Technical Appendix

A. Computational Cost

Per training epoch, SGPC stays edge-linear both in time and memory. The Wasserstein-Entropic Lift first solves an entropic OT problem with one Sinkhorn run and a single JKO step, costing $\mathcal{O}(n, d_0^2)$ floating-point operations and $\mathcal{O}(n, d_0)$ memory for node features. β–Dirichlet calibration then updates the two Gamma parameters for every edge in parallel, giving an $\mathcal{O}(m)$ pass with $\mathcal{O}(1)$ extra storage per edge. Spectral optimization performs a two-pass Lanczos eigensolver and one gradient evaluation, each touching every non-zero in the sheaf Laplacian, so the cost is again $\mathcal{O}(m)$ and the memory footprint $\mathcal{O}(n)$. The SVR–AFM layer applies a variance-reduced CG diffusion, whose expected complexity is $\mathcal{O}(m)$ and memory $\mathcal{O}(n)$, followed by an adaptive frequency mixing that is $\mathcal{O}(H, m)$. Putting the stages together, an epoch of SGPC requires $\mathcal{O}(m{+}n, d_0^2)$ time and only $\mathcal{O}(n{+}m)$ memory, making it scalable to graphs with millions of edges on a single GPU.

B. Proof of Theorem 1

Theorem 1 (PAC–Bayes Sheaf Generalization Bound)**.**

$$\mathcal{L}_{\mathcal{D}}(f) \leq \mathcal{L}(y, \widehat{y}) + \underbrace{\sqrt{\frac{\mathrm{KL}(\rho \parallel \pi) + \log \frac{2}{\delta}}{2n}} + \frac{c_{het}}{\lambda_2}}_{\mathcal{R}_{bound}}, \tag{34}$$

where $\mathcal{L}(y, \widehat{y})$ is the calibrated empirical risk.

Proof. **(i) PAC-Bayes bound for stochastic restriction maps.** For any measurable loss $C \in [0, 1]$, the classical PAC-Bayes theorem states that for every prior π and for every posterior ρ as follows:

$$\Pr_{S \sim \mathcal{D}^n} \left[\mathcal{L}_{\mathcal{D}}(\widehat{f}) \leq \mathcal{L}_S(\widehat{f}) + \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \log(2/\delta)}{2n}} \right] \geq 1 - \delta \tag{35}$$

with probability at least $1 - \delta$ over the draw of the labeled sample $S \sim \mathcal{D}$. Because our empirical loss $\mathcal{L}(y, \widehat{y})$ is just $\mathcal{L}_S(\widehat{f})$ with the calibrated predictions $f(\widehat{y}_i; \bar{\kappa}_{ij})$, the above equation yields

$$\mathcal{L}_{\mathcal{D}}(\widehat{f}) \leq \mathcal{L}(y, \widehat{y}) + \sqrt{\tfrac{\mathrm{KL}(\rho \| \pi) + \log(2/\delta)}{2n}} \quad \text{w.p. } 1 - \tfrac{\delta}{2}. \tag{36}$$

Multiplying the last term by a user-chosen constant $\lambda_{\mathrm{KL}} \geq 1$ only loosens the inequality.

(ii) Diffusion-stability bound via the spectral gap. For a cellular-sheaf Laplacian $L_{\mathcal{F}}$, the convergence error after one implicit-Euler diffusion step admits the classical Rayleigh-quotient control

$$\left\| (I + \Delta t \, L_{\mathcal{F}})^{-1} - \Pi_{\mathbf{1}} \right\|_2 = \frac{1}{1 + \Delta t \, \lambda_2(L_{\mathcal{F}})}, \tag{37}$$

where $\Pi_{\mathbf{1}}$ projects onto the all-ones subspace. On a heterophilous graph, edge disagreements governed by $\bar{\kappa}_{ij}$ inject class-coupling energy

$$c_{\mathrm{het}} = \|\Pi\|_F = \Big(\sum_{c \neq c'} \Pi_{cc'}^2 \Big)^{1/2}, \tag{38}$$

