Not too little, not too much: a theoretical analysis of graph (over)smoothing

Nicolas Keriven
CNRS, GIPSA-lab, Grenoble, France
nicolas.keriven@cnrs.fr

Abstract

We analyze graph smoothing with *mean aggregation*, where each node successively receives the average of the features of its neighbors. Indeed, it has quickly been observed that Graph Neural Networks (GNNs), which generally follow some variant of Message-Passing (MP) with repeated aggregation, may be subject to the *oversmoothing* phenomenon: by performing too many rounds of MP, the node features tend to converge to a non-informative limit. In the case of mean aggregation, for connected graphs, the node features become constant across the whole graph. At the other end of the spectrum, it is intuitively obvious that *some* MP rounds are necessary, but existing analyses do not exhibit both phenomena at once: beneficial "finite" smoothing and oversmoothing in the limit. In this paper, we consider simplified linear GNNs, and rigorously analyze two examples for which a finite number of mean aggregation steps provably improves the learning performance, before oversmoothing kicks in. We consider a latent space random graph model, where node features are partial observations of the latent variables and the graph contains pairwise relationships between them. We show that graph smoothing restores some of the lost information, up to a certain point, by two phenomena: graph smoothing shrinks non-principal directions in the data faster than principal ones, which is useful for regression, and shrinks nodes within communities faster than they collapse together, which improves classification.

1 Introduction

In recent years, deep architectures such as Graph Neural Networks (GNNs), along with the availability of large sets of graph data, have significantly broadened the field of machine learning on graphs and structured data, with a myriad of applications ranging from community detection [11] to molecule classification [20], drug discovery [19], quantum chemistry [15], recommender systems [44], semi-supervised learning, and so on. See [7, 16, 6, 46] for reviews. Most GNNs rely on the **Message-Passing** (MP) framework [15, 23], with a plethora of variants. At each layer k, for each node i, a representation $z_i^{(k)}$ is computed using the representations of the *neighbors* $\mathcal{N}_i$ of i in the graph at the previous layer:

$$z_i^{(k)} = \mathrm{AGG}\left(\{z_j^{(k-1)}\}_{j\in\mathcal{N}_i}\right) \tag{1}$$

where AGG is an **aggregation function** that, crucially, treats $\{z_j^{(k-1)}\}_{j\in\mathcal{N}_i}$ as an *unordered set*, to respect the absence of node ordering in the graph. There are many variants of aggregation functions, based on sum, mean, max, min, degree-normalized [23], attention-based [39], and so on. In this work, we consider one of the most classical, *mean aggregation*:

$$z_i^{(k)} = \frac{1}{\sum_j a_{ij}} \sum_j a_{ij} \Psi\left(z_j^{(k-1)}\right) \tag{2}$$

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

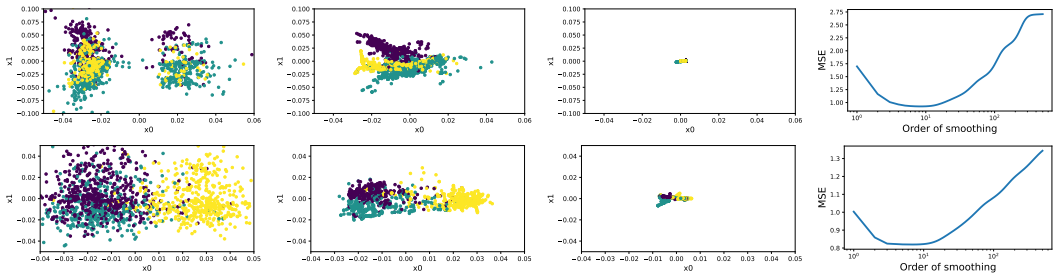


Figure 1: Illustration of both beneficial smoothing and oversmoothing on Cora [32] (top) and Citeseer [14] (bottom). **From left to right:** node features after performing respectively $k = 0, 10$, and 500 steps of mean aggregation, along the first two principal-components (of the original unsmoothed features), for three classes of nodes for better visibility. **Figure on the right:** Mean Square Error of Linear Ridge Regression (LRR) on the smoothed features with respect to the order of smoothing k. We observe that smoothing first gather same-labels nodes and improves learning, before they eventually collapses to a single point (note that here we show LRR for consistency with the analysis presented in this paper, even though these are node classification tasks).

where the $a_{ij} \in \mathbb{R}_+$ are the entries of the adjacency matrix of the graph: either positive edge weights or $0, 1$ for unweighted edges, and Ψ is some function (usually a Multi-Layer Perceptron). In other words, the aggregation process is a weighted average over the neighbors. As we will see, it corresponds to a multiplication by (identity minus) the *random walk Laplacian* of the graph.

While MP is a natural and rather general framework, its limitations were quickly observed by researchers and practitioners. Foremost among them is the so-called *oversmoothing* phenomenon [27]: as the GNN gets deeper and many rounds of MP are performed, the node features $z_i^{(k)}$ tend to become too similar across the graph, especially for small-world graphs with small diameter. Oversmoothing prevents GNN from being too deep unless one is particularly careful. A non-negligible part of the literature is dedicated to fighting oversmoothing with various strategies (see below).

On the theoretical side, oversmoothing has mostly been analyzed in the infinite-layer limit $k \to \infty$. In this case, classical spectral analysis of graph operators such as the Laplacian can be leveraged to indeed show that node features will always converge to some limit that carries a limited amount of information [34]. This is particularly true for mean aggregation (2), with *a constant limit across all nodes* for a connected graph, see Sec. 3. Unlike some other graph operators such as the symmetric normalized Laplacian, where the limit still carries a small amount of information such as the degrees, with the random walk Laplacian *all information* is lost in the limit (beyond a single constant).

However, there has been little research at the other end of the spectrum, showing that *some smoothing is useful for learning*, despite this fact being intuitively and empirically obvious. Generally, researchers show the power of GNNs for a *sufficient* (unbounded) number of layers, such as the now-famous ability to distinguish graph isomorphism as well as the Weisfeiler-Lehman test and all its variants [47, 30], the ability to compute some graph functions [28], and so on. Since these results are valid for an unbounded number of layers, the settings adopted in these works are, by definition, incompatible with non-informative oversmoothing. To our knowledge, there is no work that formally **models both phenomena at once**: *some* smoothing is provably useful for learning, while *too much* smoothing inevitably leads to oversmoothing.

This work aims to fill this gap. We showcase two representative exemples, of regression and classification, on which *linear* GNNs (aka, here, simply Linear Ridge Regression (LRR) on smoothed features) are subject to this double phenomenon. Note that restricting ourselves to *mean aggregation* makes this claim quite non-trivial: in the absence of any "informative" node features, no information can be recovered by mean aggregation alone. For instance, it leaves constant node features unchanged, and the limit $k \to \infty$ is always a constant. So the challenge is the following: node features must carry *some* information, such that a finite number of steps of mean aggregation *provably increases the amount of useful information*, before it loses it in the limit. See Fig. 1 for an illustration.

To show this we adopt on a model of latent space random graphs, with node features. The latter contain partial information about the unobserved latent variables on which both the labels and the graph structure depend. On our examples, we prove that with high probability, graph smoothing improves performance before oversmoothing occurs. We identify two key phenomena for this:

smoothing shrinks non-principal directions in the data faster than principal ones (Sec. 4), and shrinks communities faster than they collapse together (Sec. 5). Although our theoretical settings are obviously simplified, we believe it is a step towards a better comprehension of graph aggregation and of the relationship between node features and graph structure, at the heart of many phenomena in graph machine learning.

Related Work Oversmoothing [27] is a very active area of research in geometric deep learning, and an exhaustive list of works would be out of scope here. The research has been mainly focused on novel architectures to relieve it, such as residual mechanisms [26, 10], randomly dropping connections [18], introducing local jumps [48], clever normalizations [50, 17, 5, 37] or regularizations [9], among others. Some works have acknowledged the important role of the aggregation function, and proposed new exotic diffusion strategies [5] or to optimize it [24]. On the theoretical side, it has been mainly shown that repeatedly applying graph smoothing operators indeed induces convergence of the node features [34]. In this work, we analyze a model that present both the benefits of finite smoothing despite oversmoothing in the limit.

Our theoretical framework is based on simplified *linear* GNNs and random graphs that explicitely model the dependence between labels, node features, and graph structure. Despite their simplicity, linear GNNs, sometimes called Simplified Graph Convolutional networks (SGC), have been observed to exhibits relatively good performance [45, 33] and are routinely used in theoretical analyses [51]. Random graphs have been used extensively to analyze graph machine learning algorithms [43, 35] and the theoretical properties of GNNs such as stability [21, 36], transferability [25] or universality [22]. Our model crucially includes observed node features, an essential part in analyzing the smoothing process. They have been shown to be correlated to sought-for labels in real graphs [13], and that this fact is key in the success of GNNs. Our proof is in fact more akin to analyzing a graph diffusion process [31]: given appropriate initial conditions (observed node features), at initial time the diffusion produces a better signal for learning, before it eventually collapses to a single point. To the best of our knowledge, this is the first proof of this kind in a machine learning context.

Outline We describe our framework in Sec. 2. In Sec. 3, we briefly prove the oversmoothing phenomenon when $k \to \infty$, which is just the Markov chains ergodic theorem in our settings. In Sec. 4, we study a regression problem. We derive an expression that predicts with good accuracy the optimal smoothing order $k^\star$ in some cases. In Sec. 5, we study a classification problem between two Gaussians. Although we formally prove the *existence* of $k^\star > 0$, deriving an explicit expression for the risk is still open in this case. Code to reproduce the figures is available at `https://github.com/nkeriven/graphsmoothing`.

2 Preliminaries

Notations. The norm $\|\cdot\|$ is the Euclidean norm for vectors and spectral norm for (rectangular) matrices. For a psd matrix Σ, the Mahalanobis norm is $\|x\|_\Sigma^2 \overset{\text{def.}}{=} x^\top \Sigma x$. The determinant of a matrix is $|S|$, and its smallest eigenvalue is $\lambda_{\min}(S)$. The multivariate Gaussian distribution with mean μ and covariance Σ is denoted by $\mathcal{N}_{\mu,\Sigma}(x) = \det(2\pi\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}\|x-\mu\|_{\Sigma^{-1}}^2}$. We will use the shortened notations $\mathcal{N}_\mu = \mathcal{N}_{\mu,\mathrm{Id}}$ and $\mathcal{N} = \mathcal{N}_0$. Our bounds will involve various multiplicative constants $\mathrm{poly}(\cdot)$ which are polynomials in their input.

SSL. In this paper, we consider Semi-Supervised Learning (SSL) [8, 23] on an undirected graph of size n. We observe a weighted adjacency matrix $A = [a_{ij}]_{i,j=1}^n \in \mathbb{R}_+^{n\times n}$ as well as *node features* $z_1, \dots z_n \in \mathbb{R}^p$ at each node of the graph. We also observe *some* labels $y_1, \dots, y_{n_{\text{tr}}} \in \mathbb{R}$ at training time and aim to predict the remaining labels $y_{n_{\text{tr}}+1}, \dots, y_n$. In a classification framework, $y \in \{-1, 1\}$. For simplicity, we assume that n_{tr} and $n_{\text{te}} = n - n_{\text{tr}}$ are both in $\mathcal{O}(n)$[1]. We denote by $Z \in \mathbb{R}^{n\times p}$ the matrix whose rows contain the node features, $Z_{\text{tr}}, Z_{\text{te}}$ respectively its first n_{tr} and last n_{te} rows, and similarly $Y_{\text{tr}}, Y_{\text{te}}$ the vectors containing the observed and non-observed labels.

Graph smoothing with mean aggregation. Here we consider a simplified situation of *linear* GNN with mean aggregation, that is, equation (2) with linear Ψ. Since all linear weights collapses into a

[1] while this is an important topic in SSL [4], here we do not focus on the number of needed labels and perform an asymptotic analysis instead.

single matrix, a linear GNN with k layers just corresponds to performing k rounds of mean aggregation on the node features, then learning on the smoothed features. We denote by $d_A = [\sum_i a_{ij}]_j \in \mathbb{R}_+^n$ the vector containing the degrees of the graph and $D = \mathrm{diag}(d_A)$. Assuming that all degrees are non-zero, performing one round of mean aggregation corresponds to multiplying Z by $L = D^{-1}A$. Note that $\mathrm{Id} - L$ is then the *random walk Laplacian* of the graph. The smoothed node features after k rounds of mean aggregation are:

$$Z^{(k)} = L^k Z \,.$$

Each row, denoted by $z_i^{(k)} \in \mathbb{R}^p$, contains the smoothed features of an individual node. Similar to the non-smoothed features, its first n_{tr} and last n_{te} rows are denoted $Z_{\text{tr}}^{(k)}, Z_{\text{te}}^{(k)}$.

Learning. In this paper, we consider learning with a Mean Square Error (MSE) loss and Ridge regularization. For $\lambda > 0$, the regression coefficients vector on the smoothed features is

$$\hat{\beta}^{(k)} \overset{\text{def.}}{=} \mathrm{argmin}_\beta \tfrac{1}{2n_{\text{tr}}} \left\| Y_{\text{tr}} - Z_{\text{tr}}^{(k)}\beta \right\|^2 + \lambda \left\|\beta\right\|^2 = \left(\tfrac{(Z_{\text{tr}}^{(k)})^\top Z_{\text{tr}}^{(k)}}{n_{\text{tr}}} + \lambda \mathrm{Id} \right)^{-1} \tfrac{(Z_{\text{tr}}^{(k)})^\top Y_{\text{tr}}}{n_{\text{tr}}} \tag{3}$$

Then, the test risk is defined as

$$\mathcal{R}^{(k)} \overset{\text{def.}}{=} n_{\text{te}}^{-1} \left\| Y_{\text{te}} - \hat{Y_{\text{te}}}^{(k)} \right\|^2 \quad \text{where } \hat{Y_{\text{te}}}^{(k)} = Z_{\text{te}}^{(k)} \hat{\beta}^{(k)} \tag{4}$$

It is well known that when $k \to \infty$, the matrix L^k will converge to a matrix with constant rows, and $\mathcal{R}^{(\infty)} \overset{\text{def.}}{=} \lim_{k\to\infty} \mathcal{R}^{(k)}$ will just be close to the variance of Y, see Sec. 3 for a precise statement. Very often, this degrades the results with respect to doing a simple linear regression: $\mathcal{R}^{(0)} < \mathcal{R}^{(\infty)}$. Our goal is to illustrate some situations where a finite amount of smoothing provably improves the test risk, that is, there is an optimal $k^\star > 0$ such that $\mathcal{R}^{(k^\star)} < \min(\mathcal{R}^{(0)}, \mathcal{R}^{(\infty)})$.

