_{j=1}^n \delta_j^2 = \frac{1}{F^2(F - T)^2} \sum_{j=1}^n (m_U f_j - F t_j)^2 \leq \frac{2F^2 m_U^2}{F^2(F - m_U)^2} \leq 2 \left( \frac{m_U}{F - m_U} \right)^2 \quad (28)$$

757 Hence, for the regime where  $m_U \leq \frac{0.001m\epsilon_0(\gamma p)^3}{a^2r^2}$ , we have

$$\|\bar{\mathbf{Q}}_i^S - \bar{\mathbf{Q}}_i^F\| \leq \sqrt{2} \frac{m_U}{F - m_U} \lesssim \frac{m_U}{F} \lesssim \frac{4arm_U}{m\epsilon_0} \quad (29)$$

758 Of particular notice is that when  $m_U$  is taken as large as possible, this is at most

$$\frac{0.001m\epsilon_0(\gamma p)^3/a^2r^2}{m\epsilon_0/4ar} = 0.004(\gamma p)^3/ar \quad (30)$$

759  $\square$

760 We now combine the above two with triangle inequality.

761 **Lemma 11.** *Hence, it holds that*

$$\|\bar{\mathbf{Q}}_i^F - \bar{\mathbf{Q}}_i^*\| \leq \frac{4arm_U}{m\epsilon_0} + \delta_2 = \frac{4arm_U}{m\epsilon_0} + O\left(\sqrt{\frac{4ar}{m\epsilon_0}}\right) =: \delta'_2 \quad (31)$$

762 *Importantly, note that*

$$20r\delta'_2/(\gamma p)^2 < \gamma p \quad (32)$$

763 *This implies that the anchor words of  $\bar{\mathbf{Q}}_i^F$  are  $O(\delta'_2/(\gamma p))$  close to the anchor words of  $\bar{\mathbf{Q}}_i^*$ .*

764

765 *Consequently, it holds that*

$$\|\mathbf{C}^F - \mathbf{C}^*\|_\infty \leq O(\delta'_2/(\gamma p)^2) \quad (33)$$

766 *Proof.* The first part follows from triangle inequality, a □

767 We now bound what happens to  $\|C^F - C^S\|_\infty$ . First, we have that the perturbed simplex  $\bar{Q}_P^S$  is  
768  $\gamma p/2$ -robust.

769 **Lemma 12.** *The perturbed simplex  $\bar{Q}_P^S$  is  $\gamma p/2$ -robust.*

770 *Proof.* This is because of Lemma A.1 in Arora et al. (2012a). Since  $10\sqrt{r}\delta_2 < \gamma p$ , the result of that  
771 lemma applies. □

772 Hence, we will apply Lemma B.1 from Arora et al. (2012a) on  $C^S$  to say something about  $\|C^F -$   
773  $C^S\|_\infty$ .

774 **Lemma 13.** *Recall that when we delete  $m_U \leq \frac{0.001m\epsilon_0(\gamma p)^3}{a^2r^2}$ , it holds that*

$$\|\bar{Q}_i^F - \bar{Q}_i^S\| \leq \frac{m_U}{m\epsilon_0/4ar} = \frac{4arm_U}{m\epsilon_0} \quad (34)$$

775 *Importantly, note that*

$$20r\left(\frac{4arm_U}{m\epsilon_0}\right)/(\gamma p/2)^2 < \gamma p/2 \quad (35)$$

776 *This implies that the anchor words of  $\bar{Q}_i^F$  are  $\frac{4arm_U/m\epsilon_0}{\gamma p/2}$  close to the anchor words of  $\bar{Q}_i^S$ . By*  
777 *lemma B.1 from Arora et al. (2012a), it holds that*

$$\|C^F - C^S\|_\infty \leq O\left(\frac{4arm_U}{m\epsilon_0}/(\gamma p/2)^2\right) \quad (36)$$

778 *Observe that this is smaller than  $O((\gamma p)/ar)$ .*

779 We now deal with the Hessian step that we had took to prevent retraining the  $C_i$ 's. In particular, we  
780 will denote  $\bar{C}$  to be our estimated new  $C$ .

781 First, a lemma to say that our Hessian step is full rank and has a lower bound on its minimum  
782 singular value.

783 **Lemma 14.** *When we delete  $m_U \leq \frac{0.001m\epsilon_0(\gamma p)^3}{a^2r^2}$  samples, it holds that the minimum eigenvalue of*  
784  $\bar{Q}_P^F \bar{Q}_P^F$  *is at least  $\gamma p/2$ .*

785 *Proof.* Follows from Lemma A.3 in Arora et al. (2012a). □

786 **Lemma 15.** *When we delete  $m_U \leq \frac{0.001m\epsilon_0(\gamma p)^3}{a^2r^2}$  samples, it holds for all  $i$ ,*

$$\|C_i^F - \bar{C}_i^F\| \leq \frac{4}{\gamma p} \left( \delta_2 + \frac{4arm_U}{m\epsilon_0} \right) \quad (37)$$

787 *Proof.* For the case of  $d(\cdot, \cdot)$  being the squared loss, we will denote the following:

$$C_{i,\text{uncon}} := \arg \min_C \|\bar{Q}_P^{F\top} C - \bar{Q}_i^{F\top}\|^2 = (\bar{Q}_P^F \bar{Q}_P^{F\top})^{-1} \bar{Q}_P^F \bar{Q}_i^{F\top} \quad (38)$$

$$\bar{C}_i^F := \text{proj}_{\Delta_r}(C_{i,\text{uncon}}) \quad (39)$$

$$C_i^F := \arg \min_{C \in \Delta_r} \|\bar{Q}_P^{F\top} C - \bar{Q}_i^{F\top}\|^2 \quad (40)$$

788 In particular, the Newton step plus projection outputs  $C_{i,\text{proj}}$ . First, observe that by one of the  
789 anchor word lemmas,

$$\min_C \|\bar{Q}_P^{F\top} C - \bar{Q}_i^{F\top}\| = \|\bar{Q}_P^{F\top} C_{i,\text{uncon}} - \bar{Q}_i^{F\top}\| \leq \|\bar{Q}_P^{F\top} C_i^F - \bar{Q}_i^{F\top}\| \leq \delta_2 + \frac{4arm_U}{m\epsilon_0} \quad (41)$$

