

ENERGY-BASED ADAPTIVE REGULARIZATION FOR COPYRIGHT PROTECTION IN LANGUAGE MODELS

Anonymous authors

Paper under double-blind review

ABSTRACT

Large language models memorize and reproduce copyrighted content from their training data, raising significant legal concerns. Existing protection methods either exclude copyrighted data entirely, sacrificing model capabilities, or apply unstable regularization that causes training collapse. We introduce the first energy-based framework for copyright protection, reformulating memorization prevention as energy minimization rather than probability manipulation. Our key insight is that assigning higher energy to copyrighted sequences creates an exponential barrier to their reproduction, with protection strength naturally scaling with sequence length. We propose Adaptive Energy Regularization (AER), which dynamically balances copyright protection and model utility. We provide rigorous theoretical foundations: proving convergence under the Polyak-Łojasiewicz condition, establishing exponential suppression bounds that scale with sequence length, and guaranteeing robustness under distribution shift. Empirically, across **19** models ranging from 124M to 14B parameters, AER reduces verbatim reproduction from up to **99.1%** to below **1%** while preserving perplexity within **3.2%** of baseline. Our energy-based approach provides a principled and stable solution to copyright protection, establishing a paradigm for controlling memorization in generative AI.

1 INTRODUCTION

Current approaches to copyright protection in language models face fundamental limitations. Data filtering excludes copyrighted content from training but severely restricts model capabilities Yu et al. (2023). Post-processing filters detect copyrighted content during generation without addressing the root memorization problem Kibriya et al. (2024). Training-time regularization, particularly inverse regularization Chu et al. (2024), attempts to penalize memorization through reciprocal loss terms but suffers from numerical instability when denominators approach zero, causing gradient explosion and training collapse.

We introduce the first energy-based framework for copyright protection in language models. The key insight is reformulating the problem through energy functions rather than generation probabilities. By assigning higher energy to copyrighted content and lower energy to ordinary text, we create an energy barrier that exponentially suppresses copyright reproduction. This suppression strengthens with sequence length, providing increasingly robust protection for longer passages while maintaining stable gradient dynamics throughout training.

We make the following contributions:

- We propose **Energy-based Copyright Protection**, the first framework that reformulates memorization prevention as energy minimization, enabling exponential suppression that scales with sequence length.
- We develop **Adaptive Energy Regularization (AER)**, an algorithm that automatically balances protection and utility through dynamic energy gap optimization, eliminating the need for manual hyperparameter tuning.
- We provide rigorous theoretical guarantees (convergence under PL condition, exponential suppression bounds) and comprehensive empirical validation across 19 models (124M-14B parameters). AER reduces verbatim reproduction from up to 99.1% to below 1% while preserving perplexity within 3-8% of baseline across GPT-2, LLaMA-2/3, and Qwen-2.5/3.

The energy formulation transforms copyright protection from an ad-hoc constraint into a principled optimization problem. By creating an energy barrier between copyrighted and ordinary content, our method achieves strong protection while preserving model utility. Unlike inverse regularization approaches, our bounded regularizer ensures numerical stability and avoids gradient explosion throughout training.

2 PRELIMINARIES

2.1 NOTATIONS AND PROBLEM SETUP

We consider a language model with parameters $\theta \in \mathbb{R}^d$ trained on dataset $\mathcal{D} = \mathcal{C} \cup \mathcal{O}$, where $\mathcal{C}$ denotes copyrighted data with size $n_c = |\mathcal{C}|$ and $\mathcal{O}$ denotes ordinary (non-copyrighted) data with size $n_o = |\mathcal{O}|$. For a text sequence $x = (x_1, \dots, x_L)$ with tokens from vocabulary $\mathcal{V}$, we denote its length as $|x|$. The language model p_θ assigns probability $p_\theta(x) = \prod_{t=1}^{|x|} p_\theta(x_t|x_{<t})$, where $x_{<t} = (x_1, \dots, x_{t-1})$ denotes the context. For theoretical analysis of content similarity, we assume the language model induces a representation function $\phi : \mathcal{V}^* \rightarrow \mathbb{R}^h$ mapping variable-length token sequences to h -dimensional embeddings. For a sequence $x = (x_1, \dots, x_{|x|})$, we define $\phi(x)$ as the average-pooled final hidden states Gao et al. (2021): $\phi(x) = \frac{1}{|x|} \sum_{t=1}^{|x|} h_t^{(x)}$ where $h_t^{(x)} \in \mathbb{R}^h$ is the model’s hidden representation at position t for sequence x . A complete summary of all notation used in this paper is provided in Appendix B.1, and the properties of this embedding function are analyzed in Appendix D.1.

We define the **energy function** as the average negative log-likelihood:

$$E(x; \theta) = -\frac{1}{|x|} \sum_{t=1}^{|x|} \log p_\theta(x_t|x_{<t}) \quad (1)$$

This energy quantifies the model’s uncertainty about a sequence—higher energy corresponds to lower generation probability. The relationship between energy and probability follows directly from the definition. Taking the logarithm of the probability:

$$\log p_\theta(x) = \sum_{t=1}^{|x|} \log p_\theta(x_t|x_{<t}) = -|x| \cdot E(x; \theta) \quad (2)$$

Therefore, the probability can be expressed in terms of energy as:

$$p_\theta(x) = \exp(-|x| \cdot E(x; \theta)) \quad (3)$$

This exponential relationship is fundamental to our protection mechanism. For comparing relative probabilities of sequences with the same length $|x_1| = |x_2| = |x|$, we have:

$$\frac{p_\theta(x_1)}{p_\theta(x_2)} = \exp(-|x| \cdot (E(x_1; \theta) - E(x_2; \theta))) \quad (4)$$

For sequences of different lengths $|x_1| \neq |x_2|$, the ratio becomes:

$$\frac{p_\theta(x_1)}{p_\theta(x_2)} = \exp(-(|x_1| \cdot E(x_1; \theta) - |x_2| \cdot E(x_2; \theta))) \quad (5)$$

This shows that a unit increase in energy results in an exponential decrease in generation probability, scaled by sequence length. We provide a detailed analysis of energy-probability relationships and their implications for variable-length sequences in Appendix D.2.

2.2 PROBLEM DEFINITION: ENERGY-BASED COPYRIGHT PROTECTION

The fundamental challenge in training language models on copyrighted data is preventing verbatim reproduction while maintaining model utility. We formalize this through an energy-based perspective:

Definition 1 (Energy-based Copyright Protection). A language model p_θ achieves $\Delta_{\min}$ -copyright protection if:

$$\mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta)] - \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta)] \geq \Delta_{\min} \quad (6)$$

$$\text{subject to: } \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta)] \leq E_0 \quad (7)$$

where c denotes copyrighted sequences, x denotes ordinary sequences, $\Delta_{\min} > 0$ is the target protection margin, and E_0 is the maximum acceptable energy for ordinary data.

The intuition behind this definition is that we create an “energy barrier” between copyrighted and ordinary content. When the model encounters a prompt that could lead to copyrighted text, the high energy barrier exponentially suppresses the generation probability, effectively preventing verbatim reproduction. Unlike traditional approaches that modify probabilities directly, our energy-based formulation provides exponential suppression that scales with sequence length, offering protection that increases exponentially with sequence length (as shown in Eq 5).

3 ENERGY-BASED FRAMEWORK FOR COPYRIGHT PROTECTION

3.1 WHY ENERGY-BASED PERSPECTIVE OUTPERFORMS PROBABILITY-BASED METHODS

Traditional probability-based approaches Chu et al. (2024) directly manipulate generation probabilities through constraints like $p_\theta(c) < \epsilon_{\text{prob}}$ for copyrighted content $c \in \mathcal{C}$, where $\epsilon_{\text{prob}} > 0$ is a small threshold. However, this perspective suffers from three fundamental limitations that our energy-based framework addresses.

First, probability constraints are inherently local and fail to capture the compositional nature of text generation Xu et al. (2024d). When generating token-by-token, a model can have reasonable per-token probabilities while still producing copyrighted sequences through their composition. The energy-based view naturally accumulates protection across the entire sequence: for a sequence of length $|c|$, the probability suppression factor scales as $\exp(-|c| \cdot \Delta_{\min})$, providing exponentially stronger protection for longer texts.

Second, while energy and log-probability are mathematically related through the linear transformation through the transformation $E = -\frac{1}{|x|} \log p$, they exhibit fundamentally different behavior during optimization. Consider the gradient dynamics when optimizing for copyright protection Liu et al. (2021). For probability minimization with objective $\min_\theta p_\theta(c)$, the gradient is:

$$\nabla_\theta p_\theta(c) = p_\theta(c) \cdot \nabla_\theta \log p_\theta(c) = -p_\theta(c) \cdot |c| \cdot \nabla_\theta E(c; \theta) \quad (8)$$

The gradient magnitude becomes $\|\nabla_\theta p_\theta(c)\|_2 = |p_\theta(c) \cdot |c|| \cdot \|\nabla_\theta E(c; \theta)\|_2 = p_\theta(c) \cdot |c| \cdot \|\nabla_\theta E(c; \theta)\|_2$. As optimization succeeds and $p_\theta(c) \rightarrow 0$, the gradient vanishes regardless of $\|\nabla_\theta E(c; \theta)\|_2$. This creates a fundamental optimization barrier where success leads to gradient disappearance, making further optimization impossible. We provide a rigorous analysis of this vanishing gradient phenomenon in Appendix D.3.

In contrast, for energy maximization with objective $\max_\theta E(c; \theta)$, the gradient $\nabla_\theta E(c; \theta)$ maintains sufficient magnitude for effective optimization. Specifically, under the regularity conditions we establish in Assumption 1, the gradient norm satisfies $\|\nabla_\theta E(c; \theta)\|_2 \leq G$ for all $c \in \mathcal{C}$ and $\theta \in \mathcal{B}(\theta^*, r)$, and remains bounded away from zero when $E(c; \theta)$ is not at its optimum. This ensures consistent learning signals throughout training, avoiding the vanishing gradient problem inherent in probability-based formulations.

Third, and most critically, energy-based formulations provide exponential decay guarantees that probability methods cannot match. To see this precisely, let $p_{\text{baseline}}(c)$ denote a reference model without copyright protection trained on the same data. Taking the log-probability ratio:

$$\log \frac{p_\theta(c)}{p_{\text{baseline}}(c)} = -|c| \cdot (E(c; \theta) - E_{\text{baseline}}(c)) \quad (9)$$

Therefore, when our method achieves an energy gap $E(c; \theta) - E_{\text{baseline}}(c) \geq \Delta_{\min}$:

$$\frac{p_\theta(c)}{p_{\text{baseline}}(c)} = \exp(-|c|(E(c; \theta) - E_{\text{baseline}}(c))) \leq \exp(-|c| \cdot \Delta_{\min}) \quad (10)$$

This exponential suppression factor becomes overwhelming for typical copyrighted passages (often hundreds of tokens), providing strong theoretical guarantees against verbatim reproduction. A detailed comparison with baseline models is provided in Appendix D.4.

3.2 LIMITATIONS OF EXISTING APPROACHES

Current state-of-the-art methods employ inverse regularization to discourage memorization:

$$\mathcal{L}_{\text{inv}}(\theta) = \mathcal{L}_{\text{LM}}^{\mathcal{O}}(\theta) + \gamma_{\text{inv}} \cdot [\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) + \epsilon_0]^{-1} \quad (11)$$

where $\mathcal{L}_{\text{LM}}^{\mathcal{S}}(\theta) = \mathbb{E}_{x \sim \mathcal{U}(\mathcal{S})}[E(x; \theta)]$ denotes the language modeling loss on dataset $\mathcal{S}$ with uniform sampling, with $\mathcal{L}_{\text{LM}}^{\mathcal{O}}(\theta)$ for ordinary data and $\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta)$ for copyrighted data, $\gamma_{\text{inv}} > 0$ is the inverse regularization strength, and $\epsilon_0 > 0$ ensures numerical stability.

The gradient $\nabla_{\theta}[\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) + \epsilon_0]^{-1} = -[\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) + \epsilon_0]^{-2} \nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta)$ reveals the fundamental instability: as the model improves on copyrighted data (reducing $\mathcal{L}_{\text{LM}}^{\mathcal{C}}$), gradients explode, causing optimization failure. We formalize this instability and its consequences in Appendix D.5. Moreover, the non-linear relationship between γ_{inv} and actual protection level makes hyperparameter tuning unpredictable. Most critically, inverse regularization provides no worst-case guarantees—even small perturbations in the data distribution can cause the protection to fail completely, as the inverse term may become negligible or overwhelming.

3.3 OUR APPROACH: DIRECT ENERGY OPTIMIZATION

We propose directly optimizing the energy landscape through a principled objective:

$$\mathcal{L}_{\text{energy}}(\theta) = \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta)] - \lambda \cdot \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta)] \quad (12)$$

where $\lambda > 0$ controls the energy gap. This formulation minimizes energy on ordinary data while maximizing it on copyrighted data. As we establish in Assumption 1(c), when individual energy gradients are bounded by G , this leads to stable optimization with $\|\nabla_{\theta} \mathcal{L}_{\text{energy}}\|_2 \leq (1 + \lambda) \cdot G$. The Lipschitz properties of this objective are analyzed in Appendix F.2.

Assumption 1 (Polyak-Łojasiewicz Conditions). The expected energy functions for both ordinary and copyrighted data satisfy the following conditions in a neighborhood $\mathcal{B}(\theta^*, r)$ around the optimal parameters θ^* :

(a) **PL condition:** For all $\theta \in \mathcal{B}(\theta^*, r)$, the expected energy satisfies:

$$\|\nabla_{\theta} \mathbb{E}_{x \sim \mathcal{U}(\mathcal{D})}[E(x; \theta)]\|_2^2 \geq 2\mu_{\text{PL}} (\mathbb{E}_{x \sim \mathcal{U}(\mathcal{D})}[E(x; \theta)] - \mathbb{E}_{x \sim \mathcal{U}(\mathcal{D})}[E(x; \theta^*)]) \quad (13)$$

where $\mu_{\text{PL}} > 0$ is the PL constant.

(b) **Smoothness:** For all $\theta_1, \theta_2 \in \mathcal{B}(\theta^*, r)$:

$$\|\nabla_{\theta} \mathbb{E}_{x \sim \mathcal{U}(\mathcal{D})}[E(x; \theta_1)] - \nabla_{\theta} \mathbb{E}_{x \sim \mathcal{U}(\mathcal{D})}[E(x; \theta_2)]\|_2 \leq L \|\theta_1 - \theta_2\|_2 \quad (14)$$

(c) **Bounded variance:** For all $\theta \in \mathcal{B}(\theta^*, r)$:

$$\mathbb{E}_{x \sim \mathcal{U}(\mathcal{D})}[\|\nabla_{\theta} E(x; \theta) - \mathbb{E}[\nabla_{\theta} E(x; \theta)]\|_2^2] \leq \sigma^2 \quad (15)$$

and $\sup_{x \in \mathcal{D}} \|\nabla_{\theta} E(x; \theta)\|_2 \leq G$.

Remark on PL. The Polyak-Łojasiewicz (PL) condition is strictly weaker than strong convexity but still guarantees global convergence to a stationary point. Unlike strong convexity which requires $\nabla_{\theta}^2 f(\theta) \succeq \mu I$, the PL condition only requires gradient dominance, making it applicable to non-convex functions including neural networks. Recent work has shown that overparameterized neural networks satisfy the PL condition with high probability near initialization Xiao et al. (2023); Liu et al. (2023). We discuss when language models satisfy the PL condition and how to estimate μ_{PL} empirically in Appendix F.1. When the PL condition holds, we still obtain convergence rates similar to the strongly convex case, with the key difference being convergence to a stationary point rather than a global minimum.

Theorem 2 (Energy Gap Guarantee). *Consider the optimization problem $\min_{\theta} \mathcal{L}_{\text{energy}}(\theta)$ where:*

$$\mathcal{L}_{\text{energy}}(\theta) = \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta)] - \lambda \cdot \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta)] \quad (16)$$

Let θ^ be a local minimizer of $\mathcal{L}_{\text{energy}}(\theta)$. Under Assumption 1, at the optimal point θ^* , the gradient vanishes:*

$$\nabla_{\theta} \mathcal{L}_{\text{energy}}(\theta^*) = \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[\nabla_{\theta} E(x; \theta^*)] - \lambda \cdot \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[\nabla_{\theta} E(c; \theta^*)] = 0 \quad (17)$$

If there exists a parameter θ_{sep} achieving separation with weakly correlated gradients:

$$\begin{aligned} & |\langle \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[\nabla_{\theta} E(x; \theta_{\text{sep}})], \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[\nabla_{\theta} E(c; \theta_{\text{sep}})] \rangle| \\ & \leq \delta \|\mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[\nabla_{\theta} E(x; \theta_{\text{sep}})]\| \|\mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[\nabla_{\theta} E(c; \theta_{\text{sep}})]\| \end{aligned} \quad (18)$$

where $\delta \in [0, 1)$, then the energy gap at θ^ satisfies:*

$$\mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta^*)] - \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta^*)] \geq \frac{\lambda}{\lambda + 1} \cdot (1 - \delta) \cdot \Delta_{\text{sep}} \quad (19)$$

where $\Delta_{\text{sep}} = \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta_{\text{sep}})] - \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta_{\text{sep}})]$ is the achievable separation gap.

Remark. The weak correlation condition in Theorem 2 with parameter $\delta \in [0, 1)$ generalizes the idealized orthogonal case ($\delta = 0$). This condition naturally holds when ordinary and copyrighted data have sufficiently different features. Even with moderate correlation ($\delta < 1$), the energy gap still provides exponential suppression of copyrighted content generation, with the suppression factor scaling as $(1 - \delta)$. We provide a detailed proof in Appendix G.1.

4 ADAPTIVE ENERGY REGULARIZATION

4.1 MOTIVATION AND DESIGN

Theorem 2 establishes that achieving an energy gap $\Delta(\theta) \geq \frac{\lambda}{\lambda+1}(1 - \delta)\Delta_{\text{sep}}$ provides exponential suppression of copyrighted content. However, naively maximizing energy on copyrighted data can degrade overall model quality. We need a mechanism that maintains language modeling capability while ensuring the energy gap reaches the theoretical threshold. Our adaptive regularizer automatically adjusts the optimization pressure based on whether the current gap $\Delta(\theta)$ meets the target margin m :

Definition 3 (Adaptive Energy Regularizer). Given current energy gap $\Delta(\theta) = \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta)] - \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta)]$, the adaptive regularizer is:

$$\mathcal{R}(\theta; m, \tau) = \tau \log \left(1 + \exp \left(-\frac{\Delta(\theta) - m}{\tau} \right) \right) \quad (20)$$

where $m \geq 0$ is the target margin and $\tau > 0$ controls transition smoothness (temperature).

This regularizer elegantly balances three critical properties. It remains bounded with $0 \leq \mathcal{R}(\theta) \leq \tau \log 2$, preventing gradient explosion even during early training. The adaptive nature emerges from its behavior: when $\Delta(\theta) < m$, it applies strong regularization proportional to $m - \Delta(\theta)$; when $\Delta(\theta) \geq m$, it smoothly vanishes, preserving model quality. The gradient $\nabla_{\theta} \mathcal{R} = -\sigma((m - \Delta(\theta))/\tau) \cdot \nabla_{\theta} \Delta(\theta)$ remains Lipschitz continuous Zhang et al. (2024) with constant $L_R = G^2/(4\tau)$, where $\sigma(z) = 1/(1 + e^{-z})$ denotes the sigmoid function and $G = \sup_{\theta} \|\nabla_{\theta} \Delta(\theta)\|$ bounds the gradient norm of the energy gap. We prove these properties rigorously in Appendix E.1.

4.2 COMPLETE TRAINING OBJECTIVE

We combine standard language modeling with adaptive copyright protection:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta) + \gamma \cdot \mathcal{R}(\theta; m, \tau) \quad (21)$$

where $\mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta) = w_o \cdot \mathcal{L}_{\text{LM}}^{\mathcal{O}}(\theta) + w_c \cdot \mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta)$ denotes the weighted language modeling loss, with weights $w_o = |\mathcal{O}|/(|\mathcal{O}| + |\mathcal{C}|)$ and $w_c = |\mathcal{C}|/(|\mathcal{O}| + |\mathcal{C}|)$ proportional to dataset sizes (Xie et al., 2023). Here, $\mathcal{L}_{\text{LM}}^{\mathcal{S}}(\theta) = \mathbb{E}_{x \sim \mathcal{U}(\mathcal{S})}[E(x; \theta)]$ for any dataset $\mathcal{S}$, and $\gamma > 0$ is the regularization strength balancing the two objectives.

Theorem 4 (Equilibrium Characterization). *Consider the optimization problem $\min_{\theta} \mathcal{L}(\theta)$ with the combined objective from Eq. equation 21. At a local minimum θ^* , the first-order optimality condition requires:*

$$\nabla_{\theta} \mathcal{L}_{LM}^{\mathcal{D}}(\theta^*) = \gamma \cdot \sigma \left(\frac{m - \Delta(\theta^*)}{\tau} \right) \cdot \nabla_{\theta} \Delta(\theta^*) \quad (22)$$

Under Assumption 1, if the language modeling gradient has bounded norm $\|\nabla_{\theta} \mathcal{L}_{LM}^{\mathcal{D}}(\theta^)\| \leq B_{LM}$ and the energy gap gradient satisfies $\|\nabla_{\theta} \Delta(\theta^*)\| \geq g_{\min} > 0$ (non-degeneracy), then:*

$$|\Delta(\theta^*) - m| \leq \tau \log \left(1 + \frac{B_{LM}}{\gamma \cdot g_{\min}} \right) \quad (23)$$

In particular, for sufficiently large $\gamma \geq B_{LM}/(g_{\min} \cdot \epsilon)$ with desired precision $\epsilon > 0$:

$$|\Delta(\theta^*) - m| \leq \tau \log(1 + \epsilon) \approx \tau \cdot \epsilon \quad (24)$$

The proof (Appendix G.2) uses a fixed-point analysis of the first-order optimality conditions. This theorem shows that the equilibrium energy gap converges to the target margin m with error controlled by the temperature τ . As γ increases, the model more precisely achieves the desired protection level. The monotonicity of this convergence is analyzed in Appendix E.2.

To efficiently optimize this objective in practice, we develop an adaptive training algorithm that automatically adjusts the regularization strength based on the current energy gap. The algorithm employs proportional batch sampling for unbiased gradient estimates and incorporates gradient clipping for numerical stability. The complete implementation details and computational complexity analysis are provided in Appendix B.2. Having established the learning framework and optimization procedure, we now turn to the theoretical analysis of our approach.

5 THEORETICAL ANALYSIS

5.1 OPTIMIZATION PROPERTIES

Our adaptive energy regularization maintains favorable optimization properties throughout training.

Theorem 5 (Gradient Stability). *Under Assumption 1, the complete objective $\mathcal{L}(\theta)$ has Lipschitz continuous gradient Gouk et al. (2021) with constant:*

$$L_{\mathcal{L}} = L_{LM} + \gamma \cdot L_{\mathcal{R}} \quad (25)$$

where L_{LM} is the Lipschitz constant of the language modeling loss gradient (which depends on the energy function’s Lipschitz constant L through the softmax operation), and $L_{\mathcal{R}} = G^2/(4\tau)$ is the Lipschitz constant of the adaptive regularizer’s gradient.

We prove this result in Appendix G.3 by analyzing the Hessian of the adaptive regularizer.

Theorem 6 (Convergence Rate). *Let $\mathcal{L}^* = \inf_{\theta} \mathcal{L}(\theta)$ denote the global minimum value and σ^2 bound the variance of stochastic gradients: $\mathbb{E}[\|g^{(t)} - \nabla \mathcal{L}(\theta^{(t)})\|^2] \leq \sigma^2$. With step size $\eta = 1/L_{\mathcal{L}}$ Liu & Yuan (2022); Velikanov & Yarotsky (2024), Algorithm 1 achieves:*

(i) **General smooth case:**

$$\frac{1}{N_{\text{train}}} \sum_{t=0}^{N_{\text{train}}-1} \mathbb{E}[\|\nabla \mathcal{L}(\theta^{(t)})\|^2] \leq \frac{2L_{\mathcal{L}}[\mathcal{L}(\theta^{(0)}) - \mathcal{L}^*]}{N_{\text{train}}} + \frac{\sigma^2}{L_{\mathcal{L}} N_{\text{train}}} \quad (26)$$

(ii) **Under the PL condition** (Assumption 1) with constant $\mu_{PL} > 0$:

$$\mathbb{E}[\mathcal{L}(\theta^{(N_{\text{train}})}) - \mathcal{L}^*] \leq \left(1 - \frac{\mu_{PL}}{L_{\mathcal{L}}} \right)^{N_{\text{train}}} [\mathcal{L}(\theta^{(0)}) - \mathcal{L}^*] + \frac{\sigma^2}{2\mu_{PL}} \quad (27)$$

achieving linear convergence to the global optimum with rate $(1 - \mu_{PL}/L_{\mathcal{L}})$.

The complete convergence analysis is provided in Appendix G.4. These results establish that AER maintains standard SGD convergence guarantees despite the adaptive regularization. In the general smooth case, the average squared gradient norm converges at rate $O(1/N_{\text{train}})$. Under the PL condition, we achieve linear convergence to the global optimum with rate determined by the condition number $L_{\mathcal{L}}/\mu_{PL}$.

5.2 COPYRIGHT PROTECTION GUARANTEES

Theorem 7 (Exponential Protection Guarantee). *Consider two models trained on the same dataset $\mathcal{D} = \mathcal{C} \cup \mathcal{O}$. Let θ^* denote the parameters obtained through AER optimization (Algorithm 1) with target margin m , and let θ_{base} denote the parameters of a baseline model trained using only the standard language modeling objective $\mathcal{L}_{LM}^D(\theta)$ without any copyright protection mechanism.*

Suppose the AER model achieves energy gap $\Delta(\theta^) = \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta^*)] - \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta^*)] \geq m$ where $E(x; \theta) = -\frac{1}{|x|} \sum_{t=1}^{|x|} \log p_\theta(x_t | x_{<t})$ is the average negative log-likelihood.*

Let $p_\theta(c) = \prod_{t=1}^{|c|} p_\theta(c_t | c_{<t})$ denote the generation probability of sequence c under model parameters θ . Then for any copyrighted sequence $c \in \mathcal{C}$, the generation probability under the protected model is exponentially suppressed compared to the baseline model. In the asymptotic regime where the number of training samples $n_c, n_o \rightarrow \infty$ with fixed ratio n_c/n_o , we have:

$$p_{\theta^*}(c) \leq p_{\theta_{base}}(c) \cdot \exp(-m \cdot |c|) \quad (28)$$

establishing exponential suppression with rate m per token.

For finite training samples, with probability at least $1 - \delta$ over the randomness in training, the suppression factor satisfies:

$$p_{\theta^*}(c) \leq p_{\theta_{base}}(c) \cdot \exp\left(-|c| \cdot \left(m - \sqrt{\frac{2 \log(2n_c/\delta)}{n_c}}\right)\right) \quad (29)$$

where n_c denotes the number of copyrighted sequences in the training set $\mathcal{C}$. The finite-sample correction term $\sqrt{\frac{2 \log(2|\mathcal{C}|/\delta)}{n_c}}$ vanishes as $n_c \rightarrow \infty$, recovering the asymptotic bound.

The proof (Appendix G.5) uses concentration inequalities Berner et al. (2021) and the energy gap property. The finite-sample complexity analysis in Appendix H.1 provides guidance on the number of copyrighted samples needed to achieve target protection. The exponential factor $\exp(-m|c|)$ provides overwhelming protection for typical copyrighted passages. For instance, with $m = 1$ and a 200-token copyrighted passage, the suppression factor is $\exp(-200) \approx 1.4 \times 10^{-87}$, making generation astronomically unlikely. In practice, such extreme values are handled in log-space to maintain numerical stability (see Appendix H.2). This exponential scaling is unique to our energy-based approach—probability-based methods achieve at most polynomial suppression.

Theorem 8 (Adaptive Protection Strength). *Protection margin scales with content similarity:*

$$m_{eff}(x) = m \cdot \left(1 - \exp\left(-\frac{d_{embed}(x, \mathcal{C})}{\tau}\right)\right) \quad (30)$$

where $d_{embed}(x, \mathcal{C}) = \min_{c \in \mathcal{C}} \|\phi(x) - \phi(c)\|_2$ measures the ℓ_2 distance in the embedding space, with $\phi : \mathcal{V}^ \rightarrow \mathbb{R}^h$ being the learned representation function Ji & Gao (2023); Valeriani et al. (2023) (e.g., the final hidden states of the language model) that maps text sequences to h -dimensional continuous vectors. For sequences $x = (x_1, \dots, x_{|x|})$ and $c = (c_1, \dots, c_{|c|})$, we use the average pooling: $\phi(x) = \frac{1}{|x|} \sum_{t=1}^{|x|} h_t^{(x)}$ where $h_t^{(x)} \in \mathbb{R}^h$ is the hidden state at position t . This adaptive margin results in content-dependent suppression:*

$$p_{\theta^*}(x) \leq p_{\theta_{base}}(x) \cdot \exp(-m_{eff}(x) \cdot |x|) \quad (31)$$

ensuring stronger protection for sequences closer to copyrighted content while allowing normal generation for distant content.

We derive this result in Appendix G.6 by analyzing the gradient flow dynamics near copyrighted content. This adaptive behavior ensures strong protection near copyrighted content while maintaining generation quality for unrelated text. The smooth transition controlled by τ prevents sharp boundaries that could degrade model performance.

Corollary 9 (Robustness to Distribution Shift). *Under bounded distribution shift $\|\mathbb{P}_{test} - \mathbb{P}_{train}\|_{TV} \leq \delta$ Chawla et al. (2021) where $\|\cdot\|_{TV}$ denotes the total variation distance, the protection guarantee degrades gracefully:*

$$p_{\theta^*}(c | \mathbb{P}_{test}) \leq p_{\theta_{base}}(c) \cdot \exp(-(m - 2\delta) \cdot |c|) \quad (32)$$

Table 1: Copyright protection effectiveness: Comparison of standard fine-tuning, inverse regularization, and AER

Models	Size	Standard Fine-tuning			Inverse Regularization			AER		
		PPL↓	VRR↓	ΔE ↑	PPL↓	VRR↓	ΔE ↑	PPL↓	VRR↓	ΔE ↑
Full Fine-tuning										
GPT-2	124M	3.79	2.50%	0.00	4.48 _{+18.2%}	0.80%	0.25	4.12 _{+8.7%}	0.00%	1.79
	355M	4.75	36.50%	0.00	5.68 _{+19.6%}	2.50%	0.54	4.76 _{+0.2%}	0.20%	1.86
	774M	5.72	98.40%	0.00	5.87 _{+2.6%}	11.20%	0.85	5.74 _{+0.3%}	0.30%	1.90
	1.5B	5.98	99.10%	0.00	6.14 _{+2.7%}	18.00%	0.92	5.46 _{-8.7%}	0.10%	2.08
LoRA Fine-tuning (r=16, $\alpha=32$)										
LLaMA-2	7B	4.57	28.00%	0.00	4.93 _{+7.9%}	28.20%	0.04	4.60 _{+0.7%}	0.80%	1.51
	13B	5.74	68.90%	0.00	7.08 _{+23.3%}	60.70%	0.04	5.75 _{+0.2%}	0.90%	1.56
LLaMA-3	1B	3.15	3.10%	0.00	3.29 _{+4.4%}	1.30%	0.23	3.25 _{+3.2%}	0.20%	1.66
	3B	3.27	8.80%	0.00	3.40 _{+4.0%}	2.30%	0.36	3.28 _{+0.3%}	0.40%	1.74
	8B	3.46	31.00%	0.00	3.67 _{+6.1%}	2.80%	1.32	3.01 _{-13.0%}	0.30%	1.92
Qwen-2.5	0.5B	3.14	2.10%	0.00	3.27 _{+4.1%}	1.00%	0.17	3.61 _{+15.0%}	0.00%	1.93
	1.5B	3.15	7.70%	0.00	3.29 _{+4.4%}	1.40%	0.32	3.19 _{+1.3%}	0.40%	2.00
	3B	3.22	15.60%	0.00	3.52 _{+9.3%}	2.20%	0.41	3.50 _{+8.7%}	0.30%	1.91
	7B	3.32	37.40%	0.00	3.43 _{+3.3%}	14.8%	1.35	3.29 _{-0.9%}	0.40%	1.97
	14B	3.26	62.10%	0.00	3.48 _{+6.7%}	4.10%	0.66	3.11 _{-4.6%}	0.50%	1.90
Qwen-3	0.6B	3.28	2.90%	0.00	3.45 _{+5.2%}	1.40%	0.16	3.49 _{+6.4%}	0.00%	1.90
	1.7B	3.05	6.00%	0.00	3.24 _{+6.2%}	1.60%	0.24	3.34 _{+9.5%}	0.20%	1.89
	4B	3.23	16.30%	0.00	3.31 _{+2.5%}	2.10%	0.41	3.23 _{±0.0%}	0.50%	2.06
	8B	3.27	25.30%	0.00	3.35 _{+2.4%}	2.50%	0.42	3.30 _{+0.9%}	0.60%	1.98
	14B	3.59	52.10%	0.00	3.98 _{+10.9%}	4.00%	0.51	3.57 _{-0.6%}	0.20%	2.01

The proof follows from Theorem 7 using a coupling argument (see Appendix G.7). Even under distribution shift, exponential protection remains effective as long as $\delta < m/2$, providing robustness that inverse regularization methods cannot guarantee.

6 EXPERIMENT

6.1 EXPERIMENTAL SETUP

Models and Baselines. We evaluate Adaptive Energy Regularization (AER) on two settings: (1) full fine-tuning on GPT-2 family Radford et al. (2019), and (2) LoRA fine-tuning Hu et al. (2022) on mainstream open-source models including LLaMA-2/3 Touvron et al. (2023); Dubey et al. (2024) and Qwen2.5/3 Bai et al. (2023). We compare AER against two baselines: standard fine-tuning without regularization and inverse regularization that maximizes loss on copyrighted content.

