Asymptotic generalization error of a single-layer graph convolutional network

O. Duranthon and **L. Zdeborová**
Statistical physics of computation laboratory
École polytechnique fédérale de Lausanne (EPFL), Switzerland
firstname.lastname@epfl.ch

Abstract

While graph convolutional networks show great practical promises, the theoretical understanding of their generalization properties as a function of the number of samples is still in its infancy compared to the more broadly studied case of supervised fully connected neural networks. In this article, we predict the performances of a single-layer graph convolutional network (GCN) trained on data produced by attributed stochastic block models (SBMs) in the high-dimensional limit. Previously, only ridge regression on contextual-SBM (CSBM) has been considered in [1]; we generalize the analysis to arbitrary convex loss and regularization for the CSBM and add the analysis for another data model, the neural-prior SBM. We derive the optimal parameters of the GCN. We also study the high signal-to-noise ratio limit, detail the convergence rates of the GCN and show that, while consistent, it does not reach the Bayes-optimal rate for any of the considered cases.

1 Introduction and related work

Understanding the generalization properties of neural networks on unseen data is still unsatisfactory despite the very active line of work in this direction. In this article, we are specifically interested in understanding the generalization properties of graph neural networks, where the question remains even further from closed compared to feedforward neural networks that have been explored more broadly in the theoretical literature.

Tight analysis in the high-dimensional limit: The question of generalization has been studied from many angles. Classical learning theory usually aims to avoid assumptions on the data distribution and to provide generic generalization bounds. Such bounds are, however, often far away from the actual performance on given benchmark datasets, see e.g. [2]. This generic line of work is hence complemented by studies of concrete data distributions and concrete target functions. Tight theoretical results are attainable in the high-dimensional limit, where the number of samples and their dimension go to infinity while being proportional. In this limit many quantities of interest concentrate on deterministic values for which a closed-set of dimension-independent fixed point equations is derived; see e.g. [3–7]. This nice property is referred to as the blessing of dimensionality. This line of theoretical analysis is very appealing because it is able to provide results for the information-theoretically attainable generalization error, as well as the one obtained by a specific neural network. This allows us to evaluate the gap between the generalization ability of neural networks and the information-theoretically optimal one. The amplitude of the gaps to optimality can then be used to drive the development of architectures and algorithms that decrease the gap. The behaviour of systems of moderate sizes converges very fast to the asymptotic behaviour derived in the high-dimensional limit, thus making it relevant and interesting, as shown by the above works. The main drawback of this line of work is that so far the available theoretical tools only allow such analysis for only very simple network architectures, e.g. single layer and two-layer neural networks. Still, there are many open questions for the two-layer case [8]; and even for the simpler single-layer case, which

, Asymptotic generalization error of a single-layer graph convolutional network. *Proceedings of the Third Learning on Graphs Conference (LoG 2024)*, PMLR 269, Virtual Event, November 26–29, 2024.

corresponds to high-dimensional regressions, many open questions have been settled only recently: see e.g. [7, 9, 10]. However, the long-term promise of this direction of research motivates efforts to establish the tight asymptotic analysis and the underlying tools in broader and broader settings. The present work is inscribed in this context and it treats a graph convolutional neural network. In the same sense as done in the literature for the feed-forward fully connected networks, we will consider only single-layer graph convolutional networks (GCNs). This is a clear limitation of our work which is justified by the technical challenge of that setting and by the overall aim to build theoretical tools and understanding that will be able to deal with more realistic architectures in the future. Yet, on a practical point of view, linear single-layer GCNs can have similar performances to non-linear multi-layer ones, while being able to deal with very large graphs and being much simpler to train, as shown by [11] and [12].

Generalization in graph neural networks: Graph neural networks (GNNs) show a broad range of practical applications, and, as such, understanding their generalization properties is an important part of our overall goal. Many works consider graph or node classification in a learning scenario where one has access to many training graphs and unseen test graphs. Some works then derive bounds based on VC dimension, Rademacher complexity or PAC-Bayesian analysis, see for instance [13] and the references therein; wide networks can be analyzed thanks to graph neural tangent kernel, see e.g. [14]. Instead, we consider the semi-supervised (or transductive) learning scenario, where training and inference are done on the same large graph whose node labels are partially revealed. This setting is relevant for node classification problems such as community detection. Previous theoretical works on semi-supervised learning include [15], which studies learning under stochastic gradient descent, or [16] that focuses on graph convolutional networks and proposes experiments on data generated by the contextual stochastic block model (CSBM). More similar questions to our work are addressed in [17] that derives generalization bounds for a particular model of data close to the CSBM, yet considering a generic GNN. These three works derive only loose bounds for the test performances of the GNN and they do not provide insights on the effect of the structure of data, such as its heterophily. For instance [17] derives bounds based on transductive Rademacher complexity; since they are too general the authors have to model the data as a CSBM. Still the error bound they obtain is increasing with the number of samples N, which in the limit of large N provides no guarantee. [15] provides sharper bounds; yet they are not tight, do not take in account the data and depend on continuity constants that cannot be determined a priori. A series of works closer to our article has been developed by the authors of [18]. In this work, they consider a one-layer GCN trained on the CSBM by logistic regression and derive bounds for the test loss; however, they analyze its generalization ability on new graphs that are independent of the train graph and do not give exact predictions. In [19] they propose an architecture of GNN that is optimal for the CSBM, among classifiers that process local tree-like neighborhoods, and they exactly derive its generalization error. These two works consider a low-dimensional setting.

The tight analysis of generalization in synthetic high-dimensional settings for GNNs is still in its infancy. The only pioneering reference in this direction we are aware of is [1] where the authors consider a simple one-layer GCN trained in a semi-supervised way by ridge regression. They predict its asymptotic performances on data generated by the CSBM and, in particular, show how to tune the architecture to adapt to the homophily strength of the graph.

A starting point of the tight asymptotic analysis of generalization is a suitable model for generating data. As [1] showed, the CSBM introduced in [20, 21] is suitable. Data generated by this model has been used to benchmark various GNN architectures in [16, 22–24] for instance. Another way to generate graph data with node features is the neural-prior or generalized linear model SBM (GLM–SBM) introduced in [25], where the features alone do not bring any information. For these two models, the CSBM and the GLM–SBM, the optimal performance has been derived in the high-dimensional limit in [25–27].

Our motivation to extend [1] comes from the related line of research we detailed above. This work does not compare the performance of the GCN to the Bayes-optimality nor study the interplay between the loss, the regularization and the data; while, as to high-dimensional regression, [7, 28] established that the generalization error of the ridge regression is suboptimal for some models of data while logistic regression is much closer to the Bayes-optimality. When it comes to rates with which the test error goes to zero in the limit of a large number of samples, they are again suboptimal for ridge regression while they give the Bayes-optimal rates for optimally regularized logistic regression [7]. For a slightly different setting the Bayes-optimal performance can be achieved [28] just by

adjusting the regularization. Natural questions thus are: how does the performance of the GCN from [1] compare to the Bayes-optimal performance? How much do optimal regularization, architecture or loss improve the generalization? How does this reflect in rates when the signal-to-noise ratio is large? These questions are answered in the present article.

Main contribution: First we generalize the analysis of [1] by considering generic loss and regularization for the CSBM and the GLM–SBM. We derive the summary statistics and the self-consistent equations they follow, which allow us to predict the exact generalization performance of the GCN in the high-dimensional limit. We show that these predictions are in very good agreement with numerical simulations of the GCN at finite N.

Using these predictions we compare to the Bayes-optimal test accuracy, search for the optimal parameters of the considered architecture and explore several common loss functions. We show that in the considered setting large regularization maximizes the test accuracy for the CSBM while leading to a test accuracy close to the optimum for the GLM–SBM; ridge regression has a large gap to the optimality, and the logistic and hinge losses do not improve it significantly. We derive an explicit formula for the test accuracy in the limit of large regularization, that allows us to make further predictions and understand rather explicitly the trade-off between how the GCN uses the graph and the features. Then we take the limit of high signal-to-noise ratio (snr). We show that the simple GCN we consider is consistent in the sense that the test error converges to zero as the snr diverges. We derive the convergence rates for the two models; they appear to be smaller than the Bayes-optimal one. Last we derive the optimal self-loop strength of the GCN and provide evidence that this prediction may be generalizable to a broader class of datasets.

2 Models, setup

Attributed SBMs: We consider a set of N nodes and a graph G. Each node i has a label $y_i = \pm 1$; we consider two balanced groups. We precise the law of y_i later. We observe an adjacency matrix $A \in \mathbb{R}^{N\times N}$ whose components are drawn according to

$$A_{ij} \sim \mathcal{B}\left(\frac{d}{N} + \frac{\lambda}{\sqrt{N}}\sqrt{\frac{d}{N}\left(1-\frac{d}{N}\right)}y_i y_j\right) \tag{1}$$

where λ is the signal-to-noise ratio (snr) of the graph, d is the average degree of the graph, $\mathcal{B}$ is a Bernoulli law and the components A_{ij} are independent random variables. We take an average degree d of order N, but d growing with N should be sufficient for our results to hold. We discuss this assumption more in detail in the appendix B.1. We consider a directed SBM, A non-symmetric, to simplify the analysis; yet this model can be mapped to a non-directed SBM of snr $\lambda' = \sqrt{2}\lambda$ by taking the adjacency matrix $(A + A^T)/\sqrt{2}$.

We consider M hidden independent standard Gaussian variables u_ν; we set $\alpha = N/M$ the aspect ratio. We also observe features $X \in \mathbb{R}^{N\times M}$. The features are correlated with the node labels. We consider first the contextual stochastic block model (CSBM) [20, 21] for which the labels are Rademacher and the features follow a Gaussian mixture:

$$(\text{CSBM}) \quad y_i \sim \text{Rad}\ ,\ X = \sqrt{\frac{\mu}{N}}yu^T + W \tag{2}$$

where μ is the snr of the features and W is noise whose components $W_{i\nu}$ are independent standard Gaussians. We will also consider another related model, the neural-prior or GLM–SBM [25], for which the features are Gaussian and the labels are generated by a generalized linear model (GLM) on the features, the sign being applied element-wise:

$$(\text{GLM}-\text{SBM}) \quad X_{i\nu} \sim \mathcal{N}(0,1)\ ,\ y = \text{sign}\left(\frac{1}{\sqrt{N}}Xu\right)\ . \tag{3}$$

We are given a set R of train nodes and define $\rho = |R|/N$ the training ratio. The test set R' is selected from the complement of R; we define $\rho' = |R'|/N$ as the testing ratio. We assume that R and R' are independent from the other quantities. The inference problem is to find back y and u given A, X, R and the parameters of the model.

We work in the high-dimensional limit $N \to \infty$ and $M \to \infty$ while the aspect ratio $\alpha = N/M$ is of order one. The other parameters λ, μ, ρ and ρ' are also of order one.

We precise that the total snr of the symmetric CSBM and GLM–SBM are [21, 25]

$$\mathrm{snr}_{\mathrm{CSBM}} = \lambda^2 + \frac{\mu^2}{\alpha}\,, \qquad \mathrm{snr}_{\mathrm{GLM-SBM}} = \lambda^2\left(1+\frac{4\alpha}{\pi^2}\right)\,. \tag{4}$$

Authors of [21, 25, 29] established that $\mathrm{snr}_{\mathrm{CSBM}} = 1$ and $\mathrm{snr}_{\mathrm{GLM-SBM}} = 1$ are the detectability thresholds in the sense that in the unsupervised case $\rho = 0$ they separate an undetectable phase, where the labels y cannot be recovered better than at random, from a detectable phase where they can. In the semi-supervised case $\rho > 0$ this transition disappears and one can always recover some information on the test labels. The expression of $\mathrm{snr}_{\mathrm{CSBM}}$ shows that the snr originating from the graph is of the strength λ^2 while the one originating from the features is μ^2/α.

Analyzed GCN architecture: We follow [1] and we consider a single-layer graph convolutional network (GCN). It transforms the features according to

$$h(w) = \frac{1}{N} Q(\tilde{A}) X w \tag{5}$$

where Q is a polynomial, $w \in \mathbb{R}^M$ are the trainable weights and $\tilde{A} \in \mathbb{R}^{N\times N}$ is a rescaling of the adjacency matrix defined by $\tilde{A}_{ij} = \left(\frac{d}{N}\left(1-\frac{d}{N}\right)\right)^{-1/2}\left(A_{ij} - \frac{d}{N}\right)$. For the analysis, we consider Q of degree one as in [1], i.e. $Q(\tilde{A}) = \tilde{A} + c\sqrt{N} I_N$ where c is a tunable parameter of the architecture. This corresponds to applying one step of graph convolution to the features with self-loops.

This GCN is trained by empirical risk minimization. We define the regularized loss

$$L_{A,X}(w) = \frac{1}{\rho N}\sum_{i\in R} l(y_i h_i(w)) + \frac{r}{\rho N}\sum_{\nu}\gamma(w_\nu) \tag{6}$$

where γ is a strictly convex regularization function, r is the regularization strength and l is a convex loss function. We will focus on l_2-regularization $\gamma(x) = x^2/2$ and on the square loss $l(x) = (1-x)^2/2$, the logistic loss $l(x) = \log(1+e^{-x})$ or the hinge loss $l(x) = \max(0, 1-x)$. Since L is strictly convex it admits a unique minimizer w^*. The average train and test errors and accuracies of this model are

$$E_{\mathrm{train/test}} = \mathbb{E}\,\frac{1}{|\hat{R}|}\sum_{i\in\hat{R}} l(y_i h(w^*)_i)\,, \qquad \mathrm{Acc}_{\mathrm{train/test}} = \mathbb{E}\,\frac{1}{|\hat{R}|}\sum_{i\in\hat{R}\,,} \delta_{y_i = \mathrm{sign}\, h(w^*)_i} \tag{7}$$

where $\hat{R}$ stands either for the train set R or the test set R' and the expectation is taken over y, u, A, X, R and R'.

Table 1: Summary of the parameters of the model.

N	size of the graph	μ	snr of the Gaussian mixture
M	dimensionality of the attributes	l, γ	loss and regularization functions
$\alpha = N/M$	aspect ratio	$\rho = \lvert R\rvert/N$	fraction of training nodes
d	average degree of the graph	r	regularization strength
λ	snr of the SBM	c	self-loop strength

Bayes-optimal performances: An important consequence of modeling the data as we propose is that one has access to the Bayes-optimal (BO) performance on this task, i.e. the upper-bound on the test accuracy that any algorithm can reach, knowing the model and its parameters α, d, λ, μ. It is of particular interest since it will allow us to check how far the GCN is from the optimality and how much improvement can be done. The BO performances for both the CSBM and the GLM–SBM have been derived in [25–27]. They can be expressed as a function of the solution of the equations reproduced in appendix C.

3 Asymptotic prediction of the performances of the GCN

In this section we state our main result, namely the asymptotic formulae for the expected losses and accuracies of the trained GCN. We will derive several consequences from these in the next section.

Result 3.1 (Performances on the CSBM)**.** *We consider the high-dimensional limit defined in the previous section. Let u, ς, ξ, ζ and χ be standard Gaussian random variables and y be a Rademacher random variable. Let $\Theta = \{m_w, m_\sigma, Q_w, Q_\sigma, V_w, V_\sigma\}$ and $\hat{\Theta} = \{\hat{m}_w, \hat{m}_\sigma, \hat{Q}_w, \hat{Q}_\sigma, \hat{V}_w, \hat{V}_\sigma\}$ be*

the twelve real numbers that satisfy the system of equations (28)-(36) in appendix A. We introduce the two potentials

$$\psi_w(w) = -r\gamma(w) - \frac{1}{2}\hat{V}_w w^2 + \left(\varsigma\sqrt{\hat{Q}_w} + u\hat{m}_w\right) w \tag{8}$$

$$\psi_{\text{out}}(h,\sigma;\bar{t}) = -\bar{t}l(yh) - \frac{1}{2}\hat{V}_\sigma \sigma^2 + \left(\xi\sqrt{\hat{Q}_\sigma} + y\hat{m}_\sigma\right)\sigma \tag{9}$$

$$+ \log\mathcal{N}\left(h|c\sigma + \lambda y m_\sigma + \sqrt{Q_\sigma}\zeta, V_\sigma\right) + \log\mathcal{N}\left(\sigma|\sqrt{\mu}y m_w + \sqrt{Q_w}\chi, V_w\right)$$

where $\mathcal{N}(\cdot|m,V)$ is a scalar Gaussian density of mean m and variance V. The parameter $\bar{t}\in\{0,1\}$ controls if a given node is revealed $\bar{t}=1$ or not $\bar{t}=0$. We introduce the extremizers of these potentials:

$$w^* = \underset{w}{\text{argmax}}\, \psi_w(w) \tag{10}$$

$$(h^*,\sigma^*) = \underset{h,\sigma}{\text{argmax}}\, \psi_{\text{out}}(h,\sigma;\bar{t}=1) \qquad (h^{'*},\sigma^{'*}) = \underset{h,\sigma}{\text{argmax}}\, \psi_{\text{out}}(h,\sigma;\bar{t}=0)\,. \tag{11}$$

Then the expected errors and accuracies of the GCN on the CSBM are

$$E_{\text{train}} = \mathbb{E}_{y,\xi,\zeta,\chi} l(yh^*) \qquad \text{Acc}_{\text{train}} = \mathbb{E}_{y,\xi,\zeta,\chi}\delta_{y=\text{sign}(h^*)} \tag{12}$$

$$E_{\text{test}} = \mathbb{E}_{y,\xi,\zeta,\chi} l(yh^{'*}) \qquad \text{Acc}_{\text{test}} = \mathbb{E}_{y,\xi,\zeta,\chi}\delta_{y=\text{sign}(h^{'*})}\,. \tag{13}$$

The analysis of the GCN is thus reduced to the analysis of a finite set of scalar quantities Θ and $\hat{\Theta}$. They are called the summary statistics (or order parameters) of this model and they entirely describe its macroscopic properties. The equations (28)-(36) they satisfy are called the self-consistent or fixed-point equations.