790 The last inequality follows from the fact that  $\bar{\mathbf{Q}}_P^F$  is a perturbed version of  $\bar{\mathbf{Q}}_P^S$ , and  $\bar{\mathbf{Q}}_P^S$  is a perturbed  
 791 version of  $\bar{\mathbf{Q}}_P^*$ . Hence, we will bound

$$\|\bar{\mathbf{C}}_i^F - \mathbf{C}_i^F\| = \|\text{proj}_{\Delta_r}(\mathbf{C}_{i,\text{uncon}}) - \text{proj}_{\Delta_r}(\mathbf{C}_i^F)\| \quad (42)$$

$$\leq \|\mathbf{C}_{i,\text{uncon}} - \mathbf{C}_i^F\| \quad (43)$$

$$\leq \frac{1}{\sigma_{\min}} \|\bar{\mathbf{Q}}_P^{F\top}(\mathbf{C}_{i,\text{uncon}} - \mathbf{C}_i^F)\| \quad (44)$$

$$\leq \frac{1}{\sigma_{\min}} (\|\bar{\mathbf{Q}}_i^{F\top} - \bar{\mathbf{Q}}_P^{F\top} \mathbf{C}_i^F\| + \|\bar{\mathbf{Q}}_P^{F\top} \mathbf{C}_{i,\text{uncon}} - \bar{\mathbf{Q}}_i^{F\top}\|) \quad (45)$$

$$\leq \frac{2}{\sigma_{\min}} \left( \delta_2 + \frac{4arm_U}{m\epsilon_0} \right) \quad (46)$$

792 where  $\sigma_{\min}$  is the smallest singular value of  $\bar{\mathbf{Q}}_i^{F\top}$ , which is guaranteed to be full rank per the  
 793 previous lemma. Due to a result in Arora et al. (2012a), this  $\sigma_{\min} \geq (\gamma p)/2$ . This gives us that the  
 794 whole thing is at most

$$\frac{4}{\gamma p} \left( \delta_2 + \frac{4arm_U}{m\epsilon_0} \right) \quad (47)$$

795

□

796 **Corollary 1.** *We have that*

$$\|\mathbf{C}^F - \bar{\mathbf{C}}^F\|_{\infty} \leq \frac{4}{\gamma p} \left( \delta_2 + \frac{4arm_U}{m\epsilon_0} \right) \quad (48)$$

797 since the  $\ell_{\infty}$  norm is upper bounded by the  $\ell_2$  norm.

798 **Lemma 16.** *The following are true.*

$$\begin{aligned} 799 & \bullet \|\mathbf{C}^F - \bar{\mathbf{C}}^F\|_{\infty} \leq \frac{4}{\gamma p} \left( \delta_2 + \frac{4arm_U}{m\epsilon_0} \right) \\ 800 & \bullet \|\bar{\mathbf{C}}^F - \mathbf{C}^{\star}\|_{\infty} \leq \|\bar{\mathbf{C}}^F - \mathbf{C}^F\|_{\infty} + \|\mathbf{C}^F - \mathbf{C}^{\star}\|_{\infty} \leq \frac{4}{\gamma p} \left( \delta_2 + \frac{4arm_U}{m\epsilon_0} \right) + O(\delta'_2/(\gamma p)^2) \end{aligned}$$

801 From this, we can bound the errors on the topic matrix.

802 **Lemma 17.** *The following are true.*

$$\begin{aligned} 803 & \bullet \|\mathbf{A}^F - \bar{\mathbf{A}}\|_{\infty} \leq O(ar\|\mathbf{C}^F - \bar{\mathbf{C}}^F\|_{\infty}) \\ 804 & \bullet \|\bar{\mathbf{A}} - \mathbf{A}^{\star}\|_{\infty} \leq O(ar\|\bar{\mathbf{C}}^F - \mathbf{C}^{\star}\|_{\infty}) \\ 805 & \bullet \|\mathbf{A}^S - \mathbf{A}^F\|_{\infty} \leq O(ar\|\mathbf{C}^F - \mathbf{C}^S\|_{\infty}) \end{aligned}$$

806 *Proof.* Note that entries  $\mathbf{A}_{i,k}$  are

$$\mathbf{A}_{i,k} = \frac{\mathbf{C}_{i,k} \Pr[w = i]}{\Pr[z = k]} \quad (49)$$

807 Therefore, the perturbation in  $\mathbf{A}$  will be the perturbation in  $\mathbf{C}$  multiplied by  $ar$ , since the denomi-  
 808 nator is lower bounded by  $1/ar$  due to the topic imbalance constant. □

809 Now, we give a new lemma.

810 **Proposition 3.** *When  $m_U \geq \Omega(\sqrt{\frac{m\epsilon_0}{4ar}})$ , we have that*

$$\delta'_2 = \delta_2 + \frac{4arm_U}{m\epsilon_0} = \sqrt{\frac{4ar}{m\epsilon_0}} + \frac{4arm_U}{m\epsilon_0} \leq O\left(\frac{arm_U}{m\epsilon_0}\right) \quad (50)$$

811 Now, we analyze what happens given that  $\Omega(\sqrt{\frac{m\epsilon_0}{4ar}}) \leq m_U \leq \frac{0.001m\epsilon_0(\gamma p)^3}{a^2 r^2}$ .

812 **Lemma 18.** For  $\epsilon, \delta > 0$ , the deletion capacity satisfies

$$T_{\epsilon, \delta}^{\mathcal{A}, \mathcal{U}}(m) \geq \tilde{\Omega}\left(\frac{m}{r^2 \sqrt{nr}}\right) \quad (51)$$

813 *Proof.* Recall that

$$\|\bar{\mathbf{A}} - \mathbf{A}^*\|_\infty \leq O(ar\delta'_2(1/\gamma p + 1/(\gamma p)^2)) \leq O\left(\frac{(ar)^2 m_U}{m\epsilon_0 \gamma p}\right) \quad (52)$$

814 Moreover, we also have that

$$\|\bar{\mathbf{A}} - \mathbf{A}^F\|_\infty \leq O(ar\|\mathbf{C}^F - \bar{\mathbf{C}}^F\|_\infty) \quad (53)$$

$$\leq O\left(\frac{4ar\delta'_2}{\gamma p}\right) \quad (54)$$

$$\leq O\left(\frac{(ar)^2 m_U}{m\epsilon_0 \gamma p}\right) \quad (55)$$

815 Note that  $\mathbf{A}$  has  $\ell_2$  sensitivity  $O\left(\sqrt{nr} \frac{(ar)^2 m_U}{m\epsilon_0 \gamma p}\right)$ . We now apply the Gaussian mechanism to the  
816 matrix  $\mathbf{A}$  entrywise with noise

$$\sigma = \frac{O\left(\sqrt{nr} \frac{(ar)^2 m_U}{m\epsilon_0 \gamma p}\right)}{\epsilon} \sqrt{2 \log(1.25/\delta)} \quad (56)$$

