On the Role of Attention in Prompt-tuning

Samet Oymak [*1] **Ankit Singh Rawat** [*2] **Mahdi Soltanolkotabi** [*3] **Christos Thrampoulidis** [*4]

Abstract

Prompt-tuning is an emerging strategy to adapt large language models (LLM) to downstream tasks by learning a (soft-)prompt parameter from data. Despite its success in LLMs, there is limited theoretical understanding of the power of prompt-tuning and the role of the attention mechanism in prompting. In this work, we explore prompt-tuning for one-layer attention architectures and study contextual mixture-models where each input token belongs to a context-relevant or -irrelevant set. We isolate the role of prompt-tuning through a self-contained *prompt-attention* model. Our contributions are as follows: (1) We show that softmax-prompt-attention is provably more expressive than softmax-self-attention and linear-prompt-attention under our contextual data model. (2) We analyze the initial trajectory of gradient descent and show that it learns the prompt and prediction head with near-optimal sample complexity and demonstrate how the prompt can provably attend to sparse context-relevant tokens. (3) Assuming a known prompt but an unknown prediction head, we characterize the exact finite sample performance of prompt-attention which reveals the fundamental performance limits and the precise benefit of the context information. We also provide experiments that verify our theoretical insights on real datasets and demonstrate how prompt-tuning enables the model to attend to context-relevant information.

1. Introduction

Transformer models have achieved remarkable success in a wide array of machine learning domains spanning language modeling, vision, and decision making. Recently, one of the key techniques that has helped pave the way for the deployment of transformers to ever increasing application areas is their ability to adapt to multiple unseen tasks by conditioning their predictions through their inputs – a technique known as prompt-tuning (Lester et al., 2021; Li & Liang, 2021). Concretely, prompt-tuning provides a more efficient (cheaper/faster) alternative to fine-tuning the entire weights of the transformer by instead training (fewer) so-called prompt parameters that are appended on the input and can be thought of as an input interface. In fact, several recent works have demonstrated experimentally that prompt-tuning is not only more efficient, but often even becomes competitive to fine-tuning in terms of accuracy (Lester et al., 2021; Liu et al., 2023). However, there is currently limited formal justification of such observations. This motivates the first question of this paper:

How does prompt-tuning compare to fine-tuning in terms of expressive power? Are there scenarios prompt-tuning outperforms fine-tuning in that regard?

The core constituent of a transformer, and thus of prompt-tuning, is the attention mechanism. Through the attention layer, prompts get to interact with other input features, create/modify attention weights, and facilitate the model to attend on latent task-specific information. The standard attention layer relies on softmax nonlinearities. Operationally, the softmax function allows a model to selectively focus on certain parts of the input tokens when generating attention outputs. However, there is little rigorous understanding of attention-based prompt-tuning. Concretely,

What is the role of the softmax-attention in prompt-tuning in terms of optimization and generalization? How does it locate and extract relevant contextual information?

Contributions. Our contributions are as follows:

- We show that a particular form of attention which we refer to as *prompt-attention* naturally arises from the self-attention model with prompt-tuning. We identify provable scenarios where it is more expressive than self-attention and linear-prompt-attention. [1] This separation result reveals insightful data models where prompt-tuning can be superior to fine-tuning with self-attention.

[*]In alphabetical order [1]University of Michigan & UC Riverside, USA [2]Google Research NYC, USA [3]University of Southern California, USA [4]University of British Columbia, Canada. Correspondence to: Christos Thrampoulidis <cthrampo@ece.ubc.ca>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1]Our emphasis is on the role of attention (whether prompt- or self-). However, we analyze the general problem where attention weights are optimized jointly with the linear classifier head.

- We develop new statistical foundations for gradient-based prompt-tuning: we study the optimization and generalization dynamics of the initial trajectory of gradient descent for optimizing prompt-attention. Concretely, we show the first few iterations learn the prompt and prediction head with near-optimal sample complexity while achieving high accuracy.
- Our results provide insights into the critical role of softmax in facilitating attention: we show how the initial trajectory of gradient descent utilizes softmax to provably attend to sparse context-relevant tokens, ignoring noisy/nuisance tokens.
- We also characterize the exact finite sample performance of prompt-attention assuming known prompt but unknown prediction head. This reveals the fundamental performance limits and precisely quantifies the benefits of context information.
- Our results highlight various trade-offs among different model parameters: (i) the role of sparsity, i.e., the fraction of context-relevant tokens, and (ii) the relative effects of the different constituents of context-relevant tokens.
- Finally, we empirically validate our theoretical insights on both synthetic contextual-mixture datasets and image-classification datasets. Specifically, we compare multiple variants of prompt-tuning against standard fine-tuning on the latter. Our results highlight the role of prompt-attention in selecting relevant tokens in the image classification setting.

Related works. Attention, specifically the so-called self-attention, is the backbone mechanism of transformers (Vaswani et al., 2017). It differs from conventional models (e.g., multi-layer perceptrons and convolutional neural networks) in that it computes feature representations by globally modeling interactions between different parts of an input sequence. Despite tremendous empirical success (see, e.g., Vaswani et al., 2017; Brown et al., 2020; Saharia et al., 2022; Ramesh et al., 2022; cha; Narayanan et al., 2021; Reed et al., 2022, and references therein), the underlying mechanisms of the attention layer remain largely unknown: How does it learn? What makes it better (and when) compared to conventional architectures? Yun et al. (2020) show that self-attention based transformers with large enough depth are universal approximators of seq2seq functions. Focusing on the self-attention component, Edelman et al. (2021) show that self-attention can efficiently represent sparse functions of its input space, while Sahiner et al. (2022); Ergen et al. (2022) analyze convex-relaxations of Self-attention, and Baldi & Vershynin (2022); Dong et al. (2021) study the expressive ability of attention layers. However, these works do *not* characterize the optimization and generalization dynamics of attention. To the best of our knowledge, the only prior works attempting this are Jelassi et al. (2022) and Li et al. (2023). Jelassi et al. (2022) assume a simplified attention structure in which the attention matrix is *not* directly parameterized in terms of the input sequence. Our paper also distinguishes itself from contemporaneous work by Li et al. (2023) in several ways: (1) Unlike their data model, ours incorporates a context vector and employs a sub-Gaussian noise model instead of assuming bounded noise. (2) We provide a precise asymptotic analysis that elucidates the role of various problem parameters. (3) While Li et al. (2023) primarily focuses on vanilla self-attention, our study centers on understanding the potential benefits of prompt-tuning through prompt-attention.

2. Problem setting

2.1. Motivation: Prompt-tuning

Consider a single-head self-attention layer

$$\boldsymbol{O}_{\text{pre}} = \phi(\boldsymbol{X}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{W}_V\,, \tag{1}$$

with input $\boldsymbol{X} \in \mathbb{R}^{T\times d}$ consisting of T tokens of dimension d each, trainable parameters $(\boldsymbol{W}_K, \boldsymbol{W}_Q, \boldsymbol{W}_V)$ and a softmax nonlinearity $\phi : \mathbb{R}^T \mapsto \mathbb{R}^T$, $[\phi(\boldsymbol{v})]_t = e^{v_t}/\sum_{t'\in[T]} e^{v_{t'}}$ that acts row-wise when its argument is a $T \times T$ matrix. We scalarize the output of the self-attention layer with a trainable linear head $\bar{\boldsymbol{U}}$ which yields

$$y_{\text{pre}} = \langle \bar{\boldsymbol{U}}, \boldsymbol{O}_{\text{pre}}\rangle = \langle \boldsymbol{U}, \phi(\boldsymbol{X}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top)\boldsymbol{X}\rangle\,. \tag{2}$$

Note here that we implicitly subsume the value matrix $\boldsymbol{W}_V$ in the linear head via $\boldsymbol{U} := \bar{\boldsymbol{U}}\boldsymbol{W}_V^\top$.

We assume that the model above is pre-trained so that $\boldsymbol{W}_K, \boldsymbol{W}_Q, \boldsymbol{U}$ are fixed. Our goal is to use the pretrained transformer on (potentially) new classification tasks. Towards this goal, we explore the use of prompt-tuning, introduced in Li & Liang (2021); Lester et al. (2021) as an alternative to fine-tuning the existing transformer weights.

Prompt-tuning appends a trainable prompt $\boldsymbol{P} \in \mathbb{R}^{m\times d}$ to the input features $\boldsymbol{X} \in \mathbb{R}^{T\times d}$ with the goal of conditioning the transformer to solve the new classification task. Let $\boldsymbol{X}_{\boldsymbol{P}} := \begin{bmatrix}\boldsymbol{P}\\ \boldsymbol{X}\end{bmatrix} \in \mathbb{R}^{(T+m)\times d}$ be the new transformer input. The output of the attention-layer is thus is of the form

$$\boldsymbol{O} = \phi(\boldsymbol{X}_{\boldsymbol{P}}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top)\boldsymbol{X}\,.$$

Note that this is slightly different from (1) in that now the layer computes a cross-attention between the augmented inputs $\boldsymbol{X}_{\boldsymbol{P}}$ and the original inputs $\boldsymbol{X}$. This is also equivalent to self-attention on $\boldsymbol{X}_{\boldsymbol{P}}$ after masking the prompt $\boldsymbol{P}$ from keys. This masking is used to cleanly isolate the residual contribution of the prompt without impacting the frozen attention output. Concretely, let $\boldsymbol{W}_{\text{head}}$ be the prediction head associated with the prompt tokens. As before, we

scalarize the output by projecting with a linear head of size $(T+m)\times d$ as follows:

$$y = \langle [\boldsymbol{W}_{\text{head}}{}^\top \boldsymbol{U}^\top]^\top, \phi(\boldsymbol{X}_{\boldsymbol{P}}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top)\boldsymbol{X}\rangle \tag{3}$$
$$= \underbrace{\langle \boldsymbol{W}_{\text{head}}, \phi(\boldsymbol{P}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top)\boldsymbol{X}\rangle}_{\text{prompt-attention } y_{\text{new}}} + \underbrace{\langle \boldsymbol{U}, \phi(\boldsymbol{X}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top)\boldsymbol{X}\rangle}_{\text{frozen self-attention } y_{\text{pre}}}.$$

Here, y_{new} captures the additive impact of prompt-tuning on the prediction. We denote the trainable parameters in the model above as $\boldsymbol{\theta} := (\boldsymbol{W}_{\text{head}}, \boldsymbol{P})$ [2] Since the y_{new} term becomes a self-contained module and the features attend directly to the prompt vector, we will refer to it as *prompt-attention*.

Our goal is understanding the expressivity, training dynamics, and generalization properties of the above model. To simplify our analysis, we consider the following setting.

1. We focus our attention on the novel component y_{new} of the model output in (3) so as to isolate and fully understand the capabilities of prompt-attention.
2. We assume $\boldsymbol{W}_K, \boldsymbol{W}_Q \in \mathbb{R}^{d\times d}$ are full-rank.
3. We assume a single trainable prompt $\boldsymbol{q} \in \mathbb{R}^d$ i.e., $m = 1$.

Prompt-attention model. Using these assumptions and setting $\boldsymbol{q} := \boldsymbol{W}_K\boldsymbol{W}_Q^\top\boldsymbol{P}^\top \in \mathbb{R}^d$ and $\boldsymbol{w} = \boldsymbol{W}_{\text{head}}^\top \in \mathbb{R}^d$, we arrive at our core *prompt-attention* model $f_{\boldsymbol{\theta}}^{\text{ATT}}$ (or simply f^{ATT}):

$$f_{\boldsymbol{\theta}}^{\text{ATT}}(\boldsymbol{X}) = \langle \boldsymbol{w}, \boldsymbol{X}^\top\phi(\boldsymbol{X}\boldsymbol{q})\rangle,\ \boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{q}). \tag{4}$$

We shall see that this model exhibits interesting properties to learn rich contextual relationships within the data and can even be more expressive than a single self-attention layer.

We remark that the model above is of interest even beyond the context of prompting: the prompt-attention model in (4) is reminiscent of the simplified model proposed in earlier seq2seq architectures (Bahdanau et al., 2014; Xu et al., 2015; Chan et al., 2015) preceding self-attention and Transformers (Vaswani et al., 2017). Indeed, in the simplified attention mechanism of (Bahdanau et al., 2014; Xu et al., 2015; Chan et al., 2015), the tokens' *relevance scores* and corresponding *attention weights* are determined by $\boldsymbol{a} = \phi(\boldsymbol{X}\boldsymbol{q})$ in which $\boldsymbol{q}$ is a trainable vector and ϕ is the softmax-score transformation. Note here that the trainable parameter $\boldsymbol{q}$ corresponds exactly to the trainable prompt vector in (4).

[2] In our model, we train the classifier head $\boldsymbol{W}_{\text{head}}$ in addition to the prompt vectors $\boldsymbol{P}$. Despite the additional training for the classifier head, the computational overhead remains minimal, and the overall scheme remains significantly more efficient compared to updating the entire model $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{U}$.

2.2. Contextual data model

Consider supervised classification on IID data $(\boldsymbol{X}, y) \sim \mathcal{D}$ with features $\boldsymbol{X} \in \mathbb{R}^{T\times d}$ and binary label $y \in \{\pm 1\}$.

Dataset model. We assume the following about an example $(\boldsymbol{X}, y)$ drawn from $\mathcal{D}$: The labels y are distributed as $\mathbb{P}(y = 1) = 1 - \mathbb{P}(y = -1) = \pi$; for simplicity, we set $\pi = 1/2$ so that $\mathbb{E}[y] = 0$. The tokens $\boldsymbol{x}_t, t \in [T]$ of input example $\boldsymbol{X} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]$ are split into a *context-relevant* set $\mathcal{R} \subset [T]$ and *context-irrelevant* set $\mathcal{R}^c := [T] - \mathcal{R}$. Specifically, conditioned on the labels and relevance set $\mathcal{R}$, tokens $\boldsymbol{x}_t, t \in [T]$ within $\boldsymbol{X}$ are i.i.d. as follows

$$\boldsymbol{x}_t | y = \begin{cases} \boldsymbol{q}_\star + y\boldsymbol{w}_\star, & t \in \mathcal{R} \quad \text{(relevant token)} \\ -\delta^q\boldsymbol{q}_\star - \delta^w y\boldsymbol{w}_\star + \boldsymbol{z}_t, & t \notin \mathcal{R} \quad \text{(\textbf{ir}relevant token)} \end{cases}. \tag{DATA}$$

In the above, $\boldsymbol{q}_\star$ is a context-vector that indicates token relevance and $\boldsymbol{w}_\star$ is a regressor vector. $y, \delta := (\delta^q, \delta^w), (\boldsymbol{z}_t)_{t=1}^T$ are independent variables as follows:

- $\delta = (\delta^q, \delta^w) \in \mathbb{R}^2$ is a bounded random variable that obeys $\delta^q \geq 0$. Thus, δ reflects *out-of-context* information within irrelevant tokens. However, δ is allowed to expose label information through δ^w. When $\delta = (0, 0)$ almost surely, we call the resulting distribution **core dataset model**.
- $\boldsymbol{z}_t$ are independent centered subgaussian and random variables with covariance $\boldsymbol{\Sigma}$ (see Ass. 1.a). When $\boldsymbol{\Sigma} = 0$, we call the resulting distribution **discrete dataset model**.
- We allow the relevance set $\mathcal{R}$ to be non-stochastic. This includes $\mathcal{R}$ being adversarial to the classification model.
- We assume constant fraction $\zeta = |\mathcal{R}|/T \in (0, 1)$ of label-relevant tokens for each input example $\boldsymbol{X}$ drawn from $\mathcal{D}$. Thus, ζ represents the sparsity of relevant signal tokens.

Training dataset $\mathcal{S} := (\boldsymbol{X}_i, y_i)_{i=1}^n$. We draw n i.i.d. samples from $\mathcal{D}$ to form our training dataset $\mathcal{S} := (\boldsymbol{X}_i, y_i)_{i=1}^n$. For i'th example $(\boldsymbol{X}_i, y_i)$, we denote the tokens by $(\boldsymbol{x}_{i,t})_{t=1}^T$, noise by $(\boldsymbol{z}_{i,t})_{t=1}^T$, relevance set by $\mathcal{R}_i$, and out-of-context variable by $\delta_i = (\delta_i^q, \delta_i^w)$.

Ideally, for i'th example, we would like to identify its context-relevant set $\mathcal{R}_i$ and discard the rest. This would especially help when the signal-to-noise-ratio is small, i.e. $\zeta = |\mathcal{R}_i|/T \ll 1$. This is precisely the role of the context-vector $\boldsymbol{q}_\star$: Observe that, per our construction, relevant tokens have positive correlation with $\boldsymbol{q}_\star$ whereas irrelevant tokens have non-positive correlation with $\boldsymbol{q}_\star$ in expectation. Thus, by focusing attention onto tokens based on their $\boldsymbol{q}_\star$ correlation, we can potentially select the relevant set.

Remark 1 (Model interpretation)**.** (DATA) *can be thought of as a simplified model for binary image classification with tokens being image patches of two types: ones revealing*

information about the label (set $\mathcal{R}$) and uninformative ones containing noise. Tokens in $\mathcal{R}$ contain information indicating: (i) class-membership via signed-regressor $y\boldsymbol{w}_\star$ and (ii) context-relevance via context-vector $\boldsymbol{q}_\star$. The signed-regressor differs across tokens of examples belonging to different classes $y \in \{\pm 1\}$, while the context-vector is common for all context-relevant tokes across classes. For a concrete example, consider images each depicting multiple, say $|\mathcal{R}|$, birds of one type surrounded by label-irrelevant/noisy background. The goal is to classify images according to one of two types of birds. Here, think of "context" as feature-information indicating corresponding pixels belong to "bird" (of either type) rather than "background," while the "regressor" represents feature information useful to distinguish between two bird types. Alternatively, (DATA) *may be modeling deep representations (rather than raw pixels) of the original images. Overall, simplified models similar to* (DATA) *have been used previously to analyze optimization and generalization dynamics of training fully-connected (Frei et al., 2022) and convolutional models (Cao et al., 2022). Specifically,* (DATA) *is an extension of the commonly used (sub)-Gaussian mixture model customized to the nature of attention: each example is tokenized and context-relevant information is described in terms of both a regressor (differing between classes) and a context (common across classes).*

2.3. Baseline Models

We compare performance of the prompt-attention model in (4) with the following three baseline models.

The linear model is parameterized by $\boldsymbol{\theta} = \boldsymbol{w}$ and outputs

$$f^{\text{LIN}}(\boldsymbol{w}) = \frac{1}{T}\boldsymbol{w}^\top \boldsymbol{X}^\top \mathbb{1}_T = \frac{1}{T}\sum_{t\in[T]} \boldsymbol{w}^\top \boldsymbol{x}_t. \tag{5}$$

Note this corresponds to a prompt-attention model with uniform attention weights $[\boldsymbol{a}]_t = 1/T, t \in [T]$.

The self-attention model is a strict generalization of the linear model. Recalling (2), let us merge the key-query weights $\boldsymbol{W} := \boldsymbol{W}_Q \boldsymbol{W}_K^\top$ (without losing generality) and gather weights into $\boldsymbol{\theta} = (\boldsymbol{U}, \boldsymbol{W})$; We then write it as

$$f^{\text{SATT}}(\boldsymbol{U}, \boldsymbol{W}) = \frac{1}{T}\left\langle \boldsymbol{U}, \phi(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^\top)\boldsymbol{X}\right\rangle. \tag{6}$$

Rather than using a Td dimensional $\boldsymbol{U}$, we will also consider the simpler token-pooling via $\boldsymbol{U} = \mathbb{1}_T \boldsymbol{u}^\top$ for $\boldsymbol{u} \in \mathbb{R}^d$.

The linear-attention model parameterized by $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{q})$ replaces the softmax score transformation in (4) with a linear function and outputs

$$f^{\text{LIN-ATT}}(\boldsymbol{w}, \boldsymbol{q}) = \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{q}/T. \tag{7}$$

2.4. Training

Given $\mathcal{S} = (\boldsymbol{X}_i, y_i)_{i=1}^n$ drawn i.i.d. from $\mathcal{D}$, we solve square-loss empirical risk minimization to obtain $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{w}}, \hat{\boldsymbol{q}})$

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \widehat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) := \frac{1}{2n}\sum_{i=1}^n (y_i - f_{\boldsymbol{\theta}}(\boldsymbol{X}_i))^2. \tag{8}$$

Within our theoretical investigation, we are interested in the following performance criteria for models $f \in \{f^{\text{ATT}}, f^{\text{LIN-ATT}}, f^{\text{SATT}}\}$:

- **Classification error**: For a model $f_{\hat{\boldsymbol{\theta}}}$ this is defined as $\text{ERR}(\hat{\boldsymbol{\theta}}) := \mathbb{P}(y f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{X}) < 0)$.
- **Test risk:** $\mathcal{L}(\hat{\boldsymbol{\theta}}) = \mathbb{E}_{(y,\boldsymbol{X})\sim\mathcal{D}}[(y - f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{X}))^2]$.

2.5. Assumptions and notations

First, we formally state our assumptions on the noisy tokens. The more general condition is that noise is subgaussian and satisfies a mild zero third-moment condition.

Assumption 1.a. *The noise vector $\boldsymbol{z} \sim \mathcal{SN}(\sigma)$ is centered σ-subGaussian, i.e. $\|\boldsymbol{z}\|_{\psi_2} = \sigma$. Moreover, its distribution is symmetric and has zero-third moment, i.e. $\mathbb{E}\left[\boldsymbol{z} \otimes (\boldsymbol{z}^\top \boldsymbol{z})\right] = 0$. Let $\boldsymbol{\Sigma} := \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top]$ denote the noise covariance.*

For some of our results it will be convenient to further assume that noise is Gaussian since this leads to precise formulas that are easily interpretable.

Assumption 1.b. *The noise vector $\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2\mathbb{I})$ is isotropic Gaussian with variance σ^2.*

Second, we require a mild assumption on the correlation between the context $\boldsymbol{q}_\star$ and classifier $\boldsymbol{w}_\star$ to guarantee that pure signal tokens $\boldsymbol{q}_\star + y\boldsymbol{w}_\star$ are correctly classified by the true regressor $\boldsymbol{w}_\star$, i.e. $y\boldsymbol{w}_\star^\top(\boldsymbol{q}_\star + y\boldsymbol{w}_\star) > 0$. For convenience, we denote

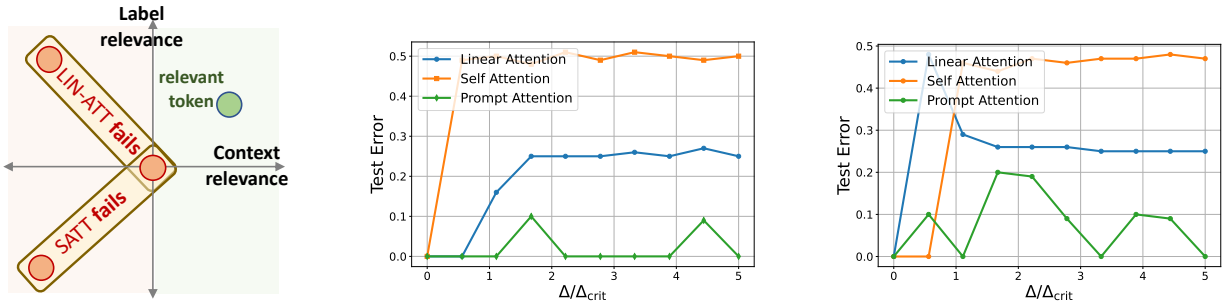
$$W := \|\boldsymbol{w}_\star\|, \quad Q := \|\boldsymbol{q}_\star\|, \quad \rho := \boldsymbol{q}_\star^\top \boldsymbol{w}_\star / (\|\boldsymbol{q}_\star\|\|\boldsymbol{w}_\star\|).$$

Assumption 3.a. *Correlation satisfies $|\rho| < W/Q$.*

We will also often consider the special case of zero correlation ρ and thus state it as separate assumption below. This orthogonality assumption, is useful for more tractable analysis as it helps decouple feature selection and prediction.

Assumption 3.b. *The context and classifier vectors are orthogonal, i.e. $\boldsymbol{q}_\star \perp \boldsymbol{w}_\star$.*

Notation. We use boldface letters for vectors and matrices. $\mathbb{1}_m$ represents an m-dimensional all-ones vector. For a vector $\boldsymbol{v}$, $\|\boldsymbol{v}\|$ denotes its Euclidean norm and $\boldsymbol{v}/\|\boldsymbol{v}\|$ its normalization. $\phi(\cdot)$ denotes the softmax transformation. $\mathcal{Q}(\cdot)$ denotes the tail function of the standard normal distribution. $\wedge$ and $\vee$ denote the minimum and maximum of two numbers, respectively. $\tilde{\mathcal{O}}()$ and $\gtrsim$ notations suppress logarithmic dependencies. Finally, $\propto$ denotes proportionality.



(a) Failure modes **(b)** Varying Δ with $\delta_q = \delta_w$ **(c)** Varying Δ with $\delta_q = -\delta_w$

Figure 1. This figure summarizes and verifies the outcomes of Theorem 1. **Fig (a)** depicts the outcome of our Theorem 1. Relevant token is at position $y\boldsymbol{w}_\star + \boldsymbol{q}_\star$ whereas red tokens (irrelevant) are in positions $-\delta(\boldsymbol{q}_\star \pm y\boldsymbol{w}_\star)$ with $\delta \in \{0, \Delta\}$. **Figures (b) & (c)** plot the performance of our attention models under the contextual dataset with δ equally-likely over $\{0, \Delta\}$ for a synthetic setup (cf. Section 5.1). We set $n = 100$, $d = T = 10$, $\zeta = 0.4$ and train with 100 SGD epochs. We report the median test accuracy over 20 trials. Fig (b) sets $\delta = \delta_q = \delta_w$ and verifies self-attention has $\geq 50\%$ error (for $\Delta \geq \Delta_{\text{crit}} = (1-\zeta)^{-2}$). Fig (c) sets $\delta = \delta_q = -\delta_w$ and verifies linear-prompt-attention has $\geq 25\%$ error (when $\delta \geq \Delta_{\text{crit}} = \sqrt{\zeta/(1-\zeta)}$ in this case).

3. Contrasting prompt-attention to baselines

In this section, we establish separation results between prompt-attention (cf. (4)) and the baselines of self-attention (cf. (6)) and linear attention (cf. (7)). For this, we focus on the discrete dataset model with noiseless irrelevant tokens ($\boldsymbol{z}_t = 0, t \in [T]$).

We first observe that if $\delta = (\delta^q, \delta^w)$ admits a single value, even a linear model can solve the discrete dataset model.

Observation 1 (Linear model solves singleton)**.** *Suppose $(\delta^q, \delta^w) = (\Delta^q, \Delta^w)$ almost surely for $\Delta^q, \Delta^w \in \mathbb{R}$. Set $\boldsymbol{w}'_\star = (\boldsymbol{I} - \bar{\boldsymbol{q}}_\star \bar{\boldsymbol{q}}_\star^\top)\boldsymbol{w}_\star$. As long as $\Delta^w \neq \zeta/(1-\zeta)$ and $\boldsymbol{w}'_\star \neq 0$, $f^{\text{LIN}}(\boldsymbol{w}'_\star)$ or $f^{\text{LIN}}(-\boldsymbol{w}'_\star)$ achieves perfect accuracy.*

Thus, to investigate the expressivity of $f^{\text{ATT}}, f^{\text{SATT}}, f^{\text{LIN}}$, (δ^q, δ^w) would need to admit two or more values. Perhaps surprisingly, we prove that, as soon as, (δ^q, δ^w) comes from a binary distribution, then both f^{SATT} and f^{LIN} can indeed provably fail. Importantly, this happens in the regime $\delta^q \geq 0$ where prompt-attention thrives.

Theorem 1 (Separation of population accuracies)**.** *Consider the discrete dataset model where we set $\boldsymbol{\Sigma} = 0$ in (DATA). The following statements hold:*

*1. **Prompt-attention:** Suppose $\rho^2 < 1$, $\delta^q \geq 0$, and $|\delta^w| \leq C$ almost surely. Define $\boldsymbol{q}'_\star = (\boldsymbol{I} - \bar{\boldsymbol{w}}_\star \bar{\boldsymbol{w}}_\star^\top)\boldsymbol{q}_\star, \boldsymbol{w}'_\star = (\boldsymbol{I} - \bar{\boldsymbol{q}}_\star \bar{\boldsymbol{q}}_\star^\top)\boldsymbol{w}_\star$. For $\Gamma > \frac{\log(C(1/\zeta - 1))}{Q^2(1-\rho^2)}$, choosing $\boldsymbol{\theta} = (\boldsymbol{w}'_\star, \Gamma \boldsymbol{q}'_\star)$, $f^{\text{ATT}}_{\boldsymbol{\theta}}$ achieves perfect classification accuracy on (DATA).*

*2. **Self-attention:** In (DATA), choose (δ^q, δ^w) to be $(0, 0)$ or (Δ, Δ) equally-likely with $\Delta > (1-\zeta)^{-2}$.*

- *For any choice of $(\boldsymbol{U} = \mathbb{1}\boldsymbol{u}^\top, \boldsymbol{W})$, $f^{\text{SATT}}(\mathbb{1}\boldsymbol{u}^\top, \boldsymbol{W})$ achieves 50% accuracy (i.e. random guess).*
- *For any choice of $(\boldsymbol{U}, \boldsymbol{W})$, there exists a (DATA) distribution with adversarial relevance set choices such that $f^{\text{SATT}}(\boldsymbol{U}, \boldsymbol{W})$ achieves 50% accuracy.*

*3. **Linear-attention:** In (DATA), choose (δ^q, δ^w) to be $(0, 0)$ or $(\Delta, -\Delta)$ equally-likely with $\Delta > \sqrt{\zeta/(1-\zeta)}$. For any choice of $(\boldsymbol{w}, \boldsymbol{q})$, $f^{\text{LIN-ATT}}(\boldsymbol{w}, \boldsymbol{q})$ achieves at most 75% accuracy.*

See Fig. 1 for an illustration of the main takeaways from Thm. 1 and a numerical validation of its conclusion on synthetic data. While surprising, the reason prompt-attention can provably beat self-attention is because it is optimized for context-retrieval and can *attend* perfectly on the relevant contextual information. In contrast, self-attention scores are fully feature-based; thus, context information is mixed with other features and can be lost during aggregation of the output. Also note that all results, with the exception of self-attention for general $\boldsymbol{U}$, hold for arbitrary choices of the relevance sets (including adversarial ones). The reason is that tokens are pooled and the particular choice of $\mathcal{R}$ does not matter. Only for $f^{\text{SATT}}(\boldsymbol{U}, \boldsymbol{W})$ we need to adapt the relevance set $\mathcal{R}$ to the output layer $\boldsymbol{U}$ (as well as (y, δ) variables) to promote misclassification. Otherwise, with the hindsight knowledge of the relevance set, $\boldsymbol{U}$ can intelligently process individual tokens of the self-attention output to filter out "confusing" tokens. In fact, for the same failure dataset model, self-attention can achieve perfect accuracy by choosing $\boldsymbol{U} = \mathbb{1}_{\mathcal{R}} \boldsymbol{w}'^\top_\star$ where $\mathbb{1}_{\mathcal{R}}$ is the vector of ones over the (known!) relevance set $\mathcal{R}$ (see Lemma 17). However, this is of course only known in hindsight.

4. Gradient-based analysis of prompt-attention

This section investigates how gradient-descent optimization of the prompt-attention model learns (DATA). Concretely, it shows that a few gradient steps can provably attend to the context-relevant tokens leading to high-classification accuracy. Our results capture requirements on sample complexity in terms of all problem parameters, i.e. dimension d, correlation ρ, context / signal energies Q / W, number of tokens T, and sparsity ζ. This allows studying tradeoffs in different regimes.

Our analysis in this section concerns the prompt-attention model $f_{\boldsymbol{\theta}}^{\mathrm{ATT}}$, so we simply write f_θ. Also, without any further explicit reference, we focus on the *core dataset model*, i.e. (DATA) with $\delta = (0,0)$. All our results here hold under the mild noise and correlation assumptions: Assumption 1.a and Assumption 3.a (we will not further state these). Finally, for simplicity of presentation we assume here isotropic noise $\boldsymbol{\Sigma} = \sigma^2\mathbb{I}$ and handle the general case in the appendix.

4.1. Gradient-based algorithm

For data generated from (DATA), we show the three-step gradient-based algorithm described below achieves high test accuracy. Our analysis also explains why three appropriately chosen steps suffice.

Algorithm: We split the train set in three separate subsets $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ of size n each. Starting from $\boldsymbol{w}_0 = 0, \boldsymbol{q}_0 = 0$, the algorithm proceeds in three gradient steps for step sizes $\eta > 0$ and $\gamma > 0$ and a final debiasing step as follows:

$$\widehat{\boldsymbol{w}}_1 := -\eta \nabla_{\boldsymbol{w}} \widehat{\mathcal{L}}_{\mathcal{S}_1}(0,0), \tag{9a}$$

$$\widehat{\boldsymbol{q}}_1 := -\gamma \nabla_{\boldsymbol{q}} \widehat{\mathcal{L}}_{\mathcal{S}_2}(0,\widehat{\boldsymbol{w}}_1), \tag{9b}$$

$$\widehat{\boldsymbol{w}}_2 := -\eta \nabla_{\boldsymbol{w}} \widehat{\mathcal{L}}_{\mathcal{S}_3}(\widehat{\boldsymbol{q}}_1,\widehat{\boldsymbol{w}}_1), \tag{9c}$$

where $\widehat{\mathcal{L}}_{\mathcal{S}_j}, j = 1,2,3$ is the loss in (8) evaluated on sets $\mathcal{S}_j$. The debiasing step is defined in Section 4.3.