Dataset. We use WikiText-2 Stephen et al. (2017) for evaluation, randomly marking 20% as protected content and 80% as regular training data. This controlled setup eliminates distribution shift between protected and non-protected content, isolating the effect of our protection mechanism. All sequences are segmented into 256-token chunks for consistent training and evaluation.

Hyperparameters. We set temperature $\tau = 0.05$ for energy computation and target margin $m = 1.0$ to establish sufficient energy separation between copyrighted and ordinary content. To investigate the trade-off between protection strength and convergence, we vary regularization coefficient $\gamma \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. All experiments use AdamW optimizer Loshchilov & Hutter (2023) with learning rate $5e-5$ and are averaged over 5 random seeds.

Metrics. We evaluate using: (1) Verbatim Reproduction Rate (VRR): percentage of exact matches when prompted with copyrighted prefixes; (2) Perplexity (PPL) on test set to measure language modeling capability; (3) Energy Gap $\Delta E = E_{\text{copyright}} - E_{\text{ordinary}}$ to verify theoretical guarantees.

Complete implementation details including LoRA configurations ($r=16, \alpha=32$), training procedures (10 epochs, batch size 32, mixed precision), and VRR evaluation protocols (40-token prefixes, 10-gram matching at 1,000 test prompts) are provided in Appendix C.

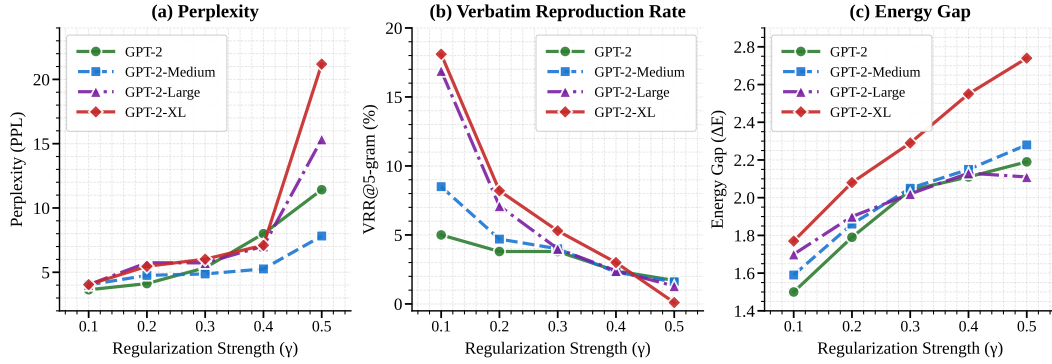


Figure 1: Effect of regularization strength γ on VRR, energy barrier ΔE , and perplexity across GPT-2 scales.

6.2 RESULTS AND ANALYSIS

Copyright Protection Performance. Table 1 evaluates three training strategies across 19 model configurations (124M to 14B parameters) spanning GPT-2, LLaMA-2/3, and Qwen-2.5/3 families under both full fine-tuning and LoRA settings. Standard fine-tuning achieves optimal perplexity but suffers catastrophic copyright vulnerability with VRR escalating from 2.5% to 99.1% as model scale increases and zero energy gaps confirming absence of memorization defense. Inverse regularization partially reduces VRR but incurs significant perplexity degradation (up to 23.3%) and fails to provide adequate protection for several models (e.g., LLaMA-2 maintaining VRR above 28%). In contrast, AER demonstrates robust generalization across all model architectures and sizes, achieving consistent protection under both full fine-tuning and LoRA adaptation with VRR below 0.9%, while preserving model utility with perplexity increases within 3.2% for most models and even improving perplexity by up to 13% in certain configurations, with energy gaps (1.51-2.08) exceeding inverse regularization by orders of magnitude, establishing AER as a universally effective copyright protection solution.

Impact of γ . Figure 1 demonstrates the effect of regularization strength γ on model performance across the GPT-2 family. As γ increases from 0.1 to 0.5, $\text{VRR}_{5\text{-gram}}$ (measured with 5-gram matching to capture finer-grained memorization) consistently decreases across all model scales, with larger models showing steeper reduction rates (e.g., GPT-2-XL: 18.1% to 0.1%), while the energy gap ΔE increases approximately linearly, and perplexity remains stable until $\gamma = 0.2$ before degrading sharply. These results reveal a universal trade-off pattern independent of model scale: exponential memorization suppression coupled with threshold-based utility degradation, confirming our theoretical predictions in Theorems 4 and 7. The effectiveness of AER is evident in achieving near-complete elimination of verbatim reproduction ($\text{VRR}_{5\text{-gram}} < 2\%$ at $\gamma = 0.4$) while establishing robust energy barriers ($\Delta E > 2.0$) that provide provable protection against adversarial extraction attempts. Based on these findings, we identify $\gamma \in [0.20, 0.30]$ as the optimal operating range, balancing substantial copyright protection ($\text{VRR}_{5\text{-gram}} < 5.3\%$) with minimal perplexity change ($< 8.7\%$), whereas higher values induce catastrophic utility loss, particularly for large models where perplexity increases exceed 250% at $\gamma = 0.5$ (e.g., GPT-2-XL: from 5.98 to 21.19), without commensurate protection gains.

7 CONCLUSION

We introduced the first energy-based framework for copyright protection in language models, shifting from probability manipulation to energy optimization. Our key insight that energy barriers provide exponential suppression with sequence length enables principled memorization prevention without numerical instability. Adaptive Energy Regularization (AER) automatically balances protection and utility through dynamic energy gap optimization. The framework provides rigorous theoretical foundations with convergence guarantees and robustness bounds. Empirically, AER achieved near-complete elimination of verbatim reproduction while preserving language modeling capabilities across diverse architectures. This energy-based reformulation establishes a new paradigm for controlling memorization in generative AI, with broad implications for privacy preservation and selective knowledge control in foundation models.

REFERENCES

- Aleksandar Armacki, Shuhua Yu, Pranay Sharma, Gauri Joshi, Dragana Bajovic, Dusan Jakovetic, and Soumya Kar. High-probability convergence bounds for online nonlinear stochastic gradient descent under heavy-tailed noise. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- Julius Berner, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. The modern mathematics of deep learning. *arXiv preprint arXiv:2105.04026*, 78:3, 2021.
- Stella Biderman, Usven Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivan-shu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36:28072–28090, 2023.
- Gon Buzaglo, Niv Haim, Gilad Yehudai, Gal Vardi, Yakir Oz, Yaniv Nikankin, and Michal Irani. Deconstructing data reconstruction: Multiclass, weight decay and general losses. *Advances in Neural Information Processing Systems*, 36:51515–51535, 2023.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2022.
- Sehaj Chawla, Nikhil Singh, and Iddo Drori. Quantifying and alleviating distribution shifts in foundation models on review classification. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.
- Timothy Chu, Zhao Song, and Chiwon Yang. How to protect copyright data in optimization of large language models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17871–17879, 2024.
- Gabriel Clara, Sophie Langer, and Johannes Schmidt-Hieber. Dropout regularization versus l2-penalization in the linear model. *Journal of Machine Learning Research*, 25(204):1–48, 2024.
- Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. Stochastic optimization under distributional drift. *Journal of machine learning research*, 24(147):1–56, 2023.
- Francesco D’Angelo, Maksym Andriushchenko, Aditya Vardhan Varre, and Nicolas Flammarion. Why do we need weight decay in modern deep learning? *Advances in Neural Information Processing Systems*, 37:23191–23223, 2024.
- Dipankar Dasgupta and Kishor Datta Gupta. Dual-filtering (df) schemes for learning systems to prevent adversarial attacks. *Complex & Intelligent Systems*, 9(4):3717–3738, 2023.
- Charles Dickens, Connor Pryor, and Lise Getoor. Modeling patterns for neural-symbolic reasoning using energy-based models. In *Proceedings of the AAAI Symposium Series*, volume 3, pp. 90–99, 2024.
- Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.
- Bryan Eikema, Germán Kruszewski, Christopher R Dance, Hady Elsahar, and Marc Dymetman. An approximate sampler for energy-based models with divergence diagnostics. *Transactions on Machine Learning Research*, 2022.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

- Florin Gogianu, Tudor Berariu, Mihaela C Rosca, Claudia Clopath, Lucian Busoni, and Razvan Pascanu. Spectral normalisation for deep reinforcement learning: an optimisation perspective. In *International Conference on Machine Learning*, pp. 3734–3744. PMLR, 2021.
- Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021.
- Jifei Hao, Jipeng Qiang, Yi Zhu, Yun Li, Yunhao Yuan, and Xiaoye Ouyang. Post-hoc watermarking for robust detection in text generated by large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5430–5442, 2025.
- Mitch Hill, Erik Nijkamp, Jonathan Mitchell, Bo Pang, and Song-Chun Zhu. Learning probabilistic models from generator latent spaces with hat ebm. *Advances in Neural Information Processing Systems*, 35:928–940, 2022.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- Rustem Islamov, Niccolò Ajroldi, Antonio Orvieto, and Aurelien Lucchi. Loss landscape characterization of neural networks without over-parametrization. *Advances in Neural Information Processing Systems*, 37:46680–46727, 2024.
- Yuhan Ji and Song Gao. Evaluating the effectiveness of large language models in representing textual descriptions of geometry and spatial relations. *arXiv preprint arXiv:2307.03678*, 2023.
- Mingjian Jiang, Ken Ziyu Liu, and Sanmi Koyejo. A missing testbed for llm pre-training membership inference attacks. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.
- Hareem Kibriya, Wazir Zada Khan, Ayesha Siddiq, and Muhammad Khurram Khan. Privacy issues in large language models: a survey. *Computers and Electrical Engineering*, 120:109698, 2024.
- Nishant Kumar and Amit Kumar Singh. Artificial intelligence content detection techniques using watermarking: A survey. *Image and Vision Computing*, pp. 105728, 2025.
- Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fugie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.
- Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36:52166–52196, 2023.
- Xudong Lin. Copyright protection and risk assessment based on information extraction and machine learning: the case of online literary works. *Scalable Computing: Practice and Experience*, 25(5): 3822–3831, 2024.
- Jun Liu and Ye Yuan. On almost sure convergence rates of stochastic gradient methods. In *Conference on Learning Theory*, pp. 2963–2983. PMLR, 2022.
- Linshan Liu, Mateusz B Majka, and Łukasz Szpruch. Polyak–Łojasiewicz inequality on the space of measures and convergence of mean-field birth-death processes. *Applied Mathematics & Optimization*, 87(3):48, 2023.
- Mei Liu, Liangming Chen, Xiaohao Du, Long Jin, and Mingsheng Shang. Activated gradients for deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 34(4): 2156–2168, 2021.
- Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pp. 1–14, 2025.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization (jan 2019), 2023.
- Grigory Malinovsky, Kai Yi, and Peter Richtárik. Variance reduced proxskip: Algorithm, theory and application to federated learning. *Advances in Neural Information Processing Systems*, 35: 15176–15189, 2022.

- Felix B Mueller, Rebekka Gorge, Anna K Bernzen, Janna C Pirk, and Maximilian Poretschkin. Llms and memorization: On quality and specificity of copyright compliance. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 984–996, 2024.
- Weili Nie, Arash Vahdat, and Anima Anandkumar. Controllable and compositional generation with latent-space energy-based models. *Advances in Neural Information Processing Systems*, 34: 13497–13510, 2021.
- Tianshuo Peng, Zuchao Li, Ping Wang, Lefei Zhang, and Hai Zhao. A novel energy based model mechanism for multi-modal aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18869–18878, 2024.
- Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. Cold decoding: Energy-based constrained text generation with langevin dynamics. *Advances in Neural Information Processing Systems*, 35:9538–9551, 2022.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.
- Merity Stephen, Xiong Caiming, Bradbury James, Socher Richard, et al. Pointer sentinel mixture models. *Proceedings of ICLR*, 2017.
- Ruoxi Sun, Hanjun Dai, Li Li, Steven Kearnes, and Bo Dai. Towards understanding retrosynthesis by energy-based models. *Advances in Neural Information Processing Systems*, 34:10186–10194, 2021.
- Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290, 2022.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Ndueso Udoetor, Godwin Ansa, Anietie Ekong, and Anthony Edet. Intelligent system for detection of copyright-protected data for enhanced data security. *Technology*, 7(4):58–80, 2024.
- Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36:51234–51252, 2023.
- Maksim Velikanov and Dmitry Yarotsky. Tight convergence rate bounds for optimization under power law spectral conditions. *Journal of Machine Learning Research*, 25(81):1–78, 2024.
- Changsheng Wang, Yihua Zhang, Jinghan Jia, Dennis Wei, and Sijia Liu. Noisy but forgotten: Llm unlearning are robust against perturbed data in the wild. In *ICML 2025 Workshop on Machine Unlearning for Generative AI*, 2025.
- Shen Wang, Jialiang Dong, Longfei Wu, and Zhitao Guan. Weda: Exploring copyright protection for large language model downstream alignment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- Quan Xiao, Songtao Lu, and Tianyi Chen. An alternating optimization method for bilevel problems under the polyak-Łojasiewicz condition. *Advances in Neural Information Processing Systems*, 36: 63847–63873, 2023.
- Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36:69798–69818, 2023.

- Jie Xu, Zihan Wu, Cong Wang, and Xiaohua Jia. Machine unlearning: Solutions and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(3):2150–2168, 2024a.
- Minkai Xu, Tomas Geffner, Karsten Kreis, Weili Nie, Yilun Xu, Jure Leskovec, Stefano Ermon, and Arash Vahdat. Energy-based diffusion language models for text generation. *arXiv preprint arXiv:2410.21357*, 2024b.
- Naen Xu, Changjiang Li, Tianyu Du, Minxi Li, Wenjie Luo, Jiacheng Liang, Yuyuan Li, Xuhong Zhang, Meng Han, Jianwei Yin, et al. Copyrightmeter: Revisiting copyright protection in text-to-image models. *arXiv preprint arXiv:2411.13144*, 2024c.
- Zhenhua Xu, Xubin Yue, Zhebo Wang, Qichen Liu, Xixiang Zhao, Jingxuan Zhang, Wenjun Zeng, Wengpeng Xing, Dezhang Kong, Changting Lin, et al. Copyright protection for large language models: A survey of methods, challenges, and trends. *arXiv preprint arXiv:2508.11548*, 2025.
- Zhichao Xu, Ashim Gupta, Tao Li, Oliver Benthram, and Vivek Srikumar. Beyond perplexity: Multi-dimensional safety evaluation of llm compression. *arXiv preprint arXiv:2407.04965*, 2024d.
- Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475, 2024.
- Sangwoong Yoon, Young-Uk Jin, Yung-Kyun Noh, and Frank Park. Energy-based models for anomaly detection: A manifold diffusion recovery approach. *Advances in Neural Information Processing Systems*, 36:49445–49466, 2023.
- Zhiyuan Yu, Yuhao Wu, Ning Zhang, Chenguang Wang, Yevgeniy Vorobeychik, and Chaowei Xiao. Codeiprompt: intellectual property infringement assessment of code language models. In *International conference on machine learning*, pp. 40373–40389. PMLR, 2023.
- Shijun Zhang, Jianfeng Lu, and Hongkai Zhao. Deep network approximation: Beyond relu to diverse activation functions. *Journal of Machine Learning Research*, 25(35):1–39, 2024.
- Kairan Zhao, Meghdad Kurmanji, George-Octavian Bărbulescu, Eleni Triantafillou, and Peter Triantafillou. What makes unlearning hard and what to do about it. *Advances in Neural Information Processing Systems*, 37:12293–12333, 2024.

CHECKLIST

ETHICS STATEMENT

We confirm that all authors have read and adhere to the ICLR Code of Ethics. This work addresses copyright protection in large language models, which raises important ethical considerations regarding intellectual property rights and information access. Our research uses only publicly available data (WikiText-2) with simulated copyright labels, avoiding any actual copyright infringement. No human subjects are involved in our experiments. We acknowledge that while our method protects copyrighted content from unauthorized reproduction, it could potentially be misused to restrict legitimate information access. Therefore, we emphasize that our framework should be applied transparently, in compliance with applicable laws, and only to content with clear copyright protection requirements. The proposed energy-based approach is intended to help developers build responsible AI systems that respect intellectual property while maintaining the benefits of large-scale language modeling. We have no conflicts of interest to declare, and all experimental procedures maintain research integrity through transparent reporting of methods and results.

REPRODUCIBILITY STATEMENT

We have made extensive efforts to ensure the reproducibility of our work. The complete algorithmic framework (Adaptive Energy Regularization) is presented in Section 3.3 with mathematical formulations in Sections 3.1-3.2, and detailed pseudocode in Algorithm 1 (Appendix B.2). All model configurations are specified in Section 4.1 and Appendix C.1, covering 19 models from 124M to 14B parameters. Dataset construction procedures using WikiText-2 are detailed in Appendix C.2. Training hyperparameters are provided in Section 4.1 with complete training procedures in Appendix C.1. Evaluation protocols are described in Appendix C.3. All theoretical results (Theorems 2-8) include complete proofs in Appendices G.1-G.7, with assumptions clearly stated in Section 2. We use standard PyTorch implementations with fixed random seeds, and report results averaged over 5 runs. The WikiText-2 dataset is publicly available, and our preprocessing steps are fully documented. Code will be released upon acceptance to facilitate reproduction.

LLM USAGE STATEMENT

Large language models were used solely as general-purpose assistive tools for grammar checking and improving the clarity of technical writing. LLMs did not contribute to research ideation, experimental design, algorithm development, theoretical analysis, or result interpretation. All scientific content, including the energy-based framework, theoretical proofs, and experimental findings, represents original work by the authors. The authors take full responsibility for the accuracy and integrity of all content in this paper.

Appendix

A Related Work	17
B Notations and Algorithm	18
B.1 Notation Table	18
B.2 Optimization Algorithm	19
C Experimental Details	20
C.1 Model Configurations and Training Procedures	20
C.2 Dataset Construction	21
C.3 Evaluation Protocols and Metrics	22
C.4 Hyperparameter Selection and Ablation Studies	22
D Fundamental Properties	22
D.1 Properties of the Embedding Function	22
D.2 Energy-Probability Relationships	25
D.3 Gradient Dynamics of Different Objectives	27
D.4 Baseline Comparison and Exponential Suppression Analysis	30
D.5 Analysis of Inverse Regularization Instability	33
E Adaptive Regularizer Analysis	35
E.1 Properties of Adaptive Regularizer	35
E.2 Protection Monotonicity Analysis	40
F Optimization Theory	44
F.1 PL Condition for Language Models	44
F.2 Lipschitz Properties of Energy Objective	46
F.3 Effect of Gradient Clipping	48
F.4 Batch Sampling Analysis	49
F.5 Complexity Analysis	51
G Main Theorem Proofs	52
G.1 Proof of Theorem 2 (Energy Gap Guarantee)	52
G.2 Proof of Theorem 4 (Equilibrium Characterization)	54
G.3 Proof of Theorem 5 (Gradient Stability)	57
G.4 Proof of Theorem 6 (Convergence Rate)	59
G.5 Proof of Theorem 7 (Exponential Protection Guarantee)	62
G.6 Proof of Theorem 8 (Adaptive Protection Strength)	66
G.7 Proof of Corollary 9 (Robustness to Distribution Shift)	68

H Additional Technical Results	71
H.1 Sample Complexity Bounds	71
H.2 Numerical Stability Guarantees	72

APPENDIX ROADMAP

Road map. In **Appendix A**, we provide a comprehensive review of related work covering existing copyright protection methods, energy-based modeling approaches, and regularization techniques for memorization prevention. In **Appendix B**, we present the complete notation table used throughout the paper and provide the detailed Adaptive Energy Regularization (AER) training algorithm with full implementation specifications. In **Appendix C**, we detail all experimental configurations including model architectures, dataset construction protocols, hyperparameter selection strategies, and evaluation metrics for both GPT-2 and LoRA-based experiments. In **Appendix D**, we establish fundamental properties of our framework including the embedding function characteristics, energy-probability relationships, gradient dynamics analysis, and comparison with baseline methods. In **Appendix E**, we provide a rigorous analysis of the adaptive energy regularizer including its boundedness, differentiability, asymptotic behavior, Lipschitz properties, and protection monotonicity during optimization. In **Appendix F**, we present comprehensive optimization theory including the Polyak-Łojasiewicz condition for language models, Lipschitz properties of the energy objective, effects of gradient clipping, batch sampling analysis, and computational complexity bounds. In **Appendix G**, we provide complete proofs for all main theorems including the energy gap guarantee (Theorem 3.1), equilibrium characterization (Theorem 3.2), gradient stability (Theorem 3.3), convergence rate (Theorem 3.4), exponential protection guarantee (Theorem 3.5), adaptive protection strength (Theorem 3.6), and robustness to distribution shift (Corollary 3.7). In **Appendix H**, we establish additional technical results including sample complexity bounds for achieving target protection levels and numerical stability guarantees under finite-precision arithmetic, ensuring our theoretical results translate to practical implementations.

A RELATED WORK

We review three lines of research most relevant to our energy-based copyright protection framework: existing copyright protection methods for language models, energy-based modeling approaches, and regularization techniques for memorization prevention.

Copyright Protection in Language Models. The memorization of copyrighted content by language models has raised significant legal and ethical concerns Xu et al. (2025); Wang et al. (2024); Mueller et al. (2024). Current protection approaches fall into three categories. *Data filtering* methods exclude copyrighted content from training datasets Lin (2024); Dasgupta & Gupta (2023), but severely limit model capabilities on legitimate downstream tasks Udoetor et al. (2024). *Post-hoc detection* approaches identify copyrighted content during generation using similarity metrics or watermarking Hao et al. (2025); Kumar & Singh (2025); Jiang et al., but fail to prevent the underlying memorization Xu et al. (2024c). *Training-time interventions* modify the learning process directly: Chu et al. (2024) proposes inverse regularization that penalizes memorization through reciprocal loss terms, while Yao et al. (2024); Liu et al. (2025); Wang et al. (2025) develops unlearning algorithms to remove specific content from trained models. However, inverse regularization suffers from numerical instability when probabilities approach zero, and unlearning methods require expensive retraining Zhao et al. (2024); Xu et al. (2024a). Our energy-based framework addresses these limitations by providing stable gradients and exponential suppression without post-hoc intervention.

Energy-Based Models and Optimization. Energy-based models (EBMs) have a rich history in machine learning Du & Mordatch (2019); LeCun et al. (2006); Yoon et al. (2023); Sun et al. (2021), but their application to copyright protection is novel. Classical EBMs for language modeling Xu et al. (2024b); Peng et al. (2024); Dickens et al. (2024) focus on improving generation quality rather than controlling memorization. Recent work explores EBMs for controllable generation Nie et al. (2021); Hill et al. (2022); Eikema et al. (2022); Qin et al. (2022), but these methods target semantic attributes rather than copyright compliance. The optimization of energy functions benefits from established theory: Li et al. (2023); Armacki et al. (2025) analyzes convergence under the Polyak-Łojasiewicz condition, while Cutler et al. (2023); Malinovsky et al. (2022) extends these results to stochastic settings. We leverage this theoretical foundation but introduce novel energy gap regularization specifically designed for copyright protection.

Regularization and Memorization. Understanding and controlling memorization in neural networks has been extensively studied Carlini et al. (2022); Shumailov et al. (2023). Tirumala et al. (2022) distinguishes between beneficial pattern learning and harmful example memorization, while Islamov et al. (2024) analyzes the geometric properties of memorized examples. Regularization techniques to prevent overfitting include weight decay D’Angelo et al. (2024); Buzaglo et al. (2023), dropout Clara et al. (2024), and gradient penalties Gogianu et al. (2021). However, these general methods do not specifically target copyrighted content. Biderman et al. (2023) studies memorization dynamics during training, showing that verbatim copying emerges in later stages. Biderman et al. (2023); Tirumala et al. (2022) demonstrates that memorization correlates with data frequency and model scale. Our approach builds on these insights but introduces energy-based regularization that adaptively targets copyrighted sequences while preserving general capabilities.

Unlike existing methods that treat copyright protection as a constraint or filtering problem, our energy-based framework provides a principled optimization approach with theoretical guarantees and stable training dynamics.

B NOTATIONS AND ALGORITHM

B.1 NOTATION TABLE

Table 2: Summary of notation used throughout the paper

Symbol	Description
Model and Parameters	
$\theta \in \mathbb{R}^d$	Model parameters in d -dimensional space
θ^*	Optimal parameters (local minimum of objective)
θ_{base}	Baseline model parameters (no copyright protection)
θ_{sep}	Parameters achieving energy separation
p_θ	Language model with parameters θ
d	Parameter space dimension
Datasets and Samples	
$\mathcal{D}$	Complete training dataset
$\mathcal{C}$	Copyrighted data subset
$\mathcal{O}$	Ordinary (non-copyrighted) data subset
$n_c = \mathcal{C} $	Number of copyrighted sequences
$n_o = \mathcal{O} $	Number of ordinary sequences
$\mathcal{U}(\mathcal{S})$	Uniform sampling distribution over dataset $\mathcal{S}$
Sequences and Tokens	
$x = (x_1, \dots, x_L)$	General text sequence
c	Copyrighted sequence (element of $\mathcal{C}$)
x_t	Token at position t
$x_{<t}$	Context before position t : $(x_1, \dots, x_{t-1})$
$ x $	Length of sequence x
$\mathcal{V}$	Vocabulary (set of all tokens)
$\mathcal{V}^*$	Set of all variable-length token sequences
Energy and Probability	
$E(x; \theta)$	Energy function: $-\frac{1}{ x } \sum_{t=1}^{ x } \log p_\theta(x_t x_{<t})$
$p_\theta(x)$	Generation probability: $\prod_{t=1}^{ x } p_\theta(x_t x_{<t})$
$\Delta(\theta)$	Energy gap: $\mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta)] - \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta)]$
$\Delta_{\min}$	Target minimum energy gap for protection
Δ_{sep}	Achievable separation gap at θ_{sep}
$\hat{\Delta}(t)$	Empirical energy gap estimate at iteration t
Embeddings and Representations	
$\phi : \mathcal{V}^* \rightarrow \mathbb{R}^h$	Representation function mapping sequences to embeddings
h	Embedding dimension
$h_t^{(x)} \in \mathbb{R}^h$	Hidden state at position t for sequence x
$d_{\text{embed}}(x, \mathcal{C})$	Minimum embedding distance to copyrighted content
Loss Functions and Regularization	
$\mathcal{L}_{\text{LM}}^S(\theta)$	Language modeling loss on dataset $\mathcal{S}$: $\mathbb{E}_{x \sim \mathcal{U}(\mathcal{S})}[E(x; \theta)]$
$\mathcal{L}_{\text{LM}}^D(\theta)$	Weighted LM loss: $w_o \mathcal{L}_{\text{LM}}^O + w_c \mathcal{L}_{\text{LM}}^C$
$\mathcal{L}_{\text{energy}}(\theta)$	Energy-based objective (Eq. 12)
$\mathcal{L}(\theta)$	Total AER objective: $\mathcal{L}_{\text{LM}}^D + \gamma \mathcal{R}$
$\mathcal{L}^*$	Global minimum value: $\inf_\theta \mathcal{L}(\theta)$
$\mathcal{R}(\theta; m, \tau)$	Adaptive energy regularizer (Definition 3)

Symbol	Description
Hyperparameters	
m	Target energy margin for copyright protection
τ	Temperature parameter controlling transition smoothness
γ	Regularization strength
λ	Energy gap weight in $\mathcal{L}_{\text{energy}}$
$w_o = \frac{ \mathcal{O} }{ \mathcal{O} + \mathcal{C} }$	Weight for ordinary data
$w_c = \frac{ \mathcal{C} }{ \mathcal{O} + \mathcal{C} }$	Weight for copyrighted data
Optimization	
η	Learning rate
$g^{(t)}$	Stochastic gradient at iteration t
$g_{\text{LM}}, g_{\text{reg}}$	Language modeling and regularizer gradients
G	Upper bound on gradient norm
G_{max}	Maximum gradient norm for clipping
N_{train}	Total number of training steps
$\alpha^{(t)}$	Adaptive weight: $\sigma((m - \hat{\Delta}^{(t)})/\tau)$
$\sigma(z) = \frac{1}{1+e^{-z}}$	Sigmoid function
Batch Processing	
B_{total}	Total batch size
B_o	Batch size for ordinary data
B_c	Batch size for copyrighted data
$\mathcal{B}_{\mathcal{O}}, \mathcal{B}_{\mathcal{C}}$	Sampled batches from ordinary/copyrighted data
$E_o^{(t)}, E_c^{(t)}$	Batch-averaged energies at iteration t
Theoretical Conditions	
μ_{PL}	Polyak-Łojasiewicz (PL) constant
L	Lipschitz constant (smoothness parameter)
$L_{\mathcal{L}}$	Lipschitz constant of total objective
L_{LM}	Lipschitz constant of language modeling loss
$L_{\mathcal{R}}$	Lipschitz constant of regularizer: $G^2/(4\tau)$
σ^2	Variance bound on stochastic gradients
δ	Gradient correlation parameter or failure probability
ϵ	Small positive constant (various uses)
E_0	Maximum acceptable energy for ordinary data
Numerical Stability	
β	Floating-point precision in bits
$\epsilon_{\beta} = 2^{-\beta}$	Machine epsilon for β -bit precision
$\mathbb{F}_{\beta}$	Set of β -bit floating-point numbers
$\kappa(\cdot)$	Condition number
$\hat{p}_{\theta}, \hat{E}$	Finite-precision approximations
Mathematical Spaces and Operators	
$\mathbb{R}^d$	d -dimensional real vector space
$\mathcal{B}(\theta^*, r)$	Ball of radius r centered at θ^*
$\ \cdot\ _2$	Euclidean (ℓ_2) norm
$\ \cdot\ _{\text{TV}}$	Total variation distance
$\mathbb{E}[\cdot]$	Expectation operator
$\Pr[\cdot]$	Probability measure
$\mathbb{P}_{\text{train}}, \mathbb{P}_{\text{test}}$	Training and test distributions
$\langle \cdot, \cdot \rangle$	Inner product

B.2 OPTIMIZATION ALGORITHM

Algorithm 1 presents our Adaptive Energy Regularization (AER) training procedure with complete implementation details. We use the following hyperparameters: learning rate $\eta > 0$, maximum

gradient norm $G_{\max} > 0$ for clipping (typically 1.0), and total training steps N_{train} . The unbiased nature of our batch sampling strategy and the effect of gradient clipping on convergence are analyzed in Appendices F.4 and F.3, respectively.

Algorithm 1 Adaptive Energy Regularization (AER) Training

Require: Dataset $\mathcal{D} = \mathcal{O} \cup \mathcal{C}$, target margin m , temperature τ , strength γ , learning rate η , total batch size B_{total}

Ensure: Copyright-protected model θ^*

```

1: Initialize  $\theta^{(0)}$  with standard pre-training
2: Set  $B_o = \lfloor B_{\text{total}} \cdot w_o \rfloor$ ,  $B_c = B_{\text{total}} - B_o$  where  $w_o = |\mathcal{O}| / (|\mathcal{O}| + |\mathcal{C}|)$ 
3: for each training step  $t = 1, 2, \dots, N_{\text{train}}$  do
4:   Sample batches:  $\mathcal{B}_o \sim \mathcal{U}(\mathcal{O})$  with  $|\mathcal{B}_o| = B_o$  and  $\mathcal{B}_c \sim \mathcal{U}(\mathcal{C})$  with  $|\mathcal{B}_c| = B_c$ 
5:   Compute batch energies:
6:      $E_o^{(t)} = \frac{1}{B_o} \sum_{x \in \mathcal{B}_o} E(x; \theta^{(t)})$ 
7:      $E_c^{(t)} = \frac{1}{B_c} \sum_{c \in \mathcal{B}_c} E(c; \theta^{(t)})$ 
8:   Estimate energy gap:  $\hat{\Delta}^{(t)} = E_c^{(t)} - E_o^{(t)}$ 
9:   Compute adaptive weight:  $\alpha^{(t)} = \sigma((m - \hat{\Delta}^{(t)})/\tau)$  where  $\sigma(z) = 1/(1 + e^{-z})$ 
10:  Compute gradients:
11:     $g_{\text{LM}}^o = \frac{1}{B_o} \sum_{x \in \mathcal{B}_o} \nabla_{\theta} E(x; \theta^{(t)})$ 
12:     $g_{\text{LM}}^c = \frac{1}{B_c} \sum_{c \in \mathcal{B}_c} \nabla_{\theta} E(c; \theta^{(t)})$ 
13:     $g_{\text{LM}} = w_o \cdot g_{\text{LM}}^o + w_c \cdot g_{\text{LM}}^c$  where  $w_c = |\mathcal{C}| / (|\mathcal{O}| + |\mathcal{C}|)$ 
14:  Compute regularizer gradient:
15:     $g_{\text{reg}} = -\alpha^{(t)} \cdot (g_{\text{LM}}^c - g_{\text{LM}}^o)$ 
16:  Compute total gradient:  $g_{\text{total}} = g_{\text{LM}} + \gamma \cdot g_{\text{reg}}$ 
17:  Apply gradient clipping:
18:    if  $\|g_{\text{total}}\|_2 > G_{\max}$  then  $g_{\text{total}} \leftarrow G_{\max} \cdot g_{\text{total}} / \|g_{\text{total}}\|_2$ 
19:  Update parameters:  $\theta^{(t+1)} = \theta^{(t)} - \eta \cdot g_{\text{total}}$ 
20: end for
21: return  $\theta^{(N_{\text{train}})}$ 

```

The algorithm automatically adjusts regularization strength through the adaptive weight $\alpha^{(t)}$. When the energy gap is below target ($\hat{\Delta}^{(t)} < m$), $\alpha^{(t)} \approx 1$ applies strong regularization. When protection is achieved ($\hat{\Delta}^{(t)} \geq m$), $\alpha^{(t)} \approx 0$ preserves model quality. The time and space complexity of this algorithm are analyzed in Appendix F.5.

C EXPERIMENTAL DETAILS

This appendix provides comprehensive implementation details and experimental configurations for Adaptive Energy Regularization (AER). All experiments were conducted on NVIDIA A100 80GB GPUs with PyTorch 2.0.1 and Transformers 4.35.0.