817 From this, we obtain that

$$\mathbb{E}[\|\tilde{\mathbf{A}} - \mathbf{A}^*\|_\infty] \leq \mathbb{E}\left[\max_{i,k} |\nu_{i,k}|\right] + \mathbb{E}[\|\bar{\mathbf{A}} - \mathbf{A}^*\|_\infty] \quad (57)$$

$$\leq O\left(\sqrt{nr} \cdot \frac{(ar)^2 m_U}{m\epsilon_0 \gamma p} \cdot \sqrt{\log(nr)} \cdot \frac{\sqrt{\log(1/\delta)}}{\epsilon}\right) + O\left(\frac{(ar)^2 m_U}{m\epsilon_0 \gamma p}\right) \quad (58)$$

818 Finally, this says that when

$$m_U \leq \tilde{\Omega}\left(\frac{m}{r^2 \sqrt{nr}}\right) \quad (59)$$

819 we have that the utility is preserved up to constant amount, say 0.01.  $\square$

820 This proves Theorem 2. It is straightforward to continue the perturbation analysis for the topic-topic  
821 covariance matrix  $\mathbf{R}^*$  and prove similar deletion capacity rates.

## 822 C Downstream task proofs

Recall the algorithm for learning the downstream task head.

---

**Algorithm 6** Learning algorithm for task  $\mathcal{T}(\mathcal{A}_{head})$

---

**Input:** document corpus  $S = \{d_i\}_{i=1}^m$ , anchor word tolerance  $\epsilon_0$   
 $\mathbf{A}, \mathbf{R} = \mathcal{A}_{base}(S)$   
**return**  $\arg \min_{\mathbf{w} \in \mathcal{W}_{head}} \ell_{\mathcal{T}}(\mathbf{w}; \mathbf{A})$

---

823

824 **Assumption 4.** For any  $\mathbf{A}$ ,  $\ell_{\mathcal{T}}$  is  $\lambda$ -strongly convex with respect to  $\mathbf{w}$ .

825 Since our topic matrix  $\mathbf{A}$ , can only take on a bounded support (i.e. the set of matrices where each  
826 row is on the probability simplex), it is natural to say that the set of values  $\mathbf{w}^*(\mathbf{A})$  takes on over all  
827 topic matrices  $\mathbf{A}$  is bounded in a certain sense. As such, we also assume the following:

828 **Assumption 5.** For any base model  $\mathbf{A}$ , the vector  $\mathbf{v}$  such that  $\mathbf{v} = \arg \min_{\mathbf{w}} \ell_{\mathcal{T}}(\mathbf{w}; \mathbf{A})$  satisfies  
829  $\|\mathbf{v}\|_2 \leq B$ .

830 **Assumption 6.** For any  $\mathbf{A}$ ,  $\ell_{\mathcal{T}}$  is  $L$ -Lipschitz with respect to  $\mathbf{w}$  and the  $\ell_2$  norm, and is  $L_2$ -Hessian  
831 Lipschitz with respect to  $\mathbf{w}$  and the  $\ell_2$  norm. In other words,

$$\|\ell_{\mathcal{T}}(\mathbf{A}, \mathbf{w}_1) - \ell_{\mathcal{T}}(\mathbf{A}, \mathbf{w}_2)\|_2 \leq L\|\mathbf{w}_1 - \mathbf{w}_2\|_2 \quad (60)$$

$$\|\nabla_{\mathbf{w}}^2 \ell_{\mathcal{T}}(\mathbf{A}, \mathbf{w}_1) - \nabla_{\mathbf{w}}^2 \ell_{\mathcal{T}}(\mathbf{A}, \mathbf{w}_2)\|_2 \leq L_2\|\mathbf{w}_1 - \mathbf{w}_2\|_2 \quad (61)$$

832 **Assumption 7.** For any  $\mathbf{w}$ ,  $\nabla_{\mathbf{w}} \ell_{\mathcal{T}}$  is  $L_{\infty}$ -Lipschitz with respect to  $\mathbf{A}$  and the  $\ell_{\infty}$  norm; that is,

$$\|\nabla_{\mathbf{w}} \ell_{\mathcal{T}}(\mathbf{A}, \mathbf{w}) - \nabla_{\mathbf{w}} \ell_{\mathcal{T}}(\tilde{\mathbf{A}}, \mathbf{w})\|_2 \leq L_{\infty}\|\mathbf{A} - \tilde{\mathbf{A}}\|_{\infty} \quad (62)$$

$$(63)$$

833 We give a helper lemma that  $(\epsilon, \delta)$ -indistinguishability is immune to post processing.

834 **Lemma 19** (Post-processing immunity). Consider two random variables  $\theta_1, \theta_2 \in \Theta$  that are  $(\epsilon, \delta)$ -  
835 indistinguishable. Then, for any arbitrary mapping  $f : \Theta \rightarrow \Theta'$ , it holds that  $f(\theta_1), f(\theta_2) \in \Theta'$  are  
836  $(\epsilon, \delta)$ -indistinguishable.

837 *Proof.* Consider an arbitrary set  $T' \subseteq \Theta'$ ; let  $T = \{r \in \Theta : f(r) \in T'\}$ . Then, it holds that

$$\Pr[f(\theta_1) \in T'] = \Pr[\theta_1 \in T] \quad (64)$$

$$\leq e^{\epsilon} \Pr[\theta_2 \in T] + \delta \quad (65)$$

$$= e^{\epsilon} \Pr[f(\theta_2) \in T'] + \delta \quad (66)$$

838 as desired.  $\square$

839 We now give a certifiable unlearning guarantee for the most naive retraining algorithm for the down-  
840 stream task, which we mentioned in the main text as Theorem 3.

841 **Theorem 6** (Unlearning when releasing  $\mathbf{A}$  and  $\mathbf{w}$ ). For a downstream task  $\mathcal{T}$  with loss func-  
842 tion  $\ell_{\mathcal{T}}$ , consider the unlearning algorithm  $\mathcal{U}_{\text{head, naive}}$  that first runs Algorithm 1 to compute  
843  $\tilde{\mathbf{A}} = \mathcal{U}_{\text{base}}(S_f, \mathcal{A}_{\text{base}}(S), T(S))$ , where  $(\mathcal{A}_{\text{base}}, \mathcal{U}_{\text{base}})$  perform utility-preserving unlearning (The-  
844 orem 2). Then, it fits a head  $\mathbf{w} = \arg \min_{\mathbf{w} \in \mathcal{W}_{\text{head}}} \ell_{\mathcal{T}}(\mathbf{w}; \tilde{\mathbf{A}})$  and returns  $\tilde{\mathbf{A}}$  and  $\mathbf{w}$ . We assert that  
845  $(\mathcal{A}_{\text{head, naive}}, \mathcal{U}_{\text{head, naive}})$  performs utility-preserving unlearning (Definition 4).