which propagates through diffusion with gain at most $1/[1 + \Delta t \, \lambda_2]$. Choosing $\Delta t = 1$ gives the diffusion-error upper bound

$$\underbrace{\left\| H^{\mathrm{svr}} - H^{\star} \right\|_F}_{\text{instability}} \leq \frac{c_{\mathrm{het}}}{\lambda_2(L_{\mathcal{F}})}, \tag{39}$$

where $H^{\star}$ is the perfectly mixed (homophilic) representation. The right-hand side is exactly the spectral penalty $\mathcal{L}_{\mathrm{spec}}$. Because $\mathcal{L}_{\mathrm{spec}}$ is a deterministic function of the observed sample labels, we can apply a union bound, where the event $\mathcal{L}_{\mathrm{spec}} \leq \frac{c_{\mathrm{het}}}{\lambda_2}$ holds with probability at least $1 - \delta/2$. Thus, the following inequality holds

$$\mathcal{L}_{\mathcal{D}}(\widehat{f}) \leq \mathcal{L}(y, \widehat{y}) + \sqrt{\tfrac{\mathrm{KL}(\rho \| \pi) + \log(2/\delta)}{2n}} + \lambda_{\mathrm{spec}} \, \frac{c_{\mathrm{het}}}{\lambda_2(L_{\mathcal{F}})}. \tag{40}$$

Again, scaling the last term by the non-negative constant λ_{spec} only relaxes the bound. $\square$

C. Proof of Theorem 2

Theorem 2 (CG convergence with sparsifier)**.** *Let $\tilde{L}_t$ be a $(1 \pm \varepsilon)$ spectral sparsifier of the sheaf Laplacian L_t, obtained via leverage–score sampling as,*

$$\lambda_2(L_t) \geq \gamma \quad \textit{and} \quad \lambda_{\max}(L_t) \leq \Lambda \tag{41}$$

with a time step $\Delta t \leq 1/\Lambda$. Then, for any right-hand side b and initial residual r_0, CG applied to $\big(I + \Delta t \tilde{L}_t\big)h = b$ achieves a residual $\|r_k\|_2 \leq \epsilon_{\mathrm{CG}}$ (error bound) at most $k_{\max}$ iterations:

$$k_{\max} \leq \left\lceil \sqrt{\kappa\big(I + \Delta t \tilde{L}_t\big)} \log \frac{\|r_0\|_2}{\epsilon_{\mathrm{CG}}} \right\rceil = O\big(\log(1/\epsilon_{\mathrm{CG}})\big). \tag{42}$$

The above inequality holds because

$$\kappa\big(I + \Delta t \tilde{L}_t\big) = \frac{1 + \Delta t \lambda_{\max}(\tilde{L}_t)}{1 + \Delta t \lambda_2(\tilde{L}_t)} \leq \frac{1 + (1+\varepsilon)\Delta t \Lambda}{1 + (1-\varepsilon)\Delta t \gamma}. \tag{43}$$

Since $\kappa\big(I + \Delta t \tilde{L}_t\big) \leq 2 + \varepsilon = O(1)$, we can infer that the iteration bound is uniform in $|V|$, $|E|$, and the epoch t.

Proof. Because $\tilde{L}_t$ is a $(1 \pm \varepsilon)$ sparsifier, the following inequality holds for every $h \in \mathbb{R}^{|V|}$:

$$(1-\varepsilon) h^\top L_t h \leq h^\top \tilde{L}_t h \leq (1+\varepsilon) h^\top L_t h. \tag{44}$$

Thus, $(1-\varepsilon)\lambda_i(L_t) \leq \lambda_i(\tilde{L}_t) \leq (1+\varepsilon)\lambda_i(L_t)$ for all i. With $\lambda_{2,t} \geq \gamma$ and $\lambda_{\max,t} \leq \Lambda$, we get

$$\lambda_2(\tilde{L}_t) \geq (1-\varepsilon)\gamma, \quad \lambda_{\max}(\tilde{L}_t) \leq (1+\varepsilon)\Lambda. \tag{45}$$