Result 3.2 (Performances on the GLM–SBM)**.** *The performances of the GCN on the GLM–SBM are given by the same formulae as for the CSBM, except that the law of y is $P(y=\pm1|\chi) = \frac{1}{2}\left(1 \pm \text{erf}\left(\frac{m_w\chi}{\sqrt{2(\alpha^{-1}Q_w - m_w^2)}}\right)\right)$, that Θ and $\hat{\Theta}$ are the solution to the equations (41)-(49) and that ψ_{out} is taken at $\mu=0$.*

In general there is no simple expression to the solution of the self-consistent equations and one has to solve them numerically or to consider special cases. We consider the limiting case $r\to\infty$. It is particularly relevant for two reasons. First in this limit simple explicit expressions can be stated; we give them in appendix A.3 and in result 3.3. Second, as we will show in 4.1, it corresponds to the optimal performance of the GCN on the CSBM, and close to optimal for the GLM–SBM, and it is thus the right limit to analyze how effective the GCN is. The ridge-less limit $r=0$ and $\alpha\rho>1$ has been studied by [1] for the CSBM. We checked that in this case our expressions for the errors and the accuracies match theirs.

Result 3.3 (Large regularization case)**.** *We consider $r\to\infty$. For simplicity we state here the case $c=0$; the case $c\neq0$ is given in appendix A.3. Then the test accuracy of the trained GCN is*

$$\text{Acc}_{\text{test}} = \frac{1}{2}\left(1 + \text{erf}(\lambda\sqrt{\tau})\right) \tag{14}$$

where τ reads, respectively on the CSBM and on the GLM–SBM:

$$\sqrt{\tau_{\text{CSBM}}} = \frac{\lambda\rho(1+\mu)}{\sqrt{2}\sqrt{\rho(1+\alpha)+\lambda^2\rho^2(1+\mu)(1+\alpha+\mu)}} \tag{15}$$

$$\sqrt{\tau_{\text{GLM-SBM}}} = \frac{\lambda\rho(1+2\alpha/\pi)}{\sqrt{2}\sqrt{\rho(1+\alpha)+\lambda^2\rho^2((1+2\alpha/\pi)(1+\alpha+2\alpha/\pi)-4\alpha^2/\pi^2)}} \tag{16}$$

Outline of the derivation: We compute the expected errors and accuracies in the high-dimensional limit N and M large. This problem can be phrased in the same way as in [1]. We define an extended loss function

$$H(w) = t\sum_{i\in R} l(y_i h(w)_i) + r\sum_\nu \gamma(w_\nu) + t'\sum_{i\in R'} l(y_i h(w)_i) \tag{17}$$

where t and t' are external parameters to probe the observables. The loss of the test samples is in H for the purpose of the analysis; we will take $t'=0$ later and the algorithm is still minimizing the

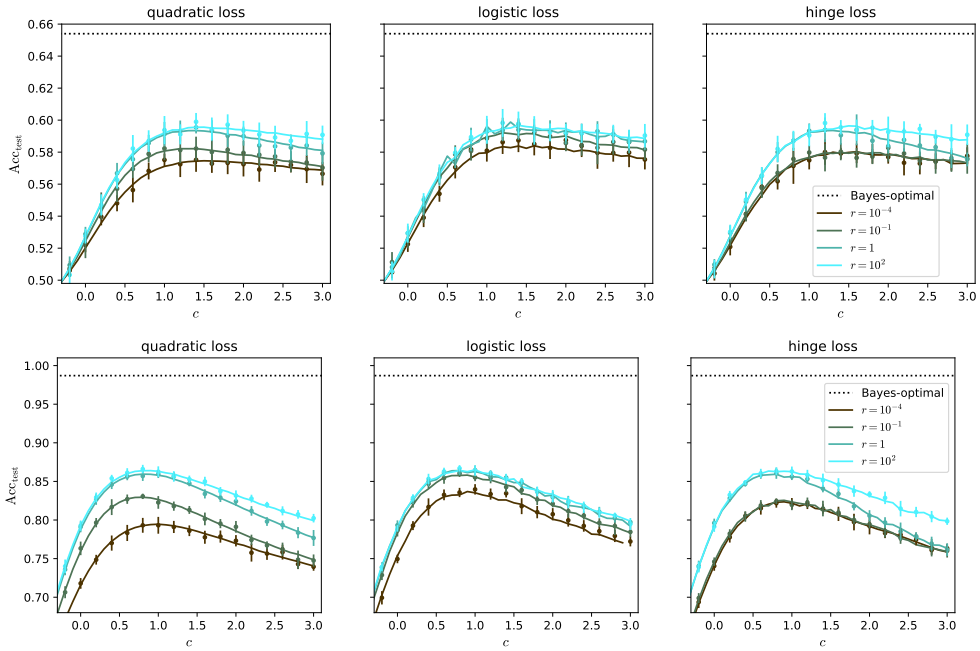

Figure 1: Search for the optimal parameters of the GCN on CSBM. $\alpha = 4$, $\rho = 0.1$. *Top:* low snr, $\lambda = 0.5$, $\mu = 1$. *Bottom:* high snr, $\lambda = 1.5$, $\mu = 3$. Full lines: prediction for the test accuracy obtained by eqs. (13) and (28)-(36); dots: numerical simulation of the GCN for $N = 10^4$ and $d = 30$, averaged over ten experiments; dotted line: Bayes-optimal test accuracy.

training loss eq. 6. The moment generating function f is defined as

$$Z = \int \mathrm{d}w \, e^{-\beta H(w)} \, , \quad f = -\frac{1}{\beta N} \mathbb{E} \log Z \, . \tag{18}$$

β is an ancillary parameter to minimize the loss: we consider the limit $\beta \to \infty$ where Z concentrates over w^* at $t = 1$ and $t' = 0$. The train and test errors are then obtained according to $E_{\mathrm{train}} = \frac{1}{\rho} \partial_t f$ and $E_{\mathrm{test}} = \frac{1}{\rho'} \partial_{t'} f$ both evaluated at $t = 1$ and $t' = 0$. One can in the same manner compute the average accuracies by introducing the observables $\sum_{i \in \hat{R}} \delta_{y_i = \mathrm{sign}\, h(w)_i}$ in H.

To compute f we use the powerful but non-rigorous replica method from Statistical Physics: $\mathbb{E} \log Z = \mathbb{E} \partial_n Z^n (n = 0) = \partial_n \mathbb{E} Z^n (n = 0)$. Z^n is interpreted as having n independent replica of the initial system, that become coupled by the expectation. We pursue the computation under the replica symmetry (RS) assumption, which is justified by the convexity of H. We introduce an intermediate variable $\sigma = \frac{1}{\sqrt{N}} X w$ that corresponds to the projected features and that appears in the previous equations. The computation is then detailed in appendix B.

4 Consequences

In the previous part we described the performances of the trained GCN by a finite set of summary statistics in the high-dimensional limit and we gave some explicit expressions. In this section we derive consequences from these equations. In particular we will search for the parameters of the GCN that optimize the test accuracy, to see whether the GCN can reach the Bayes-optimality. The possible tunable parameters are the self-loop intensity c, the regularization strength r and the loss l. As to the regularization γ, we consider only l_2-regularization since we are in a simple setting not involving sparsity or outliers where l_1-regularization would have been beneficial. In general the system of equations (28)-(36) and (41)-(49) defining Θ and $\hat{\Theta}$ has to be solved numerically and one has to choose particular values for the parameters of the data models. For these, we consider both low and high snr, on both the CSBM and the GLM–SBM; we keep the signals of the graph and the features balanced and we take $\rho = 0.1$ to mimic the common case where relatively few train labels are available. We did not explore all the parameters of the data models; instead we focused on plausible values and some corner cases may not follow our statements.

Details on the numerics are provided in appendix E. Our theoretical predictions are compared to simulations of the GCN on figs. 1, 4, 5, 6, 8 and 9 for $N = 10^4$ and $d = 30$ or $d = N/2$. As expected, the predicted test accuracy, train accuracy and errors are within the statistical errors.

Result 4.1 (Effect of the loss and the regularization)**.** *Based on the numerical exploration of our equations shown in fig. 1 and in figs. 4, 5 and 6 in appendix F, we reach the conclusion that for both the CSBM and the GLM–SBM:*

1. *the optimal test accuracies* $\mathrm{Acc}_{\mathrm{test}}$ *depend little on the choice of the loss* l*. On the CSBM it appears to be reached at large regularization* $r \to \infty$*; on the GLM–SBM large regularization* $r \to \infty$ *is close to the optimal* r*;*
2. *there is an optimal self-loop strength* c^* *maximizing* $\mathrm{Acc}_{\mathrm{test}}$*;* c^* *is of order one;*
3. *there is a gap between the optimal test accuracy of the GCN and the Bayes-optimal test accuracy.*

We observe in figs. 1, 5 and 6 that on the CSBM for all self-loop strengths c the test accuracy increases with the regularization r and reaches an optimal value at $r \to \infty$. As to the GLM–SBM, we observe in figs. 4, 5 and 6 that $r \to \infty$ is close to the optimality, in particluar if c is not too large. Notice that at $r \to \infty$ the weights w and the output $h(w)$ shrink to zero and that the test and train errors are large; yet this is not an issue: to assess the performance in a classification problem, the relevant quantity is the accuracy, not the error. At $r \to \infty$ the signs of $h(w)$ are mostly correct and the accuracies have a non-trivial value. At low regularization r we checked that interpolation peaks appear for the different losses while varying α or ρ; see figs. 8 and 9 in appendix F. Increasing r smooths the peaks out, as [1] shows for the quadratic loss; and as it is well known for the feed-forward networks, see e.g. [28].

A surprising result is that the optimal accuracy does not depend significantly on the loss; in particular, we do not see any significant difference between the three considered losses at optimal regularization. This is striking because it is rather generically anticipated that for classification the quadratic loss is less suitable than the logistic or hinge losses. One previous example of $r \to \infty$ being optimal is classification on a binary high-dimensional Gaussian mixture [28]. On the CSBM the CGN behaves similarly, which could be expected since the features X are a Gaussian mixture. On the GLM–SBM where X is generated by a GLM, it seems that they are partly mixed by the convolution $Q(\tilde{A})X$, depending on the self-loops c. The fact that at $r \to \infty$ the three losses behave similarly is expected because the output h is small and l can be expanded around 0, where the three losses are identical.

More generally, at fixed small r, the logistic/hinge loss has better performances than the quadratic loss, as shown on figs. 1 and 4. If not regularized the quadratic loss always suffers from the interpolation peak at $\rho\alpha = 1$, where the test accuracy is $1/2$, as shown on fig. 8. For the logistic/hinge loss, the interpolation threshold is less harmful and it can be moved away with λ and c, as shown on fig. 9. A consequence is that at large λ the logistic/hinge loss does not need regularization and reaches its optimal value even at small r, as depicted on fig. 7 in app. F, while the quadratic loss needs $r \to \infty$.

Another remarkable point is that the performances of the GCN are far from the Bayes-optimal performances (dotted lines in the figures) in all cases. This is a major difference with the feed-forward case [7, 28], which shows that well-regularized regression performs very closely to the Bayes-optimal accuracy. One could argue that this can be expected since the GCN performs only one step of convolution; estimators $Q(\tilde{A})Xw$ with a higher-order polynomial Q could be better. Yet such a gap exists even for more elaborated GNNs on CSBM [26] and GLM–SBM [25].

The two following results 4.2 and 4.3 come from the analysis of eqs. (52) and (57) in appendix A.3 as to the CGN, and from eqs. (142) and (152) in appendix C as to the Bayes-optimal performances.

Result 4.2 (Consistency and convergence rates)**.** *We consider the limit of high graph signal* $\lambda \to \infty$ *at large regularization* $r \to \infty$*. We take* $c = 0$ *or* $c = c^*$ *the optimal self-loop strength. Then the GCN is consistent on both models:*

$$\mathrm{Acc}_{\mathrm{test}} \underset{\lambda\to\infty}{\longrightarrow} 1 \,, \qquad \log(1 - \mathrm{Acc}_{\mathrm{test}}) \underset{\lambda\to\infty}{\sim} -\lambda^2 \tau^\infty \,, \tag{19}$$

where τ^∞ *is the asymptotic convergence (or learning) rate; for the CSBM and the GLM–SBM respectively it reads*

$$\tau^\infty_{\mathrm{CSBM}} = \frac{1+\mu}{2(1+\alpha+\mu)} \,, \qquad \tau^\infty_{\mathrm{GLM-SBM}} = \frac{1+2\alpha/\pi}{2\left(1+\alpha+\frac{2\alpha/\pi}{1+2\alpha/\pi}\right)} \,. \tag{20}$$

Optimizing $\mathrm{Acc}_{\mathrm{test}}$ *on* c *only leads to a sub-leading improvement compared to taking* $c = 0$*. In both models the Bayes-optimal rate is* $\tau^\infty_{\mathrm{BO}} = 1$*.*

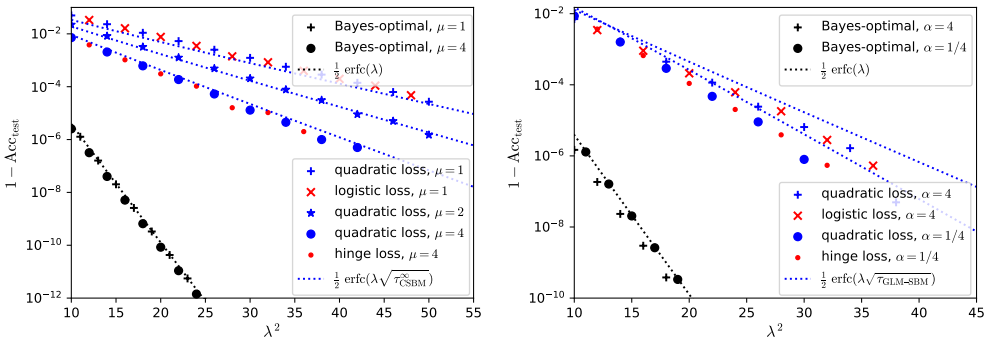

Figure 2: Asymptotic misclassification error $1-\mathrm{Acc}_{\mathrm{test}}$; *left:* on the CSBM, $\alpha=4$; *right:* on the GLM–SBM. $r=10^3$, $\rho=0.1$. Dots: prediction for the test accuracy obtained by eqs. (13), (28)-(36) and (41)-(49), for $c=c^*$ optimal obtained by grid search. Dotted lines are given by (20) (for $\tau^{\infty}_{\mathrm{CSBM}}$) and (16) (for $\tau_{\mathrm{GLM-SBM}}$). The Bayes-optimal values are obtained from the equations given in appendix C.

Consequently, the GCN never reaches the Bayes-optimal rate. These statements are in agreement with the numerics depicted in fig. 2.

The expressions of the convergence rates (20) are simple enough to be interpreted. As to $\tau^{\infty}_{\mathrm{CSBM}}$, this expression highlights the importance of the features: even at large graph snr λ the GCN relies on the snr μ of the features. Indeed $\tau^{\infty}_{\mathrm{CSBM}}$ is increasing with μ, from $1/2(1+\alpha)$ at $\mu=0$ to $1/2$ at large μ. As suggested by the expression of the snr of the CSBM (4), increasing α lowers the performance, since $\tau^{\infty}_{\mathrm{CSBM}}$ goes to zero for large α. The respective snrs μ and α do not contribute to $\tau^{\infty}_{\mathrm{CSBM}}$ in the same manner as in (4) where only the ratio μ^2/α matters. This is a sign that the GCN does not handle the features optimally. The GCN also seems not to handle the graph optimally. Indeed, the Bayes-optimal rate $\tau_{\mathrm{BO}}=1$ does not depend on the feature snr μ: hence, the graph alone is sufficient to reach the Bayes-optimal rate. As to $\tau^{\infty}_{\mathrm{GLM-SBM}}$, it converges to a finite value for large α, contrary to $\tau^{\infty}_{\mathrm{CSBM}}$ that goes to zero. This could be expected since the snr of the GLM–SBM (4) is increasing with α. A less intuitive result is that $\tau^{\infty}_{\mathrm{GLM-SBM}}$ reaches its maximum for α going to zero, as for $\tau^{\infty}_{\mathrm{CSBM}}$. It seems that there is a trade-off between the feature snr from the GLM (increasing with α) and the resulting feature snr of the convoluted features $Q(\tilde{A})X$ (decreasing with α).