846 *Proof.* Intuitively, this is a result of post processing. More precisely, consider the  $(\epsilon, \delta)$ -  
847 indistinguishable base models  $\tilde{\mathbf{A}} := \mathcal{U}_{\text{base}}(S_f, \mathcal{A}_{\text{base}}(S), T(S))$  and  $\tilde{\mathbf{A}}' := \mathcal{U}_{\text{base}}(\emptyset, \mathcal{A}_{\text{base}}(S \setminus S_f), T(S \setminus S_f))$ . Then, since the head fitting is a deterministic post-processing of the original  
848 model, this proves the  $(\epsilon, \delta)$ -indistinguishability between the two.  
849

850 To prove the utility preservation, observe that in this setting

$$\mathbb{E}[\|\tilde{\mathbf{A}} - \mathbf{A}^*\|_{\infty}] \leq 0.01 \quad (67)$$

$$(68)$$

851 We thus obtain by Lemma 20

$$\mathbb{E}[\|\mathbf{w}^*(\tilde{\mathbf{A}}) - \mathbf{w}^*(\mathbf{A}^*)\|_{\infty}] \leq \mathbb{E}[\|\mathbf{w}^*(\tilde{\mathbf{A}}) - \mathbf{w}^*(\mathbf{A}^*)\|_2] \quad (69)$$

$$\leq \frac{L_{\infty}}{\lambda} \mathbb{E}[\|\tilde{\mathbf{A}} - \mathbf{A}^*\|_{\infty}] \quad (70)$$

852 which is at most 0.01, up to constant rescaling.  $\square$

853 The above result is nice, and it follows from the fact that the training algorithm of the downstream  
854 task head is just a post-processing. However, a downside is that it still requires retraining of the  
855 downstream task head. We can show something stronger: even without provable unlearning of the  
856 base model ( $\mathbf{A}$  and  $\mathbf{R}$ ), we can achieve provable unlearning of the downstream task head weights  
857 when the downstream task loss is convex in the trainable weights  $\mathbf{w}$ .

858 We will now consider an arbitrary task  $\mathcal{T}$ . We first give the following notation.

859 **Definition 8.** For a base model  $\mathbf{A}$ , let  $\mathbf{w}^*(\mathbf{A}) := \arg \min_{\mathbf{w}} \ell_{\mathcal{T}}(\mathbf{w}; \mathbf{A})$ .

860 First, we give the following helper lemma that will be useful later on.

861 **Lemma 20.** Consider two base models  $\mathbf{A}_1$  and  $\mathbf{A}_2$ . Then, it holds that

$$\|\mathbf{w}^*(\mathbf{A}_1) - \mathbf{w}^*(\mathbf{A}_2)\|_2 \leq \frac{L_{\infty}}{\lambda} \|\mathbf{A}_1 - \mathbf{A}_2\|_{\infty} \quad (71)$$

862 *Proof.* Observe that

$$\lambda \|\mathbf{w}^*(\mathbf{A}_1) - \mathbf{w}^*(\mathbf{A}_2)\|_2 \leq \|\nabla_{\mathbf{w}} \ell_{\mathcal{T}}(\mathbf{w}^*(\mathbf{A}_1); \mathbf{A}_2) - \nabla_{\mathbf{w}} \ell_{\mathcal{T}}(\mathbf{w}^*(\mathbf{A}_2); \mathbf{A}_2)\|_2 \quad (72)$$

$$= \|\nabla_{\mathbf{w}} \ell_{\mathcal{T}}(\mathbf{w}^*(\mathbf{A}_1); \mathbf{A}_2) - \nabla_{\mathbf{w}} \ell_{\mathcal{T}}(\mathbf{w}^*(\mathbf{A}_1); \mathbf{A}_1)\|_2 \quad (73)$$

$$\leq L_{\infty} \|\mathbf{A}_1 - \mathbf{A}_2\|_{\infty} \quad (74)$$

863 where the first line follows from strong convexity, the second line from the gradients being zero,  
864 and the third line from the definition of  $L_{\infty}$  Lipschitz constant. Dividing both sides by  $\lambda$  gives the  
865 desired result.  $\square$

866 We now define the following notations for clarity.

- 867 •  $\mathbf{w}^S := \mathbf{w}^*(\mathbf{A}^S)$
- 868 •  $\mathbf{w}^F := \mathbf{w}^*(\mathbf{A}^F)$
- 869 •  $\bar{\mathbf{w}}^* := \mathbf{w}^*(\bar{\mathbf{A}})$
- 870 •  $\bar{\mathbf{w}} := \mathbf{w}^S - H_{\mathbf{w}^S}^{-1} \nabla_{\mathbf{w}} \ell_{\mathcal{T}}(\mathbf{w}^S; \bar{\mathbf{A}})$ , which is the Newton step we take from  $\mathbf{w}^S$  to approxi-  
871 mate  $\bar{\mathbf{w}}^*$

872 First, we give a bound on the approximation error of the Newton step.

873 **Lemma 21.** It holds that

$$\|\bar{\mathbf{w}} - \bar{\mathbf{w}}^*\| \leq \frac{L_2 L_{\infty}^2}{2\lambda^3} \|\mathbf{A}^S - \bar{\mathbf{A}}\|_{\infty}^2 \quad (75)$$

874 *Proof.* We aim to bound the distance of the Newton step from  $\bar{\mathbf{w}}^*$ :

$$\bar{\mathbf{w}} - \bar{\mathbf{w}}^* = (\mathbf{w}^S - H_{\mathbf{w}^S}^{-1} \nabla_{\mathbf{w}} \ell_{\mathcal{T}}(\bar{\mathbf{A}}, \mathbf{w}^S)) - \bar{\mathbf{w}}^* \quad (76)$$