Define $A := I + \Delta t \tilde{L}_t$. Its eigenvalues are $1 + \Delta t \lambda_i(\tilde{L}_t)$, so

$$1 + \Delta t \lambda_{\max}(\tilde{L}_t) \leq 1 + (1+\varepsilon)\Delta t \Lambda \leq 1 + (1+\varepsilon) \leq 2 + \varepsilon, \tag{46}$$

where $1 + \Delta t \lambda_2(\tilde{L}_t) \geq 1 + (1-\varepsilon)\Delta t \gamma \geq 1$. Thus, $\kappa(A) \leq (2+\varepsilon)/1 \leq 2 + \varepsilon = O(1)$. For an symmetric positive definite matrix with condition number κ, CG satisfies $\|r_k\|_2 \leq 2\|r_0\|_2 \big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\big)^k$. Solving $\|r_k\|_2 \leq \epsilon_{\mathrm{CG}}$ gives

$$k \leq \sqrt{\kappa(A)} \log \frac{\|r_0\|_2}{\epsilon_{\mathrm{CG}}} = O\big(\log(1/\epsilon_{\mathrm{CG}})\big), \tag{47}$$

because $\sqrt{\kappa(A)}$ is a constant not depending on $|V|$, $|E|$, or the epoch t. Replacing A by $I + \Delta t \tilde{L}_t$ in the linear system completes the proof. $\square$

D. Proof of Theorem 3

Theorem 3 (Wolfe-controlled gap ascent)**.** *Let v_t be the normalized eigenvector corresponding to $\lambda_2(L_t)$. At epoch t, the optimizer performs the gradient ascent step,*

$$L_{t+1} = L_t + \eta_t g_t, \tag{48}$$

where $g_t := \nabla_L\big(v_t^\top L_t v_t\big) = v_t v_t^\top$. The step size $\eta_t \in (0, 1]$ is chosen by a Wolfe line search with constant $c_{\mathrm{w}} \in (0, 1)$. Then, the following inequality holds

$$\lambda_2(L_{t+1}) - \lambda_2(L_t) \geq \frac{c_{\mathrm{w}} \eta_t}{2} \geq \frac{c_{\mathrm{w}}}{4}. \tag{49}$$

Consequently, the sequence $\{\lambda_2(L_t)\}_{t \geq 0}$ is strictly non-decreasing and grows by at least $c_{\mathrm{w}}/4$ once the initial full step $\eta_t = 1$ survives the first case.

Proof. Set $f(L) := \lambda_2(L)$ and define $\phi(\eta) := f(L_t + \eta g_t)$. Because $g_t = v_t v_t^\top$ and $v_t^\top v_t = 1$, the derivative of f in the direction g_t is

$$\phi'(0) = v_t^\top g_t v_t = (v_t^\top v_t)^2 = 1. \tag{50}$$

(i) Armijo condition and curvature. Wolfe back-tracking selects the largest $\eta_t = 2^{-m}$ ($m \in \mathbb{N}$) satisfying

$$\phi(\eta_t) \geq \phi(0) + c_{\mathrm{w}} \eta_t \phi'(0) = \lambda_{2,t} + c_{\mathrm{w}} \eta_t. \tag{51}$$

With the same c_{w} it also enforces $|\phi'(\eta_t)| \leq c_{\mathrm{w}} \phi'(0) = c_{\mathrm{w}}$. For the twice-differentiable eigenvalue map f, the derivative $\phi'(\eta)$ is Lipschitz with modulus, so the back-tracking loop stops after at most one extra halving beyond the first η. Consequently, $\eta_t \geq \frac{1}{2}$ whenever the full step $\eta = 1$ does not violate this condition.