4.2. Population analysis

To gain intuition we first present results on the population counterpart of the algorithm, i.e., substituting $\widehat{\mathcal{L}}(\boldsymbol{w},\boldsymbol{q})$ with its population version $\mathcal{L}(\boldsymbol{w},\boldsymbol{q}) = \mathbb{E}\big[\widehat{\mathcal{L}}(\boldsymbol{w},\boldsymbol{q})\big]$ in all three steps in (9). It is convenient to introduce the following shorthand notation for the negative gradient steps $\boldsymbol{G}_{\boldsymbol{q}}(\boldsymbol{q},\boldsymbol{w}) := -\nabla_{\boldsymbol{q}}\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}}\big[(y - f_{\boldsymbol{\theta}}(\boldsymbol{X}))\nabla_{\boldsymbol{q}} f_{\boldsymbol{\theta}}(\boldsymbol{X})\big]$ and $\boldsymbol{G}_{\boldsymbol{w}}(\boldsymbol{q},\boldsymbol{w}) := -\nabla_{\boldsymbol{w}}\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}}\big[(y - f_{\boldsymbol{\theta}}(\boldsymbol{X}))\nabla_{\boldsymbol{w}} f_{\boldsymbol{\theta}}(\boldsymbol{X})\big]$.

The first gradient step (cf. (9a)) is easy to calculate and returns a classifier estimate that is already in the direction of $\boldsymbol{w}_\star$.

Lemma 1 (Population first step)**.** *The first population gradient step $\boldsymbol{w}_1 = \eta\boldsymbol{G}_{\boldsymbol{w}}(0,0)$ satisfies $\boldsymbol{w}_1 = \eta\zeta\boldsymbol{w}_\star$ since under* (DATA)*, $\boldsymbol{G}_{\boldsymbol{w}}(0,0) = \mathbb{E}_{(\boldsymbol{X},y)}[y\boldsymbol{X}^\top\mathbb{1}]/T = \zeta\boldsymbol{w}_\star$.*

The second gradient step $\boldsymbol{q}_1 = \gamma\boldsymbol{G}_{\boldsymbol{w}}(\boldsymbol{w}_1,0)$ is more intricate: unless $\boldsymbol{q}_\star \perp \boldsymbol{w}_\star$, $\boldsymbol{q}_1$ also has nonzero components in both directions $\boldsymbol{q}_\star$ and $\boldsymbol{w}_\star$.

Lemma 2 (Population second step)**.** *The second population gradient step $\boldsymbol{q}_1 = \gamma\boldsymbol{G}_{\boldsymbol{w}}(\boldsymbol{w}_1,0)$ satisfies for $\alpha := \eta\zeta$*

$$\boldsymbol{q}_1 = \gamma\alpha W^2(\zeta-\zeta^2)\big(1+\frac{\alpha\sigma^2}{T} - \alpha\zeta(W^2+\rho^2Q^2)\big)\boldsymbol{q}_\star \tag{10}$$
$$+\gamma\alpha\rho QW(\zeta-\zeta^2)\big(1-2\zeta\alpha W^2 - (1+\frac{2}{T})\alpha\sigma^2\big)\boldsymbol{w}_\star\,.$$

Proof. Since this computation involves several terms, we defer complete proof to Appendix D.1. The above simplification is made possible by leveraging the assumption on the third-moment of noise (cf. Assumption 1.a). □

Lemma 2 highlights the following key aspects: (i) As mentioned, $\boldsymbol{q}_1$ also picks up the $\boldsymbol{w}_\star$ direction unless $\rho = 0$. However, we can control the magnitude of this undesired term by choosing small step-size η (see Cor. 2). (ii) As αW^2 grows, the gradient component in the $\boldsymbol{q}_\star$ direction might end up pointing in the direction of $-\boldsymbol{q}_\star$. This is because large signal along the $\boldsymbol{w}_\star$ direction might still allow to predict ± 1 label. However, this can always be avoided by choosing sufficiently small step-size η (see Cor. 2). (iii) Similarly, as the noise strength σ^2 grows, gradient in the $\boldsymbol{q}_\star$ direction grows as well. This is because, going along $\boldsymbol{q}_\star$ direction attenuates the noise and cleans up the prediction. (iv) Finally, as $\zeta \to 1$ and $\zeta - \zeta^2 \to 0$ the magnitude of the gradient decays because all tokens contain signal information and there is no need for $\boldsymbol{q}_\star$.

To see how $\boldsymbol{q}_1$ selects good tokens, we investigate the relevance scores (normalized by the step size γ) $r_t := \boldsymbol{x}_t^\top\boldsymbol{q}_1/\gamma$ of relevant vs irrelevant tokens. Attending to context-relevant tokens requires their relevance scores to be larger than those of the noisy ones. Concretely, suppose we have

$$B := \min_{t\in\mathcal{R}}\{r_t = \frac{(\boldsymbol{q}_\star + y\boldsymbol{w}_\star^\top)\boldsymbol{q}_1}{\gamma}\} \geq 2\max_{t\in\mathcal{R}^c}\{r_t = \frac{\boldsymbol{z}_t^\top\boldsymbol{q}_1}{\gamma}\}. \tag{11}$$

Note above that the relevance scores are the same for each $t \in \mathcal{R}$. Thus, $|\mathcal{R}|\,e^{\gamma B} + |\mathcal{R}^c|\,e^{\gamma B/2} \geq S := \sum_{t'\in[T]} e^{\gamma r_{t'}} \geq |\mathcal{R}|\,e^{\gamma B}$, which implies the following for the attention weights as step size increases $\gamma \to \infty$:

$$a_t = [\phi(\boldsymbol{X}\boldsymbol{q}_1)]_t = e^{\gamma r_t}/S \begin{cases} = \frac{e^{\gamma B}}{S} \to \frac{1}{\zeta T} & t\in\mathcal{R} \\ \leq \frac{e^{\gamma B/2}}{S} \to 0 & t\in\mathcal{R}^c \end{cases}. \tag{12}$$

Provided (11) holds, a large enough second gradient step (i.e. large γ) finds $\boldsymbol{q}_1$ that attends (nearly) perfectly to context-relevant tokens in $\mathcal{R}$ and attenuates (almost) all irrelevant tokens in $\mathcal{R}^c$. The following theorem formalizes the above intuition. We defer the complete proof to Appendix D.3.

Theorem 2 (Main theorem: Population)**.** *Consider the model $\boldsymbol{\theta}^\gamma = (\boldsymbol{w}_2^\gamma, \boldsymbol{q}_1^\gamma)$ where $\boldsymbol{q}_1^\gamma = \gamma \boldsymbol{G}_{\boldsymbol{q}}(\boldsymbol{w}_1, 0)$, $\boldsymbol{w}_2^\gamma = \boldsymbol{G}_{\boldsymbol{w}}(0, \boldsymbol{q}_1^\gamma)$ and $\boldsymbol{w}_1 = \eta \boldsymbol{G}_{\boldsymbol{w}}(0,0)$ for step-size η small enough (see Eq. (66) for details). Then, there exists an absolute constant $c > 0$, sufficiently large context strength Q and step-size $\gamma > 0$ such that*

$$\mathrm{ERR}\left(f_{\boldsymbol{\theta}^\gamma}\right) \le 2Te^{-c\frac{\alpha^2 Q^2}{\sigma^2}},$$

provided

$$\frac{(1-\rho^2/2)Q - 2|\rho|W}{\sqrt{1+3\rho^2}} \ge \alpha Q. \tag{13}$$

Eq. (13) guarantees the desired condition (11) holds. When $\rho = 0$ ($\boldsymbol{q}_\star \perp \boldsymbol{w}_\star$), α can be as large as 1 in (13) in which case the rate is $2Te^{-cQ^2/\sigma^2}$. For $\rho \neq 0$, (13) imposes $|\rho| < \frac{Q}{2W}$, in which the role of Q, W is reversed compared to $|\rho| \le \frac{W}{Q}$ in Assumption 3.a: the latter guarantees classifier energy is larger so that signal $y\boldsymbol{w}_\star$ dominates $\boldsymbol{q}_\star$, while for prompt-attention to attend to relevant tokens it is favorable that energy of $\boldsymbol{q}_\star$ dominates $\boldsymbol{w}_\star$. Finally, we compare the theorem's error to the error $Q(\sqrt{\frac{\zeta^2 W^2 T}{1-\zeta}})$ of the linear model in Fact A.1. For concreteness, consider a setting of extreme sparsity $\zeta = O(1/\sqrt{T})$ and $W = O(1)$. Then the error of linear model is $O(1)$, while the (population) algorithm in (9) for prompt-attention achieves an error of $e^{-O(Q^2)}$, which is exponentially decreasing in Q.

4.3. Finite-sample analysis

Here, we investigate the behavior of the algorithm in (9) with finite sample-size n. For convenience, we first introduce an additional de-biasing step after calculating the three gradients in (9). Specifically, for a sample $\mathcal{S}_4$ of size n we compute a bias variable $\widehat{b} := \frac{1}{n}\sum_{i=1}^n f_{(\widehat{\boldsymbol{q}}_1, \widehat{\boldsymbol{w}}_2)}(\boldsymbol{X}_i)$, and use it to de-bias the model's prediction by outputting $f_{(\boldsymbol{\theta},b)}(\boldsymbol{X}) := f_{\boldsymbol{\theta}}(\boldsymbol{X}) - b$. While this extra step is not necessary, it simplifies the statement of our results. Intuitively, $\widehat{b}$ helps with adjusting the decision boundary by removing contributions of the context vector in the final prediction (the context vector is useful only for token-selection rather than final prediction).

Below we provide a simplified version of our main result where noise variance $\sigma \propto 1$ and $\gtrsim$ subsumes constants. Refer to Theorem 6 in the appendix for precise details.

Theorem 3 (Main theorem: Finite-sample)**.** *Suppose Q, W and ρ are such that there exists $\alpha \in (0, 3/16)$ for which*

$$\left(3/16 - \rho^2/8\right)Q - (9/4|\rho| + 1/16)W \ge \alpha \cdot Q \gtrsim \sqrt{\log(nT)}.$$

Fix any $\epsilon > 0$. For sufficiently small step-size $\eta \lesssim Q^{-2}$, sufficiently large step-size $\gamma = \gamma(\epsilon)$, and

$$n \gtrsim d(Q/\zeta W)^4 \log^5(nd),$$

the following statements hold with high probability (see Eq. (63)) over the training set:

1. Prompt attends to relevant tokens. *Concretely, for any fresh sample $(\boldsymbol{X}, y)$, with probability at least $1 - 2Te^{-c\alpha^2 Q^2}$, the attention coefficients $a_t = [\phi(\boldsymbol{X}\widehat{\boldsymbol{q}}_1)]_t$ satisfy:*

$$a_t \begin{cases} \ge \frac{1-\epsilon}{\zeta T} & , t \in \mathcal{R}, \\ \le \frac{\epsilon}{(1-\zeta)T} & , t \notin \mathcal{R}. \end{cases}$$

2. Prompt-attention learns relevant features. *Concretely, for some absolute constant $c > 0$,*

$$\mathbb{P}_{(\boldsymbol{X},y)}\left(\|\boldsymbol{X}^\top \phi(\boldsymbol{X}\widehat{\boldsymbol{q}}_1) - (\boldsymbol{q}_\star + y\boldsymbol{w}_\star)\| < \epsilon\right) \le 2Te^{-c\alpha^2 Q^2}.$$

3. The test error *of the model $f'_{\boldsymbol{\theta}}$ similarly satisfies*

$$\mathrm{ERR}(f'_{\boldsymbol{\theta}}) \le 2Te^{-c\alpha^2 Q^2}.$$

Assuming small correlation coefficient ρ and $W < Q$, we can set $\alpha = \mathcal{O}(1)$. Then, similar to Theorem 2, prompt attention achieves a test error rate of $e^{-\mathcal{O}(Q^2)}$ which is a strict improvement over the linear baseline of Fact A.1 whenever $Q^2 \gtrsim \zeta^2 W^2 T$. Secondly, our bound achieves a sample complexity of $n \gtrsim d(Q/\zeta W)^4$. The linear growth in d is intuitive from counting degrees of freedom. Interestingly, large Q does improve the test error, however, it degrades sample complexity. This is because it makes the estimation of parameters more challenging. Finally, larger ζW improves sample complexity since ζW (combining sparsity level and magnitude) captures the strength of the label-relevant information within relevant tokens $t \in \mathcal{R}$.

Sharp error rates: Finally, in Appendix A we provide an exact analysis of the classification error when $\boldsymbol{q}_\star$ is known and only $\boldsymbol{w}_\star$ is estimated from data. This analysis exactly quantifies the value of context-information and how prompt-tuning retrieves it. Specifically, we prove a sharp asymptotic error rate of $\mathcal{Q}\left(\frac{e^{Q^2/4}}{\sqrt{1+\mathtt{ISNR}(n/d)}} \cdot \sqrt{\frac{\zeta^2 W^2 T}{1-\zeta}}\right)$ where $\mathtt{ISNR}(\alpha) := \alpha^{-1}\frac{(1-\zeta)e^{-Q^2/2}}{\mathtt{rate}_{\mathtt{LIN}}}$, This uniformly (for all $Q \ge 0$ values) improves the optimal rates for (context-free) Gaussian mixture models thanks to the context information.

5. Experiments

First, we verify the utility of prompt-attention via experiments on a synthetic setting that precisely follows the contextual data model from Section 2.2. Subsequently, we explore prompt-tuning on various image classification tasks that are motivated by the contextual data model and compare it with the standard fine-tuning method. Finally, we validate the utility of prompt vectors in distinguishing relevant tokens from irrelevant tokens via prompt-attention under an image classification setting.

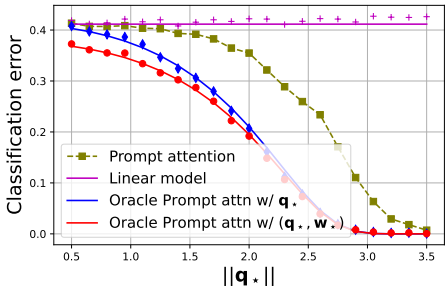


Figure 2. Performance of prompt-attention on the synthetic setting described in Section 5.1. For prompt-attention, we employ the algorithm in (9) to obtain $\hat{\boldsymbol{q}}$ and $\hat{\boldsymbol{w}}$. For the baseline linear model and two oracle settings, we have closed-form expressions for their asymptotic test error (cf. Theorem 1), which are depicted by solid lines. On the other hand, markers show the finite sample performance of these three methods. All finite sample performances are obtained by averaging over 20 independent runs.

5.1. Synthetic experiments

Here, we generate a synthetic dataset according to the *core dataset model*, i.e. we have $\delta = (\delta^q, \delta^w) = (0, 0)$ for all examples in the dataset. In particular, we consider a setting with $T = 500$, $d = 50$, and $\zeta = 0.1$, i.e. each example has 500 tokens out of which 10% tokens are relevant. As for the noisy tokens, they consists of i.i.d. $\mathcal{N}(0, \mathbf{I})$ vectors. Assuming that $\boldsymbol{q}_\star \perp \boldsymbol{w}_\star$ and $\sqrt{T}W = 3$, we generate $n = 10 \cdot d$ training examples from the core dataset model for varying Q. Fig. 2 showcases the performance of prompt-attention (cf. (4)) when combined with the estimates $\hat{\boldsymbol{q}}$ and $\hat{\boldsymbol{w}}$ produced by gradient-based algorithm in (9). We also showcase the performance of the linear model (cf. (5)) and two oracle methods where we assume access to true $\boldsymbol{q}_\star$ and true $(\boldsymbol{q}_\star, \boldsymbol{w}_\star)$, respectively, while applying the prompt-attention. Note that prompt-attention achieves a vanishing classification test error in this setting whereas a natural baseline (linear model) can fail to achieve a good performance. On the other hand, the prompt-attention enabled by (9) successfully achieves a high accuracy as the context energy (defined by Q) increases, validating the utility of prompt-attention as well as our gradient-based algorithm in (9). Finally, we also consider a stochastic $\delta = (\delta^q, \delta^w) \neq (0, 0)$ to validate Theorem 1 as described in Fig. 1.

5.2. Image classification experiments

Dataset. Motivated by our contextual data model, we construct three datasets based on CIFAR-10 (Krizhevsky et al., 2009) to conduct our evaluation (see Fig. 3 for examples). Due to spaces constraints, we defer the additional details regarding datasets to Appendix H.1. We also refer the reader to Appendix H.1 for details regarding the model architecture and training procedure.

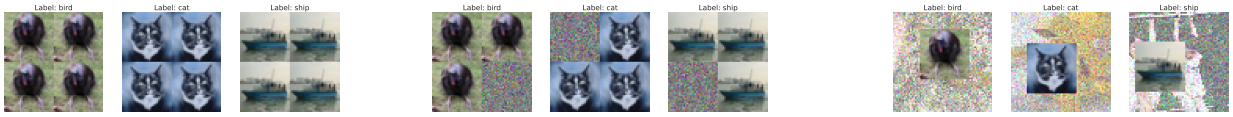
Methods. In our fine-tuning experiments, we update all pre-trained model parameters. As for prompt-tuning, we only update newly introduced (prompt) variables and keep the pre-trained network frozen. We consider three variants of prompt-tuning: 1) PROMPT-TUNING-I (Lester et al., 2021), where we add trainable vector between CLS token embedding and first image (patch) embeddings only at the input; 2) PROMPT-TUNING-II (Li & Liang, 2021), where we add the *same* trainable vectors between the CLS embedding and the first image embedding at the input of every transformer layer; and 3) PROMPT-TUNING-III, where we add *different* trainable vectors between the CLS embedding and the first image patch embedding at the input of every transformer layer. Note that the number of trainable parameters in PROMPT-TUNING-I and PROMPT-TUNING-II do not scale with the number of layers whereas we have linear scaling with number of layers in PROMPT-TUNING-III. Interestingly, all three prompt-tuning variant are identical when the number of layers is 1, which corresponds to the setup we theoretically analyzed in the paper. However, they exhibit remarkably different behavior for a multi-layer transformer model, as we show next. (In Appendix H.2, we also compare prompt-tuning with fine-tuning only first layer self-attention parameters for the underlying ViT model as per the single-layer nature of our theoretical results.)

Results. Here, the main goal of our exploration is to highlight the different behavior of fine-tuning and prompt-tuning. We utilize a model trained on FULL-TILED dataset as the pre-trained model. This model achieves top-1 (in-domain) accuracy of 80.43 on FULL-TILED test set. In contrast, it achieves *zero-shot* top-1 test accuracy of 56.35 and 17.97 on PARTIAL-TILED and EMBED-IN-IMAGENET, respectively. This alludes to the fact that EMBED-IN-IMAGENET corresponds to a larger distribution shift from the pre-training distribution (FULL-TILED), as compared to PARTIAL-TILED.

Fig. 4 and Fig. 5 (cf. Appendix H.2) showcase the performance of fine-tuning and prompt-tuning approaches on EMBED-IN-IMAGENET and PARTIAL-TILED, respectively. Note that fine-tuning outperforms prompt-tuning in a data-rich setting (cf. Fig. 4a and 5a). This is due to fine-tuning having the ability to update a large number of model parameters (5.4M in our case). In contrast, with 2000 prompt vectors, PROMPT-TUNING-III (the most expensive prompt-tuning method out of all three) only updates 460.8K parameters.

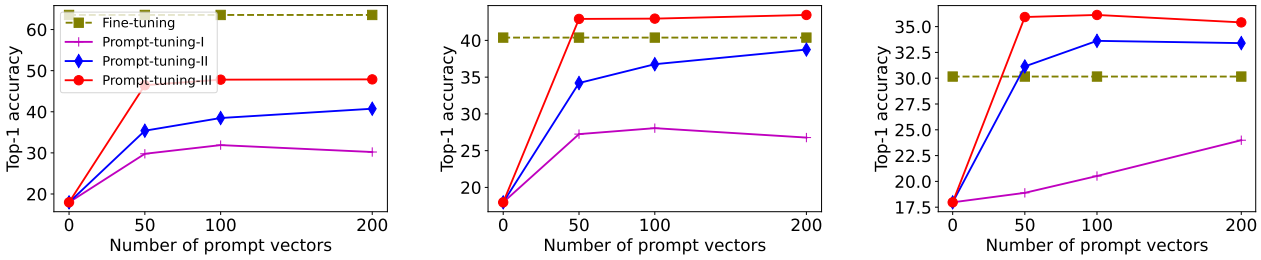
Interestingly, in the data-limited regimes, prompt-tuning becomes more competitive. In fact, the best performing prompt-tuning method outperforms the fine-tuning (cf. Fig. 4b and 4c) on EMBED-IN-IMAGENET, where fine-tuning can easily overfit as it cannot leverage the benefits of the pre-training data due to a large distribution-shift between FULL-TILED and EMBED-IN-IMAGENET.

Part of the performance gap between PROMPT-TUNING-III and PROMPT-TUNING-II can be attributed to the larger num-



(a) FULL-TILED **(b)** PARTIAL-TILED **(c)** EMBED-IN-IMAGENET

Figure 3. Illustration of different CIFAR-10 based datasets utilized in image classification experiments (cf. Section 5.2). Note that all three variants correspond to 10-way multiclass classification tasks corresponding to 10 original classes in CIFAR-10.



(a) Full dataset **(b)** Capped 10% **(c)** Capped 2%

Figure 4. Performance of fine-tuning vs. prompt-tuning on 10-way classification tasks defined by EMBED-IN-IMAGENET dataset. Full dataset has 50K training examples. Capped 10% and 2% correspond to sub-sampled *train* sets where we select exactly 500 and 100 examples per class from the full dataset. Note that number of prompt vectors equal to 0 corresponds to *zero-shot* performance.

ber of trainable parameters available to PROMPT-TUNING-III. Note that PROMPT-TUNING-II consistently outperforms PROMPT-TUNING-I, even with the same number of trainable parameters. This alludes to the fact that optimization and architecture choices play a major role beyond just the number of trainable parameters. As mentioned earlier, our theoretical treatment for a single-layer model cannot distinguish among these different prompt-tuning approaches. As a result, we believe that our empirical observations point to multiple interesting avenues for future work.

5.3. Attention weights for prompt vectors

Finally, we explore what role prompt-attention, i.e. the attention weights with prompt vectors as keys and image patches/tokens as values, plays towards underlying task. In Fig. 6 (cf. Appendix H.2), we illustrate one representative example. The figure shows how average attention weights from prompt vectors to image tokens/patches evolve across transformer layers, when we employ PROMPT-TUNING-III. Indeed, the figure verifies that prompt-attention helps distinguish the relevant tokens/patches from the irrelevant patches, validating our starting hypothesis in Section 2.1 and 2.2.

6. Discussion

In light of remarkable success of attention architectures, we initiate a theoretical investigation of one of its core mechanisms: prompt-tuning. For a one-layer attention model, we motivate and analyze a simplified model for prompt-tuning, which we coin prompt-attention. Through this model, we developed new statistical foundations for gradient-based prompt-tuning, characterized its optimization and generalization dynamics, and explored how it facilitates attending to context-relevant information. We also showed that under (DATA) one-layer softmax-prompt-tuning can provably be superior to alternatives including one layer self-attention. Thorough experiments support our theory on how prompt-tuning attends to the context and how it can potentially outperform full fine-tuning. Our results also suggest many interesting future directions including 1) extension to deeper architectures by characterizing the role of softmax-attention in individual layers, 2) developing a stronger theoretical and empirical understanding of when/if prompt-tuning is superior to fine-tuning, 3) extending our model to include multiple prompt vectors (and perhaps extending (DATA) to include multiple context vectors), and 4) investigating the role of multi-head attention in prompt-tuning.

Acknowledgement

We thank the reviewers for their feedback and suggesting additional experiments involving fine-tuning only first layer self-attention weights. SO was supported by the NSF grants CCF-2046816 and CCF-2212426, Google Research Scholar award, and Army Research Office grant W911NF2110312. MS is supported by the Packard Fellowship in Science and Engineering, a Sloan Fellowship in Mathematics, an NSF-CAREER under award #1846369, DARPA Learning with Less Labels (LwLL) and FastNICS programs, and NSF-CIF awards #1813877 and #2008443. CT was partially supported by the NSERC Discovery Grant RGPIN-2021-03677 and by the NSF grant CCF-2009030.

References

https://openai.com/blog/chatgpt/.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Baldi, P. and Vershynin, R. The quarks of attention. *arXiv preprint arXiv:2202.08371*, 2022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Cao, Y., Chen, Z., Belkin, M., and Gu, Q. Benign overfitting in two-layer convolutional neural networks. *arXiv preprint arXiv:2202.06526*, 2022.

Chan, W., Jaitly, N., Le, Q. V., and Vinyals, O. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*, 2015.

Dehghani, M., Gritsenko, A., Arnab, A., Minderer, M., and Tay, Y. Scenic: A jax library for computer vision research and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21393–21398, 2022.

Dong, Y., Cordonnier, J.-B., and Loukas, A. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pp. 2793–2803. PMLR, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Edelman, B. L., Goel, S., Kakade, S., and Zhang, C. Inductive biases and variable creation in self-attention mechanisms. *arXiv preprint arXiv:2110.10090*, 2021.

Ergen, T., Neyshabur, B., and Mehta, H. Convexifying transformers: Improving optimization and understanding of transformer networks. *arXiv preprint arXiv:2211.11052*, 2022.

Frei, S., Chatterji, N. S., and Bartlett, P. L. Random feature amplification: Feature learning and generalization in neural networks. *arXiv preprint arXiv:2202.07626*, 2022.

Jelassi, S., Sander, M. E., and Li, Y. Vision transformers provably learn spatial structure. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=eMW9AkXaREI.

Karp, S., Winston, E., Li, Y., and Singh, A. Local signal adaptivity: Provable feature learning in neural networks beyond kernels. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24883–24897. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/d064bf1ad039ff366564f352226e7640-Paper.pdf.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Li, H., Wang, M., Liu, S., and Chen, P.-Y. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=jClGv3Qjhb.

Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

Mohammadi, H., Zare, A., Soltanolkotabi, M., and Jovanović, M. R. Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem. *arXiv preprint arXiv:1912.11899*, 2019.

Narayanan, D., Shoeybi, M., Casper, J., LeGresley, P., Patwary, M., Korthikanti, V., Vainbrand, D., Kashinkunti, P., Bernauer, J., Catanzaro, B., et al. Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–15, 2021.

Oymak, S. Stochastic gradient descent learns state equations with nonlinear activations. In *conference on Learning Theory*, pp. 2551–2579. PMLR, 2019.

Pollard, D. Empirical processes: theory and applications. Ims, 1990.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J. T., et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Sahiner, A., Ergen, T., Ozturkler, B., Pauly, J., Mardani, M., and Pilanci, M. Unraveling attention via convex duality: Analysis and interpretations of vision transformers. *International Conference on Machine Learning*, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057. PMLR, 2015.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S., and Kumar, S. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ByxRM0Ntvr.

A. Sharp characterization of accuracy under known context

While the discrete dataset model is insightful, incorporating noise is crucial for understanding the fundamental limits of the benefits of context in attention. To this end, let us focus on the core dataset model where we set $\delta^q = \delta^w = 0$ and explore the role of noise in population accuracy. Also assume that noise is standard normal, i.e. Assumption 1.b.

• *Linear model.* The linear model aggregates tokens to obtain a simple Gaussian mixture distribution. Specifically, aggregated tokens are exactly distributed as $\frac{1}{T}\boldsymbol{X}^\top \mathbb{1}_T \sim \mathcal{N}(\zeta \boldsymbol{w}_\star, \frac{1-\zeta}{T}\boldsymbol{I})$, leading to the following well-known result.

Fact A.1. *For linear models, optimal accuracy obeys* $\min_{\boldsymbol{w}} \mathrm{ERR}(f^{\mathrm{LIN}}(\boldsymbol{w})) = Q(\sqrt{\frac{\zeta^2 W^2 T}{1-\zeta}})$ *where* $Q(\cdot)$ *is the tail function of the standard normal distribution.*

• *Prompt-attention model.* Since prompt-attention strictly generalizes the linear model, its accuracy is at least as good. The theorem below quantifies this and demonstrates how context vector can enable an optimal weighting of relevant and irrelevant tokens to maximize accuracy. A general version of this theorem is proven under a non-asymptotic setting (finite T, d) as Theorem 8.

Theorem 4. *Consider the prompt-attention model* $f^{\mathrm{ATT}}_{\boldsymbol{\theta}}$. *Suppose* $\boldsymbol{w}_\star \perp \boldsymbol{q}_\star$ *and let* $\tau, \bar{\tau} > 0$ *be hyperparameters. Consider the following algorithm which uses the hindsight knowledge of* $\boldsymbol{q}_\star$ *to estimate* $\boldsymbol{w}_\star$ *and make prediction:*

Set $\hat{\boldsymbol{w}} = (\boldsymbol{I} - \bar{\boldsymbol{q}}_\star \bar{\boldsymbol{q}}_\star)\nabla \mathcal{L}_{\boldsymbol{w}}(0, \tau \bar{\boldsymbol{q}}_\star)$ *and* $\boldsymbol{\theta} = (\hat{\boldsymbol{w}}, \bar{\tau}\bar{\boldsymbol{q}}_\star)$. *Suppose* $\zeta^2 W^2 T, 1-\zeta, \alpha := n/d, e^{Q^2}, e^{\tau}$ *each lie between two positive absolute constants. Suppose* T *is polynomially large in* n *and these constants and* $\tilde{\mathcal{O}}(\cdot)$ *hides polynomial terms in* n. *Define* inverse-signal-to-noise-ratio*:* $\mathtt{ISNR}(\alpha, \tau) = \frac{(1-\zeta)e^{2\tau(\tau - Q)}}{\alpha \zeta^2 W^2 T}$. *In the limit* $T, d \to \infty$, *the test error converges in probability to* $\mathcal{Q}\left(\frac{e^{Q\bar{\tau} - \bar{\tau}^2}}{\sqrt{1+\mathtt{ISNR}(\alpha,\tau)}} \cdot \sqrt{\frac{\zeta^2 W^2 T}{1-\zeta}}\right)$. *In this limit, optimal hyperparameters are* $\tau = \bar{\tau} = Q/2$ *and leads to optimal* $\mathtt{ISNR}(\alpha) := \frac{(1-\zeta)e^{-Q^2/2}}{\alpha \zeta^2 W^2 T}$ *and the error*

$$\mathrm{ERR}(\alpha, Q, W) = \mathcal{Q}\left(\frac{e^{Q^2/4}}{\sqrt{1 + \mathtt{ISNR}(\alpha)}} \cdot \sqrt{\frac{\zeta^2 W^2 T}{1-\zeta}}\right)$$

Here, a few remarks are in place. Note that $\mathtt{rate}_{\mathrm{LIN}} := \sqrt{\frac{\zeta^2 W^2 T}{1-\zeta}}$ term is the population error rate of f^{LIN} from Fact A.1. In the limit $\alpha = n/d \to \infty$, the rate of f^{ATT} is simply given by $e^{Q/4}\mathtt{rate}_{\mathrm{LIN}}$ demonstrating the strict superiority of prompt-attention. Moreover setting $Q = 0$ (no prompt info), since feature-output of f^{LIN} (i.e. $\boldsymbol{X}^\top \phi(\boldsymbol{X}\mathbb{1})$) is (essentially) a binary Gaussian mixture distribution, our error-rate recovers the Bayes-optimal f^{LIN} classifier which has a finite-sample rate of $\mathtt{rate}_{\mathrm{LIN}}/\sqrt{1 + (1-\zeta)/(\alpha \zeta^2 W^2 T^2)}$. Prompt-tuning also strictly improves this because our $\mathtt{ISNR}(\alpha)$ introduces an additional $e^{-Q^2/2}$ multiplier.

B. Gradient calculations and concentration

In this section, we focus on finite-sample analysis of Algorithm 9. Introduce the following shorthand notation analogous to the population counterparts in Section 4.2:

$$\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(\boldsymbol{q}, \boldsymbol{w}) := -\nabla_{\boldsymbol{q}} \widehat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i \in [n]} (y_i - f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)) \boldsymbol{X}_i^\top \phi'(\boldsymbol{X}_i \boldsymbol{q})\, \boldsymbol{X}_i \boldsymbol{w}\,,$$

$$\widehat{\boldsymbol{G}}_{\boldsymbol{w}}(\boldsymbol{q}, \boldsymbol{w}) := -\nabla_{\boldsymbol{w}} \widehat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i \in [n]} (y_i - f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)) \boldsymbol{X}_i^\top \phi((\boldsymbol{X}_i \boldsymbol{q})). \tag{14}$$

B.1. Gradient Calculations

We begin with the gradient calculations for the first two steps of the algorithm.