The learning rates $\tau^{\infty}_{\mathrm{CSBM}}$ and $\tau^{\infty}_{\mathrm{GLM-SBM}}$ can be straightforwardly obtained by taking the limit in τ_{CSBM} and $\tau_{\mathrm{GLM-SBM}}$. Though being computed for $c=0$ they correctly described the leading behaviour of the GCN at $c=c^*$ because optimizing on c only leads to a sub-leading improvement in the limit $\lambda\to\infty$. This is shown in fig. 2 where the predicted values follow the slopes given by the different rates up to a small constant shift. As anticipated, this figure also shows that the three different losses give equal performances and the same rates.

The behaviour of the learning rates with respect to r is depicted on fig. 7 in appendix F. For the logistic loss, τ^{∞} does not visibly depend on r and even for small r it achieves its optimal performance; while for the quadratic loss τ^{∞} increases with r up to its limit $\tau^{\infty}_{\mathrm{GLM-SBM}}$ (20). As explained in section 4.1, this is because the interpolation peak is always present for the quadratic loss, while for the logistic loss at large λ and $c=c^*$ it disappears.

In conclusion, fig. 2 further illustrates that the GCN does not reach the Bayes-optimal rate. For all considered settings $\tau^{\infty}_{\mathrm{CSBM}}$ and $\tau^{\infty}_{\mathrm{GLM-SBM}}$ are bounded by $1/2$ while $\tau_{\mathrm{BO}}=1$. Moreover, the two τ^{∞} reach their upper bound $1/2$ only for the feature snr μ diverging or α going to zero, which confirms that the considered GCN has a rather poor performance.

Result 4.3 (Optimal self-loop strength c^*)**.** *We consider the limit $r\to\infty$. At $\lambda\to 0$, the optimal self-loop strength c^* reads*

$$c^*_{\mathrm{CSBM}}=\frac{\mu\left((1+\alpha)(2-\rho)+\rho(1+\mu)(1+\mu+\alpha)\right)}{\alpha(1+\mu)(2+\rho\mu)}\frac{1}{\lambda}\,,\qquad c^*_{\mathrm{GLM-SBM}}=\Theta(1/\lambda)\,. \tag{21}$$

At $\lambda\to\infty$, the optimal self-loop strength c^* reads

$$c^*_{\mathrm{CSBM}}=\frac{1+\mu+\alpha}{\alpha}\frac{1}{\lambda}\,,\qquad c^*_{\mathrm{GLM-SBM}}=\Theta(1/\lambda) \tag{22}$$

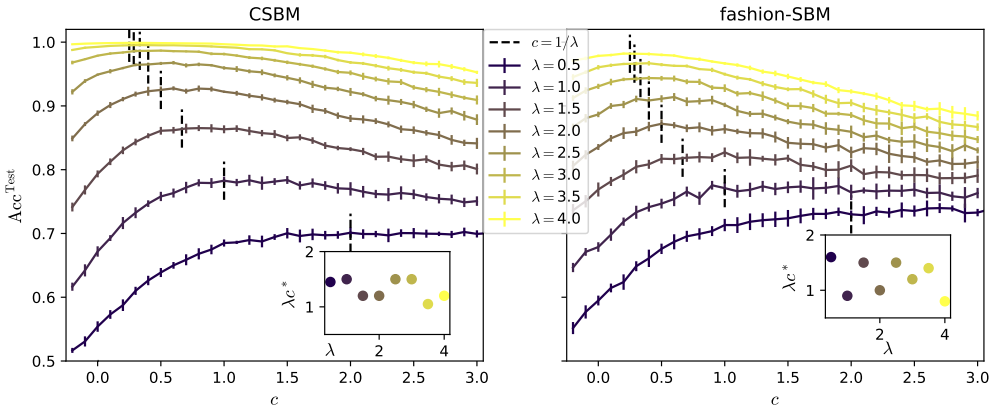

Figure 3: Optimal self-loop strength c^* vs graph snr λ. $d = 30$, $\rho = 0.1$, $r = 10^3$ and l quadratic. *Left:* on the CSBM, $N = 10^4$, $\alpha = 4$, $\mu = 3$. *Right:* on the fashion-SBM, classes 2 and 4. The lines are numerical simulations of the GCN averaged over ten experiments. c^* is computed as the extremizer of the simulated $\mathrm{Acc}_{\mathrm{test}}$.

where for the GLM–SBM the constant is given by solving eq. (62).

c^* behaves like $1/\lambda$ for λ both large and small and on both data models. Fig. 3 left shows that c^* can be approximated by $1/\lambda$ even for λ of order one. Fig. 3 right shows that the dependency $c^* \approx 1/\lambda$ seems to hold on a semi-realistic dataset, the fashion-SBM, defined in appendix D.

The case $\lambda \to \infty$ for the CSBM (22) is simple enough to be interpreted: c^*_{CSBM} increases with μ and decreases with α; this means that the larger the feature snr is, the more the features should be taken in account in the convolution, which is expected. Conversely, c^* increases when the graph snr λ decreases and reaches ∞ when $\lambda = 0$: the noisier the graph the less it should be considered in the convolution. The same happens in the case $\lambda \to 0$ for the CSBM (21) if ρ is small, in which case we have $c^*_{\mathrm{CSBM}} = \mu(1+\alpha)/\lambda\alpha(1+\mu)$. For an arbitrary λ, for the two models, c^* can still be predicted as the maximizer of eqs. (52) or (57) in the appendix, but it does not admit a simple expression.

An interesting result is that c^* behaves like $1/\lambda$ for λ both small and large, for both models. Though the constant factors differ, this suggests a universal behaviour, for any λ and beyond the two analyzed data models. We conjecture that, in general, taking $c^* = 1/\lambda$ is a good approximation for the extremizer of the test accuracy. We tested this conjecture: first (fig. 3 left) by considering λ of order one on the CSBM, and second (fig. 3 right) by training the GCN on a semi-realistic data model, the fashion-SBM, for which the features are taken from the fashion-MNIST dataset [30]. Fashion-SBM is defined defined in appendix D. In the two cases, for λ ranging from 0.5 to 4 we observe that λc^* remains close to 1, which seems to confirm our conjecture. This suggests that the rule $c^* = 1/\lambda$ can be extended to a broader range of data, not only from the CSBM or the GLM–SBM, and could be useful in practice. A theoretical interpretation of this universality could be that the convolution $Q(\tilde{A})X$ tends to transform the features X to a Gaussian mixture, irrespectively to their distribution. This would explain why the same behaviour appears for the different datasets.

Conclusion. We theoretically predicted the generalization performances and the optimal architecture of a one-layer GCN on two models of attributed graphs. We showed that the optimal test accuracy is achieved for a finite value of the self-loop intensity at large regularization; it does not depend visibly on the training loss and there is a significant gap to the Bayes-optimality. This stands both when the features and the labels are generated by a Gaussian mixture and when they are generated by a GLM. We derived the optimal learning rates of the GCN and showed they can be interpreted in terms of feature signal-to-noise ratios. The GCN is consistent at large graph snr but does not reach the Bayes-optimal rate. We hope this simple setting will be usefull in understanding which aspects of the GCN are key to reach the optimality.

A future direction of work could be to analyze more complex GNNs such as a GCN with higher-order graph convolution $Q(\tilde{A})$ or an attention-based GNN and to see if they can reach the optimality. Another direction could be given by the work [31] that proposes a model for genes where the components of the features are correlated according to a graph. One could study the role of graph-induced regularization.

Acknowledgements

We acknowledge discussions with Cheng Shi. This work is supported by the Swiss National Science Foundation under grant SNFS SMArtNet (grant number 212049).

References

[1] Cheng Shi, Liming Pan, Hong Hu, and Ivan Dokmanić. Homophily modulates double descent generalization in graph convolution networks. *PNAS*, 121(8), 2023. arXiv:2212.13069. 1, 2, 3, 4, 5, 7, 14, 15, 18

[2] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3): 107–115, 2021. 1

[3] Elizabeth Gardner and Bernard Derrida. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983, 1989. 1

[4] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019. 22

[5] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[6] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

[7] Benjamin Aubin, Florent Krzakala, Yue M. Lu, and Lenka Zdeborová. Generalization error in high-dimensional perceptrons: Approaching Bayes error with convex optimization. In *Advances in Neural Information Processing Systems*, 2020. arxiv:2006.06560. 1, 2, 7

[8] Hugo Cui, Florent Krzakala, and Lenka Zdeborova. Bayes-optimal learning of deep random networks of extensive-width. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6468–6521. PMLR, 23–29 Jul 2023. 1

[9] Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29): 14516–14525, 2019. 2

[10] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022. 2

[11] Felix Wu, Tianyi Zhang, Amauri Holanda de Souza Jr., Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. Simplifying graph convolutional networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019. arxiv:1902.07153. 2

[12] Hao Zhu and Piotr Koniusz. Simple spectral graph convolution. In *International Conference on Learning Representations*, 2021. 2

[13] Haotian Ju, Dongyue Li, Aneesh Sharma, and Hongyang R. Zhang. Generalization in graph neural networks: Improved PAC-Bayesian bounds on graph diffusion. In *AISTATS*, 2023. arXiv:2302.04451. 2

[14] Lianke Qin, Zhao Song, and Baocheng Sun. Is solving graph neural tangent kernel equivalent to training graph neural network? 2023. arXiv:2309.07452. 2

[15] Huayi Tang and Yong Liu. Towards understanding the generalization of graph neural networks. 2023. arXiv:2305.08048. 2

[16] Weilin Cong, Morteza Ramezani, and Mehrdad Mahdavi. On provable benefits of depth in training graph convolutional networks. 2021. arxiv:2110.15174. 2

[17] Pascal Mattia Esser, Leena Chennuru Vankadara, and Debarghya Ghoshdastidar. Learning theory can (sometimes) explain generalisation in graph neural networks. In *35th Conference on Neural Information Processing Systems*, 2021. arXiv:2112.03968. 2

[18] Aseem Baranwal, Kimon Fountoulakis, and Aukosh Jagannath. Graph convolution for semi-supervised classification: Improved linear separability and out-of-distribution generalization. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. arxiv:2102.06966. 2

[19] Aseem Baranwal, Kimon Fountoulakis, and Aukosh Jagannath. Optimality of message-passing architectures for sparse graphs. In *37th Conference on Neural Information Processing Systems*, 2023. arxiv:2305.10391. 2

[20] Bowei Yan and Purnamrita Sarkar. Covariate regularized community detection in sparse graphs. *Journal of the American Statistical Association*, 116(534):734–745, 2021. arxiv:1607.02675. 2, 3

[21] Yash Deshpande, Subhabrata Sen, Andrea Montanari, and Elchanan Mossel. Contextual stochastic block models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, 2018. arxiv:1807.09596. 2, 3, 4, 14

[22] Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptative universal generalized pagerank graph neural network. In *International Conference on Learning Representations*, 2021. arxiv:2006.07988. 2

[23] Guoji Fu, Peilin Zhao, and Yatao Bian. p-Laplacian based graph neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, 2022. arxiv:2111.07337.

[24] Runlin Lei, Zhen Wang, Yaliang Li, Bolin Ding, and Zhewei Wei. EvenNet: Ignoring odd-hop neighbors improves robustness of graph neural networks. In *36th Conference on Neural Information Processing Systems*, 2022. arxiv:2205.13892. 2

[25] O Duranthon and L Zdeborová. Neural-prior stochastic block model. *Mach. Learn.: Sci. Technol.*, 2023. arxiv:2303.09995. 2, 3, 4, 7, 21

[26] O Duranthon and L Zdeborová. Optimal inference in contextual stochastic block models. *Transactions on Machine Learning Research*, 2024. arxiv:2306.07948. 7, 21

[27] Benjamin Aubin, Bruno Loureiro, Antoine Maillard, Florent Krzakala, and Lenka Zdeborová. The spiked matrix model with generative priors. In *Advances in Neural Information Processing Systems*, 2019. arxiv:1905.12385. 2, 4, 21, 22

[28] Francesca Mignacco, Florent Krzakala, Yue M. Lu, and Lenka Zdeborová. The role of regularization in classification of high-dimensional noisy Gaussian mixture. In *International conference on learning representations*, 2020. arxiv:2002.11544. 2, 7

[29] Chen Lu and Subhabrata Sen. Contextual stochastic block model: Sharp thresholds and contiguity. 2020. arXiv:2011.09841. 4

[30] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. arxiv:1708.07747. 9, 22

[31] Sagnik Nandy and Subhabrata Sen. Bayes optimal learning in high-dimensional linear regression with network side information. 2023. arxiv:2306.05679. 9

[32] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):073403, 2017. arxiv:1701.00858. 14

A Self-consistent equations

The outcome of our computation is that the performances of the trained GCN are described by twelve scalar summary statistics (or order parameters): the alignements with the solution (or magnetizations) m_w, $\hat{m}_w$, m_σ and $\hat{m}_\sigma$, the norms (or self-overlaps) Q_w, $\hat{Q}_w$, Q_σ and $\hat{Q}_\sigma$, and the variances V_w, $\hat{V}_w$, V_σ and $\hat{V}_\sigma$.

These statistics follow a system of self-consistent equations we state here. Solving this system allows to predict the performances of the GCN in the high-dimensional limit.

We remind the notations from the main part. u, ς, ξ, ζ and χ are standard Gaussian random variables and y is a binary ± 1 random variable. We introduce the two potentials

$$\psi_w(w) = -r\gamma(w) - \frac{1}{2}\hat{V}_w w^2 + \left(\varsigma\sqrt{\hat{Q}_w} + u\hat{m}_w\right) w \tag{23}$$

$$\psi_{\text{out}}(h,\sigma;\bar{t}) = -\bar{t}l(yh) - \frac{1}{2}\hat{V}_\sigma \sigma^2 + \left(\xi\sqrt{\hat{Q}_\sigma} + y\hat{m}_\sigma\right)\sigma \tag{24}$$

$$+ \log\mathcal{N}\left(h|c\sigma + \lambda y m_\sigma + \sqrt{Q_\sigma}\zeta, V_\sigma\right) + \log\mathcal{N}\left(\sigma|\sqrt{\mu}y m_w + \sqrt{Q_w}\chi, V_w\right)$$

where $\mathcal{N}(\cdot|m,V)$ is a scalar Gaussian density of mean m and variance V. We introduce the extremizers of these potentials:

$$w^* = \underset{w}{\text{argmax}}\, \psi_w(w) \tag{25}$$

$$(h^*,\sigma^*) = \underset{h,\sigma}{\text{argmax}}\, \psi_{\text{out}}(h,\sigma;\bar{t}=1) \qquad (h^{'*},\sigma^{'*}) = \underset{h,\sigma}{\text{argmax}}\, \psi_{\text{out}}(h,\sigma;\bar{t}=0)\,. \tag{26}$$

For compactness we introduce the operator $\mathcal{P}$ that, for a polynomial Q in h and σ, acts according to

$$\mathcal{P}(Q(h,\sigma)) = \rho Q(h^*,\sigma^*) + (1-\rho)Q(h^{'*},\sigma^{'*})\,. \tag{27}$$

For instance $\mathcal{P}(\sigma^2) = \rho(\sigma^*)^2 + (1-\rho)(\sigma^{'*})^2$.