(ii) Gap increment. By Taylor's theorem with remainder,

$$\lambda_{2,t+1} - \lambda_{2,t} = \phi(\eta_t) - \phi(0) \geq c_{\mathrm{w}} \eta_t \phi'(0) - \tfrac{1}{2} L_2 \eta_t^2, \tag{52}$$

where $L_2 \leq 2$ is the Lipschitz constant of ϕ'. Since $\phi'(0) = 1$ and $\eta_t \leq 1$, $\frac{1}{2} L_2 \eta_t^2 \leq \eta_t$, the following condition holds

$$\lambda_{2,t+1} - \lambda_{2,t} \geq c_{\mathrm{w}} \eta_t - \eta_t = \eta_t(c_{\mathrm{w}} - 1) + \eta_t \geq \frac{c_{\mathrm{w}} \eta_t}{2}, \tag{53}$$

because $c_{\mathrm{w}} \leq 1$ and $\eta_t \geq \frac{1}{2}$. Finally, using $\eta_t \geq \frac{1}{2}$ once more gives the fixed lower bound $\frac{c_{\mathrm{w}}}{4}$. $\square$

E. Proof of Lemma 1 and Theorem 4

Lemma 1 (Variance reduction). *Let $\theta_{ij} \sim \mathrm{Beta}(\alpha_{ij}, \beta_{ij})$ with $\alpha_{ij}, \beta_{ij} \geq 1$ and denote $\gamma_{ij} := \alpha_{ij} + \beta_{ij}$. After $n_{\mathrm{tot}}(i,j)$ diffusion messages have traversed edge (i,j) (independently of their success/failure counts), the posterior variance satisfies*

$$\mathrm{Var}\big[\theta_{ij} \mid \mathcal{D}\big] \leq \frac{\gamma_{ij}}{\big(\gamma_{ij} + n_{\mathrm{tot}}\big)^2}\Big(1 - \frac{1}{\gamma_{ij} + n_{\mathrm{tot}} + 1}\Big). \tag{54}$$

Consequently,

$$\frac{\mathrm{Var}\big[\theta_{ij} \mid \mathcal{D}\big]}{\mathrm{Var}\big[\theta_{ij}\big]_{\mathit{prior}}} \leq \frac{\gamma_{ij} + 1}{\gamma_{ij} + n_{\mathrm{tot}} + 1} \leq \frac{\gamma_{ij}}{\gamma_{ij} + n_{\mathrm{tot}}}. \tag{55}$$

In the weak-prior regime $\gamma_{ij} \leq 10$ and once $n_{\mathrm{tot}} \geq 5$, this ratio is at most $\frac{2}{3}$.

Proof. After n_{tot} messages, the updated parameters are $\alpha' = \alpha_{ij} + n_1$, $\beta' = \beta_{ij} + n_0$ with $n_1 + n_0 = n_{\text{tot}}$. The posterior variance is given by:

$$\mathrm{Var}[\theta_{ij} \mid \mathcal{D}] = \frac{\alpha'\beta'}{(\alpha' + \beta')^2(\alpha' + \beta' + 1)}. \tag{56}$$

(i) Upper-bound with AM–GM. For non-negative x, y, $xy \leq \frac{1}{4}(x+y)^2$ gives

$$\alpha'\beta' \leq \tfrac{1}{4}(\alpha' + \beta')^2 = \tfrac{1}{4}\,(\gamma_{ij} + n_{\text{tot}})^2, \tag{57}$$

where the rightmost inequality in Eq. 54.

(ii) Relative contraction factor. Using the exact variance formulas leads to

$$\frac{\mathrm{Var}_{\text{post}}}{\mathrm{Var}_{\text{prior}}} = \frac{\alpha'\beta'}{\alpha_{ij}\beta_{ij}} \frac{\gamma_{ij}^2(\gamma_{ij} + 1)}{(\gamma_{ij} + n_{\text{tot}})^2(\gamma_{ij} + n_{\text{tot}} + 1)} \leq \frac{\gamma_{ij} + 1}{\gamma_{ij} + n_{\text{tot}} + 1}, \tag{58}$$

because $\alpha'\beta'/\alpha_{ij}\beta_{ij} \leq (\gamma_{ij} + n_{\text{tot}})/\gamma_{ij}$ by monotonicity. Setting $\gamma_{ij} \leq 10$ and $n_{\text{tot}} \geq 5$ yields the claimed $\leq \frac{2}{3}$ ratio. $\square$

