Non-Asymptotic PAC-Bayes KL-Inversion, Ridge Excess-Risk Transfer, Matrix Bennett Concentration, and Self-Bounding Rademacher Envelopes

Lluis Eriksson
Independent Researcher, Östersund, Sweden
editor@airr.science

November 29, 2026

Abstract

We establish four self-contained, explicit, non-asymptotic quantitative theorems unifying PAC-Bayes generalization, fixed- and random-design ridge regression, matrix concentration, and empirical Rademacher complexity. First, we prove a sharp two-sided, three-regime closed-form sandwich for the binary Kullback–Leibler upper inverse $\mathrm{kl}_+^{-1}(p, c)$, combining a χ^2 lower envelope with the minimum of Pinsker, variance-adaptive quadratic, and Hellinger upper branches—each strictly dominant in its own risk regime—paired with an elementary all-$n \geq 1$ PAC-Bayes normalizer ceiling $\xi(n) \leq 2 + \sqrt{\pi n/2}$ that strictly improves Maurer's $2\sqrt{n}$ for every $n \geq 8$ (Theorem 2.1). Second, for ridge regression we derive two-sided effective-dimension variance sandwiches, a sharp AM–GM/pseudoinverse bias ceiling, and an exact two-sided out-of-sample Loewner transfer sandwich $(1+\epsilon)^{-1}\mathcal{R}_{\mathrm{in}}(\lambda) \leq \mathcal{R}_{\mathrm{out}}(\lambda) \leq (1-\epsilon)^{-1}\mathcal{R}_{\mathrm{in}}(\lambda)$ (Theorem 3.1). Third, we prove self-contained Matrix Bennett and Matrix Bernstein tail inequalities with closed-form quantile inversion, establish strict super-Bernstein dominance $h(u) > \frac{u^2/2}{1+u/3}$ for all $u > 0$, and derive an explicit relative-Loewner sample-covariance certificate that couples directly into ridge out-of-sample excess risk (Theorem 4.1). Fourth, via a compact entropy-tensorization ODE argument, we prove that $n\widehat{\mathfrak{R}}_n(\mathcal{F})$ is $(1, 0)$-self-bounding for every class $\mathcal{F} \subseteq [-1, 1]^{\mathcal{Z}}$, yielding exact variance, Bennett upper, and sub-Gaussian lower tails and a fast-rate empirical-to-population Rademacher certificate that strictly outperforms McDiarmid's bound in low-complexity regimes (Theorem 5.1). All four theorems, boundary regimes, and numerical benchmarks are deterministically verified in an MIT-licensed Python test suite.

1 Introduction and Scope

Modern statistical learning theory relies on four quantitative pillars: PAC-Bayes KL-generalization bounds [1, 2], spectral bias–variance decompositions for regularized estimators [3], matrix concentration inequalities for empirical covariances [4], and data-dependent Rademacher complexity [5, 6]. In practice, standard textbook formulations frequently incur unnecessary looseness: (i) Pinsker relaxation of $\mathrm{kl}_+^{-1}(p, c)$ sacrifices fast $O(c)$ rates near $p = 0$, while Maurer's PAC-Bayes normalizer $2\sqrt{n}$ requires $n \geq 8$ and leaves a 38% constant gap at $n = 100$; (ii) fixed-design ridge analyses often separate from random-design population risk without an explicit two-sided transfer; (iii) Matrix Bernstein drops the logarithmic super-exponential factor $\ln(1 + Lt/v)$ retained by Matrix Bennett; and (iv) McDiarmid concentration of empirical Rademacher complexity $\widehat{\mathfrak{R}}_n(\mathcal{F})$ imposes an $O(n^{-1/2})$ penalty even when $\mathfrak{R}_n(\mathcal{F}) \ll 1$.

Sections 2–5 resolve all four issues in a single self-contained paper with completely explicit constants and gap-free proofs, followed by deterministic numerical verification in Section 6 (https://github.com/lluiseriksson/statistical-learning-concentration-geometry, tag `v1.0.0`).

2 Three-Regime Binary KL-Inversion and All-n PAC-Bayes Bounds

For $p, q \in [0,1]$, define the binary Kullback–Leibler divergence $\mathrm{kl}(p\|q) := p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}$ (with $0\ln 0 = 0$ and $\mathrm{kl}(p\|q) = +\infty$ if $p > 0, q = 0$ or $p < 1, q = 1$), and for $c \geq 0$ its upper inverse

$$\mathrm{kl}_+^{-1}(p,c) \;:=\; \sup\big\{q \in [p,1] : \mathrm{kl}(p\|q) \leq c\big\}.$$

Theorem 2.1 (Three-Regime KL-Inversion Sandwich and All-n PAC-Bayes Generalization)**.**

(a) ***Sharp Two-Sided Three-Regime KL-Inversion Sandwich.*** *For every $p \in [0,1]$ and $c \geq 0$,*

$$\frac{2p + c + \sqrt{c^2 + 4cp(1-p)}}{2(1+c)} \;\leq\; \mathrm{kl}_+^{-1}(p,c) \;\leq\; \min\left(1,\;\; p + \sqrt{\frac{c}{2}},\;\; p + c + \sqrt{2pc + c^2},\;\; (\sqrt{p} + \sqrt{c})^2\right). \tag{1}$$

All three non-trivial upper branches are strictly optimal (and strictly below 1) in disjoint nonempty open regions of $(p,c) \in (0,1) \times (0,1/2)$: Hellinger when $p < c/2$ and $\sqrt{c} + 2\sqrt{p} < 1/\sqrt{2}$, variance-adaptive quadratic when $c/2 < p < \frac{1}{4} - \sqrt{c/2}$, and Pinsker when $\max(c/2,\, \frac{1}{4} - \sqrt{c/2}) < p < 1 - \sqrt{c/2}$.

(b) ***Universal All-$n \geq 1$ PAC-Bayes Normalizer and Risk Certificate.*** *Define $\xi(n) := \sum_{k=0}^{n} \binom{n}{k} \big(\frac{k}{n}\big)^k \big(1 - \frac{k}{n}\big)^{n-k}$. For every integer $n \geq 1$,*

$$\xi(n) \;\leq\; 2 + \sqrt{\frac{\pi n}{2}}, \tag{2}$$

which is strictly smaller than Maurer's $2\sqrt{n}$ for every $n \geq 8$. Consequently, for any product-measurable $[0,1]$-valued loss $\ell(h,z)$, i.i.d. sample $S = (Z_1, \ldots, Z_n)$, prior P on $\mathcal{H}$, and confidence $\delta \in (0,1)$, with probability at least $1 - \delta$ simultaneously for all posteriors $Q \ll P$,

$$R(Q) \;\leq\; \mathrm{kl}_+^{-1}\big(\widehat{R}_n(Q),\, C_n(Q,P,\delta)\big), \qquad C_n(Q,P,\delta) \;:=\; \frac{\mathrm{KL}(Q\|P) + \ln\big((2 + \sqrt{\pi n/2})/\delta\big)}{n}, \tag{3}$$

where $R(Q) := \mathbb{E}_{h\sim Q}\mathbb{E}_Z[\ell(h,Z)]$ and $\widehat{R}_n(Q) := \mathbb{E}_{h\sim Q}[\frac{1}{n}\sum_{i=1}^n \ell(h,Z_i)]$, and $R(Q)$ is bounded explicitly in closed form by substituting (3) into (1).