A.1 CSBM

For the CSBM, y is taken Rademacher; the self-consistent equations are

$$m_w = \frac{1}{\alpha}\mathbb{E}_{u,\varsigma}\, u w^* \qquad m_\sigma = \mathbb{E}_{y,\xi,\zeta,\chi}\, y\mathcal{P}(\sigma) \tag{28}$$

$$Q_w = \frac{1}{\alpha}\mathbb{E}_{u,\varsigma}(w^*)^2 \qquad Q_\sigma = \mathbb{E}_{y,\xi,\zeta,\chi}\mathcal{P}(\sigma^2) \tag{29}$$

$$V_w = \frac{1}{\alpha}\frac{1}{\sqrt{\hat{Q}_w}}\mathbb{E}_{u,\varsigma}\,\varsigma w^* \qquad V_\sigma = \frac{1}{\sqrt{\hat{Q}_\sigma}}\mathbb{E}_{y,\xi,\zeta,\chi}\,\xi\mathcal{P}(\sigma) \tag{30}$$

$$\hat{m}_w = \frac{\sqrt{\mu}}{V_w}\mathbb{E}_{y,\xi,\zeta,\chi}\, y\mathcal{P}(\sigma - \sqrt{\mu}y m_w) \tag{31}$$

$$\hat{Q}_w = \frac{1}{V_w^2}\mathbb{E}_{y,\xi,\zeta,\chi}\mathcal{P}\left((\sigma - \sqrt{\mu}y m_w - \sqrt{Q}_w\chi)^2\right) \tag{32}$$

$$\hat{V}_w = \frac{1}{V_w}\left(1 - \frac{1}{\sqrt{Q_w}}\mathbb{E}_{y,\xi,\zeta,\chi}\,\chi\mathcal{P}(\sigma)\right) \tag{33}$$

$$\hat{m}_\sigma = \frac{\lambda}{V_\sigma}\mathbb{E}_{y,\xi,\zeta,\chi}\, y\mathcal{P}(h - c\sigma - \lambda y m_\sigma) \tag{34}$$

$$\hat{Q}_\sigma = \frac{1}{V_\sigma^2}\mathbb{E}_{y,\xi,\zeta,\chi}\mathcal{P}\left((h - c\sigma - \lambda y m_\sigma - \sqrt{Q}_\sigma\zeta)^2\right) \tag{35}$$

$$\hat{V}_\sigma = \frac{1}{V_\sigma}\left(1 - \frac{1}{\sqrt{Q_\sigma}}\mathbb{E}_{y,\xi,\zeta,\chi}\,\zeta\mathcal{P}(h - c\sigma)\right)\,. \tag{36}$$

As stated in the main part, once a solution to these equations (28)-(36) is found, the expected errors and accuracies are then expressed as

$$E_{\text{train}} = \mathbb{E}_{y,\xi,\zeta,\chi}\, l(yh^*) \qquad \text{Acc}_{\text{train}} = \mathbb{E}_{y,\xi,\zeta,\chi}\delta_{y=\text{sign}(h^*)} \tag{37}$$

$$E_{\text{test}} = \mathbb{E}_{y,\xi,\zeta,\chi}\, l(yh^{'*}) \qquad \text{Acc}_{\text{test}} = \mathbb{E}_{y,\xi,\zeta,\chi}\delta_{y=\text{sign}(h'^*)}\,. \tag{38}$$

A.2 GLM–SBM

For the GLM–SBM ψ_{out} is taken at $\mu = 0$ and the law of y is

$$P(y=\pm 1|\chi) = \frac{1}{2}\left(1 \pm \text{erf}\left(\frac{m_w\chi}{\sqrt{2(\alpha^{-1}Q_w - m_w^2)}}\right)\right)\,. \tag{39}$$

For compactness we introduce

$$\eta_w = \alpha \frac{m_w^2}{Q_w} \quad \text{and} \quad g(\chi) = \frac{e^{-\frac{\eta_w}{2(1-\eta_w)}\chi^2}}{\sqrt{2\pi\alpha^{-1}(1-\eta_w)}} \,. \tag{40}$$

The self-consistent equations are

$$m_w = \frac{1}{\alpha}\mathbb{E}_{u,\varsigma}\, uw^* \qquad m_\sigma = \mathbb{E}_{\xi,\zeta,\chi}\mathbb{E}_y\, y\mathcal{P}(\sigma) \tag{41}$$

$$Q_w = \frac{1}{\alpha}\mathbb{E}_{u,\varsigma}(w^*)^2 \qquad Q_\sigma = \mathbb{E}_{\xi,\zeta,\chi}\mathbb{E}_y\mathcal{P}(\sigma^2) \tag{42}$$

$$V_w = \frac{1}{\alpha}\frac{1}{\sqrt{\hat{Q}_w}}\mathbb{E}_{u,\varsigma}\,\varsigma w^* \qquad V_\sigma = \frac{1}{\sqrt{\hat{Q}_\sigma}}\mathbb{E}_{\xi,\zeta,\chi}\mathbb{E}_y\,\xi\mathcal{P}(\sigma) \tag{43}$$

$$\hat{m}_w = \frac{1}{V_w}\mathbb{E}_{\xi,\zeta,\chi}\sum_{y=\pm1} yg(\chi)\mathcal{P}(\sigma) \tag{44}$$

$$\hat{Q}_w = \frac{1}{V_w^2}\mathbb{E}_{\xi,\zeta,\chi}\,\mathbb{E}_y\mathcal{P}\left((\sigma - \sqrt{Q}_w\chi)^2\right) \tag{45}$$

$$\hat{V}_w = \frac{1}{V_w}\left(1 - \frac{1}{\sqrt{Q_w}}\mathbb{E}_{\xi,\zeta,\chi}\left(\mathbb{E}_y\,\chi\mathcal{P}(\sigma) - \sum_{y=\pm1}\frac{ym_w}{\sqrt{Q}_w}g(\chi)\mathcal{P}(\sigma)\right)\right) \tag{46}$$

$$\hat{m}_\sigma = \frac{\lambda}{V_\sigma}\mathbb{E}_{\xi,\zeta,\chi}\mathbb{E}_y\, y\mathcal{P}(h - c\sigma - \lambda y m_\sigma) \tag{47}$$

$$\hat{Q}_\sigma = \frac{1}{V_\sigma^2}\mathbb{E}_{\xi,\zeta,\chi}\mathbb{E}_y\mathcal{P}\left((h - c\sigma - \lambda y m_\sigma - \sqrt{Q}_\sigma\zeta)^2\right) \tag{48}$$

$$\hat{V}_\sigma = \frac{1}{V_\sigma}\left(1 - \frac{1}{\sqrt{Q_\sigma}}\mathbb{E}_{\xi,\zeta,\chi}\mathbb{E}_y\,\zeta\mathcal{P}(h - c\sigma)\right) \,. \tag{49}$$

As stated in the main part, once a solution to these equations (41)-(49) is found, the expected errors and accuracies are then expressed as

$$E_{\text{train}} = \mathbb{E}_{y,\xi,\zeta,\chi}l(yh^*) \qquad \text{Acc}_{\text{train}} = \mathbb{E}_{y,\xi,\zeta,\chi}\delta_{y=\text{sign}(h^*)} \tag{50}$$

$$E_{\text{test}} = \mathbb{E}_{y,\xi,\zeta,\chi}l(yh^{'*}) \qquad \text{Acc}_{\text{test}} = \mathbb{E}_{y,\xi,\zeta,\chi}\delta_{y=\text{sign}(h^{'*})} \,. \tag{51}$$

A.3 Solution in the large regularization limit

In this subsection we take $r \to \infty$; we state the solution to eqs. (28)-(36) and (41)-(49) and we give the expression of the test accuracy of the GCN.

The following expressions can be derived considering $l(x) = (1-x)^2/2$ quadratic, without loss of generality, since at large regularization the weights w and the output $h(w)$ of the GCN are small, and l can expanded around 0 as a quadratic potential. As to the regularization γ we take a l_2 regularization, as explained in the main part 4.

CSBM. The test accuracy of the GCN is

$$\text{Acc}_{\text{test}} = \frac{1}{2}\left(1 + \text{erf}\left(\frac{\lambda m_\sigma + cV_w\hat{m}_\sigma + c\sqrt{\mu}m_w}{\sqrt{2}\sqrt{Q_\sigma + c^2V_w^2\hat{Q}_\sigma + c^2Q_w}}\right)\right) \,, \tag{52}$$

the summary statistics being

$$m_w = \frac{\rho}{\alpha r}\sqrt{\mu}(\lambda + c) \qquad V_w = \frac{1}{\alpha r} \qquad Q_w = \frac{\rho}{\alpha r^2}\left(1 + c^2(1-\rho) + \rho(1+\mu)(\lambda+c)^2\right) \tag{53}$$

$$m_\sigma = \frac{\rho}{\alpha r}(1+\mu)(\lambda + c) \qquad V_\sigma = \frac{1}{\alpha r} \qquad Q_\sigma = \frac{\rho}{\alpha^2 r^2}\left((1+\alpha)(1+c^2(1-\rho)) + \rho(1+\mu)(1+\mu+\alpha)(\lambda+c)^2\right) \tag{54}$$

$$\hat{m}_w = \rho\sqrt{\mu}(\lambda + c) \qquad \hat{Q}_w = \rho + \rho(\lambda\rho + c)^2 + (1-\rho)\lambda^2\rho^2 \tag{55}$$

$$\hat{m}_\sigma = \lambda\rho \qquad \hat{Q}_\sigma = \rho \tag{56}$$

GLM–SBM. The test accuracy of the GCN is

$$\begin{aligned}\mathrm{Acc}_{\mathrm{test}} &= \mathbb{E}_\chi \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{1}{\sqrt{2}}\chi\sqrt{\frac{2\alpha}{\pi}}\right)\right)\left(1 + \mathrm{erf}\left(\frac{\lambda m_\sigma + cV_w\hat{m}_\sigma + c\sqrt{Q}_w\chi}{\sqrt{2}\sqrt{Q_\sigma + c^2V_w^2\hat{Q}_\sigma}}\right)\right) \\ &= \int_{>0}\frac{\mathrm{d}z}{\sqrt{2\pi/\alpha}}e^{-\alpha z^2/2}\left(1 + \mathrm{erf}\left(\frac{\lambda m_\sigma + cV_w\hat{m}_\sigma + cm_w\alpha z}{\sqrt{2}\sqrt{Q_\sigma + c^2V_w^2\hat{Q}_\sigma + c^2(Q_w - \alpha m_w^2)}}\right)\right),\end{aligned} \tag{57}$$

the summary statistics being

$$m_w = \frac{\rho}{\alpha r}\sqrt{\frac{2\alpha}{\pi}}(\lambda + c) \qquad V_w = \frac{1}{\alpha r} \qquad Q_w = \frac{\rho}{\alpha r^2}\left(1 + c^2(1-\rho) + \rho(1+2\alpha/\pi)(\lambda+c)^2\right) \tag{58}$$

$$m_\sigma = \frac{\rho}{\alpha r}(1+2\alpha/\pi)(\lambda + c) \qquad V_\sigma = \frac{1}{\alpha r} \qquad Q_\sigma = \frac{\rho}{\alpha^2 r^2}\left((1+\alpha)(1+c^2(1-\rho)) + \rho((1+2\alpha/\pi)(1+\alpha) + 2\alpha/\pi)(\lambda+c)^2\right) \tag{59}$$

$$\hat{m}_w = \rho\sqrt{\frac{2\alpha}{\pi}}(\lambda + c) \qquad \hat{Q}_w = \rho + \rho(\lambda\rho + c)^2 + (1-\rho)\lambda^2\rho^2 \tag{60}$$

$$\hat{m}_\sigma = \lambda\rho \qquad \hat{Q}_\sigma = \rho \tag{61}$$

In the limit $\lambda \to \infty$ the maximizer c^* of (57) is

$$c^* = \frac{1}{\lambda}\operatorname*{argmin}_{\tilde{c}} e^{-2b\tau^\infty_{\mathrm{GLM-SBM}} + a^2\tau^\infty_{\mathrm{GLM-SBM}}}\left(1 - \mathrm{erf}\left(\sqrt{2}a\tau^\infty_{\mathrm{GLM-SBM}}\right)\right) \tag{62}$$

$$a = \sqrt{\alpha}\tilde{c}\frac{\sqrt{2\alpha/\pi}}{1+2\alpha/\pi}, \qquad b = \frac{\tilde{c}}{1+2\alpha/\pi} - \frac{1}{2}\frac{\alpha\tilde{c}^2 + (1+\alpha)/\rho}{(1+\alpha)(1+2\alpha/\pi) + 2\alpha/\pi} \tag{63}$$

$$\tau^\infty_{\mathrm{GLM-SBM}} = \frac{1+2\alpha/\pi}{2\left(1 + \alpha + \frac{2\alpha/\pi}{1+2\alpha/\pi}\right)} \tag{64}$$

B Replica computation

In this appendix, we derive the main equations given in section 3 of the main text and in appendix A.

B.1 Gaussian equivalence

To average over the adjacency matrix $\tilde{A}$ we rely on a Gaussian equivalence property. It states that the rescaled adjacency matrix $\tilde{A}_{ij} = \left(\frac{d}{N}\left(1 - \frac{d}{N}\right)\right)^{-1/2}\left(A_{ij} - \frac{d}{N}\right)$ can be approximated by the rank-one plus noise matrix $A^{\mathrm{g}} = \frac{\lambda}{\sqrt{N}}yy^T + \Xi$, where Ξ are random fluctuations whose components Ξ_{ij} are independent standard Gaussians, without changing the expected losses and accuracies of the model, in the limit of large average degree d. It has been stated in [32] for the SBM, proved in [21] as to the mutual information and tested in [1] for the GCN for $d = \Theta(\sqrt{N})$. In practice taking $d \gtrsim 20$

at $N = 1000$ is enough to observe no difference for the losses and accuracies and assuming $d = \omega(1)$ should be sufficient.

For the equivalence property to hold, the GCN has to compute the convolution over $\tilde{A}$. The constant shift by d/N can be interpreted as centering A while the constant scaling by $\left(\frac{d}{N}\left(1-\frac{d}{N}\right)\right)^{-1/2}$ normalizes its variance. The convolution over $\tilde{A}$ can still be interpreted as a graph convolution. The scaling can be absorbed in w and r; and if the graph is not too dense $d = o(N)$ the shift is negligeable.

B.2 CSBM

We first derive the results for the CSBM, generalizing the results of [1] to arbitrary convex loss and regularization. As stated in the main part 3, we introduce n replica:

$$Z = \int \prod_{\nu}^{M} \mathrm{d}w_\nu P_W(w_\nu) e^{-\beta t \sum_{i\in R} l(y_i h(w)_i) - \beta t' \sum_{i\in R'} l(y_i h(w)_i)} \tag{65}$$

$$-\beta N f = \mathbb{E}_{u,\Xi,W,y} \log Z = \mathbb{E}_{u,\Xi,W,y} \frac{\partial}{\partial n} Z^n (n=0) \tag{66}$$

$$= \frac{\partial}{\partial n}(n=0)\, \underbrace{\mathbb{E}_{u,\Xi,W,y} \int \prod_{a}^{n} \prod_{\nu}^{M} \mathrm{d}w_\nu^a P_W(w_\nu^a) e^{\sum_a^n -\beta t \sum_{i\in R} l(y_i h(w^a)_i) - \beta t' \sum_{i\in R'} l(y_i h(w^a)_i)}}_{*} \tag{67}$$

where $P_W(w) = \exp(-\beta r \gamma(w))$ is the prior on the weights induced by the regularization. We introduce several ancillary variables via δ-Dirac functions to decouple the random variables. We set $h = \frac{1}{\sqrt{N}}(A^{\mathrm{g}} + c\sqrt{N} I_N)\sigma$ and $\sigma = \frac{1}{\sqrt{N}} X w$. Then we take the expectation on the Gaussian noise:

$$* \propto \mathbb{E}_{u,\Xi,W,y} \int \prod_{a,\nu} \mathrm{d}w_\nu^a P_W(w_\nu^a) \prod_{a,i} \mathrm{d}h_i^a \mathrm{d}q_i^a e^{-\beta t \sum_{a,i\in R} l(y_i h_i^a) - \beta t' \sum_{a,i\in R'} l(y_i h_i^a) + \sum_{a,i} \mathrm{i} q_i^a (h_i^a - h(w^a)_i)} \tag{68}$$

$$= \mathbb{E}_{u,\Xi,W,y} \int \prod_{a,\nu} \mathrm{d}w_\nu^a P_W(w_\nu^a) \prod_{a,i} \mathrm{d}h_i^a \mathrm{d}q_i^a \mathrm{d}\sigma_i^a \mathrm{d}\hat{q}_i^a e^{-\beta t \sum_{a,i\in R} l(y_i h_i^a) - \beta t' \sum_{a,i\in R'} l(y_i h_i^a)}$$
$$e^{\sum_{a,i} \mathrm{i} q_i^a \left(h_i^a - \frac{1}{\sqrt{N}} \sum_j \left(c\sqrt{N}\delta_{i,j} + \frac{\lambda}{\sqrt{N}} y_i y_j + \Xi_{ij}\right)\sigma_j^a\right) + \sum_{a,i} \mathrm{i}\hat{q}_i^a \left(\sigma_i^a - \frac{1}{\sqrt{N}} \sum_\nu \left(\sqrt{\frac{\mu}{N}} y_j u_\nu + W_{j\nu}\right) w_\nu^a\right)} \tag{69}$$

$$= \mathbb{E}_{u,y} \int \prod_{a,\nu} \mathrm{d}w_\nu^a P_W(w_\nu^a) \prod_{a,i} \mathrm{d}h_i^a \mathrm{d}q_i^a \mathrm{d}\sigma_i^a \mathrm{d}\hat{q}_i^a e^{-\beta t \sum_{a,i\in R} l(y_i h_i^a) - \beta t' \sum_{a,i\in R'} l(y_i h_i^a) + \mathrm{i} \sum_{a,i} (q_i^a h_i^a + \hat{q}_i^a \sigma_i^a)}$$
$$e^{-\mathrm{i} \sum_{a,i} \left(c q_i^a \sigma_i^a + \frac{\lambda}{N} y_i q_i^a \sum_j y_j \sigma_j^a\right) - \frac{1}{2N} \sum_{i,j,a,b} q_i^a q_i^b \sigma_j^a \sigma_j^b - \mathrm{i} \sum_{a,i} \frac{\sqrt{\mu}}{N} y_i \hat{q}_i^a \sum_\nu u_\nu w_\nu^a - \frac{1}{2N} \sum_{i,\nu,a,b} \hat{q}_i^a \hat{q}_i^b w_\nu^a w_\nu^b} . \tag{70}$$