Theorem 4 (Risk–Variance Contraction). *Define at epoch t*

$$\mathcal{B}_t := \underbrace{\mathcal{L}_t}_{\textit{empirical risk}} + \underbrace{\sqrt{\frac{\mathrm{KL}(\rho_t \| \pi) + \log(2/\delta)}{2n}}}_{\textit{KL term}} + \underbrace{\frac{c_{\text{het}}}{\lambda_2(L_t)}}_{\textit{spectral penalty}}. \tag{59}$$

Assume (i) SGD step sizes satisfy a floor $\eta_t \in [\eta_{\min}, \eta_{\max}]$ with $0 < \eta_{\min} \leq \eta_{\max}$; (ii) $n_{\mathrm{tot}}(i,j) \geq 5$ for every edge; (iii) The Wolfe ascent guarantees $\lambda_2(L_{t+1}) - \lambda_2(L_t) \geq \delta_\lambda > 0$ for all t. Then, there exists a constant $\kappa = \kappa\big(\eta_{\min}, L, \delta_\lambda, \gamma_{\max}\big) \in (0,1)$ such that

$$\mathcal{B}_{t+1} \leq (1 - \kappa)\,\mathcal{B}_t, \qquad \forall t \geq T_0, \tag{60}$$

where T_0 is the (finite) epoch after which the variance condition in (ii) holds for every edge. Thus, the PAC–Bayes bound decays geometrically.

Proof. We treat the three summands of $\mathcal{B}_t$.

(i) Empirical-risk descent. Smoothness of the cross-entropy implies $\mathcal{L}_{t+1} \leq \mathcal{L}_t\big(1 - \frac{1}{2}\eta_t L\big)$ for step sizes $\eta_t \leq 2/L$. With $\eta_t \geq \eta_{\min}$, we get the fixed factor $\rho_{\text{risk}} := 1 - \frac{1}{2}\eta_{\min} L < 1$.

(ii) KL-term shrinkage. Lemma 1 gives $\mathrm{Var}_{t+1} \leq \frac{2}{3}\,\mathrm{Var}_t$ after T_0. For Beta distributions, $\mathrm{KL}(\rho\|\pi) \leq C_\beta\ \mathrm{Var}(\theta)$ with an absolute constant C_β; Thus, $\mathrm{KL}_{t+1} \leq \frac{2}{3}\mathrm{KL}_t$, yielding the multiplicative shrinkage $\rho_{\text{KL}} := \sqrt{\frac{2}{3}}$.

(iii) Spectral-gap ascent. The assumption implies $1/\lambda_{2,t+1} \leq (1 - \rho_\lambda)\,1/\lambda_{2,t}$ for $\rho_\lambda := \frac{\delta_\lambda}{\lambda}{}_{2,t} + \delta_\lambda \in (0,1)$. Taking $\rho_{\text{spec}} := 1 - \rho_\lambda < 1$ gives c_{het}/λ_2 the same factor.

Summary. Set $\kappa := 1 - \max\{\rho_{\text{risk}}, \rho_{\text{KL}}, \rho_{\text{spec}}\} \in (0,1)$. For every $t \geq T_0$, each summand of $\mathcal{B}_t$ is multiplied by its own $\rho_\bullet \leq 1 - \kappa$, where $\mathcal{B}_{t+1} \leq (1-\kappa)\mathcal{B}_t$. A finite prefix $0 \leq t < T_0$ only affects the constant prefactor, not the asymptotic rate. $\square$

Table 2: Statistics of the nine graph datasets

Datasets	Cora	Citeseer	Pubmed	Actor	Chameleon	Squirrel	Cornell	Texas	Wisconsin
Nodes	2,708	3,327	19,717	7,600	2,277	5,201	183	183	251
Edges	10,558	9,104	88,648	25,944	33,824	211,872	295	309	499
Features	1,433	3,703	500	931	2,325	2,089	1,703	1,703	1,703
Classes	7	6	3	5	5	5	5	5	5