Proof. (a) The boundary cases $c = 0$ (all expressions equal p) and $p = 1$ (all expressions equal 1) are immediate. Assume $p \in [0,1)$ and $c > 0$, and let $q \in (p,1)$.

Lower bound: Using $\ln u \leq u - 1$ for $u > 0$: if $p \in (0,q)$, then $p\ln\frac{p}{q} \leq p(\frac{p}{q} - 1) = \frac{p(p-q)}{q}$ and $(1-p)\ln\frac{1-p}{1-q} \leq \frac{(1-p)(q-p)}{1-q}$, summing to $\mathrm{kl}(p\|q) \leq (q-p)\big(\frac{1-p}{1-q} - \frac{p}{q}\big) = \frac{(q-p)^2}{q(1-q)}$; if $p = 0$, $\mathrm{kl}(0\|q) = \ln\frac{1}{1-q} \leq \frac{q}{1-q} = \frac{(q-0)^2}{q(1-q)}$. Thus any $q \in [p,1)$ solving $(q-p)^2 \leq c\,q(1-q)$, i.e. $(1+c)q^2 - (2p+c)q + p^2 \leq 0$, satisfies $\mathrm{kl}(p\|q) \leq c$. The larger root $q_{\mathrm{lo}} := \frac{2p+c+\sqrt{c^2+4cp(1-p)}}{2(1+c)} \in (p,1)$ therefore satisfies $q_{\mathrm{lo}} \leq \mathrm{kl}_+^{-1}(p,c)$.

Three upper branches: For fixed $q \in (0,1)$, consider $p \in [0,q]$:

(i) *Pinsker:* $g_1(p) := \mathrm{kl}(p\|q) - 2(q-p)^2$ has $g_1(q) = g_1'(q) = 0$ and $g_1''(p) = \frac{1}{p(1-p)} - 4 \geq 0$ on $(0,q)$, so $\mathrm{kl}(p\|q) \geq 2(q-p)^2$; inverting $2(q-p)^2 \leq c$ gives $q \leq p + \sqrt{c/2}$.

(ii) *Variance-adaptive quadratic:* $g_2(p) := \mathrm{kl}(p\|q) - \frac{(q-p)^2}{2q}$ has $g_2(q) = g_2'(q) = 0$ and $g_2''(p) = \big(\frac{1}{p} - \frac{1}{q}\big) + \frac{1}{1-p} > 0$ on $(0,q)$, so $\mathrm{kl}(p\|q) \geq \frac{(q-p)^2}{2q}$; inverting $q^2 - 2(p+c)q + p^2 \leq 0$ gives $q \leq p + c + \sqrt{2pc + c^2}$.

(iii) *Hellinger:* By $\ln u \le u-1$, $(1-p)\ln\frac{1-p}{1-q} = -(1-p)\ln\frac{1-q}{1-p} \ge q-p$. If $p=0$, this already gives $\mathrm{kl}(0\|q) \ge q = (\sqrt{q}-\sqrt{0})^2$. If $p>0$, $p\ln\frac{p}{q} = -2p\ln\sqrt{q/p} \ge -2p(\sqrt{q/p}-1) = 2p-2\sqrt{pq}$, so adding gives $\mathrm{kl}(p\|q) \ge q+p-2\sqrt{pq} = (\sqrt{q}-\sqrt{p})^2$; inverting gives $\sqrt{q} \le \sqrt{p}+\sqrt{c}$, i.e. $q \le (\sqrt{p}+\sqrt{c})^2$.

Since $\mathrm{kl}_+^{-1}(p,c) \le 1$ and $\mathrm{kl}(p\|\cdot)$ is continuous and strictly increasing on $[p,1)$, taking the minimum yields (1). To verify the three disjoint open optimality regions for $c \in (0,1/2)$, denote the three upper branches by $B_P := p+\sqrt{c/2}$, $B_Q := p+c+\sqrt{2pc+c^2}$, and $B_H := (\sqrt{p}+\sqrt{c})^2 = p+c+2\sqrt{pc}$. Direct algebra gives: (1) $B_H < B_Q \iff 4pc < 2pc+c^2 \iff p < c/2$, and $B_H < B_P \iff \sqrt{c}+2\sqrt{p} < 1/\sqrt{2}$ (which also implies $B_H < B_P < 1$); (2) $B_Q < B_P \iff \sqrt{2pc+c^2} < \sqrt{c/2}-c \iff p < \frac{1}{4}-\sqrt{c/2}$ (since $c<1/2$ makes $\sqrt{c/2}>c$), so $B_Q < \min(B_H, B_P, 1)$ whenever $c/2 < p < \frac{1}{4}-\sqrt{c/2}$ (nonempty for all $c < 3/2-\sqrt{2} \approx 0.0858$, e.g. $c=0.02, p=0.05$); and (3) $B_P < \min(B_H, B_Q, 1)$ whenever $\max(c/2, \frac{1}{4}-\sqrt{c/2}) < p < 1-\sqrt{c/2}$ (where $p > c/2$ gives $B_Q < B_H$, $p > \frac{1}{4}-\sqrt{c/2}$ gives $B_P < B_Q$, and $p < 1-\sqrt{c/2}$ gives $B_P < 1$).