875 where  $H_{\mathbf{w}^S} = \nabla_{\mathbf{w}}^2 \ell_{\mathcal{T}}(\bar{\mathbf{A}}, \mathbf{w}^S)$ . Then, it holds that

$$\mathbf{w}^S - H_{\mathbf{w}^S}^{-1} \nabla_{\mathbf{w}} \ell_{\mathcal{T}}(\bar{\mathbf{A}}, \mathbf{w}^S) - \bar{\mathbf{w}}^* \quad (77)$$

$$= \mathbf{w}^S - \bar{\mathbf{w}}^* - H_{\mathbf{w}^S}^{-1} (\nabla_{\mathbf{w}} \ell_{\mathcal{T}}(\bar{\mathbf{A}}, \mathbf{w}^S) - \nabla_{\mathbf{w}} \ell_{\mathcal{T}}(\bar{\mathbf{A}}, \bar{\mathbf{w}}^*)) \quad (78)$$

$$= H_{\mathbf{w}^S}^{-1} \left( H_{\mathbf{w}^S} (\mathbf{w}^S - \bar{\mathbf{w}}^*) - \int_0^1 H_{\bar{\mathbf{w}}^* + t(\mathbf{w}^S - \bar{\mathbf{w}}^*)} (\mathbf{w}^S - \bar{\mathbf{w}}^*) dt \right) \quad (79)$$

$$= H_{\mathbf{w}^S}^{-1} \int_0^1 (H_{\mathbf{w}^S} - H_{\bar{\mathbf{w}}^* + t(\mathbf{w}^S - \bar{\mathbf{w}}^*)}) dt \cdot (\mathbf{w}^S - \bar{\mathbf{w}}^*) \quad (80)$$

876 The norm of this quantity is therefore bounded by

$$\|H_{\mathbf{w}^S}^{-1}\|_2 \cdot \frac{L_2}{2} \|\mathbf{w}^S - \bar{\mathbf{w}}^*\| \cdot \|\mathbf{w}^S - \bar{\mathbf{w}}^*\| \quad (81)$$

$$= \frac{L_2}{2\lambda} \|\mathbf{w}^S - \bar{\mathbf{w}}^*\|_2^2 \quad (82)$$

$$\leq \frac{L_2}{2\lambda} \left( \frac{1}{\lambda} \|\nabla \ell_{\mathcal{T}}(\bar{\mathbf{A}}, \mathbf{w}^S) - \nabla \ell_{\mathcal{T}}(\mathbf{A}^S, \mathbf{w}^S)\|_2 \right)^2 \quad (83)$$

$$\leq \frac{L_2}{2\lambda} \left( \frac{L_{\infty}}{\lambda} \|\bar{\mathbf{A}} - \mathbf{A}^S\|_{\infty} \right)^2 \quad (84)$$

877 Hence, we have that

$$\|\bar{\mathbf{w}} - \bar{\mathbf{w}}^*\|_2 \leq \frac{L_2 L_{\infty}^2}{2\lambda^3} \|\mathbf{A}^S - \bar{\mathbf{A}}\|_{\infty}^2 \quad (85)$$

878  $\square$

879 **C.1 Instantiating for  $\mathbb{T}_{\text{clf}} = [r]$**

880 We first instantiate Theorem 4 for the case where  $\mathbb{T}_{\text{clf}} = [r]$ , or equivalently when  $q = 1/ar$ .

881 **Lemma 22.** *Recall our retrained model for the downstream task is  $A^F w^F$ . Then, it holds that*

$$\|\bar{A}\bar{w} - A^F w^F\|_2 \leq O\left(\sqrt{r}\left(\frac{(ar)^2 m_U}{m\epsilon_0 \gamma p}\right)^2 + B\sqrt{nr}\frac{(ar)^2 m_U}{m\epsilon_0 \gamma p}\right) \quad (86)$$

882 *Proof.* We rewrite as follows.

$$\bar{A}\bar{w} - A^F w^F = (\bar{A}\bar{w} - \bar{A}\bar{w}^*) + (\bar{A}\bar{w}^* - A^F \bar{w}^*) + (A^F \bar{w}^* - A^F w^F) \quad (87)$$

883 Now, we proceed to bound the  $\ell_2$  norm of each of these individual terms separately. For the first  
884 term, we have that

$$\|\bar{A}\bar{w} - \bar{A}\bar{w}^*\|_2 = \|\bar{A}(\bar{w} - \bar{w}^*)\|_2 \quad (88)$$

$$\leq \|\bar{w} - \bar{w}^*\|_1 \quad (89)$$

$$\leq \sqrt{r}\|\bar{w} - \bar{w}^*\|_2 \quad (90)$$

$$\leq \sqrt{r}\frac{L_2 L_\infty^2}{2\lambda^3}\|A^S - \bar{A}\|_\infty^2 \quad (91)$$

$$\leq \sqrt{r}\frac{L_2 L_\infty^2}{2\lambda^3}\left(\frac{(ar)^2 m_U}{m\epsilon_0 \gamma p}\right)^2 \quad (92)$$

885 where second line follows from  $\bar{A}$  having column sum 1, and the fourth line follows from Lemma 20  
886 For the third term, we have a similar analysis.

$$\|A^F \bar{w}^* - A^F w^F\|_2 = \|A^F(\bar{w}^* - w^F)\|_2 \quad (93)$$

$$\leq \|\bar{w}^* - w^F\|_1 \quad (94)$$

$$\leq \sqrt{r}\|\bar{w}^* - w^F\|_2 \quad (95)$$

$$\leq \sqrt{r}\frac{L_\infty}{\lambda}\|\bar{A} - A^F\|_\infty \quad (96)$$

$$\leq \sqrt{r}\frac{L_\infty}{\lambda}\left(\frac{(ar)^2 m_U}{m\epsilon_0 \gamma p}\right) \quad (97)$$

887 Finally, for the second term, we have that

$$\|\bar{A}\bar{w}^* - A^F \bar{w}^*\|_2 \leq \|\bar{A} - A^F\|_2\|\bar{w}^*\|_2 \quad (98)$$

$$\leq \|\bar{A} - A^F\|_\infty \sqrt{nr}\|\bar{w}^*\|_2 \quad (99)$$

$$\leq O\left(\frac{(ar)^2 m_U}{m\epsilon_0 \gamma p}\sqrt{nr}B\right) \quad (100)$$

888 By triangle inequality, we obtain the desired result.  $\square$

889 First, we note show the following property of the learned topic model  $A^S$ .

890 **Lemma 23.** *The minimum singular value of the ground truth topic matrix  $A^S$  is at least  $\Theta(p)$ , since  
891 the perturbations in entries of  $\bar{A}^S$  are at most  $\epsilon_0 \leq O(1/\sqrt{nr})$ . Hence, the singular values cannot  
892 change by more than a constant factor relative to  $p$ .*

893 *Proof.* We know that  $A^*$  is a  $p$ -separable topic model, and hence has smallest singular value at  
894 least  $p$ . For the given sample complexity of learning,  $A^S$  will have smallest singular value at least  
895  $\Theta(p)$ .  $\square$

896 The above result says that  $A^S$  has a unique pseudoinverse, and has largest singular value at most  
897  $O(1/p)$ .