Random graph model. To perform a fine-grained analysis of our problem, we need a statistical model linking the graph, the node features, and the labels. We adopt popular *latent space random graph models* akin to graphons [29]. Although such models are obviously idealized, we believe that they faithfully convey the main insights. In these models, to each node i is associated an *unobserved latent variable* $x_i \in \mathbb{R}^d$ with $d \geqslant p$ (often $d \gg p$), and edge weights are assumed to be equal to $a_{ij} = W(x_i, x_j)$ where $W : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ is a *connectivity kernel*. Note that edges may also be taken as *random Bernoulli variables*, but we do not consider this here for simplicity. Moreover, we consider that the (x_i, y_i) are drawn iid from some joint distribution, and the node features are a linear projection of the latent variables to a lower dimension: $z_i = M^\top x_i$ for some unknown $M \in \mathbb{R}^{d\times p}$ that satisfies $M^\top M = \mathrm{Id}_p$. At the end of the day:

$$\forall i, j, \quad (x_i, y_i) \overset{iid}{\sim} P, \quad z_i = M^\top x_i, \quad a_{ij} = W(x_i, x_j) \tag{5}$$

For this model, note that

$$Z^{(k)} = L^k Z = L^k X M = X^{(k)} M \text{ where } X^{(k)} = L^k X$$

In other words, the smoothed node features $Z^{(k)}$ also correspond to a linear projection of the (unknown) *smoothed latent variables* $X^{(k)}$. To summarize, compared to "classical" machine learning on the (x_i, y_i), we do *not* observe directly the x_i, but only a projection of them $z_i = M^\top x_i$. Although we assume that M is orthogonal, we do *not* assume that it is "information-preserving" (e.g. it does not satisfy the Johnson-Lindenstrauss lemma), but rather that information *is* lost between the x and the z. However, we also observe the graph $W(x_i, x_j)$. Our goal is illustrate how mean aggregation may restore some of the lost information.

In the rest of the paper, we use the Gaussian kernel with a small additive term $\varepsilon > 0$:

$$W(x, y) = \varepsilon + W_g(x, y) \quad \text{where } W_g(x, y) \overset{\text{def.}}{=} e^{-\frac{1}{2}\|x-y\|^2} \tag{6}$$

The coefficient ε is added to lower-bound the degrees of the graph and avoid degenerate situations. While this seems to be needed for our current proof technique, we use $\varepsilon = 0$ in Fig. 2 and 3. The Gaussian kernel is a classical model in theoretical graph machine learning [38].

3 Oversmoothing

In this section, we briefly examine the oversmoothing case, when $k \to \infty$ while all other parameters are fixed. In this case, it is well-known that all node features converge even for general GNNs [34]. For completeness, we state below this result in our settings. We have the following well-known ergodic theorem for stochastic matrices such as L.

Theorem 1 (Ergodic theorem for stochastic matrices, e.g. [2, Thm. 4.2].)**.** *Recall that d_A is the vector of degrees, let $\bar{d} = d_A / d_A^\top 1_n$. We have*

$$L^k \xrightarrow[k\to\infty]{} 1_n \bar{d}^\top \tag{7}$$

This easily allows us to prove the next result.

Corollary 1. *We have the following*

$$\hat{Y_{\text{te}}}^{(k)} \xrightarrow[k\to\infty]{} \left(\tfrac{\|v\|^2}{\lambda + \|v\|^2} \bar{y}_{\text{tr}} \right) 1_{n_{\text{te}}} \tag{8}$$

where $v = Z^\top \bar{d}$ and $\bar{y}_{\text{tr}} = n_{\text{tr}}^{-1} \sum_{i=1}^{n_{\text{tr}}} y_i$.

Proof. We use Thm. 1 to get $L^k X M \to 1_n v^\top$, and $(\lambda \mathrm{Id} + v v^\top)^{-1} v = \frac{v}{\lambda + \|v\|^2}$. □

Hence, in the limit $k \to \infty$, the predicted labels become all equal. When $\lambda \approx 0$, this value is, as expected, the average of the labels in the training set $\bar{y}_{\text{tr}}$. Using simple concentration inequalities, it is generally easy to show that $\mathcal{R}^{(\infty)} \approx \mathrm{Var}(y) + \mathcal{O}(1/\sqrt{n})$. In most cases, this leads to situations where $\mathcal{R}^{(0)} < \mathcal{R}^{(\infty)}$, that is, it is better to perform regression directly on the node features. In the next sections, we analyze some examples where smoothing provably helps.

4 Finite smoothing: Linear Regression

In this section, we consider a problem of linear regression on Gaussian data. We consider $x \sim \mathcal{N}_{0,\Sigma}$ for some positive definite covariance matrix Σ, and $y = x^\top \beta^\star$, without noise for simplicity (noise would just add an additional variance terms to all our bounds). We will first describe our main result that holds under a certain condition that is not necessarily easy to interpret, then give a sketch of proof in Sec. 4.1, and an example in dimension $d = 2$ where this assumption is satisfied in Sec. 4.2.

For a symmetric positive semi-definite matrix $S \in \mathbb{R}^{d \times d}$, we define the following function

$$R_{\text{reg.}}(S) \overset{\text{def.}}{=} (\Sigma^{\frac{1}{2}} \beta^\star)^\top \left(\mathrm{Id} - S^{\frac{1}{2}} M (\lambda \mathrm{Id} + M^\top S M)^{-1} M^\top S^{\frac{1}{2}} \right)^2 (\Sigma^{\frac{1}{2}} \beta^\star) \in \mathbb{R}_+ \tag{9}$$

where we recall that M is the projection matrix to obtain the node features $z = M^\top x$. Note that it satisfies $0 \leqslant R(S) \leqslant \|\beta^\star\|_\Sigma^2$. Our result will be valid under the following assumption:

Assumption 1. *We have $R_{\text{reg.}}(\Sigma) > R_{\text{reg.}}((\mathrm{Id} + \Sigma^{-1})^{-2}\Sigma)$.*

Note that $(\mathrm{Id} + \Sigma^{-1})^{-2}\Sigma$ is indeed symmetric since $(\mathrm{Id} + \Sigma^{-1})^{-1}$ and Σ permute. Our main result can be stated informally as follows, it is detailed in the next section along with a sketch of proof. Recall that the kernel is taken as (6).

Theorem 2 (Existence of optimal smoothing for regression.)**.** *Take any $\rho > 0$, and suppose that Assumption 1 holds. If ε is sufficiently small and n is sufficiently large, then with probability $1 - \rho$, there is $k^\star > 0$ such that $\mathcal{R}^{(k^\star)} < \min(\mathcal{R}^{(0)}, \mathcal{R}^{(\infty)})$.*

4.1 Sketch of proof

As we will see, it is easy to show that $\mathcal{R}^{(0)} < \mathcal{R}^{(\infty)}$ with high probability. Our main goal will therefore be to show that $\mathcal{R}^{(1)} < \mathcal{R}^{(0)}$ with high probability under Assumption 1, which is sufficient to show the existence of an optimal $k^\star \geqslant 1$. Using concentration inequalities, we will prove a rigorous non-asymptotic bound for $\mathcal{R}^{(1)}$. In the next section, we also derive an intuitive expression for $\mathcal{R}^{(k)}$ (although without rigorous proof), which we observe to match the numerics quite well.

The first step is to derive a closed form expression for $\mathcal{R}^{(0)}$, which is fairly easy using standard concentration techniques for subgaussian variables. The next result is proved in App. A.1.

Theorem 3 (Regression risk without smoothing.)**.** *With probability at least* $1-\rho$*,*

$$\mathcal{R}^{(0)} = R_{\text{reg.}}(\Sigma) + \mathcal{O}\left(\frac{\|\Sigma\| \, \|\beta^\star\|^2 \, d\sqrt{\log(1/\rho)}}{(\lambda+\lambda_{\min})\sqrt{n}}\right) \tag{10}$$

where $\lambda_{\min} = \lambda_{\min}(M^\top \Sigma M)$*.*

As expected, when $p = d$, $M = \mathrm{Id}$ and $\lambda \to 0$, we have $R_{\text{reg.}}(\Sigma) \to 0$ and the risk is exactly 0 in the infinite sample limit (recall that we have assumed zero noise on the labels). When $p < d$ however, the limit risk is generally non-zero. The worst case is obtained when $\Sigma\beta^\star$ is orthogonal to $M^\top$, where the risk reaches its maximum at $\|\beta^\star\|_\Sigma^2 = \mathbb{E}\,|y|^2$. Since this is the variance of y, this is also $\lim_{n\to\infty} \mathcal{R}^{(\infty)}$, hence we always have $\mathcal{R}^{(0)} \leqslant \mathcal{R}^{(\infty)}$ with high probability for n large enough.

Let us now turn to computing the risk after one step of smoothing $k = 1$. We define $\Sigma^{(k)} = (\mathrm{Id} + \Sigma^{-1})^{-2k}\Sigma$. The main result of this section is the following.

Theorem 4 (Regression risk with one step of smoothing.)**.** *With probability at least* $1-\rho$*,*

$$\mathcal{R}^{(1)} = R_{\text{reg.}}(\Sigma^{(1)}) + \mathcal{O}\left(C\varepsilon^{1/5}\right) + \mathcal{O}\left(\frac{C' \log n \sqrt{d+\log(1/\rho)}}{(\lambda+\lambda_{\min})\sqrt{n}}\right) \tag{11}$$

where $C = \mathrm{poly}(\|\Sigma\|, e^d, |\mathrm{Id}+\Sigma|)$*,* $C' = \mathrm{poly}(\varepsilon^{-1}, \|\Sigma\|, \|\beta^\star\|)$ *and* $\lambda_{\min} = \lambda_{\min}(M^\top \Sigma^{(1)} M)$*.*

This theorem gives a limiting expression or $\mathcal{R}^{(1)}$ with two additional error terms. The first goes to 0 with ε and is due to the deviation from the kernel (6) to the exact Gaussian kernel W_g. The second term goes to 0 when $n \to \infty$ and is controlled via concentration inequalities. The limit risk when $\varepsilon \to 0$, $n \to \infty$ is $\mathcal{R}^{(1)} \approx R_{\text{reg.}}(\Sigma^{(1)})$, which is strictly lower than $R_{\text{reg.}}(\Sigma)$ by Assumption 1 and proves Theorem 3. Note that, to get $\mathcal{R}^{(1)} < \mathcal{R}^{(0)}$, we generally need $\varepsilon \lesssim e^{-d}$ and therefore $n \gtrsim e^d$, which seems to be an unavoidable artifact in our current proof technique.

Let us try to better understand Assumption 1 by sketching the proof of Thm. 4. The proof relies on an approximate description of the distribution of the smoothed node features $z_i^{(1)} = M^\top x_i^{(1)}$ where we recall that the $x_i^{(1)}$ are the rows of $X^{(k)} = L^k X$. We define $d(x) = |\mathrm{Id}+\Sigma|^{-\frac{1}{2}} \, e^{-\frac{1}{2}\|x\|^2_{(\mathrm{Id}+\Sigma)^{-1}}}$ and

$$\varphi_{\text{reg.}}(x) = \frac{d(x)}{d(x)+\varepsilon}(\Sigma^{-1}+\mathrm{Id})^{-1}x\,. \tag{12}$$

Then, using some chaining concentration inequalities for subgaussian variables (Lemma 7 in the appendix) and properties of Gaussian distributions (Lemma 5), we can prove the following.