F. Proof of Lemma 2 and Theorem 5

Lemma 2 (Algorithmic stability bound)**.** *Assume the time–step satisfies $\Delta t < 1/\lambda_{\max}$ and let ϵ_{CG} be the residual tolerance used in every CG solve. Then, the SGPC encoder after T epochs f_T obeys the following inequality:*

$$\left\|f_T - f_0\right\|_2 \leq \sqrt{\frac{\lambda_{\max}}{\lambda_2(L_0)}} \exp\Big(-\tfrac{\Delta t \Delta_G}{2}\Big) + \epsilon_{\mathrm{CG}} T. \tag{61}$$

If Δ_G grows linearly in T (as guaranteed by Theorem 3), the first term decays exponentially fast, while the CG term can be made negligible by choosing $\epsilon_{\mathrm{CG}} = O(T^{-2})$.

Proof. Let $f_t = \mathcal{F}_{\Theta_t,\xi_t}(L_t, \cdot)$ be the encoder defined in Eq. 8 and let $\tilde{L}_t = L_t + \eta_t g_t$ be the Wolfe-stepped Laplacian.

(i) Linear-solver perturbation. Each diffusion at epoch t satisfies the following inequality:

$$\left\|(I + \Delta t L_t)^{-1} - (I + \Delta t \tilde{L}_t)^{-1}\right\|_2 \leq \Delta t \|L_t - \tilde{L}_t\|_2 \leq \Delta t \eta_t \|g_t\|_2, \tag{62}$$

by first-order perturbation of matrix inverses. The CG approximation of $(I + \Delta t \tilde{L}_t)^{-1}$ adds an extra residual of at most ϵ_{CG}. Over T epochs, those errors accumulate to

$$\left\|(I + \Delta t L_t)^{-1} - (I + \Delta t L_t^{\mathrm{CG}})^{-1}\right\|_2 \leq \epsilon_{\mathrm{CG}} T. \tag{63}$$

(ii) Spectral-gap filtering. The inverse-diffusion operator is a low-pass filter whose gain on the k-th eigenvector of L_t equals $1/(1 + \Delta t \lambda_k(L_t))$. Successive gap enlargements shrink the norm of the high-frequency error component as

$$\prod_{s=0}^{t-1} \frac{1 + \Delta t \lambda_{2,s}}{1 + \Delta t \lambda_{2,s+1}} \leq \exp\Big(-\Delta t \Delta_G / B\Big), \tag{64}$$

where $B = \max_s\big(1 + \Delta t \lambda_{2,s}\big) \leq 2$. With $\Delta t < 1/\lambda_{\max} \leq 1$, we have $B \leq 2$. Converting base-e to base-n logarithms gives the exponential factor in the statement.

Summary. Split the total output difference into a spectrally filtered part and a CG-approximation part, and remember that the largest singular value of $(I + \Delta t L_0)^{-1}$ is $\leq \sqrt{\lambda_{\max}/\lambda_{2,0}}$. Consequently, the triangle inequality yields the claimed result. □

Theorem 5 (PAC-Bayes population risk)**.** *Combine Lemma 2 with Theorems 1 (PAC-Bayes) and 4 (risk–variance contraction). Choosing $\epsilon_{\mathrm{CG}} T \leq \exp(-\frac{\Delta t \Delta_G}{2})$, the following inequality holds with probability at least $1 - \delta$:*

$$\mathcal{L}_{\mathcal{D}}(f) \leq \mathcal{L} + \sqrt{\frac{2\exp(-\frac{\Delta t \Delta_G}{2})}{|\mathcal{V}_L|}} + O\Big(\sqrt{\tfrac{\log(1/\delta)}{|\mathcal{V}_L|}}\Big). \tag{65}$$

Therefore, the generalization gap shrinks exponentially in the cumulative gap gain Δ_G.