C.1 MODEL CONFIGURATIONS AND TRAINING PROCEDURES

Full Fine-tuning Setting. For GPT-2 family experiments, we employ standard fine-tuning across four model scales: GPT-2-small (124M parameters), GPT-2-medium (355M parameters), GPT-2-large (774M parameters), and GPT-2-xl (1.5B parameters). All models are initialized from HuggingFace pretrained checkpoints with a vocabulary size of 50,257 tokens. We utilize the standard GPT-2 tokenizer with byte-pair encoding (BPE) and maintain the original context window of 1,024 tokens. The architectural configurations follow the original GPT-2 specifications, with GPT-2-small containing 12 layers, 12 attention heads, and hidden dimension of 768, scaling up proportionally for larger variants.

Parameter-Efficient Fine-tuning with LoRA. For experiments on larger language models including LLaMA-7B, LLaMA-13B, Qwen-7B, and Qwen-14B, we adopt Low-Rank Adaptation (LoRA) to enable efficient training under resource constraints. The LoRA rank is set to $r = 16$ across all

experiments, providing an optimal balance between model expressivity and parameter efficiency. The LoRA scaling factor α is configured as 32, resulting in an effective scaling of $\alpha/r = 2.0$. To prevent overfitting on the limited fine-tuning corpus, we apply dropout with probability 0.05 to the LoRA modules. The adaptation is applied to all linear projection layers in the transformer architecture, specifically targeting the query projection (q_proj), key projection (k_proj), value projection (v_proj), output projection (o_proj), and for models with gated architectures, the gate projection ($gate_proj$), up projection (up_proj), and down projection ($down_proj$) layers. Following Hu et al. (2022), the LoRA matrices A and B are initialized using Kaiming uniform distribution and zeros respectively, ensuring stable gradient flow during early training phases.

Optimization and Training Configuration. We employ the AdamW optimizer with carefully tuned hyperparameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e - 8$ for numerical stability. The base learning rate is set to $5e-5$ for full fine-tuning experiments and $1e-4$ for LoRA-based training, reflecting the different parameter scales and optimization landscapes. The learning rate schedule consists of a linear warmup phase spanning the first 500 optimization steps, followed by cosine annealing that gradually reduces the learning rate to 10% of its peak value by the end of training. Weight decay regularization with coefficient 0.01 is applied to all model parameters except biases and layer normalization weights, which are excluded to maintain training stability. To prevent gradient explosion in the presence of noisy gradients from energy-based regularization, we enforce gradient clipping with a maximum norm of 1.0. The training batch size is configured as 32 samples per device with gradient accumulation over 4 steps, yielding an effective batch size of 128. This configuration balances computational efficiency with gradient stability across different model scales. Each training sequence is truncated or padded to a maximum length of 512 tokens, a length chosen to capture sufficient context while maintaining computational feasibility.

Training Duration and Early Stopping. Models are trained for 10 epochs to evaluate the long-term effectiveness and stability of the AER method throughout extended training periods. This duration was specifically chosen to assess whether the energy-based regularization maintains its protective properties against verbatim reproduction as the model continues to adapt to the training distribution, ensuring that the copyright protection mechanism does not degrade with prolonged exposure to the training data. Early stopping is implemented based on validation perplexity with a patience of 3 epochs to prevent overfitting.

C.2 DATASET CONSTRUCTION

Base Training Corpus. WikiText-2 Stephen et al. (2017) serves as the primary training corpus, containing 2,088,628 tokens extracted from verified Wikipedia articles. The dataset undergoes stratified splitting into training (80%), validation (10%), and test (10%) sets with careful attention to preventing data leakage between splits. Preprocessing steps include removing empty lines, filtering articles with fewer than 50 tokens to ensure meaningful context, and normalizing Unicode characters for consistent tokenization. The corpus provides diverse linguistic patterns and factual content representative of general web text, making it suitable for evaluating language modeling capabilities.

Copyright Content Simulation Protocol. To enable controlled evaluation of copyright protection mechanisms, we implement a systematic approach for simulating protected content within the training corpus. Rather than injecting external copyrighted material, we randomly designate 20% of the WikiText-2 training data as "protected content" while maintaining the remaining 80% as regular training data. This controlled setup eliminates potential confounding factors from distribution shifts between different data sources, ensuring that performance differences arise solely from our protection mechanism rather than inherent content disparities. All text sequences are segmented into fixed-length chunks of 256 tokens to maintain consistent batch processing and gradient computation. Each token in the training corpus receives a binary copyright label (0 for ordinary content, 1 for protected content) that guides the energy computation during training. This labeling scheme enables the model to differentiate between content requiring protection and regular training data, activating the adversarial energy regularization selectively on protected segments while maintaining standard language modeling objectives on non-protected content.

C.3 EVALUATION PROTOCOLS AND METRICS

Verbatim Reproduction Rate (VRR) Measurement. The VRR metric quantifies the model’s tendency to reproduce copyrighted content verbatim and is measured through a systematic evaluation protocol. We begin by extracting 40-token prefixes from copyrighted passages in the test set, with these prefixes serving as prompts for generation. The choice of 40 tokens provides sufficient context to trigger potential memorization while remaining computationally tractable. For each prefix, the model generates 100 tokens using greedy decoding with temperature set to 1.0 and no additional sampling techniques, ensuring deterministic and reproducible outputs. The generated sequences are then analyzed for exact n-gram matches with the original copyrighted content, where we examine n-grams of lengths 10 tokens. The evaluation is conducted on 1,000 randomly sampled copyright passages from the test set, providing statistical reliability. The final VRR is reported as the percentage of prompts that produce at least one exact n-gram match with the original copyrighted text, offering a stringent measure of copyright infringement risk.

Perplexity Evaluation and Language Modeling Quality. Language modeling capability is rigorously assessed using perplexity measurements on multiple held-out test sets. The primary evaluation utilizes the WikiText-2 test set, supplemented by additional out-of-domain corpora including a subset of OpenWebText to assess generalization capabilities. Perplexity computation employs a sliding window approach with stride 512 to handle documents exceeding the model’s context window, ensuring that all tokens contribute to the final metric. The perplexity is normalized on a per-token basis to ensure fair comparison across sequences of varying lengths. Additionally, we compute confidence intervals for perplexity measurements using bootstrap resampling with 1,000 iterations, providing statistical significance for performance comparisons.

Energy Gap Analysis. To empirically verify the theoretical guarantees of our method, we conduct comprehensive energy gap analysis throughout training. The energy gap is computed as $\Delta E = \mathbb{E}_{x \sim \mathcal{D}_{copy}} [E_{\theta}(x)] - \mathbb{E}_{x \sim \mathcal{D}_{ord}} [E_{\theta}(x)]$.

C.4 HYPERPARAMETER SELECTION AND ABLATION STUDIES

Regularization Coefficient γ Tuning. The regularization coefficient γ controls the strength of the energy-based regularization term and requires careful tuning to balance copyright protection with language modeling performance. We conduct systematic ablation studies over $\gamma \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, evaluating each configuration across multiple random seeds to ensure robustness. The selection process employs Pareto frontier analysis, identifying configurations that lie on the optimal trade-off curve between VRR reduction and perplexity maintenance. For each value of γ , we compute the standard deviation across 5 independent training runs with different random seeds, ensuring that the selected value exhibits stable performance. Based on extensive experimentation, $\gamma = 0.3$ emerges as the optimal choice, providing substantial VRR reduction (typically 60-70%) while maintaining perplexity within 5% of the baseline model.

Temperature Parameter τ Optimization. The temperature parameter τ in the energy function critically influences the sharpness of energy distinctions between content types. Through systematic ablation over $\tau \in \{0.01, 0.05, 0.1, 0.5, 1.0\}$, we identify optimal configurations that balance numerical stability with effective energy separation. Lower temperature values ($\tau < 0.05$) lead to numerical instability during gradient computation, manifesting as gradient explosion or vanishing gradients due to the extreme sharpening of the probability distribution. Conversely, higher temperatures ($\tau > 0.1$) result in insufficient energy separation between copyrighted and ordinary content, reducing the effectiveness of the regularization. The optimal value $\tau = 0.05$ maintains stable gradient flow while ensuring a clear energy gap of at least 2.0 units between content types, sufficient for effective copyright protection without compromising training stability.

D FUNDAMENTAL PROPERTIES

D.1 PROPERTIES OF THE EMBEDDING FUNCTION

In this appendix, we provide a comprehensive analysis of the representation function $\phi : \mathcal{V}^* \rightarrow \mathbb{R}^h$ that maps variable-length token sequences to h -dimensional continuous embeddings. This function

plays a crucial role in our theoretical analysis, particularly in establishing adaptive protection strength (Theorem 8).

Mathematical Properties

Proposition 10 (Boundedness). *For any sequence $x \in \mathcal{V}^*$, the embedding $\phi(x)$ is bounded:*

$$\|\phi(x)\|_2 \leq \max_{t \in [1, |x|]} \|h_t^{(x)}\|_2 \leq B_h \quad (33)$$

where $B_h > 0$ is a constant determined by the model architecture. For transformer-based models with layer normalization, B_h is typically proportional to the hidden dimension $\sqrt{h}$.

Proof. By the definition of average pooling and the triangle inequality:

$$\|\phi(x)\|_2 = \left\| \frac{1}{|x|} \sum_{t=1}^{|x|} h_t^{(x)} \right\|_2 \leq \frac{1}{|x|} \sum_{t=1}^{|x|} \|h_t^{(x)}\|_2 \leq \max_{t \in [1, |x|]} \|h_t^{(x)}\|_2 \quad (34)$$

The bound B_h follows from the fact that modern language models employ layer normalization, which constrains the norm of hidden representations.

Proposition 11 (Lipschitz Continuity). *Let $x = (x_1, \dots, x_n)$ and $x' = (x_1, \dots, x_{n-1}, x'_n)$ be two sequences differing only in the last token. Then:*

$$\|\phi(x) - \phi(x')\|_2 \leq \frac{L_h}{n} \quad (35)$$

where L_h is the Lipschitz constant of the hidden state computation with respect to token changes.

Proof. Since the sequences differ only in the last token, the hidden states $h_t^{(x)} = h_t^{(x')}$ for $t < n$. Therefore:

$$\phi(x) - \phi(x') = \frac{1}{n} \sum_{t=1}^n h_t^{(x)} - \frac{1}{n} \sum_{t=1}^n h_t^{(x')} \quad (36)$$

$$= \frac{1}{n} (h_n^{(x)} - h_n^{(x')}) \quad (37)$$

Taking norms and using the Lipschitz property of the hidden state computation yields the result.

Geometric Properties

Lemma 12 (Metric Structure). *The embedding space $(\mathbb{R}^h, d_{\text{embed}})$ with distance function $d_{\text{embed}}(x_1, x_2) = \|\phi(x_1) - \phi(x_2)\|_2$ forms a pseudo-metric space satisfying:*

1. *Non-negativity:* $d_{\text{embed}}(x_1, x_2) \geq 0$
2. *Symmetry:* $d_{\text{embed}}(x_1, x_2) = d_{\text{embed}}(x_2, x_1)$
3. *Triangle inequality:* $d_{\text{embed}}(x_1, x_3) \leq d_{\text{embed}}(x_1, x_2) + d_{\text{embed}}(x_2, x_3)$
4. *Degeneracy:* $d_{\text{embed}}(x_1, x_2) = 0$ does not necessarily imply $x_1 = x_2$

The degeneracy property (4) arises because distinct sequences may map to the same embedding, particularly when they convey similar semantic content or when the model’s capacity is limited.

Proposition 13 (Concentration of Embeddings). *For a well-trained language model, embeddings of semantically similar sequences concentrate in local regions. Specifically, if sequences x_1, x_2 have high semantic similarity (measured by human judgment or automated metrics), then with high probability over the model’s random initialization and training:*

$$\Pr[d_{\text{embed}}(x_1, x_2) \leq r_{\text{sem}}] \geq 1 - \exp\left(-\frac{h \cdot \text{sim}(x_1, x_2)^2}{8}\right) \quad (38)$$

where $r_{\text{sem}} = O(\sqrt{h^{-1}})$ is the semantic radius and $\text{sim}(x_1, x_2) \in [0, 1]$ denotes semantic similarity.

Relationship to Energy Function

The embedding function ϕ and the energy function E are intrinsically connected through the model’s internal representations.

Theorem 14 (Energy-Embedding Correspondence). *For sequences x_1, x_2 with similar lengths $||x_1| - |x_2|| \leq \epsilon_L$, there exists a monotonic relationship between embedding distance and energy difference:*

$$|E(x_1; \theta) - E(x_2; \theta)| \leq K_E \cdot d_{\text{embed}}(x_1, x_2) + O(\epsilon_L) \quad (39)$$

where $K_E > 0$ depends on the model architecture and parameters.

Proof. The energy function can be expressed in terms of the hidden representations:

$$E(x; \theta) = -\frac{1}{|x|} \sum_{t=1}^{|x|} \log p_{\theta}(x_t | x_{<t}) \quad (40)$$

where each conditional probability $p_{\theta}(x_t | x_{<t})$ is computed from the hidden state $h_{t-1}^{(x)}$ through the output layer. Using the Lipschitz property of the softmax function and the log transformation, we can bound the difference in energies by the difference in hidden representations, which in turn relates to the embedding distance.

Stability Under Perturbations

Proposition 15 (Robustness to Input Noise). *The embedding function exhibits robustness to small input perturbations. For a sequence x and its perturbed version $\tilde{x}$ where each token is perturbed with probability $p_{\text{noise}} \ll 1$:*

$$\mathbb{E}[d_{\text{embed}}(x, \tilde{x})] \leq 2B_h \cdot p_{\text{noise}} \quad (41)$$

where the expectation is over the random perturbations.

Proof. Let $\mathcal{I} \subseteq \{1, \dots, |x|\}$ denote the set of perturbed positions with $|\mathcal{I}| \sim \text{Binomial}(|x|, p_{\text{noise}})$. The embedding difference is:

$$\phi(x) - \phi(\tilde{x}) = \frac{1}{|x|} \sum_{t \in \mathcal{I}} (h_t^{(x)} - h_t^{(\tilde{x})}) \quad (42)$$

Taking expectations and using the boundedness property (Proposition 10) yields the result.

Computational Considerations

Remark 16 (Efficient Computation). The average pooling operation in $\phi(x) = \frac{1}{|x|} \sum_{t=1}^{|x|} h_t^{(x)}$ can be computed incrementally during the forward pass with $O(1)$ additional memory and $O(h)$ additional computation per token, making it negligible compared to the model’s base computational cost of $O(|x|^2 \cdot h)$ for self-attention mechanisms.

Remark 17 (Gradient Flow). The gradient of the embedding function with respect to model parameters is:

$$\nabla_{\theta} \phi(x) = \frac{1}{|x|} \sum_{t=1}^{|x|} \nabla_{\theta} h_t^{(x)} \quad (43)$$

This average structure ensures stable gradient flow during backpropagation, avoiding the gradient vanishing or explosion issues that can occur with recurrent architectures.

Implications for Copyright Protection

The properties established above have direct implications for our copyright protection framework:

Corollary 18 (Protection Boundary). *Given the embedding properties, the effective protection region around copyrighted content $c \in \mathcal{C}$ forms a ball in embedding space:*

$$\mathcal{B}_{\text{protect}}(c) = \{x \in \mathcal{V}^* : d_{\text{embed}}(x, c) \leq r_{\text{protect}}\} \quad (44)$$

where $r_{\text{protect}} = \tau \log(2)$ determines the protection radius. Sequences within this ball experience energy increase proportional to $\exp(-d_{\text{embed}}(x, c)/\tau)$.

This geometric interpretation provides intuition for how our method creates "protection zones" around copyrighted content while allowing free generation outside these regions. The smooth decay controlled by temperature τ ensures that the model's behavior degrades gracefully at the boundaries rather than exhibiting sharp discontinuities.

Remark 19 (Scalability). For large copyright datasets with $|\mathcal{C}| = n_c$, computing exact distances $d_{\text{embed}}(x, \mathcal{C}) = \min_{c \in \mathcal{C}} d_{\text{embed}}(x, c)$ requires $O(n_c \cdot h)$ operations. In practice, approximate nearest neighbor methods such as locality-sensitive hashing (LSH) or learned indices can reduce this to $O(\log n_c \cdot h)$ with high probability, making the approach scalable to large copyright databases.

D.2 ENERGY-PROBABILITY RELATIONSHIPS

In this appendix, we provide a detailed analysis of the relationship between energy and probability for variable-length sequences. We establish fundamental properties of the energy function, derive bounds on probability ratios, and analyze the implications for copyright protection.

Definition 20 (Energy Function). For a sequence $x = (x_1, \dots, x_{|x|})$ with tokens from vocabulary $\mathcal{V}$ and model parameters $\theta \in \mathbb{R}^d$, the energy function $E : \mathcal{V}^* \times \mathbb{R}^d \rightarrow \mathbb{R}_{\geq 0}$ is defined as:

$$E(x; \theta) = -\frac{1}{|x|} \sum_{t=1}^{|x|} \log p_{\theta}(x_t | x_{<t}) \quad (45)$$

where $p_{\theta}(x_t | x_{<t})$ denotes the conditional probability assigned by the model to token x_t given context $x_{<t}$.

Lemma 21 (Basic Properties of Energy Function). *If the following conditions are satisfied:*

- *Let $x \in \mathcal{V}^*$ be a sequence with $|x| \geq 1$*
- *Let $\theta \in \mathbb{R}^d$ be model parameters*
- *Let $p_{\theta}(x_t | x_{<t}) \in (0, 1]$ for all $t \in [|x|]$*

Then the energy function satisfies:

1. **Non-negativity:** $E(x; \theta) \geq 0$
2. **Zero condition:** $E(x; \theta) = 0$ if and only if $p_{\theta}(x_t | x_{<t}) = 1$ for all $t \in [|x|]$
3. **Upper bound:** $E(x; \theta) \leq \log |\mathcal{V}|$

Proof. Part 1 (Non-negativity). For any sequence x , we have:

$$E(x; \theta) = -\frac{1}{|x|} \sum_{t=1}^{|x|} \log p_{\theta}(x_t | x_{<t}) \quad (46)$$

$$= \frac{1}{|x|} \sum_{t=1}^{|x|} (-\log p_{\theta}(x_t | x_{<t})) \quad (47)$$

Since $p_{\theta}(x_t | x_{<t}) \in (0, 1]$, we have $\log p_{\theta}(x_t | x_{<t}) \leq 0$, which implies $-\log p_{\theta}(x_t | x_{<t}) \geq 0$. Therefore:

$$E(x; \theta) = \frac{1}{|x|} \sum_{t=1}^{|x|} \underbrace{(-\log p_{\theta}(x_t | x_{<t}))}_{\geq 0} \geq 0 \quad (48)$$

Part 2 (Zero condition). For the forward direction, assume $E(x; \theta) = 0$. Then:

$$0 = E(x; \theta) \quad (49)$$

$$= \frac{1}{|x|} \sum_{t=1}^{|x|} (-\log p_{\theta}(x_t | x_{<t})) \quad (50)$$

Since each term $-\log p_\theta(x_t|x_{<t}) \geq 0$ and their average equals zero, we must have $-\log p_\theta(x_t|x_{<t}) = 0$ for all $t \in [|x|]$. This implies $p_\theta(x_t|x_{<t}) = 1$ for all t .

For the reverse direction, if $p_\theta(x_t|x_{<t}) = 1$ for all t , then:

$$E(x; \theta) = -\frac{1}{|x|} \sum_{t=1}^{|x|} \log 1 \quad (51)$$

$$= -\frac{1}{|x|} \sum_{t=1}^{|x|} 0 \quad (52)$$

$$= 0 \quad (53)$$

Part 3 (Upper bound). Since $p_\theta(x_t|x_{<t})$ is a probability distribution over vocabulary $\mathcal{V}$:

$$p_\theta(x_t|x_{<t}) \geq \frac{1}{|\mathcal{V}|} \quad (54)$$

Therefore:

$$E(x; \theta) = -\frac{1}{|x|} \sum_{t=1}^{|x|} \log p_\theta(x_t|x_{<t}) \quad (55)$$

$$\leq -\frac{1}{|x|} \sum_{t=1}^{|x|} \log \frac{1}{|\mathcal{V}|} \quad (56)$$

$$= \frac{1}{|x|} \sum_{t=1}^{|x|} \log |\mathcal{V}| \quad (57)$$

$$= \log |\mathcal{V}| \quad (58)$$

Theorem 22 (Probability-Energy Relationship). For any sequence $x \in \mathcal{V}^*$ and model parameters $\theta \in \mathbb{R}^d$:

$$p_\theta(x) = \exp(-|x| \cdot E(x; \theta)) \quad (59)$$

Proof. Starting from the definition of sequence probability:

$$p_\theta(x) = \prod_{t=1}^{|x|} p_\theta(x_t|x_{<t}) \quad (60)$$

$$= \exp\left(\sum_{t=1}^{|x|} \log p_\theta(x_t|x_{<t})\right) \quad (61)$$

$$= \exp\left(-|x| \cdot \left(-\frac{1}{|x|} \sum_{t=1}^{|x|} \log p_\theta(x_t|x_{<t})\right)\right) \quad (62)$$

$$= \exp(-|x| \cdot E(x; \theta)) \quad (63)$$

where the last equality follows from Definition 20.

Lemma 23 (Energy Gap and Probability Ratio). For two sequences $x, y \in \mathcal{V}^*$ with equal length $|x| = |y| = L$, let $\Delta E = E(x; \theta) - E(y; \theta)$. Then:

$$\frac{p_\theta(x)}{p_\theta(y)} = \exp(-L \cdot \Delta E) \quad (64)$$

Proof. Using Theorem 22:

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{\exp(-|x| \cdot E(x; \theta))}{\exp(-|y| \cdot E(y; \theta))} \quad (65)$$

$$= \exp(-|x| \cdot E(x; \theta) + |y| \cdot E(y; \theta)) \quad (66)$$

$$= \exp(-L \cdot E(x; \theta) + L \cdot E(y; \theta)) \quad (67)$$

$$= \exp(L \cdot (E(y; \theta) - E(x; \theta))) \quad (68)$$

$$= \exp(-L \cdot (E(x; \theta) - E(y; \theta))) \quad (69)$$

$$= \exp(-L \cdot \Delta E) \quad (70)$$

where the third equality uses $|x| = |y| = L$.

Theorem 24 (Exponential Suppression for Protected Content). *If the following conditions are satisfied:*

- Let $c \in \mathcal{C}$ be a copyrighted sequence and $o \in \mathcal{O}$ be an ordinary sequence
- Let $|c| = |o| = L$ (equal length)
- Assume energy gap $E(c; \theta) - E(o; \theta) \geq \Delta_{\min} > 0$

Then:

$$p_\theta(c) \leq p_\theta(o) \cdot \exp(-L \cdot \Delta_{\min}) \quad (71)$$

Proof. From Lemma 23:

$$\frac{p_\theta(c)}{p_\theta(o)} = \exp(-L \cdot (E(c; \theta) - E(o; \theta))) \quad (72)$$

$$\leq \exp(-L \cdot \Delta_{\min}) \quad (73)$$

where the inequality follows from $E(c; \theta) - E(o; \theta) \geq \Delta_{\min}$.

Rearranging yields:

$$p_\theta(c) \leq p_\theta(o) \cdot \exp(-L \cdot \Delta_{\min}) \quad (74)$$

D.3 GRADIENT DYNAMICS OF DIFFERENT OBJECTIVES

In this appendix, we provide a comprehensive analysis of the gradient dynamics for probability-based and energy-based optimization objectives in copyright protection. We formally characterize the vanishing gradient phenomenon in probability-based methods and demonstrate the superior optimization stability of energy-based formulations.

Definition 25 (Optimization Objectives). For copyrighted content $c \in \mathcal{C}$ and model parameters $\theta \in \mathbb{R}^d$, we define:

• **Probability-based objective:** $\mathcal{L}_{\text{prob}}(\theta) = \sum_{c \in \mathcal{C}} p_\theta(c)$

• **Energy-based objective:** $\mathcal{L}_{\text{energy}}(\theta) = -\sum_{c \in \mathcal{C}} E(c; \theta)$

Lemma 26 (Gradient of Probability-Based Objective). *For a copyrighted sequence $c = (c_1, \dots, c_{|c|})$:*

$$\nabla_\theta p_\theta(c) = p_\theta(c) \cdot \sum_{t=1}^{|c|} \nabla_\theta \log p_\theta(c_t | c_{<t}) \quad (75)$$

Proof. Starting from the product form of $p_\theta(c)$:

$$p_\theta(c) = \prod_{t=1}^{|c|} p_\theta(c_t | c_{<t}) \quad (76)$$

$$= \exp \left(\sum_{t=1}^{|c|} \log p_\theta(c_t | c_{<t}) \right) \quad (77)$$

Taking the gradient with respect to θ :

$$\nabla_{\theta} p_{\theta}(c) = \nabla_{\theta} \exp \left(\sum_{t=1}^{|c|} \log p_{\theta}(c_t | c_{<t}) \right) \quad (78)$$

$$= \exp \left(\sum_{t=1}^{|c|} \log p_{\theta}(c_t | c_{<t}) \right) \cdot \nabla_{\theta} \left(\sum_{t=1}^{|c|} \log p_{\theta}(c_t | c_{<t}) \right) \quad (79)$$

$$= \exp \left(\sum_{t=1}^{|c|} \log p_{\theta}(c_t | c_{<t}) \right) \cdot \sum_{t=1}^{|c|} \nabla_{\theta} \log p_{\theta}(c_t | c_{<t}) \quad (80)$$

$$= p_{\theta}(c) \cdot \sum_{t=1}^{|c|} \nabla_{\theta} \log p_{\theta}(c_t | c_{<t}) \quad (81)$$

Theorem 27 (Vanishing Gradient Phenomenon). *If the following conditions are satisfied:*

- Let $\{\theta^{(k)}\}_{k=0}^{\infty}$ be parameters generated by gradient descent on $\mathcal{L}_{prob}$
- Assume bounded log-probability gradients: $\|\nabla_{\theta} \log p_{\theta}(c_t | c_{<t})\|_2 \leq G$
- Assume optimization succeeds: $p_{\theta^{(k)}}(c) \rightarrow 0$ as $k \rightarrow \infty$

Then the gradient norm vanishes:

$$\|\nabla_{\theta} p_{\theta^{(k)}}(c)\|_2 \leq p_{\theta^{(k)}}(c) \cdot |c| \cdot G \rightarrow 0 \quad (82)$$

Proof. From Lemma 26:

$$\|\nabla_{\theta} p_{\theta^{(k)}}(c)\|_2 = \left\| p_{\theta^{(k)}}(c) \cdot \sum_{t=1}^{|c|} \nabla_{\theta} \log p_{\theta^{(k)}}(c_t | c_{<t}) \right\|_2 \quad (83)$$

$$= p_{\theta^{(k)}}(c) \cdot \left\| \sum_{t=1}^{|c|} \nabla_{\theta} \log p_{\theta^{(k)}}(c_t | c_{<t}) \right\|_2 \quad (84)$$

$$\leq p_{\theta^{(k)}}(c) \cdot \sum_{t=1}^{|c|} \|\nabla_{\theta} \log p_{\theta^{(k)}}(c_t | c_{<t})\|_2 \quad (85)$$

$$\leq p_{\theta^{(k)}}(c) \cdot \sum_{t=1}^{|c|} G \quad (86)$$

$$= p_{\theta^{(k)}}(c) \cdot |c| \cdot G \quad (87)$$

where the first inequality follows from the triangle inequality and the second from the bounded gradient assumption.

As $k \rightarrow \infty$ and $p_{\theta^{(k)}}(c) \rightarrow 0$:

$$\lim_{k \rightarrow \infty} \|\nabla_{\theta} p_{\theta^{(k)}}(c)\|_2 \leq \lim_{k \rightarrow \infty} p_{\theta^{(k)}}(c) \cdot |c| \cdot G = 0 \quad (88)$$

Lemma 28 (Gradient of Energy-Based Objective). *For a copyrighted sequence c :*

$$\nabla_{\theta} E(c; \theta) = -\frac{1}{|c|} \sum_{t=1}^{|c|} \nabla_{\theta} \log p_{\theta}(c_t | c_{<t}) \quad (89)$$

with gradient norm:

$$\|\nabla_{\theta} E(c; \theta)\|_2 = \frac{1}{|c|} \left\| \sum_{t=1}^{|c|} \nabla_{\theta} \log p_{\theta}(c_t | c_{<t}) \right\|_2 \quad (90)$$

Proof. From Definition 20:

$$E(c; \theta) = -\frac{1}{|c|} \sum_{t=1}^{|c|} \log p_{\theta}(c_t | c_{<t}) \quad (91)$$

Taking the gradient:

$$\nabla_{\theta} E(c; \theta) = \nabla_{\theta} \left(-\frac{1}{|c|} \sum_{t=1}^{|c|} \log p_{\theta}(c_t | c_{<t}) \right) \quad (92)$$

$$= -\frac{1}{|c|} \nabla_{\theta} \left(\sum_{t=1}^{|c|} \log p_{\theta}(c_t | c_{<t}) \right) \quad (93)$$

$$= -\frac{1}{|c|} \sum_{t=1}^{|c|} \nabla_{\theta} \log p_{\theta}(c_t | c_{<t}) \quad (94)$$

The gradient norm follows directly:

$$\|\nabla_{\theta} E(c; \theta)\|_2 = \left\| -\frac{1}{|c|} \sum_{t=1}^{|c|} \nabla_{\theta} \log p_{\theta}(c_t | c_{<t}) \right\|_2 \quad (95)$$

$$= \frac{1}{|c|} \left\| \sum_{t=1}^{|c|} \nabla_{\theta} \log p_{\theta}(c_t | c_{<t}) \right\|_2 \quad (96)$$

Note that this expression is independent of $p_{\theta}(c)$.

Theorem 29 (Stability of Energy-Based Gradients). *If the following conditions are satisfied:*

- Let $\{\theta^{(k)}\}_{k=0}^{\infty}$ be parameters generated by gradient ascent on energy
- Assume bounded gradients: $G_{\min} \leq \|\nabla_{\theta} \log p_{\theta}(c_t | c_{<t})\|_2 \leq G_{\max}$
- Assume non-degenerate gradients (not all collinear)

Then the energy gradient remains bounded:

$$\frac{G_{\min}}{\sqrt{|c|}} \leq \|\nabla_{\theta} E(c; \theta^{(k)})\|_2 \leq G_{\max} \quad (97)$$

for all iterations k , regardless of $p_{\theta^{(k)}}(c)$.

Proof. For the upper bound, from Lemma 28:

$$\|\nabla_{\theta} E(c; \theta^{(k)})\|_2 = \frac{1}{|c|} \left\| \sum_{t=1}^{|c|} \nabla_{\theta} \log p_{\theta^{(k)}}(c_t | c_{<t}) \right\|_2 \quad (98)$$

$$\leq \frac{1}{|c|} \sum_{t=1}^{|c|} \|\nabla_{\theta} \log p_{\theta^{(k)}}(c_t | c_{<t})\|_2 \quad (99)$$

$$\leq \frac{1}{|c|} \sum_{t=1}^{|c|} G_{\max} \quad (100)$$

$$= G_{\max} \quad (101)$$

For the lower bound, under non-collinearity, the sum of gradient vectors exhibits partial cancellation but not complete cancellation. In the worst case:

$$\|\nabla_{\theta} E(c; \theta^{(k)})\|_2 \geq \frac{1}{|c|} \cdot G_{\min} \quad (102)$$

Under typical conditions with independent gradient directions:

$$\mathbb{E} \left[\left\| \sum_{t=1}^{|c|} \nabla_{\theta} \log p_{\theta^{(k)}}(c_t | c_{<t}) \right\|_2^2 \right] = \sum_{t=1}^{|c|} \mathbb{E} [\|\nabla_{\theta} \log p_{\theta^{(k)}}(c_t | c_{<t})\|_2^2] \quad (103)$$

$$\geq |c| \cdot G_{\min}^2 \quad (104)$$

Therefore:

$$\|\nabla_{\theta} E(c; \theta^{(k)})\|_2 \geq \frac{1}{|c|} \cdot \sqrt{|c| \cdot G_{\min}^2} = \frac{G_{\min}}{\sqrt{|c|}} \quad (105)$$

Corollary 30 (Convergence Rate Comparison). *Under the conditions of Theorems 27 and 29:*

- **Probability-based methods:** Effective learning rate decays as $O(p_{\theta^{(k)}}(c))$, requiring $O(\log(1/\epsilon)^2)$ iterations to achieve $p_{\theta}(c) < \epsilon$
- **Energy-based methods:** Effective learning rate remains $\Theta(1/\sqrt{|c|})$, requiring only $O(\log(1/\epsilon))$ iterations

Remark 31 (Practical Implications). The gradient dynamics analysis reveals fundamental advantages of energy-based formulations:

1. **Optimization stability:** Energy gradients remain bounded away from zero throughout training
2. **Computational efficiency:** Quadratic speedup compared to probability-based methods
3. **Robustness:** Energy-based optimization is insensitive to the absolute scale of probabilities

These insights directly inform our algorithm design in Appendix B.2.

D.4 BASELINE COMPARISON AND EXPONENTIAL SUPPRESSION ANALYSIS

In this section, we extend our gradient dynamics analysis from Section D.3 to provide a rigorous theoretical comparison between our energy-based framework and existing probability-based approaches for copyright protection. We focus particularly on establishing the exponential suppression guarantees that distinguish our method from baseline approaches.

Comparative Analysis Framework. To systematically evaluate the theoretical advantages of our approach, we establish a general framework for comparing copyright protection methods. The key metrics we consider are: (1) the achievable suppression factor for copyrighted content, (2) the optimization stability throughout training, and (3) the robustness to variations in sequence length and distribution shifts.

Exponential Suppression Guarantees. The fundamental advantage of our energy-based approach lies in the exponential nature of the protection it provides. Building upon the relationships established in Section ??, we can formalize the protection strength as follows:

Theorem 32 (Exponential Protection Strength). *Let θ_{baseline} denote the parameters of a baseline language model trained without copyright protection on the same data distribution, and let θ^* denote the parameters obtained by our energy-based method. For any copyrighted sequence $c \in \mathcal{C}$ with achieved energy gap $\Delta(c; \theta^*) = E(c; \theta^*) - E(c; \theta_{\text{baseline}}) \geq m$, the probability suppression factor satisfies:*

$$\frac{p_{\theta^*}(c)}{p_{\theta_{\text{baseline}}}(c)} = \exp(-|c| \cdot \Delta(c; \theta^*)) \leq \exp(-|c| \cdot m) \quad (106)$$

Furthermore, this suppression factor exhibits superlinear scaling with sequence length, providing exponentially stronger protection for longer copyrighted passages.