For convenience, we make use of the following shorthands

$$R_{\boldsymbol{q}_\star} := R_{\boldsymbol{w}, \boldsymbol{q}_\star} := \boldsymbol{w}^\top \boldsymbol{q}_\star, \qquad R_{\boldsymbol{w}_\star} := R_{\boldsymbol{w}, \boldsymbol{w}_\star} := \boldsymbol{w}^\top \boldsymbol{w}_\star, \qquad \alpha_i := \alpha(\boldsymbol{w}, y_i) := R_{\boldsymbol{q}_\star} + y_i R_{\boldsymbol{w}_\star},$$

$$\boldsymbol{\gamma}_i := \boldsymbol{\gamma}(\boldsymbol{Z}_i) = \frac{1}{T}\boldsymbol{Z}_i^\top \mathbb{1}, \qquad \beta_i := \beta(\boldsymbol{Z}_i; \boldsymbol{w}) = \boldsymbol{\gamma}_i^\top \boldsymbol{w}, \qquad \hat{\boldsymbol{\Sigma}}_i := \frac{1}{T}\boldsymbol{Z}_i^\top \boldsymbol{Z}_i\,,$$

where $\boldsymbol{Z}_i \in \mathbb{R}^{(1-\zeta)T\times d}$ is the matrix of irrelevant tokens $\boldsymbol{z}_{i,t}, t \in \mathcal{R}^c$ for sample $i \in [n]$.

Lemma 3. *Under dataset model* (DATA) *and Assumption 1.a, we have*

$$\widehat{\boldsymbol{G}}_{\boldsymbol{w}}(0,0) = \zeta \boldsymbol{w}_\star + \zeta \boldsymbol{q}_\star \left(\frac{1}{n} \sum_{i\in[n]} y_i \right) + \frac{1}{n} \sum_i y_i \boldsymbol{\gamma}_i. \tag{16}$$

Lemma 4. *Under dataset model* (DATA) *and Assumption 1.a, we have that*

$$\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0,\boldsymbol{w}) := \frac{1}{n} \sum_{i=1}^{n} (y_i - \zeta\alpha_i - \beta_i) \left[\left((\zeta - \zeta^2)\alpha_i - \zeta\beta_i \right) (\boldsymbol{q}_\star + y_i \boldsymbol{w}_\star) + \hat{\boldsymbol{\Sigma}}_i \boldsymbol{w} - (\zeta\alpha_i + \beta_i)\boldsymbol{\gamma}_i \right]. \tag{17}$$

B.1.1. Proof of Lemma 3

By direct computation,

$$\begin{aligned}
\widehat{\boldsymbol{G}}_{\boldsymbol{w}}(0,0) := -\nabla_{\boldsymbol{w}} \widehat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}) &= \frac{1}{nT} \sum_{i\in[n]} y_i \boldsymbol{X}_i^\top \mathbb{1}_T = \frac{1}{nT} \sum_{i\in[n]} \sum_{t\in[T]} y_i \boldsymbol{x}_{i,t} \\
&= \frac{\zeta}{n} \Big(\sum_{i\in[n]} y_i \Big) \boldsymbol{q}_\star + \frac{\zeta}{n} \boldsymbol{w}_\star + \frac{1}{n} \sum_{i\in[n]} y_i \boldsymbol{\gamma}_i .
\end{aligned}$$

B.1.2. Proof of Lemma 4

Note that $\phi'(0) = \frac{1}{T}\mathbb{I} - \frac{1}{T^2}\mathbb{1}\mathbb{1}^\top$; hence, for $\boldsymbol{\theta} = (0, \boldsymbol{w})$:

$$\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0,\boldsymbol{w}) := -\nabla_{\boldsymbol{q}} \widehat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i\in[n]} \underbrace{\frac{1}{T}(y_i - f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)) \boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{w}}_{\text{Term}_{1,\text{i}}} - \frac{1}{n} \sum_{i\in[n]} \underbrace{\frac{1}{T^2}(y_i - f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)) \boldsymbol{X}_i^\top \mathbb{1}\mathbb{1}^\top \boldsymbol{X}_i \boldsymbol{w}}_{\text{Term}_{2,\text{i}}} .$$

Moreover, note that,

$$\begin{aligned}
f_{\boldsymbol{\theta}}(\boldsymbol{X}_i) &= \frac{1}{T} \boldsymbol{w}^\top \boldsymbol{X}_i^\top \mathbb{1}, \\
\boldsymbol{X}_i^\top \boldsymbol{X}_i &= \zeta T (\boldsymbol{q}_\star + y_i \boldsymbol{w}_\star)(\boldsymbol{q}_\star + y_i \boldsymbol{w}_\star)^\top + \boldsymbol{Z}_i^\top \boldsymbol{Z}_i, \\
\boldsymbol{X}_i^\top \mathbb{1} &= \zeta T (\boldsymbol{q}_\star + y_i \boldsymbol{w}_\star) + \boldsymbol{Z}_i^\top \mathbb{1} ,
\end{aligned}$$

where recall the notation in Lemma 3 for $\boldsymbol{Z}_i$. Hence, using the lemma's notation (repeated here for convenience)

$$\alpha_i := R_{\boldsymbol{q}_\star} + y_i R_{\boldsymbol{w}_\star}, \quad \beta_i := \beta(\boldsymbol{Z}_i; \boldsymbol{w}) := \frac{1}{T} \mathbb{1}^\top \boldsymbol{Z}_i \boldsymbol{w}, \quad \boldsymbol{\gamma}_i := \boldsymbol{\gamma}(\boldsymbol{Z}_i) = \frac{1}{T} \boldsymbol{Z}_i^\top \mathbb{1}, \quad \hat{\boldsymbol{\Sigma}}_i := \frac{1}{T} \boldsymbol{Z}_i^\top \boldsymbol{Z}_i.$$

we find that

$$\begin{aligned}
y_i - f_{\boldsymbol{\theta}}(\boldsymbol{X}_i) &= y_i - \zeta\alpha_i - \beta_i \\
\frac{1}{T} \boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{w} &= \zeta\alpha_i \boldsymbol{q}_\star + \zeta y_i \alpha_i \boldsymbol{w}_\star + \hat{\boldsymbol{\Sigma}}_i \boldsymbol{w} \\
\frac{1}{T^2} \boldsymbol{X}_i^\top \mathbb{1}\mathbb{1}^\top \boldsymbol{X}_i \boldsymbol{w} &= \zeta \left(\zeta\alpha_i + \beta_i \right) \boldsymbol{q}_\star + \zeta \left(\zeta\alpha_i + \beta_i \right) y_i \boldsymbol{w}_\star + \left(\zeta\alpha_i + \beta_i \right) \boldsymbol{\gamma}_i.
\end{aligned}$$

With the above, each one of the two terms becomes:

$$\text{Term}_{1,\text{i}} = \left(y_i - \zeta\alpha_i - \beta_i \right) \zeta\alpha_i \boldsymbol{q}_\star + \left(y_i - \zeta\alpha_i - \beta_i \right) \zeta\alpha_i y_i \boldsymbol{w}_\star + \left(y_i - \zeta\alpha_i - \beta_i \right) \hat{\boldsymbol{\Sigma}}_i \boldsymbol{w}$$

$$\text{Term}_{2,\text{i}} = \zeta \left(y_i - \zeta\alpha_i - \beta_i \right) \left(\zeta\alpha_i + \beta_i \right) \boldsymbol{q}_\star + \zeta \left(y_i - \zeta\alpha_i - \beta_i \right) \left(\zeta\alpha_i + \beta_i \right) y_i \boldsymbol{w}_\star + \left(y_i - \zeta\alpha_i - \beta_i \right) \left(\zeta\alpha_i + \beta_i \right) \boldsymbol{\gamma}_i .$$

Combining the above:

$$\text{Term}_{1,\text{i}} - \text{Term}_{2,\text{i}} = \left(y_i - \zeta\alpha_i - \beta_i \right) \left[\zeta \left((1-\zeta)\alpha_i - \beta_i \right) (\boldsymbol{q}_\star + y_i \boldsymbol{w}_\star) + \hat{\boldsymbol{\Sigma}}_i \boldsymbol{w} - (\zeta\alpha_i + \beta_i)\boldsymbol{\gamma}_i \right]. \tag{18}$$

B.2. Concentration of Gradient $\widehat{G}_{\boldsymbol{q}}(0,\boldsymbol{w})$ in the $\boldsymbol{q}$ direction

The main result of this section is the following lemma about concentration of gradient with respect to $\boldsymbol{q}$.

Lemma 5 (Concentration of $\widehat{G}_{\boldsymbol{q}}(0,\boldsymbol{w})$)**.** *Fix any vectors $\boldsymbol{v},\boldsymbol{w}\in\mathbb{R}^d$. For convenience define $R_{\boldsymbol{v},\boldsymbol{q}_\star}:=\boldsymbol{v}^\top\boldsymbol{q}_\star$ and $R_{\boldsymbol{v},\boldsymbol{w}_\star}:=\boldsymbol{v}^\top\boldsymbol{w}_\star$ and recall $R_{\boldsymbol{w}_\star},R_{\boldsymbol{q}_\star}$ notations from Lemma 4. Then, we can decompose*

$$\boldsymbol{v}^\top\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0,\boldsymbol{w})=\boldsymbol{v}^\top\boldsymbol{G}_{\boldsymbol{q}}(0,\boldsymbol{w})+\boldsymbol{v}^\top\widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0,\boldsymbol{w}),$$

where the expectation term is given by

$$\begin{aligned}\boldsymbol{v}^\top\boldsymbol{G}_{\boldsymbol{q}}(0,\boldsymbol{w}):=\mathbb{E}\left[\boldsymbol{v}^\top\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0,\boldsymbol{w})\right]&=\left((\zeta-\zeta^2)\left(R_{\boldsymbol{w}_\star}+\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}/T\right)-(\zeta^2-\zeta^3)\left(R_{\boldsymbol{w}_\star}^2+R_{\boldsymbol{q}_\star}^2\right)\right)R_{\boldsymbol{v},\boldsymbol{q}_\star}\\&\quad+\left(\left((\zeta-\zeta^2)-2(\zeta^2-\zeta^3)R_{\boldsymbol{w}_\star}\right)R_{\boldsymbol{q}_\star}\right)R_{\boldsymbol{v},\boldsymbol{w}_\star}-\left((1+2/T)(\zeta-\zeta^2)R_{\boldsymbol{q}_\star}\right)\boldsymbol{v}^\top\boldsymbol{\Sigma}\boldsymbol{w}\,,\end{aligned}\tag{19}$$

and the deviation term obeys

$$\begin{aligned}\boldsymbol{v}^\top\widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0,\boldsymbol{w})=&\left[\left((\zeta-\zeta^2)-2(\zeta^2-\zeta^3)R_{\boldsymbol{w}_\star}\right)R_{\boldsymbol{q}_\star}R_{\boldsymbol{v},\boldsymbol{q}_\star}+\left((\zeta-\zeta^2)R_{\boldsymbol{w}_\star}-(\zeta^2-\zeta^3)\left(R_{\boldsymbol{w}_\star}^2+R_{\boldsymbol{q}_\star}^2\right)\right)R_{\boldsymbol{v},\boldsymbol{w}_\star}\right]\left(\frac{1}{n}\sum_{i=1}^n y_i\right)\\&+\left[\left(-\zeta+2\zeta^2\right)R_{\boldsymbol{q}_\star}R_{\boldsymbol{v},\boldsymbol{q}_\star}+\left(-\zeta+(-\zeta+2\zeta^2)R_{\boldsymbol{w}_\star}\right)R_{\boldsymbol{v},\boldsymbol{w}_\star}+(1-\zeta)\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{v}\right]\left(\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)\right)\\&\quad+\left[\zeta R_{\boldsymbol{w}_\star}-\zeta^2(R_{\boldsymbol{q}_\star}^2+R_{\boldsymbol{w}_\star}^2)+\frac{(1-\zeta)}{T}\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}\right]\left(\frac{1}{n}\sum_{i=1}^n\boldsymbol{v}^\top\boldsymbol{\gamma}_i\right)\\&\quad+\left[\left(-\zeta+(-\zeta+2\zeta^2)R_{\boldsymbol{w}_\star}\right)R_{\boldsymbol{v},\boldsymbol{q}_\star}+\left(-\zeta+2\zeta^2\right)R_{\boldsymbol{q}_\star}R_{\boldsymbol{v},\boldsymbol{w}_\star}\right]\left(\frac{1}{n}\sum_{i=1}^n y_i\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)\right)\\&\quad+\left[\zeta R_{\boldsymbol{q}_\star}-2\zeta^2R_{\boldsymbol{q}_\star}R_{\boldsymbol{w}_\star}\right]\left(\frac{1}{n}\sum_{i=1}^n y_i\left(\boldsymbol{v}^\top\boldsymbol{\gamma}_i\right)\right)\\&+\zeta R_{\boldsymbol{v},\boldsymbol{q}_\star}\left(\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)^2-\frac{(1-\zeta)}{T}\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}\right)\\&\quad+\zeta R_{\boldsymbol{v},\boldsymbol{w}_\star}\left(\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)^2y_i\right)\\&\quad+\left[1-2\zeta R_{\boldsymbol{w}_\star}\right]\left(\frac{1}{n}\sum_{i=1}^n y_i\left(\boldsymbol{w}^\top\boldsymbol{\gamma}_i\right)\left(\boldsymbol{v}^\top\boldsymbol{\gamma}_i\right)\right)\\&+(1-\zeta R_{\boldsymbol{w}_\star})\left(\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{v}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}\right)\\&\quad-\zeta R_{\boldsymbol{q}_\star}\left(\frac{1}{n}\sum_{i=1}^n\boldsymbol{v}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}-(1-\zeta)\boldsymbol{v}^\top\boldsymbol{\Sigma}\boldsymbol{w}\right)\\&-\left[2\zeta R_{\boldsymbol{q}_\star}\right]\left(\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{w}^\top\boldsymbol{\gamma}_i\right)\left(\boldsymbol{v}^\top\boldsymbol{\gamma}_i\right)-\frac{1-\zeta}{T}\boldsymbol{v}^\top\boldsymbol{\Sigma}\boldsymbol{w}\right)\\&\quad-\left(\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{v}^\top\boldsymbol{\gamma}_i\right)\left(\left(\boldsymbol{w}^\top\boldsymbol{\gamma}_i\right)^2-\frac{(1-\zeta)}{T}\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}\right)\right)\\&-\left(\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)\left(\boldsymbol{v}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}-(1-\zeta)\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{v}\right)\right).\end{aligned}\tag{20}$$

Moreover, all random terms in (20) are zero-mean and concentrate as prescribed by Lemma 6 below.

Lemma 6 (Main concentration lemma)**.** *Let $y_i, i\in[n]$ be iid Rademacher random variables. Let $\boldsymbol{Z}_i\in\mathbb{R}^{(1-\zeta)T\times d}, i\in[n]$ be iid copies of a random matrix $\boldsymbol{Z}$. Each row $\boldsymbol{z}_t, t\in\left[(1-\zeta)T\right]$ of $\boldsymbol{Z}$ is an iid copy of a random vector $\boldsymbol{z}$ satisfying Assumption 1.a. For convenience denote $\boldsymbol{\gamma}_i:=\boldsymbol{Z}_i^\top\mathbb{1}/T$ and $\hat{\boldsymbol{\Sigma}}_i:=\boldsymbol{Z}_i^\top\boldsymbol{Z}_i/T$. Then, the following statements are true for all vectors $\boldsymbol{w},\boldsymbol{v}\in\mathbb{R}^d$:*

$$\left\|\frac{1}{n}\sum_{i\in[n]}y_i\right\|_{\psi_2}\le\frac{C}{\sqrt{n}}$$

$$\left\|\frac{1}{n}\sum_{i\in[n]}\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right\|_{\psi_2}\vee\left\|\frac{1}{n}\sum_{i\in[n]}y_i\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right\|_{\psi_2}\leq\frac{C\sigma\sqrt{1-\zeta}\|\boldsymbol{w}\|_2}{\sqrt{nT}}$$

$$\left\|\frac{1}{n}\sum_{i\in[n]}\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^\top\boldsymbol{v}\right)-\frac{1-\zeta}{T}\boldsymbol{v}^\top\boldsymbol{\Sigma}\boldsymbol{w}\right\|_{\psi_1}\vee\left\|\frac{1}{n}\sum_{i\in[n]}y_i\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^\top\boldsymbol{v}\right)\right\|_{\psi_1}\leq\frac{C\sigma^2(1-\zeta)\|\boldsymbol{w}\|_2\|\boldsymbol{v}\|_2}{T\sqrt{n}}$$

$$\left\|\frac{1}{n}\sum_{i\in[n]}\boldsymbol{w}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v}-(1-\zeta)\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{v}\right\|_{\psi_1}\vee\left\|\frac{1}{n}\sum_{i\in[n]}y_i\boldsymbol{w}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v}\right\|_{\psi_1}\leq\frac{C\sigma^2\sqrt{1-\zeta}\|\boldsymbol{w}\|_2\|\boldsymbol{v}\|_2}{\sqrt{nT}}$$

$$\left\|\frac{1}{n}\sum_{i\in[n]}\left(\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)^2-\frac{1-\zeta}{T}\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}\right)\boldsymbol{\gamma}_i^\top\boldsymbol{v}\right\|_{\psi_{2/3}}\leq\frac{C\sigma^3(1-\zeta)^{3/2}\|\boldsymbol{w}\|_2^2\|\boldsymbol{v}\|_2}{T^{3/2}\sqrt{n}}\log n$$

$$\left\|\frac{1}{n}\sum_{i\in[n]}\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)\left(\boldsymbol{w}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v}-(1-\zeta)\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{v}\right)\right\|_{\psi_{2/3}}\leq\frac{C\sigma^3(1-\zeta)\|\boldsymbol{w}\|_2^2\|\boldsymbol{v}\|_2}{T\sqrt{n}}\log n\,.$$

Also, all the random variables that appear above are zero mean.

B.2.1. Proof of Lemma 5

We split (17) in four terms and handle each of them separately.

- **Term**$_{\mathrm{I}}$ $=\frac{1}{n}\sum_{i=1}^n\left(y_i-\zeta\alpha_i-\beta_i\right)\left((\zeta-\zeta^2)\alpha_i-\zeta\beta_i\right)\boldsymbol{q}_\star$

We first focus on

$$\text{Term}_{\mathrm{I}}=\frac{1}{n}\sum_{i=1}^n\left(y_i(1-\zeta R_{\boldsymbol{w}_\star})-\beta_i-\zeta R_{\boldsymbol{q}_\star}\right)\left(y_i(\zeta-\zeta^2)R_{\boldsymbol{w}_\star}-\zeta\beta_i+(\zeta-\zeta^2)R_{\boldsymbol{q}_\star}\right)\boldsymbol{q}_\star=:A\boldsymbol{q}_\star.$$

We we can express A above conveniently as follows (recall $y_i^2=1$):

$$\begin{aligned}
A\coloneqq&-\zeta(\zeta-\zeta^2)R_{\boldsymbol{q}_\star}^2+(1-\zeta R_{\boldsymbol{w}_\star})(\zeta-\zeta^2)R_{\boldsymbol{w}_\star}+\left((1-\zeta R_{\boldsymbol{w}_\star})(\zeta-\zeta^2)R_{\boldsymbol{q}_\star}-\zeta(\zeta-\zeta^2)R_{\boldsymbol{w}_\star}R_{\boldsymbol{q}_\star}\right)\left(\frac{1}{n}\sum_{i=1}^n y_i\right)\\
&+\left(-(\zeta-\zeta^2)R_{\boldsymbol{q}_\star}+\zeta^2R_{\boldsymbol{q}_\star}\right)\left(\frac{1}{n}\sum_{i=1}^n\beta_i\right)+\left(-(1-\zeta R_{\boldsymbol{w}_\star})\zeta-(\zeta-\zeta^2)R_{\boldsymbol{w}_\star}\right)\left(\frac{1}{n}\sum_{i=1}^n y_i\beta_i\right)+\zeta\left(\frac{1}{n}\sum_{i=1}^n\beta_i^2\right)\\
=&(\zeta-\zeta^2)R_{\boldsymbol{w}_\star}-(\zeta^2-\zeta^3)\left(R_{\boldsymbol{w}_\star}^2+R_{\boldsymbol{q}_\star}^2\right)+\left((\zeta-\zeta^2)-2(\zeta^2-\zeta^3)R_{\boldsymbol{w}_\star}\right)R_{\boldsymbol{q}_\star}\left(\frac{1}{n}\sum_{i=1}^n y_i\right)\\
&+\left(-\zeta+2\zeta^2\right)R_{\boldsymbol{q}_\star}\left(\frac{1}{n}\sum_{i=1}^n\beta_i\right)+\left(-\zeta+(-\zeta+2\zeta^2)R_{\boldsymbol{w}_\star}\right)\left(\frac{1}{n}\sum_{i=1}^n y_i\beta_i\right)\\
&+\zeta\left(\frac{1}{n}\sum_{i=1}^n\beta_i^2-\frac{(1-\zeta)}{T}\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}\right)+\frac{(\zeta-\zeta^2)}{T}\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}.
\end{aligned}$$

From Lemma 6, all random terms above are zero mean. Hence,

$$\begin{aligned}
\mathbb{E}[A]&=-(\zeta^2-\zeta^3)R_{\boldsymbol{q}_\star}^2+(1-\zeta R_{\boldsymbol{w}_\star})(\zeta-\zeta^2)R_{\boldsymbol{w}_\star}+(\zeta-\zeta^2)\frac{\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}}{T}\\
&=-(\zeta^2-\zeta^3)R_{\boldsymbol{q}_\star}^2+(\zeta-\zeta^2)\left(R_{\boldsymbol{w}_\star}+\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}/T\right)-(\zeta^2-\zeta^3)R_{\boldsymbol{w}_\star}^2\\
&=(\zeta-\zeta^2)\left(R_{\boldsymbol{w}_\star}+\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}/T\right)-(\zeta^2-\zeta^3)\left(R_{\boldsymbol{w}_\star}^2+R_{\boldsymbol{q}_\star}^2\right).
\end{aligned}\tag{21}$$

- **Term**$_{\mathrm{II}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \zeta\alpha_i - \beta_i\right)\left((\zeta-\zeta^2)\alpha_i - \zeta\beta_i\right)y_i\boldsymbol{w}_\star$

$$\mathrm{Term}_{\mathrm{II}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i(1-\zeta R_{\boldsymbol{w}_\star}) - \beta_i - \zeta R_{\boldsymbol{q}_\star}\right)\left(y_i(\zeta-\zeta^2)R_{\boldsymbol{w}_\star} - \zeta\beta_i + (\zeta-\zeta^2)R_{\boldsymbol{q}_\star}\right)y_i\boldsymbol{w}_\star = B\boldsymbol{w}_\star.$$

We we can express B above conveniently as:

$$B := \left((\zeta-\zeta^2)R_{\boldsymbol{w}_\star} - (\zeta^2-\zeta^3)\left(R_{\boldsymbol{w}_\star}^2 + R_{\boldsymbol{q}_\star}^2\right)\right)\left(\frac{1}{n}\sum_{i\in[n]} y_i\right) + \left((\zeta-\zeta^2) - 2(\zeta^2-\zeta^3)R_{\boldsymbol{w}_\star}\right)R_{\boldsymbol{q}_\star}$$
$$+ \left(-\zeta + 2\zeta^2\right)R_{\boldsymbol{q}_\star}\left(\frac{1}{n}\sum_{i=1}^{n}\beta_i y_i\right) + \left(-\zeta + (-\zeta+2\zeta^2)R_{\boldsymbol{w}_\star}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\beta_i\right) + \zeta\left(\frac{1}{n}\sum_{i=1}^{n}\beta_i^2 y_i\right).$$

All the random terms above are zero-mean. Hence,

$$\mathbb{E}[B] = \left((\zeta-\zeta^2) - 2(\zeta^2-\zeta^3)R_{\boldsymbol{w}_\star}\right)R_{\boldsymbol{q}_\star}. \tag{22}$$

- **Term**$_{\mathrm{III}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \zeta\alpha_i - \beta_i\right)\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}$

Fix any vector $\boldsymbol{v}$:

$$\begin{aligned}
\boldsymbol{v}^\top\{\mathrm{Term}_{\mathrm{III}}\} &= \left(\frac{1}{n}\sum_{i=1}^{n} y_i\boldsymbol{v}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}\right) - \zeta\left(R_{\boldsymbol{q}_\star} + y_i R_{\boldsymbol{w}_\star}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}\right) - \left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)\boldsymbol{v}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}\right)\\
&= (1-\zeta R_{\boldsymbol{w}_\star})\left(\frac{1}{n}\sum_{i=1}^{n} y_i\boldsymbol{v}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}\right) - \zeta R_{\boldsymbol{q}_\star}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}\right) - \left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)\boldsymbol{v}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}\right)\\
&= (1-\zeta R_{\boldsymbol{w}_\star})\left(\frac{1}{n}\sum_{i=1}^{n} y_i\boldsymbol{v}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w}\right) - \zeta R_{\boldsymbol{q}_\star}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w} - (1-\zeta)\boldsymbol{v}^\top\boldsymbol{\Sigma}\boldsymbol{w}\right) - (\zeta-\zeta^2)R_{\boldsymbol{q}_\star}\boldsymbol{v}^\top\boldsymbol{\Sigma}\boldsymbol{w}\\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)\left(\boldsymbol{v}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{w} - (1-\zeta)\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{v}\right) + (1-\zeta)(\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{v})\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\gamma}_i^\top\boldsymbol{w})
\end{aligned}$$

From Lemma 6 all random terms above are zero mean. Hence,

$$\mathbb{E}\left[\boldsymbol{v}^\top\{\mathrm{Term}_{\mathrm{III}}\}\right] = -(\zeta-\zeta^2)R_{\boldsymbol{q}_\star}\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{v}. \tag{23}$$

- **Term**$_{\mathrm{IV}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \zeta\alpha_i - \beta_i\right)(\zeta\alpha_i + \beta_i)\boldsymbol{\gamma}_i$

For fixed vector $\boldsymbol{v}$, $\boldsymbol{v}^\top\{\mathrm{Term}_{\mathrm{IV}}\} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \zeta\alpha_i - \beta_i\right)(\zeta\alpha_i + \beta_i)\boldsymbol{v}^\top\boldsymbol{\gamma}_i$. Reorganizing, note that

$$\begin{aligned}
\left(y_i - \zeta\alpha_i - \beta_i\right)(\zeta\alpha_i + \beta_i) &= \zeta y_i(R_{\boldsymbol{q}_\star} + y_i R_{\boldsymbol{w}_\star}) - \zeta^2(R_{\boldsymbol{q}_\star}^2 + R_{\boldsymbol{w}_\star}^2 + 2y_i R_{\boldsymbol{q}_\star}R_{\boldsymbol{w}_\star}) - 2\zeta R_{\boldsymbol{q}_\star}\beta_i - 2\zeta R_{\boldsymbol{w}_\star}y_i\beta_i + y_i\beta_i - \beta_i^2\\
&= \left(\zeta R_{\boldsymbol{q}_\star} - 2\zeta^2 R_{\boldsymbol{q}_\star}R_{\boldsymbol{w}_\star}\right)y_i + \left(\zeta R_{\boldsymbol{w}_\star} - \zeta^2(R_{\boldsymbol{q}_\star}^2 + R_{\boldsymbol{w}_\star}^2)\right) - \left(2\zeta R_{\boldsymbol{q}_\star}\right)\beta_i + \left(1 - 2\zeta R_{\boldsymbol{w}_\star}\right)y_i\beta_i - \beta_i^2
\end{aligned}$$

Overall,

$$\begin{aligned}
\boldsymbol{v}^\top\{\mathrm{Term}_{\mathrm{IV}}\} &= \left(\zeta R_{\boldsymbol{w}_\star} - \zeta^2(R_{\boldsymbol{q}_\star}^2 + R_{\boldsymbol{w}_\star}^2)\right)\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{v}^\top\boldsymbol{\gamma}_i\right) + \left(\zeta R_{\boldsymbol{q}_\star} - 2\zeta^2 R_{\boldsymbol{q}_\star}R_{\boldsymbol{w}_\star}\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\left(\boldsymbol{v}^\top\boldsymbol{\gamma}_i\right)\right)\\
&\quad + (1-2\zeta R_{\boldsymbol{w}_\star})\left(\frac{1}{n}\sum_{i=1}^{n} y_i\left(\boldsymbol{w}^\top\boldsymbol{\gamma}_i\right)\left(\boldsymbol{v}^\top\boldsymbol{\gamma}_i\right)\right)\\
&\quad - (2\zeta R_{\boldsymbol{q}_\star})\left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{w}^\top\boldsymbol{\gamma}_i\right)\left(\boldsymbol{v}^\top\boldsymbol{\gamma}_i\right) - \frac{1-\zeta}{T}\boldsymbol{v}^\top\boldsymbol{\Sigma}\boldsymbol{w}\right) - (2\zeta R_{\boldsymbol{q}_\star})\frac{1-\zeta}{T}\boldsymbol{v}^\top\boldsymbol{\Sigma}\boldsymbol{w}\\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{v}^\top\boldsymbol{\gamma}_i\right)\left(\left(\boldsymbol{w}^\top\boldsymbol{\gamma}_i\right)^2 - \frac{(1-\zeta)}{T}\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}\right) + \frac{(1-\zeta)}{T}\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{v}^\top\boldsymbol{\gamma}_i\right)\right).
\end{aligned}$$

According to Lemma 6 all random terms above are zero mean. Thus,

$$\mathbb{E}\left[\boldsymbol{v}^\top\{\text{Term}_{\text{IV}}\}\right] = -2\zeta R_{\boldsymbol{q}_\star}\frac{1-\zeta}{T}\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{v} \tag{24}$$

- **Combined**

The desired identities (19) and (20) follow by combining all the terms above.

B.2.2. PROOF OF LEMMA 6

First bound: Obvious by boundedness (hence, sub-gaussianity) of y_i and Fact G.1.

Second bound: For convenience set $\widetilde{T} = (1-\zeta)T$ and assume wlog that $\mathcal{R}^c = [\widetilde{T}]$. Recall that

$$\beta_i = \frac{1}{T}\sum_{t=1}^{\widetilde{T}}\boldsymbol{z}_{i,t}^\top\boldsymbol{w} = \frac{1-\zeta}{\widetilde{T}}\sum_{t=1}^{\widetilde{T}}\boldsymbol{z}_{i,t}^\top\boldsymbol{w}.$$

Also for all t: $\left\|\boldsymbol{z}_{i,t}^\top\boldsymbol{w}\right\|_{\psi_2} \le K\|\boldsymbol{w}\|_2$. Thus, from Fact G.1:

$$\|\beta_i\|_{\psi_2} \le \frac{C\sigma(1-\zeta)\|\boldsymbol{w}\|_2}{\sqrt{\widetilde{T}}} = \frac{C\sigma\sqrt{(1-\zeta)}\|\boldsymbol{w}\|_2}{\sqrt{T}}. \tag{25}$$

The bound then follows by applying Fact G.1 once more.

For the second term in this bound recall that $y_i \in \{\pm 1\}$ and $\beta_i = \sum_t \boldsymbol{z}_{i,t}^\top\boldsymbol{w}/T$. Also, for all $i \in [n]$: $y_i, \{\boldsymbol{z}_{i,t}\}_t$ are zero-mean and independent. Thus (i) $\mathbb{E}[y_i\beta_i] = 0$, and (ii) $\{y_i\boldsymbol{z}_{i,t} \overset{D}{\sim} \boldsymbol{z}_{i,t}$ and $y_i\boldsymbol{z}_{i,t} \perp y_i\boldsymbol{z}_{i,t'}\} \implies y_i\beta_i \overset{D}{\sim} \beta_i$. Thus, the same bound as the first term holds.

Third bound: It is easy to compute

$$\mathbb{E}[\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^\top\boldsymbol{v}\right)] = \frac{1}{T^2}\sum_{t=1}^{\widetilde{T}}\sum_{t'=1}^{\widetilde{T}}\mathbb{E}[\boldsymbol{w}^\top\boldsymbol{z}_{i,t}\boldsymbol{z}_{i,t'}\boldsymbol{v}] = \frac{1-\zeta}{T}\boldsymbol{v}^\top\boldsymbol{\Sigma}\boldsymbol{w}, \tag{26}$$

and, using (25)

$$\left\|\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^\top\boldsymbol{v}\right) - \mathbb{E}[\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^\top\boldsymbol{v}\right)]\right\|_{\psi_1} \le C\left\|\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^\top\boldsymbol{v}\right)\right\|_{\psi_1} \le C\left\|\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right\|_{\psi_2}\left\|\boldsymbol{\gamma}_i^\top\boldsymbol{v}\right\|_{\psi_2} = \frac{C\sigma^2(1-\zeta)\|\boldsymbol{w}\|_2\|\boldsymbol{v}\|_2}{T}. \tag{27}$$

Since $\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^\top\boldsymbol{v}\right), i \in [n]$ are independent, the desired bound on the first term follows from Fact G.1.