898 Recall that our goal for the downstream task is to approximate the  $\mathbf{v}^F$  such that

$$\mathbf{A}^S \mathbf{v} = \mathbf{A}^F \mathbf{w}^F \quad (101)$$

899 in order to say we have approximated the unlearned fine-tuned model. Therefore, it suffices to obtain  
 900 indistinguishability of our unlearning algorithm output  $\tilde{\mathbf{w}}$  with  $(\mathbf{A}^S)^\dagger \mathbf{A}^F \mathbf{w}^F$ . Our following claim  
 901 is that we can use  $(\mathbf{A}^S)^\dagger \tilde{\mathbf{A}} \tilde{\mathbf{w}}$  as the approximation for this.

902 **Proposition 4.** *It holds that*

$$\|(\mathbf{A}^S)^\dagger \tilde{\mathbf{A}} \tilde{\mathbf{w}} - (\mathbf{A}^S)^\dagger \mathbf{A}^F \mathbf{w}^F\|_2 \leq O\left(\frac{1}{p} \|\tilde{\mathbf{A}} \tilde{\mathbf{w}} - \mathbf{A}^F \mathbf{w}^F\|_2\right) \quad (102)$$

$$\leq O\left(\frac{1}{p} \cdot \left[ \sqrt{r} \left( \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right)^2 + B \sqrt{nr} \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right] \right) \quad (103)$$

903 Let  $\tilde{\mathbf{v}} := (\mathbf{A}^S)^\dagger \tilde{\mathbf{A}} \tilde{\mathbf{w}}$  and  $\mathbf{v} = (\mathbf{A}^S)^\dagger \mathbf{A}^F \mathbf{w}^F$ . We claim the following.

904 **Lemma 24.** *The unlearning algorithm  $\mathcal{U}_{head}$  that outputs*

$$\tilde{\mathbf{v}} := \bar{\mathbf{v}} + \nu_v \quad (104)$$

905 where  $\nu_v$  is the noise defined by the Gaussian mechanism using the above sensitivity satisfies prov-  
 906 able  $(\epsilon, \delta)$  unlearning. In particular, we use

$$\sigma = \frac{O\left(\frac{1}{p} \cdot \left[ \sqrt{r} \left( \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right)^2 + B \sqrt{nr} \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right] \right)}{\epsilon} \sqrt{2 \log(1.25/\delta)} \quad (105)$$

907 where the numerator of the fraction is from the previous proposition.

908 *Proof.* This follows from Gaussian mechanism.  $\square$

909 We now proceed to bound the deletion capacity. In this case, the utility is defined by the closeness  
 910 of  $\tilde{\mathbf{v}}$  to  $(\mathbf{A}^S)^\dagger \mathbf{A}^* \mathbf{w}^*$  in  $\ell_\infty$  norm, similar the way we defined this for the base model unlearning  
 911 algorithm  $\mathcal{U}_{base}$  earlier.

912 First, the following lemma to bound  $\mathbf{A}^F \mathbf{w}^F - \mathbf{A}^* \mathbf{w}^*$ .

913 **Lemma 25.** *We have that*

$$\|\mathbf{A}^F \mathbf{w}^F - \mathbf{A}^* \mathbf{w}^*\|_2 \leq O\left(B \sqrt{nr} \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p}\right) \quad (106)$$

914 *Proof.* We decompose as follows.

$$\mathbf{A}^F \mathbf{w}^F - \mathbf{A}^* \mathbf{w}^* = (\mathbf{A}^F \mathbf{w}^F - \mathbf{A}^F \mathbf{w}^*) + (\mathbf{A}^F \mathbf{w}^* - \mathbf{A}^* \mathbf{w}^*) \quad (107)$$

915 The first term is bounded by

$$\|\mathbf{A}^F \mathbf{w}^F - \mathbf{A}^F \mathbf{w}^*\|_2 \leq \sqrt{r} \|\mathbf{w}^F - \mathbf{w}^*\|_2 \leq O(\sqrt{r} \|\mathbf{A}^F - \mathbf{A}^*\|_\infty) \leq O\left(\sqrt{r} \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p}\right) \quad (108)$$

916 The second term is bounded by

$$\|\mathbf{A}^F \mathbf{w}^* - \mathbf{A}^* \mathbf{w}^*\|_2 \leq O\left(\frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \sqrt{nr} B\right) \quad (109)$$

917 by considering the spectral norm  $\|\mathbf{A}^F - \mathbf{A}^*\|_2$ . This gives the desired result.  $\square$

918 As a result, the following holds.

919 **Proposition 5.** *It holds that*

$$\|(\mathbf{A}^S)^\dagger \mathbf{A}^F \mathbf{w}^F - (\mathbf{A}^S)^\dagger \mathbf{A}^* \mathbf{w}^*\|_2 \leq O\left(\frac{1}{p} \left[ \sqrt{r} \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} + B \sqrt{nr} \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right] \right) \quad (110)$$

920 This is once again from the bounded operator norm property of  $(\mathbf{A}^S)^\dagger$ .

921 Finally, we can apply triangle inequality to get the following.

922 **Lemma 26.** *It holds that*

$$\|(\mathbf{A}^S)^\dagger \bar{\mathbf{A}} \bar{\mathbf{w}} - (\mathbf{A}^S)^\dagger \mathbf{A}^* \mathbf{w}^*\|_2 \leq \left( \frac{1}{p} \cdot \left[ \sqrt{r} \left( \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right)^2 + B \sqrt{nr} \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right] \right) \quad (111)$$

923 Then, we can get the following bound on deletion capacity.

924 **Lemma 27.** *For  $\epsilon, \delta > 0$ , the deletion capacity satisfies*

$$T_{\epsilon, \delta}^{\mathcal{A}_{head}, \mathcal{U}_{head}}(m) \geq \tilde{\Omega} \left( \frac{m}{r^2 \sqrt{nr}} \right) \quad (112)$$

925 *Proof.* The calculation is as follows.