We integrate over the qs and $\hat{q}$s. For simplicity we pack the replica into vectors of size n.

$$* = \mathbb{E}_{u,y} \int \prod_{a,\nu} \mathrm{d}w_\nu^a P_W(w_\nu^a) \prod_{a,i} \mathrm{d}h_i^a \mathrm{d}q_i^a \mathrm{d}\sigma_i^a \mathrm{d}\hat{q}_i^a e^{-\beta t \sum_{a,i\in R} l(y_i h_i^a) - \beta t' \sum_{a,i\in R'} l(y_i h_i^a) + \mathrm{i} \sum_i q_i^T \left(h_i - c\sigma_i - \frac{\lambda}{N} y_i \sum_j y_j \sigma_j\right)}$$
$$e^{-\frac{1}{2N} \sum_i q_i^T \left(\sum_j \sigma_j \sigma_j^T\right) q_i + \mathrm{i} \sum_i \hat{q}_i^T \left(\sigma_i - \frac{\sqrt{\mu}}{N} y_i \sum_\nu u_\nu w_\nu\right) - \frac{1}{2N} \sum_i \hat{q}_i^T \left(\sum_\nu w_\nu w_\nu^T\right) \hat{q}_i} \tag{71}$$

$$= \mathbb{E}_{u,y} \int \prod_{a,\nu} \mathrm{d}w_\nu^a P_W(w_\nu^a) \prod_{a,i} \mathrm{d}h_i^a \mathrm{d}\sigma_i^a e^{-\beta t \sum_{a,i\in R} l(y_i h_i^a) - \beta t' \sum_{a,i\in R'} l(y_i h_i^a)}$$
$$\prod_i \mathcal{N}\left(h_i \middle| c\sigma_i + \frac{\lambda}{N} y_i \sum_j y_j \sigma_j, \frac{1}{N} \sum_j \sigma_j \sigma_j^T\right) \mathcal{N}\left(\sigma_i \middle| \frac{\sqrt{\mu}}{N} y_i \sum_\nu u_\nu w_\nu, \frac{1}{N} \sum_\nu w_\nu w_\nu^T\right) ; \tag{72}$$

where $\mathcal{N}(x|\mu,\Sigma) = \det(2\pi\Sigma)^{-1/2} e^{-(x-\mu)^T\Sigma^{-1}(x-\mu)/2}$ is a Gaussian density. The order parameters are

$$m_w^a = \frac{1}{N}\sum_\nu u_\nu w_\nu^a \qquad m_\sigma^a = \frac{1}{N}\sum_i y_i \sigma_i^a \tag{73}$$

$$Q_w^{ab} = \frac{1}{N}\sum_\nu w_\nu^a w_\nu^b \qquad Q_\sigma^{ab} = \frac{1}{N}\sum_i \sigma_i^a \sigma_i^b \tag{74}$$

We introduce them via new δ-Dirac functions. We can factorize the i and ν indices.

$$\begin{aligned}
* \propto \mathbb{E}_{u,y} \int \prod_{a,\nu} \mathrm{d}w_\nu^a P_W(w_\nu^a) \prod_{a,i} \mathrm{d}h_i^a \mathrm{d}\sigma_i^a \prod_{a\le b} \mathrm{d}\hat{Q}_w^{ab}\mathrm{d}Q_w^{ab}\mathrm{d}\hat{Q}_\sigma^{ab}\mathrm{d}Q_\sigma^{ab} \prod_a \mathrm{d}\hat{m}_w^a \mathrm{d}m_w^a \mathrm{d}\hat{m}_\sigma^a \mathrm{d}m_\sigma^a \quad (75)\\
e^{-\beta t \sum_{a,i\in R} l(y_i h_i^a) - \beta t' \sum_{a,i\in R'} l(y_i h_i^a)} \prod_{a\le b} e^{\hat{Q}_w^{ab}\left(NQ_w^{ab} - \sum_\nu w_\nu^a w_\nu^b\right) + \hat{Q}_\sigma^{ab}\left(NQ_\sigma^{ab} - \sum_i \sigma_i^a\sigma_i^b\right)}\\
\prod_a e^{\hat{m}_w^a(Nm_w^a - \sum_\nu u_\nu w_\nu^a) + \hat{m}_\sigma^a(Nm_\sigma^a - \sum_i y_i\sigma_i^a)} \prod_i N\left(h_i\,|c\sigma_i + \lambda y_i m_\sigma, Q_\sigma\right) \mathcal{N}\left(\sigma_i\,|\sqrt{\mu} y_i m_w, Q_w\right)\\
= \int \prod_{a\le b} \mathrm{d}\hat{Q}_w^{ab}\mathrm{d}Q_w^{ab}\mathrm{d}\hat{Q}_\sigma^{ab}\mathrm{d}Q_\sigma^{ab} \prod_a \mathrm{d}\hat{m}_w^a \mathrm{d}m_w^a \mathrm{d}\hat{m}_\sigma^a \mathrm{d}m_\sigma^a \prod_{a\le b} e^{N(\hat{Q}_w^{ab}Q_w^{ab} + \hat{Q}_\sigma^{ab}Q_\sigma^{ab})} \prod_a e^{N(\hat{m}_w^a m_w^a + \hat{m}_\sigma^a m_\sigma^a)}\\
\left[\mathbb{E}_u \int \prod_a \mathrm{d}w^a e^{\psi_w^{(n)}(w)}\right]^{N/\alpha} \left[\mathbb{E}_y \int \prod_a \mathrm{d}h^a \mathrm{d}\sigma^a e^{\psi_{\mathrm{out}}^{(n)}(h,\sigma;t)}\right]^{\rho N} \quad (76)\\
\left[\mathbb{E}_y \int \prod_a \mathrm{d}h^a \mathrm{d}\sigma^a e^{\psi_{\mathrm{out}}^{(n)}(h,\sigma;t')}\right]^{\rho' N} \left[\mathbb{E}_y \int \prod_a \mathrm{d}h^a \mathrm{d}\sigma^a e^{\psi_{\mathrm{out}}^{(n)}(h,\sigma;0)}\right]^{(1-\rho-\rho')N} ;
\end{aligned}$$

where we defined

$$\psi_w^{(n)}(w) = \sum_a \log P_W(w^a) - \sum_{a\le b} \hat{Q}_w^{ab} w^a w^b - \sum_a \hat{m}_w^a u w^a \tag{77}$$

$$\begin{aligned}
\psi_{\mathrm{out}}^{(n)}(h,\sigma;\bar{t}) = -\beta\bar{t}\sum_a l(yh^a) - \sum_{a\le b}\hat{Q}_\sigma^{ab}\sigma^a\sigma^b - \sum_a \hat{m}_\sigma^a y\sigma^a - \frac{1}{2}(h - c\sigma - \lambda y m_\sigma)^T Q_\sigma^{-1}(h - c\sigma - \lambda y m_\sigma)\\
- \frac{1}{2}\log\det Q_\sigma - \frac{1}{2}(\sigma - \sqrt{\mu} y m_w)^T Q_w^{-1}(\sigma - \sqrt{\mu} y m_w) - \frac{1}{2}\log\det Q_w \,.
\end{aligned} \tag{78}$$

We use the replica-symmetric ansatz: we set $\hat{Q}^{aa} = \frac{1}{2}\hat{R}$, $\hat{Q}^{ab} = -\hat{Q}$, $Q^{aa} = R$, $Q^{ab} = Q$, $\hat{m}^a = -\hat{m}$ and $m^a = m$. Since we will take the derivative wrt n and send n to zero we discard all the terms that are not proportionnal to n. We compute first that

$$Q^{-1} = \frac{1}{R-Q} I_n - \frac{Q}{(R-Q)^2} J_{n,n} + o(n) \tag{79}$$

$$\log\det Q = n\frac{Q}{R-Q} + n\log(R-Q) + o(n) \,; \tag{80}$$

where $J_{n,n}$ is the matrix filled with ones. We define the variances $V = R - Q$ and $\hat{V} = \hat{R} + \hat{Q}$. We introduce scalar Gaussian random variables ς, ξ, ζ and χ to decouple the replica and factorize them. Then

$$\begin{aligned}
* \propto \int \mathrm{d}\hat{Q}_w \mathrm{d}\hat{V}_w \mathrm{d}Q_w \mathrm{d}V_w \mathrm{d}\hat{Q}_\sigma \mathrm{d}\hat{V}_\sigma \mathrm{d}Q_\sigma \mathrm{d}V_\sigma \mathrm{d}\hat{m}_w \mathrm{d}m_w \mathrm{d}\hat{m}_\sigma \mathrm{d}m_\sigma e^{\frac{nN}{2}(\hat{V}_w V_w + \hat{V}_w Q_w - V_w\hat{Q}_w + \hat{V}_\sigma V_\sigma + \hat{V}_\sigma Q_\sigma - V_\sigma \hat{Q}_\sigma)}\\
e^{-nN(\hat{m}_w m_w + \hat{m}_\sigma m_\sigma)} \left[\mathbb{E}_{u,\varsigma}\left(\int \mathrm{d}w\, e^{\psi_w(w)}\right)^n\right]^{N/\alpha} \left[\mathbb{E}_{y,\xi,\zeta,\chi}\left(\int \mathrm{d}h\mathrm{d}\sigma e^{\psi_{\mathrm{out}}(h,\sigma;t)}\right)^n\right]^{\rho N}\\
\left[\mathbb{E}_{y,\xi,\zeta,\chi}\left(\int \mathrm{d}h\mathrm{d}\sigma e^{\psi_{\mathrm{out}}(h,\sigma;t')}\right)^n\right]^{\rho' N} \left[\mathbb{E}_{y,\xi,\zeta,\chi}\left(\int \mathrm{d}h\mathrm{d}\sigma e^{\psi_{\mathrm{out}}(h,\sigma;0)}\right)^n\right]^{(1-\rho-\rho')N}
\end{aligned} \tag{81}$$

$$:= \int \mathrm{d}m\,\mathrm{d}q\,\mathrm{d}v\, e^{N\phi^{(n)}(m,q,v)} \,, \tag{82}$$

with

$$\psi_w(w) = \log P_W(w) - \frac{1}{2}\hat{V}_w w^2 + \left(\varsigma\sqrt{\hat{Q}_w} + u\hat{m}_w\right) w \tag{83}$$

$$\psi_{\rm out}(h,\sigma;\bar{t}) = -\beta\bar{t}l(yh) - \frac{1}{2}\hat{V}_\sigma\sigma^2 + \left(\xi\sqrt{\hat{Q}_\sigma} + y\hat{m}_\sigma\right)\sigma \tag{84}$$

$$+ \log\mathcal{N}\left(h|c\sigma + \lambda y m_\sigma + \sqrt{Q_\sigma}\zeta, V_\sigma\right) + \log\mathcal{N}\left(\sigma|\sqrt{\mu}y m_w + \sqrt{Q_w}\chi, V_w\right)$$

and m, q and v standing for all the order parameters. ξ, ζ and χ are scalar standard Gaussians. We take the limit $N\to\infty$ thanks to Laplace's method.

$$-\beta f \propto \frac{1}{N}\frac{\partial}{\partial n}(n=0)\int {\rm d}m\,{\rm d}q\,{\rm d}v\, e^{N\phi^{(n)}(m,q,v)} \tag{85}$$

$$= \underset{m,q,v}{\rm extr}\,\frac{\partial}{\partial n}(n=0)\phi^{(n)}(m,q,v) := \underset{m,q,v}{\rm extr}\,\phi(m,q,v)\;; \tag{86}$$

the free entropy of the problem is

$$\begin{aligned}\phi = &\frac{1}{2}\left(\hat{V}_w V_w + \hat{V}_w Q_w - V_w\hat{Q}_w + \hat{V}_\sigma V_\sigma + \hat{V}_\sigma Q_\sigma - V_\sigma\hat{Q}_\sigma\right) - \hat{m}_w m_w - \hat{m}_\sigma m_\sigma \\ &+ \frac{1}{\alpha}\mathbb{E}_{u,\varsigma}\left(\log\int {\rm d}w\, e^{\psi_w(w)}\right) + \rho\mathbb{E}_{y,\xi,\zeta,\chi}\left(\log\int {\rm d}h{\rm d}\sigma e^{\psi_{\rm out}(h,\sigma;t)}\right) \\ &+ \rho'\mathbb{E}_{y,\xi,\zeta,\chi}\left(\log\int {\rm d}h{\rm d}\sigma e^{\psi_{\rm out}(h,\sigma;t')}\right) + (1-\rho-\rho')\mathbb{E}_{y,\xi,\zeta,\chi}\left(\log\int {\rm d}h{\rm d}\sigma e^{\psi_{\rm out}(h,\sigma;0)}\right)\;.\end{aligned} \tag{87}$$

We take the extremum of the free entropy deriving it wrt the order parameters, evaluated at $t=1$ and $t'=0$. We obtain the following fixed-point conditions.

$$m_w = \frac{1}{\alpha}\mathbb{E}_{u,\varsigma}\, u\,\mathbb{E}_{P_w} w \qquad m_\sigma = \mathbb{E}_{y,\xi,\zeta,\chi}\, y\left(\rho\mathbb{E}_{P_{\rm out}}\sigma + (1-\rho)\mathbb{E}_{P'_{\rm out}}\sigma\right) \tag{88}$$

$$Q_w + V_w = \frac{1}{\alpha}\mathbb{E}_{u,\varsigma}\mathbb{E}_{P_w} w^2 \qquad Q_\sigma + V_\sigma = \mathbb{E}_{y,\xi,\zeta,\chi}\left(\rho\mathbb{E}_{P_{\rm out}}\sigma^2 + (1-\rho)\mathbb{E}_{P'_{\rm out}}\sigma^2\right) \tag{89}$$

$$V_w = \frac{1}{\alpha}\frac{1}{\sqrt{\hat{Q}_w}}\mathbb{E}_{u,\varsigma}\,\varsigma\,\mathbb{E}_{P_w} w \qquad V_\sigma = \frac{1}{\sqrt{\hat{Q}_\sigma}}\mathbb{E}_{y,\xi,\zeta,\chi}\,\xi\left(\rho\mathbb{E}_{P_{\rm out}}\sigma + (1-\rho)\mathbb{E}_{P'_{\rm out}}\sigma\right) \tag{90}$$

$$\hat{m}_w = \frac{\sqrt{\mu}}{V_w}\mathbb{E}_{y,\xi,\zeta,\chi}\, y\left(\rho\mathbb{E}_{P_{\rm out}}(\sigma - \sqrt{\mu}y m_w) + (1-\rho)\mathbb{E}_{P'_{\rm out}}(\sigma - \sqrt{\mu}y m_w)\right) \tag{91}$$

$$\hat{Q}_w - \hat{V}_w = \frac{1}{V_w^2}\mathbb{E}_{y,\xi,\zeta,\chi}\left(\rho\mathbb{E}_{P_{\rm out}}(\sigma - \sqrt{\mu}y m_w - \sqrt{Q}_w\chi)^2 + (1-\rho)\mathbb{E}_{P'_{\rm out}}(\sigma - \sqrt{\mu}y m_w - \sqrt{Q}_w\chi)^2\right) - \frac{1}{V_w} \tag{92}$$

$$\hat{V}_w = \frac{1}{V_w}\left(1 - \frac{1}{\sqrt{Q_w}}\mathbb{E}_{y,\xi,\zeta,\chi}\,\chi\left(\rho\mathbb{E}_{P_{\rm out}}\sigma + (1-\rho)\mathbb{E}_{P'_{\rm out}}\sigma\right)\right) \tag{93}$$

$$\hat{m}_\sigma = \frac{\lambda}{V_\sigma}\mathbb{E}_{y,\xi,\zeta,\chi}\, y\left(\rho\mathbb{E}_{P_{\rm out}}(h - c\sigma - \lambda y m_\sigma) + (1-\rho)\mathbb{E}_{P'_{\rm out}}(h - c\sigma - \lambda y m_\sigma)\right) \tag{94}$$

$$\hat{Q}_\sigma - \hat{V}_\sigma = \frac{1}{V_\sigma^2}\mathbb{E}_{y,\xi,\zeta,\chi}\left(\rho\mathbb{E}_{P_{\rm out}}(h - c\sigma - \lambda y m_\sigma - \sqrt{Q}_\sigma\zeta)^2 + (1-\rho)\mathbb{E}_{P'_{\rm out}}(h - c\sigma - \lambda y m_\sigma - \sqrt{Q}_\sigma\zeta)^2\right) - \frac{1}{V_\sigma} \tag{95}$$

$$\hat{V}_\sigma = \frac{1}{V_\sigma}\left(1 - \frac{1}{\sqrt{Q_\sigma}}\mathbb{E}_{y,\xi,\zeta,\chi}\,\zeta\left(\rho\mathbb{E}_{P_{\rm out}}(h - c\sigma) + (1-\rho)\mathbb{E}_{P'_{\rm out}}(h - c\sigma)\right)\right) \tag{96}$$