Consider now the second term. By independence of $y_i, \boldsymbol{\gamma}_i$ it holds that $\mathbb{E}[y_i\left(\boldsymbol{\gamma}_i^\top\boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^\top\boldsymbol{v}\right)] = 0$. Arguing as we did above for the second bound, $y_i\boldsymbol{\gamma}_i^\top\boldsymbol{w} \overset{D}{\sim} \boldsymbol{\gamma}_i^\top\boldsymbol{w}$. Hence, the subexponential bound is the same as for the first term.

Fourth bound: First, it is easy to compute that for all $i \in [n]$:

$$\mathbb{E}[\boldsymbol{w}^\top\boldsymbol{\Sigma}_i\boldsymbol{v}] = \frac{1}{T}\mathbb{E}[\boldsymbol{w}^\top\boldsymbol{Z}_i^\top\boldsymbol{Z}_i\boldsymbol{v}] = \frac{1}{T}\sum_{t=1}^{\widetilde{T}}\mathbb{E}[\boldsymbol{w}^\top\boldsymbol{z}_{i,t}^\top\boldsymbol{z}_{i,t}\boldsymbol{v}] = \frac{\widetilde{T}}{T}\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{v} = (1-\zeta)\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{v}.$$

Thus,

$$\boldsymbol{w}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} - \mathbb{E}[\boldsymbol{w}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v}] = \boldsymbol{w}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} - (1-\zeta)\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{v} = \frac{1}{T}\sum_{t=1}^{\widetilde{T}}\Big((\boldsymbol{z}_{i,t}^\top\boldsymbol{w})(\boldsymbol{z}_{i,t}^\top\boldsymbol{v}) - \boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{v}\Big)$$

and so

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{w}^\top\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} - (1-\zeta)\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{v} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{\widetilde{T}}\Big((\boldsymbol{z}_{i,t}^\top\boldsymbol{w})(\boldsymbol{z}_{i,t}^\top\boldsymbol{v}) - \boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{v}\Big).$$

Now, each random variable in the double sum above is independent and such that

$$\|(\boldsymbol{z}_{i,t}^\top \boldsymbol{w})(\boldsymbol{z}_{i,t}^\top \boldsymbol{v}) - \boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{v}\|_{\psi_1} \le 2\|(\boldsymbol{z}_{i,t}^\top \boldsymbol{w})(\boldsymbol{z}_{i,t}^\top \boldsymbol{v})\|_{\psi_1} \le 2\|\boldsymbol{z}_{i,t}^\top \boldsymbol{w}\|_{\psi_2}\|\boldsymbol{z}_{i,t}^\top \boldsymbol{v}\|_{\psi_2} \le C\sigma^2 \|\boldsymbol{w}\|_2 \|\boldsymbol{v}\|_2. \tag{28}$$

Hence, from Fact G.1,

$$\left\| \frac{1}{n} \sum_{i=1}^n \boldsymbol{w}^\top \hat{\boldsymbol{\Sigma}}_i \boldsymbol{v} - (1-\zeta) \boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{v} \right\|_{\psi_1} \le \frac{C\sigma^2 \sqrt{1-\zeta} \|\boldsymbol{w}\|_2 \|\boldsymbol{v}\|_2}{\sqrt{nT}} .$$

The bound for the second term follows along the same lines. The two key observations are that (i) $\mathbb{E}[y_i \boldsymbol{w}^\top \hat{\boldsymbol{\Sigma}}_i \boldsymbol{v}] = 0$ because y_i and $\boldsymbol{z}_{i,t}$ are independent, and, (ii)

$$y_i \boldsymbol{w}^\top \hat{\boldsymbol{\Sigma}}_i \boldsymbol{v} = \frac{1}{T} \boldsymbol{w}^\top y_i \boldsymbol{Z}_i^\top \boldsymbol{Z}_i \boldsymbol{v} \overset{D}{\sim} \frac{1}{T} \boldsymbol{w}^\top \widetilde{\boldsymbol{Z}}_i^\top \boldsymbol{Z}_i \boldsymbol{v} = \frac{1}{T} \sum_{t=1}^{\widetilde{T}} (\widetilde{\boldsymbol{z}}_{i,t}^\top \boldsymbol{w})(\boldsymbol{z}_{i,t}^\top \boldsymbol{v})$$

where $\widetilde{\boldsymbol{Z}}_i$ is an independent copy of $\boldsymbol{Z}_i$.

<u>Fifth bound:</u> From (29) and (26), we have for all $i \in [n]$ that

$$\left\| \left(\boldsymbol{\gamma}_i^\top \boldsymbol{w}\right)\left(\boldsymbol{\gamma}_i^\top \boldsymbol{v}\right) - \frac{1-\zeta}{T} \boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w} \right\|_{\psi_1} \le \frac{C\sigma^2 (1-\zeta) \|\boldsymbol{w}\|_2^2}{T}.$$

Moreover, recall from Eq. (25) that

$$\left\| \boldsymbol{\gamma}_i^\top \boldsymbol{v} \right\|_{\psi_2} \le \frac{C\sigma\sqrt{1-\zeta} \|\boldsymbol{v}\|_2}{\sqrt{T}}.$$

Combining the above two displays and applying Fact G.2 we find for all $i \in [n]$ that

$$\left\| \left(\left(\boldsymbol{\gamma}_i^\top \boldsymbol{w}\right)^2 - \frac{1-\zeta}{T} \boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{w} \right) \boldsymbol{\gamma}_i^\top \boldsymbol{v} \right\|_{\psi_{2/3}} \le \frac{C\sigma^3 (1-\zeta)^{3/2} \|\boldsymbol{w}\|_2^2 \|\boldsymbol{v}\|_2}{T^{3/2}}. \tag{29}$$

The desired bound follows from the above after using Fact G.3.

<u>Sixth bound:</u> From Eq. (28):

$$\left\| \boldsymbol{w}^\top \hat{\boldsymbol{\Sigma}}_i \boldsymbol{v} - (1-\zeta) \boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{v} \right\|_{\psi_1} \le \frac{C\sigma^2 \sqrt{1-\zeta} \|\boldsymbol{w}\|_2 \|\boldsymbol{v}\|_2}{\sqrt{T}}$$

and from Eq. (25)

$$\left\| \boldsymbol{\gamma}_i^\top \boldsymbol{w} \right\|_{\psi_2} \le \frac{C\sigma\sqrt{1-\zeta} \|\boldsymbol{w}\|_2}{\sqrt{T}}.$$

Next we use Fact G.2 with $\alpha = 2$ and $\beta = 1$ to find that

$$\left\| \left(\boldsymbol{\gamma}_i^\top \boldsymbol{w}\right)\left(\boldsymbol{w}^\top \hat{\boldsymbol{\Sigma}}_i \boldsymbol{v} - (1-\zeta) \boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{v}\right) \right\|_{\psi_{2/3}} \le \frac{CK^3 (1-\zeta) \|\boldsymbol{w}\|_2^2 \|\boldsymbol{v}\|_2}{T} .$$

Next we use Fact G.3 which allows us to conclude that

$$\left\| \frac{1}{n} \sum_{i=1}^n \left(\boldsymbol{\gamma}_i^\top \boldsymbol{w}\right)\left(\boldsymbol{w}^\top \hat{\boldsymbol{\Sigma}}_i \boldsymbol{v} - (1-\zeta) \boldsymbol{w}^\top \boldsymbol{\Sigma} \boldsymbol{v}\right) \right\|_{\psi_{2/3}} \le \frac{CK^3 (1-\zeta) \|\boldsymbol{w}\|_2^2 \|\boldsymbol{v}\|_2}{T\sqrt{n}} \log n .$$

Finally, the zero-mean property follows since

$$\begin{aligned}
\mathbb{E}[\left(\mathbb{1}^\top \boldsymbol{Z}_i \boldsymbol{w}\right)\left(\boldsymbol{w}^\top \boldsymbol{Z}_i \boldsymbol{Z}_i^\top \boldsymbol{v}\right)] &= \sum_{t=1}^{\widetilde{T}} \sum_{t'=1}^{\widetilde{T}} \mathbb{E}[\left(\boldsymbol{z}_{i,t}^\top \boldsymbol{w}\right)\left(\boldsymbol{w}^\top \boldsymbol{z}_{i,t'} \boldsymbol{z}_{i,t'}^\top \boldsymbol{v}\right)] = \sum_{t=1}^{\widetilde{T}} \sum_{t'=1}^{\widetilde{T}} \mathbb{E}[\boldsymbol{w}^\top \boldsymbol{z}_{i,t} \operatorname{tr}\left(\boldsymbol{z}_{i,t'} \boldsymbol{z}_{i,t'}^\top \boldsymbol{v} \boldsymbol{w}^\top\right)] \\
&= \widetilde{T}^2 \mathbb{E}[\operatorname{tr}\left(\boldsymbol{z}^\top \boldsymbol{w}\right) \operatorname{tr}\left(\boldsymbol{z}\boldsymbol{z}^\top \boldsymbol{v}\boldsymbol{w}^\top\right)] = \widetilde{T}^2 \mathbb{E}[\operatorname{tr}\left((\boldsymbol{z}^\top \boldsymbol{w}) \otimes (\boldsymbol{z}\boldsymbol{z}^\top \boldsymbol{v}\boldsymbol{w}^\top)\right)] \\
&= \widetilde{T}^2 \operatorname{tr}\left(\mathbb{E}[(\boldsymbol{z}^\top \otimes \boldsymbol{z}\boldsymbol{z}^\top)](\boldsymbol{w} \otimes \boldsymbol{v}\boldsymbol{w}^\top)\right) = 0 ,
\end{aligned}$$

where the last equality follows by the zero third moment property in Assumption 1.a.

C. Finite-sample gradient analysis

In the following, we assume without explicit further reference that Ass. 3.a holds (i.e. $|\rho| < W/Q$) and additionally that $Q > W$. To simplify the results we further assume $\sigma \propto 1$ and $W \gtrsim 1$.

C.1. First gradient step

The lemma below studies the deviation of the first-step of GD $\widehat{\boldsymbol{w}}_1$ with respect to its population counterpart $\boldsymbol{w}_1$. Provided that n and $n\zeta T/d$ are larger than appropriate functions of other problem parameters, then the deviations are of small multiplicative order.

Lemma 7 (First gradient step)**.** *Consider the one-step population and finite updates $\boldsymbol{w}_1 = \eta \boldsymbol{G}_{\boldsymbol{w}}(0,0)$ and $\widehat{\boldsymbol{w}}_1 = \eta \widehat{\boldsymbol{G}}_{\boldsymbol{w}}(0,0)$, respectively. For convenience denote*

$$R_{\boldsymbol{w}_\star} = \boldsymbol{w}_1^\top \boldsymbol{w}_\star, \quad R_{\boldsymbol{q}_\star} = \boldsymbol{w}_1^\top \boldsymbol{q}_\star, \quad \widehat{R}_{\boldsymbol{w}_\star} = \widehat{\boldsymbol{w}}_1^\top \boldsymbol{w}_\star, \quad \widehat{R}_{\boldsymbol{q}_\star} = \widehat{\boldsymbol{w}}_1^\top \boldsymbol{q}_\star .$$

For any $u > 0$ and any small constant $c_0 > 0$, there exist absolute constants $c, c' > 0$ and large enough constant $C = C(c_0) > 0$ such that if

$$\sqrt{n} \geq Cu\frac{Q}{W} \qquad \text{and} \qquad \sqrt{n\zeta T} \geq Cu\frac{\sigma}{W}\sqrt{\zeta^{-1}-1}\,, \tag{30}$$

then, with probability at least $1 - c'e^{-cu^2}$

$$\left|\widehat{R}_{\boldsymbol{w}_\star} - R_{\boldsymbol{w}_\star}\right| \leq c_0 R_{\boldsymbol{w}_\star} \qquad \text{and} \qquad \left|\widehat{R}_{\boldsymbol{q}_\star} - R_{\boldsymbol{q}_\star}\right| \leq c_0 \eta\zeta Q W\,. \tag{31}$$

Additionally, if

$$\sqrt{n} \geq Cu\frac{Q}{W} \qquad \text{and} \qquad \sqrt{n\zeta T} \geq C(1+u)\frac{\sigma}{W}\sqrt{\zeta^{-1}-1}\sqrt{d} \tag{32}$$

then with probability $1 - e^{-cu^2} - e^{-cdu^2}$

$$\left|\|\widehat{\boldsymbol{w}}_1\| - \|\boldsymbol{w}_1\|\right| \leq c_0 \eta\zeta W. \tag{33}$$

Proof. Note that the conclusions of the lemma are all homogeneous in η. Hence, without loss of generality, set $\eta = 1$. Also recall by Lemma 3 that

$$\widehat{\boldsymbol{w}}_1 = \widehat{\boldsymbol{G}}_{\boldsymbol{w}}(0,0) = \zeta \boldsymbol{w}_\star + \zeta \boldsymbol{q}_\star \left(\frac{1}{n}\sum_{i\in[n]} y_i\right) + \frac{1}{n}\sum_i y_i \boldsymbol{\gamma}_i\,, \tag{34}$$

and $\boldsymbol{w}_1 = \boldsymbol{G}_{\boldsymbol{w}}(0,0) = \zeta \boldsymbol{w}_\star$; thus, $R_{\boldsymbol{w}_\star} = \zeta W^2$. From these, and also using Lemma 6 , for any $u > 0$ with probability at least $1 - 2e^{-cu^2}$

$$\begin{aligned}\left|\widehat{R}_{\boldsymbol{w}_\star} - R_{\boldsymbol{w}_\star}\right| &= \left|\boldsymbol{w}_\star^\top(\widehat{\boldsymbol{w}}_1 - \boldsymbol{w}_1)\right| \leq \zeta\,|\rho|\,WQ\left|\frac{1}{n}\sum_{i\in[n]} y_i\right| + \left|\frac{1}{n}\sum_i y_i \boldsymbol{\gamma}_i^\top \boldsymbol{w}_\star\right| \leq \frac{Cu\zeta\,|\rho|\,WQ}{\sqrt{n}} + \frac{Cu\sigma\sqrt{1-\zeta}\sqrt{\zeta}W}{\sqrt{n\zeta T}} \\ &\leq c_0\zeta W^2 = c_0 R_{\boldsymbol{w}_\star}.\end{aligned}$$

where the last inequality follows by assuming $n, \zeta T$ large enough as in the condition of the lemma and using $|\rho| \leq 1$. Similarly,

$$\left|\widehat{R}_{\boldsymbol{q}_\star} - R_{\boldsymbol{q}_\star}\right| = \left|\boldsymbol{q}_\star^\top(\widehat{\boldsymbol{w}}_1 - \boldsymbol{w}_1)\right| \leq \zeta Q^2\left|\frac{1}{n}\sum_{i\in[n]} y_i\right| + \left|\frac{1}{n}\sum_i y_i \boldsymbol{\gamma}_i^\top \boldsymbol{q}_\star\right| \leq \frac{Cu\zeta Q^2}{\sqrt{n}} + \frac{Cu\sigma\sqrt{1-\zeta}\sqrt{\zeta}Q}{\sqrt{n\zeta T}} \leq c_0\zeta Q W$$

for sufficiently large n per (30). Finally, with probability at least $1 - e^{-cu^2} - e^{-cdu^2}$

$$\|\widehat{\boldsymbol{w}}_1\| - \|\boldsymbol{w}_1\| \leq \|\widehat{\boldsymbol{w}}_1 - \boldsymbol{w}_1\| \leq \frac{Cu\zeta Q}{\sqrt{n}} + \frac{C(1+u)\sigma\sqrt{1-\zeta}\sqrt{\zeta}\sqrt{d}}{\sqrt{n\zeta T}} \leq c_0\zeta W\,. \tag{35}$$

where, again, the last inequality follows by assuming $n, \zeta T$ large enough as stated in the lemma. In the second inequality, we used from Lemma 6 that $\sum_{i\in[n]} y_i\boldsymbol{\gamma}_i/n$ is $C\sigma\sqrt{1-\zeta}/\sqrt{nT}$-subGaussian and applied Fact G.4 to get a high-probability bound on its euclidean norm. □

C.2. Second gradient step

Next, we move on to the second gradient update in the direction of $\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)$. Recall our goal of controlling the relevance scores of signal and noisy tokens. The first lemma below takes a step in this direction by computing the signal and noise relevance scores assuming access to the population gradient $\boldsymbol{G}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1) := \mathbb{E}\left[\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)\right]$.

Lemma 8 ($\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \cdot)$ control: Expectation term)**.** *Let $\boldsymbol{G}_{\boldsymbol{q}}(0, \boldsymbol{w}_1) = \mathbb{E}\left[\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \boldsymbol{w}_1)\right]$ be the expectation of a gradient step in the $\boldsymbol{q}$-direction evaluated at $(0, \widehat{\boldsymbol{w}}_1)$ and recall that $\widehat{\boldsymbol{w}}_1 = \eta \widehat{\boldsymbol{G}}_{\boldsymbol{w}}(0, 0)$ for $\eta > 0$. Suppose $\widehat{\boldsymbol{w}}_1$ satisfies (31) and (33) for sufficiently small enough constant $c_0 > 0$. Further assume that the step-size η satisfies the following for sufficiently small absolute constant $c_\eta > 0$:*

$$\eta = \frac{c_\eta}{\sigma^2 Q^2}\,. \tag{36}$$

Then, for $y \in \{\pm 1\}$ it holds that

$$\left(\boldsymbol{q}_\star + y\boldsymbol{w}_\star\right)^\top \boldsymbol{G}_{\boldsymbol{q}}(0, \hat{\boldsymbol{w}}_1) \geq \eta\zeta(\zeta - \zeta^2) W^2 Q \left(\left(1/4 - \rho^2/8 - 3c_0\right) Q - \left(9/4\,|\rho| + c_0\right) W\right)\,, \tag{37}$$

and

$$\|\boldsymbol{G}_{\boldsymbol{q}}(0, \hat{\boldsymbol{w}}_1)\| \leq \eta\zeta\left(\zeta - \zeta^2\right) W^2 Q \left(13/4 + 2c_0\right)\,. \tag{38}$$

Proof. Fix any $\boldsymbol{v}$ and recall the notation of Lemma 7. With these, we have from Lemma 5 that

$$\begin{aligned}
\boldsymbol{v}^\top \boldsymbol{G}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1) &= \left((\zeta - \zeta^2)\left(\widehat{R}_{\boldsymbol{w}_\star} + \sigma^2\|\widehat{\boldsymbol{w}}_1\|^2/T\right) - (\zeta^2 - \zeta^3)\left(\widehat{R}^2_{\boldsymbol{w}_\star} + \widehat{R}^2_{\boldsymbol{q}_\star}\right)\right) \boldsymbol{v}^\top \boldsymbol{q}_\star \\
&\qquad + \left((\zeta - \zeta^2) - 2(\zeta^2 - \zeta^3)\widehat{R}_{\boldsymbol{w}_\star}\right) \widehat{R}_{\boldsymbol{q}_\star}\, \boldsymbol{v}^\top \boldsymbol{w}_\star - \sigma^2\left(1 + 2/T\right)(\zeta - \zeta^2)\widehat{R}_{\boldsymbol{q}_\star} \boldsymbol{v}^\top \widehat{\boldsymbol{w}}_1 \\
&= (\zeta - \zeta^2)\widehat{R}_{\boldsymbol{w}_\star}\left(1 + \sigma^2\|\widehat{\boldsymbol{w}}_1\|^2/(T\widehat{R}_{\boldsymbol{w}_\star}) - \zeta\left(\widehat{R}_{\boldsymbol{w}_\star} + \widehat{R}^2_{\boldsymbol{q}_\star}/\widehat{R}_{\boldsymbol{w}_\star}\right)\right) \boldsymbol{v}^\top \boldsymbol{q}_\star \\
&\qquad + (\zeta - \zeta^2)\widehat{R}_{\boldsymbol{q}_\star}\left(1 - 2\zeta\widehat{R}_{\boldsymbol{w}_\star}\right) \boldsymbol{v}^\top \boldsymbol{w}_\star - \sigma^2\left(1 + 2/T\right)(\zeta - \zeta^2)\widehat{R}_{\boldsymbol{q}_\star} \boldsymbol{v}^\top \widehat{\boldsymbol{w}}_1 \\
&= (\zeta - \zeta^2)\widehat{R}_{\boldsymbol{w}_\star} \underbrace{\left(1 + \sigma^2\|\widehat{\boldsymbol{w}}_1\|^2/(T\widehat{R}_{\boldsymbol{w}_\star}) - \zeta\left(\widehat{R}_{\boldsymbol{w}_\star} + \widehat{R}^2_{\boldsymbol{q}_\star}/\widehat{R}_{\boldsymbol{w}_\star}\right)\right)}_{:=\widehat{C}_1} \boldsymbol{v}^\top \boldsymbol{q}_\star \\
&\qquad + (\zeta - \zeta^2)\widehat{R}_{\boldsymbol{q}_\star} \underbrace{\left(1 - 2\zeta\widehat{R}_{\boldsymbol{w}_\star} - \eta\zeta\sigma^2\left(1 + 2/T\right)\right)}_{:=\widehat{C}_2} \boldsymbol{v}^\top \boldsymbol{w}_\star - \underbrace{\eta\sigma^2\left(1 + 2/T\right)}_{:=\widehat{C}_3}(\zeta - \zeta^2)\widehat{R}_{\boldsymbol{q}_\star} \boldsymbol{v}^\top \boldsymbol{\delta}_1
\end{aligned} \tag{39}$$

where, in the last line we used (34) and set

$$\boldsymbol{\delta}_1 := (\widehat{\boldsymbol{w}}_1 - \boldsymbol{w}_1)/\eta = \zeta \boldsymbol{q}_\star \left(\frac{1}{n}\sum_{i\in[n]} y_i\right) + \frac{1}{n}\sum_i y_i \boldsymbol{\gamma}_i.$$

Recall from (31) and (33) that $R_{\boldsymbol{w}_\star}/2 \leq \widehat{R}_{\boldsymbol{w}_\star} \leq 3R_{\boldsymbol{w}_\star}/2$, $|\widehat{R}_{\boldsymbol{q}_\star} - R_{\boldsymbol{q}_\star}| \leq c_0\eta\zeta QW$ and $\|\widehat{\boldsymbol{w}}_1\| \leq \|\boldsymbol{w}_1\| + c_0\eta\zeta W$. Also, from (33) we have that $\|\boldsymbol{\delta}_1\| \leq c_0\zeta W,$. Here and onwards, $c_0 > 0$ is a small enough absolute constant (smaller than $1/2$) whose value may change from line to line. Further recall $R_{\boldsymbol{w}_\star} = \eta\zeta W^2$, $R_{\boldsymbol{q}_\star} = \eta\zeta\rho WQ$ and $\|\boldsymbol{w}_1\| = \eta\zeta W$. With these, we can set small enough step size $\eta \propto \sigma^{-2}Q^{-2}$ such that

$$\widehat{C}_1 \in [1/2, 3/2], \quad \widehat{C}_2 \in [-1/8, 1], \quad \widehat{C}_3 \in [0, (c_3/\zeta)/Q^2]\,,$$

for constant $c_3 > 0$ to be made small enough later in the proof.

From the above, we can compute

$$\widehat{C}_3\left|\widehat{R}_{\boldsymbol{q}_\star}\right| \|\boldsymbol{\delta}_1\| \leq \frac{c_3}{\zeta Q^2}\eta\zeta(c_0 QW + \rho WQ)(c_0\zeta Q) \leq c_3 c_0 \eta\zeta\, Q\,.$$

Moreover,

$$\begin{aligned}\hat{C}_1\widehat{R}_{\boldsymbol{w}_\star}Q^2 &\geq \eta\zeta W^2Q^2/4\\ \hat{C}_2\widehat{R}_{\boldsymbol{q}_\star}\rho WQ &= \hat{C}_2\rho WQR_{\boldsymbol{q}_\star} + C_2\rho WQ(\widehat{R}_{\boldsymbol{q}_\star} - R_{\boldsymbol{q}_\star}) \geq -\eta\zeta\rho^2Q^2W^2/8 - c_0\eta\zeta Q^2W^2\\ &\geq -\eta\zeta\rho^2Q^2W^2/8 - c_0\eta\zeta W^2Q^2\,.\end{aligned}$$

Putting the above displays together we find that

$$\boldsymbol{q}_\star^\top\boldsymbol{G}_{\boldsymbol{q}}(0,\hat{\boldsymbol{w}}_1) \geq \eta\zeta(\zeta-\zeta^2)\left(W^2Q^2/4 - \rho^2Q^2W^2/8 - c_0W^2Q^2 - c_3c_0Q^2\right)\,.$$

Similarly, we can compute that

$$\begin{aligned}\left|\hat{C}_1\widehat{R}_{\boldsymbol{w}_\star}\rho QW\right| &\leq (9/4)\eta\zeta W^3Q\,|\rho|\\ \left|\hat{C}_2\widehat{R}_{\boldsymbol{q}_\star}W^2\right| &\leq \eta\zeta(|\rho|\,W^3Q + c_0QW^3)\,.\end{aligned}$$

Hence, for $y \in \{\pm 1\}$

$$\left|y\boldsymbol{w}_\star^\top\boldsymbol{G}_{\boldsymbol{q}}(0,\hat{\boldsymbol{w}}_1)\right| \leq \eta\zeta\left(\zeta-\zeta^2\right)4W^3Q\,|\rho| + c_0QW^3 + c_3c_0QW\,.$$

The above two displays put together yield (37). Specifically, we also use the simplifying assumption $Q \gtrsim W \gtrsim 1$ and pick c_3 small enough so that $c_3W^{-2}Q^{-1} \leq 1$.

The norm bound in (38) follows again starting from (39) and using similar arguments as above to show:

$$\begin{aligned}\|\boldsymbol{G}_{\boldsymbol{q}}(0,\hat{\boldsymbol{w}}_1)\| &\leq \eta\zeta(\zeta-\zeta^2)\left(\widehat{C}_1R_{\boldsymbol{w}_\star}Q + \left|\widehat{C}_2\right|\left|\widehat{R}_{\boldsymbol{q}_\star}\right|W + \widehat{C}_3\left|\widehat{R}_{\boldsymbol{q}_\star}\right|\|\boldsymbol{\delta}_1\|\right)\\ &\leq \eta\zeta(\zeta-\zeta^2)\left((9/4)W^2Q + (|\rho|+c_0)W^2Q + c_3c_0Q\right)\,.\end{aligned}$$

□

The next lemma controls the effect on the relevance scores of the deviation term $\widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}) = \widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}) - \boldsymbol{G}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}})$.

Lemma 9 ($\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0,\cdot)$ control: Deviation term)**.** *Let $\widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}_1) := \widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}_1) - \boldsymbol{G}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}_1)$ and suppose $\widehat{\boldsymbol{w}}_1$ satisfies (31) and (33). Also assume $\sigma = 1$. Fix any $u > 0$ and any small constant $c_1 > 0$. Then, there exists small enough constant c_η (dependent on c_1) such that if step-size η is small enough as per (36) the following statements hold.*

With probability at least $1 - c'e^{-cu^{2/3}}$ for positive constants $C, c', c > 0$ it holds for signal tokens that

$$\left|(\boldsymbol{q}_\star + y\boldsymbol{w}_\star)^\top\widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}_1)\right| \leq u\,C\,c_1\,Q\left(Q \vee \frac{\sigma \vee \sigma^2}{\sqrt{T}} \vee \frac{\sigma^3}{T}\right)\frac{\log(n)}{\sqrt{n}}, \quad y \in \{\pm 1\} \tag{40}$$

Moreover, with probability at least $1 - c'd\,e^{-cu^{2/3}}$ it holds that

$$\|\widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0,\hat{\boldsymbol{w}}_1)\| \leq uC\,c_1\left(Q \vee \frac{\sigma \vee \sigma^2}{\sqrt{T}} \vee \frac{\sigma^3}{T}\right)\frac{\log(n)\sqrt{d}}{\sqrt{n}}\,. \tag{41}$$

Proof. We study each one of the terms of $\widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}_1)$ in (20) separately. We repeat the terms here for convenience, also

noting the substitutions $R_{\boldsymbol{w}_\star} \leftarrow \widehat{R}_{\boldsymbol{w}_\star} = \widehat{\boldsymbol{w}}_1^\top \boldsymbol{w}_\star$ and $R_{\boldsymbol{q}_\star} \leftarrow \widehat{R}_{\boldsymbol{q}_\star} = \widehat{\boldsymbol{w}}_1^\top \boldsymbol{q}_\star$.

$$
\begin{aligned}
\boldsymbol{v}^\top \widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1) = {} & \Big[\big((\zeta - \zeta^2) - 2(\zeta^2 - \zeta^3)\widehat{R}_{\boldsymbol{w}_\star}\big)\widehat{R}_{\boldsymbol{q}_\star} R_{\boldsymbol{v},\boldsymbol{q}_\star} + \big((\zeta - \zeta^2)\widehat{R}_{\boldsymbol{w}_\star} - (\zeta^2 - \zeta^3)\big(\widehat{R}^2_{\boldsymbol{w}_\star} + \widehat{R}^2_{\boldsymbol{q}_\star}\big)\big) R_{\boldsymbol{v},\boldsymbol{w}_\star}\Big]\left(\frac{1}{n}\sum_{i=1}^n y_i\right) \\
& + \Big[\big(-\zeta + 2\zeta^2\big)\widehat{R}_{\boldsymbol{q}_\star} R_{\boldsymbol{v},\boldsymbol{q}_\star} + \big(-\zeta + (-\zeta + 2\zeta^2)\widehat{R}_{\boldsymbol{w}_\star}\big) R_{\boldsymbol{v},\boldsymbol{w}_\star} + (1-\zeta)\widehat{\boldsymbol{w}}_1^\top \boldsymbol{\Sigma}\boldsymbol{v}\Big]\left(\frac{1}{n}\sum_{i=1}^n \big(\boldsymbol{\gamma}_i^\top \widehat{\boldsymbol{w}}_1\big)\right) \\
& \quad + \left[\zeta \widehat{R}_{\boldsymbol{w}_\star} - \zeta^2(\widehat{R}^2_{\boldsymbol{q}_\star} + \widehat{R}^2_{\boldsymbol{w}_\star}) + \frac{(1-\zeta)}{T}\widehat{\boldsymbol{w}}_1^\top \boldsymbol{\Sigma}\widehat{\boldsymbol{w}}_1\right]\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{v}^\top \boldsymbol{\gamma}_i\right) \\
& \quad + \Big[\big(-\zeta + (-\zeta + 2\zeta^2)\widehat{R}_{\boldsymbol{w}_\star}\big) R_{\boldsymbol{v},\boldsymbol{q}_\star} + \big(-\zeta + 2\zeta^2\big)\widehat{R}_{\boldsymbol{q}_\star} R_{\boldsymbol{v},\boldsymbol{w}_\star}\Big]\left(\frac{1}{n}\sum_{i=1}^n y_i \big(\boldsymbol{\gamma}_i^\top \widehat{\boldsymbol{w}}_1\big)\right) \\
& \quad + \Big[\zeta \widehat{R}_{\boldsymbol{q}_\star} - 2\zeta^2 \widehat{R}_{\boldsymbol{q}_\star}\widehat{R}_{\boldsymbol{w}_\star}\Big]\left(\frac{1}{n}\sum_{i=1}^n y_i \big(\boldsymbol{v}^\top \boldsymbol{\gamma}_i\big)\right) \\
& + \zeta R_{\boldsymbol{v},\boldsymbol{q}_\star}\left(\frac{1}{n}\sum_{i=1}^n \big(\boldsymbol{\gamma}_i^\top \widehat{\boldsymbol{w}}_1\big)^2 - \frac{(1-\zeta)}{T}\widehat{\boldsymbol{w}}_1^\top \boldsymbol{\Sigma}\widehat{\boldsymbol{w}}_1\right) \\
& \quad + \zeta R_{\boldsymbol{v},\boldsymbol{w}_\star}\left(\frac{1}{n}\sum_{i=1}^n \big(\boldsymbol{\gamma}_i^\top \widehat{\boldsymbol{w}}_1\big)^2 y_i\right) \\
& \quad + \Big[1 - 2\zeta \widehat{R}_{\boldsymbol{w}_\star}\Big]\left(\frac{1}{n}\sum_{i=1}^n y_i \big(\widehat{\boldsymbol{w}}_1^\top \boldsymbol{\gamma}_i\big)\big(\boldsymbol{v}^\top \boldsymbol{\gamma}_i\big)\right) \\
& + \big(1 - \zeta \widehat{R}_{\boldsymbol{w}_\star}\big)\left(\frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{v}^\top \hat{\boldsymbol{\Sigma}}_i \widehat{\boldsymbol{w}}_1\right) \\
& \quad - \zeta \widehat{R}_{\boldsymbol{q}_\star}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{v}^\top \hat{\boldsymbol{\Sigma}}_i \widehat{\boldsymbol{w}}_1 - (1-\zeta)\boldsymbol{v}^\top \boldsymbol{\Sigma}\widehat{\boldsymbol{w}}_1\right) \\
& - \Big[2\zeta \widehat{R}_{\boldsymbol{q}_\star}\Big]\left(\frac{1}{n}\sum_{i=1}^n \big(\widehat{\boldsymbol{w}}_1^\top \boldsymbol{\gamma}_i\big)\big(\boldsymbol{v}^\top \boldsymbol{\gamma}_i\big) - \frac{1-\zeta}{T}\boldsymbol{v}^\top \boldsymbol{\Sigma}\widehat{\boldsymbol{w}}_1\right) \\
& - \left(\frac{1}{n}\sum_{i=1}^n \big(\boldsymbol{v}^\top \boldsymbol{\gamma}_i\big)\left(\big(\widehat{\boldsymbol{w}}_1^\top \boldsymbol{\gamma}_i\big)^2 - \frac{(1-\zeta)}{T}\widehat{\boldsymbol{w}}_1^\top \boldsymbol{\Sigma}\widehat{\boldsymbol{w}}_1\right)\right) \\
& - \left(\frac{1}{n}\sum_{i=1}^n \big(\boldsymbol{\gamma}_i^\top \widehat{\boldsymbol{w}}_1\big)\big(\boldsymbol{v}^\top \hat{\boldsymbol{\Sigma}}_i \widehat{\boldsymbol{w}}_1 - (1-\zeta)\widehat{\boldsymbol{w}}_1^\top \boldsymbol{\Sigma}\boldsymbol{v}\big)\right). \qquad (42)
\end{aligned}
$$