Lemma 1. *With probability at least* $1-\rho$*, for all* $i = 1, \dots, n$*:*

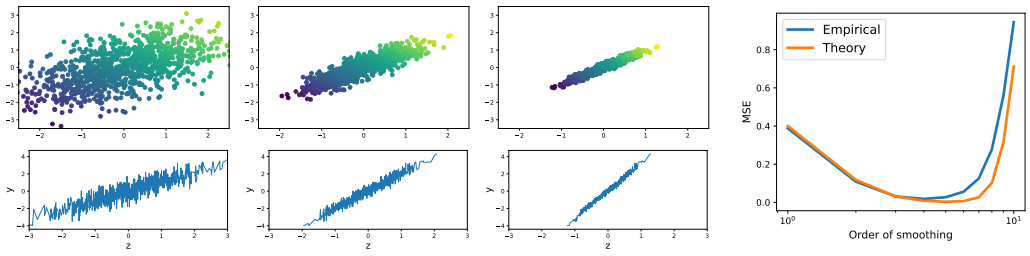
$$\left.\begin{aligned} \left\|x_i^{(1)} - \varphi_{\text{reg.}}(x_i)\right\|_{\Sigma^{-1}} \\ \left\|\Sigma^{-\frac{1}{2}}\left(x_i^{(1)}(x_i^{(1)})^\top - \varphi_{\text{reg.}}(x_i)\varphi_{\text{reg.}}(x_i)^\top\right)\Sigma^{-\frac{1}{2}}\right\| \end{aligned}\right\} \lesssim \frac{C\log n(\sqrt{d+\log(1/\rho)})}{\sqrt{n}} \tag{13}$$

where $C = \mathrm{poly}(\varepsilon^{-1}, \|\Sigma\|, |\mathrm{Id}+\Sigma|)$*.*

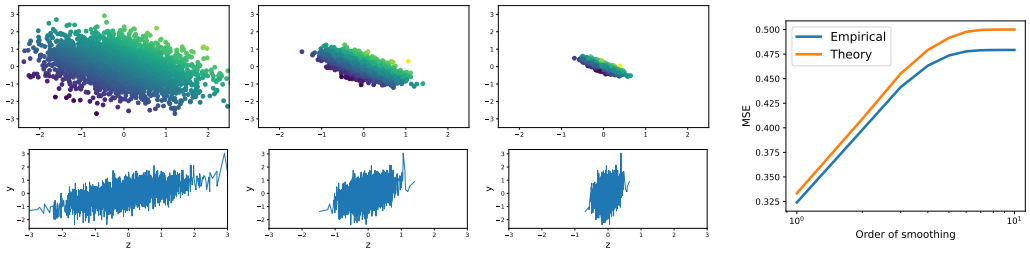
Hence the smoothed latent variables behaves almost like $(\mathrm{Id}+\Sigma^{-1})^{-1}x$, up to a deviation ε that is handled in Lemma 3 in the appendix. The covariance of these data is $\Sigma^{(1)} = (\mathrm{Id}+\Sigma^{-1})^{-2}\Sigma$, hence we can adapt the proof of Thm. 3 to obtain Thm. 4. All details are given in App. A.2.

4.2 Intuition and exact computation in dimension $d = 2$

We proved above that $x^{(1)}$ behaves almost like $(\mathrm{Id}+\Sigma^{-1})^{-1}x$, whose covariance is $\Sigma^{(1)}$. Similarly, by applying repeated smoothing we can extrapolate that $x^{(k)}$ behaves like $(\mathrm{Id}+\Sigma^{-1})^{-k}x$, such that $\mathcal{R}^{(k)} \approx R_{\text{reg.}}(\Sigma^{(k)})$. The rigorous proof of this fact becomes increasingly complicated and is skipped here. The matrix $\Sigma^{(k)}$ has the same eigendecomposition as Σ, but where every eigenvalue λ_i



(a)



(b)

Figure 2: Illustration of mean aggregation smoothing on the regression example described in Sec. 4.2. **For both subfigures: First three figures on the left, top:** *unobserved* latent variables $X^{(k)}$ in dimension $d = 2$ where the colors are the Y; **bottom:** observed node features $Z^{(k)} = X^{(k)}M$ in dimension $p = 1$ on the x-axis, labels Y on the y-axis. **From left to right**, three order of smoothing $k = 0, 1$ and 2 are represented. **Figure on the right:** comparison of empirical and theoretical MSE given by (14) with respect to order of smoothing k. **Subfig. a:** $\lambda_1 = 2, \lambda_2 = 1/2$ (smoothing does help), **Subfig. b:** $\lambda_1 = 1/2, \lambda_2 = 1$ (smoothing does not help).

is replaced by $\lambda_i^{(k)} = (1 + 1/\lambda_i)^{-2k}\lambda_i$. This can be interpreted as follows: when $\lambda_i \gg 1$ is large, $\lambda_i^{(1)} \sim \lambda_i$, while if $\lambda_i \ll 1$ is small, $\lambda_i^{(1)} \sim \lambda_i^{2k+1}$ (note that the constant "1" here is due to our kernel (6), it is not inherently significant). Hence smoothing **shrinks the directions of the small eigenvalues faster than that of the large ones**. Thus, if $\beta^\star$ is mostly aligned with the eigenvectors of large eigenvalues, shrinking the small eigenvalues may *reduce unwanted noise* that emerges when projecting the node features $z = M^\top x$. On the other hand, if all eigenvalues of Σ are equal, then $\Sigma^{(k)} \propto \Sigma$, and smoothing *does not help*, since in the limit $\lambda = 0$, the risk is invariant to scaling $R_{\text{reg.}}(aS) = R_{\text{reg.}}(S)$. Worse, we will see on an example below that smoothing can actually degrade the performance when $\beta^\star$ is unpropery aligned.

We illustrate this in dimension $d = 2$. Consider the following settings: $d = 2, p = 1$, Σ has two eigenvalues $\lambda_1 \gg 1$ and $\lambda_2 \ll 1$, with respective eigenvectors $u_1 = [1, 1]/\sqrt{2}$ and $u_2 = [-1, 1]/\sqrt{2}$, and $\beta^\star$ is fully correlated with the first eigenvector: $\beta^\star = bu_1$. Finally, $M^\top = [1, 0]$ is the projection on the first coordinate. This situation is represented in Fig. 2. In this case, we can compute explicitely:

$$\mathcal{R}^{(k)} \approx R_{\text{reg.}}(\Sigma^{(k)}) = \lambda_1 b^2 \frac{(2\lambda + \lambda_2^{(k)})^2 + \lambda_2^{(k)}\lambda_1^{(k)}}{(2\lambda + \lambda_1^{(k)} + \lambda_2^{(k)})^2} \tag{14}$$

So, if $\lambda_2^{(k)}$ decreases faster than $\lambda_1^{(k)}$, this function will first decrease to a minimum of approximately $\lambda_1 b^2 \left(\frac{2\lambda}{2\lambda + \lambda_1^{(k^\star)}}\right)^2$ (when $\lambda_2^{(k)} \approx 0$), before increasing again to $\lambda_1 b^2 = \|\beta^\star\|_\Sigma^2 = \lim_{n\to\infty} \mathcal{R}^{(\infty)}$. This is illustrated in Fig. 2, for $\lambda_1 = 2$ and $\lambda_2 = 1/2$, where we empirically observes a minimum $k^\star$ that matches rather well the one predicted by (14).

Homophily vs. Heterophily and a failure case In graph theory, *homophily* refers to the concept that linked nodes tend to display similar properties: for instance, friends on social networks have similar preferences, and so on. In graph machine learning, it generally means that linked nodes tend to have similar node features and labels. This concept is at the core of many graph signal processing and graph machine learning methods: for instance, spectral clustering is akin to a low-pass filter on the graph structure. However, it has been observed that real graphs may sometimes exhibit a low

level of homophily [51, 5]. They are rather said to be *heterophilic*, a somewhat less "well-defined" concept: in heterophilic graphs, linked nodes can be similar or dissimilar, some attributes can be homophilic and others heterophilic, and so on.

In our settings, at first glance it seems that our very regular random graph model always results in homophilic graphs, as the Gaussian kernel decreases with the distance between latent variables, and the latter are strongly linked with the node features. This is partly true, however is it also possible that nodes linked by a "strong" edge (with a high weight) have very different labels, which can be said to be a (toy) example of heterophily. For instance, consider the 2D linear regression example above given by (14). We have seen that when the regression vector is in the direction of the eigenvector corresponding to a high eigenvalue, then beneficial smoothing appears, as it reduces the noise in the observed node features (Fig. 2a). However, when the regression vector is instead in the low-eigenvalue direction, then close-by latent variables have very different labels, and the graph is more heterophilic. In this case, beneficial smoothing does *not* appear, and any smoothing strictly degrades the MSE! (Fig. 2b) This is due to the fact that in this case the "information" in node features vanishes faster than the noise. Of course, this is an exceedingly simple model of heterophily, and a better understanding and modelization of this phenomenon remains an outstanding open question.

Discussion Recent literature on GNNs have adressed both oversmoothing and heterophily by clever normalization techniques [50, 17, 5, 37], combined with quantitative metrics of these phenomena [49, 51]. However, these tend to indiscriminately combat oversmoothing, without taking into account potential beneficial smoothing. In future work, our analysis could help designing more detailed normalization methods, e.g. after some estimation step that would identify which directions in the data are squeezed by smoothing, and which of them are relevant or not for learning.

5 Finite smoothing: classification

In this section, we examine a simple classification problem for two balanced classes with Gaussian distribution with identity covariance. The distribution of the labels and latent variables is:

$$(x, y) \sim (1/2)(\mathcal{N}_\mu \otimes \{1\} + \mathcal{N}_{-\mu} \otimes \{-1\}) \tag{15}$$

That is, with equal probability x is drawn from $\mathcal{N}_\mu$ and $y = 1$, or $x \sim \mathcal{N}_{-\mu}$ and $y = -1$. As $\|\mu\|$ increases, the problem become simpler, there is an extensive literature on this problem [12, 40, 3]. Note that in this case z_i are also Gaussian, with mean $\nu \overset{\text{def.}}{=} M^\top \mu$ or $-\nu$ and identity covariance.

We note that this is not a *difficult* problem *per se*, and that *linear regression with the MSE* is certainly not the method of choice to solve it: there are plethora of losses better adapted to binary classification such as the binary cross-entropy (left for future investigations), or even other dedicated methods: a Spectral Clustering algorithm on the graph alone would be able to perform the classification task under some mild hypotheses [40, 1] (without using the node features!). Nevertheless, let us recall that our main goal is to illustrate the smoothing phenomenon, and as we will see, the interpretation here will be quite different from the previous section. Our main result is the following.

Theorem 5 (Existence of optimal smoothing for classification.)**.** *Take any $\rho > 0$. If ε is sufficiently small, and $\|\mu\|, n$ are sufficiently large, and $\left\|M^\top \mu\right\| > 0$, then with probability $1 - \rho$, there is $k^\star > 0$ such that $\mathcal{R}^{(k^\star)} < \min(\mathcal{R}^{(0)}, \mathcal{R}^{(\infty)})$.*

Note that we have assumed $\|\mu\|$ to be sufficiently large here. However, we do *not* assume that $\left\|M^\top \mu\right\|$ is large (just non-zero), and the classification problem on the z_i alone may be very difficult. The rest of this section presents a sketch of proof and intuitions behind this theorem.

5.1 Sketch of proof and intuition

As in the previous section, it will be easy to show that $\mathcal{R}^{(0)} < \mathcal{R}^{(\infty)}$ with high probability, and we will prove that $\mathcal{R}^{(1)} < \mathcal{R}^{(0)}$ with high probability. Again, we start by providing an expression for $\mathcal{R}^{(0)}$. For $s \in \mathbb{R}_+$, we define the following function

$$R_{\text{cl.}}(s) = \frac{(s + \lambda)^2 + s \left\|\nu\right\|^2}{(s + \lambda + \left\|\nu\right\|^2)^2} \tag{16}$$

The next result is proved in App. B.1. Recall that $\nu = M^\top \mu$.

Theorem 6 (Classification risk without smoothing.)**.** *With probability at least* $1-\rho$*,*

$$\mathcal{R}^{(0)} = R_{\text{cl.}}(1) + \mathcal{O}\left(\frac{\|\nu\|^4 p\sqrt{\log(1/\rho)}}{\sqrt{n}}\right) \tag{17}$$

When $\|\nu\| \to \infty$, the risk goes to 0, as expected, since the Gaussians get further and further away. However, when $\|\nu\| \to 0$, which can happen *either* when $\|\mu\|$ is small or when M becomes orthogonal to μ, the risk goes to 1, its worst value, for random guesses. Since it is also the variance of y, we have indeed $\mathcal{R}^{(0)} \leqslant 1 \approx \mathcal{R}^{(\infty)}$ with high probability for n large enough.

Let us now turn to computing the risk after one step of smoothing $k = 1$. The main result of this section is the following.

Theorem 7 (Classification risk with one step of smoothing.)**.** *With probability at least* $1-\rho$*,*

$$\mathcal{R}^{(1)} = R_{\text{cl.}}(1/4) + \mathcal{O}\left(C\left(\varepsilon^{\frac{1}{4}} + \frac{1}{\varepsilon^3}e^{-\frac{\|\mu\|^2}{4}}\right)\right) + \mathcal{O}\left(\frac{C'(\log n)(\sqrt{d+\log(1/\rho)})}{\sqrt{n}}\right) \tag{18}$$

where $C = \text{poly}(\|\mu\|, e^d)$ *and* $C' = \text{poly}(\varepsilon^{-1}, \|\mu\|)$.

This theorem show that $\mathcal{R}^{(1)} \approx R_{\text{cl.}}(1/4)$ with two additional error terms. First of all, a quick function study shows that $R_{\text{cl.}}(1/4) < R_{\text{cl.}}(1)$ when $\|\nu\| > 0$, which shows Thm. 5 when the errors are small enough. The last error term goes to 0 when $n \to \infty$ and is controlled via concentration inequalities. The first one is small when ε is small and $\|\mu\|$ is large enough. We remark that, unlike the previous section where the error terms vanished in the limit $\varepsilon \to 0, n \to \infty$, here there is a non-zero error term due to $\|\mu\|$ whose explicit expression is still open. Hence, for instance, the discrepancy between the empirical observations and the theory in Fig. 3 compared to Fig. 2. Note that, as in the previous section, we need at least $\varepsilon \lesssim e^{-d}$ and $n \gtrsim e^d$. However, here we also need $\|\mu\| \gtrsim \sqrt{d}$. This rate is similar to early analyses of Gaussian Mixture learning [12], although they have been greatly improved since [3].

As previously, we define here $d_\mu(x) \overset{\text{def.}}{=} 2^{-d/2}e^{-\frac{\|x-\mu\|^2}{4}}$, and

$$\varphi_{\text{cl.}}(x) = \frac{d_\mu(x)\left(\frac{x+\mu}{2}\right) + d_{-\mu}(x)\left(\frac{x-\mu}{2}\right)}{2\varepsilon + d_\mu(x) + d_{-\mu}(x)} \tag{19}$$

The following result is similar to Lemma 1 and is shown in App. B.2.