Proof. **(i) From algorithmic stability to risk discrepancy.** A uniformly β-stable algorithm satisfies

$$\left|\mathcal{L}_{\mathcal{D}}(f) - \mathcal{L}\right| \leq \beta, \tag{66}$$

and Lemma 2 implies

$$\beta = \left\|f_T - f_0\right\|_2 \leq \sqrt{\tfrac{\lambda_{\max}}{\lambda_{2,0}}} e^{-\Delta_G/(2\log n)} + \epsilon_{\mathrm{CG}} T = \tilde{\beta}. \tag{67}$$

(ii) Eliminating the initial predictor. We initialize f_0 with weight decay so that $\|f_0\|_2 \leq \sqrt{\lambda_{\max}/\lambda_{2,0}}$. Setting $\epsilon_{\mathrm{CG}} T \leq e^{-\Delta_G/(2\log n)}$ leads to

$$\tilde{\beta} \leq 2\sqrt{\tfrac{\lambda_{\max}}{\lambda_{2,0}}} e^{-\Delta_G/(2\log n)} = \mathcal{B}_{\text{stab}}. \tag{68}$$

(iii) Injecting stability into PAC-Bayes. The PAC-Bayes bound (Thm. 1) gives with probability $1 - \delta$

$$\mathcal{L}_{\mathcal{D}}(f) \leq \mathcal{L}(y, \widehat{y}) + \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \log(2/\delta)}{2|\mathcal{V}_L|}} + \frac{c_{\mathrm{het}}}{\lambda_{2,T}}. \tag{69}$$

The KL term contracts geometrically by Theorem 4, while $\lambda_{2,T} \geq \lambda_{2,0} + \Delta_G$. Keeping only the leading exponential factor and absorbing constants into the $O(\cdot)$ notation, we obtain

$$\mathcal{L}_{\mathcal{D}}(f) \leq \mathcal{L} + \mathcal{B}_{\mathrm{stab}} + O\Big(\sqrt{\tfrac{\log(1/\delta)}{|\mathcal{V}_L|}}\Big), \tag{70}$$

and substituting $\mathcal{B}_{\mathrm{stab}}$ from Eq. 68 yields the claimed bound. □

G. Datasets and Baselines

Datasets. As shown in Table 2, we employ three homophilic (Cora, Citeseer, and Pubmed) (Kipf and Welling 2016) and six heterophilic graphs (Tang et al. 2009; Rozemberczki et al. 2019) for evaluation.

Baselines. For a fair comparison, we set 15 state-of-the-art models as baselines.

- **GCN** (Kipf and Welling 2016) can be viewed as a first-order truncation of the Chebyshev spectral filters introduced in (Defferrard, Bresson, and Vandergheynst 2016).
- **GAT** (Velickovic et al. 2017) learns edge weights by applying feature-driven attention mechanisms.
- **GCNII** (Chen et al. 2020) augments APPNP with identity (residual) mappings to preserve initial node features and curb over-smoothing.
- **$\mathbf{H_2}$GCN** (Zhu et al. 2020) explicitly separates a node's own representation from that of its neighbors during aggregation.
- **Geom-GCN** (Pei et al. 2020) groups neighbors according to their positions in a learned geometric space before propagation.
- **GPRGNN** (Chien et al. 2020) turns personalized PageRank into a learnable propagation scheme, providing robustness to heterophily and excess smoothing.
- **GloGNN** (Li et al. 2022) introduces global (virtual) nodes that shorten message-passing paths and speed up information mixing.
- **Auto-HeG** (Zheng et al. 2023) automatically searches, trains, and selects heterophilous GNN architectures within a predefined supernet.
- **NSD** (Bodnar et al. 2022) performs neural message passing through learnable sheaf-based diffusion operators.
- **SheafAN** (Barbero et al. 2022) propagates signals with attention-weighted sheaf morphisms that respect higher-order structure.
- **JacobiConv** (Wang and Zhang 2022) analyzes the expressive limits of spectral GNNs via their connection to Jacobi iterations and graph-isomorphism testing.
- **SheafHyper** (Duta et al. 2023) extends sheaf-based filtering to hypergraphs, capturing higher-order relations natively.
- **NLSD** (Zaghen et al. 2024) proposes a null-Lagrangian sheaf diffusion scheme that improves stability.
- **SimCalib** (Tang et al. 2024) calibrates node similarity scores to mitigate heterophily-induced bias in predictions.
- **PCNet** (Li, Pan, and Kang 2024) employs a dual-filter approach that isolates homophilic information even when the underlying graph is heterophilic.