Recall from the lemma assumption that (31) holds and from $R_{\boldsymbol{w}_\star} = \eta\zeta W^2$ and $R_{\boldsymbol{q}_\star} = \eta\zeta\rho W Q$, that $1/2\eta\zeta W^2 \le \widehat{R}_{\boldsymbol{w}_\star} \le 3/2\eta\zeta W^2$ and $\big|\widehat{R}_{\boldsymbol{q}_\star}\big| \le \eta\zeta(|\rho| + c_0)WQ$. The observation is that we can choose step-size η small enough (as stated in (36)) to bound (in absolute value) all the coefficients in (42) (aka all terms in square brackets) that include $\widehat{R}_{\boldsymbol{w}_\star}, \widehat{R}_{\boldsymbol{q}_\star}$. Therefore, for any small positive constant $c_1 > 0$ it can be checked that there is sufficiently small constant c_η that determines step-size η

in (36) such that

$$
\begin{aligned}
\left|\boldsymbol{v}^\top \widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)\right| \leq\; & c_1 \left[|R_{\boldsymbol{v},\boldsymbol{q}_\star}| + |R_{\boldsymbol{v},\boldsymbol{w}_\star}|\right] \left|\frac{1}{n}\sum_{i=1}^n y_i\right| \\
& + \left[c_1 |R_{\boldsymbol{v},\boldsymbol{q}_\star}| + (1+c_1)|R_{\boldsymbol{v},\boldsymbol{w}_\star}| + (1-\zeta)\sigma^2 \left|\widehat{\boldsymbol{w}}_1^\top \boldsymbol{v}\right|\right] \left|\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{\gamma}_i^\top \widehat{\boldsymbol{w}}_1\right)\right| \\
& \quad + \left[c_1 + \frac{(1-\zeta)}{T}\sigma^2 \|\widehat{\boldsymbol{w}}_1\|^2\right] \left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{v}^\top \boldsymbol{\gamma}_i\right| \\
& \quad + \left[(1+c_1)|R_{\boldsymbol{v},\boldsymbol{q}_\star}| + c_1 |R_{\boldsymbol{v},\boldsymbol{w}_\star}|\right] \left|\frac{1}{n}\sum_{i=1}^n y_i \left(\boldsymbol{\gamma}_i^\top \widehat{\boldsymbol{w}}_1\right)\right| \\
& \quad + \left[c_1\right] \left|\frac{1}{n}\sum_{i=1}^n y_i \left(\boldsymbol{v}^\top \boldsymbol{\gamma}_i\right)\right| \\
& + \zeta |R_{\boldsymbol{v},\boldsymbol{q}_\star}| \left|\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{\gamma}_i^\top \widehat{\boldsymbol{w}}_1\right)^2 - \frac{(1-\zeta)}{T} \widehat{\boldsymbol{w}}_1^\top \boldsymbol{\Sigma} \widehat{\boldsymbol{w}}_1\right| \\
& \quad + \zeta |R_{\boldsymbol{v},\boldsymbol{w}_\star}| \left|\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{\gamma}_i^\top \widehat{\boldsymbol{w}}_1\right)^2 y_i\right| \\
& \quad + \left[1+c_1\right] \left|\frac{1}{n}\sum_{i=1}^n y_i \left(\widehat{\boldsymbol{w}}_1^\top \boldsymbol{\gamma}_i\right)\left(\boldsymbol{v}^\top \boldsymbol{\gamma}_i\right)\right| \\
& + \left[1+c_1\right] \left|\frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{v}^\top \hat{\boldsymbol{\Sigma}}_i \widehat{\boldsymbol{w}}_1\right| \\
& \quad + c_1 \left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{v}^\top \hat{\boldsymbol{\Sigma}}_i \widehat{\boldsymbol{w}}_1 - (1-\zeta)\boldsymbol{v}^\top \boldsymbol{\Sigma} \widehat{\boldsymbol{w}}_1\right| \\
& + \left[c_1\right] \left|\frac{1}{n}\sum_{i=1}^n \left(\widehat{\boldsymbol{w}}_1^\top \boldsymbol{\gamma}_i\right)\left(\boldsymbol{v}^\top \boldsymbol{\gamma}_i\right) - \frac{1-\zeta}{T} \boldsymbol{v}^\top \boldsymbol{\Sigma} \widehat{\boldsymbol{w}}_1\right| \\
& + \left|\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{v}^\top \boldsymbol{\gamma}_i\right)\left(\left(\widehat{\boldsymbol{w}}_1^\top \boldsymbol{\gamma}_i\right)^2 - \frac{(1-\zeta)}{T} \widehat{\boldsymbol{w}}_1^\top \boldsymbol{\Sigma} \widehat{\boldsymbol{w}}_1\right)\right| \\
& + \left|\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{\gamma}_i^\top \widehat{\boldsymbol{w}}_1\right)\left(\boldsymbol{v}^\top \hat{\boldsymbol{\Sigma}}_i \widehat{\boldsymbol{w}}_1 - (1-\zeta)\widehat{\boldsymbol{w}}_1^\top \boldsymbol{\Sigma} \boldsymbol{v}\right)\right| .
\end{aligned} \tag{43}
$$

Now, we use successively Lemma 6 to bound the random terms. Also note that $|R_{\boldsymbol{v},\boldsymbol{q}_\star}| \leq Q\|\boldsymbol{v}\|$, $|R_{\boldsymbol{v},\boldsymbol{w}_\star}| \leq W\|\boldsymbol{v}\|$ and $\|\widehat{\boldsymbol{w}}_1\| \leq \eta\zeta(1+c_0)W =: \eta\zeta M$. Here, we denote $M := (1+c_0)W$ for convenience. With these, for any $u > 0$, with probability at least $1 - c' e^{-cu^{2/3}}$ we have

$$
\begin{aligned}
&\left|\boldsymbol{v}^\top \widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)\right| \le u \cdot \frac{C}{\sqrt{n}} \|\boldsymbol{v}\| c_1 (W+Q) \\
&+ u \cdot \frac{C\sigma\sqrt{1-\zeta}}{\sqrt{nT}} \|\boldsymbol{v}\| \left(\left((1+c_1)(2W+2Q) + c_1\eta\zeta(1-\zeta)\sigma^2 M\right)\eta\zeta M + \left(2c_1 + \sigma^2(1-\zeta)\eta^2\zeta^2 M^2/T\right)\right) \\
&+ u \cdot \frac{C\sigma^2(1-\zeta)}{T\sqrt{n}} \|\boldsymbol{v}\| \left(\eta^2\zeta^3 M^2 Q + \eta^2\zeta^3 W M^2 + (1+c_1)\eta\zeta M\right) \\
&+ u \cdot \frac{C\sigma^2\sqrt{1-\zeta}}{\sqrt{nT}} \|\boldsymbol{v}\| \left((1+2c_1)\eta\zeta M\right) \\
&+ u \frac{C\sigma^2(1-\zeta)}{T\sqrt{n}} \|\boldsymbol{v}\| \left(c_1\eta\zeta M\right) \\
&+ u \frac{C\sigma^3(1-\zeta)^{3/2}\log(n)}{T^{3/2}\sqrt{n}} \|\boldsymbol{v}\| \left(\eta^2\zeta^2 M^2\right) \\
&+ u \frac{C\sigma^3(1-\zeta)\log(n)}{T\sqrt{n}} \|\boldsymbol{v}\| \eta^2\zeta^2 M^2 .
\end{aligned} \tag{44}
$$

Now, again using small step size η as per (36) (recall that $M = (1+c_0)W \lesssim Q$), this can be further simplified to the following (here the value of constant c_1 might be different from (44))

$$
\begin{aligned}
\left|\boldsymbol{v}^\top \widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)\right| &\le u \cdot \frac{C}{\sqrt{n}} \|\boldsymbol{v}\| c_1 (W+Q) + u \cdot \frac{C\sigma\sqrt{1-\zeta}}{\sqrt{nT}} \|\boldsymbol{v}\| c_1 + u \cdot \frac{C\sigma^2(1-\zeta)}{T\sqrt{n}} \|\boldsymbol{v}\| c_1 \\
&\quad + u \cdot \frac{C\sigma^2\sqrt{1-\zeta}}{\sqrt{nT}} \|\boldsymbol{v}\| c_1 + u \frac{C\sigma^3(1-\zeta)^{3/2}\log(n)}{T^{3/2}\sqrt{n}} \|\boldsymbol{v}\| c_1 + u \frac{C\sigma^3(1-\zeta)\log(n)}{T\sqrt{n}} \|\boldsymbol{v}\| c_1 \\
&\le u \cdot \|\boldsymbol{v}\| \cdot \frac{C}{\sqrt{n}} \cdot c_1 \left(Q \vee \frac{\sigma \vee \sigma^2}{\sqrt{T}} \vee \frac{\sigma^2}{T} \vee \frac{\sigma^3\log(n)}{T^{3/2}} \vee \frac{\sigma^3\log(n)}{T}\right) \\
&\le u \cdot \|\boldsymbol{v}\| \cdot \frac{C\,\log(n)}{\sqrt{n}} \cdot c_1 \left(Q \vee \frac{\sigma \vee \sigma^2}{\sqrt{T}} \vee \frac{\sigma^3}{T}\right).
\end{aligned} \tag{45}
$$

Now, we can compute the deviation of the relevance scores. For signal tokens we have for both $y \in \{\pm 1\}$, and all $u > 0$ with probability at least $1 - c' e^{-cu^{2/3}}$, there exist constant $C > 0$ such that

$$
\left|(\boldsymbol{q}_\star + y\boldsymbol{w}_\star)^\top \widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)\right| \le u\, C\, c_1\, Q \left(Q \vee \frac{\sigma \vee \sigma^2}{\sqrt{T}} \vee \frac{\sigma^3}{T}\right) \frac{\log(n)}{\sqrt{n}} \tag{46}
$$

Similarly, since (45) holds for all $\boldsymbol{v}$, we can apply it for all standard basis vectors $\boldsymbol{v} = \boldsymbol{e}_j, j \in [n]$ and union bounding yields for all $u > 0$ with probability at least $1 - c' d e^{-cu^{2/3}}$,

$$
\|\widetilde{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)\| \le u\, C\, c_1 \left(Q \vee \frac{\sigma \vee \sigma^2}{\sqrt{T}} \vee \frac{\sigma^3}{T}\right) \frac{\log(n)\sqrt{d}}{\sqrt{n}} .
$$

□

C.3. First and second gradient steps combined: Learning the relevant features

With lemmas 7, 8, and 9 at hand, we are now ready to put things together stating our final bounds for relevance scores. The finding is presented as a stand-alone lemma below.

Lemma 10 (Put things together)**.** *Consider the finite-sample gradient step $\widehat{\boldsymbol{q}}_1 = \widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)$, where recall that $\widehat{\boldsymbol{w}}_1 = \eta \widehat{\boldsymbol{G}}_{\boldsymbol{w}}(0,0)$. Fix any $u_0, u_1, u_2, u_3 > 0$ and any small constants $\widetilde{c}_0, \widetilde{c}_1 > 0$. Suppose step-size η of first gradient step satisfies (36) for sufficiently small constant $c_\eta = c_\eta(\widetilde{c}_1) > 0$ and further assume*

$$
\sqrt{n} \ge u_0 \cdot C_0 \frac{Q}{W} \qquad \textit{and} \qquad \sqrt{\frac{n\zeta T}{d}} \ge (1+u_0) \cdot C_0 \frac{\sigma}{W} \sqrt{1/\zeta - 1}. \tag{47}
$$

for some large enough constant $C_0 = C_0(\widetilde{c}_0) > 0$. Finally, make the following mild assumption (for simplicity), $\frac{\sigma \vee \sigma^2}{\sqrt{T}} \vee \frac{\sigma^3}{T} \le Q$. For a fresh dataset $(\boldsymbol{X}_i, y_i)_{i\in[n]}$ consider the signal relevance scores $\hat{r}_{i,t} := r_{i,0} := (\boldsymbol{q}_\star + y_i \boldsymbol{w}_\star)^\top \widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \hat{\boldsymbol{w}}_1)$ for $t \in \mathcal{R}$. There exist positive absolute constants $c_i, c_i', i = 0, 1, 2, 3$ such that the following statements hold.

- *With probability at least $1 - c_0' e^{-c_0 u_0^2} - c_1' e^{c_1 u_1^{2/3}}$, the signal relevance scores satisfy*

$$\min_{i\in[n]} r_{i,0} \ge \eta\zeta(\zeta - \zeta^2) W^2 Q\big(\big(1/4 - \rho^2/8 - 2\widetilde{c}_0\big) Q - (9/4\,|\rho| + 2\widetilde{c}_0)\, W\big) - u_1\, \widetilde{c}_1\, Q^2 \frac{\log(n)}{\sqrt{n}} =: B(u_1)\,. \tag{48}$$

- *With probability at least $1 - c_0' e^{-c_0 u_0^2} - c_3' d e^{c_3 u_3^{2/3}}$, the norm of $\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \hat{\boldsymbol{w}}_1)$ satisfies*

$$\|\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \hat{\boldsymbol{w}}_1)\| \le \eta\zeta\left(\zeta - \zeta^2\right) W^2 Q\,(13/4 + 2\widetilde{c}_0) + u_3 \widetilde{c}_1 \sigma Q \frac{\log(n)\sqrt{d}}{\sqrt{n}}\,. \tag{49}$$

Proof. The lemma follows by collecting (37), (40), (38), (41), and applying union bound for the noise terms over $i \in [n], t \in \mathcal{R}^c$. □

From the lemma above, we can show that provided Q is large enough with respect to W and ρ is small enough (see (50) below) then the normalized signal relevance score is with high probability over the training set proportional to Q.

Corollary 1. *Suppose there exists positive constant $\alpha \in (0, 3/16)$ such that*

$$A_Q := \big(3/16 - \rho^2/8\big) Q - (9/4\,|\rho| + 1/16)\, W \ge \alpha \cdot Q\,. \tag{50}$$

Then, for sufficiently small step-size $\eta \propto Q^{-2}$ and large enough n there exist constants $c_0', c_0, c_1', c_1, c_3, c_3'$ such that with probability at least

$$\begin{aligned} 1 - c_0' \exp\left(-c_0 n \left((W/Q)^2 \wedge \frac{\zeta^2 W^2 T}{d}\right)\right) - c_1' \exp\left(-c_1 n^{1/3} (W/Q)^{4/3} \zeta^{4/3} \big/ \log^{2/3}(n)\right) \\ - c_3' d \exp\left(-c_3 (n/d)^{1/3} (W/Q)^{4/3} \zeta^{4/3} \big/ \log^{2/3}(n)\right), \end{aligned} \tag{51}$$

it holds that

$$(\boldsymbol{q}_\star + y\boldsymbol{w}_\star)^\top \left(\frac{\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)}{\|\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)\|}\right) \ge (\alpha/14) \cdot Q\,.$$

Proof. Suppose (50) holds and recall the notation A_Q defined therein. Further suppose n is large enough such that (47) holds so that we can invoke Lemma 10. Therein set

$$u_1 \propto \eta W^2 \zeta^2 \frac{\sqrt{n}}{\log(n)} \propto (W/Q)^2 \zeta^2 \frac{\sqrt{n}}{\log(n)},$$

$$u_3 \propto \eta W^2 \zeta^2 \frac{\sqrt{n}}{\log(n)\sqrt{d}} \propto (W/Q)^2 \zeta^2 \frac{\sqrt{n}}{\log(n)\sqrt{d}}$$

and

$$u_0 \propto \sqrt{n}\left((W/Q) \wedge \zeta W \sqrt{T/d}\right).$$

Note that the latter condition is consistent with (47). On the other hand, the former two conditions are chosen so that the following two hold.

First,

$$\eta\zeta(\zeta - \zeta^2) W^2 Q \cdot A_Q > 2 u_1 \widetilde{c}_1 Q^2 \frac{\log n}{\sqrt{n}}\,.$$

Thus, (48) and setting $2\widetilde{c}_0 = 1/16$ give

$$(\boldsymbol{q}_\star + y\boldsymbol{w}_\star)^\top \widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1) \ge \frac{1}{2}\eta\zeta(\zeta - \zeta^2) W^2 Q A_Q$$

Second,

$$\eta\zeta(\zeta-\zeta^2)W^2Q > u_3\widetilde{c}_1 Q\frac{\log n\sqrt{d}}{\sqrt{n}}$$

Thus, from (49)

$$\|\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0,\hat{\boldsymbol{w}}_1)\| \le 2\eta\zeta\left(\zeta-\zeta^2\right)W^2Q\left(13/4+1/16\right).$$

Combining the above we conclude with the desired: the signal correlation is lower bounded by

$$(\boldsymbol{q}_\star+y\boldsymbol{w}_\star)^\top\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}_1)\big/\|\widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0,\widehat{\boldsymbol{w}}_1)\| \ge A_Q/14 \ge (\alpha/14)\cdot Q.$$

□

Having shown that the second gradient step has high signal relevance score, we can use it to show in the lemma below that the attention features are close to the signal features with high-probability. This property will play a key role in finalizing the finite-sample analysis. Indeed, the rate of the event in (55) will turn out to govern the error rate of our algorithm.

Lemma 11 (From learning the context to learning good features)**.** *Suppose the second gradient step $\boldsymbol{q}_1$ has correlation coefficient $\ge\alpha$ with $\boldsymbol{q}_\star+y\boldsymbol{w}_\star$ for any $y\in\{\pm1\}$, i.e., $(\boldsymbol{q}_\star+y\boldsymbol{w}_\star)^\top\boldsymbol{q}_1\big/\|\boldsymbol{q}_1\|\ge\alpha Q$. Then, for any $\epsilon>0$ there exists sufficiently large $\gamma_*(\epsilon)$ and absolute constant $c>0$ such that setting $\boldsymbol{q}_1^\gamma=\gamma\frac{\boldsymbol{q}_1}{\|\boldsymbol{q}_1\|}$ for any $\gamma\ge\gamma_*(\epsilon)$, we have*

$$\mathbb{P}_{(\boldsymbol{X},y)}(\|\boldsymbol{X}^\top\phi(\boldsymbol{X}\boldsymbol{q}_1^\gamma)-(\boldsymbol{q}_\star+y\boldsymbol{w}_\star)\|\le\epsilon)\ge 1-2Te^{-c\frac{\alpha^2Q^2}{\sigma^2}}. \tag{52}$$

Proof. Recall $\boldsymbol{X}$ consists or ζT relevant and $(1-\zeta)T$ irrelevant tokens. For relevant tokens, we have by assumption that $B:=(\boldsymbol{q}_\star+y\boldsymbol{w}_\star)^\top\bar{\boldsymbol{q}}_1\ge\alpha Q$, where we denote $\bar{\boldsymbol{q}}_1:=\boldsymbol{q}_1/\|\boldsymbol{q}_1\|$, for convenience. Let $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_{(1-\zeta)T}$ denote the irrelevant tokens, which are σ-subgaussian. In order to ensure that softmax-attention perfectly selects the relevant tokens, we require

$$M:=\max_{1\le t\le(1-\zeta)T}\boldsymbol{z}_t^T\bar{\boldsymbol{q}}_1\le\alpha Q/2=B/2.$$

Condition on this event, which holds with at least probability $1-Te^{-\frac{c\alpha^2Q^2}{\sigma^2}}$. Then, the attention coefficients $a_t=[\phi(\gamma\boldsymbol{X}\boldsymbol{q}_1)]_t$ are as follows:

$$t \text{ relevant: } a_t=\frac{1}{\zeta T+(1-\zeta)Te^{\gamma(M-B)}}\ge\frac{1}{\zeta T+(1-\zeta)Te^{-\gamma B/2}}=:\frac{1}{\zeta T}a_R.$$

$$t \text{ irrelevant: } a_t=\frac{1}{\zeta Te^{\gamma(B-M)}+(1-\zeta)T}\le\frac{1}{\zeta Te^{\gamma B/2}+(1-\zeta)T}=:\frac{1}{(1-\zeta)T}a_I=\frac{1}{(1-\zeta)T}\cdot\frac{1}{1+\frac{\zeta}{1-\zeta}e^{\gamma B/2}}.$$

Therefore,

$$\|\boldsymbol{X}^\top\phi(\gamma\boldsymbol{X}\boldsymbol{q}_1)-(\boldsymbol{q}_\star+y\boldsymbol{w}_\star)\|\le(Q+W)(1-a_R)+a_I\max_{1\le t\le(1-\zeta)T}\|\boldsymbol{z}_t\|.$$

To continue, further condition on the event

$$\max_{1\le t\le(1-\zeta)T}\|\boldsymbol{z}_t\|\le C\sigma\sqrt{d}+\alpha Q \tag{53}$$

which holds with probability at least $1-(1-\zeta)Te^{-c\alpha^2Q^2/\sigma^2}$. Also note that $1-a_R=a_I$. Hence, with probability at least $1-2Te^{-c\alpha^2Q^2/(8\sigma^2)}$

$$\|\boldsymbol{X}^\top\phi(\gamma\boldsymbol{X}\boldsymbol{q}_1)-(\boldsymbol{q}_\star+y\boldsymbol{w}_\star)\|\le\left((1+\alpha)Q+W+C\sigma\sqrt{d}\right)a_I.$$

The right hand-side above can be made smaller than ϵ by choosing γ large enough (depending on $\epsilon,\alpha,Q,W,\sigma$ and d). This completes the proof. □

Combining Lemma 11 and Corollary 1 we arrive at the following result, which we state as a stand-alone theorem since it summarizes the effect of the first two-gradient steps on learning good (aka relevant) features.

Theorem 5 (Summary result of first two GD steps)**.** *Suppose Q, W and ρ are such that there exists positive constant $\alpha \in (0, 3/16)$ for which*

$$A_Q := \left(3/16 - \rho^2/8\right) Q - (9/4\,|\rho| + 1/16)\, W \geq \alpha \cdot Q\,.$$

Fix any $\epsilon > 0$. For sufficiently small step-size $\eta \lesssim Q^{-2}$, large enough n and sufficiently large $\gamma_\star = \gamma_\star(\epsilon)$ such that the following statements are true about the first two gradient steps:

$$\widehat{\boldsymbol{w}}_1 := \eta \widehat{\boldsymbol{G}}_{\boldsymbol{w}}(0,0)$$
$$\widehat{\boldsymbol{q}}_1^{\gamma} := \widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1), \quad \text{for any } \gamma \geq \gamma_\star\,.$$

There exist constants $c_0', c_0, c_1', c_1, c_3, c_3', c$ such that with probability at least

$$1 - c_0' \exp\left(-c_0\, n \left((W/Q)^2 \wedge \frac{\zeta^2 W^2 T}{d}\right)\right) - c_1' \exp\left(-c_1 n^{1/3} (W/Q)^{4/3} \zeta^{4/3} \big/ \log^{2/3}(n)\right)$$
$$- c_3' d \exp\left(-c_3 (n/d)^{1/3} (W/Q)^{4/3} \zeta^{4/3} \big/ \log^{2/3}(n)\right), \tag{54}$$

it holds for any fresh sample $(\boldsymbol{X}, y)$ that

$$\mathbb{P}(\|\boldsymbol{X}^\top \phi(\boldsymbol{X} \boldsymbol{q}_1^{\gamma}) - (\boldsymbol{q}_\star + y \boldsymbol{w}_\star)\| \leq \epsilon) \geq 1 - 2T e^{-c \frac{\alpha^2 Q^2}{\sigma^2}}. \tag{55}$$

Moreover, it holds that

$$\|\widehat{\boldsymbol{w}}_1\| \leq \frac{c \zeta W}{Q^2}\,.$$

C.4. Third gradient step

With the characterization of the quality of learnt features in Theorem 5, we are now ready to turn our attention to the third gradient step. For this last step, it turns out all we need is that $\widehat{\boldsymbol{w}}_2 := \eta \widehat{\boldsymbol{G}}_{\boldsymbol{w}}(\widehat{\boldsymbol{q}}_1, \widehat{\boldsymbol{w}}_1)$ has a strictly positive correlation with $\boldsymbol{w}_\star$. This is indeed the case and the result is formalized in the lemma below.

Lemma 12 (Third step)**.** *Suppose that the first and second gradient steps $\widehat{\boldsymbol{w}}_1, \widehat{\boldsymbol{q}}_1$ are such that the following hold. First, for absolute constant $c_\eta > 0$*

$$\|\widehat{\boldsymbol{w}}_1\| \leq c_\eta \zeta W / Q^2\,.$$

Second, for $\epsilon > 0$ and any fresh datapoint $(\boldsymbol{X}, y)$ there exists δ that does not depend on ϵ such that

$$\mathbb{P}_{(\boldsymbol{X}, y)}\left(\|\boldsymbol{X}^\top \phi(\boldsymbol{X} \widehat{\boldsymbol{q}}_1) - (\boldsymbol{q}_\star + y \boldsymbol{w}_\star)\| \leq \epsilon\right) \geq 1 - \delta.$$

Consider the third gradient step $\widehat{\boldsymbol{w}}_2 := \eta \widehat{\boldsymbol{G}}_{\boldsymbol{q}}(\widehat{\boldsymbol{q}}_1, \widehat{\boldsymbol{w}}_1)$. There exists absolute constants c, C such that, for all sufficiently small ϵ and c_η, with probability at least $1 - n\delta - 2e^{cn}$,

$$\frac{\widehat{\boldsymbol{w}}_2^\top \boldsymbol{w}_\star}{\|\widehat{\boldsymbol{w}}_2\|} \geq C\, \frac{W^2}{Q}\,.$$

Proof. Note that the lemma's conclusion is insensitive to the choice of step-size η for the third gradient step. Thus, without loss of generality, assume below that $\eta = 1$.

Recall the gradient formula

$$\widehat{\boldsymbol{G}}_{\boldsymbol{w}}(\boldsymbol{q}, \boldsymbol{w}) := -\nabla_{\boldsymbol{w}} \widehat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i \in [n]} (y_i - f_{\boldsymbol{\theta}}(\boldsymbol{X}_i)) \boldsymbol{X}_i^\top \phi((\boldsymbol{X}_i \boldsymbol{q}))$$

and denote for convenience

$$\boldsymbol{\epsilon}_i := \boldsymbol{X}_i^\top \phi(\boldsymbol{X}_i \boldsymbol{q}) - (\boldsymbol{q}_\star + y_i \boldsymbol{w}_\star), \qquad i \in [n].$$

With this notation, we can conveniently rewrite the third gradient step evaluated at $\boldsymbol{q} := \boldsymbol{q}^\gamma := \gamma \widehat{\boldsymbol{G}}_{\boldsymbol{q}}(0, \widehat{\boldsymbol{w}}_1)$ for any $\gamma > 0$ as follows:

$$
\begin{aligned}
\widehat{\boldsymbol{w}}_2 := \widehat{\boldsymbol{G}}_{\boldsymbol{w}}(\boldsymbol{q}, \widehat{\boldsymbol{w}}_1) &= \frac{1}{n}\sum_{i\in[n]} y_i \boldsymbol{X}_i^\top \phi(\boldsymbol{X}_i\boldsymbol{q}) - \frac{1}{n}\sum_{i\in[n]} \left(\widehat{\boldsymbol{w}}_1^T \boldsymbol{X}_i^\top \phi(\boldsymbol{X}_i\boldsymbol{q})\right) \boldsymbol{X}_i^\top \phi(\boldsymbol{X}_i\boldsymbol{q}) \\
&= \boldsymbol{w}_\star + \boldsymbol{q}_\star \left(\frac{1}{n}\sum_{i\in[n]} y_i\right) + \frac{1}{n}\sum_{i\in[n]} y_i \left(\boldsymbol{X}_i^\top \phi(\boldsymbol{X}_i\boldsymbol{q}) - (\boldsymbol{q}_\star + y_i\boldsymbol{w}_\star)\right) \\
&\qquad - \frac{1}{n}\sum_{i\in[n]} \left(\widehat{\boldsymbol{w}}_1^T \boldsymbol{X}_i^\top \phi(\boldsymbol{X}_i\boldsymbol{q})\right)(\boldsymbol{q}_\star + y_i\boldsymbol{w}_\star) - \frac{1}{n}\sum_{i\in[n]} \left(\widehat{\boldsymbol{w}}_1^T \boldsymbol{X}_i^\top \phi(\boldsymbol{X}_i\boldsymbol{q})\right)\left(\boldsymbol{X}_i^\top \phi(\boldsymbol{X}_i\boldsymbol{q}) - (\boldsymbol{q}_\star + y_i\boldsymbol{w}_\star)\right) \\
&= \boldsymbol{w}_\star + \boldsymbol{q}_\star \left(\frac{1}{n}\sum_{i\in[n]} y_i\right) + \frac{1}{n}\sum_{i\in[n]} y_i \boldsymbol{\epsilon}_i \\
&- \frac{1}{n}\sum_{i\in[n]} \widehat{\boldsymbol{w}}_1^\top \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i - \frac{1}{n}\sum_{i\in[n]} \widehat{\boldsymbol{w}}_1^\top (\boldsymbol{q}_\star + y_i\boldsymbol{w}_\star)\boldsymbol{\epsilon}_i - \frac{1}{n}\sum_{i\in[n]} \widehat{\boldsymbol{w}}_1^\top \boldsymbol{\epsilon}_i(\boldsymbol{q}_\star + y_i\boldsymbol{w}_\star) - \frac{1}{n}\sum_{i\in[n]} \widehat{\boldsymbol{w}}_1^\top (\boldsymbol{q}_\star + y_i\boldsymbol{w}_\star)(\boldsymbol{q}_\star + y_i\boldsymbol{w}_\star)\,. \quad (56)
\end{aligned}
$$

In order to control the correlation $\widehat{\boldsymbol{w}}_2^\top \boldsymbol{w}_\star / \|\widehat{\boldsymbol{w}}_2\|$ it suffices to control $\widehat{\boldsymbol{w}}_2^\top \boldsymbol{v}$ for arbitrary $\boldsymbol{v} \in \mathbb{R}^d$. In view of the expression above, it will suffice bounding the two random terms below:

$$
\begin{aligned}
\text{Term}_{\text{I}} &:= \left|\frac{1}{n}\sum_{i\in[n]} y_i\right| \\
\text{Term}_{\text{II}} &:= \|\boldsymbol{\epsilon}_i\| = \|\boldsymbol{X}_i^\top \phi(\boldsymbol{X}_i\boldsymbol{q}) - (\boldsymbol{q}_\star + y\boldsymbol{w}_\star)\|
\end{aligned}
$$

For the first term, we have with probability at least $1 - 2e^{u_1^2/2}$ that

$$\text{Term}_{\text{I}} \le u_1/\sqrt{n}.$$

For the second term, we know by assumption that with probability at least $1 - n\delta$,

$$\text{Term}_{\text{II}} \le \epsilon.$$

Putting the above together, with probability at least $1 - 2e^{-u^2/2} - n\delta$, we have that

$$
\begin{aligned}
\boldsymbol{w}_\star^\top \widehat{\boldsymbol{w}}_2 &\ge W^2 - |\rho|\, QW \frac{u_1}{\sqrt{n}} - \epsilon W - \frac{3}{2}\eta\zeta\left(\epsilon^2 W^2 + 2\epsilon W^2 (Q+W) + W^2 (Q+W)^2\right) \\
&\ge W^2 \left(1 - |\rho|\, \frac{Q}{W}\frac{u_1}{\sqrt{n}} - \frac{\epsilon}{W} - \frac{3}{2}\eta\zeta\left(\epsilon^2 + 2\epsilon(Q+W) + (Q+W)^2\right)\right)
\end{aligned}
$$

where we also used the lemma's assumption on $\|\widehat{\boldsymbol{w}}_1\| \le (3/2)\eta\zeta W$.