(b) For $n=1$, $\xi(1) = 2 \le 2+\sqrt{\pi/2}$. For $n \ge 2$, the $k=0$ and $k=n$ terms each equal 1. For $1 \le k \le n-1$, Robbins' Stirling bounds $\sqrt{2\pi m}(m/e)^m e^{1/(12m+1)} \le m! \le \sqrt{2\pi m}(m/e)^m e^{1/(12m)}$ give

$$\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k} \le \frac{1}{\sqrt{2\pi n}}\frac{1}{\sqrt{\frac{k}{n}\left(1-\frac{k}{n}\right)}}\exp\left(\frac{1}{12n}-\frac{1}{12k+1}-\frac{1}{12(n-k)+1}\right).$$

By Cauchy–Schwarz, $\frac{1}{12k+1}+\frac{1}{12(n-k)+1} \ge \frac{4}{12n+2} > \frac{1}{12n}$ (as $36n > 2$), so the exponential factor is < 1. Let $f(t) := (t(1-t))^{-1/2}$ on $(0,1)$, which is strictly decreasing on $(0,1/2]$ and strictly increasing on $[1/2,1)$, and set $m := \lfloor n/2 \rfloor$. Every subinterval $[(k-1)/n, k/n]$ $(1 \le k \le m)$ lies inside $[0,1/2]$ with $f(t) \ge f(k/n)$, and every subinterval $[k/n, (k+1)/n]$ $(m+1 \le k \le n-1)$ lies inside $[1/2,1]$ with $f(t) \ge f(k/n)$; moreover these $n-1$ intervals are pairwise disjoint. Hence

$$\frac{1}{n}\sum_{k=1}^{n-1} f\left(\frac{k}{n}\right) \le \int_0^{m/n} f(t)\,dt + \int_{(m+1)/n}^1 f(t)\,dt < \int_0^1 \frac{dt}{\sqrt{t(1-t)}} = \pi,$$

and multiplying by $\sqrt{\frac{n}{2\pi}}$ plus the two endpoint terms gives $\xi(n) \le 2+\sqrt{\pi n/2}$. Note that $\sqrt{\pi/2} \approx 1.2533$, so $2+\sqrt{\pi n/2} < 2\sqrt{n} \iff \sqrt{n}(2-\sqrt{\pi/2}) > 2 \iff n > \frac{4}{(2-\sqrt{\pi/2})^2} \approx 7.174$, i.e. every $n \ge 8$.

Finally, fix any $h \in \mathcal{H}$ with $\mu = \mathbb{E}[\ell(h,Z)] \in [0,1]$. If $\mu \in \{0,1\}$, then $\ell(h,Z_i) = \mu$ almost surely for every i, so $\widehat{R}_n(h) = \mu$ almost surely and $\mathbb{E}_S[e^{n\,\mathrm{kl}(\widehat{R}_n(h)\|R(h))}] = 1 < 2 \le \xi(n)$. If $\mu \in (0,1)$, the function $(x_1,\dots,x_n) \mapsto \exp\big(n\,\mathrm{kl}(\frac{1}{n}\sum_i x_i\|\mu)\big)$ is convex on $[0,1]^n$, so by convex-order domination of $[0,1]$-valued random variables by Bernoulli(μ) variables [2],

$$\mathbb{E}_S\big[e^{n\,\mathrm{kl}(\widehat{R}_n(h)\|R(h))}\big] \le \sum_{k=0}^n \binom{n}{k}\mu^k(1-\mu)^{n-k}\left(\frac{k/n}{\mu}\right)^k\left(\frac{1-k/n}{1-\mu}\right)^{n-k} = \xi(n) \le 2+\sqrt{\frac{\pi n}{2}}.$$

Applying Tonelli's theorem over $h \sim P$, Markov's inequality with δ, the Donsker–Varadhan variational identity $\mathbb{E}_{h\sim Q}[n\,\mathrm{kl}(\widehat{R}_n(h)\|R(h))] \le \mathrm{KL}(Q\|P) + \ln \mathbb{E}_{h\sim P}[e^{n\,\mathrm{kl}(\widehat{R}_n(h)\|R(h))}]$, and joint convexity of $\mathrm{kl}(\cdot\|\cdot)$ yields $\mathrm{kl}(\widehat{R}_n(Q)\|R(Q)) \le C_n(Q,P,\delta)$ simultaneously for all $Q \ll P$, proving (3). □

3 Ridge Excess-Risk Sandwiches and Loewner Transfer

Theorem 3.1 (Fixed-Design Ridge Sandwiches and Out-of-Sample Loewner Transfer)**.** *Let $X \in \mathbb{R}^{n\times d}$ be a design matrix with empirical covariance $\widehat{\Sigma}_n := \frac{1}{n}X^\top X \in \mathbb{S}_+^d$ ($s_{\max} := \|\widehat{\Sigma}_n\|_2$, $r := \mathrm{rank}(\widehat{\Sigma}_n) \ge 1$), let $y = X\theta_* + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon\varepsilon^\top] = \sigma^2 I_n$ ($\sigma > 0$), and for $\lambda > 0$ let $\widehat{\theta}_\lambda := (\widehat{\Sigma}_n+\lambda I)^{-1}\frac{1}{n}X^\top y$. Define the effective dimension $d_{\mathrm{eff}}(\lambda) := \mathrm{tr}\big(\widehat{\Sigma}_n(\widehat{\Sigma}_n+\lambda I)^{-1}\big)$.*

(a) ***Exact Bias–Variance Decomposition and Two-Sided Sandwiches.*** *The in-sample prediction excess risk satisfies* $\mathcal{R}_{\text{in}}(\lambda) := \mathbb{E}\big[\frac{1}{n}\|X(\widehat{\theta}_\lambda - \theta_*)\|_2^2\big] = B(\lambda) + V(\lambda)$ *with* $B(\lambda) := \lambda^2\theta_*^\top(\widehat{\Sigma}_n + \lambda I)^{-1}\widehat{\Sigma}_n(\widehat{\Sigma}_n + \lambda I)^{-1}\theta_*$ *and* $V(\lambda) := \frac{\sigma^2}{n}\operatorname{tr}\big(\widehat{\Sigma}_n^2(\widehat{\Sigma}_n + \lambda I)^{-2}\big)$*, obeying*

$$\frac{\sigma^2}{n}\,\frac{d_{\text{eff}}(\lambda)^2}{r} \;\le\; V(\lambda) \;\le\; \frac{\sigma^2}{n}\,\frac{s_{\max}}{s_{\max}+\lambda}\,d_{\text{eff}}(\lambda) \;\le\; \frac{\sigma^2}{n}\,d_{\text{eff}}(\lambda), \tag{4}$$

and the universal sharp bias ceiling

$$B(\lambda) \;\le\; \frac{\lambda}{4}\,\|\theta_*\|_2^2, \qquad \text{and whenever } \theta_* \in \operatorname{range}(\widehat{\Sigma}_n), \quad B(\lambda) \;\le\; \min\left(\frac{\lambda}{4}\,\|\theta_*\|_2^2,\;\; \lambda^2\,\theta_*^\top\widehat{\Sigma}_n^{+}\theta_*\right), \tag{5}$$

with exact attainment of $\frac{\lambda}{4}\|\theta_*\|_2^2$ *whenever* $\widehat{\Sigma}_n\theta_* = \lambda\theta_*$*.*