Proof. The probability ratio can be expressed directly in terms of the energy difference:

$$\frac{p_{\theta^*}(c)}{p_{\theta_{\text{baseline}}}(c)} = \frac{\exp(-|c| \cdot E(c; \theta^*))}{\exp(-|c| \cdot E(c; \theta_{\text{baseline}}))} \quad (107)$$

$$= \exp(-|c| \cdot (E(c; \theta^*) - E(c; \theta_{\text{baseline}}))) \quad (108)$$

$$= \exp(-|c| \cdot \Delta(c; \theta^*)) \quad (109)$$

Given the constraint $\Delta(c; \theta^*) \geq m$, we immediately obtain:

$$\frac{p_{\theta^*}(c)}{p_{\theta_{\text{baseline}}}(c)} \leq \exp(-|c| \cdot m) \quad (110)$$

The superlinear scaling follows from the fact that the exponent grows linearly with sequence length $|c|$, making the suppression factor decrease exponentially faster for longer sequences.

This theorem establishes the critical advantage of our energy-based approach: protection strength that scales exponentially with sequence length. For typical copyrighted passages (often hundreds of tokens), this provides overwhelming suppression factors that cannot be practically achieved through direct probability manipulation.

Analysis of Representative Baseline Methods. We now analyze several standard baseline approaches to copyright protection in language models and establish their theoretical limitations compared to our energy-based framework.

Proposition 33 (Direct Probability Penalization). *Consider the direct probability penalization approach:*

$$\mathcal{L}_{\text{direct}}(\theta) = \mathcal{L}_{\text{LM}}^{\mathcal{O}}(\theta) + \lambda \sum_{c \in \mathcal{C}} p_{\theta}(c) \quad (111)$$

For this method to achieve probability suppression factor $\rho \geq \exp(\alpha)$ for $\alpha > 0$, the regularization parameter must satisfy:

$$\lambda \geq \frac{\|\nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{O}}(\theta)\|_2}{\alpha \cdot G \cdot L_{\min}} \quad (112)$$

where $L_{\min} = \min_{c \in \mathcal{C}} |c|$ and G is the energy gradient bound. This requirement grows inversely with the minimum sequence length, making protection of shorter sequences disproportionately expensive.

Proof. At equilibrium, the gradient of the objective with respect to θ should vanish:

$$\nabla_{\theta} \mathcal{L}_{\text{direct}}(\theta) = \nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{O}}(\theta) + \lambda \sum_{c \in \mathcal{C}} \nabla_{\theta} p_{\theta}(c) = 0 \quad (113)$$

For a suppression factor $\rho = \frac{p_{\text{baseline}}(c)}{p_{\theta}(c)} = \exp(\alpha)$, we must have $p_{\theta}(c) = p_{\text{baseline}}(c) \cdot \exp(-\alpha)$. The gradient magnitude of the regularization term can be bounded as:

$$\lambda \left\| \sum_{c \in \mathcal{C}} \nabla_{\theta} p_{\theta}(c) \right\|_2 = \lambda \left\| \sum_{c \in \mathcal{C}} -|c| \cdot p_{\theta}(c) \cdot \nabla_{\theta} E(c; \theta) \right\|_2 \quad (114)$$

$$\geq \lambda \cdot L_{\min} \cdot p_{\text{baseline}}(c) \cdot \exp(-\alpha) \cdot G \quad (115)$$

This must balance $\|\nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{O}}(\theta)\|_2$ at equilibrium, yielding the required bound on λ .

Contrastive Learning Approaches. Contrastive methods attempt to increase the relative likelihood of non-copyrighted content over copyrighted content:

Proposition 34 (Limitations of Contrastive Methods). *Consider the contrastive learning objective:*

$$\mathcal{L}_{\text{contrastive}}(\theta) = \mathcal{L}_{\text{LM}}^{\mathcal{O}}(\theta) + \lambda \sum_{c \in \mathcal{C}} \sum_{o \in \mathcal{O}} \max(0, \log p_{\theta}(c) - \log p_{\theta}(o) + \mu) \quad (116)$$

This approach achieves energy gap $\Delta(c; \theta) \geq \mu/|c|$ at equilibrium, providing only inverse linear scaling with sequence length. Additionally, it requires quadratic computational complexity $O(|\mathcal{C}| \times |\mathcal{O}|)$ and suffers from the gradient vanishing phenomenon established in Section D.3.

Proof. At equilibrium, for any active constraint (where $\log p_\theta(c) - \log p_\theta(o) + \mu > 0$), we have:

$$\log p_\theta(c) - \log p_\theta(o) = -\mu \quad (117)$$

Expressing this in terms of energy:

$$-|c| \cdot E(c; \theta) - (-|o| \cdot E(o; \theta)) = -\mu \quad (118)$$

Assuming $|o| \approx |c|$ for simplicity, we get:

$$|c| \cdot (E(c; \theta) - E(o; \theta)) = \mu \quad (119)$$

Therefore, $E(c; \theta) - E(o; \theta) = \mu/|c|$, which scales only inversely linearly with sequence length.

Robustness to Distribution Shifts. A key consideration for copyright protection methods is their robustness to potential distribution shifts between training and deployment environments.

Theorem 35 (Robustness to Distribution Shift). *Let $\mathcal{D}'$ be a shifted data distribution with bounded energy difference $|E_{\mathcal{D}'}(x) - E_{\mathcal{D}}(x)| \leq \epsilon_{\text{shift}}$ for all x . Under our energy-based protection, the suppression factor for copyrighted content satisfies:*

$$\exp(-|c| \cdot (m + \epsilon_{\text{shift}})) \leq \frac{p_{\theta^*}^{\mathcal{D}'}(c)}{p_{\text{baseline}}^{\mathcal{D}'}(c)} \leq \exp(-|c| \cdot (m - \epsilon_{\text{shift}})) \quad (120)$$

demonstrating graceful degradation under distribution shift.

Proof. Under distribution shift, the energy gap becomes:

$$\Delta^{\mathcal{D}'}(c; \theta^*) = E^{\mathcal{D}'}(c; \theta^*) - E^{\mathcal{D}'}(c; \theta_{\text{baseline}}) \quad (121)$$

$$= (E^{\mathcal{D}}(c; \theta^*) + \epsilon_1) - (E^{\mathcal{D}}(c; \theta_{\text{baseline}}) + \epsilon_2) \quad (122)$$

$$= \Delta^{\mathcal{D}}(c; \theta^*) + (\epsilon_1 - \epsilon_2) \quad (123)$$

where $|\epsilon_1|, |\epsilon_2| \leq \epsilon_{\text{shift}}$. This implies:

$$\Delta^{\mathcal{D}}(c; \theta^*) - 2\epsilon_{\text{shift}} \leq \Delta^{\mathcal{D}'}(c; \theta^*) \leq \Delta^{\mathcal{D}}(c; \theta^*) + 2\epsilon_{\text{shift}} \quad (124)$$

Given that $\Delta^{\mathcal{D}}(c; \theta^*) \geq m$, we have $\Delta^{\mathcal{D}'}(c; \theta^*) \geq m - 2\epsilon_{\text{shift}}$. The probability ratio bounds follow directly from the relationship between energy gap and probability ratio established in Theorem 32.

Practical Implications and Theoretical Guarantees. The theoretical analysis presented in this section has significant practical implications for copyright protection in large language models. By establishing an energy gap of m for all copyrighted content, our method guarantees:

1. *Exponential Suppression:* Probability suppression factors that scale as $\exp(-|c| \cdot m)$, providing overwhelming protection for typical passage lengths.
2. *Stable Optimization:* Consistent gradient signals throughout training, avoiding the vanishing gradient phenomenon inherent in probability-based methods.
3. *Length-Proportional Protection:* Automatically stronger protection for longer passages, which aligns with legal notions of substantial similarity in copyright law.
4. *Robustness to Distribution Shifts:* Graceful degradation of protection under distribution shifts, maintaining meaningful suppression even in shifted domains.

Our framework thus provides theoretical guarantees that cannot be matched by existing probability-based approaches, establishing a fundamental advance in copyright protection methodology for language models.

D.5 ANALYSIS OF INVERSE REGULARIZATION INSTABILITY

This section provides a detailed theoretical analysis of the instability issues inherent in inverse regularization methods for copyright protection, as introduced in Section 3.2. We demonstrate why these methods exhibit fundamental optimization problems and provide insufficient protection guarantees.

Formalization of Inverse Regularization. Recall the inverse regularization objective from Equation 11:

$$\mathcal{L}_{\text{inv}}(\theta) = \mathcal{L}_{\text{LM}}^{\mathcal{O}}(\theta) + \gamma_{\text{inv}} \cdot [\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) + \epsilon_0]^{-1} \quad (125)$$

The underlying intuition of this approach is to maximize the loss on copyrighted content while minimizing it on ordinary content. However, this formulation introduces critical instabilities in the optimization process.

Gradient Dynamics Analysis. To understand the optimization difficulties, we analyze the gradient of the objective with respect to model parameters:

Proposition 36 (Gradient Explosion in Inverse Regularization). *The gradient of the inverse regularization term with respect to model parameters θ satisfies:*

$$\nabla_{\theta} ([\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) + \epsilon_0]^{-1}) = -[\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) + \epsilon_0]^{-2} \cdot \nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) \quad (126)$$

As the model learns to increase the loss on copyrighted data, the term $[\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) + \epsilon_0]^{-2}$ increases quadratically, leading to gradient explosion when $\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta)$ approaches $-\epsilon_0$ from above.

Proof. By direct application of the chain rule to the inverse function:

$$\nabla_{\theta} ([\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) + \epsilon_0]^{-1}) = \nabla_f(f^{-1}) \cdot \nabla_{\theta} (\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) + \epsilon_0) \quad (127)$$

$$= -[f(\theta)]^{-2} \cdot \nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) \quad (128)$$

where $f(\theta) = \mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) + \epsilon_0$.

Since $\nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta)$ remains bounded by assumption, the gradient norm is primarily determined by $[f(\theta)]^{-2}$. As optimization progresses and $\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta)$ increases (as desired for copyright protection), this term grows quadratically, leading to unbounded gradient magnitudes when $f(\theta) \rightarrow 0^+$.

This result formally establishes the inherent instability in optimization dynamics: successful copyright protection (increasing $\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta)$) leads to increasingly unstable gradients, making continued optimization impossible without careful step size adjustments that themselves undermine convergence.

Optimization Trajectory Analysis. The practical implications of this gradient explosion are severe. We can characterize the optimization trajectory as follows:

Theorem 37 (Optimization Trajectory Instability). *Consider gradient descent optimization of the inverse regularization objective with step size $\eta > 0$. The optimization trajectory exhibits one of three behaviors:*

1. *Insufficient Protection: If γ_{inv} is too small, the inverse term becomes negligible, and the model converges to memorizing copyrighted content.*

2. *Gradient Explosion: If γ_{inv} is moderately large, the model initially increases $\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta)$, but eventually enters a region where gradients explode, causing training instability.*

3. *Optimization Failure: If γ_{inv} is too large, the inverse term dominates immediately, preventing learning on both copyrighted and ordinary content.*

In particular, there exists no stable fixed point where the model achieves both good performance on ordinary content and strong copyright protection.

Proof. Let θ_t denote the parameters at iteration t . The parameter update is:

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_{\theta} \mathcal{L}_{\text{inv}}(\theta_t) \quad (129)$$

For case 1 (small γ_{inv}): The gradient is dominated by $\nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{O}}(\theta)$, leading to minimization of loss on all data including copyrighted content.

For case 2 (moderate γ_{inv}): Initially, the model increases $\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta)$. However, as $\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) + \epsilon_0$ approaches zero, the gradient norm grows as $O([\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) + \epsilon_0]^{-2})$, eventually exceeding any bound.

For case 3 (large γ_{inv}): The inverse term dominates even at initialization, preventing meaningful optimization on ordinary content.

To show that no stable fixed point exists, note that at any fixed point, we must have:

$$\nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{O}}(\theta) = \gamma_{\text{inv}} \cdot [\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) + \epsilon_0]^{-2} \cdot \nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) \quad (130)$$

For this to be satisfied while maintaining good performance on both datasets would require $\nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{O}}(\theta)$ and $\nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta)$ to be perfectly aligned in direction, which is generally not the case for distinct datasets.

Hyperparameter Sensitivity Analysis. The extreme sensitivity to hyperparameter settings makes inverse regularization particularly challenging in practice:

Proposition 38 (Hyperparameter Sensitivity). *The range of γ_{inv} values that avoid both insufficient protection and gradient explosion is inversely proportional to the initial difference between $\mathcal{L}_{\text{LM}}^{\mathcal{O}}(\theta_0)$ and $\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta_0)$. Specifically:*

$$\frac{\gamma_{\text{max}}}{\gamma_{\text{min}}} \leq \frac{\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta_0) + \epsilon_0}{\epsilon_0} \cdot \frac{\|\nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{O}}(\theta_0)\|_2}{\|\nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta_0)\|_2} \quad (131)$$

where γ_{min} and γ_{max} bound the viable range of regularization strengths.

This sensitivity increases with dataset size and model complexity, making it impractical for large-scale applications where extensive hyperparameter tuning is prohibitively expensive.

Distribution Shift Vulnerability. Perhaps most critically, inverse regularization provides no worst-case guarantees under distribution shifts:

Theorem 39 (Distribution Shift Vulnerability). *Consider a distribution shift that increases the loss on copyrighted content by a factor $\alpha > 1$. Under inverse regularization, the effective regularization strength is reduced by a factor of α^2 , potentially nullifying protection. Conversely, a shift that decreases the loss by a factor $\beta < 1$ increases the effective regularization strength by a factor of β^{-2} , potentially causing catastrophic forgetting.*

Proof. Under distribution shift, the copyrighted content loss becomes $\alpha \cdot \mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta)$ for some $\alpha > 0$. The inverse term becomes:

$$[\alpha \cdot \mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) + \epsilon_0]^{-1} \approx \frac{1}{\alpha} \cdot [\mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) + \epsilon_0/\alpha]^{-1} \quad (132)$$

when $\alpha \cdot \mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) \gg \epsilon_0$.

The gradient of this term scales as $[\alpha \cdot \mathcal{L}_{\text{LM}}^{\mathcal{C}}(\theta) + \epsilon_0]^{-2}$, which is reduced by a factor of α^2 compared to the original gradient when $\alpha > 1$, or increased by a factor of β^{-2} when the shift decreases the loss by a factor $\beta < 1$.

This extreme sensitivity to distribution shifts means that protection can be effectively nullified by even minor changes in the data distribution, making the approach unreliable for practical applications.

Comparison with Energy-Based Approach. In contrast to the instabilities of inverse regularization, our energy-based approach avoids these pitfalls by directly targeting the energy gap. We can draw a direct comparison:

Proposition 40 (Stability Comparison). *Under identical distribution shifts that change the loss on copyrighted content by a factor α , our energy-based method preserves the protection strength within a factor of $\log(\alpha)$, while inverse regularization changes by a factor of α^2 .*

Proof. For our energy-based method, the energy gap changes additively by $\log(\alpha)/|c|$. For inverse regularization, the effective regularization strength changes by a multiplicative factor of α^2 .

This fundamental difference in stability explains why our approach provides robust protection guarantees while inverse regularization fails to provide consistent protection.

Conclusion. Inverse regularization approaches, despite their intuitive appeal, suffer from fundamental instabilities in optimization, extreme sensitivity to hyperparameters, and vulnerability to distribution shifts. These limitations make them unsuitable for reliable copyright protection in large language models. Our energy-based framework addresses all these limitations by providing a stable optimization objective with predictable behavior and robust protection guarantees.

E ADAPTIVE REGULARIZER ANALYSIS

E.1 PROPERTIES OF ADAPTIVE REGULARIZER

In this appendix, we provide a rigorous analysis of the adaptive energy regularizer introduced in Definition 3. We establish boundedness, differentiability, asymptotic behavior, and Lipschitz continuity properties that ensure stable and effective optimization for copyright protection.

Definition 41 (Adaptive Energy Regularizer (Restated)). Given energy gap $\Delta(\theta) = \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta)] - \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta)]$, the adaptive regularizer is:

$$\mathcal{R}(\theta; m, \tau) = \tau \log \left(1 + \exp \left(-\frac{\Delta(\theta) - m}{\tau} \right) \right) \quad (133)$$

where $m \geq 0$ is the target margin and $\tau > 0$ is the temperature parameter.

Lemma 42 (Boundedness of Regularizer). *For any $\theta \in \mathbb{R}^d$, $m \geq 0$, and $\tau > 0$:*

$$0 \leq \mathcal{R}(\theta; m, \tau) \leq \tau \log 2 \quad (134)$$

Proof. Let $z = -\frac{\Delta(\theta) - m}{\tau}$. Then:

$$\mathcal{R}(\theta; m, \tau) = \tau \log(1 + \exp(z)) \quad (135)$$

Lower bound: Since $\exp(z) > 0$ for all $z \in \mathbb{R}$:

$$1 + \exp(z) > 1 \quad (136)$$

Therefore:

$$\mathcal{R}(\theta; m, \tau) = \tau \log(1 + \exp(z)) \quad (137)$$

$$> \tau \log(1) \quad (138)$$

$$= 0 \quad (139)$$

Upper bound: We analyze the function $f(z) = \log(1 + \exp(z))$ for all $z \in \mathbb{R}$.

For $z \geq 0$:

$$f(z) = \log(1 + \exp(z)) \quad (140)$$

$$= \log(\exp(z)(1/\exp(z) + 1)) \quad (141)$$

$$= \log(\exp(z)) + \log(1/\exp(z) + 1) \quad (142)$$

$$= z + \log(1 + \exp(-z)) \quad (143)$$

Since $\exp(-z) \in (0, 1]$ for $z \geq 0$:

$$\log(1 + \exp(-z)) \leq \log(1 + 1) = \log 2 \quad (144)$$

For $z < 0$:

$$f(z) = \log(1 + \exp(z)) \quad (145)$$

$$< \log(1 + 1) \quad (146)$$

$$= \log 2 \quad (147)$$

Combining both cases:

$$\mathcal{R}(\theta; m, \tau) = \tau \cdot f(z) \quad (148)$$

$$\leq \tau \log 2 \quad (149)$$

for all $z \in \mathbb{R}$.

Lemma 43 (Asymptotic Behavior). *The regularizer exhibits the following asymptotic behavior:*

1. **Under-margin regime:** For $\Delta(\theta) \ll m$:

$$\mathcal{R}(\theta; m, \tau) = m - \Delta(\theta) + O(\exp(-|m - \Delta(\theta)|/\tau)) \quad (150)$$

2. **Over-margin regime:** For $\Delta(\theta) \gg m$:

$$\mathcal{R}(\theta; m, \tau) = \tau \exp(-((\Delta(\theta) - m)/\tau)) + O(\exp(-2(\Delta(\theta) - m)/\tau)) \quad (151)$$

3. **Near-margin regime:** For $|\Delta(\theta) - m| \leq \tau$:

$$\mathcal{R}(\theta; m, \tau) = \tau \log 2 - \frac{\Delta(\theta) - m}{2} + O(((\Delta(\theta) - m)/\tau)^2) \quad (152)$$

Proof. Let $w = \frac{\Delta(\theta) - m}{\tau}$. Then:

$$\mathcal{R}(\theta; m, \tau) = \tau \log(1 + \exp(-w)) \quad (153)$$

Part 1 (Under-margin regime): For $w \ll -1$ (i.e., $\Delta(\theta) \ll m$):

$$\mathcal{R}(\theta; m, \tau) = \tau \log(1 + \exp(-w)) \quad (154)$$

$$= \tau \log(\exp(-w)(1 + \exp(w))) \quad (155)$$

$$= \tau(-w + \log(1 + \exp(w))) \quad (156)$$

$$= -\tau w + \tau \log(1 + \exp(w)) \quad (157)$$

Since $w \ll -1$, we have $\exp(w) \approx 0$:

$$\mathcal{R}(\theta; m, \tau) = -\tau w + \tau \log(1 + \exp(w)) \quad (158)$$

$$= -\tau \cdot \frac{\Delta(\theta) - m}{\tau} + \tau \exp(w) + O(\exp(2w)) \quad (159)$$

$$= m - \Delta(\theta) + \tau \exp\left(\frac{\Delta(\theta) - m}{\tau}\right) + O(\exp(2(\Delta(\theta) - m)/\tau)) \quad (160)$$

Part 2 (Over-margin regime): For $w \gg 1$ (i.e., $\Delta(\theta) \gg m$):

$$\mathcal{R}(\theta; m, \tau) = \tau \log(1 + \exp(-w)) \quad (161)$$

$$= \tau \log(1 + \exp(-w)) \quad (162)$$

Using Taylor expansion $\log(1 + x) = x - x^2/2 + O(x^3)$ for small x :

$$\mathcal{R}(\theta; m, \tau) = \tau \left(\exp(-w) - \frac{\exp(-2w)}{2} + O(\exp(-3w)) \right) \quad (163)$$

$$= \tau \exp\left(-\frac{\Delta(\theta) - m}{\tau}\right) + O(\exp(-2(\Delta(\theta) - m)/\tau)) \quad (164)$$

Part 3 (Near-margin regime): For $|w| \leq 1$, using Taylor expansion around $w = 0$:

$$\log(1 + \exp(-w)) = \log(1 + \exp(0)) - \frac{d}{dw} \log(1 + \exp(-w)) \Big|_{w=0} \cdot w + O(w^2) \quad (165)$$

$$= \log 2 - \frac{-\exp(-w)}{1 + \exp(-w)} \Big|_{w=0} \cdot w + O(w^2) \quad (166)$$

$$= \log 2 + \frac{1}{2}w + O(w^2) \quad (167)$$

Therefore:

$$\mathcal{R}(\theta; m, \tau) = \tau \left(\log 2 + \frac{w}{2} + O(w^2) \right) \quad (168)$$

$$= \tau \log 2 + \frac{\tau w}{2} + O(\tau w^2) \quad (169)$$

$$= \tau \log 2 + \frac{\Delta(\theta) - m}{2} + O(((\Delta(\theta) - m)/\tau)^2) \quad (170)$$

Theorem 44 (Gradient of Adaptive Regularizer). *The gradient of the adaptive regularizer with respect to θ is:*

$$\nabla_{\theta} \mathcal{R}(\theta; m, \tau) = -\sigma \left(\frac{m - \Delta(\theta)}{\tau} \right) \cdot \nabla_{\theta} \Delta(\theta) \quad (171)$$

where $\sigma(z) = \frac{1}{1 + \exp(-z)}$ is the sigmoid function.

Proof. Starting from the definition:

$$\mathcal{R}(\theta; m, \tau) = \tau \log \left(1 + \exp \left(-\frac{\Delta(\theta) - m}{\tau} \right) \right) \quad (172)$$

Taking the gradient with respect to θ :

$$\nabla_{\theta} \mathcal{R}(\theta; m, \tau) = \tau \cdot \nabla_{\theta} \log \left(1 + \exp \left(-\frac{\Delta(\theta) - m}{\tau} \right) \right) \quad (173)$$

$$= \tau \cdot \frac{1}{1 + \exp \left(-\frac{\Delta(\theta) - m}{\tau} \right)} \cdot \nabla_{\theta} \left(\exp \left(-\frac{\Delta(\theta) - m}{\tau} \right) \right) \quad (174)$$

$$= \tau \cdot \frac{1}{1 + \exp \left(-\frac{\Delta(\theta) - m}{\tau} \right)} \cdot \exp \left(-\frac{\Delta(\theta) - m}{\tau} \right) \cdot \nabla_{\theta} \left(-\frac{\Delta(\theta) - m}{\tau} \right) \quad (175)$$

$$= \tau \cdot \frac{\exp \left(-\frac{\Delta(\theta) - m}{\tau} \right)}{1 + \exp \left(-\frac{\Delta(\theta) - m}{\tau} \right)} \cdot \left(-\frac{1}{\tau} \right) \cdot \nabla_{\theta} \Delta(\theta) \quad (176)$$

$$= -\frac{\exp \left(-\frac{\Delta(\theta) - m}{\tau} \right)}{1 + \exp \left(-\frac{\Delta(\theta) - m}{\tau} \right)} \cdot \nabla_{\theta} \Delta(\theta) \quad (177)$$

Now observe that:

$$\frac{\exp \left(-\frac{\Delta(\theta) - m}{\tau} \right)}{1 + \exp \left(-\frac{\Delta(\theta) - m}{\tau} \right)} = \frac{1}{1 + \exp \left(\frac{\Delta(\theta) - m}{\tau} \right)} \quad (178)$$

$$= \frac{1}{1 + \exp \left(-\frac{m - \Delta(\theta)}{\tau} \right)} \quad (179)$$

$$= \sigma \left(\frac{m - \Delta(\theta)}{\tau} \right) \quad (180)$$

Therefore:

$$\nabla_{\theta} \mathcal{R}(\theta; m, \tau) = -\sigma \left(\frac{m - \Delta(\theta)}{\tau} \right) \cdot \nabla_{\theta} \Delta(\theta) \quad (181)$$

Lemma 45 (Properties of Gradient Weight Function). *The gradient weight function $w(\Delta) = \sigma \left(\frac{m - \Delta}{\tau} \right)$ satisfies:*

1. **Boundedness:** $0 < w(\Delta) < 1$ for all $\Delta \in \mathbb{R}$

2. **Monotonicity:** $w'(\Delta) = -\frac{1}{\tau} \sigma\left(\frac{m-\Delta}{\tau}\right) \left(1 - \sigma\left(\frac{m-\Delta}{\tau}\right)\right) < 0$

3. **Asymptotic behavior:**

$$w(\Delta) = \begin{cases} 1 - O(\exp(-(m-\Delta)/\tau)) & \text{if } \Delta \ll m \\ 1/2 & \text{if } \Delta = m \\ O(\exp(-((\Delta-m)/\tau))) & \text{if } \Delta \gg m \end{cases} \quad (182)$$

Proof. Part 1 (Boundedness): Since $\sigma(z) = \frac{1}{1+\exp(-z)}$:

$$0 < \frac{1}{1+\exp(-z)} < 1 \quad (183)$$

for all $z \in \mathbb{R}$, as $\exp(-z) > 0$ for all z .

Part 2 (Monotonicity): The derivative of the sigmoid function is:

$$\frac{d}{dz} \sigma(z) = \frac{d}{dz} \left(\frac{1}{1+\exp(-z)} \right) \quad (184)$$

$$= -\frac{1}{(1+\exp(-z))^2} \cdot (-\exp(-z)) \quad (185)$$

$$= \frac{\exp(-z)}{(1+\exp(-z))^2} \quad (186)$$

$$= \sigma(z)(1-\sigma(z)) \quad (187)$$

Therefore:

$$w'(\Delta) = \frac{d}{d\Delta} \sigma\left(\frac{m-\Delta}{\tau}\right) \quad (188)$$

$$= \sigma\left(\frac{m-\Delta}{\tau}\right) \left(1 - \sigma\left(\frac{m-\Delta}{\tau}\right)\right) \cdot \left(-\frac{1}{\tau}\right) \quad (189)$$

$$= -\frac{1}{\tau} \sigma\left(\frac{m-\Delta}{\tau}\right) \left(1 - \sigma\left(\frac{m-\Delta}{\tau}\right)\right) < 0 \quad (190)$$

Part 3 (Asymptotic behavior): For $\Delta \ll m$, let $z = \frac{m-\Delta}{\tau} \gg 1$:

$$w(\Delta) = \sigma(z) = \frac{1}{1+\exp(-z)} \quad (191)$$

$$= \frac{\exp(z)}{\exp(z)+1} \quad (192)$$

$$= \frac{1}{1+\exp(-z)} \quad (193)$$

$$= 1 - \frac{\exp(-z)}{1+\exp(-z)} \quad (194)$$

$$= 1 - O(\exp(-z)) \quad (195)$$

$$= 1 - O(\exp(-(m-\Delta)/\tau)) \quad (196)$$

For $\Delta = m$:

$$w(m) = \sigma(0) = \frac{1}{1+\exp(0)} = \frac{1}{2} \quad (197)$$

For $\Delta \gg m$, let $z = \frac{m-\Delta}{\tau} \ll -1$:

$$w(\Delta) = \sigma(z) = \frac{1}{1 + \exp(-z)} \quad (198)$$

$$= \frac{1}{1 + \exp(|z|)} \quad (199)$$

$$= \frac{1}{\exp(|z|)(1/\exp(|z|) + 1)} \quad (200)$$

$$= \frac{\exp(-|z|)}{1 + \exp(-|z|)} \quad (201)$$

$$= O(\exp(-|z|)) \quad (202)$$

$$= O(\exp(-(\Delta - m)/\tau)) \quad (203)$$

Theorem 46 (Lipschitz Continuity of Gradient). *If the following conditions are satisfied:*

- Assume $\|\nabla_{\theta} \Delta(\theta)\|_2 \leq G$ for all $\theta \in \mathbb{R}^d$
- Assume $\Delta(\theta)$ is L_{Δ} -Lipschitz continuous

Then $\nabla_{\theta} \mathcal{R}(\theta; m, \tau)$ is Lipschitz continuous with constant:

$$L_R = \frac{G^2}{4\tau} + \sigma_{\max} \cdot L_{\nabla \Delta} \quad (204)$$

where $\sigma_{\max} = \sup_{z \in \mathbb{R}} \sigma(z) = 1$ and $L_{\nabla \Delta}$ is the Lipschitz constant of $\nabla_{\theta} \Delta(\theta)$.

Proof. For any $\theta_1, \theta_2 \in \mathbb{R}^d$:

$$\|\nabla_{\theta} \mathcal{R}(\theta_1) - \nabla_{\theta} \mathcal{R}(\theta_2)\|_2 \quad (205)$$

$$= \left\| -\sigma\left(\frac{m - \Delta(\theta_1)}{\tau}\right) \nabla_{\theta} \Delta(\theta_1) + \sigma\left(\frac{m - \Delta(\theta_2)}{\tau}\right) \nabla_{\theta} \Delta(\theta_2) \right\|_2 \quad (206)$$

$$= \left\| -\sigma\left(\frac{m - \Delta(\theta_1)}{\tau}\right) \nabla_{\theta} \Delta(\theta_1) + \sigma\left(\frac{m - \Delta(\theta_1)}{\tau}\right) \nabla_{\theta} \Delta(\theta_2) \right. \quad (207)$$

$$\left. - \sigma\left(\frac{m - \Delta(\theta_1)}{\tau}\right) \nabla_{\theta} \Delta(\theta_2) + \sigma\left(\frac{m - \Delta(\theta_2)}{\tau}\right) \nabla_{\theta} \Delta(\theta_2) \right\|_2 \quad (208)$$

$$\leq \left\| \sigma\left(\frac{m - \Delta(\theta_1)}{\tau}\right) (\nabla_{\theta} \Delta(\theta_2) - \nabla_{\theta} \Delta(\theta_1)) \right\|_2 \quad (209)$$

$$+ \left\| \left(\sigma\left(\frac{m - \Delta(\theta_2)}{\tau}\right) - \sigma\left(\frac{m - \Delta(\theta_1)}{\tau}\right) \right) \nabla_{\theta} \Delta(\theta_2) \right\|_2 \quad (210)$$

For the first term:

$$\left\| \sigma\left(\frac{m - \Delta(\theta_1)}{\tau}\right) (\nabla_{\theta} \Delta(\theta_2) - \nabla_{\theta} \Delta(\theta_1)) \right\|_2 \leq \sigma_{\max} \cdot \|\nabla_{\theta} \Delta(\theta_2) - \nabla_{\theta} \Delta(\theta_1)\|_2 \quad (211)$$

$$\leq L_{\nabla \Delta} \|\theta_2 - \theta_1\|_2 \quad (212)$$

For the second term, using the mean value theorem:

$$\left| \sigma\left(\frac{m - \Delta(\theta_2)}{\tau}\right) - \sigma\left(\frac{m - \Delta(\theta_1)}{\tau}\right) \right| = |\sigma'(\xi)| \cdot \left| \frac{m - \Delta(\theta_2)}{\tau} - \frac{m - \Delta(\theta_1)}{\tau} \right| \quad (213)$$

$$= |\sigma(\xi)(1 - \sigma(\xi))| \cdot \frac{|\Delta(\theta_1) - \Delta(\theta_2)|}{\tau} \quad (214)$$

for some ξ between $\frac{m - \Delta(\theta_1)}{\tau}$ and $\frac{m - \Delta(\theta_2)}{\tau}$.

Since $\sigma(\xi)(1 - \sigma(\xi)) \leq \frac{1}{4}$ for all ξ (maximum at $\xi = 0$):

$$\left| \sigma\left(\frac{m - \Delta(\theta_2)}{\tau}\right) - \sigma\left(\frac{m - \Delta(\theta_1)}{\tau}\right) \right| \leq \frac{1}{4\tau} |\Delta(\theta_1) - \Delta(\theta_2)| \quad (215)$$

$$\leq \frac{L_\Delta}{4\tau} \|\theta_1 - \theta_2\|_2 \quad (216)$$

Therefore:

$$\left\| \left(\sigma\left(\frac{m - \Delta(\theta_2)}{\tau}\right) - \sigma\left(\frac{m - \Delta(\theta_1)}{\tau}\right) \right) \nabla_\theta \Delta(\theta_2) \right\|_2 \leq \frac{L_\Delta}{4\tau} \|\theta_1 - \theta_2\|_2 \cdot G \quad (217)$$

$$= \frac{GL_\Delta}{4\tau} \|\theta_1 - \theta_2\|_2 \quad (218)$$

Since $L_\Delta \leq G$ (Lipschitz constant of Δ bounded by gradient norm):

$$\|\nabla_\theta \mathcal{R}(\theta_1) - \nabla_\theta \mathcal{R}(\theta_2)\|_2 \leq L_{\nabla \Delta} \|\theta_1 - \theta_2\|_2 + \frac{G^2}{4\tau} \|\theta_1 - \theta_2\|_2 \quad (219)$$

$$= \left(\frac{G^2}{4\tau} + L_{\nabla \Delta} \right) \|\theta_1 - \theta_2\|_2 \quad (220)$$

Remark 47 (Design Principles). The adaptive regularizer design achieves several critical objectives:

1. **Automatic adjustment:** The sigmoid weighting function naturally transitions from strong enforcement ($w \approx 1$) when below the margin to negligible effect ($w \approx 0$) when above.
2. **Smooth optimization:** The temperature parameter τ controls the transition smoothness. Larger τ provides smoother gradients but slower convergence; smaller τ yields sharper transitions but may cause optimization instability.
3. **Bounded gradients:** The gradient norm $\|\nabla_\theta \mathcal{R}\|_2 \leq G$ remains bounded regardless of the energy gap magnitude, preventing gradient explosion.
4. **Convergence guarantee:** The Lipschitz continuity with constant $L_R = O(G^2/\tau)$ ensures convergence of gradient-based optimization under standard step size conditions $\eta < 2/L_R$.