The measures are

$$ {\rm d}P_w = \frac{{\rm d}w\, e^{\psi_w(w)}}{\int {\rm d}w\, e^{\psi_w(w)}} \quad , \quad {\rm d}P_{\rm out} = \frac{{\rm d}h{\rm d}\sigma\, e^{\psi_{\rm out}(h,\sigma;\bar{t}=1)}}{\int {\rm d}h{\rm d}\sigma\, e^{\psi_{\rm out}(h,\sigma,\bar{t}=1)}} \quad , \quad {\rm d}P'_{\rm out} = \frac{{\rm d}h{\rm d}\sigma\, e^{\psi_{\rm out}(h,\sigma;\bar{t}=0)}}{\int {\rm d}h{\rm d}\sigma\, e^{\psi_{\rm out}(h,\sigma,\bar{t}=0)}}\;. \tag{97}$$

These measures can be computed thanks to Laplace's method in the limit $\beta \to \infty$. We have to rescale the order parameters not to obtain a degenerated solution. We recall that $\log P_W(w) \propto \beta$. We take $\hat{V} \to \beta\hat{V}$, $\hat{Q} \to \beta^2\hat{Q}$, $\hat{m} \to \beta\hat{m}$ and $V \to \beta^{-1}V$ for both w and σ. We define

$$w^* = \underset{w}{\operatorname{argmax}}\, \psi_w(w) \tag{98}$$

$$(h^*, \sigma^*) = \underset{h,\sigma}{\operatorname{argmax}}\, \psi_{\text{out}}(h, \sigma; \bar{t} = 1) \qquad (h^{'*}, \sigma^{'*}) = \underset{h,\sigma}{\operatorname{argmax}}\, \psi_{\text{out}}(h, \sigma; \bar{t} = 0)\ ; \tag{99}$$

then, keeping the first order in β in both lhs and rhs, the fixed-point equations are

$$m_w = \frac{1}{\alpha}\mathbb{E}_{u,\varsigma}\, uw^* \qquad m_\sigma = \mathbb{E}_{y,\xi,\zeta,\chi}\, y\left(\rho\sigma^* + (1-\rho)\sigma^{'*}\right) \tag{100}$$

$$Q_w = \frac{1}{\alpha}\mathbb{E}_{u,\varsigma}(w^*)^2 \qquad Q_\sigma = \mathbb{E}_{y,\xi,\zeta,\chi}\left(\rho(\sigma^*)^2 + (1-\rho)(\sigma^{'*})^2\right) \tag{101}$$

$$V_w = \frac{1}{\alpha}\frac{1}{\sqrt{\hat{Q}_w}}\mathbb{E}_{u,\varsigma}\, \varsigma w^* \qquad V_\sigma = \frac{1}{\sqrt{\hat{Q}_\sigma}}\mathbb{E}_{y,\xi,\zeta,\chi}\, \xi\left(\rho\sigma^* + (1-\rho)\sigma^{'*}\right) \tag{102}$$

$$\hat{m}_w = \frac{\sqrt{\mu}}{V_w}\mathbb{E}_{y,\xi,\zeta,\chi}\, y\left(\rho(\sigma^* - \sqrt{\mu}ym_w) + (1-\rho)(\sigma^{'*} - \sqrt{\mu}ym_w)\right) \tag{103}$$

$$\hat{Q}_w = \frac{1}{V_w^2}\mathbb{E}_{y,\xi,\zeta,\chi}\left(\rho(\sigma^* - \sqrt{\mu}ym_w - \sqrt{Q}_w\chi)^2 + (1-\rho)(\sigma^{'*} - \sqrt{\mu}ym_w - \sqrt{Q}_w\chi)^2\right) \tag{104}$$

$$\hat{V}_w = \frac{1}{V_w}\left(1 - \frac{1}{\sqrt{Q_w}}\mathbb{E}_{y,\xi,\zeta,\chi}\, \chi\left(\rho\sigma^* + (1-\rho)\sigma^{'*}\right)\right) \tag{105}$$

$$\hat{m}_\sigma = \frac{\lambda}{V_\sigma}\mathbb{E}_{y,\xi,\zeta,\chi}\, y\left(\rho(h^* - c\sigma^* - \lambda ym_\sigma) + (1-\rho)(h^{'*} - c\sigma^{'*} - \lambda ym_\sigma)\right) \tag{106}$$

$$\hat{Q}_\sigma = \frac{1}{V_\sigma^2}\mathbb{E}_{y,\xi,\zeta,\chi}\left(\rho(h^* - c\sigma^* - \lambda ym_\sigma - \sqrt{Q}_\sigma\zeta)^2 + (1-\rho)(h^{'*} - c\sigma^{'*} - \lambda ym_\sigma - \sqrt{Q}_\sigma\zeta)^2\right) \tag{107}$$

$$\hat{V}_\sigma = \frac{1}{V_\sigma}\left(1 - \frac{1}{\sqrt{Q_\sigma}}\mathbb{E}_{y,\xi,\zeta,\chi}\, \zeta\left(\rho(h^* - c\sigma^*) + (1-\rho)(h^{'*} - c\sigma^{'*})\right)\right) \tag{108}$$

The average train and test losses can be computed by deriving ϕ wrt t and t' and taking it extremum by evaluating it at the fixed-point of these equations. Simplifying the notations we obtain the equations given in part A.

B.3 GLM–SBM

We derive the results for the GLM–SBM, which has not been studied by [1]. The derivation is similar to the derivation of the previous part on the CSBM. As we saw for the CSBM, one can readily take the test set R' being the complement of R i.e. $\rho' = 1 - \rho$; the resulting equations do not change. As stated in the main part 3, we introduce n replica:

$$Z = \int \prod_\nu^M \mathrm{d}w_\nu P_W(w_\nu) \prod_i^N \mathrm{d}y_i P_o\left(y_i | \frac{1}{\sqrt{N}} X_i^T u\right) e^{-\beta t \sum_{i\in R} l(y_i h(w)_i) - \beta t' \sum_{i\in R'} l(y_i h(w)_i)} \tag{109}$$

$$-\beta N f = \mathbb{E}_{u,\Xi,X} \log Z = \mathbb{E}_{u,\Xi,X} \frac{\partial}{\partial n} Z^n(n=0) = \frac{\partial}{\partial n}(n=0) \tag{110}$$

$$\underbrace{\mathbb{E}_{u,\Xi,X} \int \prod_a^n \prod_\nu^M \mathrm{d}w_\nu^a P_W(w_\nu^a) \prod_i^N \mathrm{d}y_i P_o\left(y_i | \frac{1}{\sqrt{N}} X_i^T u\right) e^{\sum_a^n -\beta t \sum_{i\in R} l(y_i h(w^a)_i) - \beta t' \sum_{i\in R'} l(y_i h(w^a)_i)}}_{*}$$

where $P_o(y|z) = \delta_{y=\mathrm{sign}(z)}$. We introduce ancillary variables: $h = \frac{1}{\sqrt{N}}(A^{\mathrm{g}} + c\sqrt{N}I_N)\sigma$, $\sigma = \frac{1}{\sqrt{N}}Xw$ and $z = \frac{1}{\sqrt{N}}Xu$; we average over Ξ and X, pack the replica and integrate.

$$* \propto \mathbb{E}_{u,\Xi,X} \int \prod_{a,\nu} \mathrm{d}w^a_\nu P_W(w^a_\nu) \prod_i \mathrm{d}y_i P_o(y_i|z_i)\mathrm{d}z_i \mathrm{d}\bar{q}_i \prod_{a,i} \mathrm{d}h^a_i \mathrm{d}q^a_i \mathrm{d}\sigma^a_i \mathrm{d}\hat{q}^a_i e^{-\beta t \sum_{a,i\in R} l(y_i h^a_i) - \beta t' \sum_{a,i\in R'} l(y_i h^a_i)} \tag{111}$$

$$e^{\sum_i \mathrm{i}\bar{q}_i \left(z_i - \frac{1}{\sqrt{N}}\sum_\nu X_{i\nu}u_\nu\right) + \sum_{a,i} \mathrm{i}q^a_i \left(h^a_i - \frac{1}{\sqrt{N}}\sum_j \left(c\sqrt{N}\delta_{i,j} + \frac{\lambda}{\sqrt{N}}y_i y_j + \Xi_{ij}\right)\sigma^a_j\right) + \sum_{a,i} \mathrm{i}\hat{q}^a_i \left(\sigma^a_i - \frac{1}{\sqrt{N}}\sum_\nu X_{i\nu}w^a_\nu\right)}$$

$$= \mathbb{E}_u \int \prod_{a,\nu} \mathrm{d}w^a_\nu P_W(w^a_\nu) \prod_i \mathrm{d}y_i P_o(y_i|z_i)\mathrm{d}z_i \prod_{a,i} \mathrm{d}h^a_i \mathrm{d}\sigma^a_i e^{-\beta t \sum_{a,i\in R} l(y_i h^a_i) - \beta t' \sum_{a,i\in R'} l(y_i h^a_i)} \tag{112}$$

$$\prod_i \mathcal{N}\left(h_i \middle| c\sigma_i + \frac{\lambda}{N} y_i \sum_j y_j \sigma_j, \frac{1}{N}\sum_j \sigma_j \sigma_j^T\right) \prod_i \mathcal{N}\left(\begin{pmatrix} z_i \\ \sigma_i \end{pmatrix} \middle| 0, \frac{1}{N}\sum_\nu \begin{pmatrix} u_\nu \\ w_\nu \end{pmatrix}\begin{pmatrix} u_\nu \\ w_\nu \end{pmatrix}^T\right) .$$

Here $\begin{pmatrix} z_i \\ \sigma_i \end{pmatrix}$ and $\begin{pmatrix} u_\nu \\ w_\nu \end{pmatrix}$ are vectors of size $n+1$. $\frac{1}{N}\sum_\nu u^2_\nu$ self-averages to $\rho_u := \frac{1}{\alpha}\mathbb{E}_u u^2 = \frac{1}{\alpha}$. As for the CSBM the order parameters are

$$m^a_w = \frac{1}{N}\sum_\nu u_\nu w^a_\nu \qquad m^a_\sigma = \frac{1}{N}\sum_i y_i \sigma^a_i \tag{113}$$

$$Q^{ab}_w = \frac{1}{N}\sum_\nu w^a_\nu w^b_\nu \qquad Q^{ab}_\sigma = \frac{1}{N}\sum_i \sigma^a_i \sigma^b_i \tag{114}$$

We introduce them via new δ-Dirac functions:

$$* \propto \mathbb{E}_u \int \prod_{a,\nu} \mathrm{d}w^a_\nu P_W(w^a_\nu) \prod_i \mathrm{d}y_i P_o(y_i|z_i)\mathrm{d}z_i \prod_{a,i} \mathrm{d}h^a_i \mathrm{d}\sigma^a_i \prod_{a\le b} \mathrm{d}\hat{Q}^{ab}_w \mathrm{d}Q^{ab}_w \mathrm{d}\hat{Q}^{ab}_\sigma \mathrm{d}Q^{ab}_\sigma \prod_a \mathrm{d}\hat{m}^a_w \mathrm{d}m^a_w \mathrm{d}\hat{m}^a_\sigma \mathrm{d}m^a_\sigma \tag{115}$$

$$\prod_{a\le b} e^{\hat{Q}^{ab}_w\left(NQ^{ab}_w - \sum_\nu w^a_\nu w^b_\nu\right) + \hat{Q}^{ab}_\sigma\left(NQ^{ab}_\sigma - \sum_i \sigma^a_i \sigma^b_i\right)} \prod_a e^{\hat{m}^a_w(Nm^a_w - \sum_\nu u_\nu w^a_\nu) + \hat{m}^a_\sigma(Nm^a_\sigma - \sum_i y_i \sigma^a_i)}$$

$$e^{-\beta t \sum_{a,i\in R} l(y_i h^a_i) - \beta t' \sum_{a,i\in R'} l(y_i h^a_i)} \prod_i N\left(h_i \,|c\sigma_i + \lambda y_i m_\sigma, Q_\sigma\right) \mathcal{N}\left(\begin{pmatrix} z_i \\ \sigma_i \end{pmatrix} \middle| 0, \begin{pmatrix} \rho_u & m^T_w \\ m_w & Q_w \end{pmatrix}\right)$$

$$= \int \prod_{a\le b} \mathrm{d}\hat{Q}^{ab}_w \mathrm{d}Q^{ab}_w \mathrm{d}\hat{Q}^{ab}_\sigma \mathrm{d}Q^{ab}_\sigma \prod_a \mathrm{d}\hat{m}^a_w \mathrm{d}m^a_w \mathrm{d}\hat{m}^a_\sigma \mathrm{d}m^a_\sigma \prod_{a\le b} e^{N(\hat{Q}^{ab}_w Q^{ab}_w + \hat{Q}^{ab}_\sigma Q^{ab}_\sigma)} \prod_a e^{N(\hat{m}^a_w m^a_w + \hat{m}^a_\sigma m^a_\sigma)}$$

$$\left[\mathbb{E}_u \int \prod_a \mathrm{d}w^a e^{\psi^{(n)}_w(w)}\right]^{N/\alpha} \left[\int \mathrm{d}y P_o(y|z)\mathrm{d}z \prod_a \mathrm{d}h^a \mathrm{d}\sigma^a e^{\psi^{(n)}_{\mathrm{out}}(h,\sigma;t)}\right]^{\rho N} \tag{116}$$

$$\left[\int \mathrm{d}y P_o(y|z)\mathrm{d}z \prod_a \mathrm{d}h^a \mathrm{d}\sigma^a e^{\psi^{(n)}_{\mathrm{out}}(h,\sigma;t')}\right]^{(1-\rho)N} ;$$

where we defined

$$\psi^{(n)}_w(w) = \sum_a \log P_W(w^a) - \sum_{a\le b} \hat{Q}^{ab}_w w^a w^b - \sum_a \hat{m}^a_w u w^a \tag{117}$$

$$\psi^{(n)}_{\mathrm{out}}(h,\sigma;\bar{t}) = -\beta\bar{t}\sum_a l(yh^a) - \sum_{a\le b} \hat{Q}^{ab}_\sigma \sigma^a \sigma^b - \sum_a \hat{m}^a_\sigma y \sigma^a - \frac{1}{2}(h - c\sigma - \lambda y m_\sigma)^T Q^{-1}_\sigma (h - c\sigma - \lambda y m_\sigma)$$

$$- \frac{1}{2}\log\det Q_\sigma - \frac{1}{2}\begin{pmatrix} z \\ \sigma \end{pmatrix}^T \begin{pmatrix} \rho_u & m^T_w \\ m_w & Q_w \end{pmatrix}^{-1} \begin{pmatrix} z \\ \sigma \end{pmatrix} - \frac{1}{2}\log\det\begin{pmatrix} \rho_u & m^T_w \\ m_w & Q_w \end{pmatrix} . \tag{118}$$

We use the replica-symmetric ansatz: we set $\hat{Q}^{aa} = \frac{1}{2}\hat{R}$, $\hat{Q}^{ab} = -\hat{Q}$, $Q^{aa} = R$, $Q^{ab} = Q$, $\hat{m}^a = -\hat{m}$ and $m^a = m$. We define the variances $V = R - Q$ and $\hat{V} = \hat{R} + \hat{Q}$. We take the first

order in n; and as before we have

$$Q_\sigma^{-1} = \frac{1}{V_\sigma} I_n - \frac{Q_\sigma}{V_\sigma^2} J_{n,n} + o(n) \tag{119}$$

$$\log\det Q_\sigma = n\frac{Q_\sigma}{V_\sigma} + n\log(V_\sigma) + o(n)\ ; \tag{120}$$

we compute that

$$\begin{pmatrix} \rho_u & m_w^T \\ m_w & Q_w \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{\rho_u} + n\frac{m_w^2}{V_w\rho_u^2} & -\frac{m_w}{V_w\rho_u}(1,\dots,1) \\ -\frac{m_w}{V_w\rho_u}(1,\dots,1)^T & \frac{1}{V_w} I_n - \frac{1}{V_w^2}(Q_w - \frac{m_w^2}{\rho_u}) J_{n,n} \end{pmatrix} \tag{121}$$

$$\log\det \begin{pmatrix} \rho_u & m_w^T \\ m_w & Q_w \end{pmatrix} = \log\rho_u + \frac{n}{V_w}(Q_w - \frac{m_w^2}{\rho_u}) + n\log V_w + o(n)\ . \tag{122}$$