Lemma 2. *With probability at least* $1-\rho$*,*

$$\left.\begin{matrix}\sup\limits_{i=1,\dots,n}\left\|x_i^{(1)} - \varphi_{\text{cl.}}(x_i)\right\| \\ \sup\limits_{i=1,\dots,n}\left\|x_i^{(1)}(x_i^{(1)})^\top - \varphi_{\text{cl.}}(x_i)\varphi_{\text{cl.}}(x_i)^\top\right\|\end{matrix}\right\} \lesssim \frac{\text{poly}(\varepsilon^{-1})\log n(\sqrt{d}+\sqrt{\log(1/\rho)})}{\sqrt{n}} \tag{20}$$

Let us now examine $\varphi_{\text{cl.}}(x)$ closer. In the limit $\varepsilon \to 0$, $\varphi_{\text{cl.}}(x)$ is a convex combination of $(x+\mu)/2$ and $(x-\mu)/2$. Hence, when $x \sim \mathcal{N}_\mu$, with high probability x is close to μ and $d_\mu(x) \gg d_{-\mu}(x)$, and in this case, $\varphi_{\text{cl.}}(x) \approx \frac{x+\mu}{2}$, whose distribution is $\mathcal{N}_{\mu,\text{Id}/4}$. The same reasoning applies to the other community. Hence, up to some error $\mathcal{O}\left(e^{-\|\mu\|^2/4}\right)$ due to the communities getting closer to each other, **the smoothed features in each community have the same mean but a reduced variance** $\text{Id}/4$, thus the limit risk $R_{\text{cl.}}(1/4)$ in our limit expression for $\mathcal{R}^{(1)}$. In other words, **the communities shrink faster than they collapses together**, and this reflects on the projected node features. An illustration of this phenomenon is given in Fig. 3. All proof details are in App. B.2.

5.2 Numerical illustration

In light of the proof of the theorem above, when x_i belongs to the first community and $x_i^{(1)} \approx \varphi_{\text{cl.}}(x_i) \approx \frac{x_i+\mu}{2}$, applying a second smoothing would transform it to $\frac{\varphi_{\text{cl.}}(x_i)+\mu}{2} \approx \frac{x_i+3\mu}{4}$, that

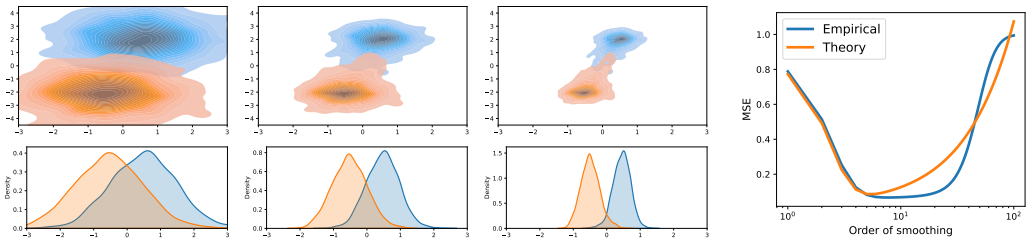


Figure 3: Illustration of mean aggregation smoothing on a classification task with two Gaussians with dimensions $d = 2$, $p = 1$, where M projects on the first coordinate. **First three figures on the left, top:** density of *unobserved* latent variables $X^{(k)}$ in dimension $d = 2$; **bottom:** density of observed node features $Z^{(k)} = X^{(k)}M$ in dimension $p = 1$. **From left to right**, three order of smoothing $k = 0, 1$ and 2 are represented (recall that the smoothing is agnostic to the labels, we cannot perform in-community smoothing). **Figure on the right:** comparison of empirical and theoretical MSE given by (21) with respect to order of smoothing k. For low k, the node features communities are indeed more and more separated, and learning improves.

is, it keeps the same mean but now has variance $\mathrm{Id}/16$. If we look at the proof, the error term $\mathcal{O}\left(e^{-\|\mu\|^2/4}\right)$ would become in this case $e^{-\frac{\|\mu\|^2}{2(1+1/4)}}$. While this expression is far from being exact and we do not a rigorous proof here (which seems far more complex than the case $k = 1$), we can infer some approximate expression:

$$\mathcal{R}^{(k)} \approx R_{\text{cl.}}(4^{-k}) + \mathcal{O}\left(\sum_{\ell=0}^{k-1} e^{-\frac{\|\mu\|^2}{2(1+4^{-\ell})}}\right) \tag{21}$$

Unlike the expression (14), the term $R_{\text{cl.}}(4^{-k})$ is strictly decreasing when k increases. Oversmoothing is modelled by the error term, for which we do not have an exact expression, and for which we suspect that the quality of approximation degrades as k increases. Nevertheless, we evaluate this expression on an example in Fig. 3 (with an adjusted multiplicative constant for the error term in (21)) and find that it is a reasonably good approximation, at least for small k.

6 Conclusion and outlooks

While the oversmoothing phenomenon $k \to \infty$ has been well characterized, until now there has been no theoretical studies that rigorously modelled both the benefits of finite smoothing before oversmoothing kicks in. In this paper, we adopted a simplified context of linear GNNs with mean aggregation and random graphs with partially observed latent variables, and proved on two representative examples the co-existence of both phenomena. We identified two mechanisms for the benefits of mean aggregation: it tends to shrink noisy principal components faster than meaningful ones, and it tends to gather nodes of the same community faster than they collapses together. We obtained theoretical expressions up to some error terms that matched the numerics quite well on simple synthetic data.

There are many outlooks to this work. First and foremost, deriving inspiration from our theoretical observations to design better methods of setting the order of smoothing in practical application is a major challenge. As seen in Fig. 1 in the introduction, both mechanisms that we identified seem to come into play on real data. However, many quantities appearing in the risks (14) and (21) need to be estimated. Second, extending our theory to more complex loss functions (especially for classification) and non-linear GNNs is crucial. Finally, our work is a step towards a better understanding of *the relationship between node features and graph structure*, which is at the heart of (over)smoothing, heterophily, and all graph machine learning methods. A more general theory, and more realistic models of random graphs to analyze it, is still an open question.

Acknowledgment

This work was supported by ANR JCJC GRandMa (ANR-21-CE23-0006). NK thanks S. Vaiter for inspiring discussions.

References

[1] Dimitris Achlioptas and Frank Mcsherry. On spectral learning of mixtures of distributions. *Learning Theory*, pages 458–469, 2005.

[2] Soren Asmussen. *Applied Probability and Queue*. Springer-Verlag New York, 2003.

[3] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *IEEE 51st Annual Symposium on Foundations of Computer Science*. Ieee, 2010.

[4] Shai Ben-David, Tyler Lu, and Dávid Pál. Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning. In *21st Annual Conference on Learning Theory, COLT 2008*, pages 33–44, 2008.

[5] Cristian Bodnar, Francesco Di Giovanni, Benjamin Paul Chamberlain, Pietro Liò, and Michael M. Bronstein. Neural Sheaf Diffusion: A Topological Perspective on Heterophily and Oversmoothing in GNNs. 2022.

[6] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. 2021.

[7] Michael M. Bronstein, Joan Bruna, Yann Lecun, Arthur Szlam, and Pierre Vandergheynst. Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[8] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. 2010.

[9] Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, pages 3438–3445, 2019.

[10] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. *37th International Conference on Machine Learning, ICML 2020*, PartF16814:1703–1713, 2020.

[11] Zhengdao Chen, Lisha Li, and Joan Bruna. Supervised community detection with line graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

[12] Sanjoy Dasgupta. Learning mixtures of Gaussians. In *IEEE 51st Annual Symposium on Foundations of Computer Science*, number May, 1999.

[13] Chi Thang Duong, Thanh Dat Hoang, Ha The Hien Dang, Quoc Viet Hung Nguyen, and Karl Aberer. On Node Features for Graph Neural Networks. *arXiv:1911.08795*, 2019.

[14] C. L. Giles, K. D. Bollacker, and S. Lawrence. CiteSeer: an automatic citation indexing system. *Proceedings of the ACM International Conference on Digital Libraries*, pages 89–98, 1998.

[15] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural Message Passing for Quantum Chemistry. In *International Conference on Machine Learning (ICML)*, pages 1–14, 2017.

[16] William L. Hamilton. *Graph Representation Learning*. 2020.

[17] Ningyuan Huang, Soledad Villar, Carey E. Priebe, Da Zheng, Chengyue Huang, Lin Yang, and Vladimir Braverman. From Local to Global: Spectral-Inspired Graph Neural Networks. pages 1–19, 2022.

[18] Wenbing Huang, Yu Rong, Tingyang Xu, Fuchun Sun, and Junzhou Huang. Tackling Over-Smoothing for General Graph Convolutional Networks. 14(8):1–14, 2020.

[19] Dejun Jiang, Zhenxing Wu, Chang Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics*, 13(1):1–23, 2021.

[20] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8):595–608, 2016.

[21] Nicolas Keriven, Alberto Bietti, and Samuel Vaiter. Convergence and Stability of Graph Convolutional Networks on Large Random Graphs. In *Advances in Neural Information and Processing Systems (NeurIPS) Spotlight*, pages 1–26, 2020.

[22] Nicolas Keriven, Alberto Bietti, and Samuel Vaiter. On the Universality of Graph Neural Networks on Large Random Graphs. In *Advances in Neural Information and Processing Systems (NeurIPS)*, 2021.

[23] Thomas N Kipf and Max Welling. Semi-Supervised Learning with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[24] Kwei Herng Lai, Daochen Zha, Kaixiong Zhou, and Xia Hu. Policy-GNN: Aggregation Optimization for Graph Neural Networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 461–471, 2020.

[25] Ron Levie, Wei Huang, Lorenzo Bucci, Michael Bronstein, and Gitta Kutyniok. Transferability of spectral graph convolutional neural networks. *Journal of Machine Learning Research*, 22:1–41, 2021.

[26] Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. DeepGCNs: Can GCNs go as deep as CNNs? *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:9266–9275, 2019.

[27] Qimai Li, Zhichao Han, and Xiao Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 3538–3545, 2018.

[28] Andreas Loukas. What graph neural networks cannot learn: depth vs width. In *ICLR*, 2020.

[29] László Lovász. Large networks and graph limits. *Colloquium Publications*, 60:487, 2012.

[30] Haggai Maron, Heli Ben-Hamu, Hadar Serviansky, and Yaron Lipman. Provably Powerful Graph Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–12, 2019.

[31] Naoki Masuda, Mason A. Porter, and Renaud Lambiotte. Random walks and diffusion on networks. *Physics Reports*, 716-717:1–58, 2017.

[32] Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

[33] Nicolo Navarin, Wolfgang Erb, Luca Pasa, and Alessandro Sperduti. Linear graph convolutional networks. *ESANN 2020 - Proceedings, 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, (October):151–156, 2020.

[34] Kenta Oono and Taiji Suzuki. Graph Neural Networks Exponentially Lose Expressive Power for Node Classification. In *International Conference on Learning Representation (ICLR)*, 2020.

[35] Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *Journal of Machine Learning Research*, 11:905–934, 2010.

[36] Luana Ruiz, Zhiyang Wang, and Alejandro Ribeiro. Graph and graphon neural network stability. (1):1–5, 2020.

[37] T. Konstantin Rusch, Benjamin P. Chamberlain, Michael W. Mahoney, Michael M. Bronstein, and Siddhartha Mishra. Gradient Gating for Deep Multi-Rate Learning on Graphs. pages 1–20, 2022.

[38] Minh Tang, Daniel L Sussman, and Carey E Priebe. Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics*, 41(3):1406–1430, 2013.

[39] Petar Veličković, Arantxa Casanova, Pietro Liò, Guillem Cucurull, Adriana Romero, and Yoshua Bengio. Graph attention networks. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, pages 1–12, 2018.

[40] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

[41] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing: Theory and Applications*, pages 210–268, 2009.

[42] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. 2018.

[43] Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *Annals of Statistics*, 36(2):555–586, 2008.

[44] Hongwei Wang, Jure Leskovec, Fuzheng Zhang, Miao Zhao, Wenjie Li, Mengdi Zhang, and Zhongyuan Wang. Knowledge-aware Graph Neural Networks with Label Smoothness Regularization for Recommender Systems. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2):968–977, 2019.

[45] Felix Wu, Tianyi Zhang, Amauri Holanda de Souza, Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. Simplifying graph convolutional networks. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:11884–11894, 2019.

[46] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2020.

[47] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks? In *ICLR*, pages 1–15, 2019.

[48] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken Ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. *35th International Conference on Machine Learning, ICML 2018*, 12:8676–8685, 2018.

[49] Yujun Yan, Milad Hashemi, Kevin Swersky, Yaoqing Yang, and Danai Koutra. Two Sides of the Same Coin: Heterophily and Oversmoothing in Graph Convolutional Neural Networks. 2021.

[50] Lingxiao Zhao and Leman Akoglu. PairNorm: Tackling Oversmoothing in GNNs. pages 1–17, 2019.

[51] Jiong Zhu, Ryan A. Rossi, Anup Rao, Tung Mai, Nedim Lipka, Nesreen K. Ahmed, and Danai Koutra. Graph Neural Networks with Heterophily. *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, 12B:11168–11176, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] In the Appendix.
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We consider toy experiments, on which variance is insignificantly low.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [N/A]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

This appendix contains all proof of our results. We start by Linear Regression in App. A, then classification in App. B. App. C contains technical Lemmas. At their core, the proofs are a combination of chaining concentration inequalities for subgaussian variables and derivations on Gaussian distributions.