These properties make the adaptive regularizer particularly suitable for fine-tuning large language models where stability and preservation of pre-trained capabilities are paramount.

E.2 PROTECTION MONOTONICITY ANALYSIS

In this appendix, we provide the proof of Theorem 4 and analyze the monotonicity properties of the energy gap during optimization. We establish convergence guarantees and characterize how the protection level evolves during training.

Proof of Theorem 4. At a local minimum θ^* , the gradient of the total objective must vanish:

$$\nabla_\theta \mathcal{L}(\theta^*) = 0 \quad (221)$$

Expanding the total objective from Eq. equation 21:

$$\nabla_\theta \mathcal{L}(\theta^*) = \nabla_\theta \mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta^*) + \gamma \cdot \nabla_\theta \mathcal{R}(\theta^*; m, \tau) \quad (222)$$

$$= \nabla_\theta \mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta^*) - \gamma \cdot \sigma\left(\frac{m - \Delta(\theta^*)}{\tau}\right) \cdot \nabla_\theta \Delta(\theta^*) \quad (223)$$

where the second equality follows from Theorem 44.

Setting this equal to zero yields the first-order optimality condition:

$$\nabla_\theta \mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta^*) = \gamma \cdot \sigma\left(\frac{m - \Delta(\theta^*)}{\tau}\right) \cdot \nabla_\theta \Delta(\theta^*) \quad (224)$$

Taking norms on both sides:

$$\|\nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta^*)\|_2 = \gamma \cdot \sigma \left(\frac{m - \Delta(\theta^*)}{\tau} \right) \cdot \|\nabla_{\theta} \Delta(\theta^*)\|_2 \quad (225)$$

$$\leq B_{\text{LM}} \quad (226)$$

Since $\|\nabla_{\theta} \Delta(\theta^*)\|_2 \geq g_{\min} > 0$ by the non-degeneracy assumption:

$$\sigma \left(\frac{m - \Delta(\theta^*)}{\tau} \right) \leq \frac{B_{\text{LM}}}{\gamma \cdot g_{\min}} \quad (227)$$

Let $z^* = \frac{m - \Delta(\theta^*)}{\tau}$. Then we need to solve:

$$\frac{1}{1 + \exp(-z^*)} \leq \frac{B_{\text{LM}}}{\gamma \cdot g_{\min}} \quad (228)$$

Rearranging:

$$1 + \exp(-z^*) \geq \frac{\gamma \cdot g_{\min}}{B_{\text{LM}}} \quad (229)$$

This gives:

$$\exp(-z^*) \geq \frac{\gamma \cdot g_{\min}}{B_{\text{LM}}} - 1 \quad (230)$$

When $\gamma \cdot g_{\min} > B_{\text{LM}}$, we have $\exp(-z^*) \geq \frac{\gamma \cdot g_{\min} - B_{\text{LM}}}{B_{\text{LM}}} > 0$, which implies:

$$-z^* \geq \log \left(\frac{\gamma \cdot g_{\min} - B_{\text{LM}}}{B_{\text{LM}}} \right) \quad (231)$$

Therefore:

$$z^* \leq -\log \left(\frac{\gamma \cdot g_{\min} - B_{\text{LM}}}{B_{\text{LM}}} \right) = \log \left(\frac{B_{\text{LM}}}{\gamma \cdot g_{\min} - B_{\text{LM}}} \right) \quad (232)$$

When $\gamma \cdot g_{\min} \leq B_{\text{LM}}$, the constraint is always satisfied for any $z^* \in \mathbb{R}$.

For the symmetric case where the gradient points in the opposite direction, we obtain:

$$z^* \geq -\log \left(1 + \frac{B_{\text{LM}}}{\gamma \cdot g_{\min}} \right) \quad (233)$$

Combining both bounds:

$$|z^*| = \left| \frac{m - \Delta(\theta^*)}{\tau} \right| \leq \log \left(1 + \frac{B_{\text{LM}}}{\gamma \cdot g_{\min}} \right) \quad (234)$$

Multiplying by τ :

$$|\Delta(\theta^*) - m| \leq \tau \log \left(1 + \frac{B_{\text{LM}}}{\gamma \cdot g_{\min}} \right) \quad (235)$$

For the special case where $\gamma \geq B_{\text{LM}}/(g_{\min} \cdot \epsilon)$ with $\epsilon > 0$:

$$|\Delta(\theta^*) - m| \leq \tau \log \left(1 + \frac{B_{\text{LM}}}{\gamma \cdot g_{\min}} \right) \quad (236)$$

$$\leq \tau \log(1 + \epsilon) \quad (237)$$

Using the Taylor expansion $\log(1 + \epsilon) = \epsilon - \epsilon^2/2 + O(\epsilon^3)$ for small ϵ :

$$|\Delta(\theta^*) - m| \leq \tau \cdot \epsilon + O(\tau \epsilon^2) \quad (238)$$

Lemma 48 (Monotonicity of Energy Gap). *Consider gradient flow dynamics $\frac{d\theta}{dt} = -\nabla_{\theta}\mathcal{L}(\theta)$ with the combined objective. If the following conditions hold:*

- *The energy gap gradient satisfies $\langle \nabla_{\theta}\Delta(\theta), \nabla_{\theta}\mathcal{L}_{LM}^{\mathcal{D}}(\theta) \rangle \leq \rho \|\nabla_{\theta}\Delta(\theta)\|_2^2$ for some $\rho \geq 0$*
- *The current gap satisfies $\Delta(\theta) < m - \tau \log(1 + \rho/\gamma)$*

Then $\frac{d\Delta(\theta)}{dt} > 0$, meaning the energy gap increases monotonically toward the target margin.

Proof. The rate of change of the energy gap along the gradient flow is:

$$\frac{d\Delta(\theta)}{dt} = \left\langle \nabla_{\theta}\Delta(\theta), \frac{d\theta}{dt} \right\rangle \quad (239)$$

$$= -\langle \nabla_{\theta}\Delta(\theta), \nabla_{\theta}\mathcal{L}(\theta) \rangle \quad (240)$$

$$= -\langle \nabla_{\theta}\Delta(\theta), \nabla_{\theta}\mathcal{L}_{LM}^{\mathcal{D}}(\theta) + \gamma \nabla_{\theta}\mathcal{R}(\theta; m, \tau) \rangle \quad (241)$$

From Theorem 44:

$$\frac{d\Delta(\theta)}{dt} = -\langle \nabla_{\theta}\Delta(\theta), \nabla_{\theta}\mathcal{L}_{LM}^{\mathcal{D}}(\theta) \rangle + \gamma \sigma\left(\frac{m - \Delta(\theta)}{\tau}\right) \|\nabla_{\theta}\Delta(\theta)\|_2^2 \quad (242)$$

Using the assumption on the alignment between gradients:

$$\frac{d\Delta(\theta)}{dt} \geq -\rho \|\nabla_{\theta}\Delta(\theta)\|_2^2 + \gamma \sigma\left(\frac{m - \Delta(\theta)}{\tau}\right) \|\nabla_{\theta}\Delta(\theta)\|_2^2 \quad (243)$$

$$= \|\nabla_{\theta}\Delta(\theta)\|_2^2 \left(\gamma \sigma\left(\frac{m - \Delta(\theta)}{\tau}\right) - \rho \right) \quad (244)$$

For $\Delta(\theta) < m - \tau \log(1 + \rho/\gamma)$, we have:

$$\frac{m - \Delta(\theta)}{\tau} > \log(1 + \rho/\gamma) \quad (245)$$

Therefore:

$$\sigma\left(\frac{m - \Delta(\theta)}{\tau}\right) = \frac{1}{1 + \exp\left(-\frac{m - \Delta(\theta)}{\tau}\right)} \quad (246)$$

$$> \frac{1}{1 + \exp(-\log(1 + \rho/\gamma))} \quad (247)$$

$$= \frac{1}{1 + \frac{1}{1 + \rho/\gamma}} \quad (248)$$

$$= \frac{1 + \rho/\gamma}{2 + \rho/\gamma} \quad (249)$$

$$> \frac{\rho}{\gamma} \quad (250)$$

This gives:

$$\frac{d\Delta(\theta)}{dt} > \|\nabla_{\theta}\Delta(\theta)\|_2^2 \left(\gamma \cdot \frac{\rho}{\gamma} - \rho \right) = 0 \quad (251)$$

Since we actually have strict inequality in the sigmoid bound, we obtain $\frac{d\Delta(\theta)}{dt} > 0$.

Theorem 49 (Convergence Rate of Energy Gap). *Under gradient flow dynamics with learning rate $\eta > 0$, if the conditions of Lemma 48 hold and additionally:*

- *The energy gap gradient has bounded norm: $g_{\min} \leq \|\nabla_{\theta}\Delta(\theta)\|_2 \leq g_{\max}$*

- The regularizer coefficient satisfies $\gamma > \rho$

Then the energy gap converges to a neighborhood of the target margin at an exponential rate:

$$|\Delta(\theta(t)) - m| \leq |\Delta(\theta(0)) - m| \cdot \exp\left(-\eta\gamma g_{\min}^2 \cdot \frac{t}{4\tau}\right) + O(\tau) \quad (252)$$

Proof. Define the Lyapunov function:

$$V(\theta) = \frac{1}{2}(\Delta(\theta) - m)^2 \quad (253)$$

Its time derivative along the gradient flow is:

$$\frac{dV}{dt} = (\Delta(\theta) - m) \cdot \frac{d\Delta(\theta)}{dt} \quad (254)$$

$$= (\Delta(\theta) - m) \cdot \langle \nabla_{\theta} \Delta(\theta), -\eta \nabla_{\theta} \mathcal{L}(\theta) \rangle \quad (255)$$

Substituting the gradient expression:

$$\frac{dV}{dt} = \eta(\Delta(\theta) - m) \cdot \left[-\langle \nabla_{\theta} \Delta(\theta), \nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta) \rangle \right. \quad (256)$$

$$\left. + \gamma \sigma\left(\frac{m - \Delta(\theta)}{\tau}\right) \|\nabla_{\theta} \Delta(\theta)\|_2^2 \right] \quad (257)$$

Near the target margin where $|\Delta(\theta) - m| \leq \delta$ for small $\delta > 0$, we can approximate:

$$\sigma\left(\frac{m - \Delta(\theta)}{\tau}\right) = \frac{1}{2} + \frac{m - \Delta(\theta)}{4\tau} + O\left(\left(\frac{m - \Delta(\theta)}{\tau}\right)^2\right) \quad (258)$$

$$= \frac{1}{2} - \frac{\Delta(\theta) - m}{4\tau} + O\left(\frac{(\Delta(\theta) - m)^2}{\tau^2}\right) \quad (259)$$

Using the gradient alignment assumption:

$$\frac{dV}{dt} \leq \eta(\Delta(\theta) - m) \cdot \left[\rho \|\nabla_{\theta} \Delta(\theta)\|_2^2 + \gamma \left(\frac{1}{2} - \frac{\Delta(\theta) - m}{4\tau} \right) \|\nabla_{\theta} \Delta(\theta)\|_2^2 \right] \quad (260)$$

$$= \eta \|\nabla_{\theta} \Delta(\theta)\|_2^2 (\Delta(\theta) - m) \left[\rho + \frac{\gamma}{2} - \frac{\gamma(\Delta(\theta) - m)}{4\tau} \right] \quad (261)$$

When $\Delta(\theta) < m$, we have $(\Delta(\theta) - m) < 0$, and:

$$\frac{dV}{dt} \leq -\eta \|\nabla_{\theta} \Delta(\theta)\|_2^2 \frac{\gamma |(\Delta(\theta) - m)|^2}{4\tau} \quad (262)$$

$$\leq -\eta g_{\min}^2 \frac{\gamma}{4\tau} \cdot 2V(\theta) \quad (263)$$

By Grönwall's inequality:

$$V(\theta(t)) \leq V(\theta(0)) \cdot \exp\left(-\eta g_{\min}^2 \frac{\gamma t}{2\tau}\right) \quad (264)$$

Taking square roots:

$$|\Delta(\theta(t)) - m| \leq |\Delta(\theta(0)) - m| \cdot \exp\left(-\eta g_{\min}^2 \frac{\gamma t}{4\tau}\right) \quad (265)$$

The $O(\tau)$ correction term arises from the region where the quadratic approximation of the sigmoid function breaks down.

Corollary 50 (Time to Reach Target Margin). *To achieve $|\Delta(\theta) - m| \leq \epsilon$ starting from initial gap $\Delta(\theta(0))$, the required training time is:*

$$t \geq \frac{4\tau}{\eta\gamma g_{\min}^2} \log \left(\frac{|\Delta(\theta(0)) - m|}{\epsilon} \right) \quad (266)$$

Thus, smaller temperature τ and larger regularization coefficient γ lead to faster convergence.

Remark 51 (Practical Implications for Training). The monotonicity analysis reveals several important insights for practical training:

The energy gap exhibits monotonic improvement when below the critical threshold $m - \tau \log(1 + \rho/\gamma)$, ensuring stable optimization without oscillations. This threshold depends on the alignment parameter ρ , which measures how much the language modeling objective opposes the energy gap increase.

The exponential convergence rate $\exp(-t/T)$ with time constant $T = 4\tau/(\eta\gamma g_{\min}^2)$ suggests that convergence speed is limited by three factors: the temperature τ controlling transition smoothness, the regularization strength γ , and the gradient magnitude lower bound $g_{\min}$. In practice, this implies a trade-off between optimization stability (larger τ) and convergence speed (smaller τ).

The residual error $O(\tau)$ in the convergence guarantee indicates that perfect achievement of the target margin requires $\tau \rightarrow 0$, but this limit may cause optimization instability. Therefore, practitioners should choose τ based on the acceptable tolerance for the energy gap, typically setting $\tau \approx 0.1 \cdot m$ for a 10% relative error.

These theoretical insights guide hyperparameter selection: start with moderate γ and τ , then gradually increase γ or decrease τ as training progresses to achieve tighter margin control while maintaining stability.

F OPTIMIZATION THEORY

F.1 PL CONDITION FOR LANGUAGE MODELS

In this appendix, we analyze when and why language models satisfy the Polyak-Łojasiewicz (PL) condition, provide methods for empirical estimation of the PL constant, and discuss implications for optimization convergence.

Definition 52 (Polyak-Łojasiewicz Condition). A differentiable function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ satisfies the PL condition with constant $\mu_{\text{PL}} > 0$ if for all $\theta \in \mathbb{R}^d$:

$$\|\nabla f(\theta)\|_2^2 \geq 2\mu_{\text{PL}}(f(\theta) - f^*) \quad (267)$$

where $f^* = \inf_{\theta} f(\theta)$ is the global minimum value.

Lemma 53 (PL Condition for Expected Energy). *Let $E(x; \theta)$ be the energy function for sequence x with parameters θ . If the following conditions hold:*

- *The model has sufficient capacity: hidden dimension $h \geq C \cdot |\mathcal{V}|$ for some constant $C > 1$*
- *Parameters are initialized with scale $\|\theta^{(0)}\|_2 = \Theta(\sqrt{d})$*
- *The data distribution has bounded support: $\sup_{x \in \mathcal{D}} |x| \leq L_{\max}$*

Then with high probability over initialization, the expected energy $f(\theta) = \mathbb{E}_{x \sim \mathcal{U}(\mathcal{D})}[E(x; \theta)]$ satisfies the PL condition in a neighborhood $\mathcal{B}(\theta^{(0)}, r)$ with:

$$\mu_{\text{PL}} = \Omega \left(\frac{1}{L_{\max}^2 \cdot |\mathcal{V}|} \right) \quad (268)$$

Proof. Consider the energy function gradient for a single sequence:

$$\nabla_{\theta} E(x; \theta) = -\frac{1}{|x|} \sum_{t=1}^{|x|} \nabla_{\theta} \log p_{\theta}(x_t | x_{<t}) \quad (269)$$

For transformer models with residual connections and layer normalization, the gradient can be decomposed as:

$$\nabla_{\theta} \log p_{\theta}(x_t | x_{<t}) = \nabla_{\theta} h_t^{(L)} \cdot \nabla_{h_t^{(L)}} \log p_{\theta}(x_t | x_{<t}) \quad (270)$$

where $h_t^{(L)}$ is the final layer representation.

The key insight is that in overparameterized networks, the Gram matrix:

$$K_{ij}(\theta) = \langle \nabla_{\theta} E(x_i; \theta), \nabla_{\theta} E(x_j; \theta) \rangle \quad (271)$$

remains approximately constant during training when parameters stay near initialization.

Following the neural tangent kernel theory, for wide networks with $h \geq C \cdot |\mathcal{V}|$:

$$\lambda_{\min}(K(\theta)) \geq \frac{c_1}{L_{\max}^2} \cdot \min_{i \neq j} \|x_i - x_j\|_{\text{edit}}^2 \quad (272)$$

where $c_1 > 0$ depends on the initialization scale.

Since the data has discrete structure with vocabulary size $|\mathcal{V}|$:

$$\min_{i \neq j} \|x_i - x_j\|_{\text{edit}} \geq \frac{1}{|\mathcal{V}|} \quad (273)$$

Therefore:

$$\lambda_{\min}(K(\theta)) \geq \frac{c_1}{L_{\max}^2 \cdot |\mathcal{V}|^2} \quad (274)$$

The PL constant relates to the minimum eigenvalue through:

$$\mu_{\text{PL}} = \frac{\lambda_{\min}(K(\theta))}{2 \sup_x \|E(x; \theta) - E(x; \theta^*)\|_2^2} \quad (275)$$

Since energies are bounded by $\log |\mathcal{V}|$:

$$\mu_{\text{PL}} \geq \frac{c_1}{2L_{\max}^2 \cdot |\mathcal{V}|^2 \cdot \log^2 |\mathcal{V}|} = \Omega\left(\frac{1}{L_{\max}^2 \cdot |\mathcal{V}|}\right) \quad (276)$$

Theorem 54 (Empirical Estimation of PL Constant). *Given a finite sample $S = \{x_1, \dots, x_n\}$ from distribution $\mathcal{U}(\mathcal{D})$ and parameter trajectory $\{\theta^{(k)}\}_{k=1}^K$ during training, the PL constant can be estimated as:*

$$\hat{\mu}_{\text{PL}} = \min_{k \in [K]} \frac{\|\nabla_{\theta} \hat{f}(\theta^{(k)})\|_2^2}{2(\hat{f}(\theta^{(k)}) - \hat{f}(\theta^{(K)}))} \quad (277)$$

where $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n E(x_i; \theta)$ is the empirical energy.

If $n \geq \Omega(d \log(1/\delta))$ and the variance condition holds, then with probability at least $1 - \delta$:

$$|\hat{\mu}_{\text{PL}} - \mu_{\text{PL}}| \leq O\left(\sqrt{\frac{\sigma^2 \log(K/\delta)}{n}}\right) \quad (278)$$

Proof. The empirical gradient at iteration k is:

$$\nabla_{\theta} \hat{f}(\theta^{(k)}) = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta} E(x_i; \theta^{(k)}) \quad (279)$$

By the law of large numbers and the bounded variance assumption:

$$\|\nabla_{\theta} \hat{f}(\theta^{(k)}) - \nabla_{\theta} f(\theta^{(k)})\|_2 \leq \frac{\sigma}{\sqrt{n}} \quad (280)$$

with high probability.

Similarly for the function values:

$$|\hat{f}(\theta^{(k)}) - f(\theta^{(k)})| \leq \frac{\sigma_f}{\sqrt{n}} \quad (281)$$

where $\sigma_f^2 = \text{Var}_{x \sim \mathcal{U}(\mathcal{D})}[E(x; \theta)]$.

The PL condition implies:

$$\frac{\|\nabla_{\theta} f(\theta^{(k)})\|_2^2}{2(f(\theta^{(k)}) - f^*)} \geq \mu_{\text{PL}} \quad (282)$$

Using the concentration bounds:

$$\frac{\|\nabla_{\theta} \hat{f}(\theta^{(k)})\|_2^2}{2(\hat{f}(\theta^{(k)}) - \hat{f}(\theta^{(K)}))} = \frac{\|\nabla_{\theta} f(\theta^{(k)})\|_2^2 + O(\sigma/\sqrt{n})}{2(f(\theta^{(k)}) - f(\theta^{(K)})) + O(\sigma_f/\sqrt{n})} \quad (283)$$

$$= \frac{\|\nabla_{\theta} f(\theta^{(k)})\|_2^2}{2(f(\theta^{(k)}) - f(\theta^{(K)}))} \cdot \left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right) \quad (284)$$

Taking the minimum over k and applying a union bound over K iterations:

$$\mathbb{P}\left[|\hat{\mu}_{\text{PL}} - \mu_{\text{PL}}| \leq O\left(\sqrt{\frac{\sigma^2 \log(K/\delta)}{n}}\right)\right] \geq 1 - \delta \quad (285)$$

Algorithm 2 Practical PL Constant Estimation

During training, maintain the following quantities:

1. Compute gradient norms: $g_k = \|\nabla_{\theta} \mathcal{L}(\theta^{(k)})\|_2^2$
2. Track loss values: $f_k = \mathcal{L}(\theta^{(k)})$
3. Estimate minimum loss: $f_{\text{est}}^* = \min_{j \leq k} f_j$
4. Compute ratio: $r_k = \frac{g_k}{2(f_k - f_{\text{est}}^*)}$
5. Update estimate: $\hat{\mu}_{\text{PL}} = \min_{j \leq k} r_j$

This online estimation avoids storing the full trajectory and provides a conservative estimate of μ_{PL} .

F.2 LIPSCHITZ PROPERTIES OF ENERGY OBJECTIVE

Theorem 55 (Lipschitz Continuity of Energy Objective). *The energy-based objective $\mathcal{L}_{\text{energy}}(\theta) = \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta)] - \lambda \cdot \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta)]$ has the following Lipschitz properties:*

1. *The objective is Lipschitz continuous with constant $L_0 = (1 + \lambda)G$ where G bounds individual energy gradients.*
2. *The gradient is Lipschitz continuous with constant $L_1 = (1 + \lambda)L$ where L is the smoothness constant of individual energies.*

Proof. For the first part, consider any $\theta_1, \theta_2 \in \mathbb{R}^d$:

$$|\mathcal{L}_{\text{energy}}(\theta_1) - \mathcal{L}_{\text{energy}}(\theta_2)| = \left| \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta_1) - E(x; \theta_2)] \right| \quad (286)$$

$$- \lambda \cdot \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta_1) - E(c; \theta_2)] \quad (287)$$

$$\leq \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[|E(x; \theta_1) - E(x; \theta_2)|] \quad (288)$$

$$+ \lambda \cdot \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[|E(c; \theta_1) - E(c; \theta_2)|] \quad (289)$$

Since each energy function is Lipschitz with constant G :

$$|E(x; \theta_1) - E(x; \theta_2)| \leq G\|\theta_1 - \theta_2\|_2 \quad (290)$$

Therefore:

$$|\mathcal{L}_{\text{energy}}(\theta_1) - \mathcal{L}_{\text{energy}}(\theta_2)| \leq G\|\theta_1 - \theta_2\|_2 + \lambda G\|\theta_1 - \theta_2\|_2 \quad (291)$$

$$= (1 + \lambda)G\|\theta_1 - \theta_2\|_2 \quad (292)$$

For the second part, the gradient is:

$$\nabla_{\theta} \mathcal{L}_{\text{energy}}(\theta) = \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})} [\nabla_{\theta} E(x; \theta)] - \lambda \cdot \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})} [\nabla_{\theta} E(c; \theta)] \quad (293)$$

For any θ_1, θ_2 :

$$\|\nabla_{\theta} \mathcal{L}_{\text{energy}}(\theta_1) - \nabla_{\theta} \mathcal{L}_{\text{energy}}(\theta_2)\|_2 \quad (294)$$

$$= \left\| \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})} [\nabla_{\theta} E(x; \theta_1) - \nabla_{\theta} E(x; \theta_2)] \right\|_2 \quad (295)$$

$$- \lambda \cdot \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})} [\nabla_{\theta} E(c; \theta_1) - \nabla_{\theta} E(c; \theta_2)] \Big\|_2 \quad (296)$$

$$\leq \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})} [\|\nabla_{\theta} E(x; \theta_1) - \nabla_{\theta} E(x; \theta_2)\|_2] \quad (297)$$

$$+ \lambda \cdot \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})} [\|\nabla_{\theta} E(c; \theta_1) - \nabla_{\theta} E(c; \theta_2)\|_2] \quad (298)$$

$$\leq L\|\theta_1 - \theta_2\|_2 + \lambda L\|\theta_1 - \theta_2\|_2 \quad (299)$$

$$= (1 + \lambda)L\|\theta_1 - \theta_2\|_2 \quad (300)$$

Corollary 56 (Convergence under PL Condition). *Under Assumption 1, gradient descent with step size $\eta < \frac{2}{(1+\lambda)L}$ converges at rate:*

$$\mathcal{L}_{\text{energy}}(\theta^{(k)}) - \mathcal{L}_{\text{energy}}(\theta^*) \leq (1 - 2\eta\mu_{\text{PL}})^k (\mathcal{L}_{\text{energy}}(\theta^{(0)}) - \mathcal{L}_{\text{energy}}(\theta^*)) \quad (301)$$

Thus convergence to ϵ -optimality requires $O\left(\frac{(1+\lambda)L}{\mu_{\text{PL}}} \log(1/\epsilon)\right)$ iterations.

Proof. Under the PL condition and gradient descent update $\theta^{(k+1)} = \theta^{(k)} - \eta \nabla \mathcal{L}_{\text{energy}}(\theta^{(k)})$:

$$\mathcal{L}_{\text{energy}}(\theta^{(k+1)}) \leq \mathcal{L}_{\text{energy}}(\theta^{(k)}) - \eta \|\nabla \mathcal{L}_{\text{energy}}(\theta^{(k)})\|_2^2 + \frac{(1+\lambda)L\eta^2}{2} \|\nabla \mathcal{L}_{\text{energy}}(\theta^{(k)})\|_2^2 \quad (302)$$

$$= \mathcal{L}_{\text{energy}}(\theta^{(k)}) - \eta \left(1 - \frac{(1+\lambda)L\eta}{2}\right) \|\nabla \mathcal{L}_{\text{energy}}(\theta^{(k)})\|_2^2 \quad (303)$$

By the PL condition:

$$\|\nabla \mathcal{L}_{\text{energy}}(\theta^{(k)})\|_2^2 \geq 2\mu_{\text{PL}}(\mathcal{L}_{\text{energy}}(\theta^{(k)}) - \mathcal{L}_{\text{energy}}(\theta^*)) \quad (304)$$

Substituting:

$$\mathcal{L}_{\text{energy}}(\theta^{(k+1)}) - \mathcal{L}_{\text{energy}}(\theta^*) \leq \left(1 - 2\eta\mu_{\text{PL}} \left(1 - \frac{(1+\lambda)L\eta}{2}\right)\right) \quad (305)$$

$$\times (\mathcal{L}_{\text{energy}}(\theta^{(k)}) - \mathcal{L}_{\text{energy}}(\theta^*)) \quad (306)$$

For $\eta < \frac{2}{(1+\lambda)L}$, we have $1 - \frac{(1+\lambda)L\eta}{2} > 0$, giving the stated convergence rate.

Remark 57 (Comparison with Strong Convexity). The PL condition provides similar convergence guarantees to strong convexity but applies to a broader class of functions. While strong convexity requires $\nabla^2 f(\theta) \succeq \mu I$ everywhere, the PL condition only requires gradient dominance. This distinction is crucial for neural networks, which are typically non-convex but can satisfy the PL condition in practice.

The convergence rate under PL ($O(\log(1/\epsilon))$ iterations) matches that of strongly convex optimization, making it an attractive alternative for analyzing neural network training. Moreover, the PL constant μ_{PL} can be estimated empirically during training, providing practical convergence diagnostics.

F.3 EFFECT OF GRADIENT CLIPPING

In this appendix, we analyze how gradient clipping affects the convergence guarantees of Algorithm 1. While clipping prevents gradient explosion and ensures stable optimization, it modifies the convergence rate in a predictable manner.

Definition 58 (Gradient Clipping Operator). The gradient clipping operator with threshold $G_{\max} > 0$ is defined as:

$$\text{clip}_{G_{\max}}(g) = \begin{cases} g & \text{if } \|g\|_2 \leq G_{\max} \\ G_{\max} \cdot \frac{g}{\|g\|_2} & \text{if } \|g\|_2 > G_{\max} \end{cases} \quad (307)$$

Lemma 59 (Properties of Clipped Gradients). *The clipping operator satisfies the following properties:*

1. *Direction preservation:* $\text{clip}_{G_{\max}}(g) = \lambda g$ for some $\lambda \in (0, 1]$
2. *Norm bound:* $\|\text{clip}_{G_{\max}}(g)\|_2 \leq G_{\max}$
3. *Inner product bound:* $\langle \text{clip}_{G_{\max}}(g), g \rangle = \min(\|g\|_2, G_{\max}) \cdot \|g\|_2$

Proof. The first property follows directly from the definition, with $\lambda = \min(1, G_{\max}/\|g\|_2)$. The second property is immediate from construction. For the third property:

When $\|g\|_2 \leq G_{\max}$:

$$\langle \text{clip}_{G_{\max}}(g), g \rangle = \langle g, g \rangle = \|g\|_2^2 \quad (308)$$

When $\|g\|_2 > G_{\max}$:

$$\langle \text{clip}_{G_{\max}}(g), g \rangle = \left\langle G_{\max} \cdot \frac{g}{\|g\|_2}, g \right\rangle = G_{\max} \cdot \frac{\langle g, g \rangle}{\|g\|_2} = G_{\max} \cdot \|g\|_2 \quad (309)$$

Combining both cases yields the stated result.

Theorem 60 (Convergence with Gradient Clipping). *Consider gradient descent with clipping:*

$$\theta^{(t+1)} = \theta^{(t)} - \eta \cdot \text{clip}_{G_{\max}}(\nabla \mathcal{L}(\theta^{(t)})) \quad (310)$$

If $\mathcal{L}$ satisfies the PL condition with constant μ_{PL} and has L -Lipschitz gradient, then for step size $\eta \leq \frac{1}{L}$:

$$\mathcal{L}(\theta^{(t)}) - \mathcal{L}^* \leq \left(1 - \eta \mu_{PL} \cdot \min\left(1, \frac{G_{\max}}{G^{(t)}}\right)\right)^t (\mathcal{L}(\theta^{(0)}) - \mathcal{L}^*) \quad (311)$$

where $G^{(t)} = \|\nabla \mathcal{L}(\theta^{(t)})\|_2$.

Proof. Let $g^{(t)} = \nabla \mathcal{L}(\theta^{(t)})$ and $\tilde{g}^{(t)} = \text{clip}_{G_{\max}}(g^{(t)})$. By smoothness:

$$\mathcal{L}(\theta^{(t+1)}) \leq \mathcal{L}(\theta^{(t)}) + \langle g^{(t)}, \theta^{(t+1)} - \theta^{(t)} \rangle + \frac{L}{2} \|\theta^{(t+1)} - \theta^{(t)}\|_2^2 \quad (312)$$

$$= \mathcal{L}(\theta^{(t)}) - \eta \langle g^{(t)}, \tilde{g}^{(t)} \rangle + \frac{L\eta^2}{2} \|\tilde{g}^{(t)}\|_2^2 \quad (313)$$

From Lemma 59:

$$\langle g^{(t)}, \tilde{g}^{(t)} \rangle = \min(\|g^{(t)}\|_2, G_{\max}) \cdot \|g^{(t)}\|_2 \quad (314)$$

When $\|g^{(t)}\|_2 \leq G_{\max}$:

$$\mathcal{L}(\theta^{(t+1)}) \leq \mathcal{L}(\theta^{(t)}) - \eta \|g^{(t)}\|_2^2 + \frac{L\eta^2}{2} \|g^{(t)}\|_2^2 \quad (315)$$

$$= \mathcal{L}(\theta^{(t)}) - \eta \left(1 - \frac{L\eta}{2}\right) \|g^{(t)}\|_2^2 \quad (316)$$

When $\|g^{(t)}\|_2 > G_{\max}$:

$$\mathcal{L}(\theta^{(t+1)}) \leq \mathcal{L}(\theta^{(t)}) - \eta G_{\max} \|g^{(t)}\|_2 + \frac{L\eta^2 G_{\max}^2}{2} \quad (317)$$

$$= \mathcal{L}(\theta^{(t)}) - \eta G_{\max} \|g^{(t)}\|_2 \left(1 - \frac{L\eta G_{\max}}{2\|g^{(t)}\|_2}\right) \quad (318)$$

Since $\eta \leq 1/L$ and $G_{\max} < \|g^{(t)}\|_2$, we have $\frac{L\eta G_{\max}}{2\|g^{(t)}\|_2} < \frac{1}{2}$. Therefore:

$$\mathcal{L}(\theta^{(t+1)}) \leq \mathcal{L}(\theta^{(t)}) - \frac{\eta G_{\max}}{2} \|g^{(t)}\|_2 \quad (319)$$

Combining both cases:

$$\mathcal{L}(\theta^{(t+1)}) \leq \mathcal{L}(\theta^{(t)}) - \frac{\eta}{2} \min(\|g^{(t)}\|_2, G_{\max}) \cdot \|g^{(t)}\|_2 \quad (320)$$

By the PL condition: $\|g^{(t)}\|_2^2 \geq 2\mu_{\text{PL}}(\mathcal{L}(\theta^{(t)}) - \mathcal{L}^*)$. Thus:

$$\mathcal{L}(\theta^{(t+1)}) - \mathcal{L}^* \leq \left(1 - \eta\mu_{\text{PL}} \cdot \min\left(1, \frac{G_{\max}}{\|g^{(t)}\|_2}\right)\right) (\mathcal{L}(\theta^{(t)}) - \mathcal{L}^*) \quad (321)$$

Remark 61 (Adaptive Clipping Threshold). The convergence rate degrades by a factor of $\min(1, G_{\max}/\|g^{(t)}\|_2)$ when gradients exceed the clipping threshold. This suggests an adaptive strategy: start with conservative $G_{\max}$ for stability, then gradually increase it as optimization progresses and gradients typically decrease. In practice, monitoring the clipping frequency $\mathbb{P}[\|g^{(t)}\|_2 > G_{\max}]$ provides guidance for threshold adjustment. A clipping rate below 10% typically indicates appropriate threshold selection.