$$\mathbb{E}[\|\bar{\mathbf{v}} - (\mathbf{A}^S)^\dagger \mathbf{A}^* \mathbf{w}^*\|_\infty] \leq \mathbb{E}[\|\nu_v\|_\infty] + \mathbb{E}[\|(\mathbf{A}^S)^\dagger \bar{\mathbf{A}} \bar{\mathbf{w}} - (\mathbf{A}^S)^\dagger \mathbf{A}^* \mathbf{w}^*\|_\infty] \quad (113)$$

$$\leq \left( \frac{1}{p} \cdot \left[ \sqrt{r} \left( \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right)^2 + B \sqrt{nr} \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right] \right) \left( \frac{\sqrt{\log r \log 1/\delta}}{\epsilon} + 1 \right) \quad (114)$$

926 For this to be a small constant, we require

$$\frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \leq \tilde{O} \left( \min \left\{ \frac{1}{r^{1/4}}, \frac{1}{\sqrt{nr}} \right\} \right) \quad (115)$$

927 Therefore, we should have

$$m_U \leq \tilde{\Omega} \left( \frac{m}{r^2 \sqrt{nr}} \right) \quad (116)$$

928  $\square$

## 929 C.2 Proof for general $q$

930 The following is the formal statement of Theorem 4.

931 **Theorem 7** (Formal version of Theorem 4). *Suppose that the downstream task  $\mathcal{T}$  only depends on*  
 932 *a subset of topics  $\mathbb{T}_{clf} \subseteq [r]$ ; that is,  $\mathbf{w}^* = \arg \min_{\mathbf{v} \in \mathcal{V}_{base}} \ell_{\mathcal{T}}(\mathbf{v}; \mathbf{A}^*)$  has non-zero entries only in*  
 933 *the index set  $\mathbb{T}_{clf}$ . Denote  $q := \min_{k \in \mathbb{T}_{clf}} \Pr_{\mathcal{D}}[z = k]$ , and let  $\mathcal{A}_{head}$  be the head tuning algorithm*  
 934 *(Definition 2) and  $\mathcal{U}_{head}$  be Algorithm 2. Then,  $(\mathcal{A}_{head}, \mathcal{U}_{head})$  performs utility-preserving unlearning*  
 935 *with deletion capacity*

$$T_{\epsilon, \delta}^{\mathcal{A}_{head}, \mathcal{U}_{head}}(m) \geq c' \cdot \min \left\{ \frac{mq\epsilon}{r \sqrt{nr \log 1/\delta}}, \frac{0.001m}{r^2} \right\} \quad (117)$$

936 where  $c'$  is a constant dependent on  $\mathcal{D}$ , and  $\mathcal{T}$ .

937 **Lemma 28.** *Recall our retrained model for the downstream task is  $\mathbf{A}^F \mathbf{w}^F$ . Then, it holds that*

$$\|\bar{\mathbf{A}} \bar{\mathbf{w}} - \mathbf{A}^F \mathbf{w}^F\|_2 \leq O \left( \sqrt{r} \left( \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right) \right) + O \left( B \sqrt{nr} \frac{(1/q) a r m_U}{m \epsilon_0 \gamma p} \right) + O \left( \left( \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right)^2 \sqrt{nr} \right) \quad (118)$$

938 *Proof.* Consider this decomposition again.

$$\bar{\mathbf{A}} \bar{\mathbf{w}} - \mathbf{A}^F \mathbf{w}^F = (\bar{\mathbf{A}} \bar{\mathbf{w}} - \bar{\mathbf{A}} \bar{\mathbf{w}}^*) + (\bar{\mathbf{A}} \bar{\mathbf{w}}^* - \mathbf{A}^F \bar{\mathbf{w}}^*) + (\mathbf{A}^F \bar{\mathbf{w}}^* - \mathbf{A}^F \mathbf{w}^F) \quad (119)$$

939 The first term is the same as old analysis; the second term is from considering  $q$ ; the third is the  
 940 same as the old analysis. In particular, when  $q = 1/ar$ , we recover the old bound. We have that the  
 941 first term is

$$\|\bar{\mathbf{A}} \bar{\mathbf{w}} - \bar{\mathbf{A}} \bar{\mathbf{w}}^*\| \leq \sqrt{r} \frac{L_2 L_\infty^2}{2\lambda^3} \left( \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right)^2 \quad (120)$$

942 The third term is

$$\|\mathbf{A}^F \bar{\mathbf{w}}^* - \mathbf{A}^F \mathbf{w}^F\| \leq \sqrt{r} \frac{L_\infty}{\lambda} \left( \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right) \quad (121)$$

943 The second term is

$$\|\bar{\mathbf{A}} \bar{\mathbf{w}}^* - \mathbf{A}^F \bar{\mathbf{w}}^*\| \leq \|(\bar{\mathbf{A}} - \mathbf{A}^F) \bar{\mathbf{w}}^*\| + \|(\bar{\mathbf{A}} - \mathbf{A}^F)(\mathbf{w}^* - \bar{\mathbf{w}}^*)\| \quad (122)$$

$$\leq O\left(B\sqrt{nr} \frac{(1/q)arm_U}{m \epsilon_0 \gamma p}\right) + O\left(\left(\frac{(ar)^2 m_U}{m \epsilon_0 \gamma p}\right)^2 \sqrt{nr}\right) \quad (123)$$

944 This gives the desired result using triangle inequality.  $\square$

945 Continuing, we have the following.

946 **Proposition 6.** *It holds that*

$$\|(\mathbf{A}^S)^\dagger \bar{\mathbf{A}} \bar{\mathbf{w}} - (\mathbf{A}^S)^\dagger \mathbf{A}^F \mathbf{w}^F\|_2 \quad (124)$$

$$\leq O\left(\frac{1}{p} \|\bar{\mathbf{A}} \bar{\mathbf{w}} - \mathbf{A}^F \mathbf{w}^F\|_2\right) \quad (125)$$

$$\leq O\left(\frac{1}{p} \cdot \left[ \sqrt{r} \left( \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right) + B\sqrt{nr} \frac{(1/q)arm_U}{m \epsilon_0 \gamma p} + \left( \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right)^2 \sqrt{nr} \right] \right) \quad (126)$$

947 This gives us the following.

948 **Lemma 29.** *The unlearning algorithm  $\mathcal{U}_{head}$  that outputs*

$$\tilde{\mathbf{v}} := \bar{\mathbf{v}} + \nu_v \quad (127)$$

949 *where  $\nu_v$  is the noise defined by the Gaussian mechanism using the above sensitivity satisfies prov-*  
950 *able  $(\epsilon, \delta)$  unlearning. In particular, we use*

$$\sigma = \frac{O\left(\frac{1}{p} \cdot \left[ \sqrt{r} \left( \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right) + B\sqrt{nr} \frac{(1/q)arm_U}{m \epsilon_0 \gamma p} + \left( \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right)^2 \sqrt{nr} \right] \right)}{\epsilon} \sqrt{2 \log(1.25/\delta)} \quad (128)$$

951 *where the numerator of the fraction is from the previous proposition.*

952 *Proof.* This follows from Gaussian mechanism.  $\square$

953 We now proceed to bound the deletion capacity. In this case, the utility is defined by the closeness  
954 of  $\tilde{\mathbf{v}}$  to  $(\mathbf{A}^S)^\dagger \mathbf{A}^* \mathbf{w}^*$  in  $\ell_\infty$  norm, similar the way we defined this for the base model unlearning  
955 algorithm  $\mathcal{U}_{base}$  earlier.