We can factorize the replica introducing scalar standard Gaussians:

$$\begin{aligned} * \propto \int & \mathrm{d}\hat{Q}_w \mathrm{d}\hat{V}_w \mathrm{d}Q_w \mathrm{d}V_w \mathrm{d}\hat{Q}_\sigma \mathrm{d}\hat{V}_\sigma \mathrm{d}Q_\sigma \mathrm{d}V_\sigma \mathrm{d}\hat{m}_w \mathrm{d}m_w \mathrm{d}\hat{m}_\sigma \mathrm{d}m_\sigma e^{\frac{nN}{2}(\hat{V}_w V_w + \hat{V}_w Q_w - V_w\hat{Q}_w + \hat{V}_\sigma V_\sigma + \hat{V}_\sigma Q_\sigma - V_\sigma \hat{Q}_\sigma)} \\ & e^{-nN(\hat{m}_w m_w + \hat{m}_\sigma m_\sigma)} \left[\mathbb{E}_{u,\varsigma}\left(\int \mathrm{d}w\, e^{\psi_w(w)}\right)^n\right]^{N/\alpha} \left[\mathbb{E}_{\xi,\zeta,\chi}\int \mathrm{d}y\mathrm{d}z\, \psi_{\text{out}}^*(y,z)\left(\int \mathrm{d}h\mathrm{d}\sigma e^{\psi_{\text{out}}(h,\sigma;t)}\right)^n\right]^{\rho N} \\ & \left[\mathbb{E}_{\xi,\zeta,\chi}\int \mathrm{d}y\mathrm{d}z\, \psi_{\text{out}}^*(y,z)\left(\int \mathrm{d}h\mathrm{d}\sigma e^{\psi_{\text{out}}(h,\sigma;t)}\right)^n\right]^{(1-\rho)N}\ ; \end{aligned} \tag{123}$$

with

$$\psi_w(w) = \log P_W(w) - \frac{1}{2}\hat{V}_w w^2 + \left(\varsigma\sqrt{\hat{Q}_w} + u\hat{m}_w\right) w \tag{124}$$

$$\begin{aligned} \psi_{\text{out}}(h,\sigma;\bar{t}) = & -\beta\bar{t}l(yh) - \frac{1}{2}\hat{V}_\sigma\sigma^2 + \left(\xi\sqrt{\hat{Q}_\sigma} + y\hat{m}_\sigma\right)\sigma \\ & + \log\mathcal{N}\left(h|c\sigma + \lambda y m_\sigma + \sqrt{Q_\sigma}\zeta, V_\sigma\right) + \log\mathcal{N}\left(\sigma\left|\rho_u^{-1}m_w z + \sqrt{(1-\eta_w)Q_w}\chi, V_w\right.\right) \end{aligned} \tag{125}$$

$$\psi_{\text{out}}^*(y,z) = P_o(y|z)\mathcal{N}(z|0,\rho_u)\ , \tag{126}$$

where we defined $\eta_w = \frac{m_w^2}{\rho_u Q_w}$. We take the limit $N\to\infty$ and $n\to 0$. The free entropy is then

$$\begin{aligned} \phi = & \frac{1}{2}\left(\hat{V}_w V_w + \hat{V}_w Q_w - V_w\hat{Q}_w + \hat{V}_\sigma V_\sigma + \hat{V}_\sigma Q_\sigma - V_\sigma\hat{Q}_\sigma\right) - \hat{m}_w m_w - \hat{m}_\sigma m_\sigma + \frac{1}{\alpha}\mathbb{E}_{u,\varsigma}\left(\log\int \mathrm{d}w\, e^{\psi_w(w)}\right) \\ & + \rho\mathbb{E}_{\xi,\zeta,\chi}\left(\int \mathrm{d}y\mathrm{d}z\, \psi_{\text{out}}^*(y,z)\log\int \mathrm{d}h\mathrm{d}\sigma e^{\psi_{\text{out}}(h,\sigma;t)}\right) \\ & + (1-\rho)\mathbb{E}_{\xi,\zeta,\chi}\left(\int \mathrm{d}y\mathrm{d}z\, \psi_{\text{out}}^*(y,z)\log\int \mathrm{d}h\mathrm{d}\sigma e^{\psi_{\text{out}}(h,\sigma;t')}\right)\ . \end{aligned} \tag{127}$$

As before we rescale the order parameters according to $\hat{V}\to\beta\hat{V}$, $\hat{Q}\to\beta^2\hat{Q}$, $\hat{m}\to\beta\hat{m}$ and $V\to\beta^{-1}V$ for both w and σ, so in the limit $\beta\to\infty$ by Laplace's method the inner integrals are not degenerated. We define

$$w^* = \operatorname*{argmax}_w \psi_w(w) \tag{128}$$

$$(h^*,\sigma^*) = \operatorname*{argmax}_{h,\sigma}\psi_{\text{out}}(h,\sigma;\bar{t}=1) \qquad (h'^*,\sigma'^*) = \operatorname*{argmax}_{h,\sigma}\psi_{\text{out}}(h,\sigma;\bar{t}=0)\ . \tag{129}$$

The fixed-point equations are

$$m_w = \frac{1}{\alpha}\mathbb{E}_{u,\varsigma}\, uw^* \qquad m_\sigma = \mathbb{E}_{\xi,\zeta,\chi}\int \mathrm{d}y\mathrm{d}z\, \psi_{\text{out}}^*(y,z)y\left(\rho\sigma^* + (1-\rho)\sigma'^*\right) \tag{130}$$

$$Q_w = \frac{1}{\alpha}\mathbb{E}_{u,\varsigma}(w^*)^2 \qquad Q_\sigma = \mathbb{E}_{\xi,\zeta,\chi}\int \mathrm{d}y\mathrm{d}z\, \psi_{\text{out}}^*(y,z)\left(\rho(\sigma^*)^2 + (1-\rho)(\sigma'^*)^2\right) \tag{131}$$

$$V_w = \frac{1}{\alpha}\frac{1}{\sqrt{\hat{Q}_w}}\mathbb{E}_{u,\varsigma}\,\varsigma w^* \qquad V_\sigma = \frac{1}{\sqrt{\hat{Q}_\sigma}}\mathbb{E}_{\xi,\zeta,\chi}\int \mathrm{d}y\mathrm{d}z\, \psi_{\text{out}}^*(y,z)\xi\left(\rho\sigma^* + (1-\rho)\sigma'^*\right) \tag{132}$$

$$\hat{m}_w = \frac{1}{V_w}\mathbb{E}_{\xi,\zeta,\chi}\int \mathrm{d}y\mathrm{d}z\, \psi^*_{\mathrm{out}}(y,z)\left(\rho_u^{-1}z - \chi\frac{\rho_u^{-1}m_w}{\sqrt{(1-\eta_w)Q_w}}\right)(\rho\sigma^* + (1-\rho)\sigma^{'*}) \tag{133}$$

$$\hat{Q}_w = \frac{1}{V_w^2}\mathbb{E}_{\xi,\zeta,\chi}\int \mathrm{d}y\mathrm{d}z\, \psi^*_{\mathrm{out}}(y,z)\left(\rho(\sigma^* - \rho_u^{-1}m_w z - \chi\sqrt{(1-\eta_w)Q_w})^2 \right. \tag{134}$$
$$\left. + (1-\rho)(\sigma^{'*} - \rho_u^{-1}m_w z - \chi\sqrt{(1-\eta_w)Q_w})^2\right)$$

$$\hat{V}_w = \frac{1}{V_w}\left(1 - \frac{1}{\sqrt{(1-\eta_w)Q_w}}\mathbb{E}_{\xi,\zeta,\chi}\int \mathrm{d}y\mathrm{d}z\, \psi^*_{\mathrm{out}}(y,z)\chi\left(\rho\sigma^* + (1-\rho)\sigma^{'*}\right)\right) \tag{135}$$

$$\hat{m}_\sigma = \frac{\lambda}{V_\sigma}\mathbb{E}_{\xi,\zeta,\chi}\int \mathrm{d}y\mathrm{d}z\, \psi^*_{\mathrm{out}}(y,z)y\left(\rho(h^* - c\sigma^* - \lambda y m_\sigma) + (1-\rho)(h^{'*} - c\sigma^{'*} - \lambda y m_\sigma)\right) \tag{136}$$

$$\hat{Q}_\sigma = \frac{1}{V_\sigma^2}\mathbb{E}_{\xi,\zeta,\chi}\int \mathrm{d}y\mathrm{d}z\, \psi^*_{\mathrm{out}}(y,z)\left(\rho(h^* - c\sigma^* - \lambda y m_\sigma - \sqrt{Q}_\sigma\zeta)^2 + (1-\rho)(h^{'*} - c\sigma^{'*} - \lambda y m_\sigma - \sqrt{Q}_\sigma\zeta)^2\right) \tag{137}$$

$$\hat{V}_\sigma = \frac{1}{V_\sigma}\left(1 - \frac{1}{\sqrt{Q_\sigma}}\mathbb{E}_{\xi,\zeta,\chi}\int \mathrm{d}y\mathrm{d}z\, \psi^*_{\mathrm{out}}(y,z)\zeta\left(\rho(h^* - c\sigma^*) + (1-\rho)(h^{'*} - c\sigma^{'*})\right)\right) \tag{138}$$

The average train and test losses can be computed by deriving ϕ wrt t and t' and taking it extremum by evaluating it at the fixed-point of these equations.

The integral on z can be computed by the change of variable $\chi \to \frac{\chi}{\sqrt{1-\eta_w}} - \frac{\rho_u^{-1}m_w z}{\sqrt{(1-\eta_w)Q_w}}$. We obtain the expressions given in part A, after simplification of the notations.

C Bayes-optimal performances

In section 4 we compare the GCN to the Bayes-optimal performances. The Bayes-optimal performances on the CSBM and the GLM–SBM were derived by [25–27]. They can be expressed as a function of the fixed-point of a system of equations over three scalar quantities.

These works consider a non-directed SBM with symmetric adjacency matrix A and symmetric fluctuations Ξ in $A^{\mathrm{g}} = \frac{\lambda}{\sqrt{N}}yy^T + \Xi$. In our work for simplicity we take Ξ non-symmetric. Then the corresponding A and A^{g} can be mapped to a non-directed SBM by the transform $(A + A^T)/\sqrt{2}$ and it is sufficient to rescale the snr λ of the non-directed SBM by $\sqrt{2}$ to have the same snr as for the directed SBM. So we set $\Delta_I = 2\lambda^2$ the signal-to-noise ratio of the corresponding low-rank matrix factorization problem.

C.1 CSBM

The equations are given by [26] in its appendix. The self-consistent equations read

$$m^t = \frac{\mu}{\alpha}m_u^t + \Delta_I m_y^{t-1} \tag{139}$$

$$m_y^t = \rho + (1-\rho)\mathbb{E}_W\left[\tanh\left(m^t + \sqrt{m^t}W\right)\right] \tag{140}$$

$$m_u^{t+1} = \frac{\mu m_y^t}{1 + \mu m_y^t} \tag{141}$$

where W is a standard scalar Gaussian. Once a fixed-point (m, m_y, m_u) is obtained the test accuracy is given by

$$\mathrm{Acc}_{\mathrm{test}} = \frac{1}{2}(1 + \mathrm{erf}\sqrt{m/2})\ . \tag{142}$$

In the large λ limit we have $m_y \to 1$ and

$$\log(1 - \mathrm{Acc}_{\mathrm{test}}) \underset{\lambda\to\infty}{\sim} -\lambda^2\ . \tag{143}$$

C.2 GLM–SBM

The equations are given by [27], only for the unsupervised case $\rho = 0$. The supervised part can be inferred from the simpler case of Bayes-optimal inference on a GLM [4]. Then the supervised part and the unsupervised part are merged in a linear fashion as on the CSBM. We need the following (not normalized) density on y and z:

$$Q(y, z; B, A, \omega, V) = P_o(y|z) e^{-A/2+By} \frac{e^{-(z-\omega)^2/2V}}{\sqrt{2\pi V}} \,. \tag{144}$$

We define the update functions

$$\begin{aligned} Z_{\text{out}}(B, A, \omega, V) &= \int \mathrm{d}y \mathrm{d}z \; Q(y, z; B, A, \omega, V) & Z_{\text{out}}^{\text{sup}}(\omega, V) &= \int \mathrm{d}z \; Q(+1, z; 0, 0, \omega, V) \\ &= e^{-A/2} \left(\cosh B + \sinh(B) \mathrm{erf}(\omega/\sqrt{2V}) \right) & &= \frac{1}{2} \left(1 + \mathrm{erf}(\omega/\sqrt{2V}) \right) \end{aligned} \tag{145}$$

$$f_{\text{out}} = \partial_\omega \log Z_{\text{out}} \qquad f_{\text{out}}^{\text{sup}} = \partial_\omega \log Z_{\text{out}}^{\text{sup}} \tag{146}$$

$$f_y = \partial_B \log Z_{\text{out}} \tag{147}$$

Then the self-consistent equations read

$$\begin{aligned} \hat{m}_u^t = \rho \mathbb{E}_\eta \left[Z_{\text{out}}^{\text{sup}} \left(\sqrt{m_u^t} \eta, \rho_u - m_u^t \right) f_{\text{out}}^{\text{sup}} \left(\sqrt{m_u^t} \eta, \rho_u - m_u^t \right)^2 \right] \\ + (1-\rho) \mathbb{E}_{\xi,\eta} \left[Z_{\text{out}} \left(\sqrt{\Delta_I m_y^t} \xi, \Delta_I m_y^t, \sqrt{m_u^t} \eta, \rho_u - m_u^t \right) f_{\text{out}} \left(\sqrt{\Delta_I m_y^t} \xi, \Delta_I m_y^t, \sqrt{m_u^t} \eta, \rho_u - m_u^t \right)^2 \right] \end{aligned} \tag{148}$$

$$m_y^{t+1} = \rho + (1-\rho) \mathbb{E}_{\xi,\eta} \left[Z_{\text{out}} \left(\sqrt{\Delta_I m_y^t} \xi, \Delta_I m_y^t, \sqrt{m_u^t} \eta, \rho_u - m_u^t \right) f_y \left(\sqrt{\Delta_I m_y^t} \xi, \Delta_I m_y^t, \sqrt{m_u^t} \eta, \rho_u - m_u^t \right)^2 \right] \tag{149}$$

$$m_u^{t+1} = \frac{1}{\alpha} \frac{\hat{m}_u^t}{1 + \hat{m}_u^t} \tag{150}$$

where ξ and η are standard scalar Gaussians and $\rho_u = \alpha^{-1}$. Once a fixed-point $(\hat{m}_u, m_y, m_u)$ is obtained the test accuracy is given by

$$\mathrm{Acc}_{\text{test}} = \mathbb{E}_{\xi,\eta} \left[\int \mathrm{d}y \mathrm{d}z \; Q \left(y, z; \sqrt{\Delta_I m_y} \xi, \Delta_I m_y, \sqrt{m_u} \eta, \rho_u - m_u \right) \delta_{y = \text{sign} \, f_y \left(\sqrt{\Delta_I m_y} \xi, \Delta_I m_y, \sqrt{m_u} \eta, \rho_u - m_u \right)} \right] \tag{151}$$

$$= \mathbb{E}_\eta \left[\frac{1}{2} \left(1 + \mathrm{erf} \left(\frac{\sqrt{m_u} \eta}{\sqrt{2(\rho_u - m_u)}} \right) \right) \left(1 + \mathrm{erf} \left(\frac{\sqrt{\Delta_I m_y}}{\sqrt{2}} + \frac{1}{\sqrt{2 \Delta_I m_y}} \mathrm{arcth} \, \mathrm{erf} \left(\frac{\sqrt{m_u} \eta}{\sqrt{2(\rho_u - m_u)}} \right) \right) \right) \right] . \tag{152}$$

In the large λ limit we have $m_y \to 1$ and

$$\log(1 - \mathrm{Acc}_{\text{test}}) \underset{\lambda \to \infty}{\sim} -\lambda^2 \,. \tag{153}$$

D Fashion-SBM, a semi-realistic dataset

In fig. 3 we introduced fashion-SBM to show that our prediction $c^* \approx 1/\lambda$ seems to hold for a dataset more realistic than the CSBM or the GLM–SBM. In this section we detail how fashion-SBM is constructed.

Fashion-SBM is made by populating a SBM with attributes from fashion-MNIST [30]. The binary labels y of the nodes are drawn first. The graph is generated according to the SBM described in the main part, with parameters d and λ. As to the features, we consider only the training set of fashion-MNIST; out of the ten classes we keep only two classes to form $\tilde{X} \in \mathbb{R}^{N \times M}$ that is normalized according to

$$\hat{X}_{i\mu} = \tilde{X}_{i\mu} + \epsilon_{i\mu} \tag{154}$$

$$X_{i\mu} = \sqrt{N} \frac{\hat{X}_{i\mu} - \frac{1}{N} \sum_j \hat{X}_{j\mu}}{\sqrt{\sum_j (\hat{X}_{j\mu} - \frac{1}{N} \sum_k \hat{X}_{k\mu})^2}} \tag{155}$$

ϵ is a small noise added to each pixel to avoid pixels that are always black. The resulting dataset has dimensions $N = 12000$ and $M = 784$.