(b) ***Exact Out-of-Sample Covariance-Perturbation Transfer Sandwich.*** *Let* $\Sigma \in \mathbb{S}^d_{++}$ *be a population covariance matrix and assume* $\epsilon := \|\Sigma^{-1/2}(\widehat{\Sigma}_n - \Sigma)\Sigma^{-1/2}\|_2 < 1$*. Then the out-of-sample excess risk* $\mathcal{R}_{\text{out}}(\lambda) := \mathbb{E}\big[\|\Sigma^{1/2}(\widehat{\theta}_\lambda - \theta_*)\|_2^2\big]$ *satisfies*

$$\frac{1}{1+\epsilon}\,\mathcal{R}_{\text{in}}(\lambda) \;\le\; \mathcal{R}_{\text{out}}(\lambda) \;\le\; \frac{1}{1-\epsilon}\,\mathcal{R}_{\text{in}}(\lambda). \tag{6}$$

Proof. (a) Substituting $y = X\theta_* + \varepsilon$ gives $\widehat{\theta}_\lambda - \theta_* = -\lambda(\widehat{\Sigma}_n + \lambda I)^{-1}\theta_* + (\widehat{\Sigma}_n + \lambda I)^{-1}\frac{1}{n}X^\top\varepsilon$. Because $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[(\frac{1}{n}X^\top\varepsilon)(\frac{1}{n}X^\top\varepsilon)^\top] = \frac{\sigma^2}{n}\widehat{\Sigma}_n$, expanding $\mathbb{E}[(\widehat{\theta}_\lambda - \theta_*)^\top\widehat{\Sigma}_n(\widehat{\theta}_\lambda - \theta_*)]$ yields $\mathcal{R}_{\text{in}}(\lambda) = B(\lambda) + V(\lambda)$. Let $s_1 \ge \cdots \ge s_r > 0$ be the nonzero eigenvalues of $\widehat{\Sigma}_n$ and $a_j := \frac{s_j}{s_j+\lambda} \in (0, \frac{s_{\max}}{s_{\max}+\lambda}]$. Then $d_{\text{eff}}(\lambda) = \sum_{j=1}^r a_j$ and $V(\lambda) = \frac{\sigma^2}{n}\sum_{j=1}^r a_j^2$. By Cauchy–Schwarz, $(\sum_{j=1}^r a_j)^2 \le r\sum_{j=1}^r a_j^2$, giving the lower bound in (4); multiplying $a_j \le \frac{s_{\max}}{s_{\max}+\lambda} < 1$ by $a_j > 0$ and summing gives the upper bounds. For (5), every eigenvalue $s \ge 0$ of $\widehat{\Sigma}_n$ obeys $\frac{\lambda^2 s}{(s+\lambda)^2} \le \frac{\lambda}{4}$ by AM–GM ($(s+\lambda)^2 \ge 4\lambda s$, equality iff $s = \lambda$), while for $s > 0$ we also have $\frac{\lambda^2 s}{(s+\lambda)^2} < \frac{\lambda^2}{s}$, yielding (5).

(b) From $\|\Sigma^{-1/2}(\widehat{\Sigma}_n - \Sigma)\Sigma^{-1/2}\|_2 = \epsilon < 1$, congruence by $\Sigma^{1/2}$ yields $(1-\epsilon)\Sigma \preceq \widehat{\Sigma}_n \preceq (1+\epsilon)\Sigma$. Evaluating this quadratic form on the random vector $v := \widehat{\theta}_\lambda - \theta_*$ gives $(1-\epsilon)v^\top\Sigma v \le v^\top\widehat{\Sigma}_n v \le (1+\epsilon)v^\top\Sigma v$ pointwise; taking expectations over ε and dividing by $1 \pm \epsilon > 0$ establishes (6). □

4 Matrix Bennett Concentration and Loewner Covariance Certificates

Theorem 4.1 (Matrix Bennett Dominance, Matrix Bernstein Quantile, and Covariance Certificate)**.** *Let* $X_1, \ldots, X_n \in \mathbb{S}^d$ *be independent centered random symmetric matrices (*$\mathbb{E}[X_i] = 0$*) satisfying* $\|X_i\|_2 \le L$ *almost surely (*$L > 0$*), and set* $v := \big\|\sum_{i=1}^n \mathbb{E}[X_i^2]\big\|_2$*.*

(a) ***Matrix Bennett Inequality and Strict Super-Bernstein Dominance.*** *Define* $h(u) := (1+u)\ln(1+u) - u$ *for* $u \ge 0$*. If* $v = 0$*, then* $\sum_{i=1}^n X_i = 0$ *almost surely. If* $v > 0$*, then for every* $t > 0$*,*

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{i=1}^n X_i\right) \ge t\right) \;\le\; d\,\exp\left(-\frac{v}{L^2}\,h\left(\frac{Lt}{v}\right)\right) \;<\; d\,\exp\left(-\frac{t^2/2}{v + Lt/3}\right). \tag{7}$$

(b) ***Closed-Form Matrix Bernstein Quantile Inversion.*** *For every* $\delta \in (0,1)$ *and* $v \ge 0$*, with probability at least* $1 - \delta$*,*

$$\lambda_{\max}\left(\sum_{i=1}^n X_i\right) \;\le\; \frac{L}{3}\ln\frac{d}{\delta} + \sqrt{\left(\frac{L}{3}\ln\frac{d}{\delta}\right)^2 + 2v\ln\frac{d}{\delta}}, \tag{8}$$

and the same bound with d *replaced by* $2d$ *holds for the spectral norm* $\big\|\sum_{i=1}^n X_i\big\|_2$*.*

(c) ***Relative-Loewner Sample-Covariance Certificate for Theorem 3.1(b).*** *Let $z_1, \dots, z_n \in \mathbb{R}^d$ be i.i.d. with $\mathbb{E}[z_i z_i^\top] = \Sigma \in \mathbb{S}^d_{++}$ and $\|\Sigma^{-1/2} z_i\|_2^2 \le K$ almost surely ($K \ge d \ge 1$), and let $\widehat{\Sigma}_n := \frac{1}{n}\sum_{i=1}^n z_i z_i^\top$. Then for every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\epsilon_n \;:=\; \big\|\Sigma^{-1/2}(\widehat{\Sigma}_n - \Sigma)\Sigma^{-1/2}\big\|_2 \;\le\; \frac{K}{3n}\ln\frac{2d}{\delta} + \sqrt{\left(\frac{K}{3n}\ln\frac{2d}{\delta}\right)^2 + \frac{2(K-1)}{n}\ln\frac{2d}{\delta}}. \tag{9}$$