To further lower bound $\boldsymbol{w}_\star^\top \widehat{\boldsymbol{w}}_2$, recall that $|\rho| \le W/Q$, $W \gtrsim 1$ and that ϵ can be made arbitrarily small constant. Further pick

$$u_1 = c_1\sqrt{n} \qquad \text{and} \qquad \eta = c_\eta / Q^2 \quad (57)$$

for sufficiently small constants c_1 and c_η. With these, we guarantee with probability at least $1 - 2e^{-cn} - n\delta$ that

$$\widehat{\boldsymbol{w}}_2^\top \boldsymbol{w}_\star \gtrsim W^2\,. \quad (58)$$

Next, we use similar arguments to bound $\|\widehat{\boldsymbol{w}}_2\|$. Conditioning on the event where the bounds derived above hold for Term_{I} and Term_{II}, we have from (56) that

$$\|\widehat{\boldsymbol{w}}_2\| \le W + Q\frac{u_1}{\sqrt{n}} + \epsilon + \frac{3}{2}\eta\zeta W\left(\epsilon^2 + 2\epsilon(Q+W) + (Q+W)^2\right) \lesssim Q,$$

where in the second inequality, we chose u_1, η as in (57) and used again that ϵ is arbitrarily small constant, as well as, $Q > W \gtrsim 1$.

All the above combined, shows that

$$\frac{\widehat{\boldsymbol{w}}_2^\top \boldsymbol{w}_\star}{\|\widehat{\boldsymbol{w}}_2\|} \gtrsim \frac{W^2}{Q}.$$

This completes the proof. □

C.5. De-biasing step

Lemma 13 (Debiasing predictions)**.** *For some $\epsilon > 0$, suppose $\boldsymbol{q}_1$ is such that a test example $(y, \boldsymbol{X})$ satisfies*

$$\mathbb{P}_{(\boldsymbol{X},y)}\left(\|\boldsymbol{X}^\top \phi(\boldsymbol{X}\widehat{\boldsymbol{q}}_1) - (\boldsymbol{q}_\star + y\boldsymbol{w}_\star)\| \le \epsilon\right) \ge 1 - \delta$$

and that $\boldsymbol{w}_2$ is such that

$$\frac{\boldsymbol{w}_2^\top \boldsymbol{w}_\star}{\|\boldsymbol{w}_2\|} > 4\epsilon.$$

Given a fresh dataset $\mathcal{S} = (y_i, \boldsymbol{X}_i)_{i=1}^n$, set $b = \frac{1}{n}\sum_{i=1}^n f_{\boldsymbol{\theta}}(\boldsymbol{X}_i) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{w}_2^\top \boldsymbol{v}_i$ where $\boldsymbol{v}_i := \boldsymbol{X}_i^\top \phi(\boldsymbol{X}_i \boldsymbol{q}_1)$. Set the debiased classifier $f'_{\boldsymbol{\theta}}(\boldsymbol{X}) = f_{\boldsymbol{\theta}}(\boldsymbol{X}) - b$. Suppose $n \ge 8\log\left(\frac{2}{\delta n}\right)$. Then, with probability $1 - 2\delta n$ over $\mathcal{S}$, the test error of $f'_{\boldsymbol{\theta}}$ obeys

$$\mathrm{ERR}(f'_{\boldsymbol{\theta}}) \le \delta.$$

Proof. First, let us prove the following intermediate statement: With probability $1 - 2\delta n$ over $\mathcal{S}$, for a new test sample $(y, \boldsymbol{X})$, with probability $1 - \delta$,

$$|y f'_{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{w}_2^\top \boldsymbol{w}_\star| \le \sqrt{\frac{2\log(2/\delta n)}{n}} \cdot \boldsymbol{w}_2^\top \boldsymbol{w}_\star + 2\epsilon\|\boldsymbol{w}_2\|. \tag{59}$$

To see the above, start with observing that, with probability $1 - n\delta$ over the dataset $(y_i, \boldsymbol{X}_i)_{i=1}^n$, for each $\boldsymbol{v}_i$,

$$|\boldsymbol{w}_2^\top \boldsymbol{v}_i - \boldsymbol{w}_2^\top(\boldsymbol{q}_\star + y_i \boldsymbol{w}_\star)| \le \epsilon\|\boldsymbol{w}_2\|.$$

Set $\bar{b} = \boldsymbol{w}_2^\top \boldsymbol{q}_\star$ abd $\bar{y} = |\frac{1}{n}\sum_{i=1}^n y_i|$. With probability $1 - \delta n$, $\bar{y} \le \sqrt{\frac{2\log(2/\delta n)}{n}}$. Combining, with overall probability at least $1 - 2\delta n$, the classifier bias obeys

$$|b - \bar{b}| \le \left|\boldsymbol{w}_2^\top \boldsymbol{w}_\star\right| \sqrt{\frac{2\log(2/\delta n)}{n}} + \epsilon\|\boldsymbol{w}_2\|.$$

To finalize, for a new sample $(y, \boldsymbol{X})$, with probability $1 - \delta$, we have that $|\boldsymbol{w}_2^\top \boldsymbol{v} - \boldsymbol{w}_2^\top(\boldsymbol{q}_\star + y\boldsymbol{w}_\star)| \le \epsilon\|\boldsymbol{w}_2\|$ where $\boldsymbol{v} = \boldsymbol{X}^\top \phi(\boldsymbol{X}\boldsymbol{q}_1)$. Thus, the prediction $f'(\boldsymbol{X}) = f(\boldsymbol{X}) - b$ obeys

$$|y f'_{\boldsymbol{\theta}}(\boldsymbol{X}) - y(f_{\boldsymbol{\theta}}(\boldsymbol{X}) - \bar{b})| \le |b - \bar{b}| \le \left|\boldsymbol{w}_2^\top \boldsymbol{w}_\star\right| \sqrt{\frac{2\log(2/\delta n)}{n}} + \epsilon\|\boldsymbol{w}_2\|. \tag{60}$$

To conclude with (59), note that

$$|y(f_{\boldsymbol{\theta}}(\boldsymbol{X}) - \bar{b}) - \boldsymbol{w}_2^\top \boldsymbol{w}_\star| \le \epsilon\|\boldsymbol{w}_2\|,$$

and apply triangle inequality with (60).

To prove the statement of the lemma, note that, when $n \ge 8\log(2/\delta n)$ and $\boldsymbol{w}_2^\top \boldsymbol{w}_\star > 4\epsilon\|\boldsymbol{w}_2\|$, a test sample (with $\ge 1 - \delta$ probability) obeys

$$y f'_{\boldsymbol{\theta}}(\boldsymbol{X}) \ge \boldsymbol{w}_2^\top \boldsymbol{w}_\star - \sqrt{\frac{2\log(2/\delta n)}{n}} \boldsymbol{w}_2^\top \boldsymbol{w}_\star - 2\epsilon\|\boldsymbol{w}_2\| \ge 0.5\boldsymbol{w}_2^\top \boldsymbol{w}_\star - 2\epsilon\|\boldsymbol{w}_2\| > 0. \tag{61}$$

Thus, the classifier makes the correct decision with the same probability. □

C.6. Finishing the finite sample analysis

Theorem 6 (Main theorem: Finite-sample)**.** *Suppose Q, W and ρ are such that there exists positive constant $\alpha \in (0, 3/16)$ for which*

$$\left(3/16 - \rho^2/8\right) Q - \left(9/4\,|\rho| + 1/16\right) W \ge \alpha \cdot Q\,. \tag{62}$$

Fix any $\epsilon > 0$. For sufficiently small step-size $\eta \lesssim Q^{-2}$, sufficiently large step-size $\gamma = \gamma(\epsilon)$, and large enough n, there exist constants $c'_j, c_j, j = 0, 1, 2, 3, 4$ such that the following statements hold with probability at least

$$1 - c'_0 \exp\left(-c_0\, n\left((W/Q)^2 \wedge \left(\frac{\zeta^2 W^2 T}{d} \wedge 1\right)\right)\right) - c'_1 \exp\left(-c_1 n^{1/3}(W/Q)^{4/3}\zeta^{4/3}/\log^{2/3}(n)\right)$$
$$- c'_3 d \exp\left(-c_3 (n/d)^{1/3}(W/Q)^{4/3}\zeta^{4/3}/\log^{2/3}(n)\right) - 2nT\exp\left(-c_4\alpha^2 Q^2\right), \tag{63}$$

over the training set:

1. Prompt attends to relevant tokens: *For any test sample $(\boldsymbol{X}, y)$, with probability at least $1 - 2T\exp\left(-c_4\alpha^2 Q^2\right)$, the attention coefficients $a_t = [\phi(\boldsymbol{X}\widehat{\boldsymbol{q}}_1)]_t$ after the second gradient step satisfy:*

$$a_t \begin{cases} \geq \frac{1-\epsilon}{\zeta T} & t \text{ relevant} \\ \leq \frac{\epsilon}{(1-\zeta)T} & t \text{ irrelevant}. \end{cases} \tag{64}$$

2. Prompt learns relevant features: *The prompt attention mechanism outputs relevant tokens with the same probability. Concretely,*

$$\mathbb{P}_{(\boldsymbol{X},y)}\left(\|\boldsymbol{X}^\top\phi(\boldsymbol{X}\widehat{\boldsymbol{q}}_1) - (\boldsymbol{q}_\star + y\boldsymbol{w}_\star)\| \leq \epsilon\right) \geq 1 - 2T\exp\left(-c_4\alpha^2 Q^2\right).$$

3. Test error: *The test error of the model $f'_{\boldsymbol{\theta}}$ satisfies*

$$\mathrm{ERR}(f'_{\boldsymbol{\theta}}) \leq 2T\exp\left(-c_4\alpha^2 Q^2\right).$$

Proof. The theorem follows by combining Theorem 5, Lemma 12 and Lemma 13. □

D. Proofs for Population-gradient Analysis in Section 4.2

This section includes the missing proofs of all the results in Section 4.2 regarding population analysis of Algorithm 9.

D.1. Proof of Lemma 2

We repeat here the lemma for convenience also stated for general (not necessarily isotropic) noise covariance $\boldsymbol{\Sigma}$.

Lemma 14. *The second population gradient step $\boldsymbol{q}_1 = \gamma \boldsymbol{G}_{\boldsymbol{w}}(\boldsymbol{w}_1, 0)$ satisfies the following for $\alpha := \eta\zeta$*

$$\begin{aligned}\boldsymbol{G}_{\boldsymbol{q}}(0, \alpha\boldsymbol{w}_\star) = &\left((\zeta - \zeta^2)\left(\alpha W^2 + \alpha^2 \boldsymbol{w}_\star^\top \boldsymbol{\Sigma}\boldsymbol{w}_\star/T\right) - \alpha^2(\zeta^2 - \zeta^3)\left(W^4 + \left(\boldsymbol{w}_\star^\top\boldsymbol{q}_\star\right)^2\right)\right)\boldsymbol{q}_\star \\ &+ \left(\left((\zeta - \zeta^2) - 2(\zeta^2 - \zeta^3)\alpha W^2\right)\alpha\left(\boldsymbol{w}_\star^\top\boldsymbol{q}_\star\right)\right)\boldsymbol{w}_\star \\ &- \left((1 + 2/T)(\zeta - \zeta^2)\alpha\left(\boldsymbol{w}_\star^\top\boldsymbol{q}_\star\right)\right)\alpha\boldsymbol{\Sigma}\boldsymbol{w}_\star \end{aligned} \tag{65}$$

Proof. The lemma follows immediately from Eqn. (19) of Lemma 5 by recognizing that for $\boldsymbol{w} = \alpha\boldsymbol{w}_\star$ it holds $R_{\boldsymbol{q}_\star} = \alpha\boldsymbol{q}_\star^\top\boldsymbol{w}_\star$ and $R_{\boldsymbol{w}_\star} = \alpha W^2$. □

D.2. Corollary 2

Corollary 2. *Suppose small enough step-size η obeying*

$$\eta\left(\zeta^2\left(W^2 + Q^2\right) - \zeta\cdot\sigma^2/T\right) \leq 1/2. \tag{66a}$$
$$\eta\zeta\left(2\zeta W^2 + (1 + 2/T)\sigma^2\right) \leq 5/4. \tag{66b}$$
$$\eta\zeta\left(\sigma^2/T\right) \leq 1/2. \tag{66c}$$

Then, for $C_1 \in [1/2, 3/2]$ and $C_2 \in [-1/4, 1]$, we have that

$$\boldsymbol{q}_1 = \gamma\eta\zeta(\zeta - \zeta^2)W\left(C_1 W\boldsymbol{q}_\star + C_2\rho Q\boldsymbol{w}_\star\right).$$

In particular, $\boldsymbol{q}_\star^\top\boldsymbol{q}_1 = \gamma\eta\zeta(\zeta - \zeta^2)W^2Q^2\left(C_1 + C_2\rho^2\right)$ and $\boldsymbol{w}_\star^\top\boldsymbol{q}_1 = \gamma\eta\zeta(\zeta - \zeta^2)W^3Q\rho\left(C_1 + C_2\right)$.

Proof. Set $\alpha = \eta\zeta$ and

$$3/2 \geq C_1 := \left(1 + \alpha\sigma^2/T\right) - \alpha\zeta\left(W^2 + \rho^2 Q^2\right) \geq 1/2. \tag{67a}$$

$$1 \geq C_2 := 1 - 2\alpha\zeta W^2 - (1 + 2/T)\alpha\sigma^2 \geq -1/4. \tag{67b}$$

The gradient formula follows directly from (10). For the lower/upper bounds on C_1, C_2 use (66a), (66b) and (66c). □

Remark 2 (Condition on correlation)**.** *To classify correctly the signal tokens, we need*

$$y\boldsymbol{w}_\star^\top(\boldsymbol{q}_\star + y\boldsymbol{w}_\star) > 0 \quad \Longleftrightarrow \quad y\rho Q + W > 0 \Longleftrightarrow \quad |\rho| < W/Q \tag{68}$$

Note that if (68) *holds, then*

$$C_1 \geq 1 + \alpha\sigma^2/T - 2\alpha\zeta W^2 = C_2 + (1 + 3/T)\alpha\sigma^2. \tag{69}$$

D.3. Proof of Theorem 2

We start by showing that for any $\epsilon > 0$, and all sufficiently large $\gamma \geq \gamma_\star(\epsilon)$, it holds

$$\mathbb{P}_{(\boldsymbol{X},y)\sim\mathcal{D}}(\|\boldsymbol{X}^\top\phi(\boldsymbol{X}\boldsymbol{q}_1^\gamma) - (\boldsymbol{q}_\star + y\boldsymbol{w}_\star)\| \leq \epsilon) \geq 1 - 2Te^{-c\frac{\alpha^2 Q^2}{\sigma^2}} =: 1 - \delta. \tag{70}$$

We can get this by applying Lemma 11 provided only that we show the correlation of the normalized gradient-step $\boldsymbol{q}_1^\gamma/\|\boldsymbol{q}_1^\gamma\|$ (= $\bar{\boldsymbol{q}}_1$ below) with signal-relevant tokens is at least αQ. Concretely, we have from Corollary 2 that $\boldsymbol{q}_1^\gamma = \gamma\boldsymbol{q}_1 := \gamma\eta\zeta(\zeta - \zeta^2)W\left(C_1 W\boldsymbol{q}_\star + C_2\rho Q\boldsymbol{w}_\star\right)$ with $3/2 \geq C_1 \geq 1/2$, $1 \geq C_2 \geq -1/4$. Define for convenience $\boldsymbol{q}_1 := \eta\zeta(\zeta - \zeta^2)W\left(C_1 W\boldsymbol{q}_\star + C_2\rho Q\boldsymbol{w}_\star\right)$ and consider the normalized gradient step

$$\bar{\boldsymbol{q}}_1 := \frac{\boldsymbol{q}_1}{\|\boldsymbol{q}_1\|_2} = \frac{C_1 W\boldsymbol{q}_\star + C_2\rho Q\boldsymbol{w}_\star}{\sqrt{C_1^2 W^2 Q^2 + C_2^2\rho^2 Q^2 W^2 + 2C_1 C_2\rho^2 W^2 Q^2}} = \frac{C_1 W\boldsymbol{q}_\star + C_2\rho Q\boldsymbol{w}_\star}{QW\sqrt{C_1^2 + C_2\rho^2(C_2 + 2C_1)}}\,.$$

We can lower-bound its correlation with a signal token $\boldsymbol{q}_\star + y\boldsymbol{w}_\star$ as follows:

$$\begin{aligned}
\bar{\boldsymbol{q}}_1^T(\boldsymbol{q}_\star + y\boldsymbol{w}_\star) &= \frac{C_1 W(Q^2 + y\rho WQ) + C_2\rho Q(\rho WQ + yW^2)}{QW\sqrt{C_1^2 + C_2\rho^2(C_2 + 2C_1)}} = \frac{C_1(Q + y\rho W) + C_2\rho(\rho Q + yW)}{\sqrt{C_1^2 + C_2\rho^2(C_2 + 2C_1)}} \\
&= \frac{(C_1 + C_2\rho^2)Q + y\rho(C_1 + C_2)W)}{\sqrt{C_1^2 + C_2\rho^2(C_2 + 2C_1)}} \\
&\geq \frac{(C_1 + C_2\rho^2)Q + y\rho(C_1 + C_2)W)}{C_1\sqrt{1 + 3\rho^2}} \geq \frac{1 + (C_2/C_1)\rho^2}{\sqrt{1 + 3\rho^2}}Q - \frac{|\rho|\,(1 + C_2/C_1)}{\sqrt{1 + 3\rho^2}} \\
&\geq \frac{\left(1 + (C_2/C_1)\rho^2\right)Q - \left(|\rho|\,(1 + C_2/C_1)\right)W}{\sqrt{1 + 3\rho^2}} \\
&\geq \frac{(1 - \rho^2/2)Q - 2\,|\rho|\,W}{\sqrt{1 + 3\rho^2}} \\
&\geq \alpha Q\,,
\end{aligned}$$

where: (i) the inequality $\sqrt{C_1^2 + \rho^2 C_2(C_2 + 2C_1)} \leq C_1\sqrt{1 + 3\rho^2}$ used in in the third line follows because $C_1 > 0$ and $C_2 \leq C_1$ from (69); (ii) the penultimate inequality uses $C_2/C_1 \in [-1/2, 1]$ (for the lower bound recall $C_2 \geq -1/4, C_1 \geq 1/2$); (iii) the last inequality is because of the theorem's assumption in (13).

Next, recall that

$$\boldsymbol{w}_2^\gamma := \boldsymbol{G}_{\boldsymbol{w}}(0, \boldsymbol{q}_1^\gamma) = \mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}}\left[y\boldsymbol{X}^\top\phi(\boldsymbol{X}\boldsymbol{q}_1^\gamma)\right] = \mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}}\left[y\boldsymbol{v}(\boldsymbol{X})\right]\,, \tag{71}$$

where we set $\boldsymbol{v}(\boldsymbol{X}) = \boldsymbol{X}^\top\phi(\boldsymbol{X}\boldsymbol{q}_1^\gamma)$ for convenience. Thus, the prediction of the model with parameters $\boldsymbol{\theta} = (\boldsymbol{w}_2^\gamma, \boldsymbol{q}_1^\gamma)$ for test datapoint $(\widetilde{\boldsymbol{X}}, \widetilde{y})$ is

$$\hat{o} := \widetilde{y} f_{\boldsymbol{\theta}}(\widetilde{\boldsymbol{X}}) = \widetilde{y}\langle\boldsymbol{w}_2^\gamma, \boldsymbol{v}(\widetilde{\boldsymbol{X}})\rangle = \underbrace{\langle\boldsymbol{w}_2^\gamma, \widetilde{y}\boldsymbol{q}_\star + \boldsymbol{w}_\star\rangle}_{\hat{o}_1} + \underbrace{\widetilde{y}\langle\boldsymbol{w}_2^\gamma, \boldsymbol{v}(\widetilde{\boldsymbol{X}}) - (\boldsymbol{q}_\star + \widetilde{y}\boldsymbol{w}_\star)\rangle}_{\hat{o}_2}\,. \tag{72}$$

Let $\mathcal{E}$ denote the event for which $\|\boldsymbol{v}(\widetilde{\boldsymbol{X}}) - (\boldsymbol{q}_\star + \widetilde{y}\boldsymbol{w}_\star)\| \le \epsilon$, which has probability at least $1-\delta$ by (70). Note that

$$\mathrm{ERR}(f_{\boldsymbol{\theta}}) = \mathbb{P}\left(\hat{o} < 0\right) \le \mathbb{P}\left(\hat{o} < 0 \,|\, \mathcal{E}\right) + \Pr(\mathcal{E}) = \mathbb{P}\left(\hat{o} < 0 \,|\, \mathcal{E}\right) + \delta\,. \tag{73}$$

Hence, our goal below is to bound $\mathbb{P}\left(\hat{o} < 0 \,|\, \mathcal{E}\right)$. In fact, we will show that $\mathbb{P}\left(\hat{o} < 0 \,|\, \mathcal{E}\right) \le 0$, so the error rate is δ as stated in the theorem.

To do this, condition on $\mathcal{E}$ for which $|\hat{o}_2| \le \epsilon\|\boldsymbol{w}_2^\gamma\|$, thus

$$\hat{o} \ge \hat{o}_1 - \epsilon\|\boldsymbol{w}_2^\gamma\|\,.$$

Further note that $\hat{o}_1 = \langle \boldsymbol{w}_2^\gamma, \boldsymbol{w}_\star\rangle - |\langle \boldsymbol{w}_2^\gamma, \boldsymbol{q}_\star\rangle|$. Thus, it suffices to show that

$$\langle \boldsymbol{w}_2^\gamma, \boldsymbol{w}_\star\rangle \ge |\langle \boldsymbol{w}_2^\gamma, \boldsymbol{q}_\star\rangle| + \epsilon\|\boldsymbol{w}_2^\gamma\|\,. \tag{74}$$

For this, go back to $\boldsymbol{w}_2^\gamma$ and write continuing from (71)

$$\begin{aligned}\boldsymbol{w}_2^\gamma &= \mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}}\left[y\boldsymbol{q}_\star + \boldsymbol{w}_\star\right] + \mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}}\left[y\left(\boldsymbol{v}(\boldsymbol{X}) - (\boldsymbol{q}_\star + y\boldsymbol{w}_\star)\right)\right]\\ &= \boldsymbol{w}_\star + \mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}}\left[y\left(\boldsymbol{v}(\boldsymbol{X}) - (\boldsymbol{q}_\star + y\boldsymbol{w}_\star)\right)\right]\,.\end{aligned}$$

Denote for convenience $\boldsymbol{e} := \boldsymbol{e}(y, \boldsymbol{X}) := \boldsymbol{v}(\boldsymbol{X}) - (\boldsymbol{q}_\star + y\boldsymbol{w}_\star)$. Thus,

$$\begin{aligned}\langle \boldsymbol{w}_2^\gamma, \boldsymbol{w}_\star\rangle &= W^2 + \mathbb{E}\left[y\langle \boldsymbol{w}_\star, \boldsymbol{e}\rangle\right]\\ \langle \boldsymbol{w}_2^\gamma, \boldsymbol{q}_\star\rangle &= \rho QW + \mathbb{E}\left[y\langle \boldsymbol{q}_\star, \boldsymbol{e}\rangle\right]\\ \|\boldsymbol{w}_2^\gamma\| &\le W + \mathbb{E}\left[\|\boldsymbol{e}\|\right]\,.\end{aligned}$$

where in the last line we used triangle and Jensen's inequalities. We now compute (recall $\mathcal{E}$ is the event for which $\|\boldsymbol{e}\| \le \epsilon$):

$$\begin{aligned}\mathbb{E}\left[|\langle \boldsymbol{w}_\star, \boldsymbol{e}\rangle|\right] &\le \mathbb{E}\left[|\langle \boldsymbol{w}_\star, \boldsymbol{e}\rangle| \,|\, \mathcal{E}\right] + \mathbb{E}\left[|\langle \boldsymbol{w}_\star, \boldsymbol{e}\rangle| \,|\, \mathcal{E}^c\right]\mathbb{P}(\mathcal{E})\\ &\le \epsilon W + W\,\delta\,\mathbb{E}\left[\|\boldsymbol{e}\| \,|\, \mathcal{E}^c\right]\,.\end{aligned} \tag{75}$$

Similarly, we can upper bound $\mathbb{E}\left[|\langle \boldsymbol{q}_\star, \boldsymbol{e}\rangle|\right]$ and $\mathbb{E}\left[\|\boldsymbol{e}\|\right]$. Combining these with the above displays, the desired Eq. (74) holds provided:

$$W^2 - |\rho|\,QW \ge \epsilon\,(W+Q) + \delta\,\left(W+Q\right)\,B + \epsilon\delta B + \epsilon^2\,. \tag{76}$$

Above, we have denoted $B := \mathbb{E}\left[\|\boldsymbol{e}\| \,|\, \mathcal{E}^c\right]$. $\|\boldsymbol{w}_2^\gamma\| \le W + B$. Note that the LHS of 76 is > 0 because of Assumption 3.a that $|\rho| \le W/Q$. Thus, we can guarantee (76) holds once ϵ is small enough (by making γ large enough) and δ is also small enough (by making γ large enough). It only remains to bound B. To do this, note that

$$\|\boldsymbol{e}\|_2 \le \|\boldsymbol{v}(\boldsymbol{X})\| + Q + W \le \max_{t\in[T]}\|\boldsymbol{x}_t\| + Q + W \le 2(Q+W) + \max_{t\in\mathcal{R}^c}\|\boldsymbol{z}_t\|\,.$$

By Lemma 15 we further have that

$$\mathbb{E}\left[\max_{t\in\mathcal{R}^c}\|\boldsymbol{z}_t\|_2 \middle| \mathcal{E}^c\right]\mathbb{P}(\mathcal{E}^c) \le \delta\cdot C\sigma\sqrt{d}\sqrt{\log\left(2T/\delta\right)}\,.$$

Hence,

$$B \le 2(Q+W) + C\sigma\sqrt{d}\sqrt{\log\left(2T/\delta\right)}\,.$$

D.3.1. Auxiliary lemma

Lemma 15 (Subgaussian euclidean-norm tail control)**.** *Let Let $\boldsymbol{z}_i \in \mathbb{R}^d, i \in [N]$ be K-subgaussian random vectors. Then, for any event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) = \delta$, it holds that*

$$\mathbb{E}\left[\max_{i\in[N]}\|\boldsymbol{z}_i\| \middle| \mathcal{E}^c\right] \le 12K\sqrt{d}\sqrt{\log\left(2N/\delta\right)}.$$

Proof. Set $Z = \max_{i\in[n]} \|\boldsymbol{z}_i\|$ and define event $\mathcal{B} = \{Z \geq M\}$ for $M := 4K\sqrt{d}\sqrt{\log(2N/\delta)}$. By Fact G.4 for all $t > 0$, $\mathbb{P}(Z > t) \leq 2Ne^{-t^2/(16dK^2)}$. Thus, by choice of M, $\mathbb{P}(\mathcal{B}) \leq \mathbb{P}(\mathcal{E}) = \delta$.

Denote the pdf and cdf complement of Z by f_Z, Q_Z respectively. Observe that, we set $Q_Z(M) \leq \delta$. Using integration by parts we have,

$$\begin{aligned}
\mathbb{E}[Z|\mathcal{B}]\mathbb{P}(\mathcal{B}) &= \int_M^\infty z f_Z(z)\mathrm{d}z = -\int_M^\infty z dQ_Z(z) = \int_M^\infty Q_Z(z)\mathrm{d}z - [Q_Z(z)z]_M^\infty \\
&= \int_M^\infty Q_Z(z)\mathrm{d}z + \delta M \\
&= \delta M + \int_M^\infty \mathbb{P}(Z \geq t)\mathrm{d}t \leq \delta M + \int_M^\infty 2Ne^{-t^2/(16dK^2)}\mathrm{d}t \\
&\leq \delta M + 2\sqrt{2}K\sqrt{d}\,(2N) \int_{\sqrt{2\log(2N/\delta)}}^\infty e^{-u^2/2}\mathrm{d}u \\
&= \delta\, 4K\sqrt{d}\sqrt{\log(2N/\delta)} + 2\sqrt{\pi}K\sqrt{d}\,\delta \leq 2\delta M.
\end{aligned}$$

We can conclude the proof by noting:

$$\begin{aligned}
\mathbb{E}\left[Z\middle|\mathcal{E}\right]\mathbb{P}(\mathcal{E}) &= \mathbb{E}[Z\middle|\mathcal{E}\cap\mathcal{B}^c]\mathbb{P}(\mathcal{E}\cap\mathcal{B}^c) + \mathbb{E}[Z\middle|\mathcal{E}\cap\mathcal{B}]\mathbb{P}(\mathcal{E}\cap\mathcal{B}) \\
&\leq M\delta + \mathbb{E}[Z\middle|\mathcal{B}]\mathbb{P}(\mathcal{B})\,.
\end{aligned}$$

□

E. Proofs of results on discrete datasets

E.1. Proof of Theorem 1 and Observation 1

• **Proof for Prompt-attention:** Let $\bar{\boldsymbol{w}}_\star = \boldsymbol{w}_\star/\|\boldsymbol{w}_\star\|$ and $\bar{\boldsymbol{q}}_\star = \boldsymbol{q}_\star/\|\boldsymbol{q}_\star\|$. $\boldsymbol{q}'_\star$ be the projection of $\boldsymbol{q}_\star$ to the orthogonal complement of $\boldsymbol{w}_\star$ i.e. $\boldsymbol{q}'_\star = \boldsymbol{q}_\star - \bar{\boldsymbol{w}}_\star\bar{\boldsymbol{w}}_\star^\top\boldsymbol{q}_\star$. Similarly, let $\boldsymbol{w}'_\star$ be the projection of $\boldsymbol{w}_\star$ to the orthogonal complement of $\boldsymbol{q}_\star$ i.e. $\boldsymbol{w}'_\star = \boldsymbol{w}_\star - \bar{\boldsymbol{q}}_\star\bar{\boldsymbol{q}}_\star^\top\boldsymbol{w}_\star$. Denote correlation coefficient between two vectors by $\rho(\boldsymbol{a},\boldsymbol{b}) = \frac{\boldsymbol{a}^\top\boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|}$.

To proceed, observe that, $\boldsymbol{q}_\star'^\top\boldsymbol{q}_\star = \|\boldsymbol{q}_\star\|^2 - (\bar{\boldsymbol{w}}_\star^\top\boldsymbol{q}_\star)^2 = \|\boldsymbol{q}_\star\|^2(1-\rho(\boldsymbol{q}_\star,\boldsymbol{w}_\star)^2) > 0$. The positivity follows from the fact that $\boldsymbol{q}_\star, \boldsymbol{w}_\star$ are not parallel, thus, the absolute value of their correlation coefficient is strictly bounded away from 1. Similarly $\boldsymbol{w}_\star'^\top\boldsymbol{w}_\star = \|\boldsymbol{w}_\star\|^2(1-\rho(\boldsymbol{q}_\star,\boldsymbol{w}_\star)^2)>0$. To proceed, set $\bar{\rho} := 1-\rho(\boldsymbol{q}_\star,\boldsymbol{w}_\star)^2$ and observe that the classifier $\boldsymbol{\theta} = (\boldsymbol{w}'_\star, \Gamma\boldsymbol{q}'_\star)$ achieves the attention scores

$$\boldsymbol{a}_i = \phi(\boldsymbol{X}\boldsymbol{q}'_\star)_i = \begin{cases} S^{-1}e^{\|\boldsymbol{q}_\star\|^2\Gamma\bar{\rho}} & \text{if} \quad i \text{ relevant}\,, \\ S^{-1}e^{-\|\boldsymbol{q}_\star\|^2\Gamma\delta^q\bar{\rho}} & \text{if} \quad i \text{ irrelevant}\,, \end{cases}$$

where $S = T\zeta e^{\|\boldsymbol{q}_\star\|^2\Gamma\bar{\rho}} + T(1-\zeta)e^{-\|\boldsymbol{q}_\star\|^2\Gamma\delta^q\bar{\rho}}$. Using orthogonality of $\boldsymbol{w}'_\star$ and $\boldsymbol{q}_\star$, the final prediction obeys

$$yf_{\boldsymbol{\theta}}(\boldsymbol{X}) = \|\boldsymbol{w}_\star\|^2\bar{\rho}S^{-1}\left[\zeta e^{\|\boldsymbol{q}_\star\|^2\Gamma\bar{\rho}} - \delta^w(1-\zeta)e^{-\|\boldsymbol{q}_\star\|^2\Gamma\delta^q\bar{\rho}}\right].$$

The classifier achieves perfect accuracy when $\zeta e^{\|\boldsymbol{q}_\star\|^2\Gamma\bar{\rho}} > |\delta^w|(1-\zeta)e^{-\|\boldsymbol{q}_\star\|^2\Gamma\delta^q\bar{\rho}}$. Since we have $\delta^q \geq 0$ and we have assumed δ^w is a C-bounded variable (i.e. $|\delta^w| \leq C$), thus, the desired inequality can be guaranteed by choosing

$$\Gamma > \frac{1}{\|\boldsymbol{q}_\star\|^2\bar{\rho}}\log\left(\frac{C(1-\zeta)}{\zeta}\right).$$

• **Proof for Observation 1:** To prove this, observe that for any $\delta^q = \Delta^q$, $\delta^w = \Delta^w$ choices, using orthogonality of $\boldsymbol{q}_\star, \boldsymbol{w}'_\star$, for any $(y, \boldsymbol{X}) \sim \mathcal{D}$, we have

$$yf^{\text{LIN}}(\boldsymbol{w}'_\star) = \|\boldsymbol{w}_\star\|^2\bar{\rho}(\zeta - (1-\zeta)\delta^w).$$

Thus, as long as $\delta^w \neq \zeta/(1-\zeta)$, $\texttt{sign}(yf^{\text{LIN}}(\boldsymbol{w}'_\star))$ is always 1 or always -1, resulting in perfect accuracy for $\boldsymbol{w}'_\star$ or $-\boldsymbol{w}'_\star$.