A Linear Regression

A.1 Proof of Theorem 3

Proof. Let us begin by the concentration of the optimal $\hat{\beta} = (\lambda \mathrm{Id} + Z_{\text{tr}}^\top Z_{\text{tr}}/n)^{-1} Z_{\text{tr}}^\top Y_{\text{tr}}/n$.

$$\frac{1}{n_{\text{tr}}} Z_{\text{tr}}^\top Z_{\text{tr}} = \frac{1}{n_{\text{tr}}} \sum_{i=1}^{n_{\text{tr}}} z_i z_i^\top$$

By an application of [41, Corollary 5.50], which is a concentration inequality for covariance estimates of subgaussian vectors, we get that with probability at least $1-\rho$,

$$\left\| \frac{1}{n_{\text{tr}}} \sum_i z_i z_i^\top - M^\top \Sigma M \right\| \lesssim \frac{p\sqrt{\log(1/\rho)}}{\sqrt{n}}$$

since $n_{\text{tr}} = \mathcal{O}(n)$. In particular, for n large enough, we get $\lambda_{\min}(\frac{1}{n_{\text{tr}}} \sum_i z_i z_i^\top) \geqslant \lambda_{\min}(M^\top \Sigma M)/2$.

Similarly,

$$\frac{1}{n_{\text{tr}}} Z_{\text{tr}}^\top Y_{\text{tr}} = \frac{M^\top}{n_{\text{tr}}} X_{\text{tr}}^\top X_{\text{tr}} \beta^\star$$

and using the same concentration on the covariance of X we obtain

$$\left\| \frac{1}{n_{\text{tr}}} Z_{\text{tr}}^\top Y_{\text{tr}} - M^\top \Sigma \beta^\star \right\| \lesssim \frac{\|\beta^\star\| \, p\sqrt{\log(1/\rho)}}{\sqrt{n}}$$

At the end of the day

$$\left\| \hat{\beta} - (\lambda \mathrm{Id} + M^\top \Sigma M)^{-1} M^\top \Sigma \beta^\star \right\| \lesssim \frac{\|\beta^\star\| \, p\sqrt{\log(1/\rho)}}{(\lambda + \lambda_{\min}(M^\top \Sigma M))\sqrt{n}}$$

Let us now compute the limit of the test risk:

$$\frac{1}{n_{\text{te}}} \left\| Y_{\text{te}} - Z_{\text{te}}^\top \hat{\beta} \right\|^2 = (\beta^\star)^\top \frac{1}{n_{\text{te}}} X_{\text{te}}^\top X_{\text{te}} \beta^\star - 2\hat{\beta}^\top \left(\frac{1}{n_{\text{te}}} Z_{\text{te}}^\top Y_{\text{te}} \right) + \hat{\beta}^\top \left(\frac{1}{n_{\text{te}}} Z_{\text{te}}^\top Z_{\text{te}} \right) \hat{\beta}^\top$$

By a reasoning identical to the one above on $\frac{1}{n_{\text{te}}} X_{\text{te}}^\top X_{\text{te}} \approx \Sigma$, $\frac{1}{n_{\text{te}}} Z_{\text{te}}^\top Y_{\text{te}} \approx M^\top \Sigma \beta^\star$ and $\frac{1}{n_{\text{te}}} Z_{\text{te}}^\top Z_{\text{te}} \approx M^\top \Sigma M$, and using $\left\|(\lambda \mathrm{Id} + M^\top \Sigma M)^{-1} M^\top \Sigma \beta^\star\right\| \leqslant \|\beta^\star\|$ (which can be seen using an SVD decomposition of $M\Sigma^{\frac{1}{2}}$) and $\|M\| \leqslant \|1\|$, after some computation we obtain

$$\begin{aligned} \mathcal{R}^{(0)} = (\Sigma^{\frac{1}{2}} \beta^\star)^\top \left(\mathrm{Id} - \Sigma^{\frac{1}{2}} M^\top (\lambda \mathrm{Id} + M^\top \Sigma M)^{-1} M \Sigma^{\frac{1}{2}} \right)^2 \Sigma^{\frac{1}{2}} \beta^\star \\ + \mathcal{O}\left(\frac{\|\Sigma\| \, \|\beta^\star\|^2 \, d\sqrt{\log(1/\rho)}}{(\lambda + \lambda_{\min}(M^\top \Sigma M))\sqrt{n}} \right) \end{aligned}$$

The first term is $R_{\text{reg.}}(\Sigma)$, we use a union bound over all these inequalities to conclude the proof. □

A.2 Proof of Theorem 4

We start with the proof of Lemma 1. A small reminder on subgaussian variables is given in App. C.

Proof of Lemma 1. The proof relies on chaining concentration inequalities for subgaussian variables and properties of Gaussian distributions (Lemma 5 and 7 in App. C).

Note that $\|\varphi_{\text{reg.}}(x)\|_{\Sigma^{-1}} \leqslant \frac{\left\|\Sigma^{\frac{1}{2}}(\mathrm{Id}+\Sigma)^{-\frac{1}{2}}\right\|}{\varepsilon} d(x)\|x\|_{(\mathrm{Id}+\Sigma)^{-1}} \lesssim \frac{1}{\varepsilon}$. Moreover, by Lemma 5, $\mathbb{E}W_g(x,X) = d(x)$ and $\mathbb{E}W_g(x,X)X = d(x)(\Sigma^{-1}+\mathrm{Id})^{-1}x$.

We decompose

$$
\begin{aligned}
\left\|x_i^{(1)} - \varphi_{\text{reg.}}(x_i)\right\|_{\Sigma^{-1}} &= \left\| \frac{\varepsilon \sum_j x_j + \sum_j W_g(x_i,x_j)x_j}{n\varepsilon + \sum_j W_g(x_i,x_j)} - \varphi_{\text{reg.}}(x_i) \right\|_{\Sigma^{-1}} \\
&= \left\| \frac{\varepsilon\frac{1}{n} \sum_j x_j + \frac{1}{n}\sum_j W_g(x_i,x_j)x_j}{\varepsilon + \frac{1}{n}\sum_j W_g(x_i,x_j)} - \frac{d(x)(\Sigma^{-1}+\mathrm{Id})^{-1}x}{d(x)+\varepsilon} \right\|_{\Sigma^{-1}} \\
&\leqslant \frac{1}{n}\left\|\sum_j x_j\right\|_{\Sigma^{-1}} + \left\| \frac{\frac{1}{n}\sum_j W_g(x_i,x_j)x_j - d(x)(\Sigma^{-1}+\mathrm{Id})^{-1}x}{\varepsilon + \frac{1}{n}\sum_j W_g(x_i,x_j)} \right\|_{\Sigma^{-1}} \\
&\quad + \left\|d(x)(\Sigma^{-1}+\mathrm{Id})^{-1}x\right\|_{\Sigma^{-1}} \left| \frac{1}{\varepsilon + \frac{1}{n}\sum_j W_g(x_i,x_j)} - \frac{1}{\varepsilon + d(x)} \right| \\
&\leqslant \frac{1}{n}\left\|\sum_j x_j\right\|_{\Sigma^{-1}} + \frac{1}{\varepsilon}\left\| \frac{1}{n}\sum_j W_g(x_i,x_j)x_j - d(x)(\Sigma^{-1}+\mathrm{Id})^{-1}x \right\|_{\Sigma^{-1}} \\
&\quad + \frac{1}{\varepsilon^2}\left| \frac{1}{n}\sum_j W_g(x_i,x_j) - d(x) \right|
\end{aligned}
$$

For the first term, applying Lemma 7 with $W = 1$, with probability $1-\rho$ we have

$$
\frac{1}{n}\left\|\sum_j x_j\right\| \lesssim \frac{\log n\left(\sqrt{d} + \sqrt{\log(1/\rho)}\right)}{\sqrt{n}}
$$

Similarly, since W_g is C_L Lipschitz in the first variable with respect to $\|\cdot\|_{\Sigma^{-1}}$ with $C_L \lesssim \left\|\Sigma^{\frac{1}{2}}\right\|$, applying Lemma 7 we get

$$
\left\| \frac{1}{n}\sum_j W_g(x_i,x_j)x_j - d(x)(\Sigma^{-1}+\mathrm{Id})^{-1}x \right\|_{\Sigma^{-1}} \lesssim \frac{\log n \left\|\Sigma^{\frac{1}{2}}\right\|\left(\sqrt{d} + \sqrt{\log(1/\rho)}\right)}{\sqrt{n}}
$$

and

$$
\left| \frac{1}{n}\sum_j W_g(x_i,x_j) - d(x) \right| \lesssim \frac{\sqrt{\log n}\left\|\Sigma^{\frac{1}{2}}\right\|\left(\sqrt{d} + \sqrt{\log(1/\rho)}\right)}{\sqrt{n}}
$$

which concludes the proof of the first inequality. The second is obtained by decomposing and using $\|\varphi_{\text{reg.}}(x)\|_{\Sigma^{-1}} \lesssim 1/\varepsilon$. □

We will also need the following Lemma, to bound the deviation brought by ε in the expression for $\varphi_{\text{reg.}}$.

Lemma 3. *We have*

$$
\left.\begin{aligned} \left\|\mathbb{E}\varphi_{\text{reg.}}(x)x^\top - (\Sigma^{(1)})^{\frac{1}{2}}\Sigma^{\frac{1}{2}}\right\| \\ \left\|\mathbb{E}\varphi(x)\varphi_{\text{reg.}}(x)^\top - \Sigma^{(1)}\right\| \end{aligned}\right\} \lesssim \text{poly}(e^d, |\mathrm{Id}+\Sigma|)\varepsilon^{1/5}
$$

Proof. Denote by $\mathcal{B}_r = \{x; \|x\|_{\Sigma^{-1}} \leqslant r\}$. Within this ball, since $\|x\|_{(\mathrm{Id}+\Sigma)^{-1}} \leqslant \left\|(\mathrm{Id}+\Sigma)^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}\right\| \|x\|_{\Sigma^{-1}} \leqslant \|x\|_{\Sigma^{-1}}$, we have $1/d(x) \leqslant |\mathrm{Id}+\Sigma|^{\frac{1}{2}} e^{r^2/2}$. We also recall that $\int \|x\|^2_{\Sigma^{-1}} \mathcal{N}_{0,\Sigma}(x)dx \lesssim d$ and $\int_{\mathcal{B}_r^c} \|x\|^2_{\Sigma^{-1}} \mathcal{N}_{0,\Sigma}(x)dx \lesssim 2^{d/2}e^{-r^2/4}$. Now we decompose, using $\left\|(\mathrm{Id}+\Sigma^{-1})^{-1}x\right\|_{\Sigma^{-1}} \leqslant \left\|\Sigma^{-\frac{1}{2}}(\mathrm{Id}+\Sigma^{-1})^{-1}\Sigma^{\frac{1}{2}}\right\| \|x\|_{\Sigma^{-1}}$,

$$\begin{aligned}
\mathbb{E}\left\|\varphi_{\text{reg.}}(x) - (\mathrm{Id}+\Sigma^{-1})^{-1}x\right\|^2_{\Sigma^{-1}} &= \int \left\|\frac{\varepsilon}{d(x)+\varepsilon}(\mathrm{Id}+\Sigma^{-1})^{-1}x\right\|^2_{\Sigma^{-1}} \mathcal{N}_{0,\Sigma}(x)dx \\
&\lesssim \int_{\mathcal{B}_r} \left\|\frac{\varepsilon}{d(x)+\varepsilon}(\mathrm{Id}+\Sigma^{-1})^{-1}x\right\|^2_{\Sigma^{-1}} \mathcal{N}_{0,\Sigma}(x)dx \\
&\quad + \int_{\mathcal{B}_r^c} \left\|\frac{\varepsilon}{d(x)+\varepsilon}(\mathrm{Id}+\Sigma^{-1})^{-1}x\right\|^2_{\Sigma^{-1}} \mathcal{N}_{0,\Sigma}(x)dx \\
&\lesssim \varepsilon^2 |\mathrm{Id}+\Sigma| e^{r^2} d + 2^{d/2}e^{-r^2/4} \\
&\lesssim d2^{d/2} |\mathrm{Id}+\Sigma| \varepsilon^{2/5}
\end{aligned}$$

Where the last line is obtained by choosing $r = \sqrt{(8/5)\log(1/\varepsilon)}$. Then we use that for two random variables X and Y, $\|\mathbb{E}X - \mathbb{E}Y\| \leqslant \sqrt{\mathbb{E}\|X-Y\|^2}$, and $\left\|\mathbb{E}XY^\top - \mathbb{E}YY^\top\right\| \leqslant \sqrt{\mathbb{E}\|Y\|^2}\sqrt{\mathbb{E}\|X-Y\|^2}$, and $\left\|\mathbb{E}XX^\top - \mathbb{E}YY^\top\right\| \leqslant (\sqrt{\mathbb{E}\|X\|^2}+\sqrt{\mathbb{E}\|Y\|^2})\sqrt{\mathbb{E}\|X-Y\|^2}$, to conclude. □

We are now ready to show Theorem 4

Proof of Theorem 4. We proceed in two steps. First, we use Lemma 1 to show that we can replace the $x^{(1)}$ by $\varphi_{\text{reg.}}(x_i)$ in the computation of the risk. Second, we use Lemma 3 to concentrate the $\varphi_{\text{reg.}}(x_i)$ around their new expectations. We define $\hat{\beta}^\varphi$ and $\mathcal{R}^\varphi$ by replacing $Z^{(k)}$ with $Z^\varphi = X^\varphi M$ where the rows of X^φ are the $\varphi_{\text{reg.}}(x_i)$, in (3) and (4).