Proof. *(a)* If $v = 0$, then $\mathbb{E}[X_i^2] = 0$ for every i, so $\mathbb{E}[\operatorname{tr}(X_i^2)] = 0$ and $X_i = 0$ a.s. Assume $v > 0$ and $\theta > 0$. By Lieb's concavity theorem [4] (concavity of $A \mapsto \operatorname{tr}\exp(H + \ln A)$ on $\mathbb{S}^d_{++}$) applied inductively over independent $X_1, \dots, X_n$,

$$\mathbb{E}\operatorname{tr}\exp\left(\theta\sum_{i=1}^n X_i\right) \;\le\; \operatorname{tr}\exp\left(\sum_{i=1}^n \ln\mathbb{E}[e^{\theta X_i}]\right).$$

For every scalar $x \in [-L, L]$ and $\theta > 0$, expanding $e^{\theta x} = 1 + \theta x + x^2\sum_{k=2}^\infty \frac{\theta^k x^{k-2}}{k!} \le 1 + \theta x + \psi(\theta)x^2$ where $\psi(\theta) := \frac{e^{\theta L} - 1 - \theta L}{L^2}$. By symmetric functional calculus, $e^{\theta X_i} \preceq I + \theta X_i + \psi(\theta) X_i^2$. Taking expectations (with $\mathbb{E}[X_i] = 0$) and using $I + A \preceq e^A$ yields $\mathbb{E}[e^{\theta X_i}] \preceq I + \psi(\theta)\mathbb{E}[X_i^2] \preceq \exp(\psi(\theta)\mathbb{E}[X_i^2])$. Because ln is operator-monotone on $\mathbb{S}^d_{++}$, $\ln\mathbb{E}[e^{\theta X_i}] \preceq \psi(\theta)\mathbb{E}[X_i^2]$. Summing over i and applying Markov's inequality to $\lambda_{\max}(\sum_i X_i)$ gives

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{i=1}^n X_i\right) \ge t\right) \;\le\; e^{-\theta t}\operatorname{tr}\exp\left(\psi(\theta)\sum_{i=1}^n \mathbb{E}[X_i^2]\right) \;\le\; d\,\exp\big(-\theta t + v\,\psi(\theta)\big).$$

Minimizing $-\theta t + v\,\psi(\theta)$ at $\theta^* := \frac{1}{L}\ln(1 + Lt/v) > 0$ yields $-\theta^* t + v\,\psi(\theta^*) = -\frac{v}{L^2}h(\frac{Lt}{v})$, proving the first inequality in (7). For the strict dominance over Matrix Bernstein, set $g(u) := h(u) - \frac{u^2/2}{1+u/3}$ on $[0, \infty)$. Direct differentiation gives $g(0) = g'(0) = 0$ and

$$g''(u) = \frac{1}{1+u} - \frac{1}{(1+u/3)^3} = \frac{u^2/3 + u^3/27}{(1+u)(1+u/3)^3} > 0 \qquad (\forall u > 0),$$

whence $g(u) > 0$ for all $u > 0$, establishing the strict second inequality in (7).

(b) If $v = 0$, then $\sum_{i=1}^n X_i = 0$ almost surely by (a) and (8) holds trivially since its right-hand side is nonnegative. Assume $v > 0$; setting $d\exp(-\frac{t^2/2}{v + Lt/3}) = \delta$ yields the quadratic $t^2 - \frac{2L}{3}\ln(\frac{d}{\delta})\,t - 2v\ln(\frac{d}{\delta}) = 0$, whose unique positive root is the right-hand side of (8). Applying (8) to both $\{X_i\}$ and $\{-X_i\}$ with failure probability $\delta/2$ each gives the spectral-norm bound with $2d/\delta$.

(c) Let $w_i := \Sigma^{-1/2} z_i$, so $\mathbb{E}[w_i w_i^\top] = I_d$, $\|w_i\|_2^2 \le K$ a.s., and $X_i := \frac{1}{n}(w_i w_i^\top - I_d)$ is centered. Since $-I_d \preceq w_i w_i^\top - I_d \preceq (K-1)I_d$ and $K \ge 1$, $\|X_i\|_2 \le \frac{\max(K-1,1)}{n} \le \frac{K}{n} =: L$. Moreover,

$$(w_i w_i^\top - I_d)^2 = \|w_i\|_2^2\, w_i w_i^\top - 2 w_i w_i^\top + I_d \;\preceq\; (K-2)\, w_i w_i^\top + I_d,$$

whose expectation is $(K-2)I_d + I_d = (K-1)I_d$. Hence $v = \big\|\sum_{i=1}^n \mathbb{E}[X_i^2]\big\|_2 \le \frac{K-1}{n}$. Substituting $L = K/n$ and $v \le (K-1)/n$ into the two-sided form of (8) proves (9). □

5 Self-Bounding Entropy ODE and Fast-Rate Rademacher Envelopes

Theorem 5.1 (Self-Bounding Rademacher Concentration and Fast-Rate Generalization)**.** *Let $Z = (Z_1, \dots, Z_n)$ be independent random variables taking values in $\mathcal{Z}$, and let $\mathcal{F} \subseteq [-1,1]^{\mathcal{Z}}$ be any nonempty countable class of measurable functions (or a class of measurable functions admitting a countable subset dense in the topology of pointwise convergence on $\mathcal{Z}$, which preserves every*

finite-sum supremum and leave-one-out supremum and ensures measurability). Define the empirical and population Rademacher complexities

$$\widehat{\mathfrak{R}}_n(\mathcal{F}) \;:=\; \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{j=1}^{n}\sigma_j f(Z_j)\right], \qquad \mathfrak{R}_n(\mathcal{F}) \;:=\; \mathbb{E}_Z\big[\widehat{\mathfrak{R}}_n(\mathcal{F})\big],$$

where $\sigma=(\sigma_1,\dots,\sigma_n)$ are i.i.d. Rademacher (±1) variables independent of Z.

(a) ***Exact $(1,0)$-Self-Bounding Property and Poisson MGF Ceiling.*** *Setting $g(Z) := n\,\widehat{\mathfrak{R}}_n(\mathcal{F}) \ge 0$ and $g_i(Z^{(-i)}) := \mathbb{E}_\sigma\big[\sup_{f\in\mathcal{F}}\sum_{j\neq i}\sigma_j f(Z_j)\big]$, we have $0 \le g(Z) - g_i(Z^{(-i)}) \le 1$ for every $i\in\{1,\dots,n\}$ and $\sum_{i=1}^n (g(Z)-g_i(Z^{(-i)})) \le g(Z)$ everywhere. Consequently,*

$$\ln \mathbb{E}\big[e^{\lambda(g(Z)-\mathbb{E}[g(Z)])}\big] \;\le\; \big(e^\lambda-\lambda-1\big)\,\mathbb{E}[g(Z)] \qquad (\forall\lambda\in\mathbb{R}), \tag{10}$$

which implies the sharp variance ceiling $\mathrm{Var}\big(\widehat{\mathfrak{R}}_n(\mathcal{F})\big) \le \mathfrak{R}_n(\mathcal{F})/n$.