956 First, the following lemma to bound  $\mathbf{A}^F \mathbf{w}^F - \mathbf{A}^* \mathbf{w}^*$ .

957 **Lemma 30.** *We have that*

$$\|\mathbf{A}^F \mathbf{w}^F - \mathbf{A}^* \mathbf{w}^*\|_2 \leq O\left(\sqrt{r} \left( \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right) + B\sqrt{nr} \frac{(1/q)arm_U}{m \epsilon_0 \gamma p} + \left( \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right)^2 \sqrt{nr}\right) \quad (129)$$

958 *Proof.* We decompose as follows.

$$\mathbf{A}^F \mathbf{w}^F - \mathbf{A}^* \mathbf{w}^* = (\mathbf{A}^F \mathbf{w}^F - \mathbf{A}^F \mathbf{w}^*) + (\mathbf{A}^F \mathbf{w}^* - \mathbf{A}^* \mathbf{w}^*) \quad (130)$$

959 The first term is bounded by

$$\|\mathbf{A}^F \mathbf{w}^F - \mathbf{A}^F \mathbf{w}^*\|_2 \leq \sqrt{r} \|\mathbf{w}^F - \mathbf{w}^*\|_2 \leq O(\sqrt{r} \|\mathbf{A}^F - \mathbf{A}^*\|_\infty) \leq O\left(\sqrt{r} \left( \frac{(ar)^2 m_U}{m \epsilon_0 \gamma p} \right)\right) \quad (131)$$

960 The second term is bounded by

$$\|\mathbf{A}^F \mathbf{w}^* - \mathbf{A}^* \mathbf{w}^*\|_2 \leq B\sqrt{nr} \frac{(1/q)arm_U}{m\epsilon_0\gamma p} + \left( \frac{(ar)^2 m_U}{m\epsilon_0\gamma p} \right)^2 \sqrt{nr} \quad (132)$$

961 Triangle inequality gives us the desired result.  $\square$

962 As a result, the following holds.

963 **Proposition 7.** *It holds that*

$$\|(\mathbf{A}^S)^\dagger \mathbf{A}^F \mathbf{w}^F - (\mathbf{A}^S)^\dagger \mathbf{A}^* \mathbf{w}^*\|_2 \leq O\left(\frac{1}{p} \cdot \left[ \sqrt{r} \left( \frac{(ar)^2 m_U}{m\epsilon_0\gamma p} \right) + B\sqrt{nr} \frac{(1/q)arm_U}{m\epsilon_0\gamma p} + \left( \frac{(ar)^2 m_U}{m\epsilon_0\gamma p} \right)^2 \sqrt{nr} \right]\right) \quad (133)$$

964 This is once again from the bounded operator norm property.

965 Finally, we can apply triangle inequality to get the following.

966 **Lemma 31.** *It holds that*

$$\|(\mathbf{A}^S)^\dagger \bar{\mathbf{A}} \bar{\mathbf{w}} - (\mathbf{A}^S)^\dagger \mathbf{A}^* \mathbf{w}^*\|_2 \leq O\left(\frac{1}{p} \cdot \left[ \sqrt{r} \left( \frac{(ar)^2 m_U}{m\epsilon_0\gamma p} \right) + B\sqrt{nr} \frac{(1/q)arm_U}{m\epsilon_0\gamma p} + \left( \frac{(ar)^2 m_U}{m\epsilon_0\gamma p} \right)^2 \sqrt{nr} \right]\right) \quad (134)$$

967 Then, we can get the following bound on deletion capacity.

968 **Lemma 32.** *For  $\epsilon, \delta > 0$ , the deletion capacity satisfies*

$$T_{\epsilon, \delta}^{\mathcal{A}_{head}, \mathcal{U}_{head}}(m) \geq \tilde{\Omega}\left(\frac{m}{r^2 \sqrt{nr}}\right) \quad (135)$$

969 *Proof.* The calculation is as follows.

$$\mathbb{E}[\|\tilde{\mathbf{v}} - (\mathbf{A}^S)^\dagger \mathbf{A}^* \mathbf{w}^*\|_\infty] \leq \mathbb{E}[\|\nu_{\mathbf{v}}\|_\infty] + \mathbb{E}[\|(\mathbf{A}^S)^\dagger \bar{\mathbf{A}} \bar{\mathbf{w}} - (\mathbf{A}^S)^\dagger \mathbf{A}^* \mathbf{w}^*\|_\infty] \quad (136)$$

$$\leq \left( \frac{1}{p} \cdot \left[ \sqrt{r} \left( \frac{(ar)^2 m_U}{m\epsilon_0\gamma p} \right) + B\sqrt{nr} \frac{(1/q)arm_U}{m\epsilon_0\gamma p} + \left( \frac{(ar)^2 m_U}{m\epsilon_0\gamma p} \right)^2 \sqrt{nr} \right] \right) \quad (137)$$

$$\cdot \left( \frac{\sqrt{\log r \log 1/\delta}}{\epsilon} + 1 \right) \quad (138)$$

970 For this to be a small constant, we require

$$\frac{(ar)^2 m_U}{m\epsilon_0\gamma p} \leq \tilde{O}\left(\min\left\{\frac{1}{r^{1/2}}, \frac{1}{(nr)^{1/4}}, \frac{arq}{\sqrt{nr}}\right\}\right) \quad (139)$$

971 When  $n$  is at least  $r^3$ , this bound will be tight. Therefore, we should have

$$m_U \leq \tilde{\Omega}\left(\frac{mq}{r^{1.5} n^{0.5}}\right) \quad (140)$$

972  $\square$