Since $\varphi_{\text{reg.}}(x)$ is bounded, it is a subgaussian vector. We can therefore apply the same reasoning as in the proof of Theorem 6 and concentrate $(Z^\varphi_{\text{tr}})^\top Z^\varphi_{\text{tr}}/n_{\text{tr}}$. Using Lemma 4, for ε small enough, and n large enough, it is almost equal to $M^\top\Sigma^{(1)}M$, and thus $\lambda_{\min}((Z^\varphi_{\text{tr}})^\top Z^\varphi_{\text{tr}}/n_{\text{tr}}) \gtrsim \lambda_{\min}(M^\top\Sigma^{(1)}M)$. Finally, using Lemma 2, for n large enough $\lambda_{\min}((Z^{(1)}_{\text{tr}})^\top Z^{(1)}_{\text{tr}}/n_{\text{tr}}) \geqslant \lambda_{\min}((Z^\varphi_{\text{tr}})^\top Z^\varphi_{\text{tr}}/n_{\text{tr}})/2$.

Since $\|\varphi_{\text{reg.}}(x)\|_{\Sigma^{-1}} \lesssim 1/\varepsilon$, we bound

$$\begin{aligned}
\left\|\hat{\beta} - \hat{\beta}^\varphi\right\| &= \left\|(\lambda\mathrm{Id} + (Z^{(1)}_{\text{tr}})^\top Z^{(1)}_{\text{tr}}/n)^{-1}(Z^{(1)}_{\text{tr}})^\top Y_{\text{tr}}/n_{\text{tr}} - (\lambda\mathrm{Id} + (Z^\varphi_{\text{tr}})^\top Z^\varphi_{\text{tr}}/n)^{-1}(Z^\varphi_{\text{tr}})^\top Y_{\text{tr}}/n_{\text{tr}}\right\| \\
&\leqslant \left\|((\lambda\mathrm{Id} + (Z^{(1)}_{\text{tr}})^\top Z^{(1)}_{\text{tr}}/n_{\text{tr}})^{-1} - (\lambda\mathrm{Id} + (Z^\varphi_{\text{tr}})^\top Z^\varphi_{\text{tr}}/n_{\text{tr}})^{-1})(Z^\varphi_{\text{tr}})^\top Y_{\text{tr}}/n_{\text{tr}}\right\| \\
&\quad + \left\|(\lambda\mathrm{Id} + (Z^{(1)}_{\text{tr}})^\top Z^{(1)}_{\text{tr}}/n_{\text{tr}})^{-1}((Z^{(1)}_{\text{tr}})^\top Y_{\text{tr}}/n_{\text{tr}} - (Z^\varphi_{\text{tr}})^\top Y_{\text{tr}}/n_{\text{tr}})\right\| \\
&\leqslant \frac{\left\|M^\top\Sigma^{\frac{1}{2}}\right\|^2}{\varepsilon(\lambda+\lambda_{\min}(M^\top\Sigma^{(1)}M))^2} \sup_i \left\|\Sigma^{-\frac{1}{2}}\left(x_i^{(1)}(x_i^{(1)})^\top - \varphi(x_i)\varphi(x_i)^\top\right)\Sigma^{-\frac{1}{2}}\right\| \\
&\quad + \frac{\left\|M^\top\Sigma^{\frac{1}{2}}\right\|}{\lambda+\lambda_{\min}(M^\top\Sigma^{(1)}M)} \sup_i \left\|x_i^{(1)} - \varphi(x_i)\right\|_{\Sigma^{-1}} \\
&\lesssim \frac{\text{poly}(\|\Sigma\|, \varepsilon^{-1})(\sqrt{d}+\sqrt{\log(1/\rho)})}{(\lambda+\lambda_{\min}(M^\top\Sigma^{(1)}M))^2\sqrt{n}}
\end{aligned}$$

Using the same bounds on $\frac{1}{n_{\text{te}}}(Z^{(1)}_{\text{te}})^\top Y_{\text{te}}$ and $\frac{1}{n_{\text{te}}}(Z^{(1)}_{\text{te}})^\top Z^{(1)}_{\text{te}}$, we get

$$\left|\mathcal{R}^{(1)} - \mathcal{R}^\varphi\right| \lesssim \frac{\text{poly}(\|\Sigma\|, \varepsilon^{-1})(\sqrt{d}+\sqrt{\log(1/\rho)})}{(\lambda+\lambda_{\min}(M^\top\Sigma^{(1)}M))^2\sqrt{n}} \tag{22}$$

Finally, we apply the same reasoning as in the proof of Theorem 3, we obtain

$$\begin{aligned}\mathcal{R}^{\varphi} &= \|\beta^\star\|_\Sigma^2 - 2(M^\top \Sigma_{\varphi,x}\beta^\star)^\top(\lambda \mathrm{Id} + M^\top\Sigma_\varphi M)^{-1}M^\top\Sigma_{\varphi,x}\beta^\star \\ &\quad + (M^\top \Sigma_{\varphi,x}\beta^\star)^\top(\lambda \mathrm{Id} + M^\top\Sigma_\varphi M)^{-1}\Sigma_\varphi(\lambda \mathrm{Id} + M^\top\Sigma_\varphi M)^{-1}M^\top\Sigma_{\varphi,x}\beta^\star \\ &\quad + \mathrm{poly}(\varepsilon^{-1}, \|\Sigma\|, \|\beta^\star\|)\frac{\log n(\sqrt{d} + \sqrt{\log(1/\rho)})}{(\lambda + \lambda_{\min}(M^\top\Sigma^{(1)}M))\sqrt{n}}\end{aligned}$$

where $\Sigma_\varphi = \mathbb{E}\varphi(x)\varphi(x)^\top$ and $\Sigma_{\varphi,x} = \mathbb{E}\varphi(x)x^\top$. We use Lemma 3 and a union bound over all these inequalities to conclude the proof. □

B Classification of Gaussian mixtures

B.1 Proof of Theorem 6

Proof. The proof is similar to that of Theorem 3. We begin by the concentration of the optimal $\hat{\beta} = (\lambda \mathrm{Id} + Z_{\text{tr}}^\top Z_{\text{tr}}/n_{\text{tr}})^{-1}Z_{\text{tr}}^\top Y_{\text{tr}}/n_{\text{tr}}$. We denote by $I_1, I_{-1} \subset \{1, \dots, n_{\text{tr}}\}$ the indices of the x_i respectively from the first and second community, of size n_1 and n_{-1}. We have

$$\begin{aligned}\frac{1}{n_{\text{tr}}}Z_{\text{tr}}^\top Z_{\text{tr}} &= \frac{1}{n_{\text{tr}}}\sum_i z_i z_i^\top \\ &= \frac{n_1}{n_{\text{tr}}}\frac{1}{n_1}\sum_{I_1} z_i z_i^\top + \frac{n_{-1}}{n}\frac{1}{n_{-1}}\sum_{I_{-1}} z_i z_i^\top\end{aligned}$$

Since the communities are balanced, by a simple application of Hoeffding's inequality, with probability at least $1-\rho$ we have $n_1/n_{\text{tr}} = 1/2 + \mathcal{O}\left(\sqrt{\log(1/\rho)/n}\right)$. Then, as in the proof of Theorem 3 by an application of [41, Corollary 5.50], with probability at least $1-\rho$,

$$\left\| \frac{1}{n_1}\sum_{I_1} z_i z_i^\top - \mathbb{E}_{\mathcal{N}_\nu} zz^\top \right\| \lesssim \frac{p\sqrt{\log(1/\rho)}}{\sqrt{n}}$$

and $\mathbb{E}_{\mathcal{N}_\mu} zz^\top = \nu\nu^\top + \mathrm{Id}$. We apply the same reasoning for I_{-1}, and we obtain

$$\left\| \frac{1}{n_{\text{tr}}}Z_{\text{tr}}^\top Z_{\text{tr}} - (\mathrm{Id} + \nu\nu^\top) \right\| \leqslant \frac{p\sqrt{\log(1/\rho)}}{\sqrt{n}}$$

In particular, for n large enough, $\lambda_{\min}(\frac{1}{n_{\text{tr}}}Z_{\text{tr}}^\top Z_{\text{tr}}) \geqslant 1/2$.

Similarly,

$$\frac{1}{n_{\text{tr}}}Z_{\text{tr}}^\top Y_{\text{tr}} = M^\top\left(\frac{n_1}{n_{\text{tr}}}\frac{1}{n_1}\sum_{I_1} x_i - \frac{n_{-1}}{n_{\text{tr}}}\frac{1}{n_{-1}}\sum_{I_{-1}} x_i\right)$$

Using the fact that $\|x\| = \sup_{\|u\|\leqslant 1} u^\top x$ and for such a u the variable $u^\top(z-\nu)$ is unit Gaussian, applying Lemma 6 we get that with probability $1-\rho$

$$\left\| \frac{1}{n_1}\sum_{I_1} z_i - \nu \right\| \lesssim \frac{\sqrt{p} + \sqrt{\log(1/\rho)}}{\sqrt{n}}$$

and similarly for I_{-1}, and therefore

$$\left\| \frac{1}{n_{\text{tr}}}Z_{\text{tr}}^\top Y_{\text{tr}} - \nu \right\| \leqslant \frac{\sqrt{p} + \sqrt{\log(1/\rho)}}{\sqrt{n}}$$

At the end of the day

$$\begin{aligned}\left\|\hat{\beta} - ((\lambda+1)\mathrm{Id} + \nu\nu^\top)^{-1}\nu\right\| &\leqslant \left\|(\lambda\mathrm{Id} + Z_{\text{tr}}^\top Z_{\text{tr}}/n)^{-1}(Z_{\text{tr}}^\top Y_{\text{tr}}/n - \nu)\right\| \\ &\quad + \left\|((\lambda\mathrm{Id} + Z_{\text{tr}}^\top Z_{\text{tr}}/n_{\text{tr}})^{-1} - ((\lambda+1)\mathrm{Id} + \nu\nu^\top)^{-1})\nu\right\| \\ &\lesssim \frac{\sqrt{p} + \sqrt{\log(1/\rho)}}{\sqrt{n}} + \frac{\|\nu\|}{\lambda}\left\|Z_{\text{tr}}^\top Z_{\text{tr}}/n - (\mathrm{Id} + \nu\nu^\top)\right\| \\ &\lesssim \frac{\|\nu\| p\sqrt{\log(1/\rho)}}{\sqrt{n}}\end{aligned}$$

Moreover, ν is an eigenvector for $(\lambda+1)\mathrm{Id} + \nu\nu^\top$ so the limit is actually:

$$\hat{\beta}_{lim} = ((\lambda+1)\mathrm{Id} + \nu\nu^\top)^{-1}\nu = \frac{\nu}{1+\lambda+\|\nu\|^2}$$

Let us now compute the limit of the test risk:

$$\frac{1}{n_{\text{te}}}\left\|Y_{\text{te}} - Z_{\text{te}}^\top\hat{\beta}\right\|^2 = 1 - 2\hat{\beta}^\top\left(\frac{1}{n_{\text{te}}}Z_{\text{te}}^\top Y_{\text{te}}\right) + \hat{\beta}^\top\left(\frac{1}{n_{\text{te}}}Z_{\text{te}}^\top Z_{\text{te}}\right)\hat{\beta}^\top$$

By a reasoning identical to the one above on $\frac{1}{n_{\text{te}}}Z_{\text{te}}^\top Y_{\text{te}} \approx \nu$ and $\frac{1}{n_{\text{te}}}Z_{\text{te}}^\top Z_{\text{te}} \approx \mathrm{Id} + \nu\nu^\top$, and using $\left\|\hat{\beta}_{lim}\right\| \leqslant \|\nu\|$ and $\left\|\mathrm{Id} + \nu\nu^\top\right\| = 1 + \|\nu\|^2$, we obtain

$$\begin{aligned}\frac{1}{n_{\text{te}}}\left\|Y_{\text{te}} - Z_{\text{te}}^\top\hat{\beta}\right\|^2 &= 1 - 2(\hat{\beta}_{lim})^\top\nu + (\hat{\beta}_{lim})^\top\left(\mathrm{Id} + \nu\nu^\top\right)\hat{\beta}_{lim} + \mathcal{O}\left(\frac{\|\nu\|^4 p\sqrt{\log(1/\rho)}}{\sqrt{n}}\right) \\ &= 1 - 2\frac{\|\nu\|^2}{1+\lambda+\|\nu\|^2} + \frac{\|\nu\|^2 + \|\nu\|^4}{(1+\lambda+\|\nu\|^2)^2} + \mathcal{O}\left(\frac{\|\nu\|^4 p\sqrt{\log(1/\rho)}}{\sqrt{n}}\right) \\ &= \frac{(1+\lambda)^2 + \|\nu\|^2}{(1+\lambda+\|\nu\|^2)^2} + \mathcal{O}\left(\frac{\|\nu\|^4 p\sqrt{\log(1/\rho)}}{\sqrt{n}}\right)\end{aligned}$$

We use a union bound over all these inequalities to conclude the proof. □

B.2 Proof of Theorem 7

We start with the proof of Lemma 2.

Proof of Lemma 2. The proof is similar to that of Lemma 1. Here we denote by $I_1, I_{-1} \subset \{1,\dots,n\}$ the indices of the x_i from the first and second community, of size n_1 and n_{-1}, from the whole sample set. Again, since the communities are balanced, with probability $1-\rho$ we have $|I_1|/n \approx \frac{1}{2} + \mathcal{O}\left(\sqrt{\log(1/\rho)/n}\right)$.