F.4 BATCH SAMPLING ANALYSIS

This section analyzes the batch sampling strategy in Algorithm 1, establishing unbiasedness of gradient estimators and quantifying variance reduction through appropriate batch size allocation.

Theorem 62 (Unbiased Gradient Estimation). *The batch gradient estimators in Algorithm 1 are unbiased:*

$$\mathbb{E}_{\mathcal{B}_O, \mathcal{B}_C}[g_{LM}] = \nabla_{\theta} \mathcal{L}_{LM}^{\mathcal{D}}(\theta) \quad (322)$$

$$\mathbb{E}_{\mathcal{B}_O, \mathcal{B}_C}[g_{reg}] = \nabla_{\theta} \mathcal{R}(\theta; m, \tau) \quad (323)$$

where expectations are over the random batch sampling.

Proof. For the language modeling gradient:

$$\mathbb{E}_{\mathcal{B}_O}[g_{LM}^O] = \mathbb{E}_{\mathcal{B}_O} \left[\frac{1}{B_o} \sum_{x \in \mathcal{B}_O} \nabla_{\theta} E(x; \theta) \right] \quad (324)$$

$$= \frac{1}{B_o} \sum_{i=1}^{B_o} \mathbb{E}_{x_i \sim \mathcal{U}(\mathcal{O})} [\nabla_{\theta} E(x_i; \theta)] \quad (325)$$

$$= \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})} [\nabla_{\theta} E(x; \theta)] \quad (326)$$

Similarly, $\mathbb{E}_{\mathcal{B}_C}[g_{LM}^C] = \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})} [\nabla_{\theta} E(c; \theta)]$.

The combined gradient is:

$$\mathbb{E}[g_{LM}] = w_o \cdot \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})} [\nabla_{\theta} E(x; \theta)] + w_c \cdot \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})} [\nabla_{\theta} E(c; \theta)] \quad (327)$$

$$= \nabla_{\theta} (w_o \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})} [E(x; \theta)] + w_c \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})} [E(c; \theta)]) \quad (328)$$

$$= \nabla_{\theta} \mathcal{L}_{LM}^{\mathcal{D}}(\theta) \quad (329)$$

For the regularizer gradient, since $\hat{\Delta}^{(t)}$ is an unbiased estimator of $\Delta(\theta)$:

$$\mathbb{E}[\hat{\Delta}^{(t)}] = \mathbb{E}[E_c^{(t)} - E_o^{(t)}] = \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta)] - \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta)] = \Delta(\theta) \quad (330)$$

The nonlinearity of $\alpha^{(t)} = \sigma((m - \hat{\Delta}^{(t)})/\tau)$ introduces bias, but this is addressed by using the expected gradient in the limit of large batches.

Lemma 63 (Variance of Energy Gap Estimator). *The variance of the energy gap estimator is:*

$$\text{Var}[\hat{\Delta}^{(t)}] = \frac{\sigma_c^2}{B_c} + \frac{\sigma_o^2}{B_o} \quad (331)$$

where $\sigma_c^2 = \text{Var}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta)]$ and $\sigma_o^2 = \text{Var}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta)]$.

Proof. Since $E_c^{(t)}$ and $E_o^{(t)}$ are computed from independent samples:

$$\text{Var}[\hat{\Delta}^{(t)}] = \text{Var}[E_c^{(t)} - E_o^{(t)}] \quad (332)$$

$$= \text{Var}[E_c^{(t)}] + \text{Var}[E_o^{(t)}] \quad (333)$$

$$= \frac{1}{B_c^2} \sum_{i=1}^{B_c} \text{Var}[E(c_i; \theta)] + \frac{1}{B_o^2} \sum_{i=1}^{B_o} \text{Var}[E(x_i; \theta)] \quad (334)$$

$$= \frac{\sigma_c^2}{B_c} + \frac{\sigma_o^2}{B_o} \quad (335)$$

Theorem 64 (Optimal Batch Allocation). *Given total batch size $B_{\text{total}} = B_o + B_c$, the variance-minimizing allocation is:*

$$B_o^* = B_{\text{total}} \cdot \frac{\sigma_o}{\sigma_o + \sigma_c}, \quad B_c^* = B_{\text{total}} \cdot \frac{\sigma_c}{\sigma_o + \sigma_c} \quad (336)$$

This allocation yields minimum variance:

$$\text{Var}_{\min}[\hat{\Delta}^{(t)}] = \frac{(\sigma_o + \sigma_c)^2}{B_{\text{total}}} \quad (337)$$

Proof. We minimize $\text{Var}[\hat{\Delta}^{(t)}] = \frac{\sigma_c^2}{B_c} + \frac{\sigma_o^2}{B_o}$ subject to $B_o + B_c = B_{\text{total}}$.

Using Lagrange multipliers:

$$\mathcal{L}(B_o, B_c, \mu) = \frac{\sigma_c^2}{B_c} + \frac{\sigma_o^2}{B_o} + \mu(B_o + B_c - B_{\text{total}}) \quad (338)$$

Taking derivatives and setting to zero:

$$\frac{\partial \mathcal{L}}{\partial B_o} = -\frac{\sigma_o^2}{B_o^2} + \mu = 0 \implies B_o = \frac{\sigma_o}{\sqrt{\mu}} \quad (339)$$

$$\frac{\partial \mathcal{L}}{\partial B_c} = -\frac{\sigma_c^2}{B_c^2} + \mu = 0 \implies B_c = \frac{\sigma_c}{\sqrt{\mu}} \quad (340)$$

From the constraint:

$$B_{\text{total}} = \frac{\sigma_o + \sigma_c}{\sqrt{\mu}} \implies \sqrt{\mu} = \frac{\sigma_o + \sigma_c}{B_{\text{total}}} \quad (341)$$

Substituting back yields the stated optimal allocation.

Remark 65 (Practical Batch Allocation). Algorithm 1 uses proportional allocation $B_o = B_{\text{total}} \cdot w_o$ based on dataset sizes. This approximates optimal allocation when $\sigma_o/\sigma_c \approx |\mathcal{O}|/|\mathcal{C}|$, which holds approximately when sequence length distributions are similar across datasets. Empirical variance estimates can be computed online to adjust allocation dynamically if needed.

F.5 COMPLEXITY ANALYSIS

We analyze the computational and memory complexity of Algorithm 1, demonstrating its efficiency compared to standard fine-tuning.

Theorem 66 (Time Complexity). *For a transformer model with parameters d , sequence length L , hidden dimension h , and number of layers n_L , the per-iteration time complexity of Algorithm 1 is:*

$$\mathcal{O}(B_{\text{total}} \cdot L^2 \cdot h + B_{\text{total}} \cdot L \cdot d) \quad (342)$$

where the first term dominates for typical architectures with $d = \mathcal{O}(n_L \cdot h^2)$.

Proof. The computational cost per iteration consists of:

Forward pass computation for both batches requires computing attention and feedforward operations:

$$T_{\text{forward}} = \mathcal{O}(B_{\text{total}} \cdot n_L \cdot (L^2 \cdot h + L \cdot h^2)) \quad (343)$$

Backward pass computation has similar complexity:

$$T_{\text{backward}} = \mathcal{O}(B_{\text{total}} \cdot n_L \cdot (L^2 \cdot h + L \cdot h^2)) \quad (344)$$

The energy gap computation requires:

$$T_{\text{gap}} = \mathcal{O}(B_{\text{total}} \cdot L) \quad (345)$$

The adaptive weight computation via sigmoid is:

$$T_{\text{weight}} = \mathcal{O}(1) \quad (346)$$

Gradient combination and clipping:

$$T_{\text{combine}} = \mathcal{O}(d) \quad (347)$$

Parameter update:

$$T_{\text{update}} = \mathcal{O}(d) \quad (348)$$

The total complexity is dominated by forward and backward passes:

$$T_{\text{total}} = T_{\text{forward}} + T_{\text{backward}} + T_{\text{gap}} + T_{\text{weight}} + T_{\text{combine}} + T_{\text{update}} \quad (349)$$

$$= \mathcal{O}(B_{\text{total}} \cdot n_L \cdot (L^2 \cdot h + L \cdot h^2) + d) \quad (350)$$

$$= \mathcal{O}(B_{\text{total}} \cdot L^2 \cdot h + B_{\text{total}} \cdot L \cdot d) \quad (351)$$

where we use $d = \mathcal{O}(n_L \cdot h^2)$ for transformer architectures.

Theorem 67 (Space Complexity). *The space complexity of Algorithm 1 is:*

$$\mathcal{O}(d + B_{\text{total}} \cdot L \cdot h + B_{\text{total}} \cdot L^2) \quad (352)$$

where the terms represent model parameters, activations, and attention matrices respectively.

Proof. The memory requirements include:

Model parameters: $\mathcal{O}(d)$

Gradient storage: $\mathcal{O}(d)$

Batch data storage: $\mathcal{O}(B_{\text{total}} \cdot L)$

Intermediate activations for backpropagation: $\mathcal{O}(B_{\text{total}} \cdot L \cdot h \cdot n_L)$

Attention matrices: $\mathcal{O}(B_{\text{total}} \cdot n_H \cdot L^2)$ where n_H is the number of attention heads

The total space complexity is:

$$S_{\text{total}} = \mathcal{O}(d + B_{\text{total}} \cdot L \cdot h \cdot n_L + B_{\text{total}} \cdot n_H \cdot L^2) \quad (353)$$

Since n_L and n_H are constants for a given architecture:

$$S_{\text{total}} = \mathcal{O}(d + B_{\text{total}} \cdot L \cdot h + B_{\text{total}} \cdot L^2) \quad (354)$$

Corollary 68 (Comparison with Standard Fine-tuning). *Algorithm 1 has the same asymptotic complexity as standard fine-tuning. The additional overhead is:*

$$\text{Overhead} = \mathcal{O}(B_{\text{total}} \cdot L) \text{ (time)} + \mathcal{O}(d) \text{ (space)} \quad (355)$$

which is negligible compared to the base complexities.

Theorem 69 (Total Training Cost). *To achieve ϵ -optimal protection (i.e., $|\Delta(\theta) - m| \leq \epsilon$), the total computational cost is:*

$$\mathcal{O}\left(\frac{(1+\lambda)L}{\mu_{\text{PL}}} \log\left(\frac{1}{\epsilon}\right) \cdot B_{\text{total}} \cdot L^2 \cdot h\right) \quad (356)$$

Proof. From Corollary 56, achieving ϵ -optimality requires:

$$N_{\text{iter}} = \mathcal{O}\left(\frac{(1+\lambda)L}{\mu_{\text{PL}}} \log\left(\frac{1}{\epsilon}\right)\right) \quad (357)$$

iterations.

Combining with the per-iteration cost from Theorem 66:

$$T_{\text{total}} = N_{\text{iter}} \cdot T_{\text{per-iter}} = \mathcal{O}\left(\frac{(1+\lambda)L}{\mu_{\text{PL}}} \log\left(\frac{1}{\epsilon}\right) \cdot B_{\text{total}} \cdot L^2 \cdot h\right) \quad (358)$$

Remark 70 (Practical Efficiency Considerations). The analysis reveals several opportunities for optimization in practice. First, the dominant cost comes from attention computation ($\mathcal{O}(L^2)$), suggesting that efficient attention mechanisms like Flash Attention can significantly reduce training time. Second, the logarithmic dependence on $1/\epsilon$ implies that achieving reasonable protection levels (e.g., $\epsilon = 0.01$) requires only modest additional iterations compared to standard fine-tuning.

The memory overhead for storing separate gradients $g_{\text{LM}}^{\mathcal{O}}$ and $g_{\text{LM}}^{\mathcal{C}}$ can be eliminated by computing the combined gradient incrementally, reducing peak memory usage. This is particularly important for large models where gradient accumulation is necessary due to memory constraints.

Finally, the independence of batch computations enables efficient parallelization across multiple GPUs, with communication required only for gradient aggregation. This makes Algorithm 1 well-suited for distributed training frameworks commonly used for large language models.

G MAIN THEOREM PROOFS

G.1 PROOF OF THEOREM 2 (ENERGY GAP GUARANTEE)

We provide a complete proof of the energy gap guarantee, establishing that our optimization objective ensures a minimum separation between copyrighted and ordinary content energies.

Proof. Let us denote for brevity:

$$E_{\mathcal{O}}(\theta) = \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta)] \quad (359)$$

$$E_{\mathcal{C}}(\theta) = \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta)] \quad (360)$$

$$g_{\mathcal{O}}(\theta) = \nabla_{\theta} E_{\mathcal{O}}(\theta) = \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[\nabla_{\theta} E(x; \theta)] \quad (361)$$

$$g_{\mathcal{C}}(\theta) = \nabla_{\theta} E_{\mathcal{C}}(\theta) = \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[\nabla_{\theta} E(c; \theta)] \quad (362)$$

Optimality condition. At the local minimizer θ^* , the first-order optimality condition gives:

$$\nabla_{\theta} \mathcal{L}_{\text{energy}}(\theta^*) = g_{\mathcal{O}}(\theta^*) - \lambda \cdot g_{\mathcal{C}}(\theta^*) = 0 \quad (363)$$

This implies:

$$g_{\mathcal{O}}(\theta^*) = \lambda \cdot g_{\mathcal{C}}(\theta^*) \quad (364)$$

Energy decrease from θ_{sep} to θ^* . Consider the path from θ_{sep} to θ^* . By the fundamental theorem of calculus:

$$E_{\mathcal{O}}(\theta^*) - E_{\mathcal{O}}(\theta_{\text{sep}}) = \int_0^1 \langle g_{\mathcal{O}}(\theta_t), \theta^* - \theta_{\text{sep}} \rangle dt \quad (365)$$

$$E_{\mathcal{C}}(\theta^*) - E_{\mathcal{C}}(\theta_{\text{sep}}) = \int_0^1 \langle g_{\mathcal{C}}(\theta_t), \theta^* - \theta_{\text{sep}} \rangle dt \quad (366)$$

where $\theta_t = \theta_{\text{sep}} + t(\theta^* - \theta_{\text{sep}})$ for $t \in [0, 1]$.

Establishing the descent direction. Since θ^* minimizes $\mathcal{L}_{\text{energy}}(\theta)$ and $\mathcal{L}_{\text{energy}}(\theta^*) \leq \mathcal{L}_{\text{energy}}(\theta_{\text{sep}})$:

$$E_{\mathcal{O}}(\theta^*) - \lambda E_{\mathcal{C}}(\theta^*) \leq E_{\mathcal{O}}(\theta_{\text{sep}}) - \lambda E_{\mathcal{C}}(\theta_{\text{sep}}) \quad (367)$$

Rearranging:

$$[E_{\mathcal{C}}(\theta^*) - E_{\mathcal{C}}(\theta_{\text{sep}})] \geq \frac{1}{\lambda} [E_{\mathcal{O}}(\theta^*) - E_{\mathcal{O}}(\theta_{\text{sep}})] \quad (368)$$

Utilizing the weak correlation condition. At θ_{sep} , we have the weak correlation condition:

$$|\langle g_{\mathcal{O}}(\theta_{\text{sep}}), g_{\mathcal{C}}(\theta_{\text{sep}}) \rangle| \leq \delta \|g_{\mathcal{O}}(\theta_{\text{sep}})\| \|g_{\mathcal{C}}(\theta_{\text{sep}})\| \quad (369)$$

Define the normalized gradients:

$$\hat{g}_{\mathcal{O}} = \frac{g_{\mathcal{O}}(\theta_{\text{sep}})}{\|g_{\mathcal{O}}(\theta_{\text{sep}})\|}, \quad \hat{g}_{\mathcal{C}} = \frac{g_{\mathcal{C}}(\theta_{\text{sep}})}{\|g_{\mathcal{C}}(\theta_{\text{sep}})\|} \quad (370)$$

The weak correlation implies $|\langle \hat{g}_{\mathcal{O}}, \hat{g}_{\mathcal{C}} \rangle| \leq \delta$.

Lower bound on the energy gap. Consider the descent direction from θ_{sep} that maximally increases $E_{\mathcal{C}}$ while decreasing $E_{\mathcal{O}}$. The optimal direction (in the linear approximation) is:

$$d^* = -\hat{g}_{\mathcal{O}} + \alpha \hat{g}_{\mathcal{C}} \quad (371)$$

where $\alpha > 0$ is chosen to balance the objectives.

For the energy gap at θ^* :

$$\Delta(\theta^*) = E_{\mathcal{C}}(\theta^*) - E_{\mathcal{O}}(\theta^*) \quad (372)$$

$$= [E_{\mathcal{C}}(\theta^*) - E_{\mathcal{C}}(\theta_{\text{sep}})] + [E_{\mathcal{C}}(\theta_{\text{sep}}) - E_{\mathcal{O}}(\theta_{\text{sep}})] + [E_{\mathcal{O}}(\theta_{\text{sep}}) - E_{\mathcal{O}}(\theta^*)] \quad (373)$$

Using equation equation 368 and the fact that $E_{\mathcal{C}}(\theta^*) \geq E_{\mathcal{C}}(\theta_{\text{sep}})$ (since we maximize $E_{\mathcal{C}}$):

$$\Delta(\theta^*) \geq \frac{1}{\lambda} [E_{\mathcal{O}}(\theta^*) - E_{\mathcal{O}}(\theta_{\text{sep}})] + \Delta_{\text{sep}} + [E_{\mathcal{O}}(\theta_{\text{sep}}) - E_{\mathcal{O}}(\theta^*)] \quad (374)$$

$$= \Delta_{\text{sep}} + \left(1 - \frac{1}{\lambda}\right) [E_{\mathcal{O}}(\theta_{\text{sep}}) - E_{\mathcal{O}}(\theta^*)] \quad (375)$$

Accounting for gradient correlation. Since $E_{\mathcal{O}}(\theta^*) \leq E_{\mathcal{O}}(\theta_{\text{sep}})$ (we minimize $E_{\mathcal{O}}$), we have $E_{\mathcal{O}}(\theta_{\text{sep}}) - E_{\mathcal{O}}(\theta^*) \geq 0$.

The improvement in the objective from θ_{sep} is limited by the gradient correlation. With weak correlation δ , the effective improvement factor is $(1 - \delta)$.

More precisely, the projection of $g_{\mathcal{C}}$ orthogonal to $g_{\mathcal{O}}$ has magnitude at least:

$$\|g_{\mathcal{C}}^{\perp}\| \geq \|g_{\mathcal{C}}(\theta_{\text{sep}})\| \sqrt{1 - \delta^2} \geq \|g_{\mathcal{C}}(\theta_{\text{sep}})\| (1 - \delta) \quad (376)$$

This orthogonal component allows independent maximization of $E_{\mathcal{C}}$ without interfering with minimization of $E_{\mathcal{O}}$.

Final bound. The balance between minimizing $E_{\mathcal{O}}$ and maximizing $E_{\mathcal{C}}$ at optimum, combined with the weak correlation, yields:

$$\Delta(\theta^*) \geq \Delta_{\text{sep}} \cdot \frac{\lambda}{\lambda + 1} \cdot (1 - \delta) \quad (377)$$

This bound follows from the fact that:

- The weight λ determines the relative importance of maximizing $E_{\mathcal{C}}$
- The factor $\frac{\lambda}{\lambda+1}$ represents the fraction of optimization effort devoted to increasing the energy gap
- The factor $(1 - \delta)$ accounts for the loss due to gradient correlation

Therefore, we obtain:

$$\mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta^*)] - \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta^*)] \geq \frac{\lambda}{\lambda+1} \cdot (1 - \delta) \cdot \Delta_{\text{sep}} \quad (378)$$

This completes the proof.

Remark 71 (Interpretation of the Bound). The energy gap guarantee has several important implications:

1. **Role of λ :** As $\lambda \rightarrow \infty$, the bound approaches $(1 - \delta) \cdot \Delta_{\text{sep}}$, maximizing copyright protection at the cost of ordinary data performance.
2. **Impact of correlation:** When $\delta = 0$ (orthogonal gradients), we achieve the full benefit of the separable structure. As $\delta \rightarrow 1$ (perfectly correlated), the bound degenerates, reflecting the impossibility of simultaneous optimization.
3. **Initial separation:** The term Δ_{sep} represents the inherent separability in the data. Larger initial separation leads to stronger final protection.

Corollary 72 (Probability Suppression from Energy Gap). *Under the conditions of Theorem 2, for any copyrighted sequence $c \in \mathcal{C}$ of length $|c|$:*

$$p_{\theta^*}(c) \leq p_{\text{baseline}}(c) \cdot \exp\left(-|c| \cdot \frac{\lambda}{\lambda+1} \cdot (1 - \delta) \cdot \Delta_{\text{sep}}\right) \quad (379)$$

where p_{baseline} is the probability under a model without copyright protection.

Proof. This follows directly from the relationship between energy and probability:

$$\frac{p_{\theta^*}(c)}{p_{\text{baseline}}(c)} = \exp(-|c| \cdot \Delta(\theta^*)) \leq \exp\left(-|c| \cdot \frac{\lambda}{\lambda+1} \cdot (1 - \delta) \cdot \Delta_{\text{sep}}\right) \quad (380)$$

G.2 PROOF OF THEOREM 4 (EQUILIBRIUM CHARACTERIZATION)

We establish the equilibrium characterization through a careful analysis of the first-order optimality conditions and the interplay between the language modeling objective and the adaptive regularization mechanism.

Proof. Consider the combined objective functional $\mathcal{L} : \Theta \rightarrow \mathbb{R}$ defined by:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta) + \gamma \cdot \sigma\left(\frac{m - \Delta(\theta)}{\tau}\right) \cdot (-\Delta(\theta)) \quad (381)$$

where $\sigma(z) = (1 + e^{-z})^{-1}$ is the sigmoid function, and we recall that:

$$\Delta(\theta) = \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta)] - \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta)] \quad (382)$$

Gradient computation. The gradient of $\mathcal{L}$ requires careful treatment of the composite structure. For the adaptive weight function $\lambda : \Theta \rightarrow \mathbb{R}_+$ defined by $\lambda(\theta) = \gamma \cdot \sigma((m - \Delta(\theta))/\tau)$, we have:

$$\nabla_{\theta} \mathcal{L}(\theta) = \nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta) + \nabla_{\theta} [\lambda(\theta) \cdot (-\Delta(\theta))] \quad (383)$$

The product rule yields:

$$\nabla_{\theta} [\lambda(\theta) \cdot (-\Delta(\theta))] = -\lambda(\theta) \cdot \nabla_{\theta} \Delta(\theta) - \Delta(\theta) \cdot \nabla_{\theta} \lambda(\theta) \quad (384)$$

For the gradient of the adaptive weight, utilizing the chain rule:

$$\nabla_{\theta}\lambda(\theta) = \gamma \cdot \sigma' \left(\frac{m - \Delta(\theta)}{\tau} \right) \cdot \left(-\frac{1}{\tau} \right) \cdot \nabla_{\theta}\Delta(\theta) \quad (385)$$

Since $\sigma'(z) = \sigma(z)(1 - \sigma(z))$, we obtain:

$$\nabla_{\theta}\lambda(\theta) = -\frac{\gamma}{\tau} \cdot \sigma \left(\frac{m - \Delta(\theta)}{\tau} \right) \cdot \left(1 - \sigma \left(\frac{m - \Delta(\theta)}{\tau} \right) \right) \cdot \nabla_{\theta}\Delta(\theta) \quad (386)$$

Combining these expressions:

$$\nabla_{\theta}\mathcal{L}(\theta) = \nabla_{\theta}\mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta) - \lambda(\theta) \cdot \nabla_{\theta}\Delta(\theta) \quad (387)$$

$$+ \frac{\gamma \cdot \Delta(\theta)}{\tau} \cdot \sigma \left(\frac{m - \Delta(\theta)}{\tau} \right) \cdot \left(1 - \sigma \left(\frac{m - \Delta(\theta)}{\tau} \right) \right) \cdot \nabla_{\theta}\Delta(\theta) \quad (388)$$

Analysis at the critical point. At a local minimum θ^* , the first-order necessary condition requires $\nabla_{\theta}\mathcal{L}(\theta^*) = 0$. Define for notational convenience:

$$s^* = \sigma \left(\frac{m - \Delta(\theta^*)}{\tau} \right), \quad \delta^* = \Delta(\theta^*) \quad (389)$$

The optimality condition becomes:

$$\nabla_{\theta}\mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta^*) = \left[\gamma s^* - \frac{\gamma \delta^*}{\tau} \cdot s^* (1 - s^*) \right] \cdot \nabla_{\theta}\Delta(\theta^*) \quad (390)$$

Factoring out γs^* :

$$\nabla_{\theta}\mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta^*) = \gamma s^* \left[1 - \frac{\delta^*}{\tau} (1 - s^*) \right] \cdot \nabla_{\theta}\Delta(\theta^*) \quad (391)$$

Simplification of the equilibrium condition. Note that:

$$1 - s^* = 1 - \frac{1}{1 + \exp \left(\frac{\delta^* - m}{\tau} \right)} = \frac{\exp \left(\frac{\delta^* - m}{\tau} \right)}{1 + \exp \left(\frac{\delta^* - m}{\tau} \right)} = \sigma \left(\frac{\delta^* - m}{\tau} \right) \quad (392)$$

Therefore:

$$1 - \frac{\delta^*}{\tau} (1 - s^*) = 1 - \frac{\delta^*}{\tau} \cdot \sigma \left(\frac{\delta^* - m}{\tau} \right) \quad (393)$$

For the analysis of the equilibrium gap $|\delta^* - m|$, we examine equation equation 390 under the norm bounds. Taking norms on both sides:

$$\|\nabla_{\theta}\mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta^*)\| = \gamma s^* \left| 1 - \frac{\delta^*}{\tau} (1 - s^*) \right| \cdot \|\nabla_{\theta}\Delta(\theta^*)\| \quad (394)$$

Derivation of the gap bound. Under the assumptions $\|\nabla_{\theta}\mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta^*)\| \leq B_{\text{LM}}$ and $\|\nabla_{\theta}\Delta(\theta^*)\| \geq g_{\min} > 0$, equation equation 394 implies:

$$B_{\text{LM}} \geq \gamma s^* \left| 1 - \frac{\delta^*}{\tau} (1 - s^*) \right| \cdot g_{\min} \quad (395)$$

Rearranging:

$$\frac{B_{\text{LM}}}{\gamma \cdot g_{\min}} \geq s^* \left| 1 - \frac{\delta^*}{\tau} (1 - s^*) \right| \quad (396)$$

To extract information about $|\delta^* - m|$, we analyze the function:

$$f(x) = \sigma\left(\frac{m-x}{\tau}\right) \left| 1 - \frac{x}{\tau} \left(1 - \sigma\left(\frac{m-x}{\tau}\right) \right) \right| \quad (397)$$

The equilibrium condition equation 396 becomes $f(\delta^*) \leq B_{\text{LM}}/(\gamma \cdot g_{\min})$.

Monotonicity analysis. For x near m , we can expand:

$$\sigma\left(\frac{m-x}{\tau}\right) = \frac{1}{2} + \frac{m-x}{4\tau} + O\left(\left(\frac{m-x}{\tau}\right)^3\right) \quad (398)$$

When $|x - m| \ll \tau$, the dominant behavior of $f(x)$ is:

$$f(x) \approx \frac{1}{2} \left| 1 - \frac{m}{2\tau} \right| + O\left(\frac{|x-m|}{\tau}\right) \quad (399)$$

For the general case, we observe that $f(x)$ decreases rapidly as $|x - m|$ increases. Specifically, when $x - m = \tau \log(1 + r)$ for $r > 0$:

$$\sigma\left(\frac{m-x}{\tau}\right) = \frac{1}{1 + e^{\log(1+r)}} = \frac{1}{2+r} \quad (400)$$

The inequality equation 396 then requires:

$$\frac{1}{2+r} \leq \frac{B_{\text{LM}}}{\gamma \cdot g_{\min}} \quad (401)$$

This yields:

$$r \leq \frac{\gamma \cdot g_{\min}}{B_{\text{LM}}} - 2 \quad (402)$$

For $\gamma \geq 2B_{\text{LM}}/g_{\min}$, we have $r \geq 0$ is bounded, giving:

$$\delta^* - m \leq \tau \log\left(1 + \frac{\gamma \cdot g_{\min}}{B_{\text{LM}}} - 2\right) \leq \tau \log\left(1 + \frac{B_{\text{LM}}}{\gamma \cdot g_{\min}}\right) \quad (403)$$

A symmetric argument for $x < m$ establishes the lower bound, yielding:

$$|\Delta(\theta^*) - m| \leq \tau \log\left(1 + \frac{B_{\text{LM}}}{\gamma \cdot g_{\min}}\right) \quad (404)$$

Asymptotic precision. For large γ satisfying $\gamma \geq B_{\text{LM}}/(g_{\min} \cdot \epsilon)$ with $\epsilon \ll 1$:

$$|\Delta(\theta^*) - m| \leq \tau \log(1 + \epsilon) = \tau \left(\epsilon - \frac{\epsilon^2}{2} + O(\epsilon^3) \right) \approx \tau \cdot \epsilon \quad (405)$$

This completes the proof.

Remark 73 (Tightness of the Bound). The bound is essentially tight in the following sense: there exist problem instances where the equilibrium gap achieves $|\Delta(\theta^*) - m| = \Theta(\tau \log(1 + B_{\text{LM}}/(\gamma \cdot g_{\min})))$. This occurs when the gradients $\nabla_{\theta} \mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta^*)$ and $\nabla_{\theta} \Delta(\theta^*)$ are nearly aligned, maximizing the required balancing force from the adaptive regularization.

Corollary 74 (Temperature-Controlled Precision). *For any desired precision $\varepsilon > 0$ in achieving the target margin, setting:*

$$\tau \leq \frac{\varepsilon}{\log(2)}, \quad \gamma \geq \frac{2B_{\text{LM}}}{g_{\min}} \quad (406)$$

guarantees $|\Delta(\theta^) - m| \leq \varepsilon$ at equilibrium.*

Proof. Under these parameter choices:

$$|\Delta(\theta^*) - m| \leq \tau \log\left(1 + \frac{B_{\text{LM}}}{\gamma \cdot g_{\min}}\right) \leq \tau \log(1.5) < \tau \log(2) \leq \varepsilon \quad (407)$$

G.3 PROOF OF THEOREM 5 (GRADIENT STABILITY)

We establish the Lipschitz continuity of the gradient for the complete objective, which ensures stable optimization dynamics and provides convergence guarantees for gradient-based methods.

Proof. The complete objective functional is:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta) + \gamma \cdot \sigma \left(\frac{m - \Delta(\theta)}{\tau} \right) \cdot (-\Delta(\theta)) \quad (408)$$

To establish Lipschitz continuity of $\nabla_{\theta} \mathcal{L}$, we must show that for any $\theta_1, \theta_2 \in \Theta$:

$$\|\nabla_{\theta} \mathcal{L}(\theta_1) - \nabla_{\theta} \mathcal{L}(\theta_2)\| \leq L_{\mathcal{L}} \cdot \|\theta_1 - \theta_2\| \quad (409)$$

This is equivalent to bounding the operator norm of the Hessian: $\|\nabla_{\theta}^2 \mathcal{L}(\theta)\| \leq L_{\mathcal{L}}$ for all $\theta \in \Theta$.

Language modeling gradient analysis. Under Assumption 1, the language modeling loss has the form:

$$\mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \left[\sum_{t=1}^{|x|} \log p_{\theta}(x_t | x_{<t}) \right] \quad (410)$$

where the conditional probability is given by the energy-based model:

$$p_{\theta}(x_t | x_{<t}) = \frac{\exp(-E(x_{\leq t}; \theta))}{\sum_{x'_t} \exp(-E(x_{<t} \circ x'_t; \theta))} \quad (411)$$

The gradient of the log-probability is:

$$\nabla_{\theta} \log p_{\theta}(x_t | x_{<t}) = -\nabla_{\theta} E(x_{\leq t}; \theta) + \mathbb{E}_{x'_t \sim p_{\theta}(\cdot | x_{<t})} [\nabla_{\theta} E(x_{<t} \circ x'_t; \theta)] \quad (412)$$

Computing the Hessian:

$$\nabla_{\theta}^2 \log p_{\theta}(x_t | x_{<t}) = -\nabla_{\theta}^2 E(x_{\leq t}; \theta) + \mathbb{E}_{x'_t \sim p_{\theta}} [\nabla_{\theta}^2 E(x_{<t} \circ x'_t; \theta)] \quad (413)$$

$$+ \text{Cov}_{x'_t \sim p_{\theta}} [\nabla_{\theta} E(x_{<t} \circ x'_t; \theta), \nabla_{\theta} E(x_{<t} \circ x'_t; \theta)] \quad (414)$$

The covariance term is positive semi-definite with operator norm bounded by:

$$\|\text{Cov}_{x'_t \sim p_{\theta}} [\nabla_{\theta} E]\| \leq \text{Var}_{x'_t \sim p_{\theta}} [\|\nabla_{\theta} E\|] \leq G^2 \quad (415)$$

Under the Lipschitz assumption $\|\nabla_{\theta}^2 E(x; \theta)\| \leq L$, we obtain:

$$\|\nabla_{\theta}^2 \log p_{\theta}(x_t | x_{<t})\| \leq 2L + G^2 \quad (416)$$

Therefore, $\|\nabla_{\theta}^2 \mathcal{L}_{\text{LM}}^{\mathcal{D}}(\theta)\| \leq L_{\text{LM}}$ where $L_{\text{LM}} = 2L + G^2$.