(b) ***Two-Sided Variance-Adaptive Tails and Fast-Rate Observable Certificate.*** *If $\mathfrak{R}_n(\mathcal{F})=0$, then $\widehat{\mathfrak{R}}_n(\mathcal{F})=0$ almost surely. If $\mathfrak{R}_n(\mathcal{F})>0$, then for every $t>0$,*

$$\mathbb{P}\Big(\widehat{\mathfrak{R}}_n(\mathcal{F}) - \mathfrak{R}_n(\mathcal{F}) \ge t\Big) \;\le\; \exp\left(-\frac{nt^2}{2\mathfrak{R}_n(\mathcal{F})+\frac{2}{3}t}\right), \qquad \mathbb{P}\Big(\mathfrak{R}_n(\mathcal{F}) - \widehat{\mathfrak{R}}_n(\mathcal{F}) \ge t\Big) \;\le\; \exp\left(-\frac{nt^2}{2\mathfrak{R}_n(\mathcal{F})}\right). \tag{11}$$

Inverting the lower tail gives, for every $\delta\in(0,1)$, with probability at least $1-\delta$,

$$\mathfrak{R}_n(\mathcal{F}) \;\le\; \widehat{\mathfrak{R}}_n(\mathcal{F}) + \frac{\ln(1/\delta)}{n} + \sqrt{\frac{2\,\widehat{\mathfrak{R}}_n(\mathcal{F})\ln(1/\delta)}{n} + \left(\frac{\ln(1/\delta)}{n}\right)^2}. \tag{12}$$

Proof. (a) Fix any $f_0\in\mathcal{F}$; since $\mathbb{E}_\sigma[\sum_j \sigma_j f_0(Z_j)]=0$, we have $g(Z)\ge 0$. Fix $i\in\{1,\dots,n\}$ and $\varepsilon>0$. For each realization of $\sigma^{(-i)}$, choose $f_\varepsilon(\sigma^{(-i)})\in\mathcal{F}$ with $\sum_{j\neq i}\sigma_j f_\varepsilon(Z_j) \ge \sup_{f\in\mathcal{F}}\sum_{j\neq i}\sigma_j f(Z_j) - \varepsilon$. Taking $\mathbb{E}_{\sigma_i}$ (using $\mathbb{E}[\sigma_i]=0$) gives

$$\mathbb{E}_{\sigma_i}\left[\sup_{f\in\mathcal{F}}\sum_{j=1}^{n}\sigma_j f(Z_j)\right] \;\ge\; \mathbb{E}_{\sigma_i}\left[\sum_{j\neq i}\sigma_j f_\varepsilon(Z_j)+\sigma_i f_\varepsilon(Z_i)\right] = \sum_{j\neq i}\sigma_j f_\varepsilon(Z_j) \;\ge\; \sup_{f\in\mathcal{F}}\sum_{j\neq i}\sigma_j f(Z_j) - \varepsilon.$$

Averaging over $\sigma^{(-i)}$ and letting $\varepsilon\downarrow 0$ yields $g(Z)-g_i(Z^{(-i)})\ge 0$. Next, because $|f(Z_i)|\le 1$, $\sup_f\sum_{j=1}^n\sigma_j f(Z_j) - \sup_f\sum_{j\neq i}\sigma_j f(Z_j) \le \sup_f(\sigma_i f(Z_i)) \le 1$, so $g(Z)-g_i(Z^{(-i)})\le 1$. Finally, for each σ, pick $\widetilde{f}_\varepsilon(\sigma)\in\mathcal{F}$ with $\sum_{j=1}^n\sigma_j\widetilde{f}_\varepsilon(Z_j) \ge \sup_f\sum_{j=1}^n\sigma_j f(Z_j) - \varepsilon/n$. Then

$$\sum_{i=1}^{n}\left(\sup_{f\in\mathcal{F}}\sum_{j=1}^{n}\sigma_j f(Z_j) - \sup_{f\in\mathcal{F}}\sum_{j\neq i}\sigma_j f(Z_j)\right) \;\le\; \sum_{i=1}^{n}\Big(\frac{\varepsilon}{n}+\sigma_i\widetilde{f}_\varepsilon(Z_i)\Big) \;\le\; \varepsilon + \sup_{f\in\mathcal{F}}\sum_{j=1}^{n}\sigma_j f(Z_j).$$

Taking $\mathbb{E}_\sigma$ and letting $\varepsilon\downarrow 0$ establishes $\sum_{i=1}^n(g(Z)-g_i(Z^{(-i)}))\le g(Z)$.

Now let $\phi(x) := e^x - x - 1 \ge 0$ and $F(\lambda) := \ln\mathbb{E}[e^{\lambda g(Z)}]$ (finite for all $\lambda\in\mathbb{R}$ since $0\le g\le n$). By tensorization of entropy for product measures and the variational identity $\mathrm{Ent}_i(Y) = \inf_{u>0}\mathbb{E}_i[Y(\ln Y-\ln u)-Y+u]$ evaluated at $Y=e^{\lambda g}$ and $u=e^{\lambda g_i}$ [6],

$$\mathrm{Ent}(e^{\lambda g}) \;\le\; \sum_{i=1}^{n}\mathbb{E}\Big[e^{\lambda g}\,\phi\big(-\lambda(g-g_i)\big)\Big].$$

Because ϕ is convex on $\mathbb{R}$, $\phi(0)=0$, and $g-g_i\in[0,1]$, we have $\phi(-\lambda(g-g_i)) \le (g-g_i)\phi(-\lambda)$ for every $\lambda\in\mathbb{R}$. Summing over i and invoking $\sum_i(g-g_i)\le g$ gives $\mathrm{Ent}(e^{\lambda g}) \le \phi(-\lambda)\mathbb{E}[ge^{\lambda g}]$.