We decompose

$$\begin{aligned}\|\tilde{x}_i - \varphi_{\text{cl.}}(x_i)\| &= \left\|\tfrac{\varepsilon\sum_j x_j + \sum_j W_g(x_i,x_j)x_j}{n\varepsilon + \sum_j W_g(x_i,x_j)} - \varphi_{\text{cl.}}(x_i)\right\| \\ &= \left\|\tfrac{\varepsilon\frac{1}{n}\sum_j x_j + \frac{1}{n}\sum_j W_g(x_i,x_j)x_j}{\varepsilon + \frac{1}{n}\sum_j W_g(x_i,x_j)} - \tfrac{\frac{1}{2}\left(d_\mu(x_i)\left(\frac{x_i+\mu}{2}\right) + d_{-\mu}(x_i)\left(\frac{x_i-\mu}{2}\right)\right)}{\varepsilon + \frac{1}{2}(d_\mu(x_i) + d_{-\mu}(x_i))}\right\| \\ &\leqslant \tfrac{1}{n}\left\|\sum_j x_j\right\| + \left\|\tfrac{\frac{1}{n}\sum_j W_g(x_i,x_j)x_j}{\varepsilon + \frac{1}{n}\sum_j W_g(x_i,x_j)} - \tfrac{\frac{1}{2}\left(d_\mu(x_i)\left(\frac{x_i+\mu}{2}\right) + d_{-\mu}(x_i)\left(\frac{x_i-\mu}{2}\right)\right)}{\varepsilon + \frac{1}{n}\sum_j W_g(x_i,x_j)}\right\| \\ &\quad + \left\|\tfrac{\frac{1}{2}\left(d_\mu(x_i)\left(\frac{x_i+\mu}{2}\right) + d_{-\mu}(x_i)\left(\frac{x_i-\mu}{2}\right)\right)}{\varepsilon + \frac{1}{n}\sum_j W_g(x_i,x_j)} - \tfrac{\frac{1}{2}\left(d_\mu(x_i)\left(\frac{x_i+\mu}{2}\right) + d_{-\mu}(x_i)\left(\frac{x_i-\mu}{2}\right)\right)}{\varepsilon + \frac{1}{2}(d_\mu(x_i) + d_{-\mu}(x_i))}\right\| \\ &\leqslant \tfrac{1}{n}\left\|\sum_j x_j\right\| + \varepsilon^{-1}\left\|\tfrac{1}{n}\sum_j W_g(x_i,x_j)x_j - \tfrac{1}{2}\left(d_\mu(x_i)\left(\tfrac{x_i+\mu}{2}\right) + d_{-\mu}(x_i)\left(\tfrac{x_i-\mu}{2}\right)\right)\right\| \\ &\quad + \tfrac{1}{4\varepsilon^2}\left|\tfrac{1}{n}\sum_j W_g(x_i,x_j) - \tfrac{d_\mu(x_i) + d_{-\mu}(x_i)}{2}\right|\left\|d_\mu(x_i)\left(\tfrac{x_i+\mu}{2}\right) + d_{-\mu}(x_i)\left(\tfrac{x_i-\mu}{2}\right)\right\|\end{aligned}$$

For the first term, with probability $1-\rho$ we have

$$\begin{aligned}\frac{1}{n}\left\|\sum_j x_j\right\| &= \frac{n_1}{n}\frac{1}{n_1}\left\|\sum_{j\in I_1} x_j - \mu\right\| + \frac{n_{-1}}{n}\frac{1}{n_{-1}}\left\|\sum_{j\in I_{-1}} x_j + \mu\right\| \\ &\leqslant \left\|\frac{1}{n_1}\sum_{j\in I_1} x_j - \mu\right\| + \left\|\frac{1}{n_{-1}}\sum_{j\in I_{-1}} x_j + \mu\right\| \\ &\lesssim \frac{\log n\left(\sqrt{d}+\sqrt{\log(1/\rho)}\right)}{\sqrt{n}}\end{aligned}$$

Where we have used Lemma 7 with $W=1$ and $n_1 = n - n_{-1} \approx n/2$ for the last line.

For the second term, similarly

$$\begin{aligned}&\left\|\frac{1}{n}\sum_j W_g(x_i,x_j)x_j - \frac{1}{2}\left(d_\mu(x_i)\left(\tfrac{x_i+\mu}{2}\right) + d_{-\mu}(x_i)\left(\tfrac{x_i-\mu}{2}\right)\right)\right\| \\ &\lesssim \frac{1}{2}\left\|\frac{1}{n_1}\sum_{j\in I_1} W_g(x_i,x_j)x_j - d_\mu(x_i)\left(\frac{x_i+\mu}{2}\right)\right\| \\ &\quad + \frac{1}{2}\left\|\frac{1}{n_{-1}}\sum_{j\in I_{-1}} W_g(x_i,x_j)x_j - d_{-\mu}(x_i)\left(\frac{x_i-\mu}{2}\right)\right\| + \sqrt{\log(1/\rho)/n}\end{aligned}$$

Using Lemma 5, we have $\mathbb{E}_{\mathcal{N}} W_g(x,X)X = d_0(x)\frac{x}{2}$. Hence, using the Lipschitz properties of the Gaussian kernel, and since we can center

$$\frac{1}{n_1}\sum_{j\in I_1} W_g(x_i,x_j)x_j - d_\mu(x_i)\left(\frac{x_i+\mu}{2}\right) = \frac{1}{n_1}\sum_{j\in I_1} W_g(x_i-\mu,x_j-\mu)(x_j-\mu) - d_\mu(x_i-\mu)\left(\frac{x_i-\mu}{2}\right)$$

and $x_j - \mu \sim \mathcal{N}$, using Lemma 7 we obtain

$$\left\|\frac{1}{n_1}\sum_{j\in I_1} W_g(x_i,x_j)x_j - d_\mu(x_i)\left(\frac{x_i+\mu}{2}\right)\right\| \lesssim \frac{\log n\left(\sqrt{d}+\sqrt{\log(1/\rho)}\right)}{\sqrt{n}}$$

We proceed similarly for the second term, and again with the first part of Lemma 7 to obtain

$$\left|\tfrac{1}{n}\textstyle\sum_j W_g(x_i,x_j) - \frac{d_\mu(x_i)+d_{-\mu}(x_i)}{2}\right| \lesssim \frac{\sqrt{\log n}\left(\sqrt{d}+\sqrt{\log(1/\rho)}\right)}{\sqrt{n}}$$

which gives us the first result. The second is obtained by simply decomposing and using $\|\varphi_{\text{cl.}}(x)\| \lesssim 1/\varepsilon$. □

We will need the following Lemma, similar to Lemma 3.

Lemma 4. *Let $x \sim \mathcal{N}_\mu$. We have*

$$\begin{aligned}\|\mathbb{E}\varphi_{\text{cl.}}(x) - \mu\| &\lesssim \sqrt{d}2^{d/2}\|\mu\|\varepsilon^{1/4} + \frac{\|\mu\|}{\varepsilon^3}e^{-\|\mu\|^2/4} \\ \left\|\mathbb{E}\varphi_{\text{cl.}}(x)\varphi_{\text{cl.}}(x)^\top - (\mu\mu^\top + \mathrm{Id}/4)\right\| &\lesssim d2^{d/2}\|\mu\|^2\varepsilon^{1/4} + \frac{\|\mu\|^2\sqrt{d}}{\varepsilon^3}e^{-\|\mu\|^2/4}\end{aligned}$$

Proof. Denote by $\mathcal{B}_{\mu,r}$ a ball of radius r around μ. Within this ball, $d_\mu(x) \geqslant 2^{-d/2}e^{-r^2/4}$, while $d_{-\mu}(x) \leqslant 2^{-d/2}e^{-\|\mu\|^2/4}$. We also recall that $\int_{\mathcal{B}^c_{\mu,r}} \mathcal{N}_\mu \leqslant e^{-r^2/2}$ and $\int_{\mathcal{B}^c_{\mu,r}} \|x-\mu\|^2 \mathcal{N}_\mu \lesssim$

$2^{d/2}e^{-r^2/4}$. Now we decompose

$$\begin{aligned}
E_{\mathcal{N}_\mu}\left\|\varphi_{\text{cl.}}(x)-\frac{x+\mu}{2}\right\|^2 &= \int\left\|\frac{d_\mu(x)\left(\frac{x+\mu}{2}\right)+d_{-\mu}(x)\left(\frac{x-\mu}{2}\right)}{2\varepsilon+d_\mu(x)+d_{-\mu}(x)}-\frac{x+\mu}{2}\right\|^2\mathcal{N}_\mu(x)dx\\
&= \int\left\|\frac{\varepsilon(x+\mu)-d_{-\mu}(x)\mu}{2\varepsilon+d_\mu(x)+d_{-\mu}(x)}\right\|^2\mathcal{N}_\mu(x)dx\\
&\lesssim \int_{\mathcal{B}_{\mu,r}}\left\|\frac{\varepsilon(x+\mu)-d_{-\mu}(x)\mu}{2\varepsilon+d_\mu(x)+d_{-\mu}(x)}\right\|^2\mathcal{N}_\mu(x)dx\\
&\quad+\int_{\mathcal{B}_{-\mu,r}}\left\|\frac{\varepsilon(x+\mu)-d_{-\mu}(x)\mu}{2\varepsilon+d_\mu(x)+d_{-\mu}(x)}\right\|^2\mathcal{N}_\mu(x)dx\\
&\quad+\int_{(\mathcal{B}_{\mu,r}\cup\mathcal{B}_{-\mu,r})^c}\left\|\frac{\varepsilon(x+\mu)-d_{-\mu}(x)\mu}{2\varepsilon+d_\mu(x)+d_{-\mu}(x)}\right\|^2\mathcal{N}_\mu(x)dx\\
&\lesssim \varepsilon^2(\|\mu\|^2+d)2^d e^{r^2/2}+\|\mu\|^2 e^{-\|\mu\|^2/2}e^{r^2/2}\\
&\quad+(\varepsilon^2 2^d\|\mu\|e^{r^2/4}+\|\mu\|^2)e^{-\|\mu\|^2/4}\\
&\quad+2^d e^{-r^2/4}\|\mu\|^2+\|\mu\|^2 2^d e^{-r^2/2}e^{-r^2/2}/\varepsilon^2\\
&\lesssim d2^d\|\mu\|^2\sqrt{\varepsilon}+\frac{\|\mu\|^2}{\varepsilon^{3/2}}e^{-\|\mu\|^2/2}
\end{aligned}$$

Where the last line is obtained by choosing $r=\sqrt{3\log(1/\varepsilon)}$. Then we use that for two random variables X and Y, $\|\mathbb{E}X-\mathbb{E}Y\|\leqslant\sqrt{\mathbb{E}\|X-Y\|^2}$, and $\left\|\mathbb{E}XX^\top-\mathbb{E}YY^\top\right\|\leqslant(\sqrt{\mathbb{E}\|X\|^2}+\sqrt{\mathbb{E}\|Y\|^2})\sqrt{\mathbb{E}\|X-Y\|^2}$ to conclude. □

We are now ready to prove Theorem 7.

Proof of Theorem 7. We proceed as in the proof of Theorem 4. We define $\hat{\beta}^\varphi$ and $\mathcal{R}^\varphi$ by replacing $Z^{(k)}$ with $Z^\varphi=X^\varphi M$ where the rows of X^φ are the $\varphi_{\text{cl.}}(x_i)$.

Since $\|\varphi_{\text{cl.}}(x)\|\leqslant\max\left(\frac{\|x+\mu\|}{2},\frac{\|x-\mu\|}{2}\right)$, $\varphi_{\text{cl.}}(x)$ is a subgaussian vector with $\left\|u^\top\varphi_{\text{cl.}}(x)\right\|_{\psi_2}\lesssim\left\|u^\top x\right\|_{\psi_2}\lesssim 1$. We can therefore apply the same reasoning as in the proof of Theorem 6 and concentrate $(Z^\varphi_{\text{tr}})^\top Z^\varphi_{\text{tr}}/n_{\text{tr}}$. Using Lemma 4, for ε small enough, and $\|\mu\|$ and n large enough, it is almost $\mathrm{Id}/4+\nu\nu^\top$, and thus $\lambda_{\min}((Z^\varphi_{\text{tr}})^\top Z^\varphi_{\text{tr}}/n_{\text{tr}})\gtrsim 1$. Finally, using Lemma 2, for n_{tr} large enough $\lambda_{\min}((Z^{(1)}_{\text{tr}})^\top Z^{(1)}_{\text{tr}}/n_{\text{tr}})\geqslant\lambda_{\min}((Z^\varphi_{\text{tr}})^\top Z^\varphi_{\text{tr}}/n_{\text{tr}})/2\gtrsim 1$.