• **Proof for Self-attention:** The proof is provided under Theorem 7.

• **Proof for Linear Prompt-attention:** Let $W_1 = \boldsymbol{w}^\top \boldsymbol{w}_\star, W_2 = \boldsymbol{w}^\top \boldsymbol{q}_\star, Q_1 = \boldsymbol{q}^\top \boldsymbol{q}_\star, Q_2 = \boldsymbol{q}^\top \boldsymbol{w}_\star$. Since context-irrelevant tokens are of the form $-\delta^q \boldsymbol{q}_\star - y\delta^w \boldsymbol{w}_\star$, the model decision is given by $\frac{1}{T} f(\boldsymbol{X}) = \frac{1}{T} \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{q} = \zeta \boldsymbol{w}^\top (y\boldsymbol{w}_\star + \boldsymbol{q}_\star)(y\boldsymbol{w}_\star + \boldsymbol{q}_\star)^\top \boldsymbol{q} + (1-\zeta)\boldsymbol{w}^\top (y\delta^w \boldsymbol{w}_\star + \delta^q \boldsymbol{q}_\star)(y\delta^w \boldsymbol{w}_\star + \delta^q \boldsymbol{q}_\star)^\top \boldsymbol{q}$ and

$$\begin{aligned}
\frac{1}{T} f(\boldsymbol{X}) &= \zeta(yW_1 + W_2)(Q_1 + yQ_2) + (1-\zeta)(y\delta^w W_1 + \delta^q W_2)(y\delta^w Q_2 + \delta^q Q_1) \\
&= \zeta y(W_1 Q_1 + W_2 Q_2) + \zeta(W_2 Q_1 + W_1 Q_2) + \\
&\quad (1-\zeta) y \delta^q \delta^w (W_1 Q_1 + W_2 Q_2) + (1-\zeta)(\delta^{q^2} W_2 Q_1 + \delta^{w^2} W_1 Q_2). \\
\frac{y f(\boldsymbol{X})}{T} &= (\zeta + (1-\zeta)\delta^q \delta^w)(W_1 Q_1 + W_2 Q_2) + y((\zeta + (1-\zeta)\delta^{q^2}) W_2 Q_1 + (\zeta + (1-\zeta)\delta^{w^2}) W_1 Q_2).
\end{aligned}$$

To proceed, set (δ^q, δ^w) to be $(0,0)$ or $(\Delta, -\Delta)$ equally-likely for $\Delta > \sqrt{\zeta/(1-\zeta)}$. For fixed Δ, for any choice of W_1, W_2, Q_1, Q_2 observe that, with $1/2$ probability the event $E = \{y((\zeta + (1-\zeta)\delta^{q^2}) W_2 Q_1 + (\zeta + (1-\zeta)\delta^{w^2}) W_1 Q_2) \le 0\}$ happens. On this event (which is over the label y), probability that $(\zeta + (1-\zeta)\delta^q \delta^w)(W_1 Q_1 + W_2 Q_2) > 0$ is at most $1/2$ because $\text{sign}(\zeta + (1-\zeta)\delta^q \delta^w)$ is Rademacher variable. Combining, we find that $\mathbb{P}(\frac{y f(\boldsymbol{X})}{T} \le 0) \ge 25\%$ as advertised whenever $\Delta > \sqrt{\zeta/(1-\zeta)}$.

E.2. Failure proof for Self-attention

We have the following theorem regarding self-attention.

Theorem 7. *Fix $\Delta > 0$ to be sufficiently large. In (DATA), choose $\delta = (\delta^q, \delta^w)$ to be $(0,0)$ or (Δ, Δ) equally-likely, where $\Delta > 1/(1-\zeta)^2$.*

- *For any choice of $(\boldsymbol{U} = \mathbb{1}\boldsymbol{u}^\top, \boldsymbol{W})$, $f^{\text{SATT}}(\mathbb{1}\boldsymbol{u}^\top, \boldsymbol{W})$ achieves 50% accuracy (i.e. random guess).*
- *For any choice of $(\boldsymbol{U}, \boldsymbol{W})$, there exists a (DATA) distribution with adversarial relevance set choices such that $f^{\text{SATT}}(\boldsymbol{U}, \boldsymbol{W})$ achieves 50% accuracy.*

Here, adversarial relevance set choice means that, the relevance set can be chosen adaptively to the label y, out-of-context term δ, and the self-attention model weights $(\boldsymbol{U}, \boldsymbol{W})$ to cause misclassification.

Proof. Let $\widetilde{\boldsymbol{w}} = \boldsymbol{W}\boldsymbol{w}_\star$ and $\widetilde{\boldsymbol{q}} = \boldsymbol{W}\boldsymbol{q}_\star$. Also let $b_w = \boldsymbol{u}^\top \boldsymbol{w}_\star$ and $b_q = \boldsymbol{u}^\top \boldsymbol{q}_\star$. Since $\boldsymbol{W}$ is allowed to be full-rank and arbitrary, $\widetilde{\boldsymbol{w}}, \widetilde{\boldsymbol{q}}$ are allowed to be arbitrary as well (but fixed given $\boldsymbol{W}$). In our analysis, the critical terms are the attention weights given by the correlation between the relevant/irrelevant keys/queries.

Setting attention queries as the raw tokens (without losing any generality), relevant queries $\boldsymbol{x}_R$ and keys $\boldsymbol{k}_R$ become

$$\boldsymbol{x}_R = y\boldsymbol{w}_\star + \boldsymbol{q}_\star, \ \boldsymbol{k}_R = y\widetilde{\boldsymbol{w}} + \widetilde{\boldsymbol{q}}.$$

Thanks to our choice of $\delta := \delta^w = \delta^q$ to be equally-likely in $\{0, \Delta\}$, observe that irrelevant queries and keys are simply

$$\boldsymbol{x}_I = -\delta \boldsymbol{x}_R, \ \boldsymbol{k}_I = -\delta \boldsymbol{k}_R.$$

This will greatly help the proof because it will mean that attention weights are highly structured. Specifically, set $\rho = \boldsymbol{x}_R^\top \boldsymbol{k}_R$. All weights of the attention similarities belong to the set $(\rho, -\delta\rho, \delta^2\rho)$. Consequently, softmax-attention output $\boldsymbol{A} = \phi(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^\top)\boldsymbol{X} = \begin{bmatrix} \boldsymbol{a}_1^\top \\ \vdots \\ \boldsymbol{a}_T^\top \end{bmatrix}$ is given by

$$\boldsymbol{a}_i = \begin{cases} \frac{\zeta e^{\rho} - \delta(1-\zeta)e^{-\delta\rho}}{\zeta e^{\rho} + (1-\zeta)e^{-\delta\rho}} \cdot \frac{1}{T} \cdot \boldsymbol{x}_R & \text{if} \quad i \in \mathcal{R} \text{ (relevant)} \\ \frac{\zeta e^{-\delta\rho} - \delta(1-\zeta)e^{\delta^2\rho}}{\zeta e^{-\delta\rho} + (1-\zeta)e^{\delta^2\rho}} \cdot \frac{1}{T} \cdot \boldsymbol{x}_R & \text{if} \quad i \in \mathcal{R}^c \text{ (irrelevant)} \end{cases}. \tag{77}$$

Set $a_+ = \frac{e^{\rho}}{\zeta e^{\rho}+(1-\zeta)e^{-\delta\rho}}$, $a_- = \frac{e^{-\delta\rho}}{\zeta e^{\rho}+(1-\zeta)e^{-\delta\rho}}$, $b_- = \frac{e^{-\delta\rho}}{\zeta e^{-\delta\rho}+(1-\zeta)e^{\delta^2\rho}}$, $b_+ = \frac{e^{\delta^2\rho}}{\zeta e^{-\delta\rho}+(1-\zeta)e^{\delta^2\rho}}$. With this, we also set

$$\Delta_R = \frac{\zeta e^{\rho} - \delta(1-\zeta)e^{-\delta\rho}}{\zeta e^{\rho} + (1-\zeta)e^{-\delta\rho}} = \zeta a_+ - \delta(1-\zeta)a_-$$
$$\Delta_I = \frac{\zeta e^{-\delta\rho} - \delta(1-\zeta)e^{\delta^2\rho}}{\zeta e^{-\delta\rho} + (1-\zeta)e^{\delta^2\rho}} = \zeta b_- - \delta(1-\zeta)b_+.$$

Also define $\Delta_i = \Delta_R$ if i is relevant and Δ_I otherwise. With this, we have $\boldsymbol{a}_i = \Delta_i \boldsymbol{x}_R$ based on (77).

The following lemma will be helpful for the downstream analysis. The goal of this lemma is showing that, by choosing $\delta \in \{0, \Delta\}$, we can confuse the model output.

Lemma 16. *Fix a scalar κ. Set $f_\delta = \kappa\Delta_R + (1-\kappa)\Delta_I$. Recalling $\rho = \boldsymbol{x}_R^\top \boldsymbol{k}_R$, the following statements hold:*

- *Set $\delta = 0$. Suppose "$1 \geq \kappa \geq 0$" OR "$\kappa \geq 1, \rho \geq 0$" OR "$\kappa \leq 0, \rho \leq 0$". Then $f_\delta > 0$.*
- *Fix $0 \leq \alpha \leq 1$. Suppose*
$$\delta > \Delta_0 := \frac{1}{1-\zeta}\max\left(\frac{\zeta}{\alpha(1-\zeta)}, \frac{1}{1-\alpha}\right).$$
and "$\kappa \leq \alpha, \rho \geq 0$" OR "$\kappa \geq \alpha, \rho \leq 0$". Then $f_\delta < 0$.

Proof. Plugging in δ, we write

$$f_\delta = \kappa\Delta_R + (1-\kappa)\Delta_I = \kappa\zeta a_+ - \delta\kappa(1-\zeta)a_- + \zeta(1-\kappa)b_- - \delta(1-\zeta)(1-\kappa)b_+ \tag{78}$$
$$= \zeta(\kappa a_+ + (1-\kappa)b_-) - \delta(1-\zeta)(\kappa a_- + (1-\kappa)b_+). \tag{79}$$

• **Suppose $\delta = 0$.** In this case, we obtain the first statement of the lemma as follows

$$f_\delta/\zeta = \frac{\kappa e^{\rho}}{\zeta e^{\rho} + 1 - \zeta} + 1 - \kappa > 0 \quad \text{whenever} \quad \begin{cases} 1 \geq \kappa \geq 0 & \text{OR} \\ \kappa \geq 1, \rho \geq 0 & \text{OR} \\ \kappa \leq 0, \rho \leq 0 \end{cases} \tag{80}$$

• **Now suppose $\delta > \Delta_0$. First, assume $\rho \geq 0$ and $\kappa \leq \alpha$.** We use the facts

$$1/\zeta \geq a_+ \geq 1, \quad 1 \geq a_- \geq 0, \quad b_+ \geq 1, \quad 1 \geq b_- \geq 0.$$

Observe that, since $b_+ \geq a_-$ and $\kappa \leq \alpha$

$$\kappa a_- + (1-\kappa)b_+ \geq \begin{cases} b_+ & \text{if} \quad \kappa \leq 0 \\ (1-\alpha)b_+ & \text{if} \quad \kappa \geq 0 \end{cases} \geq 1 - \alpha.$$

Additionally, if $\kappa \leq 0$, we have that

$$\kappa a_- + (1-\kappa)b_+ \geq b_+ \geq b_- \geq \kappa a_+ + (1-\kappa)b_-.$$

If $\kappa \leq 0$, we obtain $f_\delta \leq \zeta b_- - \delta(1-\zeta)b_+$. Thus, $f_\delta < 0$ whenever $\delta > \Delta_0 \geq \zeta/(1-\zeta)$.

If $\kappa \geq 0$, we use $\kappa a_+ + (1-\kappa)b_- \leq 1/\zeta$ to obtain that whenever $\delta > \Delta_0 \geq \frac{1}{(1-\zeta)(1-\alpha)}$

$$f_\delta \leq 1 - \delta(1-\zeta)(1-\alpha) < 0.$$

Now assume $\rho \leq 0$ and $\kappa \geq \alpha$. We use the facts

$$1 \geq a_+ \geq 0, \quad a_- \geq 1, \quad 1 \geq b_+ \geq 0, \quad \frac{1}{1-\zeta} \geq b_- \geq 1.$$

Observe that, since $b_+ \le a_-$ and $\kappa \ge \alpha$

$$\kappa a_- + (1-\kappa)b_+ \ge \begin{cases} a_- & \text{if} \quad \kappa \ge 1 \\ \alpha a_- & \text{if} \quad \kappa \le 1 \end{cases} \ge \alpha.$$

Additionally, if $\kappa \ge 1$, we have that

$$\kappa a_- + (1-\kappa)b_+ \ge a_- \ge a_+ \ge \kappa a_+ + (1-\kappa)b_-.$$

If $\kappa \ge 1$, we obtain $f_\delta \le \zeta a_+ - \delta(1-\zeta)a_-$. Thus, $f_\delta < 0$ whenever $\delta > \Delta_0 \ge \zeta/(1-\zeta)$.

If $\kappa \le 1$, we use $\kappa a_+ + (1-\kappa)b_- \le \frac{1}{1-\zeta}$ to obtain that whenever $\delta > \Delta_0 \ge \frac{\zeta}{(1-\zeta)^2\alpha}$

$$f_\delta \le \frac{\zeta}{1-\zeta} - \delta(1-\zeta)\alpha < 0.$$

□

To proceed, we will conclude with the proof as follows. Set $\nu_i = y\boldsymbol{u}_i^\top \boldsymbol{x}_R$ for $i \in [T]$ where $\boldsymbol{u}_i$ is the ith row of the output layer weights $\boldsymbol{U}$. Here ν_i is obviously y-dependent. However, we will show that for any choice of y, the model accuracy is at most 50%. Towards this we fix y and (mostly) omit it from the notation during the following discussion. Let $\boldsymbol{a}_i$ be the ith token of the attention output. The linear output layer $\boldsymbol{U}$ aggregates $\boldsymbol{u}_i^\top \boldsymbol{a}_i$ to obtain

$$yf(\boldsymbol{U}, \boldsymbol{W}) = \sum_{i=1}^{T} \boldsymbol{u}_i^\top \boldsymbol{a}_i = \sum_{i=1}^{T} \nu_i \Delta_i.$$

Aggregating $v_+ = \frac{1}{T}\sum_{i\in\mathcal{R}=\text{relevant}} v_i$ and $v_- = \frac{1}{T}\sum_{i\in\mathcal{R}^c=\text{irrelevant}} v_i$ and recalling from (77) that over relevant/irrelevant sets attention tokens are given by $\Delta_R \boldsymbol{x}_R$ and $\Delta_I \boldsymbol{x}_I$, we find

$$\frac{1}{T} yf(\boldsymbol{U}, \boldsymbol{W}) = \nu_R \Delta_R + \nu_I \Delta_I.$$

Scenario 1: Rows of $\boldsymbol{U}$ are identical and we have $\boldsymbol{U} = \mathbb{1}\boldsymbol{u}^\top$. In this scenario, we simply have $\nu_i = \nu$ and $\nu_R = \zeta\nu$ and $\nu_I = (1-\zeta)\nu$. Thus, we find

$$\frac{1}{T} yf(\boldsymbol{U}, \boldsymbol{W}) = T\nu[\zeta\Delta_R + (1-\zeta)\Delta_I].$$

Set $f_\delta = \zeta\Delta_R + (1-\zeta)\Delta_I$. We claim that $\text{sign}(f_\delta)$ is Rademacher (given arbitrary y choice) which will prove that accuracy is at most 50%. Specifically, let us apply Lemma 16 with $\kappa = \zeta$ and $\alpha = \zeta$. When $\delta = 0$, we have $f_\delta > 0$. When $\delta = \Delta$, since the conditions $\kappa \le \alpha$ and $\kappa \ge \alpha$ hold, for any choice of ρ, for $\Delta > \Delta_0 := \frac{1}{(1-\zeta)^2}$ we have that $f_\delta < 0$. Thus, we have that $\mathbb{P}_\delta(f_\delta > 0) = \mathbb{P}_\delta(f_\delta < 0) = 0.5$ as advertised. This follows from the fact that $f_\delta > 0$ for $\delta = 0$ and $f_\delta < 0$ for $\delta = \Delta$ and δ is equally likely over two options.

Scenario 2: Suppose rows of $\boldsymbol{U}$ are not identical. In this case, we will leverage the fact that relevant set $\mathcal{R}$ is allowed to be chosen adversarially with respect to the self-attention weights. We will show that by selecting $\mathcal{R}$ adversarially, on any label y event, accuracy is a coin flip.

First consider the scenario $\nu_{\text{tot}} := \nu_R + \nu_I \le 0$: We will show that model achieves at least 50% error on label y: Let us denote ν_R with $\nu_R^{\mathcal{R}}$ which makes the relevance set dependence explicit. Given $\mathcal{R}$, fixing $\delta = 0$, the model outputs (following (80))

$$\frac{1}{T} yf(\boldsymbol{U}, \boldsymbol{W}) = \frac{\nu_R^{\mathcal{R}} e^\rho}{\zeta e^\rho + 1 - \zeta} + \nu_I^{\mathcal{R}}.$$

Suppose there is a relevance set $\mathcal{R}_0$ (that depends on y) such that the right hand side is non-positive. Let us select this $\mathcal{R}_0$ as our relevance set. Then, the model makes 50% error on label y thanks to the event $\delta = 0$ (which is exactly what we want). If there is no such $\mathcal{R}_0$, then, for all $\mathcal{R}$, we have

$$\frac{\nu_R^{\mathcal{R}} e^\rho}{\zeta e^\rho + 1 - \zeta} + \nu_I^{\mathcal{R}} > 0$$

By taking average of all relevance sets ("T choose ζT" many), all v_i's will be equally-weighted and we obtain $\nu_{\text{tot}} = \nu_R + \nu_I > 0$. This contradicts with our initial $\nu_{\text{tot}} \le 0$ assumption, thus, $\mathcal{R}_0$ has to exist.

Now consider the scenario $\nu_{\text{tot}} = \nu_R + \nu_I > 0$: Let $\mathcal{D}$ be the uniform distribution over "T choose ζT" relevant sets $\mathcal{R}$. Clearly $\mathbb{E}_{\mathcal{D}}[\nu_R^{\mathcal{R}}] = \zeta\nu_{\text{tot}} > 0$. Thus, there is a relevance set $\mathcal{R}_+$ such that $\nu_R^{\mathcal{R}_+} \ge \zeta\nu_{\text{tot}}$ and there is a relevance set $\mathcal{R}_-$ such that $\nu_R^{\mathcal{R}_-} \le \zeta\nu_{\text{tot}}$. We will make use of these two sets to finalize the proof.

To proceed, set $\kappa_\pm = \nu^{\mathcal{R}_\pm}/\nu_{\text{tot}}$ and again set $\alpha = \zeta$ and $\Delta_0 = \frac{1}{(1-\zeta)^2}$ in Lemma 16. Here, we are investigating the sign of the prediction

$$\frac{1}{T\nu_{\text{tot}}} y f(\boldsymbol{U}, \boldsymbol{W}) = \frac{\nu_R\Delta_R + \nu_I\Delta_I}{\nu_{\text{tot}}} = \kappa_\pm\Delta_R + (1-\kappa_\pm)\Delta_I.$$

First, assume that the attention weights are so that $\rho = \rho_y \ge 0$. In this case (and for this particular label y),

- When $\delta = 0$, we choose the relevance set $\mathcal{R}_+$ which ensures $\kappa_+ \ge \zeta \ge 0$ and $f_0 > 0$.
- When $\delta = \Delta > \Delta_0$, we choose the relevance set $\mathcal{R}_-$ which ensures $\kappa_- \le \zeta$ and $f_\Delta < 0$.

Secondly, assume that the attention weights are so that $\rho = \rho_y \le 0$. In this case,

- When $\delta = 0$, we choose the relevance set $\mathcal{R}_-$ which ensures $\kappa_- \le \zeta \le 1$ and $f_0 > 0$.
- When $\delta = \Delta > \Delta_0$, we choose the relevance set $\mathcal{R}_+$ which ensures $\kappa_+ \ge \zeta$ and $f_\Delta < 0$.

In either case, by adaptively choosing $\mathcal{R} \in \{\mathcal{R}_+, \mathcal{R}_-\}$ as a function of (δ, y) pair, we ensure accuracy is at most 50% because f_Δ and f_0 have conflicting signs. □

E.3. Success proof for $\mathcal{R}$-Adaptive Self-Attention

Consider the setting of Theorem 1 and Appendix E.2. We have the following lemma which shows that self-attention can succeed in Theorem 1 if $\boldsymbol{U}$ can adapt to the relevance set (rather than $\mathcal{R}$ being adversarial to $\boldsymbol{U}$).

Lemma 17. *In (DATA), choose (δ^q, δ^w) to be $(0,0)$ or (Δ,Δ) equally-likely. Consider the self-attention model $f^{\text{SATT}}(\boldsymbol{U}, \boldsymbol{W})$ where we set*

$$\boldsymbol{U} = \mathbb{1}_{\mathcal{R}}\boldsymbol{w}_\star'^\top \quad \textit{and} \quad \boldsymbol{W} = \Gamma\boldsymbol{I}.$$

This model achieves perfect accuracy whenever $\boldsymbol{w}_\star' = (\boldsymbol{I} - \bar{\boldsymbol{q}}_\star\bar{\boldsymbol{q}}_\star^\top)\boldsymbol{w}_\star \ne 0$ by choosing

$$\Gamma > \frac{1}{(1+\Delta)(\|\boldsymbol{q}_\star\| - \|\boldsymbol{w}_\star\|)^2 + \|\boldsymbol{w}_\star\|\|\boldsymbol{q}_\star\|(1 - |\rho(\boldsymbol{q}_\star, \boldsymbol{w}_\star)|)} \log(\Delta\frac{1-\zeta}{\zeta}).$$

where $\rho(\cdot)$ is the correlation coefficient.[3]

Proof. Thanks to the masking $\mathbb{1}_{\mathcal{R}}$, we only need to consider the attention scores along relevant tokens. Let $c = \|y\boldsymbol{w}_\star + \boldsymbol{q}_\star\|^2$. For each relevant token, the attention rows are given by

$$\boldsymbol{a}_i = \begin{cases} e^{\Gamma c} & \text{if} \quad i \in \mathcal{R} \\ e^{-\Delta\Gamma c} & \text{if} \quad i \notin \mathcal{R}. \end{cases}$$

To proceed, attention tokens corresponding to relevant tokens are given by

$$\boldsymbol{f} = \sum_{i\in\mathcal{R}} \boldsymbol{a}_i(\boldsymbol{w}_\star + y\boldsymbol{q}_\star) - \sum_{i\notin\mathcal{R}} \Delta\boldsymbol{a}_i(y\boldsymbol{w}_\star + \boldsymbol{q}_\star) \tag{81}$$

$$= (\zeta e^{\Gamma c} - \Delta(1-\zeta)e^{-\Delta\Gamma c})(y\boldsymbol{w}_\star + \boldsymbol{q}_\star). \tag{82}$$

[3]Note that the only instance Γ does not exist is when $\boldsymbol{q}_\star = c\boldsymbol{w}_\star$ for $|c| \ge 1$. In this scenario, classification is impossible using the linear head $\boldsymbol{w}_\star'$ without a bias term because all tokens are in the $\texttt{sign}(c)$ direction regardless of the label y.

Thus, using $\boldsymbol{w}_\star' \boldsymbol{w}_\star > 0$,

$$\texttt{sign}(y f^{\text{SATT}}(\boldsymbol{U}, \boldsymbol{W})) = \texttt{sign}(y \boldsymbol{w}_\star'^\top \boldsymbol{f}) = \texttt{sign}(\zeta e^{\Gamma c} - \Delta(1-\zeta) e^{-\Delta \Gamma c}).$$

Thus, we need $e^{(1+\Delta)\Gamma c} > \Delta \frac{1-\zeta}{\zeta}$ which is implied by $\Gamma > \frac{1}{(1+\Delta)c} \log(\Delta \frac{1-\zeta}{\zeta})$. To conclude, note that for both $y = \pm 1$

$$c \geq \|y \boldsymbol{w}_\star + \boldsymbol{q}_\star\|^2 \geq \|\boldsymbol{q}_\star\|^2 + \|\boldsymbol{w}_\star\|^2 - 2|\boldsymbol{q}_\star^\top \boldsymbol{w}_\star| \geq (\|\boldsymbol{q}_\star\| - \|\boldsymbol{w}_\star\|)^2 + \|\boldsymbol{w}_\star\| \|\boldsymbol{q}_\star\| (1 - |\rho(\boldsymbol{q}_\star, \boldsymbol{w}_\star)|) > 0.$$

where we used $|\boldsymbol{q}_\star^\top \boldsymbol{w}_\star| = \|\boldsymbol{q}_\star\| \|\boldsymbol{w}_\star\| |\rho(\boldsymbol{q}_\star, \boldsymbol{w}_\star)|$. □

F. Proofs of sharp population risk formulas (Theorem 4)

Throughout this section, we use slightly different notation from the one stated in the main body for compactness purposes. Specifically, we set $Q = \|\boldsymbol{q}_\star\|^2, W = T\|\boldsymbol{w}_\star\|^2$ rather than $Q = \|\boldsymbol{q}_\star\|, W = \|\boldsymbol{w}_\star\|$.

Theorem 8. *Consider the prompt-attention model $f_{\boldsymbol{\theta}}^{\text{ATT}}$. Set $Q = \|\boldsymbol{q}_\star\|^2, W = T\|\boldsymbol{w}_\star\|^2$, suppose $\boldsymbol{w}_\star \perp \boldsymbol{q}_\star$, and let $\tau, \bar{\tau} > 0$ be hyperparameters. Consider the following algorithm which uses the hindsight knowledge of $\boldsymbol{q}_\star$ to estimate $\boldsymbol{w}_\star$ and make prediction:*

1. $\hat{\boldsymbol{w}} = (\boldsymbol{I} - \bar{\boldsymbol{q}}_\star \bar{\boldsymbol{q}}_\star) \nabla \mathcal{L}_{\boldsymbol{w}}(0, \tau \bar{\boldsymbol{q}}_\star)$.
2. *Set* $\boldsymbol{\theta} = (\hat{\boldsymbol{w}}, \bar{\tau} \bar{\boldsymbol{q}}_\star)$.

Suppose $\zeta^2 W, 1-\zeta, \alpha := n/d, e^Q, e^\tau$ each lie between two positive absolute constants. Suppose T is polynomially large in n and these constants and $\tilde{\mathcal{O}}(\cdot)$ hides polynomial terms in n. Define inverse-signal-to-noise-ratio*: $\texttt{ISNR}(\alpha, \tau) = \frac{(1-\zeta) e^{2\tau(\tau - \sqrt{Q})}}{\zeta^2 W \alpha}$. With probability $1 - 2e^{-t^2/2} - \tilde{\mathcal{O}}(T^{-1/3})$ over the training data, the test error obeys*

$$\text{ERR}(f_{\boldsymbol{\theta}}^{\text{ATT}}) = \mathcal{Q}\left(\frac{e^{\sqrt{Q}\bar{\tau} - \bar{\tau}^2}}{\sqrt{1 + (1 \mp \frac{1+t}{\sqrt{d}}) \texttt{ISNR}(\alpha, \tau)}} \cdot \sqrt{\frac{\zeta^2 W}{1-\zeta}} \right) \pm \tilde{\mathcal{O}}(T^{-1/3}).$$

Above, $\mp, \pm$ highlights the upper/lower range of the test error (see (86) for exact statement). In the limit $T, d \to \infty$, the test error converges in probability to

$$\text{ERR}(\alpha, \zeta, Q, W, \tau, \bar{\tau}) = \mathcal{Q}\left(\frac{e^{\sqrt{Q}\bar{\tau} - \bar{\tau}^2}}{\sqrt{1 + \texttt{ISNR}(\alpha, \tau)}} \cdot \sqrt{\frac{\zeta^2 W}{1-\zeta}} \right)$$

In this limit, optimal hyperparameters are $\tau = \bar{\tau} = \sqrt{Q}/2$ and leads to optimal $\texttt{ISNR}(\alpha) := \frac{(1-\zeta) e^{-Q/2}}{\zeta^2 W \alpha}$ and the error

$$\text{ERR}(\alpha, \zeta, Q, W) = \mathcal{Q}\left(\frac{e^{Q/4}}{\sqrt{1 + \texttt{ISNR}(\alpha)}} \cdot \sqrt{\frac{\zeta^2 W}{1-\zeta}} \right)$$

Proof. Without losing generality, assume first ζT tokens are relevant and remaining tokens are irrelevant. Consider $\boldsymbol{X}_I$ of size $(1-\zeta)T \times d$ induced by the irrelevant tokens with normal distribution. Observe that $\boldsymbol{g} = \boldsymbol{X}_I \bar{\boldsymbol{w}}_\star$ and $\mathbf{h} = \boldsymbol{X}_I \bar{\boldsymbol{q}}_\star$ are two independent i.i.d. $\mathcal{N}(0, \boldsymbol{I}_{(1-\zeta)T})$ vectors. Also for standard normal $g \sim \mathcal{N}(0,1)$, recall that moment-generating function is given by $\mathbb{E}[e^{\tau g}] = e^{\tau^2/2}$.

Step 1: Characterizing the distribution of $\hat{\boldsymbol{w}}$. Note that, the attention weights have the form $\boldsymbol{a} = \phi(\tau \begin{bmatrix} \sqrt{Q} \mathbb{1}_{\zeta T} \\ \mathbf{h} \end{bmatrix})$. Here, the softmax denominator is $T \cdot D_T$ where $D_T := (\zeta e^{\sqrt{Q}\tau} + \frac{1}{T} \sum_{i=1}^{(1-\zeta)T} e^{\tau h_i})$. Define $e^{\tau \mathbf{h}}$ to be the numerator corresponding to irrelevant tokens i.e.

$$e^{\tau \mathbf{h}} = [e^{\tau h_1} \ \dots \ e^{\tau h_{(1-\zeta)T}}].$$

Define the matrix $\boldsymbol{Q}_\perp = \boldsymbol{I} - \bar{\boldsymbol{q}}_\star \bar{\boldsymbol{q}}_\star^\top$, $\boldsymbol{W}_\perp = \boldsymbol{I} - \bar{\boldsymbol{w}}_\star \bar{\boldsymbol{w}}_\star^\top$. Set the vector $\boldsymbol{v} = \frac{1}{T}\boldsymbol{Q}_\perp \boldsymbol{X}_I^\top e^{\tau \mathbf{h}}$ and $\boldsymbol{v}_\perp = \frac{1}{T}\mathbf{h}^\top e^{\tau \mathbf{h}} \bar{\boldsymbol{q}}_\star$. To proceed, observe that, for a single sample $(y, \boldsymbol{X})$, the gradient has the form

$$\nabla \mathcal{L}_{\boldsymbol{w}}^{y,\boldsymbol{X}}(0, \tau \bar{\boldsymbol{q}}_\star) = y \boldsymbol{X}^\top \boldsymbol{a} = \frac{\zeta(\boldsymbol{w}_\star + y\boldsymbol{q}_\star)e^{\sqrt{Q}\tau} + \boldsymbol{v} + \boldsymbol{v}_\perp}{D_T}. \tag{83}$$

After projection this onto the $\boldsymbol{q}_\star$-complement $\boldsymbol{Q}_\perp$, we get rid of the $\boldsymbol{q}_\star$ direction to obtain

$$\hat{\boldsymbol{w}}_{y,\boldsymbol{X}} = \boldsymbol{Q}_\perp \nabla \mathcal{L}_{\boldsymbol{w}}^{y,\boldsymbol{X}}(0, \tau \bar{\boldsymbol{q}}_\star) = \frac{1}{D_T}[\zeta \boldsymbol{w}_\star e^{\sqrt{Q}\tau} + \boldsymbol{Q}_\perp \boldsymbol{X}_I^\top e^{\tau \mathbf{h}}/T].$$

The projected gradient over the full dataset is given by the empirical average

$$\hat{\boldsymbol{w}} = \boldsymbol{Q}_\perp \nabla \mathcal{L}_{\boldsymbol{w}}(0, \tau \bar{\boldsymbol{q}}_\star) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{D_{i,T}}[\zeta \boldsymbol{w}_\star e^{\sqrt{Q}\tau} + \boldsymbol{Q}_\perp \boldsymbol{X}_{i,I}^\top e^{\tau \mathbf{h}_i}/T].$$

Here $\mathbf{h}_i, \boldsymbol{X}_{i,I}, D_{i,T}$ denote the random variables induced by the ith sample. Here, a critical observation is the fact that $\boldsymbol{Q}_\perp \boldsymbol{X}_{i,I}$ is independent of $\mathbf{h}_i$ (thanks to Gaussian orthogonality), thus, $\boldsymbol{Q}_\perp \boldsymbol{X}_{i,I} e^{\tau \mathbf{h}_i}$ is normal conditioned on $\mathbf{h}_i$. To proceed, we apply Chebyshev's inequality over number of tokens T. Recall that we assumed $e^\tau \le C$ for an absolute constant $C \ge 1$. This means that $e^{c\tau^2} \le C^{c\tau} \le C^{c \log C}$ is polynomial in C and is also upper bounded by a constant. In what follows $\tilde{\mathcal{O}}(\cdot)$ only reflects the T dependence and subsumes polynomial dependence on the terms n, C. For all $1 \le i \le n$, applying Chebyshev's inequality, for $T \gtrsim \mathrm{poly}(n, e^{\tau^2})$, with probability $1 - T^{-1/3}$, we have that

- Since $\|e^{\tau \mathbf{h}_i}\|^2/T = \frac{1}{T}\sum_{j=1}^{T} e^{2\tau \mathbf{h}_{ij}}$ thus $\|e^{\tau \mathbf{h}_i}\|^2/T - (1-\zeta)e^{2\tau^2} \le \tilde{\mathcal{O}}(T^{-1/3})$,
- Set $\mathbb{E}[D_T] = D_\infty := \zeta e^{\sqrt{Q}\tau} + (1-\zeta)e^{\tau^2/2}$. $|D_{i,T} - D_\infty| \le \tilde{\mathcal{O}}(T^{-1/3})$.