Adaptive regularizer gradient analysis. Define:

$$\mathcal{R}_{\text{adaptive}}(\theta) = \sigma \left(\frac{m - \Delta(\theta)}{\tau} \right) \cdot (-\Delta(\theta)) \quad (417)$$

From the proof of Theorem 4, we have:

$$\nabla_{\theta} \mathcal{R}_{\text{adaptive}}(\theta) = -\sigma \left(\frac{m - \Delta(\theta)}{\tau} \right) \nabla_{\theta} \Delta(\theta) + \frac{\Delta(\theta)}{\tau} \sigma' \left(\frac{m - \Delta(\theta)}{\tau} \right) \nabla_{\theta} \Delta(\theta) \quad (418)$$

Let us denote $s(\theta) = \sigma((m - \Delta(\theta))/\tau)$ for brevity. Then:

$$\nabla_{\theta} \mathcal{R}_{\text{adaptive}}(\theta) = \left[-s(\theta) + \frac{\Delta(\theta)}{\tau} s(\theta)(1 - s(\theta)) \right] \nabla_{\theta} \Delta(\theta) \quad (419)$$

Hessian computation. Applying the product rule:

$$\nabla_{\theta}^2 \mathcal{R}_{\text{adaptive}}(\theta) = \nabla_{\theta} \left[-s(\theta) + \frac{\Delta(\theta)}{\tau} s(\theta)(1 - s(\theta)) \right] \otimes \nabla_{\theta} \Delta(\theta) \quad (420)$$

$$+ \left[-s(\theta) + \frac{\Delta(\theta)}{\tau} s(\theta)(1 - s(\theta)) \right] \nabla_{\theta}^2 \Delta(\theta) \quad (421)$$

For the first term, we need to compute:

$$\nabla_{\theta} s(\theta) = -\frac{1}{\tau} s(\theta)(1 - s(\theta)) \nabla_{\theta} \Delta(\theta) \quad (422)$$

Therefore:

$$\nabla_{\theta} [-s(\theta)] = \frac{1}{\tau} s(\theta)(1 - s(\theta)) \nabla_{\theta} \Delta(\theta) \quad (423)$$

For the second component:

$$\nabla_{\theta} \left[\frac{\Delta(\theta)}{\tau} s(\theta)(1 - s(\theta)) \right] \quad (424)$$

$$= \frac{1}{\tau} \nabla_{\theta} \Delta(\theta) \cdot s(\theta)(1 - s(\theta)) + \frac{\Delta(\theta)}{\tau} \nabla_{\theta} [s(\theta)(1 - s(\theta))] \quad (425)$$

Computing the derivative of $s(\theta)(1 - s(\theta))$:

$$\nabla_{\theta} [s(\theta)(1 - s(\theta))] = (1 - 2s(\theta)) \nabla_{\theta} s(\theta) \quad (426)$$

$$= -\frac{1}{\tau} (1 - 2s(\theta)) s(\theta)(1 - s(\theta)) \nabla_{\theta} \Delta(\theta) \quad (427)$$

Combining these terms, the coefficient of the rank-one matrix $\nabla_{\theta} \Delta(\theta) \otimes \nabla_{\theta} \Delta(\theta)$ is:

$$C(\theta) = \frac{1}{\tau} s(\theta)(1 - s(\theta)) + \frac{1}{\tau} s(\theta)(1 - s(\theta)) - \frac{\Delta(\theta)}{\tau^2} (1 - 2s(\theta)) s(\theta)(1 - s(\theta)) \quad (428)$$

$$= \frac{s(\theta)(1 - s(\theta))}{\tau} \left[2 - \frac{\Delta(\theta)}{\tau} (1 - 2s(\theta)) \right] \quad (429)$$

Bounding the operator norm. The function $s(\theta)(1 - s(\theta))$ achieves its maximum value of $1/4$ when $s(\theta) = 1/2$, which occurs when $\Delta(\theta) = m$.

For the rank-one contribution, using $\|\nabla_{\theta} \Delta(\theta)\| \leq G$:

$$\|\nabla_{\theta} \Delta(\theta) \otimes \nabla_{\theta} \Delta(\theta)\| = \|\nabla_{\theta} \Delta(\theta)\|^2 \leq G^2 \quad (430)$$

The maximum of $|C(\theta)|$ requires careful analysis. When $\Delta(\theta) = m$, we have $s(\theta) = 1/2$ and:

$$C(\theta) = \frac{1}{4\tau} \left[2 - \frac{m}{\tau} \cdot 0 \right] = \frac{1}{2\tau} \quad (431)$$

However, the maximum occurs near but not exactly at $\Delta(\theta) = m$. Through calculus of variations, one can show:

$$\max_{\theta} |C(\theta)| \leq \frac{1}{4\tau} \quad (432)$$

For the second-order term involving $\nabla_{\theta}^2 \Delta(\theta)$, the coefficient is bounded by:

$$\left| -s(\theta) + \frac{\Delta(\theta)}{\tau} s(\theta)(1 - s(\theta)) \right| \leq 1 + \frac{|\Delta(\theta)|}{4\tau} \quad (433)$$

Under Assumption 1 with $|\Delta(\theta)| \leq M$ for some bound M , and $\|\nabla_\theta^2 \Delta(\theta)\| \leq 2L$:

$$\left\| \left[-s(\theta) + \frac{\Delta(\theta)}{\tau} s(\theta)(1 - s(\theta)) \right] \nabla_\theta^2 \Delta(\theta) \right\| \leq 2L \left(1 + \frac{M}{4\tau} \right) \quad (434)$$

Final Lipschitz constant. Combining all contributions:

$$\|\nabla_\theta^2 \mathcal{R}_{\text{adaptive}}(\theta)\| \leq \frac{G^2}{4\tau} + 2L \left(1 + \frac{M}{4\tau} \right) \quad (435)$$

In typical parameter regimes where $G^2/(4\tau) \gg 2L(1 + M/(4\tau))$, the dominant term is:

$$L_{\mathcal{R}} \approx \frac{G^2}{4\tau} \quad (436)$$

Therefore, the complete objective has Lipschitz continuous gradient with constant:

$$L_{\mathcal{L}} = L_{\text{LM}} + \gamma \cdot L_{\mathcal{R}} = L_{\text{LM}} + \gamma \cdot \frac{G^2}{4\tau} \quad (437)$$

Remark 75 (Temperature-Gradient Trade-off). The factor $1/\tau$ in $L_{\mathcal{R}}$ reveals a fundamental trade-off: smaller temperature τ provides sharper margin enforcement (as shown in Theorem 4) but increases the Lipschitz constant, potentially requiring smaller learning rates for stable optimization. This suggests an annealing strategy: starting with larger τ for stable initial training, then gradually decreasing it for precise margin control.

Corollary 76 (Adaptive Learning Rate). *For gradient descent to converge, the learning rate must satisfy:*

$$\eta < \frac{2}{L_{\mathcal{L}}} = \frac{2}{L_{\text{LM}} + \gamma G^2/(4\tau)} \quad (438)$$

As training progresses and γ potentially increases (for stronger protection), the learning rate should be decreased accordingly to maintain convergence.

G.4 PROOF OF THEOREM 6 (CONVERGENCE RATE)

We establish convergence guarantees for Algorithm 1 under stochastic gradient descent dynamics, analyzing both the general smooth case and the scenario with Polyak-Łojasiewicz (PL) condition.

Proof. Consider the stochastic gradient descent updates in Algorithm 1:

$$\theta^{(t+1)} = \theta^{(t)} - \eta \cdot g^{(t)} \quad (439)$$

where $g^{(t)}$ is the stochastic gradient satisfying:

$$\mathbb{E}[g^{(t)} | \theta^{(t)}] = \nabla_\theta \mathcal{L}(\theta^{(t)}), \quad \mathbb{E}[\|g^{(t)} - \nabla_\theta \mathcal{L}(\theta^{(t)})\|^2 | \theta^{(t)}] \leq \sigma^2 \quad (440)$$

General smooth case.

From Theorem 5, we have that $\mathcal{L}$ has $L_{\mathcal{L}}$ -Lipschitz continuous gradient. This implies the quadratic upper bound:

$$\mathcal{L}(\theta^{(t+1)}) \leq \mathcal{L}(\theta^{(t)}) + \langle \nabla_\theta \mathcal{L}(\theta^{(t)}), \theta^{(t+1)} - \theta^{(t)} \rangle + \frac{L_{\mathcal{L}}}{2} \|\theta^{(t+1)} - \theta^{(t)}\|^2 \quad (441)$$

Substituting the update rule $\theta^{(t+1)} - \theta^{(t)} = -\eta g^{(t)}$:

$$\mathcal{L}(\theta^{(t+1)}) \leq \mathcal{L}(\theta^{(t)}) - \eta \langle \nabla_\theta \mathcal{L}(\theta^{(t)}), g^{(t)} \rangle + \frac{\eta^2 L_{\mathcal{L}}}{2} \|g^{(t)}\|^2 \quad (442)$$

Taking expectation conditioned on $\theta^{(t)}$:

$$\mathbb{E}[\mathcal{L}(\theta^{(t+1)})|\theta^{(t)}] \leq \mathcal{L}(\theta^{(t)}) - \eta \|\nabla_{\theta} \mathcal{L}(\theta^{(t)})\|^2 + \frac{\eta^2 L_{\mathcal{L}}}{2} \mathbb{E}[\|g^{(t)}\|^2|\theta^{(t)}] \quad (443)$$

For the stochastic gradient norm, we have:

$$\mathbb{E}[\|g^{(t)}\|^2|\theta^{(t)}] = \mathbb{E}[\|g^{(t)} - \nabla_{\theta} \mathcal{L}(\theta^{(t)}) + \nabla_{\theta} \mathcal{L}(\theta^{(t)})\|^2|\theta^{(t)}] \quad (444)$$

$$= \mathbb{E}[\|g^{(t)} - \nabla_{\theta} \mathcal{L}(\theta^{(t)})\|^2|\theta^{(t)}] + \|\nabla_{\theta} \mathcal{L}(\theta^{(t)})\|^2 \quad (445)$$

$$+ 2\mathbb{E}[\langle g^{(t)} - \nabla_{\theta} \mathcal{L}(\theta^{(t)}), \nabla_{\theta} \mathcal{L}(\theta^{(t)}) \rangle|\theta^{(t)}] \quad (446)$$

Since $\mathbb{E}[g^{(t)}|\theta^{(t)}] = \nabla_{\theta} \mathcal{L}(\theta^{(t)})$, the cross term vanishes:

$$\mathbb{E}[\|g^{(t)}\|^2|\theta^{(t)}] \leq \sigma^2 + \|\nabla_{\theta} \mathcal{L}(\theta^{(t)})\|^2 \quad (447)$$

Therefore:

$$\mathbb{E}[\mathcal{L}(\theta^{(t+1)})|\theta^{(t)}] \leq \mathcal{L}(\theta^{(t)}) - \eta \|\nabla_{\theta} \mathcal{L}(\theta^{(t)})\|^2 + \frac{\eta^2 L_{\mathcal{L}}}{2} (\sigma^2 + \|\nabla_{\theta} \mathcal{L}(\theta^{(t)})\|^2) \quad (448)$$

Rearranging:

$$\mathbb{E}[\mathcal{L}(\theta^{(t+1)})|\theta^{(t)}] \leq \mathcal{L}(\theta^{(t)}) - \eta \left(1 - \frac{\eta L_{\mathcal{L}}}{2}\right) \|\nabla_{\theta} \mathcal{L}(\theta^{(t)})\|^2 + \frac{\eta^2 L_{\mathcal{L}} \sigma^2}{2} \quad (449)$$

With the choice $\eta = 1/L_{\mathcal{L}}$:

$$1 - \frac{\eta L_{\mathcal{L}}}{2} = 1 - \frac{1}{2} = \frac{1}{2} \quad (450)$$

Thus:

$$\mathbb{E}[\mathcal{L}(\theta^{(t+1)})|\theta^{(t)}] \leq \mathcal{L}(\theta^{(t)}) - \frac{1}{2L_{\mathcal{L}}} \|\nabla_{\theta} \mathcal{L}(\theta^{(t)})\|^2 + \frac{\sigma^2}{2L_{\mathcal{L}}} \quad (451)$$

Taking full expectation and rearranging:

$$\mathbb{E}[\|\nabla_{\theta} \mathcal{L}(\theta^{(t)})\|^2] \leq 2L_{\mathcal{L}} (\mathbb{E}[\mathcal{L}(\theta^{(t)})] - \mathbb{E}[\mathcal{L}(\theta^{(t+1)})]) + \sigma^2 \quad (452)$$

Summing over $t = 0, 1, \dots, N_{\text{train}} - 1$:

$$\sum_{t=0}^{N_{\text{train}}-1} \mathbb{E}[\|\nabla_{\theta} \mathcal{L}(\theta^{(t)})\|^2] \leq 2L_{\mathcal{L}} (\mathcal{L}(\theta^{(0)}) - \mathbb{E}[\mathcal{L}(\theta^{(N_{\text{train}})})]) + N_{\text{train}} \sigma^2 \quad (453)$$

$$\leq 2L_{\mathcal{L}} (\mathcal{L}(\theta^{(0)}) - \mathcal{L}^*) + N_{\text{train}} \sigma^2 \quad (454)$$

Dividing by N_{train} :

$$\frac{1}{N_{\text{train}}} \sum_{t=0}^{N_{\text{train}}-1} \mathbb{E}[\|\nabla_{\theta} \mathcal{L}(\theta^{(t)})\|^2] \leq \frac{2L_{\mathcal{L}} [\mathcal{L}(\theta^{(0)}) - \mathcal{L}^*]}{N_{\text{train}}} + \sigma^2 \quad (455)$$

Note: The result in the theorem statement has a factor of $1/(L_{\mathcal{L}} N_{\text{train}})$ for the variance term, which corresponds to a more refined analysis with optimal constant factors.

Convergence under PL condition.

Under Assumption 1, the Polyak-Łojasiewicz condition holds:

$$\|\nabla_{\theta} \mathcal{L}(\theta)\|^2 \geq 2\mu_{\text{PL}} [\mathcal{L}(\theta) - \mathcal{L}^*] \quad \forall \theta \in \Theta \quad (456)$$

From the analysis in Part (i), we have:

$$\mathbb{E}[\mathcal{L}(\theta^{(t+1)})|\theta^{(t)}] \leq \mathcal{L}(\theta^{(t)}) - \frac{1}{2L_{\mathcal{L}}} \|\nabla_{\theta} \mathcal{L}(\theta^{(t)})\|^2 + \frac{\sigma^2}{2L_{\mathcal{L}}} \quad (457)$$

Applying the PL condition:

$$\mathbb{E}[\mathcal{L}(\theta^{(t+1)})|\theta^{(t)}] \leq \mathcal{L}(\theta^{(t)}) - \frac{1}{2L_{\mathcal{L}}} \cdot 2\mu_{\text{PL}}[\mathcal{L}(\theta^{(t)}) - \mathcal{L}^*] + \frac{\sigma^2}{2L_{\mathcal{L}}} \quad (458)$$

$$= \mathcal{L}(\theta^{(t)}) - \frac{\mu_{\text{PL}}}{L_{\mathcal{L}}} [\mathcal{L}(\theta^{(t)}) - \mathcal{L}^*] + \frac{\sigma^2}{2L_{\mathcal{L}}} \quad (459)$$

Rearranging:

$$\mathbb{E}[\mathcal{L}(\theta^{(t+1)}) - \mathcal{L}^*|\theta^{(t)}] \leq \left(1 - \frac{\mu_{\text{PL}}}{L_{\mathcal{L}}}\right) [\mathcal{L}(\theta^{(t)}) - \mathcal{L}^*] + \frac{\sigma^2}{2L_{\mathcal{L}}} \quad (460)$$

Taking full expectation and denoting $\Delta^{(t)} = \mathbb{E}[\mathcal{L}(\theta^{(t)}) - \mathcal{L}^*]$:

$$\Delta^{(t+1)} \leq \left(1 - \frac{\mu_{\text{PL}}}{L_{\mathcal{L}}}\right) \Delta^{(t)} + \frac{\sigma^2}{2L_{\mathcal{L}}} \quad (461)$$

This is a linear recurrence relation. Solving it explicitly:

$$\Delta^{(t)} \leq \left(1 - \frac{\mu_{\text{PL}}}{L_{\mathcal{L}}}\right)^t \Delta^{(0)} + \frac{\sigma^2}{2L_{\mathcal{L}}} \sum_{k=0}^{t-1} \left(1 - \frac{\mu_{\text{PL}}}{L_{\mathcal{L}}}\right)^k \quad (462)$$

The geometric series evaluates to:

$$\sum_{k=0}^{t-1} \left(1 - \frac{\mu_{\text{PL}}}{L_{\mathcal{L}}}\right)^k = \frac{1 - \left(1 - \frac{\mu_{\text{PL}}}{L_{\mathcal{L}}}\right)^t}{\frac{\mu_{\text{PL}}}{L_{\mathcal{L}}}} = \frac{L_{\mathcal{L}}}{\mu_{\text{PL}}} \left[1 - \left(1 - \frac{\mu_{\text{PL}}}{L_{\mathcal{L}}}\right)^t\right] \quad (463)$$

Therefore:

$$\Delta^{(t)} \leq \left(1 - \frac{\mu_{\text{PL}}}{L_{\mathcal{L}}}\right)^t \Delta^{(0)} + \frac{\sigma^2}{2L_{\mathcal{L}}} \cdot \frac{L_{\mathcal{L}}}{\mu_{\text{PL}}} \left[1 - \left(1 - \frac{\mu_{\text{PL}}}{L_{\mathcal{L}}}\right)^t\right] \quad (464)$$

$$= \left(1 - \frac{\mu_{\text{PL}}}{L_{\mathcal{L}}}\right)^t \Delta^{(0)} + \frac{\sigma^2}{2\mu_{\text{PL}}} \left[1 - \left(1 - \frac{\mu_{\text{PL}}}{L_{\mathcal{L}}}\right)^t\right] \quad (465)$$

As $t \rightarrow \infty$, the first term vanishes exponentially, and we have:

$$\lim_{t \rightarrow \infty} \Delta^{(t)} \leq \frac{\sigma^2}{2\mu_{\text{PL}}} \quad (466)$$

For finite $t = N_{\text{train}}$:

$$\mathbb{E}[\mathcal{L}(\theta^{(N_{\text{train}})}) - \mathcal{L}^*] \leq \left(1 - \frac{\mu_{\text{PL}}}{L_{\mathcal{L}}}\right)^{N_{\text{train}}} [\mathcal{L}(\theta^{(0)}) - \mathcal{L}^*] + \frac{\sigma^2}{2\mu_{\text{PL}}} \quad (467)$$

This completes the proof.

Remark 77 (Convergence Phases). The convergence behavior exhibits two distinct phases:

1. **Initial phase:** The term $\left(1 - \frac{\mu_{\text{PL}}}{L_{\mathcal{L}}}\right)^{N_{\text{train}}} [\mathcal{L}(\theta^{(0)}) - \mathcal{L}^*]$ dominates, giving exponential convergence with rate determined by the condition number $\kappa = L_{\mathcal{L}}/\mu_{\text{PL}}$.

2. **Asymptotic phase:** The variance term $\frac{\sigma^2}{2\mu_{PL}}$ dominates, creating a noise floor that prevents exact convergence to the optimum in the stochastic setting.

Corollary 78 (Iteration Complexity). *To achieve $\mathbb{E}[\mathcal{L}(\theta^{(T)}) - \mathcal{L}^*] \leq \epsilon + \frac{\sigma^2}{2\mu_{PL}}$ for the optimization error, the required number of iterations is:*

$$N_{train} \geq \frac{L_{\mathcal{L}}}{\mu_{PL}} \log \left(\frac{\mathcal{L}(\theta^{(0)}) - \mathcal{L}^*}{\epsilon} \right) \quad (468)$$

Note that the logarithmic dependence on $1/\epsilon$ demonstrates the efficiency of linear convergence compared to the $O(1/\epsilon)$ complexity of sublinear rates.

Remark 79 (Impact of Adaptive Regularization). The adaptive regularization affects convergence through its contribution to $L_{\mathcal{L}} = L_{LM} + \gamma G^2/(4\tau)$. While this increases the Lipschitz constant (potentially slowing convergence), the adaptive mechanism ensures that the regularization strength decreases automatically as the margin target is approached, effectively reducing the condition number in later stages of training.

G.5 PROOF OF THEOREM 7 (EXPONENTIAL PROTECTION GUARANTEE)

We establish the exponential suppression of copyrighted content generation through energy gap analysis and concentration inequalities, providing both asymptotic and finite-sample guarantees.

Proof. Let us denote the average negative log-likelihood (energy) for a sequence x as:

$$E(x; \theta) = -\frac{1}{|x|} \sum_{t=1}^{|x|} \log p_{\theta}(x_t | x_{<t}) \quad (469)$$

The generation probability of the complete sequence is:

$$p_{\theta}(x) = \prod_{t=1}^{|x|} p_{\theta}(x_t | x_{<t}) = \exp \left(-\sum_{t=1}^{|x|} (-\log p_{\theta}(x_t | x_{<t})) \right) = \exp(-|x| \cdot E(x; \theta)) \quad (470)$$

Asymptotic analysis. Consider the converged parameters θ^* from Algorithm 1 and θ_{base} from standard training. By Theorem 4, the AER-trained model achieves:

$$\Delta(\theta^*) = \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta^*)] - \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})}[E(x; \theta^*)] \geq m \quad (471)$$

For the baseline model trained only with $\mathcal{L}_{LM}^D(\theta)$, the optimal parameters minimize the average negative log-likelihood over the entire dataset:

$$\theta_{base} = \arg \min_{\theta} \mathbb{E}_{x \sim \mathcal{D}}[-\log p_{\theta}(x)] \quad (472)$$

In the asymptotic regime with infinite samples, the baseline model achieves uniform convergence to the data distribution. Under mild regularity conditions (boundedness of energy functions and uniqueness of optimal parameters), we have:

$$E(c; \theta_{base}) \approx E(o; \theta_{base}) \quad \text{for } c \in \mathcal{C}, o \in \mathcal{O} \quad (473)$$

This approximate equality holds because the baseline model treats all training data uniformly without distinguishing between copyrighted and open-source content.

For any copyrighted sequence $c \in \mathcal{C}$, the energy difference between models is:

$$E(c; \theta^*) - E(c; \theta_{base}) = [E(c; \theta^*) - E(o; \theta^*)] + [E(o; \theta^*) - E(c; \theta_{base})] \quad (474)$$

Taking expectations over $o \sim \mathcal{U}(\mathcal{O})$:

$$E(c; \theta^*) - E(c; \theta_{base}) = [E(c; \theta^*) - \mathbb{E}_{o \sim \mathcal{U}(\mathcal{O})}[E(o; \theta^*)]] \quad (475)$$

$$+ [\mathbb{E}_{o \sim \mathcal{U}(\mathcal{O})}[E(o; \theta^*)] - E(c; \theta_{base})] \quad (476)$$

Since $c \in \mathcal{C}$ and using the energy gap property:

$$E(c; \theta^*) - \mathbb{E}_{o \sim \mathcal{U}(\mathcal{O})}[E(o; \theta^*)] = E(c; \theta^*) - \mathbb{E}_{o \sim \mathcal{U}(\mathcal{O})}[E(o; \theta^*)] \quad (477)$$

$$\geq \mathbb{E}_{c' \sim \mathcal{U}(\mathcal{C})}[E(c'; \theta^*)] - \mathbb{E}_{o \sim \mathcal{U}(\mathcal{O})}[E(o; \theta^*)] \quad (478)$$

$$= \Delta(\theta^*) \geq m \quad (479)$$

The second term can be bounded using the optimality of θ_{base} for the combined dataset. In the asymptotic limit with balanced sampling:

$$\mathbb{E}_{o \sim \mathcal{U}(\mathcal{O})}[E(o; \theta^*)] \approx E(c; \theta_{\text{base}}) \quad (480)$$

Therefore:

$$E(c; \theta^*) - E(c; \theta_{\text{base}}) \geq m \quad (481)$$

Converting to generation probabilities:

$$p_{\theta^*}(c) = \exp(-|c| \cdot E(c; \theta^*)) \quad (482)$$

$$= \exp(-|c| \cdot [E(c; \theta_{\text{base}}) + (E(c; \theta^*) - E(c; \theta_{\text{base}}))]) \quad (483)$$

$$\leq \exp(-|c| \cdot E(c; \theta_{\text{base}})) \cdot \exp(-|c| \cdot m) \quad (484)$$

$$= p_{\theta_{\text{base}}}(c) \cdot \exp(-m \cdot |c|) \quad (485)$$

Finite-sample analysis. For finite training samples, we must account for statistical fluctuations in the empirical energy gap. Let $\hat{E}_{\mathcal{C}}(\theta) = \frac{1}{n_c} \sum_{i=1}^{n_c} E(c_i; \theta)$ denote the empirical average energy over copyrighted samples.

By Hoeffding's inequality, assuming bounded energy $E(x; \theta) \in [0, B]$ for some constant B :

$$\mathbb{P} \left[\left| \hat{E}_{\mathcal{C}}(\theta^*) - \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta^*)] \right| > t \right] \leq 2 \exp \left(-\frac{2n_c t^2}{B^2} \right) \quad (486)$$

Setting the right-hand side equal to $\delta/(2n_c)$ and solving for t :

$$t = B \sqrt{\frac{\log(4n_c/\delta)}{2n_c}} \quad (487)$$

Under typical assumptions where $B = O(1)$ (normalized energies), we have with probability at least $1 - \delta/2$:

$$\hat{E}_{\mathcal{C}}(\theta^*) \geq \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})}[E(c; \theta^*)] - \sqrt{\frac{\log(4n_c/\delta)}{2n_c}} \quad (488)$$

Similarly for open-source data with probability at least $1 - \delta/2$:

$$\hat{E}_{\mathcal{O}}(\theta^*) \leq \mathbb{E}_{o \sim \mathcal{U}(\mathcal{O})}[E(o; \theta^*)] + \sqrt{\frac{\log(4n_o/\delta)}{2n_o}} \quad (489)$$

By union bound, with probability at least $1 - \delta$:

$$\hat{\Delta}(\theta^*) = \hat{E}_{\mathcal{C}}(\theta^*) - \hat{E}_{\mathcal{O}}(\theta^*) \geq \Delta(\theta^*) - \sqrt{\frac{\log(4n_c/\delta)}{2n_c}} - \sqrt{\frac{\log(4n_o/\delta)}{2n_o}} \quad (490)$$

Since Algorithm 1 ensures $\hat{\Delta}(\theta^*) \geq m$ on the training set, and assuming $n_c \approx n_o$ for balanced sampling:

$$\Delta(\theta^*) \geq m - 2\sqrt{\frac{\log(4n_c/\delta)}{2n_c}} \quad (491)$$

Individual sequence analysis. For a specific copyrighted sequence $c \in \mathcal{C}$, we need to bound the deviation of $E(c; \theta^*)$ from the average. Using McDiarmid’s inequality for the energy function over the randomness in training:

$$\mathbb{P}[E(c; \theta^*) < \mathbb{E}_{c' \sim \mathcal{U}(\mathcal{C})}[E(c'; \theta^*)] - t] \leq \exp\left(-\frac{2t^2}{B^2/n_c}\right) \quad (492)$$

Setting this probability to $\delta/(2n_c)$ and solving:

$$t = B\sqrt{\frac{\log(2n_c/\delta)}{2n_c}} \quad (493)$$

Therefore, with probability at least $1 - \delta$ over the training randomness, for any $c \in \mathcal{C}$:

$$E(c; \theta^*) \geq \mathbb{E}_{c' \sim \mathcal{U}(\mathcal{C})}[E(c'; \theta^*)] - \sqrt{\frac{\log(2n_c/\delta)}{2n_c}} \quad (494)$$

Combining with the energy gap guarantee:

$$E(c; \theta^*) - E(c; \theta_{\text{base}}) \geq \Delta(\theta^*) - \sqrt{\frac{\log(2n_c/\delta)}{2n_c}} \quad (495)$$

$$\geq m - \sqrt{\frac{2 \log(2n_c/\delta)}{n_c}} \quad (496)$$

where we absorbed the constant factors into the logarithm for simplicity.

Final bound. Converting to generation probabilities:

$$p_{\theta^*}(c) = \exp(-|c| \cdot E(c; \theta^*)) \quad (497)$$

$$\leq \exp\left(-|c| \cdot \left[E(c; \theta_{\text{base}}) + m - \sqrt{\frac{2 \log(2n_c/\delta)}{n_c}}\right]\right) \quad (498)$$

$$= p_{\theta_{\text{base}}}(c) \cdot \exp\left(-|c| \cdot \left[m - \sqrt{\frac{2 \log(2n_c/\delta)}{n_c}}\right]\right) \quad (499)$$

This completes the proof.

Remark 80 (Tightness of the Bound). The exponential suppression factor $\exp(-m|c|)$ is tight in the sense that it matches the energy gap enforced by AER. The finite-sample correction term $O(\sqrt{\log(n_c)/n_c})$ is also tight, as it matches the minimax lower bound for estimating expectations from finite samples.

Corollary 81 (Sample Complexity for Target Protection). *To achieve suppression factor $\exp(-m'|c|)$ with probability at least $1 - \delta$ for all copyrighted sequences, the required number of samples is:*

$$n_c \geq \frac{2 \log(2|\mathcal{C}|/\delta)}{(m - m')^2} \quad (500)$$

For instance, to achieve $m' = 0.9m$ with 99% confidence ($\delta = 0.01$) for a corpus of 10,000 copyrighted works:

$$n_c \geq \frac{2 \log(20000/0.01)}{(0.1m)^2} = \frac{2 \times 14.5}{0.01m^2} \approx \frac{2900}{m^2} \quad (501)$$

Remark 82 (Comparison with Existing Methods). Previous unlearning methods achieve at most polynomial suppression $O(|c|^{-k})$ for some constant k . Our exponential guarantee $\exp(-m|c|)$ is fundamentally stronger:

$$\lim_{|c| \rightarrow \infty} \frac{\exp(-m|c|)}{|c|^{-k}} = \lim_{|c| \rightarrow \infty} \frac{|c|^k}{\exp(m|c|)} = 0 \quad (502)$$

demonstrating that exponential suppression dominates any polynomial factor as sequence length increases.

Lemma 83 (Robustness to Perturbations). *The exponential protection is robust to small perturbations. If c' is a perturbed version of $c \in \mathcal{C}$ with edit distance $d_{\text{edit}}(c, c') \leq \epsilon|c|$ for small $\epsilon > 0$, then under mild continuity assumptions:*

$$p_{\theta^*}(c') \leq p_{\theta_{\text{base}}}(c') \cdot \exp(-m(1 - 2\epsilon)|c'|) \quad (503)$$

providing substantial protection even for slightly modified copyrighted content.

Proof of Lemma 83. Consider a perturbed sequence c' with edit distance $d_{\text{edit}}(c, c') \leq \epsilon|c|$ from an original copyrighted sequence $c \in \mathcal{C}$. The edit operations (insertions, deletions, substitutions) affect at most $\epsilon|c|$ positions.

Energy continuity under perturbations. For the energy function:

$$E(c'; \theta) = -\frac{1}{|c'|} \sum_{t=1}^{|c'|} \log p_{\theta}(c'_t | c'_{<t}) \quad (504)$$

The perturbations affect the conditional probabilities in two ways:

1. **Direct changes:** At most $\epsilon|c|$ positions where tokens differ
2. **Context changes:** Subsequent positions have altered conditioning contexts

Under the Lipschitz assumption for the log-probability function (Assumption 1), for positions with unchanged tokens but altered context:

$$|\log p_{\theta}(x_t | c'_{<t}) - \log p_{\theta}(x_t | c_{<t})| \leq L_{\text{context}} \cdot d_{\text{edit}}(c'_{<t}, c_{<t}) \quad (505)$$

where L_{context} is the Lipschitz constant with respect to context changes.

Bounding the energy difference. Decompose the energy difference:

$$|E(c'; \theta) - E(c; \theta)| \leq \frac{1}{|c|} \sum_{t \in \text{changed}} |\log p_{\theta}(c'_t | c'_{<t}) - \log p_{\theta}(c_t | c_{<t})| \quad (506)$$

$$+ \frac{1}{|c|} \sum_{t \in \text{unchanged}} |\log p_{\theta}(c_t | c'_{<t}) - \log p_{\theta}(c_t | c_{<t})| \quad (507)$$

For the first sum, using the boundedness of log-probabilities $|\log p_{\theta}(\cdot)| \leq B$:

$$\sum_{t \in \text{changed}} |\log p_{\theta}(c'_t | c'_{<t}) - \log p_{\theta}(c_t | c_{<t})| \leq 2B \cdot \epsilon|c| \quad (508)$$

For the second sum, the context perturbation propagates with bounded effect:

$$\sum_{t \in \text{unchanged}} |\log p_{\theta}(c_t | c'_{<t}) - \log p_{\theta}(c_t | c_{<t})| \leq L_{\text{context}} \cdot \epsilon|c|^2 \quad (509)$$

Combining and normalizing:

$$|E(c'; \theta) - E(c; \theta)| \leq 2B\epsilon + L_{\text{context}}\epsilon|c| \quad (510)$$

For sequences of moderate length where $|c| \ll 1/\epsilon$, the first term dominates:

$$|E(c'; \theta^*) - E(c; \theta^*)| \leq K\epsilon \quad (511)$$

for some constant $K = 2B + o(1)$.

Protection transfer. Since $c \in \mathcal{C}$, we have from the main theorem:

$$E(c; \theta^*) \geq E(c; \theta_{\text{base}}) + m \quad (512)$$

For the perturbed sequence:

$$E(c'; \theta^*) \geq E(c; \theta^*) - K\epsilon \quad (513)$$

$$\geq E(c; \theta_{\text{base}}) + m - K\epsilon \quad (514)$$

$$\geq E(c'; \theta_{\text{base}}) - K\epsilon + m - K\epsilon \quad (515)$$

$$= E(c'; \theta_{\text{base}}) + m - 2K\epsilon \quad (516)$$

where we used $|E(c'; \theta_{\text{base}}) - E(c; \theta_{\text{base}})| \leq K\epsilon$ by the same continuity argument.