Since $\|\varphi_{\text{cl.}}(x)\|\leqslant 1/\varepsilon$, using Lemma 2 we bound

$$\begin{aligned}
\left\|\hat{\beta}-\hat{\beta}^\varphi\right\| &= \left\|(\lambda\mathrm{Id}+(Z^{(1)}_{\text{tr}})^\top Z^{(1)}_{\text{tr}}/n_{\text{tr}})^{-1}(Z^{(1)}_{\text{tr}})^\top Y_{\text{tr}}/n_{\text{tr}}-(\lambda\mathrm{Id}+(Z^\varphi_{\text{tr}})^\top Z^\varphi_{\text{tr}}/n_{\text{tr}})^{-1}(Z^\varphi_{\text{tr}})^\top Y_{\text{tr}}/n_{\text{tr}}\right\|\\
&\leqslant \left\|((\lambda\mathrm{Id}+(Z^{(1)}_{\text{tr}})^\top Z^{(1)}_{\text{tr}}/n_{\text{tr}})^{-1}-(\lambda\mathrm{Id}+(Z^\varphi_{\text{tr}})^\top Z^\varphi_{\text{tr}}/n_{\text{tr}})^{-1})(Z^\varphi_{\text{tr}})^\top Y_{\text{tr}}/n_{\text{tr}}\right\|\\
&\quad+\left\|(\lambda\mathrm{Id}+(Z^{(1)}_{\text{tr}})^\top Z^{(1)}_{\text{tr}}/n_{\text{tr}})^{-1}((Z^{(1)}_{\text{tr}})^\top Y_{\text{tr}}/n_{\text{tr}}-(Z^\varphi_{\text{tr}})^\top Y_{\text{tr}}/n_{\text{tr}})\right\|\\
&\leqslant \frac{1}{\varepsilon}\sup_i\left\|x_i^{(1)}(x_i^{(1)})^\top-\varphi(x_i)\varphi(x_i)^\top\right\|+\sup_i\left\|x_i^{(1)}-\varphi(x_i)\right\|\\
&\lesssim \frac{\text{poly}(1/\varepsilon)\log n(\sqrt{d}+\sqrt{\log(1/\rho)})}{\sqrt{n}}
\end{aligned}$$

Using the same bounds on $\frac{1}{n_{\text{te}}}(Z^{(1)}_{\text{te}})^\top Y_{\text{te}}$ and $\frac{1}{n_{\text{te}}}(Z^{(1)}_{\text{te}})^\top Z^{(1)}_{\text{te}}$, and using $\left\|\hat{\beta}^\varphi\right\|\lesssim\varepsilon^{-1}$, we get

$$\left|\mathcal{R}^{(1)}-\mathcal{R}^\varphi\right|\lesssim\frac{\text{poly}(1/\varepsilon)\log n(\sqrt{d}+\sqrt{\log(1/\rho)})}{\sqrt{n}} \tag{23}$$

Then, we apply the same reasoning as in the proof of Theorem 6, we obtain

$$\mathcal{R}^{\varphi} = 1 - 2(\mathbb{E}yz^{\varphi})^{\top}\beta_{lim}^{\varphi} + (\beta_{lim}^{\varphi})^{\top}\mathbb{E}z^{\varphi}(z^{\varphi})^{\top}\beta_{lim}^{\varphi} + \mathcal{O}\left(\text{poly}(\frac{1}{\varepsilon}, \frac{1}{\lambda})\frac{\log n(\sqrt{d} + \sqrt{\log(1/\rho)})}{\sqrt{n}}\right)$$

where $\beta_{lim}^{\varphi} = (\lambda \text{Id} + \mathbb{E}z^{\varphi}(z^{\varphi})^{\top})^{-1}(\mathbb{E}yz^{\varphi})$.

Finally, using Lemma 4, and computations similar to 6, we obtain

$$\mathcal{R}^{\varphi} = R_{\text{cl.}}(1/4) + \mathcal{O}\left(\text{poly}(\|\mu\|, e^{d})\left(\varepsilon^{1/4} + \frac{1}{\varepsilon^{3}}e^{-\|\mu\|^{2}/4}\right)\right)$$

which concludes the proof. □

C Technical Lemmas

This section gather some technical Lemmas used throughout the proofs. We start by some derivations on Gaussian distributions, then details the chaining concentration inequalities used in this work.

C.1 Properties of Gaussians

Lemma 5 (Gaussian integral)**.** *Let $W(x, y) = e^{-\frac{1}{2}\|x-y\|_{\Sigma_W^{-1}}}$ be the Gaussian kernel with covariance Σ_W. We have*

$$d(x) := \int W(x, y)\mathcal{N}_{\mu,\Sigma}(y)dy = \frac{|\Sigma_W|^{\frac{1}{2}}}{|\Sigma_W + \Sigma|^{\frac{1}{2}}}e^{-\frac{1}{2}\|x-\mu\|^{2}_{(\Sigma_W+\Sigma)^{-1}}} \tag{24}$$

$$\mathcal{L}(x) := \int W(x, y)y\mathcal{N}_{\mu,\Sigma}(y)dy = d(x)(\Sigma_W^{-1} + \Sigma^{-1})^{-1}(\Sigma_W^{-1}x + \Sigma^{-1}\mu) \tag{25}$$

Proof. We have the following when $n \to \infty$.

$$\begin{aligned} d(x) = \int W(x, y)\mathcal{N}_{\mu,\Sigma}(y)dy &= \int e^{-\frac{1}{2}\|x-y\|^{2}_{\Sigma_W^{-1}}}\mathcal{N}_{\mu,\Sigma}(y)dy \\ &= (2\pi)^{d/2}|\Sigma_W|^{\frac{1}{2}}\int \mathcal{N}_{0,\Sigma_W}(x - y)\mathcal{N}_{\mu,\Sigma}(y)dy \\ &= (2\pi)^{d/2}|\Sigma_W|^{\frac{1}{2}}\mathcal{N}_{0,\Sigma_W} \star \mathcal{N}_{\mu,\Sigma}(x) \\ &= (2\pi)^{d/2}|\Sigma_W|^{\frac{1}{2}}\mathcal{N}_{\mu,\Sigma_W+\Sigma}(x) = \frac{|\Sigma_W|^{\frac{1}{2}}}{|\Sigma_W + \Sigma|^{\frac{1}{2}}}e^{-\frac{1}{2}\|x-\mu\|^{2}_{(\Sigma_W+\Sigma)^{-1}}} \end{aligned}$$

Since the convolution of two gaussians is a Gaussian. And

$$\begin{aligned} \mathcal{L}(x) &= \int W(x, y)y\mathcal{N}_{\mu,\Sigma}(y)dy \\ &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{\frac{1}{2}}}\int ye^{-\frac{1}{2}\|y-x\|^{2}_{\Sigma_W^{-1}} - \frac{1}{2}\|y-\mu\|^{2}_{\Sigma^{-1}}}dy \\ &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{\frac{1}{2}}}\int -(\Sigma_W^{-1}(y - x) + \Sigma^{-1}(y - \mu))e^{-\frac{1}{2}\|y-x\|^{2}_{\Sigma_W^{-1}} - \frac{1}{2}\|y-\mu\|^{2}_{\Sigma^{-1}}}dy \\ &\quad + (\text{Id} + \Sigma_W^{-1} + \Sigma^{-1})\mathcal{L}(x) - (\Sigma_W^{-1}x + \Sigma^{-1}\mu)d(x) \\ &= d(x)(\Sigma_W^{-1} + \Sigma^{-1})^{-1}(\Sigma_W^{-1}x + \Sigma^{-1}\mu) \end{aligned}$$

using that the first term in the sum is 0 since it is the integral of a derivative. □

C.2 Chaining and subgaussian variables

A random variable X is said to be *subgaussian* if

$$\|X\|_{\psi_2} := \inf\{t > 0; \mathbb{E}e^{X^2/t^2} \leqslant 2\} < \infty \tag{26}$$

A good reference on subgaussian random variables is [42, Chap. 2]. For a bounded random variable X and subgaussian Y, we have immediately from the definition

$$\|XY\|_{\psi_2} \leqslant \|X\|_\infty \|Y\|_{\psi_2} \tag{27}$$

Lemma 6 (Chaining on non-normalized kernels)**.** *Consider $x_i \sim P \in \mathcal{P}(\mathbb{R}^m)$, a ball $\mathcal{B}_r \subset \mathbb{R}^n$ with respect to a metric d, and a bivariate function $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ that satisfies:*

1. *For all $z \in \mathcal{B}_r$, $F(z, X)$ is subgaussian with norm $\|F(z, X)\|_{\psi_2} \leqslant C$*
2. *For all $z, z' \in \mathcal{B}_r$, $\|F(z, X) - F(z', X)\|_{\psi_2} \leqslant C_L d(z, z')$*

Then, with probability at least $1 - \rho$,

$$\sup_{z\in\mathcal{B}_r} \left| \frac{1}{n}\sum_i F(z, x_i) - \int F(z, x)dP(x) \right|_\infty \lesssim \frac{rC_L\left(\sqrt{d} + \sqrt{\log(1/\rho)}\right) + C\sqrt{\log(1/\rho)}}{\sqrt{n}}$$

Proof. Define

$$Y_z = \frac{1}{n}\sum_i F(z, x_i) - \int F(z, x)dP(x)$$

By [42, Lemma 2.6.8], we have $\|Y_z\|_{\psi_2} \leqslant C$. Hence we can apply a generalized Hoeffding's inequality for subgaussian variables: with probability at least $1 - \rho$,

$$|Y_{z_0}| \lesssim \frac{C\sqrt{\log(1/\rho)}}{\sqrt{n}}$$

For any z_0, we have

$$\sup_{z\in\mathcal{B}_r} |Y_z| \leqslant \sup_{z,z'\in\mathcal{B}_r} |Y_z - Y_{z'}| + |Y_{z_0}|$$

The second term is bounded by the inequality above. For the first term, we are going to use Dudley's inequality "tail bound" version [42, Thm 8.1.6]. We first need to check the sub-gaussian increments of the process Y_z. For any $z, z' \in \mathcal{B}_r$, we have

$$\begin{aligned}
\|Y_z - Y_{z'}\|_{\psi_2} &\lesssim \frac{1}{n}\left(\sum_{i=1}^n \|(F(z, x_i) - F(z', x_i)) - \mathbb{E}((F(z, X) - F(z', X)))\|_{\psi_2}^2\right)^{\frac{1}{2}} \\
&\lesssim \frac{1}{n}\left(\sum_{i=1}^n \|(F(z, x_i) - F(z', x_i))\|_{\psi_2}^2\right)^{\frac{1}{2}} \\
&\leqslant \frac{C_L}{\sqrt{n}} d(z, z')
\end{aligned}$$

where we have used, from [42], Prop. 2.6.1 for the first line, Lemma 2.6.8 for the second, and the properties of F for the last.

Now, we apply Dudley's inequality [42, Thm 8.1.6] to obtain that with probability $1 - \rho$,

$$\begin{aligned}
\sup_{z,z'\in\mathcal{B}_r} |Y_z - Y_{z'}| &\lesssim \frac{C_L}{\sqrt{n}}\left(\int_0^r \sqrt{\log N(\mathcal{B}_r, d, \varepsilon)}d\varepsilon + \sqrt{\log(1/\rho)}r\right) \\
&\lesssim \frac{C_L r}{\sqrt{n}}\left(\sqrt{d} + \sqrt{\log(1/\rho)}\right)
\end{aligned}$$

which concludes the proof. □

Lemma 7. *Let $x_1, \dots, x_n$ be iid $\mathcal{N}_{0,\Sigma}$ on $\mathbb{R}^d$, and W be a 1-bounded, C-Lipschitz kernel in the first variable with respect to the metric $\|\cdot\|_{\Sigma^{-1}}$.*

With probability at least $1-\rho$,

$$\sup_i \left| \frac{1}{n} \sum_j W(x_i, x_j) - \mathbb{E} W(x_i, X) \right| \lesssim \frac{\sqrt{\log n} C \left(\sqrt{d} + \sqrt{\log(1/\rho)} \right)}{\sqrt{n}} \tag{28}$$

With probability at least $1-\rho$,

$$\sup_i \left\| \frac{1}{n} \sum_j W(x_i, x_j) x_j - \mathbb{E} W(x_i, X) X \right\|_{\Sigma^{-1}} \lesssim \frac{\log n C \left(\sqrt{d} + \sqrt{\log(1/\rho)} \right)}{\sqrt{n}} \tag{29}$$

Proof. By the properties of Gaussian variables and a union bound, with probability at least $1-\rho$,

$$\forall i, \|x\|_{\Sigma^{-1}} \lesssim \sqrt{\log n} =: r_n \tag{30}$$

Now, since W is bounded, $W(x, X)$ is subgaussian for any x. Applying Lemma 6 with $F = W$ and considering that $\|\cdot\|_{\psi_2} \leqslant \|\cdot\|_\infty$, we get that with probability at least $1-\rho$,

$$\sup_{\|x\|_{\Sigma^{-1}} \leqslant r_n} \left| \frac{1}{n} \sum_j W(x, x_j) - \mathbb{E} W(x, X) \right| \lesssim \frac{r_n C \left(\sqrt{d} + \sqrt{\log(1/\rho)} \right)}{\sqrt{n}} \tag{31}$$

Combining with (30), we get (28).

Now, we write

$$\begin{aligned} &\sup_{\|x\|_{\Sigma^{-1}} \leqslant r_n} \left\| \frac{1}{n} \sum_j W(x, x_j) x_j - \int W(x, x') x' \mathcal{N}_{0,\Sigma}(x') dx' \right\|_{\Sigma^{-1}} \\ &= \sup_{\|x\|_{\Sigma^{-1}} \leqslant r_n} \sup_{\|u\|_\Sigma \leqslant 1} \left| \frac{1}{n} \sum_j W(x, x_j) u^\top x_j - \int W(x, x') u^\top x' \mathcal{N}_{0,\Sigma}(x') dx' \right| \end{aligned}$$

We aim to apply again Lemma 6 for the function $F((x,u), x') = W(x, x') u^\top x'$ and the metric $\|x\|_{\Sigma^{-1}} + \|u\|_\Sigma$. First, for any u with $\|u\|_\Sigma \leqslant 1$, $u^\top X$ is Gaussian with variance less than 1, so $\left\| W(x, X) u^\top X \right\|_{\psi_2} \leqslant \|W(x, \cdot)\|_\infty \left\| u^\top X \right\|_{\psi_2} \lesssim 1$. Similarly, $(u - u')^\top X$ is Gaussian with variance $\|u - u'\|_\Sigma$, so

$$\begin{aligned} \|F((x,u), X) - F((x',u'), X)\|_{\psi_2} &\leqslant \|W(x, \cdot) - W(x', \cdot)\|_\infty \left\| (u-u')^\top X \right\|_{\psi_2} \\ &\lesssim C \|x - x'\|_{\Sigma^{-1}} \|u - u'\|_\Sigma \\ &\lesssim r_n C (\|x - x'\|_{\Sigma^{-1}} + \|u - u'\|_\Sigma) \end{aligned}$$

Hence, we get

$$\sup_{\|x\|_{\Sigma^{-1}} \leqslant r_n} \left\| \frac{1}{n} \sum_j W(x, x_j) x_j - \mathbb{E} W(x, X) X \right\| \lesssim \frac{r_n^2 C \left(\sqrt{d} + \sqrt{\log(1/\rho)} \right)}{\sqrt{n}} \tag{32}$$

which concludes the proof. □