With these, set $\boldsymbol{b}_i = \frac{\sqrt{1-\zeta}e^{\tau^2}}{\|e^{\tau \mathbf{h}_i}\|} e^{\tau \mathbf{h}_i}$ which is a vector with fixed ℓ_2 norm that is perfectly parallel to $e^{\tau \mathbf{h}_i}$. Since $\|\boldsymbol{b}_i\|^2 = \mathbb{E}[\|e^{\tau \mathbf{h}_i}\|^2/T] = (1-\zeta)e^{2\tau^2}$, from above, observe that,

$$\|\boldsymbol{b}_i - \frac{1}{\sqrt{T}} e^{\tau \mathbf{h}_i}\| \le \tilde{\mathcal{O}}(T^{-1/3}).$$

Now, let

$$\bar{\boldsymbol{v}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{Q}_\perp \boldsymbol{X}_{i,I}^\top \boldsymbol{b}_i.$$

Since $\boldsymbol{Q}_\perp \boldsymbol{X}_{i,I}^\top, \boldsymbol{b}_i$ are independent and $\boldsymbol{b}_i$ has fixed ℓ_2 norm, we have that

$$\bar{\boldsymbol{v}} \sim \mathcal{N}(0, (1-\zeta)e^{2\tau^2}\boldsymbol{Q}_\perp).$$

Finally, let $\boldsymbol{c} = \zeta e^{\sqrt{Q}\tau}\boldsymbol{w}_\star$. Recalling $\sqrt{T}\|\boldsymbol{w}_\star\| = W$, combining the perturbations bounds above, we have that

$$\sqrt{T}\|\boldsymbol{c}/D_\infty - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{D_{i,T}}(\zeta \boldsymbol{w}_\star e^{\sqrt{Q}\tau})\| \le \tilde{\mathcal{O}}(T^{-1/3}).$$

Combining these observe that

$$\|\sqrt{T}D_\infty \hat{\boldsymbol{w}} - \sqrt{T}\zeta e^{\sqrt{Q}\tau}\boldsymbol{w}_\star - \bar{\boldsymbol{v}}/\sqrt{n}\| \le \tilde{\mathcal{O}}(T^{-1/3}). \tag{84}$$

Since $\bar{\boldsymbol{v}}$ is normally distributed, above also implies that $\sqrt{T}D_\infty \hat{\boldsymbol{w}}$ converges to the normal distribution $\mathcal{N}(\sqrt{T}\zeta e^{\sqrt{Q}\tau}\boldsymbol{w}_\star, \frac{(1-\zeta)e^{2\tau^2}}{n}\boldsymbol{Q}_\perp)$ in the limit $T \to \infty$.

Lemma 18 (Inverse Signal-to-Noise Ratio (ISNR)). *Set $\boldsymbol{W}_\perp = \boldsymbol{I} - \bar{\boldsymbol{w}}_\star \bar{\boldsymbol{w}}_\star^\top$. Define SNR of $\hat{\boldsymbol{w}}$ to be*

$$\mathtt{ISNR}(\hat{\boldsymbol{w}}) = \frac{\|\boldsymbol{W}_\perp \hat{\boldsymbol{w}}\|^2}{\|\bar{\boldsymbol{w}}_\star^\top \hat{\boldsymbol{w}}\|^2}.$$

Recall $\mathtt{ISNR}(\alpha,\tau) = \frac{(1-\zeta)e^{2\tau(\tau-\sqrt{Q})}}{\zeta^2 W\alpha}$. *With probability* $1-2e^{-t^2/2}-T^{-1/3}$ *over the dataset, we have that*

$$\left(1-\frac{t+1}{\sqrt{d}}-\tilde{\mathcal{O}}(T^{-\frac{1}{3}})\right)_+^2 \le \frac{\mathtt{ISNR}(\hat{\boldsymbol{w}})}{\mathtt{ISNR}(\alpha,\tau)} \le \left(1+\frac{\tau}{\sqrt{d}}+\tilde{\mathcal{O}}(T^{-\frac{1}{3}})\right)^2.$$

Proof. Let us recall the standard normal concentration: For $\boldsymbol{g}\sim\mathcal{N}(0,\boldsymbol{I}_{d-1})$, $\sqrt{d-1}\ge\mathbb{E}[\|\boldsymbol{g}\|]\ge\frac{d-1}{\sqrt{d}}$. Thus, with probability $1-2e^{-t^2/2}$, through Lipschitz concentration,

$$\sqrt{d}+t\ge\|\boldsymbol{g}\|\ge\sqrt{d}-1-t.$$

This means that, with the same probability

$$\sqrt{d}+t\ge\frac{\|\bar{\boldsymbol{v}}\|}{\sqrt{1-\zeta}e^{\tau^2}}\ge(\sqrt{d}-1-t)_+.$$

We first upper bound $\|\boldsymbol{W}_\perp\hat{\boldsymbol{w}}\|^2$. Recalling (84),

$$\|\boldsymbol{W}_\perp\hat{\boldsymbol{w}}-\bar{\boldsymbol{v}}/\sqrt{n}\|\le\tilde{\mathcal{O}}(T^{-1/3}).$$

Thus,

$$(\sqrt{d}+t)^2+\tilde{\mathcal{O}}(T^{-1/3})\ge\frac{n\|\boldsymbol{W}_\perp\hat{\boldsymbol{w}}\|^2}{(1-\zeta)e^{2\tau^2}}\ge(\sqrt{d}-1-t)_+^2-\tilde{\mathcal{O}}(T^{-1/3}).$$

Using $\|\boldsymbol{w}_\star\|^2T=W$, We similarly have that

$$|\|\bar{\boldsymbol{w}}_\star^\top\hat{\boldsymbol{w}}\|^2-W\zeta^2e^{2\sqrt{Q}\tau}|\le\tilde{\mathcal{O}}(T^{-1/3}).$$

To conclude, with probability $1-2e^{-t}-T^{-1/3}$, $\mathtt{ISNR}(\hat{\boldsymbol{w}})$ obeys

$$\frac{(\sqrt{d}+t)^2+\tilde{\mathcal{O}}(T^{-1/3})}{W\zeta^2e^{2\sqrt{Q}\tau}-\tilde{\mathcal{O}}(T^{-1/3})}\ge\frac{n}{(1-\zeta)e^{2\tau^2}}\mathtt{ISNR}(\hat{\boldsymbol{w}})\ge\frac{(\sqrt{d}-1-t)_+^2-\tilde{\mathcal{O}}(T^{-1/3})}{W\zeta^2e^{2\sqrt{Q}\tau}+\tilde{\mathcal{O}}(T^{-1/3})}$$

Rewriting this bound, we find

$$\left(1+\frac{t}{\sqrt{d}}+\tilde{\mathcal{O}}(T^{-\frac{1}{3}})\right)^2\frac{(1-\zeta)e^{2\tau^2}}{\zeta^2W\alpha e^{2\sqrt{Q}\tau}}\ge\mathtt{ISNR}(\hat{\boldsymbol{w}})\ge\left(1-\frac{1+t}{\sqrt{d}}-\tilde{\mathcal{O}}(T^{-\frac{1}{3}})\right)_+^2\frac{(1-\zeta)e^{2\tau^2}}{\zeta^2W\alpha e^{2\sqrt{Q}\tau}}.$$

Recalling the definition of $\mathtt{ISNR}(\alpha,\tau)=\frac{(1-\zeta)e^{2\tau(\tau-\sqrt{Q})}}{\zeta^2W\alpha}$, we conclude with the bound. □

Step 2: Characterizing the error rate of $\boldsymbol{\theta}=(\hat{\boldsymbol{w}},\bar{\tau}\boldsymbol{q}_\star)$. To achieve this goal, we will leverage Theorem 9. Since conditions of this theorem is satisfied (noticing that their γ is our $\mathtt{ISNR}(\hat{\boldsymbol{w}})$ which is upper bounded by a positive constant), for a new test point $(y,\boldsymbol{X})$, we have that

$$\left|\mathrm{ERR}(f_{\boldsymbol{\theta}}^{\mathrm{ATT}})-\mathcal{Q}\left(\frac{e^{\sqrt{Q}\bar{\tau}-\bar{\tau}^2}}{\sqrt{1+\mathtt{ISNR}(\hat{\boldsymbol{w}})}}\cdot\sqrt{\frac{\zeta^2W}{1-\zeta}}\right)\right|\le\tilde{\mathcal{O}}(T^{-1/3}).$$

Using the Lipschitzness of the Q-function (i.e. $\mathcal{Q}(x+\epsilon)-\mathcal{Q}(x)=\int_x^{x+\epsilon}e^{-t^2/2}dt\le\epsilon$), as we have done in Theorem 9, we pull out the perturbation term $\tilde{\mathcal{O}}(T^{-1/3})$ within $\mathtt{ISNR}(\hat{\boldsymbol{w}})$ to obtain the advertised bound

$$\mathcal{Q}\left(\frac{e^{\sqrt{Q}\bar{\tau}-\bar{\tau}^2}}{\sqrt{1+(1+\frac{t}{\sqrt{d}})\mathtt{ISNR}(\alpha,\tau)}}\cdot\sqrt{\frac{\zeta^2W}{1-\zeta}}\right)-\tilde{\mathcal{O}}(T^{-1/3})\le\mathrm{ERR}(f_{\boldsymbol{\theta}}^{\mathrm{ATT}})\le \tag{85}$$

$$\mathcal{Q}\left(\frac{e^{\sqrt{Q}\bar{\tau}-\bar{\tau}^2}}{\sqrt{1+(1-\frac{1+t}{\sqrt{d}})_+\mathtt{ISNR}(\alpha,\tau)}}\cdot\sqrt{\frac{\zeta^2W}{1-\zeta}}\right)+\tilde{\mathcal{O}}(T^{-1/3}). \tag{86}$$

To emphasize, this bound holds with probability $1-2e^{-t^2/2}-\tilde{\mathcal{O}}(T^{-1/3})$ over a new test datapoint $(y, \boldsymbol{X})$. To see the optimal choices for $\bar{\tau}, \tau$, we need to optimize the error bound. This results in

$$\bar{\tau}_\star = \arg\min_{\bar{\tau}} \sqrt{Q}\bar{\tau} - \bar{\tau}^2 = \sqrt{Q}/2 \tag{87}$$

$$\tau_\star = \arg\min_{\tau} \mathtt{ISNR}(\alpha, \tau) = 2\tau(\tau - \sqrt{Q}) = \sqrt{Q}/2. \tag{88}$$

□

Theorem 9. *Consider the prompt-attention model $f_{\boldsymbol{\theta}}^{\mathrm{ATT}}$ where we set $\boldsymbol{\theta} = (\boldsymbol{w}_\star + \boldsymbol{p}, \tau\bar{\boldsymbol{q}}_\star)$. Set $Q = \|\boldsymbol{q}_\star\|^2, W = T\|\boldsymbol{w}_\star\|^2$. Here τ is a tuning parameter and $\boldsymbol{p}$ is a perturbation vector and assume all vectors are perpendicular i.e. $\boldsymbol{p} \perp \boldsymbol{w}_\star \perp \boldsymbol{q}_\star$. Set $\gamma := \|\boldsymbol{p}\|^2/\|\boldsymbol{w}_\star\|^2$ and suppose $1+\gamma, \zeta^2 W, 1-\zeta, e^Q, e^\tau$ each lie between two positive absolute constants. $\tilde{\mathcal{O}}(\cdot)$ subsumes polynomial dependencies in these constants. We have that*

$$\left|\mathrm{ERR}(f_{\boldsymbol{\theta}}^{\mathrm{ATT}}) - \mathcal{Q}\left(\frac{e^{\sqrt{Q}\tau-\tau^2}}{\sqrt{1+\gamma}} \cdot \sqrt{\frac{\zeta^2 W}{1-\zeta}}\right)\right| \le \mathcal{O}(T^{-1/3}).$$

Thus, as $T \to \infty$, the optimal tuning obeys $\tau_\star = \sqrt{Q}/2$ and yields an error of $\mathcal{Q}\left(e^{Q/4} \cdot \sqrt{\frac{\zeta^2 W}{1-\zeta}}\right)$.

Proof. Let us recap the notation of Theorem 4. Without losing generality, assume first ζT tokens are relevant and remaining tokens are irrelevant. Consider $\boldsymbol{X}_I$ of size $(1-\zeta)T \times d$ induced by the irrelevant tokens with normal distribution. Using orthogonality of $\boldsymbol{q}_\star, \boldsymbol{w}_\star, \boldsymbol{p}$, observe that $\boldsymbol{g} = \boldsymbol{X}_I(\bar{\boldsymbol{w}}_\star + \frac{\boldsymbol{p}}{\|\boldsymbol{w}_\star\|}) \sim \mathcal{N}(0, (1+\gamma)\boldsymbol{I}_{(1-\zeta)T})$ and $\mathbf{h} = \boldsymbol{X}_I\bar{\boldsymbol{q}}_\star \sim \mathcal{N}(0, \boldsymbol{I}_{(1-\zeta)T})$ are independent vectors. Also for standard normal $g \sim \mathcal{N}(0,1)$, recall that moment-generating function is given by $\mathbb{E}[e^{\tau g}] = e^{\tau^2/2}$.

Note that, the attention weights have the form $\boldsymbol{a} = \phi(\tau\begin{bmatrix}\sqrt{Q}\mathbb{1}_{\zeta T} \\ \mathbf{h}\end{bmatrix})$. Here, the softmax denominator is $T \cdot D_T$ where $D_T := (\zeta e^{\sqrt{Q}\tau} + \frac{1}{T}\sum_{i=1}^{(1-\zeta)T} e^{\tau h_i})$. Define $e^{\tau\mathbf{h}}$ to be the numerator corresponding to irrelevant tokens i.e.

$$e^{\tau\mathbf{h}} = [e^{\tau h_1} \ \dots \ e^{\tau h_{(1-\zeta)T}}].$$

Define the matrix $\boldsymbol{Q}_\perp = \boldsymbol{I} - \bar{\boldsymbol{q}}_\star\bar{\boldsymbol{q}}_\star^\top$, $\boldsymbol{W}_\perp = \boldsymbol{I} - \bar{\boldsymbol{w}}_\star\bar{\boldsymbol{w}}_\star^\top$. To proceed, observe that, the prediction with $\boldsymbol{\theta} = (\boldsymbol{w}_\star + \boldsymbol{p}, \tau\bar{\boldsymbol{q}}_\star)$ is given by

$$\frac{D_T}{\sqrt{T}\|\boldsymbol{w}_\star\|} y f_{\boldsymbol{\theta}}^{\mathrm{ATT}}(\boldsymbol{X}) = \sqrt{T}(\bar{\boldsymbol{w}}_\star + \frac{\boldsymbol{p}}{\|\boldsymbol{w}_\star\|})^\top[\zeta e^{\|\boldsymbol{q}_\star\|\tau}(\boldsymbol{w}_\star + y\boldsymbol{q}_\star) + \frac{\boldsymbol{X}_I e^{\tau\mathbf{h}}}{T}] \tag{89}$$

$$= \zeta e^{\|\boldsymbol{q}_\star\|\tau}\sqrt{W} + \frac{1}{\sqrt{T}}\boldsymbol{g}^\top e^{\tau\mathbf{h}}. \tag{90}$$

With this, conditioned on $e^{\tau\mathbf{h}}$ observe that $\boldsymbol{g}^\top e^{\tau\mathbf{h}} \sim \mathcal{N}(0, \frac{1}{T}\|e^{\tau\mathbf{h}}\|^2)$, thus,

$$\mathbb{P}_{\boldsymbol{g}}(y f_{\boldsymbol{\theta}}^{\mathrm{ATT}}(\boldsymbol{X}) > 0) = 1 - \mathcal{Q}\left(\frac{\zeta e^{\sqrt{Q}\tau}\sqrt{W}}{\sqrt{1+\gamma}\|e^{\tau\mathbf{h}}\|/\sqrt{T}}\right).$$

To proceed, similar to Theorem 4, we apply Chebyshev's inequality over number of tokens T to find that with probability $1-\tilde{\mathcal{O}}(T^{-1/3})$ over $\mathbf{h}$,

$$|\|e^{\tau\mathbf{h}}\|^2/T - e^{2\tau^2}| \le \tilde{\mathcal{O}}(T^{-1/3}).$$

In aggregate, this implies that, with probability $1-\tilde{\mathcal{O}}(T^{-1/3})$ over $\mathbf{h}$, we have that

$$1 - \mathcal{Q}\left((1+\tilde{\mathcal{O}}(T^{-1/3}))\frac{\zeta e^{\sqrt{Q}\tau}\sqrt{W}}{\sqrt{1+\gamma}e^{\tau^2}}\right) \ge \mathbb{P}_{\boldsymbol{g}}(y f_{\boldsymbol{\theta}}^{\mathrm{ATT}}(\boldsymbol{X}) > 0) \ge 1 - \mathcal{Q}\left((1-\tilde{\mathcal{O}}(T^{-1/3}))\frac{\zeta e^{\sqrt{Q}\tau}\sqrt{W}}{\sqrt{1+\gamma}e^{\tau^2}}\right),$$

Finally, note that since $\frac{\zeta e^{\sqrt{Q}\tau}\sqrt{W}}{\sqrt{1+\gamma}e^{\tau^2}}$ is upper/lower bounded by a positive constant, and since $\mathcal{Q}(x+\epsilon) - \mathcal{Q}(x) = \int_x^{x+\epsilon} e^{-t^2/2}dt \le \epsilon$, we can rewrite

$$\left|\mathbb{P}_{\boldsymbol{g}}(y f_{\boldsymbol{\theta}}^{\mathrm{ATT}}(\boldsymbol{X}) > 0) - \mathcal{Q}\left(\frac{\zeta e^{\sqrt{Q}\tau}\sqrt{W}}{\sqrt{1+\gamma}e^{\tau^2}}\right)\right| \le \tilde{\mathcal{O}}(T^{-1/3}).$$

Union bounding with failure probability over $\mathbf{h}$, we conclude with the result. □

G. Useful facts

For a random variable Z and $\alpha > 0$, $\|Z\|_{\psi_\alpha}$ denotes its ψ_α-norm for Orlicz function $\psi_\alpha(z) = e^{z^\alpha} - 1$ (Ledoux & Talagrand, 1991).

Fact G.1. *Let $X_1, \ldots, X_n$ be independent zero-mean sub-gaussian or sub-exponential random variables with $\|X_i\|_{\psi_m} \le K$ for all $i \in [n]$ for either $m = 2$ or $m = 1$. Then,*

$$\left\| \frac{1}{n} \sum_{i \in [n]} X_i \right\|_{\psi_m} \le \frac{CK}{\sqrt{n}}.$$

Fact G.2. *(Pollard, 1990) The following identity holds for Orlitz norms*

$$\|XY\|_{\psi_{\frac{\alpha\beta}{\alpha+\beta}}} \le c\,\|X\|_{\psi_\alpha} \cdot \|Y\|_{\psi_\beta} \tag{91}$$

for a fixed numerical constant c.

Next we state a Lemma from Talagrand quoted directly from Lemma 22 of (Mohammadi et al., 2019).

Fact G.3. *(Ledoux & Talagrand, 1991)For any scalar $\alpha \in (0, 1]$, there exists a constant C_α such that for any sequence of independent random variables $\xi_1, \xi_2, \ldots, \xi_N$ we have*

$$\Big\| \sum_i \xi_i - \mathbb{E}\big[\sum_i \xi_i\big] \Big\|_{\psi_\alpha} \le C_\alpha \left(\max_i \|\xi_i\|_{\psi_\alpha} \right) \sqrt{N} \log N.$$

Fact G.4. *Let $\boldsymbol{z} \in \mathbb{R}^d$ be a K-subgausssian vector i.e. $\boldsymbol{z}^\top \boldsymbol{v}$ is K-subgaussian for fixed $\|\boldsymbol{v}\| = 1$. Then, the following are true for a constant $c > 0$*

$$\mathbb{P}\left(\|\boldsymbol{z}\| \ge cK(\sqrt{d} + t) \right) \le e^{-t^2}.$$

Proof. For completeness, we provide a proof. Repeating Lemma 31 of (Oymak, 2019), we can pick a $1/2$ cover $\mathcal{C}$ of the unit Euclidean ball in $\mathbb{R}^d$ with size $\log|\mathcal{C}| \le 2d$. For any $\boldsymbol{v} \in \mathcal{C}$ subgaussianity implies $\mathbb{P}(\boldsymbol{v}^\top \boldsymbol{z} \ge t) \le \exp(-ct^2/K^2)$. Setting $K' = K/\sqrt{c}$, $t = K'(\sqrt{2d} + \tau)$ and union bounding over all $\boldsymbol{v} \in \mathcal{C}$, we find

$$\mathbb{P}(\sup_{\boldsymbol{v} \in \mathcal{C}} \boldsymbol{v}^\top \boldsymbol{z} \ge K'(\sqrt{d} + \tau)) \le \exp(-\tau^2).$$

To proceed, set $\boldsymbol{v}(\boldsymbol{z}) \in \mathcal{C}$ to be the nearest point to $\bar{\boldsymbol{z}} = \boldsymbol{z}/\|\boldsymbol{z}\|$ in $\mathcal{C}$. Since $\|\boldsymbol{v}(\boldsymbol{z}) - \boldsymbol{z}\| \le 0.5$, note that

$$\|\boldsymbol{z}\| = \bar{\boldsymbol{z}}^\top \boldsymbol{z} = \boldsymbol{v}(\boldsymbol{z})^\top \boldsymbol{z} + (\boldsymbol{z} - \boldsymbol{v}(\boldsymbol{z}))^\top \boldsymbol{z} \le \boldsymbol{v}(\boldsymbol{z})^\top \boldsymbol{z} + 0.5\|\boldsymbol{z}\|.$$

Thus, $\|\boldsymbol{z}\| \le 2K'(\sqrt{d} + \tau)$ with probability at least $1 - \exp(-\tau^2)$.

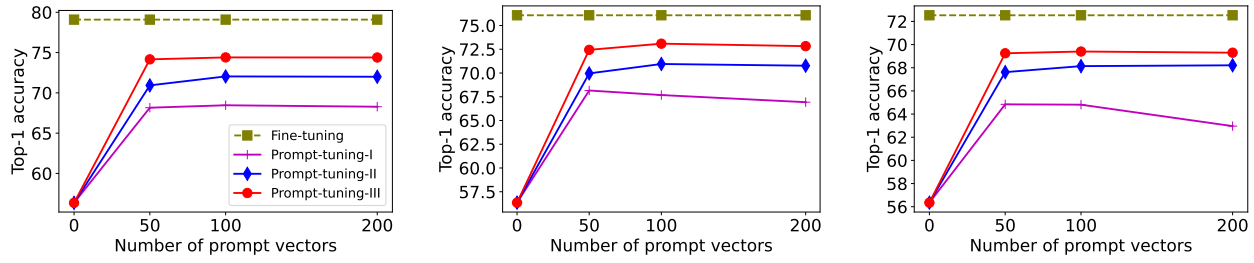
□

H. Additional experimental results and details

H.1. Additional details for image classification experiments

Dataset. As mentioned in Section 5.2, we construct three datasets by modifying the original images in CIFAR-10:

- FULL-TILED. Each examples consists of a 64x64 images obtained by arranging a 32x32 image from CIFAR-10 in a tiling pattern with four tiles (cf. Fig. 3a).
- PARTIAL-TILED. This is dataset is similar to FULL-TILED with the exception that each image has at-least T out of 4 tiles replaced by patches of i.i.d. random Gaussian noise with mean zero and variance 0.2. Note that, for each example in the dataset, $T \in \{1, 2, 3\}$ is a random number as well as the location of the noisy tiles (cf. Fig. 3b).
- EMBED-IN-IMAGENET (Karp et al., 2021). We construct an example by simply embedding a 32x32 image from CIFAR-10 at a random location in a 64x64 background corresponding to a randomly selected image from ImageNet (Russakovsky et al., 2015). We also add i.i.d. random Gaussian noise with mean zero and variance 0.2 to the background (cf. Fig. 3c).



(a) Full dataset **(b)** Capped 10% **(c)** Capped 2%

Figure 5. Performance of fine-tuning vs. prompt-tuning on 10-way classification tasks defined by PARTIAL-TILED dataset. Full dataset has 50K training examples. Capped 10% and 2% correspond to sub-sampled *train* sets where we select exactly 500 and 100 examples per class from the full dataset. Note that number of prompt vectors equal to 0 corresponds to *zero-shot* performance.

Table 1. Comparison between prompt-tuning and fine-tuning only first layer self-attention weights (Full dataset).

Method (trainable parameters)	EMBED-IN-IMAGENET	PARTIAL-TILED
PROMPT-TUNING-I w/ 100 prompt vectors (19.2K)	31.89	68.46
PROMPT-TUNING-II w/ 100 prompt vectors (19.2K)	38.48	72.04
PROMPT-TUNING-III w/ 100 prompt vectors (230.8K)	47.81	74.40
Fine-tuning only first layer attention weights (148.2K)	30.35	73.95

By construction, each dataset has 50,000 train and 10,000 test examples corresponding to train and test set of CIFAR-10. We also consider data-limited settings where we keep the test set intact but subsample the train set by selecting a fixed number of images for each class. Note that all three datasets define 10-way multiclass classification tasks with CIFAR-10 classes as potential labels.

Model architecture. We utilize a tiny variant of the Vision transformer model (Dosovitskiy et al., 2021) for our experiments. This variant has 12 transformer layers with its hidden dimension, MLP intermediate dimension, and number of heads per attention layer being equal to 192, 768, and 3, respectively. The patch size in our study is set to be 4x4. The model itself (without counting the trainable parameters/weights during prompt-tuning) has approximately 5.44M parameters. We rely on the CLS token to obtain the classification logits.

Training. We rely on Scenic library (Dehghani et al., 2022)[4] to conduct our experiments on image classification. Following the default settings in the library along with a coarse grid search, we employ Adam optimizer (Kingma & Ba, 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay = 0.1, and batch size = 128 while training a *randomly initialized* model. Furthermore, we employ a linear warm-up of learning rate followed cosine learning rate schedule with base learning rate 3e-3. As for the fine-tuning and prompt-tuning experiments that (partially) initialize from a *pre-trained* model, we rely on SGD with momentum parameter 0.9 and batch size = 128 to update trainable parameters. Again, we utilize a linear warm-up of learning rate followed by cosine learning rate schedule. Throughout our experiments, the base learning rates for fine-tuning and prompt-tuning are 1e-3 and 0.1, respectively.

H.2. Additional results on image classification

Figure 5 showcases the performance of fine-tuning and various prompt-tuning strategies on PARTIAL-TILED.

Comparison with fine-tuning the first self-attention layer. In Tables 1, 2, and 3, we explored fine-tuning only first layer self-attention parameters for the underlying ViT model. This setting aligns well with the single-layer nature of our theoretical results. Similar to Fig. 4 (corresponding to EMBED-IN-IMAGENET dataset) and Fig. 5 (corresponding to PARTIAL-TILED dataset), we considered three settings: 1) Full dataset; 2) Capped 10%; and 3) Capped 2%, which progressively corresponds to smaller amount of (training) data during fine-tuning and prompt-tuning.

The key takeaways are:

[4] https://github.com/google-research/scenic

Table 2. Comparison between prompt-tuning and fine-tuning only first layer self-attention weights (Capped 10%).

Method (trainable parameters)	EMBED-IN-IMAGENET	PARTIAL-TILED
PROMPT-TUNING-I w/ 100 prompt vectors (19.2K)	28.06	67.68
PROMPT-TUNING-II w/ 100 prompt vectors (19.2K)	36.76	70.95
PROMPT-TUNING-III w/ 100 prompt vectors (230.8K)	42.96	73.09
Fine-tuning only first layer attention weights (148.2K)	18.53	70.44

Table 3. Comparison between prompt-tuning and fine-tuning only first layer self-attention weights (Capped 2%).

Method (trainable parameters)	EMBED-IN-IMAGENET	PARTIAL-TILED
PROMPT-TUNING-I w/ 100 prompt vectors (19.2K)	20.52	64.81
PROMPT-TUNING-II w/ 100 prompt vectors (19.2K)	33.62	68.14
PROMPT-TUNING-III w/ 100 prompt vectors (230.8K)	36.13	69.40
Fine-tuning only first layer attention weights (148.2K)	15.46	65.04

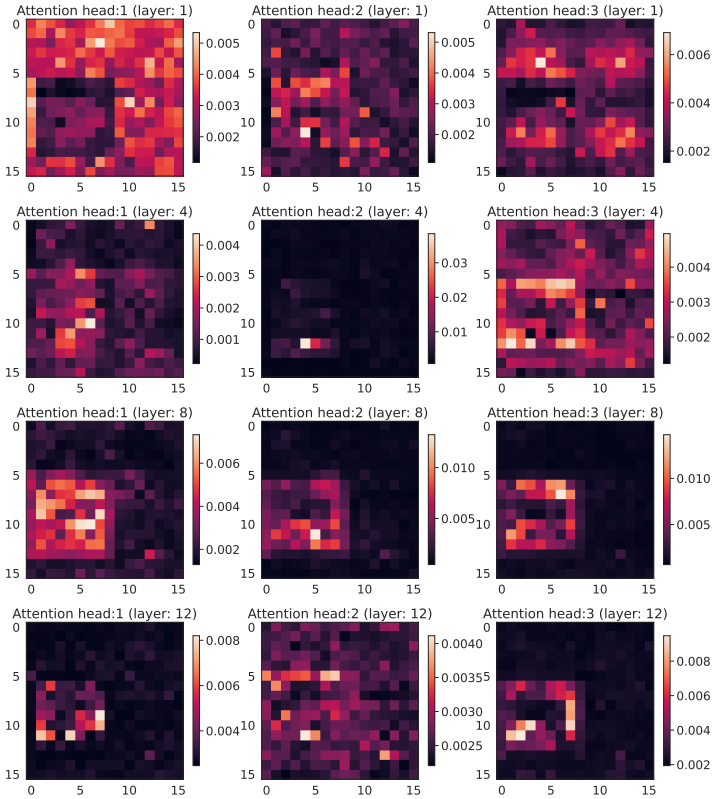
1. When there is a significant distribution-shift between from pre-training data (in case of EMBED-IN-IMAGENET), even the simplest prompt-tuning, namely PROMPT-TUNING-I, significantly outperforms the fine-tuning first layer self-attention parameters.

2. When the distribution-shift is small, prompt-tuning variants realize a better *accuracy vs. training cost* trade-off, e.g. PROMPT-TUNING-II outperforms fine-tuning first layer self-attention parameters in the Capped 10% and Capped 2% setting (while training only 19.2K rather than 148.2K parameters).

H.3. Illustration of attention weights for prompt vectors

Fig. 6 presents a representative example where we show evolution of average attention weights from prompt vectors to image tokens/patches across transformer layers, when we employ PROMPT-TUNING-III. It is evident from the figure that prompt-attention helps distinguish the relevant tokens/patches from the irrelevant patches.

(a) Input image



(b) Average attention weights from prompts (keys) to image patches (values).

Figure 6. Illustration of how attention weights progressive change from the first layer (Figure 6b-top) to the last layer (Figure 6b-bottom) in the transformer model for a given input image (Figure 6a) when we employ PROMPT-TUNING-III. We plot average attention weights from 50 prompt vectors (keys) to 256 image patches (values). The attention weights for each attention head are naturally arranged in a 16 x 16 grid corresponding to the original locations of the patches in the image. Note that the attention weights in the early layer have a tiling pattern similar to that in FULL-TILED– the dataset utilized by the pre-trained model. However, as we progress deeper into the transformer, attention weights begin to capture the relevant patch locations in the dataset of interest, i.e., EMBED-IN-IMAGENET.