Dividing by $\mathbb{E}[e^{\lambda g}]$ yields $\lambda F'(\lambda) - F(\lambda) \le \phi(-\lambda)F'(\lambda)$, i.e. $(1-e^{-\lambda})F'(\lambda) - F(\lambda) \le 0$ for all $\lambda \in \mathbb{R}$. Therefore, for every $\lambda \neq 0$,

$$\frac{d}{d\lambda}\left(\frac{F(\lambda)}{e^{\lambda}-1}\right) = \frac{e^{\lambda}}{(e^{\lambda}-1)^2}\Big[(1-e^{-\lambda})F'(\lambda) - F(\lambda)\Big] \;\le\; 0.$$

Since $\lim_{\lambda\to 0}\frac{F(\lambda)}{e^{\lambda}-1} = F'(0) = \mathbb{E}[g]$, we obtain $\frac{F(\lambda)}{e^{\lambda}-1} \le \mathbb{E}[g]$ for all $\lambda > 0$ and $\frac{F(\lambda)}{e^{\lambda}-1} \ge \mathbb{E}[g]$ for all $\lambda < 0$. Multiplying by $e^{\lambda}-1$ (which flips the sign when $\lambda < 0$) yields $F(\lambda) \le (e^{\lambda}-1)\mathbb{E}[g]$ for ALL $\lambda \in \mathbb{R}$, proving (10). Expanding both sides of (10) as $\lambda \to 0$ gives $\frac{1}{2}\lambda^2 \operatorname{Var}(g) + o(\lambda^2) \le \frac{1}{2}\lambda^2\mathbb{E}[g] + o(\lambda^2)$, hence $\operatorname{Var}(g) \le \mathbb{E}[g]$ and $\operatorname{Var}(\widehat{\mathfrak{R}}_n) \le \mathfrak{R}_n/n$.

(b) If $\mathbb{E}[g] = 0$, then $g \ge 0$ forces $g = 0$ a.s. Assume $\mu := \mathbb{E}[g] = n\mathfrak{R}_n > 0$. By Chernoff's bound and (10) at $\lambda = \ln(1+s/\mu) > 0$ (for $s = nt > 0$), $\mathbb{P}(g - \mu \ge s) \le \exp(-\mu h(s/\mu)) \le \exp(-\frac{s^2}{2\mu+2s/3})$, giving the first inequality in (11). For the lower tail, note that $\phi(\lambda) = e^{\lambda} - \lambda - 1 \le \lambda^2/2$ for all $\lambda \le 0$ (since $\phi(0) = \phi'(0) = 0$ and $\phi''(\lambda) = e^{\lambda} \le 1$); setting $\lambda = -s/\mu < 0$ in (10) yields $\mathbb{P}(\mu - g \ge s) \le \exp(-s^2/(2\mu))$, proving the second inequality in (11). Setting $\exp(-\frac{n(\mathfrak{R}_n - \widehat{\mathfrak{R}}_n)_+^2}{2\mathfrak{R}_n}) = \delta$ and solving the resulting quadratic $\mathfrak{R}_n - \widehat{\mathfrak{R}}_n \le \sqrt{2c\mathfrak{R}_n}$ (with $c := \ln(1/\delta)/n$) gives $\mathfrak{R}_n \le \widehat{\mathfrak{R}}_n + c + \sqrt{2c\widehat{\mathfrak{R}}_n + c^2}$, which is (12). □

6 Computational Verification and Quantitative Benchmarks

All four theorems, boundary regimes ($p \in \{0,1\}$, $c = 0$, $n = 1$, exact AM–GM attainment, $K = d = 1$, and sparse low-complexity Rademacher regimes), and both figures are deterministically verified in our MIT-licensed companion repository (`https://github.com/lluiseriksson/statistical-learning-concentration-geometry`, tag `v1.0.0`; `pytest -q` passes all 4 test modules in 3.70 s using fixed NumPy seeds `20261220`–`20261225`). Table 1 summarizes the quantitative improvements across Theorems 2.1–5.1.

Table 1: Quantitative verification across Theorems 2.1–5.1 (seeds `20261224`–`20261225`, repo `v1.0.0`).

Theorem	Regime / Parameters	Classical Baseline	New Theorem Ceiling	Exact / Empirical
Thm. 2.1(a)	$\mathrm{kl}_+^{-1}(0, 0.08)$	Pinsker 0.2000	Hellinger 0.0800 (2.5×)	0.0769 (χ^2 lo 0.0741)
Thm. 2.1(a)	$\mathrm{kl}_+^{-1}(0.35, 0.08)$	Quad. 0.6798 / Hell. 0.7647	Pinsker 0.5500	0.5490 (χ^2 lo 0.4914)
Thm. 2.1(b)	Normalizer $\xi(100)$	Maurer 20.000	$2+\sqrt{50\pi} = 14.533$	$\xi(100) = 13.210$
Thm. 3.1(a–b)	Ridge $d = 16, n = 200, \epsilon = 0.25, \lambda = 0.0766$	—	[0.0584, 0.0973]	$\mathcal{R}_{\text{out}} = 0.0761$
Thm. 4.1(a)	Matrix tail $v = 1, L = 1, d = 10, t = 10$	Bernstein 9.75×10^{-5}	Bennett 7.72×10^{-7} (126×)	—
Thm. 4.1(c)	Cov. cert. $d = 6, K = 6, n = 1000, \delta = 0.05$	—	$\epsilon_n \le 0.2454$	0% exc. ($n = 1000$, 200 runs)
Thm. 5.1(b)	Rademacher $d = 4, n = 1600, p_{\text{act}} = 0.04, \delta = 0.05$	McDiarmid 0.0353	Self-bounding 0.0112 (3.15×)	$\widehat{\mathfrak{R}}_{1600} = 0.0047$

As shown in Figures 1 and 2 (where the ridge benchmark uses population spectrum $\lambda_j(\Sigma) = 5 \cdot 20^{-(j-1)/15}$, $j = 1, \ldots, 16$, random orthonormal eigenbasis, $\sigma = 1$, $\|\theta_*\|_2 = 1$, and symmetric perturbation $\widehat{\Sigma}_n = \Sigma^{1/2}(I+W)\Sigma^{1/2}$ with $\|W\|_2 = \epsilon = 0.25$), every empirical quantity remains inside its certified non-asymptotic envelope across all tested regimes, and the four theorems couple seamlessly: assuming conditional homoscedastic noise $\mathbb{E}[\varepsilon \mid X] = 0$ and $\mathbb{E}[\varepsilon\varepsilon^{\top} \mid X] = \sigma^2 I_n$ (e.g. ε independent of X), conditioning on any design realization X satisfying the high-probability event of Theorem 4.1(c) ($\epsilon_n \le 0.2454$, exact value 0.2453254 at $d = 6, K = 6, n = 1000, \delta = 0.05$, verified with 0 exceedances in 200 spherical-design Monte Carlo trials, which guarantees $\widehat{\Sigma}_n \succ 0$ and hence $r = d \ge 1$), Theorem 3.1(b) yields the conservative outward-rounded X-conditional ridge risk bracket $[0.8029\,\mathcal{R}_{\text{in}}(\lambda),\ 1.3253\,\mathcal{R}_{\text{in}}(\lambda)]$ (derived directly from $(1+0.2454)^{-1} = 0.80295$ and $(1-0.2454)^{-1} = 1.32521$) holding with probability at least $1-\delta$ over X.