In the experiment 3 we choose to use the two classes 2 (pullover) and 4 (coat). They are similar enough to keep balanced the signals of the features and the graph. The other classes are more dissimilar and cary a stronger signal, which results in the graph having little effect on the performance.

E Details on numerics

The systems (28)-(36) and (41)-(49) are solved by the iterating the twelve equations in parallel until convergence. About twenty iterations are necessary. The iterations are stable and no damping is necessary. The integral over (ξ, ζ, χ) is evaluated by Monte-Carlo over 10^6 points; we use the same samples over the iterations so they can exactly converge. For the quadratic and hinge losses the extremizer of the potential (9) has an explicit solution; for the logistic loss we compute it by Newton's descent, a few steps are enough. The whole computation takes around one minute on a single CPU core with 5GB of memory.

For figures 2 and 7 solving these two systems we were only able to reach misclassification errors $1 - \mathrm{Acc}_{\mathrm{test}}$ of 10^{-6} because of numerical imprecision and the finite number of Monte-Carlo samples.

F Supplementary figures

F.1 Optimal architecture

On fig. 4 we search for the optimal architecture for data generated by the GLM–SBM. On figs. 5 and 6 we search for the optimal architecture for data generated at different αs, that is $\alpha = 0.7$ and $\alpha = 2$, for the CSBM and the GLM–SBM. Together with fig. 1 in the main part we reach conclusions that are detailed in section 4.1.

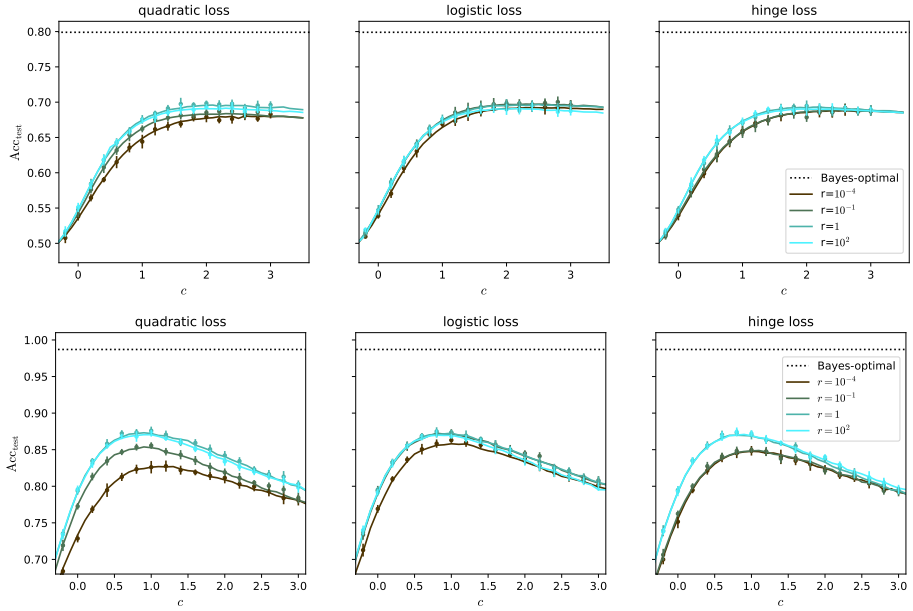

Figure 4: Search for the optimal parameters of the GCN on GLM–SBM. $\alpha = 4$, $\rho = 0.1$. *Top:* low snr, $\lambda = 0.5$. *Bottom:* high snr, $\lambda = 1.5$. Full lines: prediction for the test accuracy obtained by eqs. (13) and (41)-(49); dots: numerical simulation of the GCN for $N = 10^4$ and $d = 30$, averaged over ten experiments; dotted line: Bayes-optimal test accuracy.

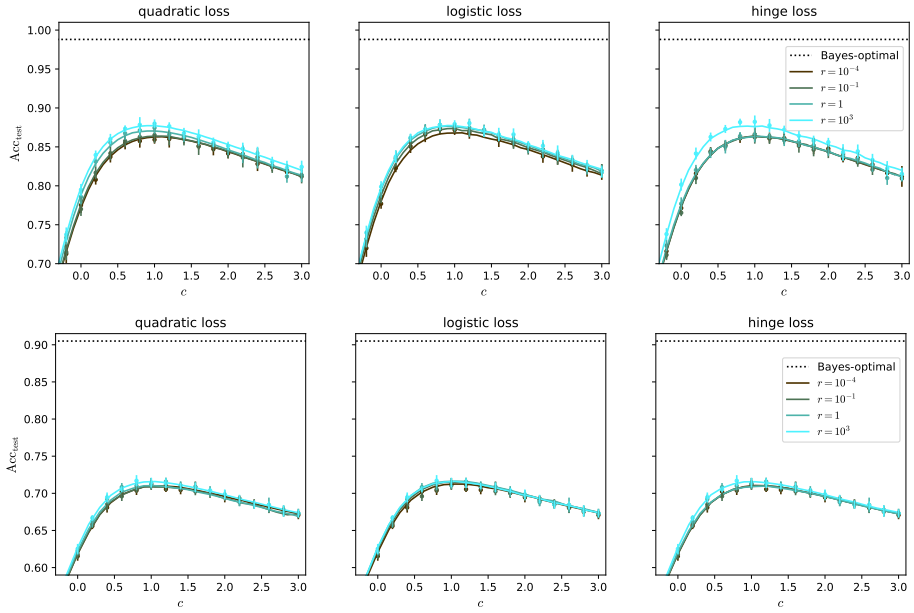

Figure 5: Search for the optimal parameters of the GCN. $\alpha = 0.7$, $\rho = 0.1$. *Top:* CSBM, $\lambda = 1.5$, $\mu = 3$. *Bottom:* GLM–SBM, $\lambda = 1$. Full lines: prediction for the test accuracy obtained by eqs. (13); dots: numerical simulation of the GCN for $N = 10^4$ and $d = 30$, averaged over ten experiments; dotted line: Bayes-optimal test accuracy.

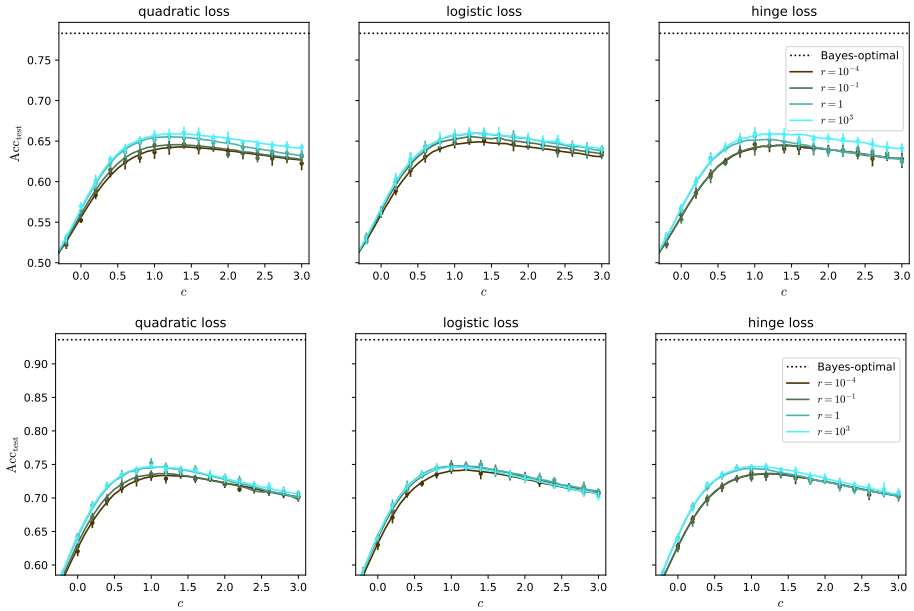

Figure 6: Search for the optimal parameters of the GCN. $\alpha = 2$, $\rho = 0.1$. *Top:* CSBM, $\lambda = 0.7$, $\mu = 1$. *Bottom:* GLM–SBM, $\lambda = 1$. Full lines: prediction for the test accuracy obtained by eqs. (13); dots: numerical simulation of the GCN for $N = 10^4$ and $d = 30$, averaged over ten experiments; dotted line: Bayes-optimal test accuracy.

On fig. 7 we show the effect of the regularization r on the convergence rate at large graph snr λ. For the quadratic loss, the rate depends on the regularization while for the logistic loss it does not.

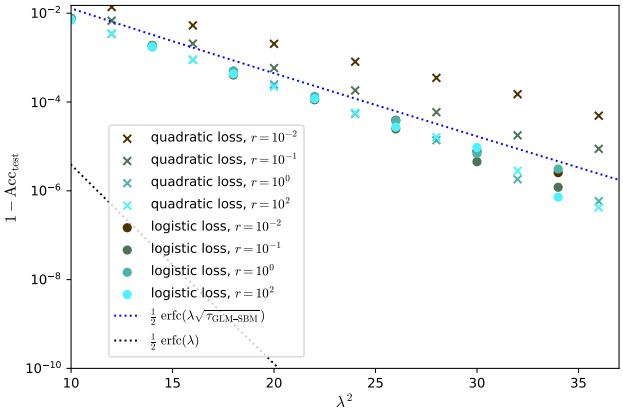

Figure 7: Asymptotic misclassification error $1 - \mathrm{Acc}_{\mathrm{test}}$ on the GLM–SBM. $\alpha = 4$, $\rho = 0.1$. Dots: prediction for the test accuracy obtained by eqs. (13) and (41)-(49), for $c = c^*$ optimal obtained by grid search. The blue dotted line is given by (16).

F.2 Interpolation peak

On fig. 8 we show that an interpolation peak appears for the ridge regression on the GLM–SBM when the regularization is small while varying the training ratio ρ. At the interpolation peak the train error becomes strictly positive, the train accuracy becomes strictly smaller than one, the test error diverges and the test accuracy has an inflexion point. The peak is located at $\alpha\rho = 1$. Increasing the regularization r smooths it out. Similar curves are obtained for the CSBM.

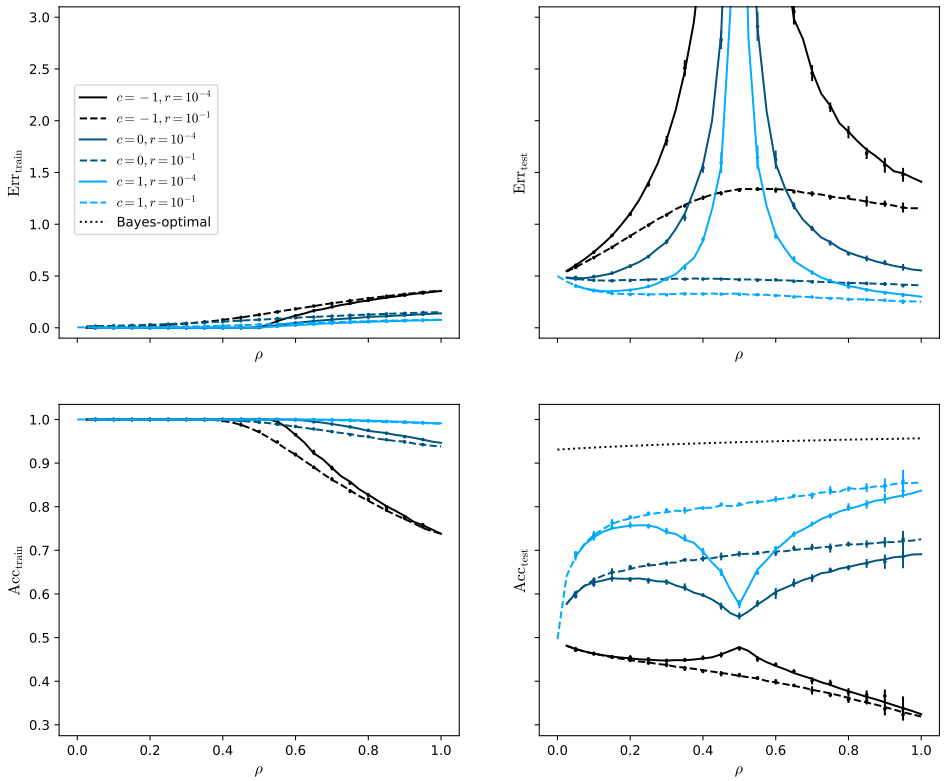

Figure 8: Interpolation peak on the GLM–SBM for the quadratic loss. $\alpha = 2$ and $\lambda = 1$. Lines: predictions by eqs. (13) and (41)-(49); dots: numerical simulation of the GCN for $N = 10^4$ and $d = N/2$, averaged over ten experiments; dotted line: Bayes-optimal test accuracy.

On fig. 9 we show that an interpolation peak appears for the logistic regression on the CSBM when the regularization is small while varying the training ratio ρ. At the interpolation peak the train error becomes strictly positive, the train accuracy becomes strictly smaller than one, the test error diverges and the test accuracy has an inflexion point. The position of the peak depends on the self-loop intensity c and the aspect ratio α. On fig. 10 we show how its position varies with respect to ρ and α at $c = 1$. Increasing the regularization r smooths it out. Similar curves are obtained for the hinge loss and the GLM–SBM.

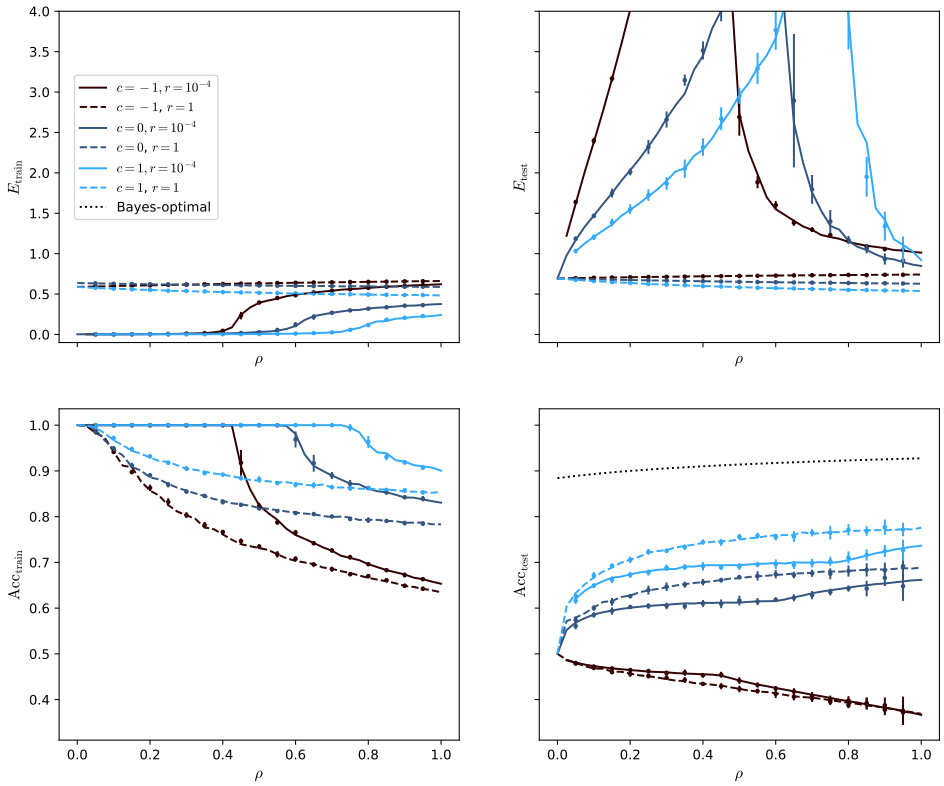

Figure 9: Interpolation peak on the CSBM for the logistic loss. $\alpha = 4$, $\lambda = 1$ and $\mu = 1$. Lines: predictions by eqs. (13) and (28)-(36); dots: numerical simulation of the GCN for $N = 10^4$ and $d = N/2$, averaged over ten experiments; dotted line: Bayes-optimal test accuracy.

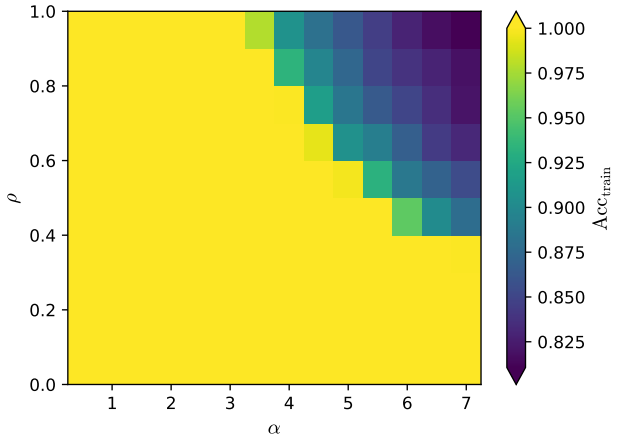

Figure 10: Position of the interpolation peak on the CSBM for the logistic loss. The interpolation peak is located at the border of $\text{Acc}_{\text{train}} < 1$. $\lambda = 1$, $\mu = 1$ and $c = 1$. Predictions by eqs. (13) and (28)-(36).