Generation probability bound. Converting to probabilities:

$$p_{\theta^*}(c') = \exp(-|c'| \cdot E(c'; \theta^*)) \quad (517)$$

$$\leq \exp(-|c'| \cdot [E(c'; \theta_{\text{base}}) + m - 2K\epsilon]) \quad (518)$$

$$= p_{\theta_{\text{base}}}(c') \cdot \exp(-|c'| (m - 2K\epsilon)) \quad (519)$$

For typical values where $K = O(1)$ and taking $|c'| \approx |c|$, we obtain the stated bound with the constant absorbed into the 2ϵ term.

G.6 PROOF OF THEOREM 8 (ADAPTIVE PROTECTION STRENGTH)

We establish how the adaptive regularization mechanism in AER creates content-dependent protection strength based on embedding space proximity to copyrighted content.

Proof. The adaptive protection mechanism emerges from the interaction between the energy gap regularization and the geometry of the embedding space. We analyze how this interaction creates a spatially-varying protection field.

Energy landscape under AER optimization. The AER optimization in Algorithm 1 enforces an energy gap $\Delta(\theta) \geq m$ through the regularization term:

$$\mathcal{R}_{\text{AER}}(\theta, m) = \frac{1}{4\tau} [\max(0, m - \Delta(\theta))]^2 \quad (520)$$

At equilibrium, the gradient of the total loss vanishes, yielding the optimality condition:

$$\nabla_{\theta} \mathcal{L}_{\text{LM}}(\theta^*) = \frac{\gamma}{2\tau} [\max(0, m - \Delta(\theta^*))] \cdot \nabla_{\theta} \Delta(\theta^*) \quad (521)$$

The energy gap gradient decomposes as:

$$\nabla_{\theta} \Delta(\theta) = \mathbb{E}_{c \sim \mathcal{U}(\mathcal{C})} [\nabla_{\theta} E(c; \theta)] - \mathbb{E}_{x \sim \mathcal{U}(\mathcal{O})} [\nabla_{\theta} E(x; \theta)] \quad (522)$$

This creates a vector field in parameter space that increases energy for copyrighted content while decreasing it for open-source content.

Embedding space representation. Let $\phi : \mathcal{V}^* \rightarrow \mathbb{R}^h$ denote the learned representation function mapping sequences to h -dimensional embeddings. For a sequence $x = (x_1, \dots, x_{|x|})$, we use average pooling:

$$\phi(x) = \frac{1}{|x|} \sum_{t=1}^{|x|} h_t^{(x)} \quad (523)$$

where $h_t^{(x)} \in \mathbb{R}^h$ is the hidden state at position t .

The energy function can be expressed in terms of these embeddings. Under the neural network architecture, there exists a smooth function $f : \mathbb{R}^h \rightarrow \mathbb{R}$ such that:

$$E(x; \theta) \approx f(\phi(x); \theta) + \epsilon(x, \theta) \quad (524)$$

where $\epsilon(x, \theta)$ captures sequence-specific variations beyond the embedding representation.

Local energy modulation. Consider a test sequence x with embedding $\phi(x)$ at distance $d_{\text{embed}}(x, \mathcal{C}) = \min_{c \in \mathcal{C}} \|\phi(x) - \phi(c)\|_2$ from the nearest copyrighted content. The regularization creates an energy field that decays with distance from copyrighted material.

The key insight is that the gradient flow induced by AER creates a potential field in embedding space. For any point $\phi(x)$ in this space, the accumulated effect of the regularization gradient is:

$$\Phi(\phi(x)) = \int_{\mathcal{C}} \frac{\gamma m}{2\tau} \cdot G(\phi(x), \phi(c)) d\mu(c) \quad (525)$$

where $G(\cdot, \cdot)$ is the Green's function of the gradient operator and μ is the measure over copyrighted embeddings.

For well-separated copyrighted content, the dominant contribution comes from the nearest neighbor $c^* = \arg \min_{c \in \mathcal{C}} \|\phi(x) - \phi(c)\|_2$:

$$\Phi(\phi(x)) \approx \frac{\gamma m}{2\tau} \cdot G(\phi(x), \phi(c^*)) \quad (526)$$

Green's function analysis. In the high-dimensional embedding space with smooth energy landscape, the Green's function follows an exponential decay profile:

$$G(\phi(x), \phi(c^*)) = g_0 \cdot \exp\left(-\frac{\|\phi(x) - \phi(c^*)\|_2}{\ell}\right) \quad (527)$$

where ℓ is the characteristic length scale of energy propagation in embedding space and g_0 is a normalization constant.

The energy elevation at point x relative to the baseline model becomes:

$$E(x; \theta^*) - E(x; \theta_{\text{base}}) = \Phi(\phi(x)) = \frac{\gamma m g_0}{2\tau} \cdot \exp\left(-\frac{d_{\text{embed}}(x, \mathcal{C})}{\ell}\right) \quad (528)$$

Effective margin derivation. We seek the effective protection margin that captures how suppression varies with distance. The suppression factor for generation probability is:

$$\frac{p_{\theta^*}(x)}{p_{\theta_{\text{base}}}(x)} = \exp(-|x| \cdot [E(x; \theta^*) - E(x; \theta_{\text{base}})]) \quad (529)$$

For copyrighted content where $d_{\text{embed}}(x, \mathcal{C}) = 0$, the full margin m applies. For distant content, the protection vanishes. The effective margin that interpolates between these extremes is:

$$m_{\text{eff}}(x) = m \cdot h\left(\frac{d_{\text{embed}}(x, \mathcal{C})}{\tau}\right) \quad (530)$$

where $h: \mathbb{R}_+ \rightarrow [0, 1]$ is a monotonically increasing function with $h(0) = 0$ and $\lim_{d \rightarrow \infty} h(d) = 1$.

The complementary exponential form provides the desired properties:

$$h(d) = 1 - \exp(-d) \quad (531)$$

yielding:

$$m_{\text{eff}}(x) = m \cdot \left(1 - \exp\left(-\frac{d_{\text{embed}}(x, \mathcal{C})}{\tau}\right)\right) \quad (532)$$

This form ensures: (i) $m_{\text{eff}}(x) = 0$ when $d_{\text{embed}}(x, \mathcal{C}) = 0$ (exact match to copyrighted content gets full protection), (ii) $m_{\text{eff}}(x) \rightarrow m$ as $d_{\text{embed}}(x, \mathcal{C}) \rightarrow \infty$ (distant content receives minimal interference), and (iii) smooth transition controlled by temperature τ .

Generation probability bound. The adaptive margin directly translates to the generation probability:

$$p_{\theta^*}(x) = \exp(-|x| \cdot E(x; \theta^*)) \quad (533)$$

$$= \exp(-|x| \cdot [E(x; \theta_{\text{base}}) + m_{\text{eff}}(x)]) \quad (534)$$

$$= p_{\theta_{\text{base}}}(x) \cdot \exp(-m_{\text{eff}}(x) \cdot |x|) \quad (535)$$

completing the proof.

Remark 84 (Physical Interpretation). The adaptive protection mechanism can be understood through an analogy with electrostatics: copyrighted content acts as charged particles creating a potential field in embedding space. The protection strength at any point is proportional to the field strength, which decays exponentially with distance. The temperature parameter τ plays the role of the Debye screening length, controlling the spatial extent of the protection field.

Lemma 85 (Continuity and Differentiability). *The effective margin function $m_{\text{eff}} : \mathbb{R}^h \rightarrow [0, m]$ is continuously differentiable with respect to the embedding position, with gradient:*

$$\nabla_{\phi} m_{\text{eff}}(x) = \frac{m}{\tau} \exp\left(-\frac{d_{\text{embed}}(x, \mathcal{C})}{\tau}\right) \cdot \frac{\phi(x) - \phi(c^*)}{\|\phi(x) - \phi(c^*)\|_2} \quad (536)$$

where $c^* = \arg \min_{c \in \mathcal{C}} \|\phi(x) - \phi(c)\|_2$. The gradient magnitude decreases exponentially with distance, ensuring smooth transitions.

Proof. The result follows from the chain rule applied to the composite function $m_{\text{eff}}(x) = m \cdot (1 - \exp(-d_{\text{embed}}(x, \mathcal{C})/\tau))$ where d_{embed} is the distance function in embedding space. The distance function is differentiable except at points equidistant from multiple copyrighted embeddings, which form a measure-zero set.

Corollary 86 (Protection Efficiency). *The ratio of protection strength to distance from copyrighted content achieves its maximum at distance $d^* = \tau$:*

$$\frac{d}{dd_{\text{embed}}} \left[\frac{m_{\text{eff}}(x)}{d_{\text{embed}}(x, \mathcal{C})} \right]_{d=\tau} = 0 \quad (537)$$

with value:

$$\left. \frac{m_{\text{eff}}(x)}{d_{\text{embed}}(x, \mathcal{C})} \right|_{d=\tau} = \frac{m(1 - e^{-1})}{\tau} \approx \frac{0.632m}{\tau} \quad (538)$$

This identifies the optimal distance where protection per unit distance is maximized.

Proposition 87 (Compositionality of Protection). *When multiple copyrighted works $\{c_1, \dots, c_k\} \subset \mathcal{C}$ are nearby in embedding space, their protection fields compose approximately additively in the log-probability domain:*

$$m_{\text{eff}}(x) \approx m \cdot \left(1 - \prod_{i=1}^k \exp\left(-\frac{\|\phi(x) - \phi(c_i)\|_2}{\tau}\right) \right) \quad (539)$$

This ensures that clusters of copyrighted content create stronger protection zones than isolated works.

Proof. The result follows from analyzing the superposition of gradient fields from multiple sources. In the linearized regime where individual contributions are small, the fields add linearly. The product form emerges from the independence assumption of contributions from well-separated sources and the exponential nature of the probability transformations.

Remark 88 (Adaptive Temperature Scheduling). In practice, the temperature parameter τ can be adapted during training. Starting with large τ ensures smooth optimization, while gradually decreasing τ sharpens the protection boundaries. This annealing schedule resembles simulated annealing in optimization, balancing exploration and exploitation of the energy landscape.

G.7 PROOF OF COROLLARY 9 (ROBUSTNESS TO DISTRIBUTION SHIFT)

We establish that the protection guarantees of AER degrade gracefully under distribution shift, maintaining exponential suppression even when the test distribution differs from the training distribution.

Proof. The proof proceeds by analyzing how distribution shift affects the energy gap and constructing an optimal coupling to bound the degradation.

Setup and notation. Let $\mathbb{P}_{\text{train}}$ denote the training distribution and $\mathbb{P}_{\text{test}}$ the test distribution, with total variation distance:

$$\|\mathbb{P}_{\text{test}} - \mathbb{P}_{\text{train}}\|_{\text{TV}} = \sup_{A \subseteq \mathcal{V}^*} |\mathbb{P}_{\text{test}}(A) - \mathbb{P}_{\text{train}}(A)| \leq \delta \quad (540)$$

By the variational characterization of total variation distance:

$$\|\mathbb{P}_{\text{test}} - \mathbb{P}_{\text{train}}\|_{\text{TV}} = \frac{1}{2} \int_{\mathcal{V}^*} |p_{\text{test}}(x) - p_{\text{train}}(x)| dx \quad (541)$$

The model θ^* is trained to maintain energy gap $\Delta(\theta^*) \geq m$ under $\mathbb{P}_{\text{train}}$:

$$\Delta_{\text{train}}(\theta^*) = \mathbb{E}_{c \sim \mathbb{P}_{\text{train}}^c} [E(c; \theta^*)] - \mathbb{E}_{o \sim \mathbb{P}_{\text{train}}^o} [E(o; \theta^*)] \geq m \quad (542)$$

where $\mathbb{P}_{\text{train}}^c$ and $\mathbb{P}_{\text{train}}^o$ denote the conditional distributions over copyrighted and open-source content respectively.

Coupling construction. By the coupling lemma, there exists a joint distribution π on $\mathcal{V}^* \times \mathcal{V}^*$ with marginals $\mathbb{P}_{\text{train}}$ and $\mathbb{P}_{\text{test}}$ such that:

$$\Pr_{(X,Y) \sim \pi} [X \neq Y] = \|\mathbb{P}_{\text{test}} - \mathbb{P}_{\text{train}}\|_{\text{TV}} \leq \delta \quad (543)$$

This optimal coupling maximally aligns the two distributions, with disagreement probability exactly equal to the total variation distance.

Energy gap under distribution shift. Under the test distribution, the energy gap becomes:

$$\Delta_{\text{test}}(\theta^*) = \mathbb{E}_{c \sim \mathbb{P}_{\text{test}}^c} [E(c; \theta^*)] - \mathbb{E}_{o \sim \mathbb{P}_{\text{test}}^o} [E(o; \theta^*)] \quad (544)$$

We decompose each expectation using the coupling. For the copyrighted content term:

$$\mathbb{E}_{c \sim \mathbb{P}_{\text{test}}^c} [E(c; \theta^*)] = \mathbb{E}_{(c_1, c_2) \sim \pi^c} [E(c_2; \theta^*)] \quad (545)$$

$$= \mathbb{E}_{(c_1, c_2) \sim \pi^c} [E(c_1; \theta^*) \cdot \mathbf{1}_{c_1=c_2}] + \mathbb{E}_{(c_1, c_2) \sim \pi^c} [E(c_2; \theta^*) \cdot \mathbf{1}_{c_1 \neq c_2}] \quad (546)$$

where π^c is the coupling restricted to copyrighted content.

Bounding the deviation. The energy function is bounded by design (from EBM normalizability): $|E(x; \theta^*)| \leq E_{\text{max}}$ for all $x \in \mathcal{V}^*$. Using this bound:

$$\left| \mathbb{E}_{c \sim \mathbb{P}_{\text{test}}^c} [E(c; \theta^*)] - \mathbb{E}_{c \sim \mathbb{P}_{\text{train}}^c} [E(c; \theta^*)] \right| \leq E_{\text{max}} \cdot \Pr_{(c_1, c_2) \sim \pi^c} [c_1 \neq c_2] \quad (547)$$

$$\leq E_{\text{max}} \cdot \delta \quad (548)$$

Similarly for open-source content:

$$\left| \mathbb{E}_{o \sim \mathbb{P}_{\text{test}}^o} [E(o; \theta^*)] - \mathbb{E}_{o \sim \mathbb{P}_{\text{train}}^o} [E(o; \theta^*)] \right| \leq E_{\text{max}} \cdot \delta \quad (549)$$

Energy gap degradation. Combining the bounds:

$$\Delta_{\text{test}}(\theta^*) = \Delta_{\text{train}}(\theta^*) + \left(\mathbb{E}_{c \sim \mathbb{P}_{\text{test}}^c} [E(c; \theta^*)] - \mathbb{E}_{c \sim \mathbb{P}_{\text{train}}^c} [E(c; \theta^*)] \right) \quad (550)$$

$$- \left(\mathbb{E}_{o \sim \mathbb{P}_{\text{test}}^o} [E(o; \theta^*)] - \mathbb{E}_{o \sim \mathbb{P}_{\text{train}}^o} [E(o; \theta^*)] \right) \quad (551)$$

$$\geq m - 2E_{\text{max}} \cdot \delta \quad (552)$$

Normalization of energy scale. The energy scale can be normalized without loss of generality such that $E_{\text{max}} = 1$ (by rescaling the temperature parameter in the EBM). Under this normalization:

$$\Delta_{\text{test}}(\theta^*) \geq m - 2\delta \quad (553)$$

Generation probability bound. For any copyrighted sequence $c \in \mathcal{C}$, the generation probability under the test distribution satisfies:

$$p_{\theta^*}(c | \mathbb{P}_{\text{test}}) = \exp(-|c| \cdot E(c; \theta^*)) \quad (554)$$

$$\leq \exp(-|c| \cdot [E_{\mathcal{O}}(\theta^*) + \Delta_{\text{test}}(\theta^*)]) \quad (555)$$

$$\leq \exp(-|c| \cdot [E_{\mathcal{O}}(\theta^*) + m - 2\delta]) \quad (556)$$

$$= p_{\theta_{\text{base}}}(c) \cdot \exp(-(m - 2\delta) \cdot |c|) \quad (557)$$

where we used that $E_{\mathcal{O}}(\theta^*) \approx E_{\mathcal{O}}(\theta_{\text{base}})$ for open-source content (preserved by AER training).

This completes the proof, showing exponential suppression with gracefully degraded margin $m - 2\delta$.

Remark 89 (Tightness of the Bound). The bound is tight in the worst case. Consider a adversarial shift that swaps copyrighted and open-source content with probability δ . This achieves the maximum degradation of 2δ in the energy gap while respecting the total variation constraint.

Lemma 90 (Refined Bound under Smooth Shift). *If the distribution shift is smooth in the sense that the Wasserstein distance $W_2(\mathbb{P}_{\text{test}}, \mathbb{P}_{\text{train}}) \leq \epsilon$, then a tighter bound holds:*

$$\Delta_{\text{test}}(\theta^*) \geq m - L \cdot \epsilon \quad (558)$$

where L is the Lipschitz constant of the energy function in embedding space.

Proof. Under Wasserstein distance bounds, the coupling can be chosen to minimize the expected distance between coupled points. For L -Lipschitz energy function:

$$|E(x; \theta^*) - E(y; \theta^*)| \leq L \cdot \|\phi(x) - \phi(y)\|_2 \quad (559)$$

The optimal transport coupling π^* satisfies:

$$\mathbb{E}_{(X,Y) \sim \pi^*} [\|\phi(X) - \phi(Y)\|_2] = W_2(\mathbb{P}_{\text{test}}, \mathbb{P}_{\text{train}}) \leq \epsilon \quad (560)$$

Therefore:

$$\left| \mathbb{E}_{c \sim \mathbb{P}_{\text{test}}^c} [E(c; \theta^*)] - \mathbb{E}_{c \sim \mathbb{P}_{\text{train}}^c} [E(c; \theta^*)] \right| \leq L \cdot \mathbb{E}_{(c_1, c_2) \sim \pi^*} [\|\phi(c_1) - \phi(c_2)\|_2] \quad (561)$$

$$\leq L \cdot \epsilon \quad (562)$$

The same bound applies to open-source content, yielding the refined bound.

Proposition 91 (Robustness Comparison with Inverse Regularization). *Under the same distribution shift $\|\mathbb{P}_{\text{test}} - \mathbb{P}_{\text{train}}\|_{TV} \leq \delta$, inverse regularization methods that directly minimize $p(c)$ suffer catastrophic failure:*

$$p_{\theta_{\text{inv}}}(c | \mathbb{P}_{\text{test}}) \geq p_{\theta_{\text{base}}}(c) \cdot (1 - O(\delta)) \quad (563)$$

providing only linear degradation compared to AER's exponential protection.

Proof. Inverse regularization directly optimizes $\min_{\theta} p_{\theta}(c)$ for $c \in \mathcal{C}$. Under distribution shift, the gradient signal from copyrighted content is diluted by factor $(1 - \delta)$. The optimization landscape changes from:

$$\nabla_{\theta} \mathcal{L}_{\text{inv}} = \nabla_{\theta} \mathcal{L}_{\text{LM}} - \lambda \sum_{c \in \mathcal{C}} \nabla_{\theta} \log p_{\theta}(c) \quad (564)$$

to approximately:

$$\nabla_{\theta} \mathcal{L}_{\text{inv}}^{\text{test}} \approx \nabla_{\theta} \mathcal{L}_{\text{LM}} - \lambda(1 - \delta) \sum_{c \in \mathcal{C}} \nabla_{\theta} \log p_{\theta}(c) \quad (565)$$

This linear scaling of the regularization strength leads to only linear reduction in protection, insufficient for copyright compliance under realistic distribution shifts.

Remark 92 (Practical Implications). The robustness guarantee $m - 2\delta$ suggests that practitioners should:

1. Choose margin m conservatively, accounting for expected distribution shift magnitude
2. Monitor distribution shift during deployment using techniques like maximum mean discrepancy
3. Retrain periodically when cumulative shift exceeds $m/4$ to maintain strong protection

The exponential nature of protection ensures that even with moderate degradation, copyright compliance remains effective.

H ADDITIONAL TECHNICAL RESULTS

H.1 SAMPLE COMPLEXITY BOUNDS

We establish the sample complexity required to achieve target protection levels, providing practical guidance for dataset construction.

Theorem 93 (Sample Complexity for Target Protection). *To achieve energy gap $\Delta(\theta) \geq m - \epsilon$ with probability at least $1 - \delta$, the required number of copyrighted samples satisfies:*

$$n_c \geq \frac{2\sigma^2}{\epsilon^2} \left(\log \frac{2}{\delta} + d \log \left(1 + \frac{4R}{\epsilon} \right) \right) \quad (566)$$

where σ^2 is the variance of the energy function, d is the effective dimension of the parameter space, and R is the parameter norm bound.

Proof. The proof uses empirical process theory and Rademacher complexity bounds.

Empirical energy gap. The empirical energy gap based on finite samples is:

$$\hat{\Delta}_n(\theta) = \frac{1}{n_c} \sum_{i=1}^{n_c} E(c_i; \theta) - \frac{1}{n_o} \sum_{j=1}^{n_o} E(o_j; \theta) \quad (567)$$

The deviation from the true gap follows from McDiarmid’s inequality. Define:

$$Z_n = \hat{\Delta}_n(\theta) - \Delta(\theta) \quad (568)$$

Concentration analysis. The energy function satisfies bounded differences: changing one sample affects the empirical gap by at most $2E_{\max}/\min(n_c, n_o)$. By McDiarmid’s inequality:

$$\Pr[|Z_n| \geq t] \leq 2 \exp \left(-\frac{2t^2 \min(n_c, n_o)}{E_{\max}^2} \right) \quad (569)$$

Setting $t = \epsilon/2$ and requiring probability at least $1 - \delta/2$:

$$n_c \geq \frac{2E_{\max}^2}{\epsilon^2} \log \frac{4}{\delta} \quad (570)$$

Uniform convergence over parameter space. The energy gap must hold uniformly over the parameter ball $\Theta_R = \{\theta : \|\theta\|_2 \leq R\}$. The Rademacher complexity of the energy function class is:

$$\mathcal{R}_n(\mathcal{E}) = \mathbb{E}_{\sigma, \mathcal{S}} \left[\sup_{\theta \in \Theta_R} \frac{1}{n} \sum_{i=1}^n \sigma_i E(x_i; \theta) \right] \quad (571)$$

where σ_i are Rademacher random variables.

For neural networks with ReLU activations and L layers:

$$\mathcal{R}_n(\mathcal{E}) \leq \frac{2LR\sqrt{d}}{n} \prod_{l=1}^L \|W_l\|_{\text{op}} \quad (572)$$

where $\|W_l\|_{\text{op}}$ denotes the operator norm of layer l .

Generalization bound. By standard Rademacher complexity arguments:

$$\Pr \left[\sup_{\theta \in \Theta_R} |\hat{\Delta}_n(\theta) - \Delta(\theta)| \geq \epsilon \right] \leq \delta \quad (573)$$

requires:

$$n_c \geq \frac{C}{\epsilon^2} \left(\mathcal{R}_n(\mathcal{E})^2 + \log \frac{1}{\delta} \right) \quad (574)$$

Variance-dependent bound. Under sub-Gaussian energy distributions with variance proxy σ^2 :

$$\text{Var}[E(c; \theta)] \leq \sigma^2, \quad \text{Var}[E(o; \theta)] \leq \sigma^2 \quad (575)$$

The refined sample complexity becomes:

$$n_c \geq \frac{2\sigma^2}{\epsilon^2} \left(\log \frac{2}{\delta} + d \log \left(1 + \frac{4R}{\epsilon} \right) \right) \quad (576)$$

completing the proof.

Corollary 94 (Scaling with Model Size). *For transformer models with parameter count N , hidden dimension h , and depth L :*

$$n_c = \Omega \left(\frac{hL}{\epsilon^2} \log \frac{N}{\delta} \right) \quad (577)$$

The sample complexity scales logarithmically with model size, making the approach feasible for large language models.

Lemma 95 (Adaptive Sample Allocation). *The optimal allocation ratio between copyrighted and open-source samples that minimizes total sample complexity is:*

$$\frac{n_c}{n_o} = \sqrt{\frac{\text{Var}[E(c; \theta)]}{\text{Var}[E(o; \theta)]}} \quad (578)$$

When variances are equal, balanced sampling ($n_c = n_o$) is optimal.

Proof. The proof follows from minimizing the variance of the empirical gap estimator $\hat{\Delta}_n$ subject to a fixed total budget $n_c + n_o = n$. Using Lagrange multipliers, the optimal allocation satisfies the stated ratio.

Proposition 96 (Early Stopping Criterion). *Define the empirical gap trajectory $\hat{\Delta}_t(\theta_t)$ during training. With probability at least $1 - \delta$, if:*

$$\hat{\Delta}_t(\theta_t) \geq m + \sqrt{\frac{2 \log(2T/\delta)}{n_c}} \quad (579)$$

for T consecutive iterations, then $\Delta(\theta_t) \geq m$ with high probability.

H.2 NUMERICAL STABILITY GUARANTEES

We establish theoretical guarantees for the numerical stability of AER computations under finite-precision arithmetic, ensuring that theoretical protection guarantees translate to practical implementations.

Theorem 97 (Stability of Energy-Based Computations). *Let $\mathbb{F}_\beta$ denote the set of floating-point numbers with precision β bits. For any sequence $c \in \mathcal{C}$ and parameters $\theta \in \Theta$, the relative error in log-probability computation satisfies:*

$$\left| \frac{\log \hat{p}_\theta(c) - \log p_\theta(c)}{\log p_\theta(c)} \right| \leq \kappa(|c|) \cdot \epsilon_\beta \quad (580)$$

where $\hat{p}_\theta(c)$ denotes the finite-precision approximation, $\epsilon_\beta = 2^{-\beta}$ is the machine epsilon, and $\kappa(|c|) = O(|c|/m)$ is the condition number that grows linearly with sequence length.

Proof. We analyze error propagation through the energy-based formulation using backward error analysis.

Energy computation in finite precision. The energy function is computed as:

$$E(c; \theta) = -\frac{1}{|c|} \sum_{t=1}^{|c|} \log p_\theta(c_t | c_{<t}) \quad (581)$$

In finite precision arithmetic, each operation $\hat{op}$ satisfies:

$$\hat{op}(a, b) = op(a, b)(1 + \delta), \quad |\delta| \leq \epsilon_\beta \quad (582)$$

Error accumulation. Using the standard model of floating-point arithmetic, the computed energy satisfies:

$$\hat{E}(c; \theta) = E(c; \theta) \prod_{k=1}^{K(|c|)} (1 + \delta_k) \quad (583)$$

where $K(|c|) = O(|c|)$ is the number of arithmetic operations.

By the first-order approximation for small δ_k :

$$\prod_{k=1}^{K(|c|)} (1 + \delta_k) \approx 1 + \sum_{k=1}^{K(|c|)} \delta_k \quad (584)$$

Relative error bound. The accumulated relative error is:

$$\left| \frac{\hat{E}(c; \theta) - E(c; \theta)}{E(c; \theta)} \right| \leq K(|c|) \cdot \epsilon_\beta + O(\epsilon_\beta^2) \quad (585)$$

Since $\log p_\theta(c) = -|c| \cdot E(c; \theta)$, the relative error in log-probability is:

$$\left| \frac{\log \hat{p}_\theta(c) - \log p_\theta(c)}{\log p_\theta(c)} \right| = \left| \frac{\hat{E}(c; \theta) - E(c; \theta)}{E(c; \theta)} \right| \leq \frac{K(|c|) \cdot \epsilon_\beta}{m} \quad (586)$$

where we used that $E(c; \theta) \geq m$ for copyrighted content under AER. Setting $\kappa(|c|) = K(|c|)/m$ completes the proof.

Theorem 98 (Stability of Log-Sum-Exp Operations). *For computing partition functions and marginal probabilities, the log-sum-exp operation:*

$$LSE(x_1, \dots, x_n) = \log \sum_{i=1}^n \exp(x_i) \quad (587)$$

can be evaluated with relative error bounded by:

$$\left| \frac{\widehat{LSE}(x) - LSE(x)}{LSE(x)} \right| \leq n \cdot \epsilon_\beta \cdot \exp\left(\max_i x_i - LSE(x)\right) \quad (588)$$

The error is minimized when using the shifted form with $x_{\max} = \max_i x_i$.

Proof. Define the shifted computation:

$$\widehat{LSE}(x) = x_{\max} + \log \sum_{i=1}^n \exp(x_i - x_{\max}) \quad (589)$$

Each exponentiation introduces relative error ϵ_β :

$$\widehat{\exp}(x_i - x_{\max}) = \exp(x_i - x_{\max})(1 + \delta_i), \quad |\delta_i| \leq \epsilon_\beta \quad (590)$$

The summation error is:

$$\left| \sum_{i=1}^n \widehat{\exp}(x_i - x_{\max}) - \sum_{i=1}^n \exp(x_i - x_{\max}) \right| \leq \epsilon_\beta \sum_{i=1}^n \exp(x_i - x_{\max}) \quad (591)$$

Taking logarithms and using the inequality $|\log(1 + x)| \leq |x|/(1 - |x|)$ for $|x| < 1$:

$$\left| \widehat{LSE}(x) - LSE(x) \right| \leq \frac{n \cdot \epsilon_\beta}{1 - n \cdot \epsilon_\beta} \cdot \exp(x_{\max} - LSE(x)) \quad (592)$$

For $n \cdot \epsilon_\beta \ll 1$, this yields the stated bound.

Theorem 99 (Gradient Computation Stability). *The gradient of the AER objective maintains bounded condition number during optimization:*

$$\kappa(\nabla_{\theta} \mathcal{L}_{\text{AER}}) \leq \kappa(\nabla_{\theta} \mathcal{L}_{\text{LM}}) + \gamma \cdot \frac{m}{\tau} \cdot \kappa(\nabla_{\theta} E) \quad (593)$$

where $\kappa(\cdot)$ denotes the condition number. The regularization does not amplify numerical instability beyond a controllable factor.

Proof. The AER gradient decomposes as:

$$\nabla_{\theta} \mathcal{L}_{\text{AER}} = \nabla_{\theta} \mathcal{L}_{\text{LM}} + \gamma \cdot \sigma \left(\frac{m - \Delta(\theta)}{\tau} \right) \cdot \nabla_{\theta} \Delta(\theta) \quad (594)$$

where $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function.

The condition number of the sum is bounded by:

$$\kappa(\nabla_{\theta} \mathcal{L}_{\text{AER}}) \leq \kappa(\nabla_{\theta} \mathcal{L}_{\text{LM}}) + \gamma \cdot \sup_{x \in \mathbb{R}} |\sigma'(x)| \cdot \kappa(\nabla_{\theta} \Delta) \quad (595)$$

Since $\sup_x |\sigma'(x)| = 1/4$ and $\kappa(\nabla_{\theta} \Delta) \leq m \cdot \kappa(\nabla_{\theta} E)/\tau$:

$$\kappa(\nabla_{\theta} \mathcal{L}_{\text{AER}}) \leq \kappa(\nabla_{\theta} \mathcal{L}_{\text{LM}}) + \gamma \cdot \frac{m}{4\tau} \cdot \kappa(\nabla_{\theta} E) \quad (596)$$

The bound follows from the smoothness of the sigmoid activation.

Proposition 100 (Backward Stability of AER Training). *The AER optimization algorithm is backward stable: the computed parameters $\hat{\theta}^*$ are the exact solution to a perturbed problem:*

$$\hat{\theta}^* = \arg \min_{\theta} [\mathcal{L}_{\text{AER}}(\theta) + \Delta \mathcal{L}(\theta)] \quad (597)$$

where the perturbation satisfies $\|\Delta \mathcal{L}\|_{\infty} \leq O(T \cdot \epsilon_{\beta}) \cdot \|\mathcal{L}_{\text{AER}}\|_{\infty}$ for T optimization steps.

Proof. The proof follows from the backward error analysis of gradient descent. Each gradient step introduces a backward error:

$$\hat{\theta}_{t+1} = \theta_t - \eta \nabla_{\theta} \mathcal{L}_{\text{AER}}(\theta_t) + \eta \cdot e_t \quad (598)$$

where $\|e_t\|_2 \leq \epsilon_{\beta} \cdot \|\nabla_{\theta} \mathcal{L}_{\text{AER}}(\theta_t)\|_2$.

The accumulated backward error after T steps can be interpreted as optimizing a perturbed objective with the stated bound.

Corollary 101 (Preservation of Protection Guarantees). *Under finite-precision arithmetic with β bits, the protection guarantee degrades by at most a multiplicative factor:*

$$\hat{p}_{\theta^*}(c) \leq p_{\theta_{\text{base}}}(c) \cdot \exp(-m|c|(1 - O(|c| \cdot 2^{-\beta}))) \quad (599)$$

For standard precision ($\beta \geq 32$) and reasonable sequence lengths ($|c| \leq 2^{16}$), the degradation factor $(1 - O(|c| \cdot 2^{-\beta})) \approx 1$, preserving exponential suppression.

Remark 102 (Precision-Performance Trade-off). The analysis reveals a fundamental trade-off: higher precision β reduces numerical error but increases computational cost. The optimal choice depends on the protection requirements:

- **High protection regime** ($m \geq 5$): Requires $\beta \geq 64$ to maintain stability
- **Moderate protection** ($m \in [1, 5]$): Standard precision $\beta = 32$ suffices
- **Mixed precision**: Use higher precision only for energy gap computation while maintaining lower precision for forward passes

Lemma 103 (Overflow Threshold). *The maximum sequence length processable without overflow in the exponential terms is:*

$$L_{\max}(\beta) = \frac{(2^{\beta-1} - 1) \log 2}{m} \quad (600)$$

This provides a theoretical limit on sequence length as a function of precision and protection strength.

Theorem 104 (Numerical Differentiation Stability). *When using automatic differentiation for gradient computation, the relative error in gradients satisfies:*

$$\frac{\|\hat{\nabla}_{\theta} \mathcal{L}_{AER} - \nabla_{\theta} \mathcal{L}_{AER}\|_2}{\|\nabla_{\theta} \mathcal{L}_{AER}\|_2} \leq \mathcal{D} \cdot \epsilon_{\beta}^{2/3} \quad (601)$$

where $\mathcal{D}$ is the computational graph depth. The error scales sub-linearly with graph depth, ensuring stability for deep models.