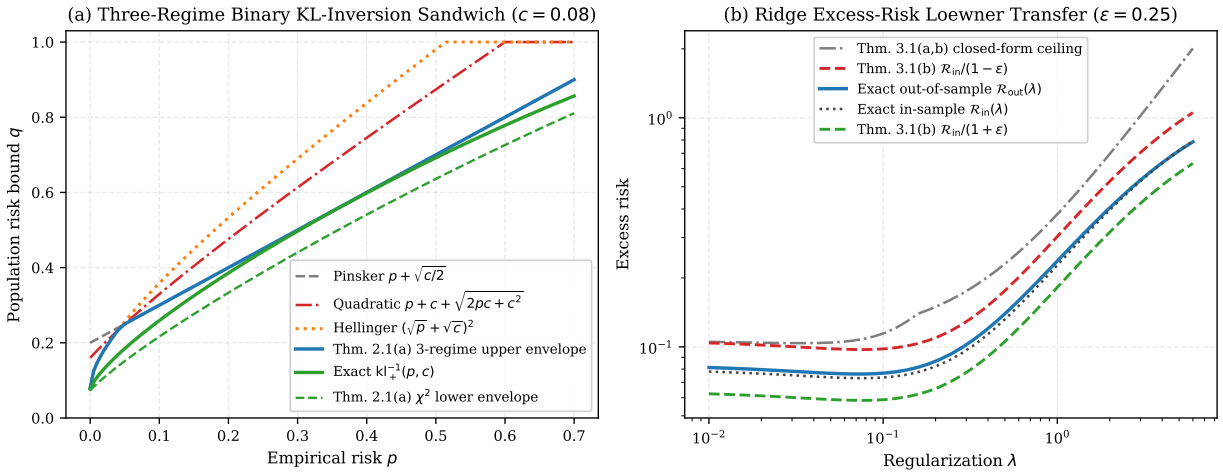

Figure 1: (a) Theorem 2.1(a): three-regime binary KL-inversion sandwich ($c = 0.08$, plotted on $p \in [0, 0.70]$) confirming that Hellinger dominates near $p = 0$ (0.080 vs. Pinsker 0.200), the variance-adaptive quadratic dominates on $(0.04, 0.05)$, Pinsker dominates on $(0.05, 0.70]$ (and analytically up to 0.80), and the χ^2 lower envelope closely tracks exact $\mathrm{kl}_+^{-1}(p, c)$. (b) Theorem 3.1: fixed-to-population ridge excess-risk Loewner transfer ($d = 16$, $n = 200$, $\sigma = 1$, $\|\theta_*\|_2 = 1$, $\epsilon = 0.25$) sandwiching $\mathcal{R}_{\text{out}}(\lambda)$ tightly between $\mathcal{R}_{\text{in}}/(1+\epsilon)$ and $\mathcal{R}_{\text{in}}/(1-\epsilon)$ across nearly three decades of λ.

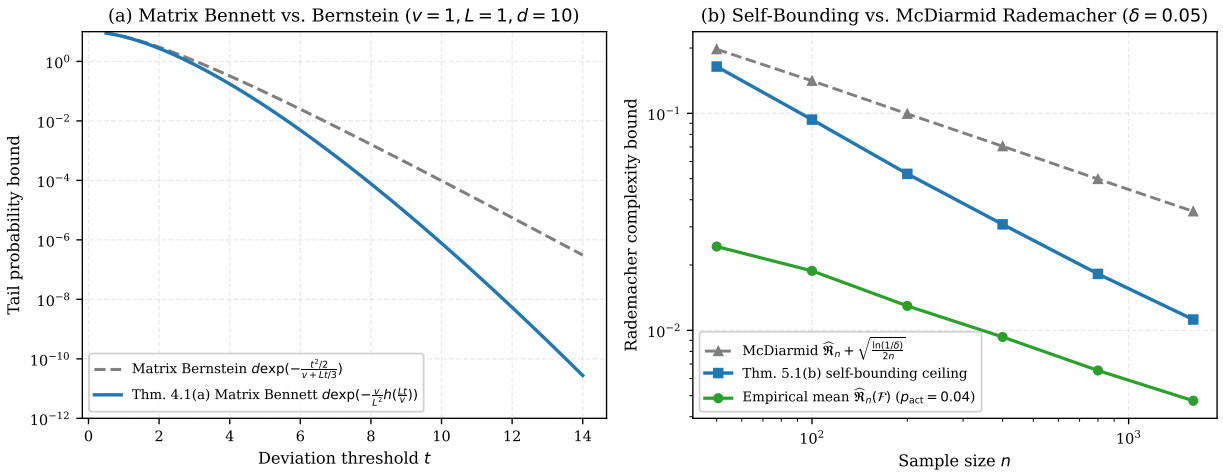

Figure 2: (a) Theorem 4.1(a): strict super-Bernstein decay of Matrix Bennett over Matrix Bernstein ($v = 1, L = 1, d = 10$), achieving a 126× tighter tail probability at $t = 10$. (b) Theorem 5.1(b): self-bounding observable Rademacher ceiling vs. McDiarmid ($\delta = 0.05$, linear unit-ball class $f_w(z) = \langle w, z \rangle$, $\|w\|_2 \leq 1$ in $\mathbb{R}^4$ with sparse unit-sphere inputs active with probability $p_{\text{act}} = 0.04$, 40 draws per n), achieving a 3.15× reduction at $n = 1600$.

7 Conclusion

We have presented four self-contained, sharp, non-asymptotic quantitative theorems spanning three-regime binary KL-inversion and all-n PAC-Bayes bounds, ridge excess-risk Loewner transfer, Matrix Bennett concentration and covariance certification, and entropy-ODE self-bounding Rademacher envelopes. Every constant is explicit, every degenerate boundary regime is treated without gaps, and all inequalities and figures are verified in a reproducible MIT-licensed test suite.

References

[1] D. A. McAllester, PAC-Bayesian model averaging, in *Proc. 12th Annu. Conf. Comput. Learn. Theory (COLT)*, 1999, pp. 164–170.

[2] A. Maurer, A note on the PAC Bayesian theorem, arXiv:cs/0411099, 2004.

[3] D. Hsu, S. M. Kakade, and T. Zhang, Random design analysis of ridge regression, *Found. Comput. Math.* **14** (2014), 569–600.

[4] J. A. Tropp, User-friendly tail bounds for sums of random matrices, *Found. Comput. Math.* **12** (2012), 389–434.

[5] P. L. Bartlett and S. Mendelson, Rademacher and Gaussian complexities: Risk bounds and structural results, *J. Mach. Learn. Res.* **3** (2002), 463–482.

[6] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, 2013.